Virtual University

Cell Biology

Fridoon Jawad Ahmad
BS& PhD (UW-Madison, USA)
Professor of Biomedical Sciences,

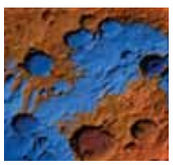
King Edward Medical University,
Lahore, Pakistan.

The majority of the contents of this document were taken from the following two sources.

Molecular Biology of the Cell. 4th edition.
Alberts B, Johnson A, Lewis J, et al.
New York: Garland Science; 2002.

Life: The Science of Biology, Seventh Edition,
by William K. Purves, David Sadava, Gordon
H. Orians, and H. Craig Heller.

1 Life and Chemistry: Small Molecules

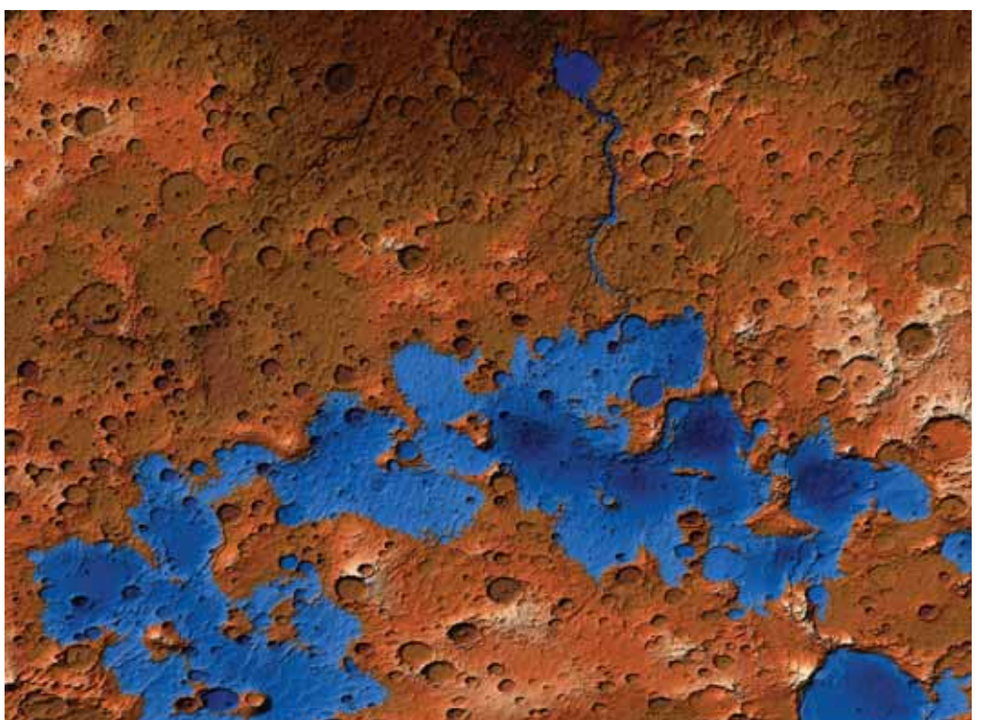


Mars today is a cold, dry place, not suitable for life as we know it. But 3 billion years ago, it was warmer and wetter. An orbiting probe from Earth recently photographed a huge dry lake bed, the size of New Mexico and Texas combined, on the Martian surface. Another probe found evidence of water trapped just below the icy surface of the Martian polar region. These discoveries by geologists have sparked the interest of biologists, for where there is water, there can be life. There is good reason to believe that life as we know it cannot exist without water.

Animals and plants that live on Earth's land masses had to evolve elaborate ways to retain the water that makes up about 70 percent of their bodies. Aquatic organisms living in water do not need these water-retention mechanisms; thus biologists have concluded that the first living things originated in a watery environment. This environment need not have been the lakes, rivers, and oceans with which we are familiar. Living organisms have been found in hot springs at temperatures above the usual boiling point of water, in a lake beneath the frozen Antarctic ice, in water trapped 2 miles below Earth's surface, in water 3 miles below the surface of the sea, in extremely acid and extremely salty water, and even in the water that cools the interiors of nuclear reactors.

With 20 trillion galaxies in the universe, each with 100 billion stars, there are many planets out there, and if our own solar system is typical, some of them have the water needed for life. As biologists contemplate how life could originate from nonliving matter, their attention focuses not just on the presence of water, but on what is dissolved in it.

A major discovery of biology is that living things are composed of the same types of chemical elements as the vast nonliving portion of the universe. This *mechanistic* view—that life is chemically based and obeys universal physicochemical laws—is a relatively recent one in human history. The concept of a "vital force" responsible for life, different from the forces found in physics and chemistry, was common in Western culture until the nineteenth century, and many people still assume such a force exists. However, most scientists adhere to a mechanistic view of life, and it is the cornerstone of medicine and agriculture.



A Grander Canyon on Mars This false color image from the Mars Global Surveyor shows in blue the dry remains of what was once a huge lake on Mars. Just as the Colorado River carved Earth's Grand Canyon, torrents of water from the lake probably carved the mile-deep canyon that is visible as a thin blue line just north of the lake bed.

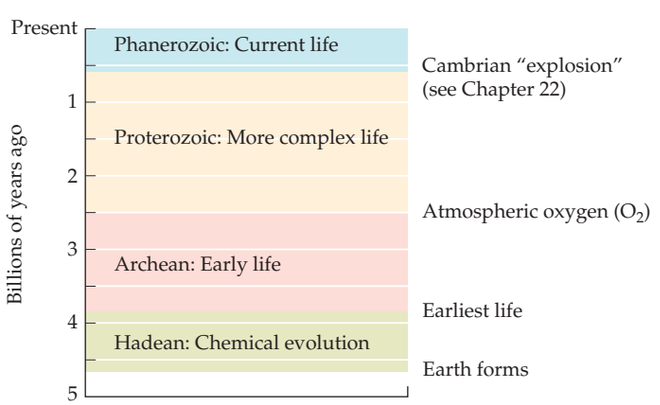
Before describing how chemical elements are arranged in living creatures, we will examine some fundamental chemical concepts. We will first address the constituents of matter: atoms. We will examine their variety, their properties, and their capacity to combine with other atoms. Then we will consider how matter changes. In addition to changes in state (solid to liquid to gas), substances undergo changes that transform both their composition and their characteristic properties. Then we will describe the structure and properties of water and its relationship to acids and bases. We will close the chapter with a consideration of characteristic groups of atoms that contribute specific properties to larger molecules of which they are part, and which will be the subject of Chapter 3.

Water and the Origin of Life's Chemistry

Astronomers believe our solar system began forming about 4.6 billion years ago when a star exploded and collapsed to form the sun, and 500 or so bodies called planetesimals collided with one another to form the inner planets, including Earth. The first chemical signatures indicating the presence of life here are about 4 billion years old. So it took 600 million years, during a geological time frame called the Hadean, for the chemical conditions on Earth to become just right for life. Key among those conditions was the presence of water.

Ancient Earth probably had a lot of water high in the atmosphere. But the new planet was hot, and this water evaporated into space. As Earth cooled, it became possible for water to remain on its surface, but where did that water come from? One current view is that comets—loose agglomerations of dust and ice that have orbited the sun since the planets formed—struck Earth repeatedly and brought not only water but other chemical components of life, such as nitrogen. As Earth cooled, chemicals from the rocks dissolved in the water and simple chemical reactions took place. Some of these reactions could have led to life, but impacts by large comets and rocky meteorites would have released enough energy to heat the developing oceans almost to boiling, thus destroying any early life. These large impacts eventually subsided, and life gained a foothold about 3.8 to 4 billion years ago. The prebiotic Hadean was over (Figure 2.1). The Archean had begun, and there has been life on Earth ever since.

In Chapter 3 we will return to the question of how the first life could have arisen from inanimate chemicals. But before doing so, we need to understand what the chemistry of life entails. Like the rest of the chemical world, living things are made up of atoms and molecules.

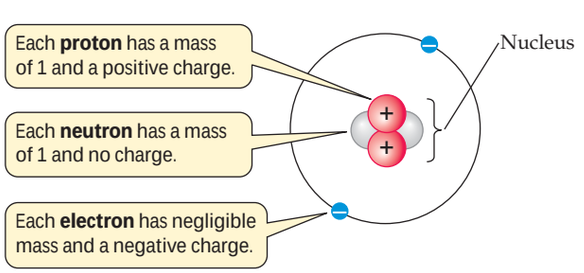


2.1 A Geological Time Scale The Hadean encompasses the time from the formation of Earth (about 4.6 billion years ago) until the earliest life appeared (about 3.8 billion years ago). During the Hadean chemical conditions evolved that were conducive to life, which was able to gain a foothold once the rain of comets and meteorites ended.

Atoms: The Constituents of Matter

More than a trillion (10^{12}) atoms could fit over the period at the end of this sentence. Each atom consists of a dense, positively charged **nucleus**, around which one or more negatively charged **electrons** move (Figure 2.2). The nucleus contains one or more **protons** and may contain one or more **neutrons**. Atoms and their component particles have volume and mass, which are properties of all matter. **Mass** measures the quantity of matter present; the greater the mass, the greater the quantity of matter.

The mass of a proton serves as a standard unit of measure: the atomic mass unit (amu), or *dalton* (named after the English chemist John Dalton). A single proton or neutron has a mass of about 1 dalton (Da), which is 1.7×10^{-24} grams (0.00000000000000000000000017 g). The mass of an electron is 9×10^{-28} g (0.0005 Da). Because the mass of an electron is negligible compare to the mass of a proton or a neutron, the



2.2 The Helium Atom This representation of a helium atom is called a Bohr model. It exaggerates the space occupied by the nucleus. In reality, although the nucleus accounts for virtually all of the atomic mass, it occupies only 1/10,000 of the atom's volume.

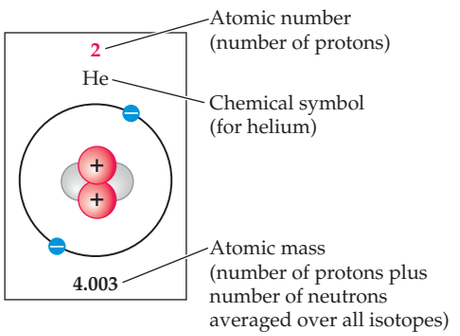
contribution of electrons to the mass of an atom can usually be ignored when measurements and calculations are made. It is electrons however, that determine how atoms will interact in chemical reactions, and we will discuss them extensively later in this chapter.

Each proton has a positive electric charge, defined as +1 unit of charge. An electron has a negative charge equal and opposite to that of a proton; thus the charge of an electron is –1 unit. The neutron, as its name suggests, is electrically neutral, so its charge is 0 unit. Unlike charges (+/–) attract each other; like charges (+/+ or –/–) repel each other. Atoms are electrically neutral: The number of protons in an atom equals the number of electrons.

An element is made up of only one kind of atom

An **element** is a pure substance that contains only one type of atom. The element hydrogen consists only of hydrogen atoms; the element iron consists only of iron atoms. The atoms of each element have certain characteristics or properties that distinguish them from the atoms of other elements. The more than 100 elements found in the universe are arranged in the periodic table (Figure 2.3). These elements are not found in equal amounts. Stars have abundant hydrogen and helium. Earth's crust, and those of the neighboring planets, are almost half oxygen, 28 percent silicon, 8 percent aluminum, 2–5 percent each of sodium, magnesium, potassium, calcium, and iron, and contain much smaller amounts of the other elements.

About 98 percent of the mass of every living organism (bacterium, turnip, or human) is composed of just six elements: carbon, hydrogen, nitrogen, oxygen, phosphorus, and sulfur. The chemistry of these six elements will be our primary con-



2.3 The Periodic Table The periodic table groups the elements according to their physical and chemical properties. Elements 1–92 occur in nature; elements above 92 were created in the laboratory.

The six elements highlighted in yellow make up 98% of the mass of most living organisms.

Vertical columns have elements with similar properties.

Elements framed in orange are present in small amounts in many organisms.

1 H 1.0079																	2 He 4.003
3 Li 6.941	4 Be 9.012											5 B 10.81	6 C 12.011	7 N 14.007	8 O 15.999	9 F 18.998	10 Ne 20.179
11 Na 22.990	12 Mg 24.305											13 Al 26.982	14 Si 28.086	15 P 30.974	16 S 32.06	17 Cl 35.453	18 Ar 39.948
19 K 39.098	20 Ca 40.08	21 Sc 44.956	22 Ti 47.88	23 V 50.942	24 Cr 51.996	25 Mn 54.938	26 Fe 55.847	27 Co 58.933	28 Ni 58.69	29 Cu 63.546	30 Zn 65.38	31 Ga 69.72	32 Ge 72.59	33 As 74.922	34 Se 78.96	35 Br 79.909	36 Kr 83.80
37 Rb 85.4778	38 Sr 87.62	39 Y 88.906	40 Zr 91.22	41 Nb 92.906	42 Mo 95.94	43 Tc (99)	44 Ru 101.07	45 Rh 102.906	46 Pd 106.4	47 Ag 107.870	48 Cd 112.41	49 In 114.82	50 Sn 118.69	51 Sb 121.75	52 Te 127.60	53 I 126.904	54 Xe 131.30
55 Cs 132.905	56 Ba 137.34	71 Lu 174.97	72 Hf 178.49	73 Ta 180.948	74 W 183.85	75 Re 186.207	76 Os 190.2	77 Ir 192.2	78 Pt 195.08	79 Au 196.967	80 Hg 200.59	81 Tl 204.37	82 Pb 207.19	83 Bi 208.980	84 Po (209)	85 At (210)	86 Rn (222)
87 Fr (223)	88 Ra 226.025	103 Lr (260)	104 Rf (261)	105 Db (262)	106 Sg (266)	107 Bh (264)	108 Hs (269)	109 Mt (268)	110 (269)	111 (272)	112 (277)	113	114 (285)	115 (289)	116	117	118 (293)
Lanthanide series		57 La 138.906	58 Ce 140.12	59 Pr 140.9077	60 Nd 144.24	61 Pm (145)	62 Sm 150.36	63 Eu 151.96	64 Gd 157.25	65 Tb 158.924	66 Dy 162.50	67 Ho 164.930	68 Er 167.26	69 Tm 168.934	70 Yb 173.04		
Actinide series		89 Ac 227.028	90 Th 232.038	91 Pa 231.0359	92 U 238.02	93 Np 237.0482	94 Pu (244)	95 Am (243)	96 Cm (247)	97 Bk (247)	98 Cf (251)	99 Es (252)	100 Fm (257)	101 Md (258)	102 No (259)		

Masses in parentheses indicate unstable elements that decay rapidly to form other elements.

Elements without a chemical symbol are as yet unnamed.

cern here, but the others are not unimportant. Sodium and potassium, for example, are essential for nerves to function; calcium can act as a biological signal; iodine is a component of a vital hormone; and plants need magnesium as part of their green pigment (chlorophyll) and molybdenum in order to incorporate nitrogen into biologically useful substances.

The number of protons identifies the element

An element is distinguished from other elements by the number of protons in each of its atoms. This number, which does not change, is called the *atomic number*. An atom of helium has 2 protons, and an atom of oxygen has 8 protons; the atomic numbers of these elements are thus 2 and 8, respectively.

Along with a definitive number of protons, every element except hydrogen has one or more neutrons in its nucleus. The *mass number* of an atom is the total number of protons and neutrons in its nucleus. The nucleus of a helium atom contains 2 protons and 2 neutrons; oxygen has 8 protons and 8 neutrons. Therefore, helium has a mass number of 4 and oxygen a mass number of 16. The mass number may be thought of as the mass of the atom in daltons.

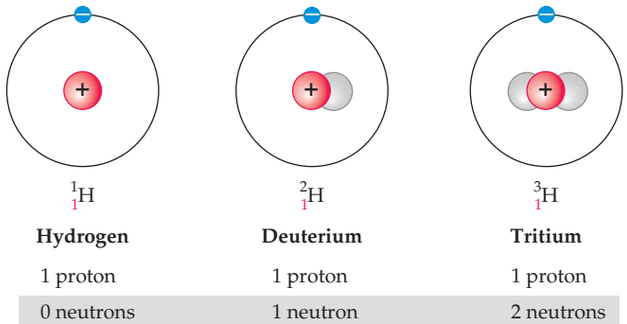
Each element has its own one- or two-letter chemical symbol. For example, H stands for hydrogen, He for helium, and O for oxygen. Some symbols come from other languages: Fe (from the Latin, *ferrum*) stands for iron, Na (Latin, *natrium*) for sodium, and W (German, *Wolfram*) for tungsten.

In text, immediately preceding the symbol for an element, the atomic number is written at the lower left and the mass number at the upper left. Thus, hydrogen, carbon, and oxygen are written as $^{1}_{1}H$, $^{12}_{6}C$, and $^{16}_{8}O$, respectively.

Isotopes differ in number of neutrons

Elements can have more than one atomic form. **Isotopes** of the same element all have the same, definitive, number of protons, but differ in the number of neutrons in the atomic nucleus.

In nature, many elements exist as several isotopes. The isotopes of hydrogen shown in Figure 2.4 have special names, but the isotopes of most elements do not have distinct names. For example, the natural isotopes of carbon are ^{12}C, ^{13}C, and ^{14}C (spoken of as carbon-12, carbon-13, and carbon-14). Most carbon atoms are ^{12}C, about 1.1 percent are ^{13}C, and a tiny fraction are ^{14}C. An element's atomic mass, or **atomic weight**,* is the average of the mass numbers of a representative sample of atoms of the element, with all isotopes in their normally occurring proportions. The atomic weight of carbon is thus calculated to be 12.011.



2.4 Isotopes Have Different Numbers of Neutrons The isotopes of hydrogen all have one proton in the nucleus, defining them as that element. Their differing mass numbers are due to different numbers of neutrons.

Some isotopes, called *radioisotopes*, are unstable and spontaneously give off energy as α (alpha), β (beta), or γ (gamma) radiation from the atomic nucleus. Such radioactive decay transforms the original atom into another atom, usually of another element. For example, carbon-14 loses a β-particle (actually an electron) to form ^{14}N. Biologists and physicians can incorporate radioisotopes into molecules and use the emitted radiation as a tag to locate those molecules or to identify changes that the molecules undergo inside the body (Figure 2.5). Three radioisotopes commonly used in this way are ^{3}H (tritium), ^{14}C (carbon-14), and ^{32}P (phosphorus-32). In addition to these applications, radioisotopes can be used to date fossils (see Chapter 22).

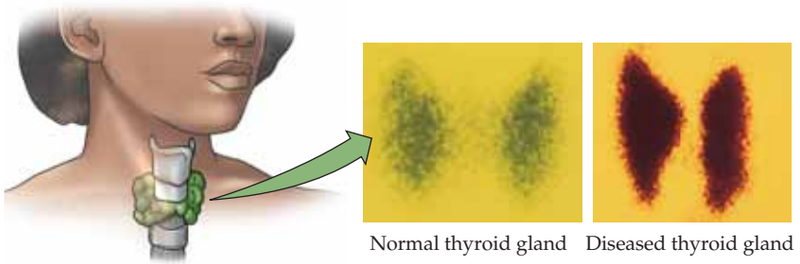
Although radioisotopes are useful for experiments and in medicine, even low doses of their radiation have the potential to damage molecules and cells. The devastating effects of radiation from nuclear weapons are well known, as are concerns about possible damage to organisms from isotopes used in nuclear power plants. In medicine, γ-radiation from ^{60}Co (cobalt-60) is used to damage or kill cancer cells.

In discussing isotopes and radioactivity, we have focused on the nucleus of the atom, but the nucleus is not directly involved in the ability of atoms to combine with other atoms. That ability is determined by the number and distribution of electrons. In the following sections, we describe some of the properties and chemical behavior of electrons.

Electron behavior determines chemical bonding

When considering atoms, biologists are concerned primarily with electrons because the behavior of electrons explains how chemical changes occur in living cells. These changes, called *chemical reactions* or just *reactions*, are changes in the atomic

*The concepts of "weight" and "mass" are not identical. Weight is the measure of the Earth's gravitational attraction for mass; on another planet, the same quantity of mass would have a different weight. On Earth, however, the term "weight" is often used as a measure of mass, and in biology one encounters the terms "weight" and "atomic weight" more frequently than "mass" and "atomic mass." Therefore, we will use "weight" for the remainder of this book.



2.5 A Radioisotope Used in Medicine The thyroid gland takes up iodine and uses it to make thyroid hormone. A patient suspected of having thyroid disease can be injected with radioactive iodine, which allows the thyroid gland to be visualized by a scanning device.

composition of substances. The characteristic number of electrons in each atom of an element determines how its atoms will react with other atoms. All chemical reactions involve changes in the relationships of electrons with one another.

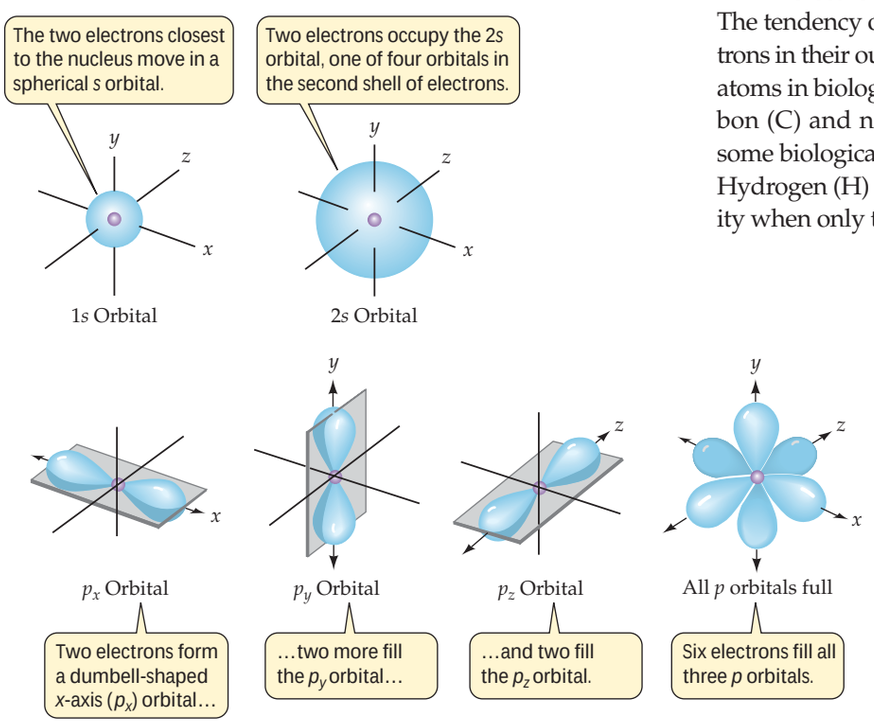
The location of a given electron in an atom at any given time is impossible to determine. We can only describe a volume of space within the atom where the electron is likely to be. The region of space where the electron is found at least 90 percent of the time is the electron's *orbital* (Figure 2.6). In an atom, a given orbital can be occupied by at most two electrons. Thus any atom larger than helium (atomic number 2) must have electrons in two or more orbitals. As Figure 2.6 shows, the different orbitals have characteristic forms and orientations in space. The orbitals, in turn, constitute a series of *electron shells*, or energy levels, around the nucleus (Figure 2.7).

- The innermost electron shell consists of only one orbital, called an *s* orbital. Hydrogen ($_1$H) has one electron in its first shell; helium ($_2$He) has two. All other elements have two first-shell electrons, as well as electrons in other shells.
- The second shell is made up of four orbitals (an *s* orbital and three *p* orbitals) and hence can hold up to eight electrons.

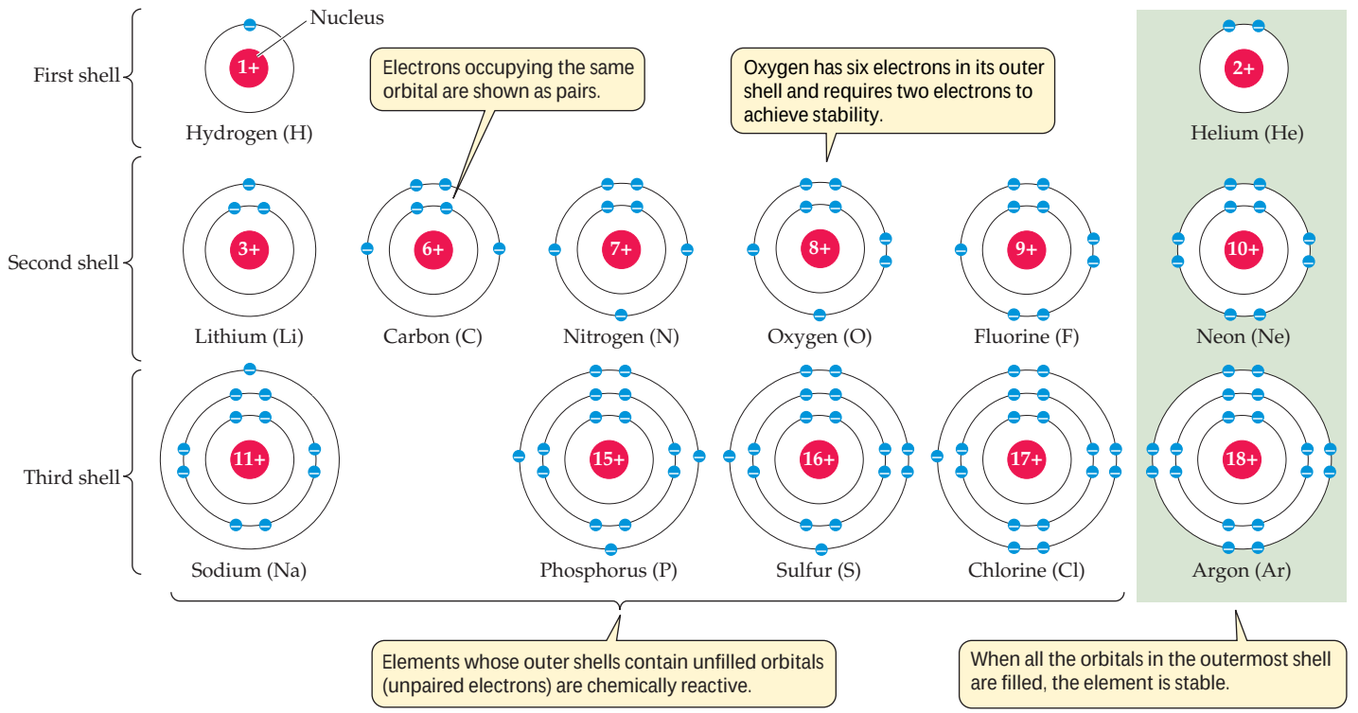
The *s* orbitals fill with electrons first, and their electrons have the lowest energy. Subsequent shells have different numbers of orbitals, but the outermost shells usually hold only eight electrons. In any atom, the outermost electron shell determines how the atom combines with other atoms—that is, how the atom behaves chemically. When an outermost shell consisting of four orbitals contains eight electrons, there are no unpaired electrons (see Figure 2.7). Such an atom is *stable* and will not react with other atoms. Examples of chemically stable elements are helium, neon, and argon.

Reactive atoms seek to attain the stable condition of having no unpaired electrons in their outermost shells. They attain this stability by sharing electrons with other atoms, or by gaining or losing one or more electrons. In either case, the atoms are bonded together. Such bonds create stable associations of atoms called molecules.

A **molecule** is two or more atoms linked by chemical bonds. The tendency of atoms in stable molecules to have eight electrons in their outermost shells is known as the octet rule. Many atoms in biologically important molecules—for example, carbon (C) and nitrogen (N)—follow the *octet rule*. However, some biologically important atoms are exceptions to the rule. Hydrogen (H) is the most obvious exception, attaining stability when only two electrons occupy its single shell.



2.6 Electron Orbitals Each orbital holds a maximum of two electrons. The *s* orbitals have a lower energy level and fill with electrons before the *p* orbitals do.

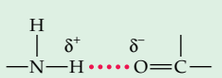
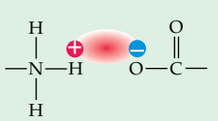
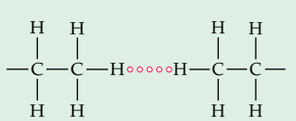
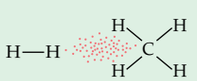


2.7 Electron Shells Determine the Reactivity of Atoms Each orbital holds a maximum of two electrons, and each shell can hold a specific maximum number of electrons. Each shell must be filled before electrons move into the next shell. The energy level of electrons is higher in shells farther from the nucleus. An atom with unpaired electrons in its outermost shell may react (bond) with other atoms.

Chemical Bonds: Linking Atoms Together

A **chemical bond** is an attractive force that links two atoms together to form a molecule. There are several kinds of chemical bonds (Table 2.1). In this section, we will first discuss co-

2.1 Chemical Bonds and Interactions

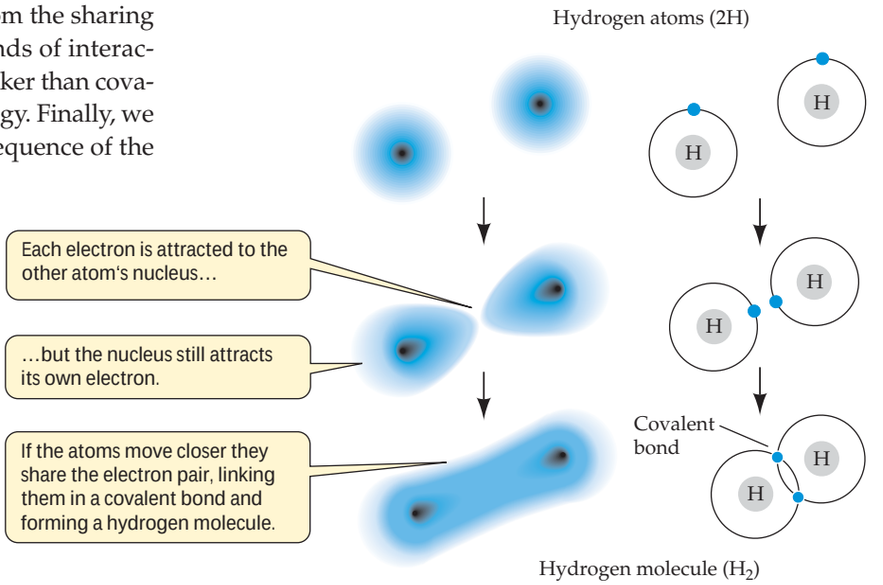
NAME	BASIS OF INTERACTION	STRUCTURE	BOND ENERGY[a] (KCAL/MOL)
Covalent bond	Sharing of electron pairs		50–110
Hydrogen bond	Sharing of H atom		3–7
Ionic bond	Attraction of opposite charges		3–7
Hydrophobic interaction	Interaction of nonpolar substances in the presence of polar substances		1–2
van der Waals interaction	Interaction of electrons of nonpolar substances		1

[a]*Bond energy* is the amount of energy needed to separate two bonded or interacting atoms under physiological conditions.

valent bonds, the strong bonds that result from the sharing of electrons. Then we will examine other kinds of interactions, including hydrogen bonds, that are weaker than covalent bonds but enormously important to biology. Finally, we will consider ionic bonding, which is a consequence of the loss or gain of electrons by atoms.

Covalent bonds consist of shared pairs of electrons

When two atoms attain stable electron numbers in their outermost shells by sharing one or more pairs of electrons, a **covalent bond** forms. Consider two hydrogen atoms in close proximity, each with a single unpaired electron in its outer shell. Each positively charged nucleus attracts the other atom's unpaired electron, but this attraction is balanced by each electron's attraction to its own nucleus. Thus the two unpaired electrons become shared by both atoms, filling the outer shells of both of them (Figure 2.8). The two atoms are thus linked by a covalent bond, and a hydrogen gas molecule (H_2) is formed.



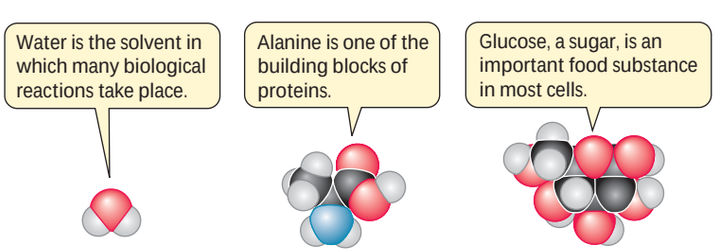
2.8 Electrons Are Shared in Covalent Bonds Two hydrogen atoms combine to form a hydrogen molecule. Each electron is attracted to both protons. A covalent bond forms when the electron orbitals of the two atoms overlap.

A molecule made up of more than one type of atom is called a compound. A molecular formula uses chemical symbols to identify the different atoms in a compound and subscript numbers to show how many of each type of atoms are present. Thus, the formula for sucrose—table sugar—is $C_{12}H_{22}O_{11}$. Each compound has a **molecular weight** (molecular mass) that is the sum of the atomic weights of all atoms in the molecule. Looking at the periodic table in Figure 2.3, you can calculate the molecular weight of table sugar to be 342. Molecular weights are usually related to a molecule's size (Figure 2.9).

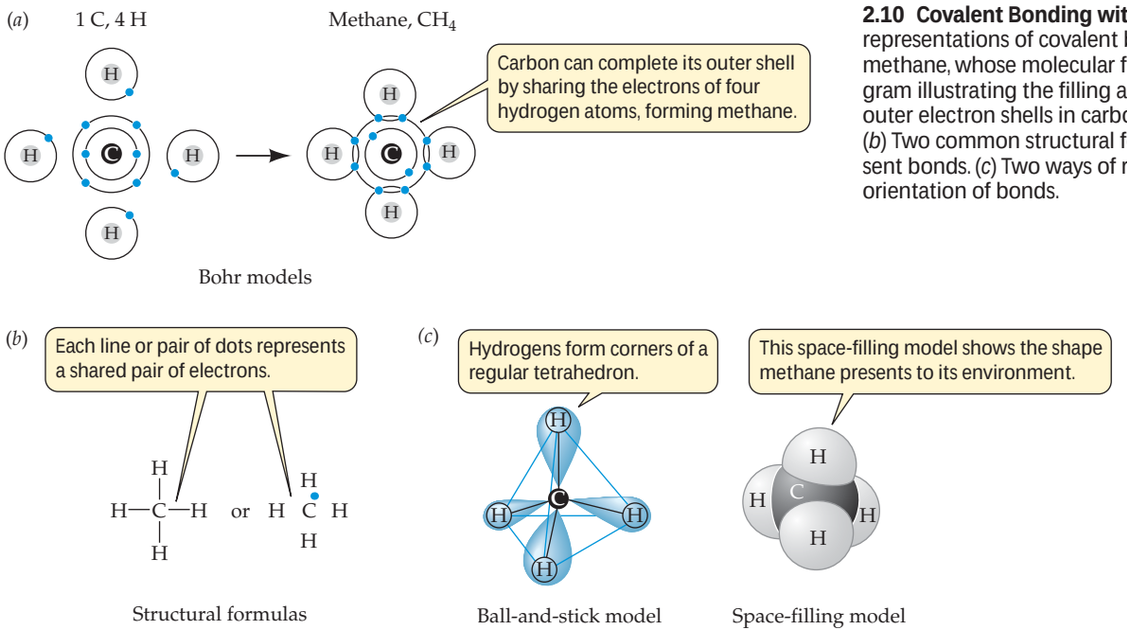
A carbon atom has a total of six electrons; two electrons fill its inner shell and four are in its outer shell. Because its outer shell can hold up to eight electrons, carbon can share electrons with up to four other atoms—it can form four covalent bonds. When an atom of carbon reacts with four hydrogen atoms, a molecule called methane (CH_4) forms (Figure 2.10*a*). Thanks to electron sharing, the outer shell of methane's carbon atom is filled with eight electrons, and the outer shell of each hydrogen atom is also filled. Four covalent bonds—each consisting of a shared pair of electrons—hold methane together. Table 2.2 shows the covalent bonding capacities of some biologically significant elements.

ORIENTATION OF COVALENT BONDS. Covalent bonds are very strong. The thermal energy that biological molecules ordinarily have at body temperature is less than 1 percent of that needed to break covalent bonds. So biological molecules, most of which are put together with covalent bonds, are quite stable. This means that their three-dimensional structures and the spaces they occupy are also stable. A second property of covalent bonds is that, for a given pair

2.9 Weights and Sizes of Atoms and Molecules The color conventions used here are standard for the atoms. (Yellow is used for sulfur and phosphorus atoms, which are not depicted.)



	Hydrogen (H)	Carbon (C)	Nitrogen (N)	Oxygen (O)	Water	Alanine	Glucose
Molecular weights	1	12	14	16	18	89	180



2.10 Covalent Bonding with Carbon Different representations of covalent bond formation in methane, whose molecular formula is CH_4. (*a*) Diagram illustrating the filling and stabilizing of the outer electron shells in carbon and hydrogen atoms. (*b*) Two common structural formulas used to represent bonds. (*c*) Two ways of representing the spatial orientation of bonds.

of atoms, they are the same in length, angle, and direction, regardless of the larger molecule of which the particular bond is a part. The four filled orbitals around the carbon nucleus of methane, for example, distribute themselves in space so that the bonded hydrogens are directed to the corners of a regular tetrahedron with carbon in the center (Figure 2.10*c*). The three-dimensional structure of carbon and hydrogen is the same in a large, complicated protein as it is in the simple methane molecule. This property of covalent bonds makes the prediction of biological structure possible.

Although the orientations of orbitals and the shapes of molecules differ depending on the types of atoms involved and how they are linked together, it is essential to remember that all molecules occupy space and have three-dimensional shapes. The shapes of molecules contribute to their biological functions, as we will see in Chapter 3.

2.2 Covalent Bonding Capabilities of Some Biologically Important Elements

ELEMENT	USUAL NUMBER OF COVALENT BONDS
Hydrogen (H)	1
Oxygen (O)	2
Sulfur (S)	2
Nitrogen (N)	3
Carbon (C)	4
Phosphorus (P)	5

MULTIPLE COVALENT BONDS. A covalent bond is represented by a line between the chemical symbols for the linked atoms:

- A single bond involves the sharing of a single pair of electrons (for example, H — H, C — H).
- A double bond involves the sharing of four electrons (two pairs) (C ═ C).

Triple bonds (six shared electrons) are rare, but there is one in nitrogen gas (N ≡ N), the chief component of the air we breathe.

UNEQUAL SHARING OF ELECTRONS. If two atoms of the same element are covalently bonded, there is an equal sharing of the pair(s) of electrons in the outer shell. However, when the two atoms are of different elements, the sharing is not necessarily equal. One nucleus may exert a greater attractive force on the electron pair than the other nucleus, so that the pair tends to be closer to that atom.

The attractive force that an atom exerts on electrons is its **electronegativity**. It depends on how many positive charges a nucleus has (nuclei with more protons are more positive and thus more attractive to electrons) and how far away the electrons are from the nucleus (closer means more electronegativity). The closer two atoms are in electronegativity, the more equal their sharing of electrons will be.

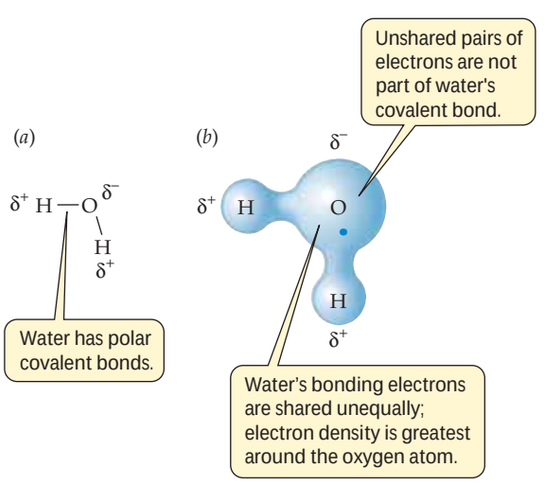
Table 2.3 shows the electronegativities of some elements important in biological systems. Looking at the table, it is obvious that two oxygen atoms, both with electronegativity of 3.5, will share electrons equally, producing what is called a *nonpolar covalent bond*. So will two hydrogen atoms (both 2.1).

2.3 Some Electronegativities

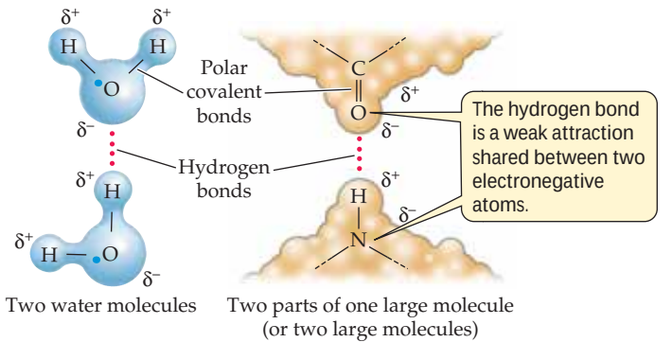
ELEMENT	ELECTRONEGATIVITY
Oxygen (O)	3.5
Chlorine (Cl)	3.1
Nitrogen (N)	3.0
Carbon (C)	2.5
Phosphorus (P)	2.1
Hydrogen (H)	2.1
Sodium (Na)	0.9
Potassium (K)	0.8

But when hydrogen bonds with oxygen to form water, the electrons involved are unequally shared: they tend to be nearer to the oxygen nucleus because it is the more electronegative of the two. The result is called a *polar* covalent bond (Figure 2.11).

Because of this unequal sharing of electrons, the oxygen end of the hydrogen–oxygen bond has a slightly negative charge (symbolized δ^- and spoken as "delta negative," meaning a partial unit of charge), and the hydrogen end is slightly positive (δ^+). The bond is **polar** because these opposite charges are separated at the two ends, or poles, of the bond. The partial charges that result from polar covalent bonds produce polar molecules or polar regions of large molecules. Polar bonds greatly influence the interactions between molecules that contain them.



2.11 The Polar Covalent Bond in the Water Molecule (*a*) A covalent bond between atoms with different electronegativities is a polar covalent bond, and has partial (δ) charges at the ends. (*b*) In water, the electrons are displaced toward the oxygen atom and away from the hydrogen atoms.



2.12 Hydrogen Bonds Can Form between or within Molecules Hydrogen bonds can form between two molecules or, if a molecule is large, between two different parts of the same molecule. Covalent and polar covalent bonds, on the other hand, are always found within molecules.

Hydrogen bonds may form within or between atoms with polar covalent bonds

In liquid water, the negatively charged oxygen (δ^-) atom of one water molecule is attracted to the positively charged hydrogen (δ^+) atoms of another water molecule. (Remember, negative charges attract positive charges.) The bond resulting from this attraction is called a **hydrogen bond**.

Hydrogen bonds are not restricted to water molecules. They may form between an electronegative atom and a hydrogen atom covalently bonded to a different electronegative atom (Figure 2.12). A hydrogen bond is a weak bond; it has about one-tenth (10%) the strength of a covalent bond between a hydrogen atom and an oxygen atom (see Table 2.1). However, where many hydrogen bonds form, they have considerable strength and greatly influence the structure and properties of substances. Later in this chapter we'll see how hydrogen bonding in water contributes to many of the properties that make water significant for living systems. Hydrogen bonds also play important roles in determining and maintaining the three-dimensional shapes of giant molecules such as DNA and proteins (see Chapter 3).

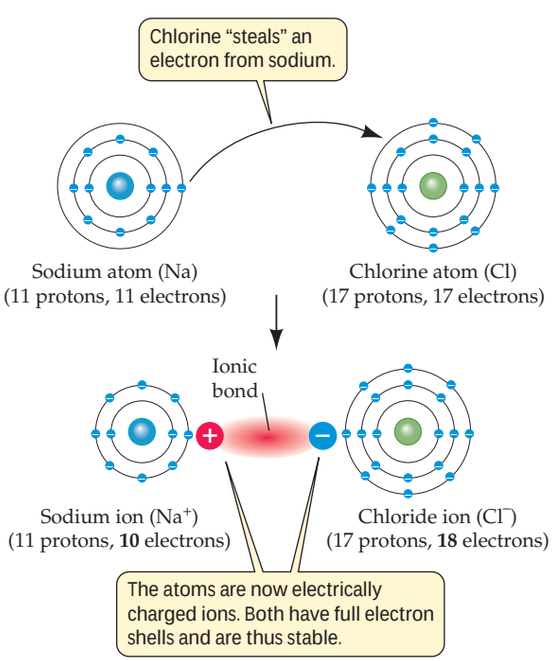
Ionic bonds form by electrical attraction

When one interacting atom is much more electronegative than the other, a complete transfer of one or more electrons may take place. Consider sodium (electronegativity 0.9) and chlorine (3.1). A sodium atom has only one electron in its outermost shell; this condition is unstable. A chlorine atom has seven electrons in its outer shell—another unstable condition. Since the electronegativities of these elements are so different, any electrons involved in bonding will tend to be much nearer to the chlorine nucleus—so near, in fact, that there is

a complete transfer of the electron from one element to the other (Figure 2.13). This reaction between sodium and chlorine makes the resulting atoms more stable. The result is two **ions**. Ions are electrically charged particles that form when atoms gain or lose one or more electrons.

- The sodium ion (Na^+) has a +1 unit charge because it has one less electron than it has protons. The outermost electron shell of the sodium ion is full, with eight electrons, so the ion is stable. Positively charged ions are called **cations**.
- The chloride ion (Cl^-) has a –1 unit charge because it has one more electron than it has protons. This additional electron gives Cl^- a stable outermost shell with eight electrons. Negatively charged ions are called **anions**.

Some elements form ions with multiple charges by losing or gaining more than one electron. Examples are Ca^{2+} (calcium ion, created from a calcium atom that has lost two electrons) and Mg^{2+} (magnesium ion). Two biologically important elements each yield more than one stable ion: Iron yields Fe^{2+} (ferrous ion) and Fe^{3+} (ferric ion), and copper yields Cu^+ (cuprous ion) and Cu^{2+} (cupric ion). Groups of covalently bonded atoms that carry an electric charge are called *complex ions;* examples include NH_4^+ (ammonium ion), SO_4^{2-} (sulfate ion), and PO_4^{3-} (phosphate ion).



2.13 Formation of Sodium and Chloride Ions When a sodium atom reacts with a chlorine atom, the more electronegative chlorine acquires a more stable, filled outer shell by obtaining an electron from the sodium. In so doing, the chlorine atom becomes a negatively charged chloride ion (Cl^-).The sodium atom, upon losing the electron, becomes a positively charged sodium ion (Na^+).

The charge from an ion radiates from it in all directions. Once formed, ions are usually stable, and no more electrons are lost or gained. Ions can form stable bonds, resulting in stable solid compounds such as sodium chloride (NaCl) and potassium phosphate (K_3PO_4).

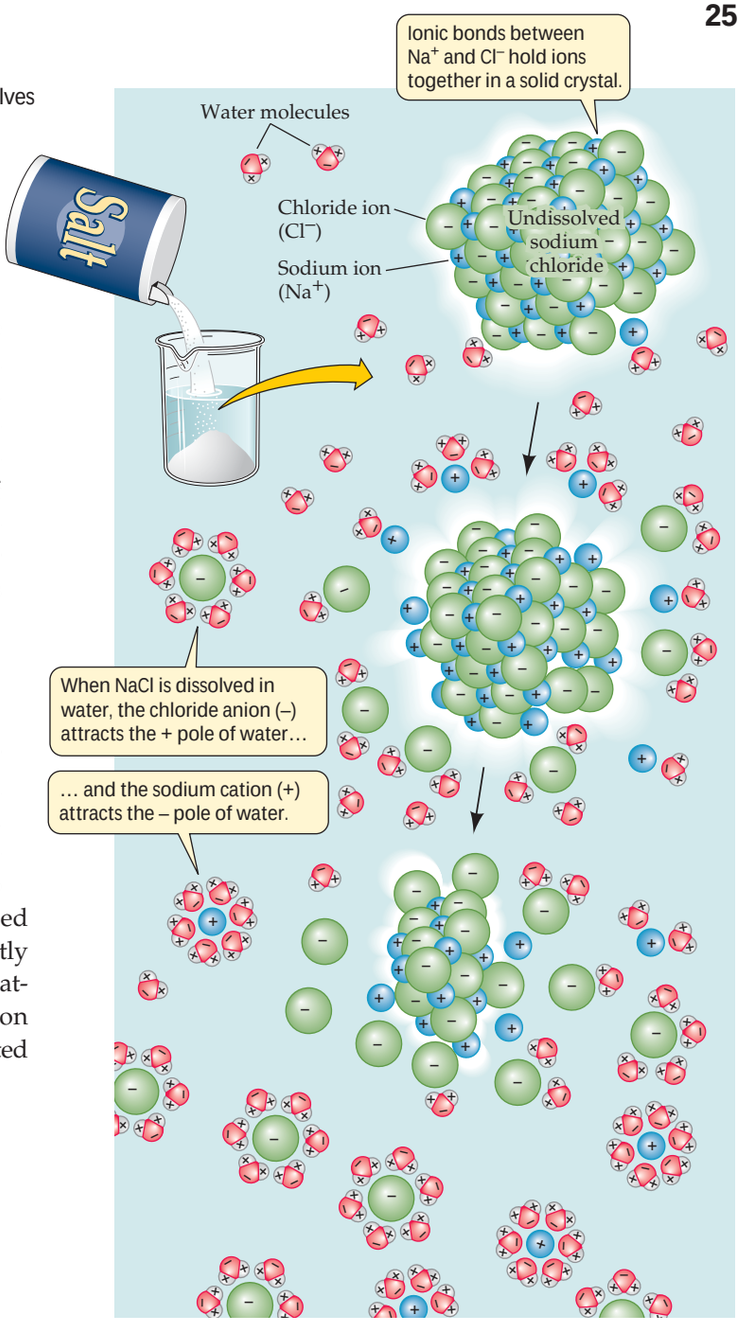
Ionic bonds are bonds formed by electrical attraction between ions bearing opposite charges. In sodium chloride—familiar to us as table salt—cations and anions are held together by ionic bonds. In solids, the ionic bonds are strong because the ions are close together. However, when ions are dispersed in water, the distance between them can be large; the strength of their attraction is thus greatly reduced. Under the conditions that exist in the cell, an ionic attraction is less than one-tenth as strong as a covalent bond that shares electrons equally (see Table 2.1).

Not surprisingly, ions with one or more units of charge can interact with polar molecules as well as with other ions. Such interaction results when table salt, or any other ionic solid, dissolves in water: "shells" of water molecules surround the individual ions, separating them (Figure 2.14). The hydrogen bond that we described earlier is a type of ionic bond because it is formed by electrical attraction. However, it is weaker than most ionic bonds because it is formed by partial charges (δ^+ and δ^-) rather than by whole-unit charges (+1 unit, –1 unit).

Polar and nonpolar substances interact best among themselves

"Like attracts like" is an old saying, and nowhere is it more true than in polar and nonpolar molecules, which tend to interact with their own kind. Just as water molecules interact with one another through polarity-induced hydrogen bonds, any molecule that is itself polar will interact with other polar molecules by weak (δ^+ to δ^-) attractions in hydrogen bonds. If a polar molecule interacts with water in this way, it is called **hydrophilic** ("water-loving").

What about nonpolar molecules? For example, carbon (electronegativity 2.5) forms nonpolar bonds with hydrogen (electronegativity 2.1). The resulting *hydrocarbon molecule*—that is, a molecule containing only hydrogen and carbon atoms—is nonpolar, and in water it tends to aggregate with other nonpolar molecules rather than with polar water. Such molecules are known as **hydrophobic** ("water-hating"), and the interactions between them are called hydrophobic interactions. It is important to realize that hydrophobic substances do not really "hate" water; they can form weak interactions with it (recall that the electronegativities of carbon and hydrogen are not exactly the same). But these interactions are far weaker than the hydrogen bonds between the water molecules, and so the nonpolar substances keep to themselves.



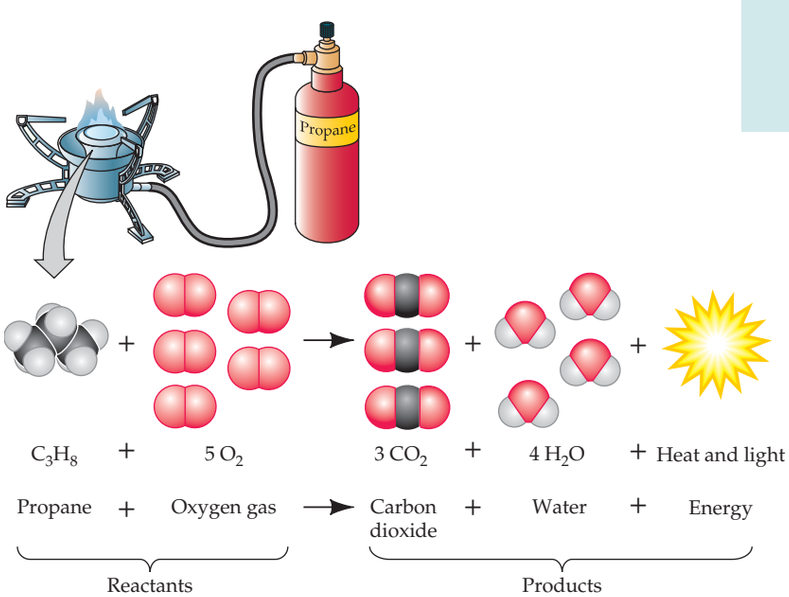
2.14 Water Molecules Surround Ions When an ionic solid dissolves in water, polar water molecules cluster around cations or anions, blocking their reassociation into a solid and forming a solution.

These weak interactions between nonpolar substances are enhanced by **van der Waals forces**, which result when two atoms of nonpolar molecules are in close proximity. These brief interactions are a result of random variations in the electron distribution in one molecule, which create an opposite charge distribution in the adjacent molecule. Although a single van der Waals interaction is brief and weak at any given site, the summation of many such interactions over the entire span of a large nonpolar molecule can produce substantial attraction. van der Waals forces are important in maintaining the structures of many biologically important substances.

Chemical Reactions: Atoms Change Partners

A **chemical reaction** occurs when atoms combine or change their bonding partners. Consider the combustion reaction that takes place in the flame of a propane stove. When propane (C_3H_8) reacts with oxygen gas (O_2), the carbon atoms become bonded to oxygen atoms instead of to hydrogen atoms, and the hydrogen atoms become bonded to oxygen instead of carbon (Figure 2.15). As the covalently bonded atoms change partners, the composition of the matter changes, and propane and oxygen gas become carbon dioxide and water. This chemical reaction can be represented by the equation

$$C_3H_8 + 5\ O_2 \rightarrow 3\ CO_2 + 4\ H_2O + \text{energy}$$



2.15 Bonding Partners and Energy May Change in a Chemical Reaction One molecule of propane reacts with five molecules of oxygen gas to give three molecules of carbon dioxide and four molecules of water. This reaction releases energy in the form of heat and light.

In this equation, the propane and oxygen are the **reactants**, and the carbon dioxide and water are the **products**. In this case, the reaction is complete: all the propane and oxygen are used up in forming the two products. The arrow symbolizes the direction of the chemical reaction. The numbers preceding the molecular formulas balance the equation and indicate how many molecules are used or are produced.

In this and all other chemical reactions, matter is neither created nor destroyed. The total number of carbons on the left equals the total number of carbons on the right. However, there is another product of this reaction: energy. The heat and light of the stove's flame reveal that the reaction of propane and oxygen releases a great deal of energy. **Energy** is defined as the capacity to do work, but on a more intuitive level, it can be thought of as the capacity for change. Chemical reactions do not create or destroy energy, but changes in energy usually accompany chemical reactions.

In the reaction between propane and oxygen, the energy that was released as heat and light was already present in the reactants in another form, called *potential chemical energy*. In some chemical reactions, energy must be supplied from the environment (for example, some substances will react only after being heated), and some of this supplied energy is stored as potential chemical energy in the bonds formed in the products.

We can measure the energy associated with chemical reactions using a unit called a **calorie** (cal). A calorie* is the amount of heat energy needed to raise the temperature of 1 gram of pure water from 14.5°C to 15.5°C. Another unit of energy that is increasingly used is the joule (J). When you compare data on energy, always compare joules to joules and calories to calories. The two units can be interconverted: 1 J = 0.239 cal, and 1 cal = 4.184 J. Thus, for example, 486 cal = 2,033 J, or 2.033 kJ. Although defined in terms of heat, the calorie and the joule are measures of any form of energy—mechanical, electric, or chemical.

Many biological reactions have much in common with the combustion of propane. The fuel is different—it is the sugar glucose, rather than propane—and the reactions proceed by many intermediate steps that permit the energy released from the glucose to be harvested and put to use by the cell. But the products are the same: carbon dioxide and water. These reactions were key to the origin of life from simpler molecules.

We will present and discuss energy changes, oxidation–reduction reactions, and several other types of chemical reactions that are prevalent in living systems in the chapters that follow.

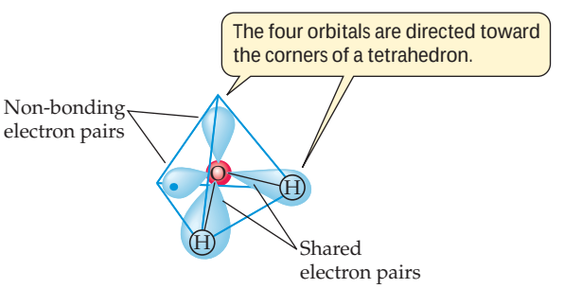
*The nutritionist's or dieter's Calorie, with a capital C, is what biologists call a kilocalorie (kcal) and is equal to 1,000 heat-energy calories.

Water: Structure and Properties

Water, like all other matter, can exist in three states: solid (ice), liquid, or gas (vapor). Liquid water is probably the medium in which life originated on Earth, and it is in water that life evolved for its first billion years. In this section, we will explore how the structure and interactions of water molecules make water essential to life.

Water has a unique structure and special properties

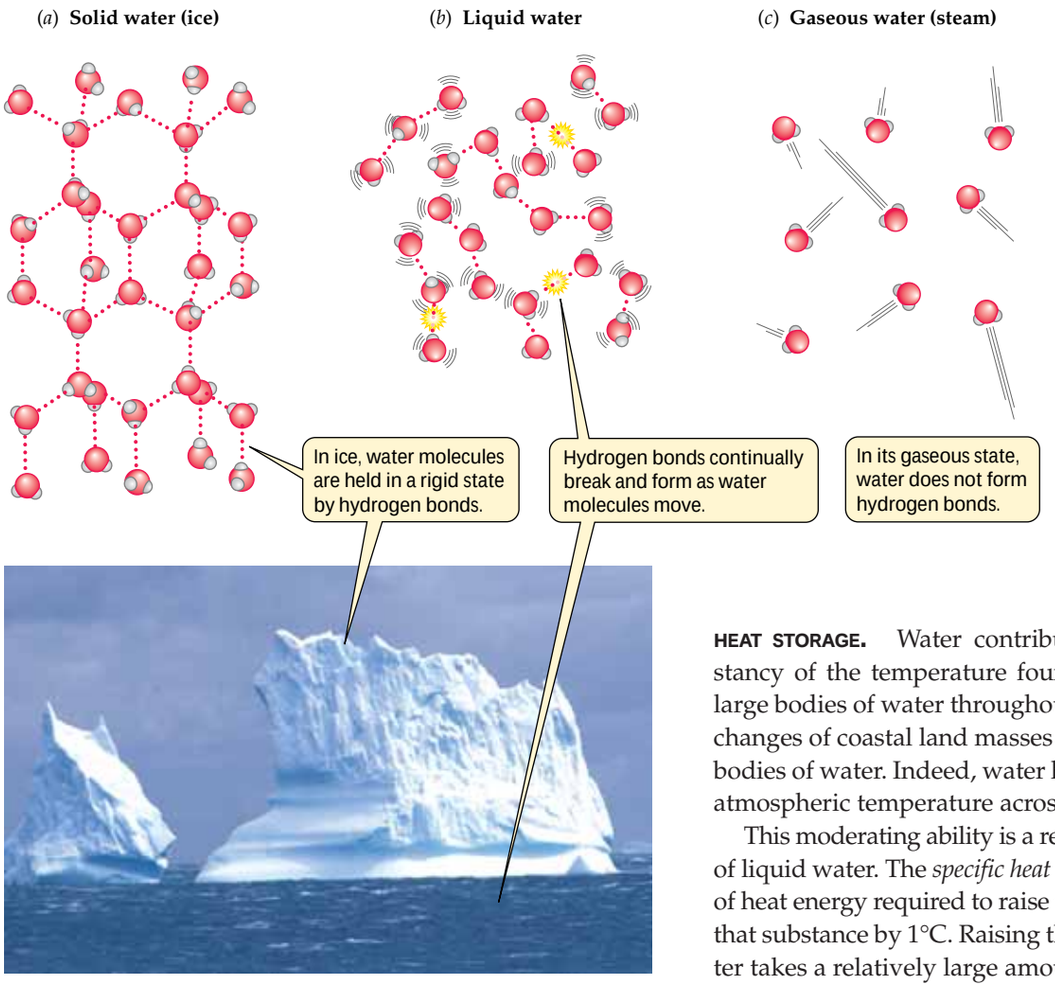
The water molecule, H_2O, has unique chemical features. As we learned in the preceding sections, water is a polar molecule that can form hydrogen bonds. In addition, the shape of water is a tetrahedron. The four pairs of electrons in the outer shell of oxygen repel one another, producing a tetrahedral shape.



These chemical features explain some of the interesting properties of water, such as the ability of ice to float, the melting and freezing temperatures of water, the ability of water to store heat, and the ability of water droplets to form. These properties are described in detail below.

ICE FLOATS. In water's solid state (ice), individual water molecules are held in place by hydrogen bonds, creating a rigid, crystalline structure in which each water molecule is hydrogen-bonded to four other water molecules (Figure 2.16*a*). Although the molecules are held firmly in place, they are not as tightly packed as they are in liquid water (Figure 2.16*b*). In other words, solid water is less dense than liquid water, which is why ice floats in water.

If ice were to sink in water, as almost all other solids do in their corresponding liquids, ponds and lakes would freeze from the bottom up, becoming solid blocks of ice in winter and killing most of the organisms living in them. Once the whole pond had frozen, its temperature could drop well below the freezing point of water. But, because ice floats, it forms a protective insulating layer on the top of the pond, reducing heat flow to the cold air above. Thus fish, plants, and other organisms in the pond are not subjected to temperatures lower than 0°C, the freezing point of pure water. The recent discovery of liquid water below the polar ice on



2.16 Hydrogen Bonds Hold Water Molecules Together
Hydrogen bonding exists between the molecules of water in both its liquid and solid states. (*a*) Solid water (ice). (*b*) Liquid water. Although more structured, ice is less dense than liquid water, so it floats. (*c*) Water forms a gas when its hydrogen bonds are broken and molecules move farther apart.

Mars has created speculation that life could exist in that environment.

MELTING AND FREEZING. Water is a moderator of temperature changes. Compared with other nonmetallic substances of the same size, molecular ice requires a great deal of heat energy to melt. Melting 1 mole (6.02×10^{23} molecules, a standard quantity; see page 28) of water requires the addition of 5.9 kJ of energy. This value is high because hydrogen bonds must be broken in order for water to change from solid to liquid. In the opposite process—freezing—a great deal of energy is lost when water is transformed from liquid to solid.

HEAT STORAGE. Water contributes to the surprising constancy of the temperature found in the oceans and other large bodies of water throughout the year. The temperature changes of coastal land masses are also moderated by large bodies of water. Indeed, water helps minimize variations in atmospheric temperature across the planet.

This moderating ability is a result of the high heat capacity of liquid water. The *specific heat* of a substance is the amount of heat energy required to raise the temperature of 1 gram of that substance by 1°C. Raising the temperature of liquid water takes a relatively large amount of heat because much of the heat energy is used to break the hydrogen bonds that hold the liquid together. Compared with other small molecules that are liquids, water has a high specific heat.

EVAPORATION AND COOLING. Water has a high heat of vaporization, which means that a lot of heat is required to change water from its liquid to its gaseous state (the process of *evaporation*). Once again, much of the heat energy is used to break hydrogen bonds. This heat must be absorbed from the environment in contact with the water. Evaporation thus has a cooling effect on the environment—whether a leaf, a forest, or an entire land mass. This effect explains why sweating cools the human body: as sweat evaporates off the skin, it uses up some of the adjacent body heat.

COHESION AND SURFACE TENSION. In liquid water, individual water molecules are free to move about. The hydrogen bonds between the molecules continually form and break. In other words, liquid water has a dynamic structure. On average, every water molecule forms 3.4 hydrogen bonds with other water molecules. This number represents fewer bonds than exist in ice, but it is still a high number. These

2.17 Surface Tension Water striders "skate" along, supported by the surface tension of the water that is their home.

hydrogen bonds explain the cohesive strength of liquid water. This cohesive strength permits narrow columns of water to stretch from the roots to the leaves of trees more than 100 meters high. When water evaporates from the leaves, the entire column moves upward in response to the pull of the molecules at the top.

Water also has a high surface tension, which means that the surface of liquid water exposed to the air is difficult to puncture. The water molecules in this surface layer are hydrogen-bonded to other water molecules below. The surface tension of water permits a container to be filled slightly above its rim without overflowing, and it permits small animals to walk on the surface of water (Figure 2.17).

Water is the solvent of life

A living organism is over 70 percent water by weight, excluding minerals such as bones. Many substances undergo reactions in this watery environment. Others do not, and thus form biological structures (such as bones).

A *solution* is produced when a substance (the *solute*) is dissolved in a liquid (the *solvent*) such as water (forming an *aqueous solution*). Many of the important molecules in biological systems are polar, and therefore are soluble in water. Reactions that take place in an aqueous solution can be studied in two ways:

- *Qualitative analysis* deals with substances dissolved in water and the chemical reactions that occur there. Qualitative analysis is the subject of much of the next few chapters.
- *Quantitative analysis* measures concentrations, or the amount of a substance in a given amount of solution. What follows is a brief introduction to some of the quantitative chemical terms you will see in this text.

Fundamental to quantitative thinking in chemistry and biology is the mole concept. A **mole** is the amount of an ion or compound (in grams) whose weight is numerically equal to its molecular weight. So a mole of table sugar ($C_{12}H_{22}O_{11}$) weighs 342 grams.

One aim of quantitative analysis is to study the behaviors of precise numbers of molecules in solution. But it is not possible to count molecules directly. Instead, chemists use a constant that relates the weight of any substance to the number of molecules of that substance. This constant is called *Avogadro's number*, which is 6.02×10^{23} molecules per mole. It allows chemists to work with moles of substances (which can be weighed out in the laboratory) instead of actual molecules. The mole concept is analogous to the concept of a dozen: We buy a dozen eggs or a dozen doughnuts, knowing that we will get 12 of whichever we buy. In the same way, when a physician injects a certain molar concentration of a drug into the bloodstream of a patient, a rough calculation can be made of the actual number of drug molecules that will interact with the patient's cells.

In the same way, chemists can dissolve a mole of sugar in water to make 1 liter of solution, knowing that the mole contains 6.02×10^{23} individual sugar molecules. This solution—1 mole of a substance dissolved in water to make 1 liter—is called a 1 molar (1 M) solution.

The many molecules dissolved in water in living tissues are not present at anything close to a 1 molar concentration. Most are in the micromolar (millionths of a mole per liter of solution; mM) to millimolar (thousandths of a mole per liter; μM) range. Some, such as hormone molecules, are even less concentrated than that. While these molarities seem to indicate very low concentrations, remember that even a 1 μM solution has 6.02×10^{17} molecules of the solute per liter.

Acids, Bases, and the pH Scale

When some substances dissolve in water, they release hydrogen ions (H^+), which are actually single, positively charged protons. These tiny bits of charged matter can attach to other molecules, and in doing so, change their properties. For example, the protons in acid rain can damage plants, and you are probably familiar with excess stomach acidity that affects digestion. In this section, we will examine the properties of substances that release H^+ (called **acids**) and substances that attach to H^+ (called **bases**). We will distinguish strong and weak acids and bases and provide a quantitative means for stating the concentration of H^+ in solutions: the pH scale.

Acids donate H^+, bases accept H^+

If hydrochloric acid (HCl) is added to water, it dissolves and ionizes, releasing the ions H^+ and Cl^-:

$$HCl \rightarrow H^+ + Cl^-$$

Because its H^+ concentration has increased, such a solution is acidic. Just like the combustion reaction of propane and oxygen (see Figure 2.15), the dissolution of HCl to form its ions is a complete reaction. HCl is therefore called a *strong acid.*

Acids *release* H^+ ions in solution. HCl is an acid, as is H_2SO_4 (sulfuric acid). One molecule of sulfuric acid may ionize to yield two H^+ and one SO_4^{2-}. Biological compounds that contain —COOH (the carboxyl group; see Figure 2.20) are also acids (such as acetic acid and pyruvic acid), because

$$—COOH \rightarrow —COO^- + H^+$$

Not all acids dissolve fully in water. For example, if acetic acid is added to water, at the end of the reaction, there are not just the two ions, but some of the original acid as well. Because the reaction is not complete, acetic acid is a *weak acid*.

Bases *accept* H^+ in solution. Just as with acids, there are strong and weak bases. If NaOH (sodium hydroxide) is added to water, it dissolves and ionizes, releasing OH^- and Na^+ ions:

$$NaOH \rightarrow Na^+ + OH^-$$

Because the concentration of OH^- increases and OH^- absorbs H^+ to form water, such a solution is basic. Because this reaction is complete, NaOH is a strong base.

Weak bases include the bicarbonate ion (HCO_3^-), which can accept a H^+ ion and become carbonic acid (H_2CO_3), and ammonia (NH_3), which can accept a H^+ and become an ammonium ion (NH_4^+). Amino groups (—NH_2) in biological molecules can also accept protons, thus acting as bases:

$$—NH_2 + H^+ \rightarrow —NH_3^+$$

The reactions between acids and bases may be reversible

When acetic acid is dissolved in water, two reactions happen. First, the acetic acid forms its ions:

$$CH_3COOH \rightarrow CH_3COO^- + H^+$$

Then, once ions are formed, they re-form acetic acid:

$$CH_3COO^- + H^+ \rightarrow CH_3COOH$$

This pair of reactions is reversible. A **reversible reaction** can proceed in either direction—left to right or right to left—depending on the relative starting concentrations of the reactants and products. The formula for a reversible reaction can be written using a double arrow:

$$CH_3COOH \rightleftharpoons CH_3COO^- + H^+$$

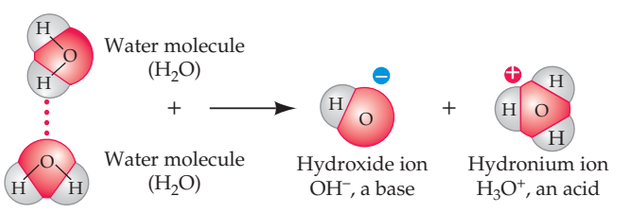
In principle, all chemical reactions are reversible. In terms of acids and bases, there are two types of reactions, depending on the extent of reversibility:

- Ionization of strong acids and bases is virtually irreversible.
- Ionization of weak acids and bases is somewhat reversible.

Many of the acid and base groups on large molecules in biological systems are weak.

Water is a weak acid

The water molecule has a slight but significant tendency to ionize into a hydroxide ion (OH^-) and a hydrogen ion (H^+). Actually, two water molecules participate in this ionization. One of the two molecules "captures" a hydrogen ion from the other, forming a hydroxide ion and a hydronium ion:



The hydronium ion is in effect a hydrogen ion bound to a water molecule. For simplicity, biochemists tend to use a modified representation of the ionization of water:

$$H_2O \rightarrow H^+ + OH^-$$

The ionization of water is important to all living creatures. This fact may seem surprising, since only about one water molecule in 500 million is ionized at any given time. But we are less surprised if we focus on the abundance of water in living systems and the reactive nature of the H^+ produced by ionization.

pH is the measure of hydrogen ion concentration

The terms "acidic" and "basic" refer only to solutions. How acidic or basic a solution is depends on the relative concentrations of H^+ and OH^- ions in it. The terms "acid" and "base" refer to compounds and ions. A compound or ion that is an acid can donate H^+; one that is a base can accept H^+.

How do we specify how acidic or basic a solution is? First, let's look at the H^+ concentrations of a few contrasting solutions. Remember that these concentrations are expressed in terms of molarity, the number of in moles of a substance in a liter of solution (see page 28). In pure water, the concentration of H^+ is 10^{-7} moles per liter (10^{-7} M). In 1 M hydrochloric acid, the H^+ concentration is 1 M; and in 1 M sodium hydroxide, the H^+ concentration is 10^{-14} M. Because

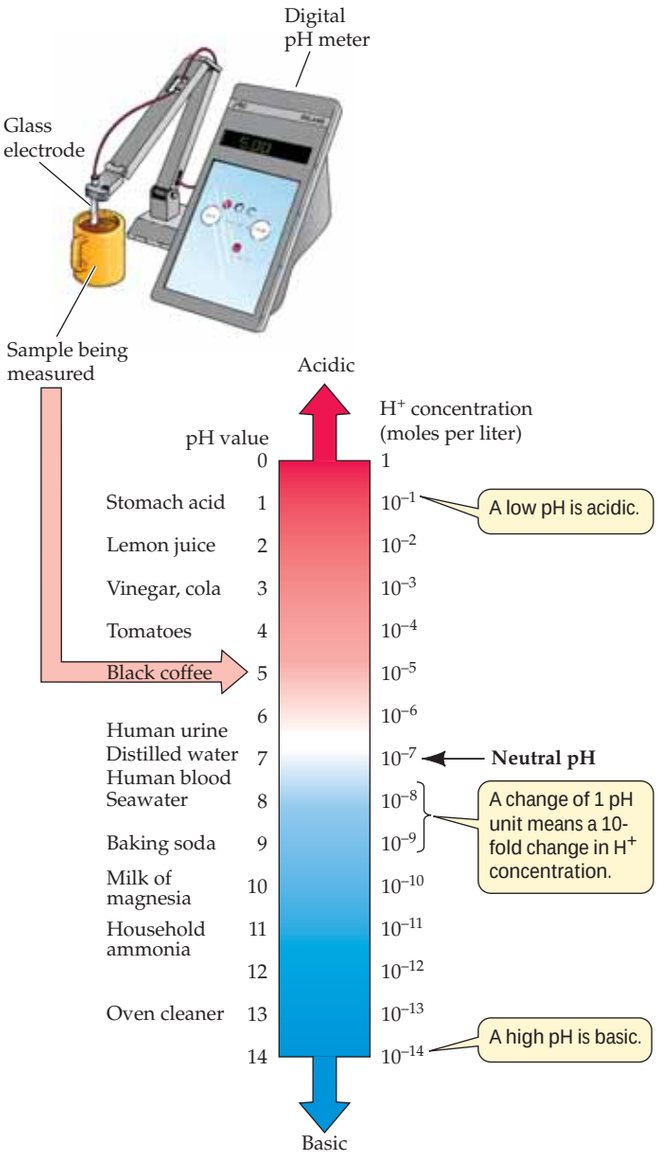
its values range so widely, the H^+ concentration itself is an inconvenient quantity to measure. It is easier to work with the logarithm of the concentration, because logarithms compress this range.

We indicate how acidic or basic a solution is by its pH. The pH value is defined as the negative logarithm of the hydrogen ion concentration in moles per liter (molar concentration). In chemical notation, molar concentration is often indicated by putting square brackets around the symbol for a substance; thus $[H^+]$ stands for the molar concentration of H^+. The equation for pH is

$$pH = -\log_{10}[H^+]$$

Since the H^+ concentration of pure water is 10^{-7} M, its pH is $-\log(10^{-7}) = -(-7)$, or 7. A smaller negative logarithm means a larger number. In practical terms, a lower pH means a higher H^+ concentration, or greater acidity. In 1 M HCl, the H^+ concentration is 1 M, so the pH is the negative logarithm of 1 ($-\log 10^0$), or 0. The pH of 1 M NaOH is the negative logarithm of 10^{-14}, or 14.

A solution with a pH of less than 7 is acidic—it contains more H^+ ions than OH^- ions. A solution with a pH of 7 is neutral, and a solution with a pH value greater than 7 is basic. Figure 2.18 shows the pH values of some common substances.



2.18 pH Values of Some Familiar Substances An electronic instrument similar to the one drawn at the top of the figure is used to measure the pH of a solution.

Buffers minimize pH change

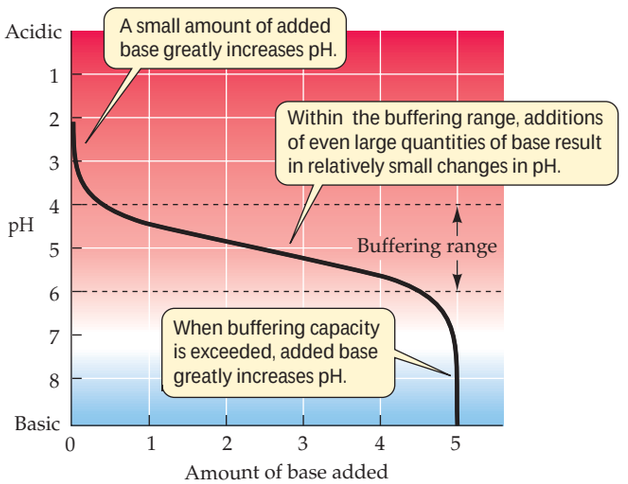
Some organisms, probably including the earliest forms of life, live in and have adapted to solutions with extremes of pH. However, most organisms control the pH of the separate compartments within their cells. The normal pH of human red blood cells, for example, is 7.4, and deviations of even a few tenths of a pH unit can be fatal. The control of pH is made possible in part by **buffers**: chemical mixtures that maintain a relatively constant pH even when substantial amounts of an acid or base are added.

A buffer is a mixture of a weak acid and its corresponding base—for example, carbonic acid (H_2CO_3) and bicarbonate ions (HCO_3^-). If an acid is added to a solution containing this buffer, not all the H^+ ions from that acid stay in solution. Instead, many of them combine with the bicarbonate ions to produce more carbonic acid. This reaction uses up some of the H^+ ions in the solution and decreases the acidifying effect of the added acid:

$$HCO_3^- + H^+ \rightleftharpoons H_2CO_3$$

If a base is added, the reaction essentially reverses. Some of the carbonic acid ionizes to produce bicarbonate ions and more H^+, which counteracts some of the added base. In this way, the buffer minimizes the effects of an added acid or base on pH. This is what happens in the blood, where this buffering system is important in preventing significant changes in pH that could disrupt the ability of the blood to function in carrying vital O_2 to tissues. A given amount of acid or base causes a smaller change in pH in a buffered solution than in an unbuffered one (Figure 2.19).

Buffers illustrate an important chemical principle in reversible reactions called the *law of mass action*. Addition of a reactant on one side of a reversible system drives the reaction in the direction that uses up that compound. In this case, addition of an acid drives the reaction in one direction; addition of a base drives it in the other direction.



2.19 Buffers Minimize Changes in pH With increasing amounts of added base, the overall slope of a graph of pH is downward. In the buffering range, however, the slope is shallow. At high and low values of pH, where the buffer is ineffective, the slopes are much steeper.

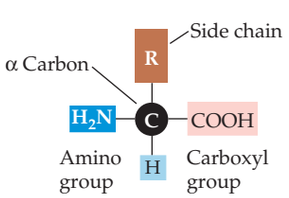
Properties of Molecules

So far, this chapter has discussed many properties of molecules, including size, polarity, solubility, and acid/base properties. Two other important properties that influence the behavior of molecules in a chemical reaction are the presence of recognizable functional groups, and existence of different isomers of molecules with the same chemical formula.

Functional groups give specific properties to molecules

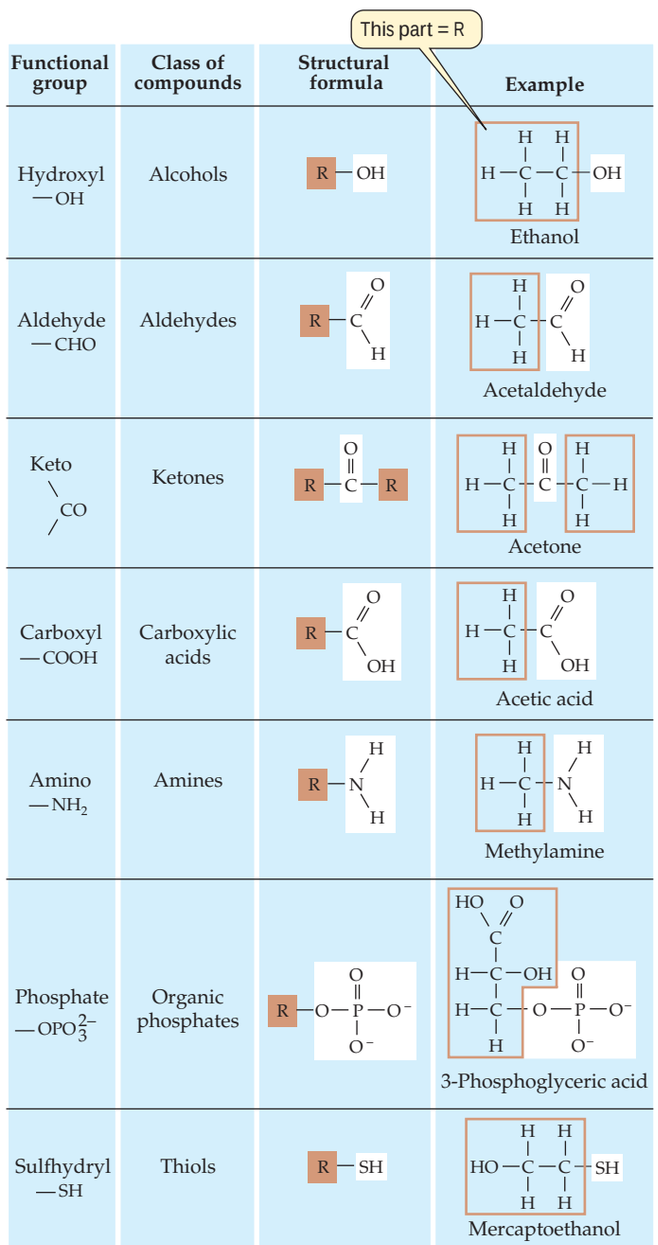
Certain small groups of atoms called **functional groups** are consistently found together in a variety of different molecules, a fact that simplifies our understanding of the reactions that molecules undergo in living cells. Each functional group has specific properties that, when attached to a larger molecule, in turn give the larger molecules specific properties. You will encounter several functional groups in your study of biology, including alcohols, aldehydes, ketones, acids, amines, phosphates, and thiols (Figure 2.20).

An important category of biological molecules containing functional groups is the amino acids, which have both a carboxyl group and an amino group attached to the same carbon atom, called the α carbon. Also attached to the α carbon atom are a hydrogen atom and a side chain, or R group, designated by the letter R:



Different side chains have different chemical compositions, structures, and properties. Each of the 20 amino acids found in proteins has a different side chain that gives it its distinctive chemical properties, as we'll see in Chapter 3.

Because they possess both carboxyl and amino groups, amino acids are simultaneously acids and bases. At the pH values commonly found in cells, both the carboxyl and the



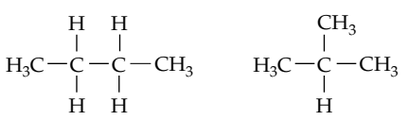
2.20 Some Functional Groups Important to Living Systems These functional groups (highlighted in white boxes) are the most common ones found in biologically important molecules. R represents the "remainder" of the molecule, which may be any of a large number of carbon skeletons or other chemical groups.

amino group are ionized: The carboxyl group has lost a proton, and the amino group has gained one.

Isomers have different arrangements of the same atoms

Isomers are molecules that have the same chemical formula but different arrangements of the atoms. (The prefix *iso-* , meaning "same," is encountered in many biological terms.) Of the different kinds of isomers, we will consider two: structural isomers and optical isomers.

Structural isomers differ in how their atoms are joined together. Consider two simple molecules, each composed of 4 carbon and 10 hydrogen atoms bonded covalently, both with the formula C_4H_{10}. These atoms can be linked in two different ways, resulting in two forms of the molecule:



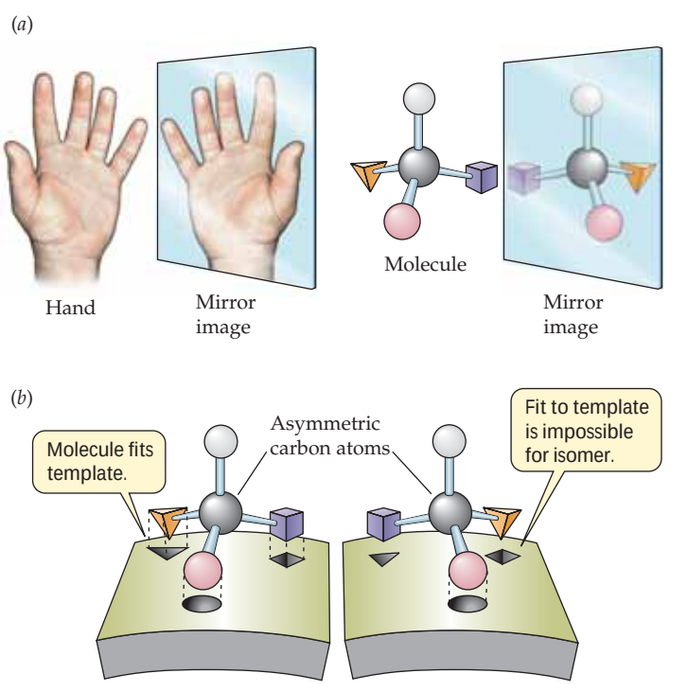
The different bonding relationships of butane and isobutane are distinguished in their structural formulas, and the two compounds have different chemical properties.

Optical isomers occur whenever a carbon atom has four different atoms or groups attached to it. This pattern allows two different ways of making the attachments, each the mirror image of the other (Figure 2.21). Such a carbon atom is an asymmetrical carbon, and the pair of compounds are optical isomers of each other. You can imagine your right and left hands as optical isomers. Just as a glove is specific for a particular hand, some biochemical molecules can interact with one optical isomer of a compound, but are unable to "fit" the other.

The α carbon in an amino acid is an asymmetrical carbon because it is bonded to four different functional groups. Therefore, amino acids exist in two isomeric forms, called D-amino acids and L-amino acids. D and L are abbreviations for the Latin terms for right (*dextro*) and left (*levo*), respectively. Only L-amino acids are commonly found in most organisms, and their presence is an important chemical "signature" for life.

Now that we have covered the major properties of all molecules, let's review them in preparation for the next chapter, which focuses on the major molecules of biological systems.

- *Molecules vary in size.* Some are small, such as H_2 and CH_4. Others are larger, such as a molecule of table sugar (sucrose, $C_{12}H_{22}O_{11}$), which has 45 atoms. Still other molecules, especially proteins such as hemoglobin (the oxygen carrier in red blood cells), are gigantic, sometimes containing tens of thousands of atoms. The formation of large molecules from simpler ones in the environment was a key precursor to the emergence of life during the Archean.
- *All molecules have a specific three-dimensional shape.* For example, the orientation of the bonding orbitals around the carbon atom gives the methane molecule (CH_4) the shape of a regular tetrahedron (see Figure 2.10*c*). In carbon dioxide (CO_2), the three atoms are in line. Larger molecules have complex shapes that result from the numbers and kinds of atoms present and the ways in which they are linked together. Some large molecules, such as hemoglobin, have compact, ball-like shapes. Others, such as the protein called keratin that makes up your hair, are long, thin, ropelike structures. Their shapes relate to the roles these molecules play in living cells.
- *Molecules are characterized by certain chemical properties* that determine their biological roles. Chemists use the characteristics of composition, structure (three-dimensional shape), reactivity, and solubility to distinguish a pure sample of one molecule from a sample of a different molecule. The presence of functional groups can impart distinctive chemical properties to a molecules, as does the physical arrangement of atoms into isomers.



2.21 Optical Isomers (*a*) Optical isomers are mirror images of each other. (*b*) Molecular optical isomers result when four different groups are attached to a single carbon atom. If a template is laid out to match the groups on one carbon atom, the groups on that molecule's mirror-image isomer cannot be rotated to fit the same template.

Between the small molecules discussed in this chapter and the world of the living cell stand the macromolecules. These huge molecules—proteins, lipids, carbohydrates, and nucleic acids—are the subject of the next chapter.

Chapter Summary

Water and the Origin of Life's Chemistry

▶ Current scientific evidence indicates that life as we know it cannot exist without water, and that life on Earth originated in the water of the planet's primordial oceans.

▶ The chemistry of life is ancient. Earth began forming about 4.6 billion years ago, and the first signs of life are 3.8–4 billion years old.

Atoms: The Constituents of Matter

▶ Matter is composed of atoms. Each atom consists of a positively charged nucleus of protons and neutrons, surrounded by electrons bearing negative charges. There are many elements in nature, but only a few of them make up the bulk of living systems. **Review Figures 2.2, 2.3**

▶ Isotopes of an element differ in their numbers of neutrons. Some isotopes are radioactive, emitting radiation as they decay. **Review Figure 2.4**

▶ Electrons are distributed in shells consisting of orbitals. Each orbital contains a maximum of two electrons. **Review Figures 2.6, 2.7. See Web/CD Activity 2.1**

▶ In losing, gaining, or sharing electrons to become more stable, an atom can combine with other atoms to form molecules. **Review Table 2.1**

Chemical Bonds: Linking Atoms Together

▶ Covalent bonds are strong bonds formed when two atomic nuclei share one or more pairs of electrons. Covalent bonds have spatial orientations that give molecules three-dimensional shapes. **Review Figures 2.8, 2.9, 2.10, Table 2.2**

▶ Nonpolar covalent bonds are formed when the electronegativities of two atoms are approximately equal. When atoms with strong electronegativity (such as oxygen) bond to atoms with weaker electronegativity (such as hydrogen), a polar covalent bond is formed, in which one end is δ^+ and the other is δ^-. **Review Figure 2.11, Table 2.3**

▶ Hydrogen bonds are weak electrical attractions that form between a δ^+ hydrogen atom in one molecule and a δ^- nitrogen or oxygen atom in another molecule or in another part of a large molecule. Hydrogen bonds are abundant in water. **Review Figure 2.12**

▶ Ions are electrically charged bodies that form when an atom gains or loses one or more electrons. Ionic bonds are electrical attractions between oppositely charged ions. Ionic bonds are strong in solids, but weaker when the ions are separated from one another in solution. **Review Figures 2.13, 2.14**

▶ Nonpolar molecules interact very little with polar molecules, including water. Nonpolar molecules are attracted to one another by very weak bonds called van der Waals forces. **See Web/CD Tutorial 2.1**

Chemical Reactions: Atoms Change Partners

▶ In chemical reactions, substances change their atomic compositions and properties. Energy is released in some reactions, whereas in others energy must be provided. Neither matter nor energy is created or destroyed in a chemical reaction, but both change form. **Review Figure 2.15**

▶ In living cells, chemical reactions take place in multiple steps so that the released energy can be harvested for cellular activities.

Water: Structure and Properties

▶ Water's molecular structure and its capacity to form hydrogen bonds give it unusual properties that are significant for life. Solid water floats in liquid water, and water gains or loses a great deal of heat when it changes its state, a property that moderates environmental temperature changes. **Review Figure 2.16**

▶ Water's high heat of vaporization assures effective cooling when water evaporates. The cohesion of water molecules permits liquid water to rise to great heights in narrow columns and produces a high surface tension.

▶ Solutions are produced when substances dissolve in water. The concentration of a solution is the amount of a given substance in a given amount of solution. Most biological substances are dissolved in water at very low concentrations.

Acids, Bases, and the pH Scale

▶ Acids are substances that donate hydrogen ions. Bases are substances that accept hydrogen ions.

▶ The pH of a solution is the negative logarithm of the hydrogen ion concentration. Values lower than pH 7 indicate an acidic solution; values above pH 7 indicate a basic solution. **Review Figure 2.18**

▶ Buffers are mixtures of weak acids and bases that limit the change in the pH of a solution when acids or bases are added. **Review Figure 2.19**

The Properties of Molecules

▶ Functional groups make up part of a larger molecule and have particular chemical properties. The consistent chemical behavior of functional groups helps us understand the properties of the molecules that contain them. **Review Figure 2.20. See Web/CD Activities 2.2, 2.3**

▶ Structural and optical isomers have the same kinds and numbers of atoms, but differ in their structures and properties. **Review Figure 2.21**

▶ Molecules vary in their size, shape, reactivity, solubility, and other chemical properties.

Self-Quiz

1. The atomic number of an element
   a. equals the number of neutrons in an atom.
   b. equals the number of protons in an atom.
   c. equals the number of protons minus the number of neutrons.
   d. equals the number of neutrons plus the number of protons.
   e. depends on the isotope.
2. The atomic weight (atomic mass) of an element
   a. equals the number of neutrons in an atom.
   b. equals the number of protons in an atom.
   c. equals the number of electrons in an atom.
   d. equals the number of neutrons plus the number of protons.
   e. depends on the relative abundances of its isotopes.

3. Which of the following statements about all the isotopes of an element is *not* true?
   a. They have the same atomic number.
   b. They have the same number of protons.
   c. They have the same number of neutrons.
   d. They have the same number of electrons.
   e. They have identical chemical properties.
4. Which of the following statements about a covalent bond is *not* true?
   a. It is stronger than a hydrogen bond.
   b. One can form between atoms of the same element.
   c. Only a single covalent bond can form between two atoms.
   d. It results from the sharing of electrons by two atoms.
   e. One can form between atoms of different elements.
5. Hydrophobic interactions
   a. are stronger than hydrogen bonds.
   b. are stronger than covalent bonds.
   c. can hold two ions together.
   d. can hold two nonpolar molecules together.
   e. are responsible for the surface tension of water.
6. Which of the following statements about water is *not* true?
   a. It releases a large amount of heat when changing from liquid into vapor.
   b. Its solid form is less dense than its liquid form.
   c. It is the most effective solvent of polar molecules.
   d. It is typically the most abundant substance in an active organism.
   e. It takes part in some important chemical reactions.
7. The following reaction occurs in the human stomach:

   $$HCl \rightarrow H^+ + Cl^-$$

   This reaction is an example of the
   a. cleavage of a covalent bond.
   b. formation of a hydrogen bond.
   c. elevation of the pH of the stomach.
   d. formation of ions by dissolving an acid.
   e. formation of polar covalent bonds.
8. The hydrogen bond between two water molecules arises because water is
   a. polar.
   b. nonpolar.
   c. a liquid.
   d. small.
   e. hydrophobic.
9. Which of the following statements about the carboxyl group is *not* true?
   a. It has the chemical formula —COOH.
   b. It is an acidic group.
   c. It can ionize.
   d. It is found in amino acids.
   e. It has an atomic weight of 75.
10. The three most abundant elements in a human skin cell are
    a. calcium, carbon, and oxygen.
    b. carbon, hydrogen, and oxygen.
    c. carbon, hydrogen, and sodium.
    d. carbon, nitrogen, and oxygen.
    e. nitrogen, hydrogen, and oxygen.

For Discussion

1. Would you expect the elemental composition of Earth's crust to be the same as that of the human body? How could you find out?
2. Some scientists and science fiction writers have envisioned life on other planets based not on carbon, as on Earth, but on silicon (Si). Using the Bohr model (see Figure 2.10*a*), draw the structure of silicon dioxide, showing electrons shared in covalent bonds.
3. Compare a covalent bond between two hydrogen atoms and a hydrogen bond between hydrogen and oxygen atoms with regards to the electrons involved, the role of polarity, and the strength of the bond.
4. The pH of the human stomach is about 2.0, while the pH of the small intestine is about 10.0. What are the hydrogen ion (H^+) concentrations inside these two organs?
5. The side chain of the amino acid glycine is simply a hydrogen atom (—H). Are there two optical isomers of glycine? Explain.

2 Life and Chemistry: Large Molecules

In 1984, a rock was found on the ice in the Allan Hills region of Antarctica. ALH 84001, as it came to be called, was a meteorite that came from Mars. We know this because the composition of the gases trapped within the rock was identical to the Martian atmosphere, which is quite different from Earth's atmosphere. Radioactive dating and mineral analyses determined that ALH 84001 was 4.5 billion years old and had been blasted off the Martian surface 16 million years ago, landing on Earth fairly recently, about 11,000 years ago.

Scientists found water trapped below the Martian meteorite's surface. This discovery was not surprising, considering that surface observations of Mars have indicated that liquid water may once have been abundant there (see Chapter 2). Because water is the *sine qua non* for life, scientists wondered whether the meteorite might contain other signs of life as well. Their analysis revealed two substances related to living systems. First, simple carbon-containing molecules called polycyclic aromatic hydrocarbons were present in small but unmistakable amounts; these substances are formed by decaying organisms, such as microbes. And second, crystals of magnetite, an iron oxide mineral made by many living things on Earth, were isolated from the interior of the rock.

ALH 84001 is not the only visitor from outer space that has been shown to contain the chemistry of life. Fragments of a meteorite that fell around the town of Murchison, Australia in 1969 were found to contain molecules that are unique to life, including purines and pyrimidines (the building blocks of DNA) and amino acids (which link together to form proteins). All of the amino acids showed a "handedness" that is unique to life.

These meteorites suggest that life is not found only on Earth, but they do not answer the question of how or where life arose from nonliving matter. We begin this chapter by presenting two hypotheses for the origin of life on Earth. After discussing these hypotheses, we take a detailed look at the four kinds of large molecules that characterize living organisms: proteins, carbohydrates, lipids (fats), and nucleic acids.

Was Life Once Here? The meteorite ALH 84001, which came from Mars and landed in Antarctica, contains the chemical signatures of life.

Theories of the Origin of Life

Living things are composed of the same elements as the inanimate universe, the 92 elements of the periodic table (see Figure 2.3). But the arrangements of these atoms into molecules in biological

systems are unique. You cannot find DNA in rocks unless it came from a once-living organism.

How life began on Earth sometime during the 600 million years of the Hadean is impossible for us to know for certain, given the vast amount of time that has passed. There are two theories of the origin of life: life from extraterrestrial sources, and chemical evolution.

Could life have come from outside Earth?

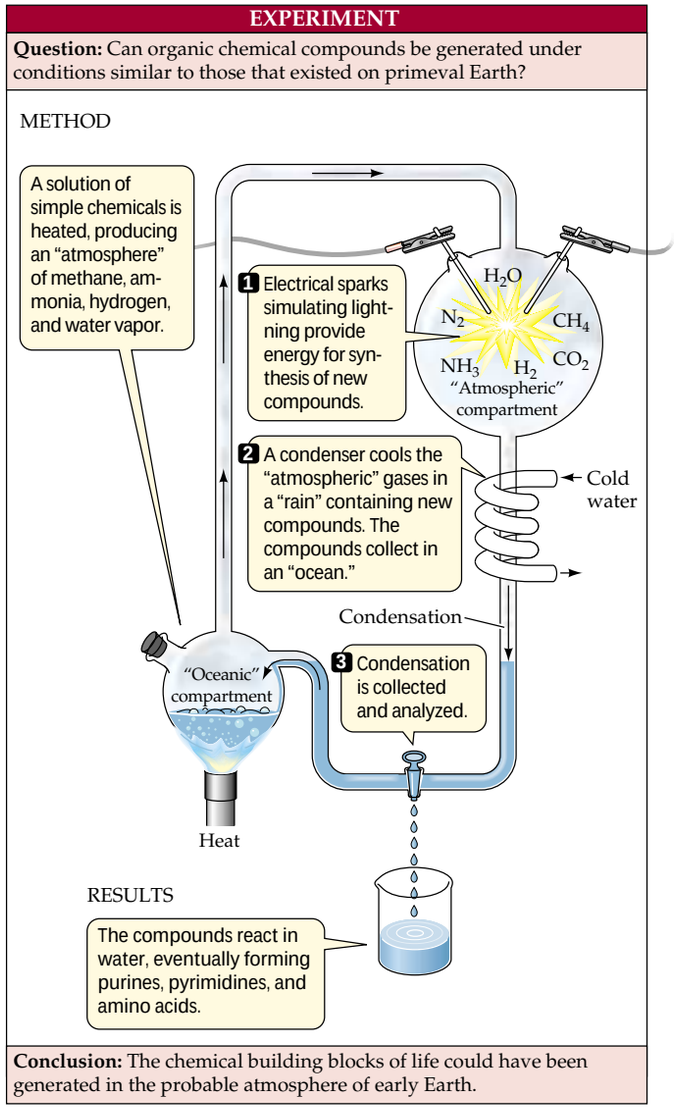
As we described in Chapter 2, comets probably brought Earth most of its water. The meteorites described at the beginning of this chapter are evidence that molecules characteristic of life may have traveled to Earth from space. Taken together, these two observations suggest that some of life's complex molecules could have come from space. Although the presence of such molecules in rocks may suggest that those rocks once harbored life, it does not prove that there were living things in the rocks when they landed on Earth. Claims that the spherical objects seen in ALH 84001 are the remnants of ancient Martian organisms are far from accepted by all scientists in the field.

Most scientists find it hard to believe that an organism in a meteorite could survive thousands of years traveling through space, followed by intense heat as it passed through Earth's atmosphere. But there is some evidence that the heat inside some meteorites may not have been severe. When weakly magnetized rock is heated, it reorients its magnetic field to align with the magnetic field around it. In the case of ALH 84001, this would have been Earth's powerful magnetic field, which would have affected the meteorite as it approached our planet. Careful measurements indicate that, while reorientation did occur at the surface of the rock, it did not occur in the inside. The scientists who took these measurements, Benjamin Weiss and Joseph Kirschvink at the California Institute of Technology, claim that the inside of ALH 84001 was never heated over 40°C on its trip to Antarctica, making a long interplanetary trip by living organisms more plausible.

Did life originate on Earth?

Both Earth and Mars once had the water and other simple molecules that could, under the right conditions, form the large molecules unique to life. The second theory of the origin of life on Earth, **chemical evolution**, holds that conditions on the primitive Earth led to the emergence of these molecules. Scientists have sought to reconstruct those primitive conditions.

Early in the twentieth century, researchers proposed that there was little oxygen gas (O_2) in Earth's first atmosphere (unlike today, when it constitutes 21 percent of our atmosphere). O_2 is thought to have accumulated in quantity about 2.5 billion years ago as the by-product of the metabolism of single-celled life forms. In the 1950s, Stanley Miller and Harold Urey set up an experimental "primitive" atmosphere, containing hydrogen gas, ammonia, methane gas, and water vapor. Through these gases, they passed a spark to simulate lightning, then cooled the system so the gases would condense and collect in a watery solution, or "ocean" (Figure 3.1). Within days, the system contained numerous complex molecules, including amino acids, purines, and pyrimidines—some of the building blocks of life.



3.1 Synthesis of Prebiotic Molecules in an Experimental Atmosphere The Miller-Urey experiment simulated possible atmospheric conditions on primitive Earth to obtain some of the molecular building blocks of biological systems.

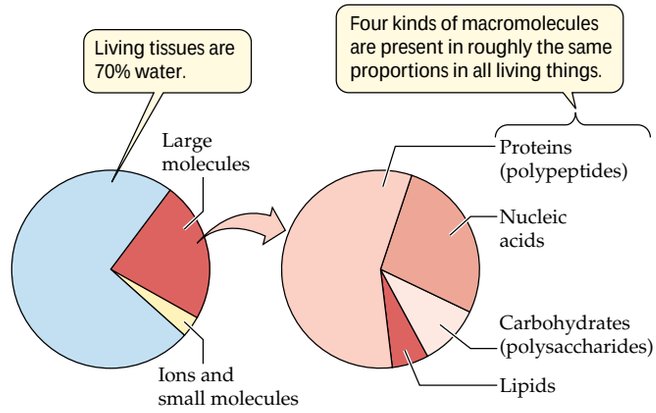
In science, an experiment and its results must be constantly reinterpreted and refined as more knowledge accumulates. The results of the Miller-Urey experiments have undergone several such refinements:

- In living organisms, many molecules have a unique three-dimensional "handedness" (see Figure 2.21). The amino acids, for example, are all in the L-configuration. But the amino acids formed in the Miller-Urey experiments were a mixture of the D- and L-forms. Recent experiments show that natural processes could have selected the L-amino acids from the mixture. Some minerals, especially calcite-based rocks, have unique crystal structures that selectively bind to D- or L-amino acids, separating the two. Such rocks were abundant during the Archean.
- Scientists' views of the Earth's original atmosphere have changed since Miller and Urey did their experiment. There is abundant evidence of major volcanic eruptions 4 billion years ago that released carbon dioxide (CO_2), nitrogen (N_2), hydrogen sulfide (H_2S), and sulfur dioxide (SO_2). Prebiotic chemistry experiments using these molecules in addition to the ones in the original "soup" have led to more diverse molecules.
- Long polymers had to be formed from simpler building blocks, called monomers. Scientists have used model systems to try to simulate conditions under which polymers could be made. Solid mineral surfaces, such as finely divided clays, seem to provide the best environment to bind monomers and allow them to polymerize.
- Miller and Urey, as well as others, suggested that life originated in hot pools at the edges of oceans. Because life has been found in many extreme environments on earth, scientists have proposed that such environments—found beneath ice, in deep-sea hydrothermal vents, and within fine clays near the shore—could be the original site of life's emergence.

In whatever way the earliest stages of chemical evolution occurred, they resulted in the emergence of monomers and polymers that have probably remained unchanged in their general structure and function for 3.8 billion years. We now turn our attention to these large molecules.

Macromolecules: Giant Polymers

The four kinds of large molecules are made the same way and they are present in roughly the same proportions in all living organisms (Figure 3.2). A protein that has a certain role in an apple tree probably has a similar role in a human being, because their basic chemistry is the same. One important advantage of this *biochemical unity* is that organisms acquire needed biochemicals by eating other organisms. When you eat an apple, the molecules you take in include carbohydrates, lipids, and proteins that can be refashioned into the special varieties of those molecules used by humans.



3.2 Substances Found in Living Tissues The substances shown here make up the nonmineral components of living tissue (bone would be an example of a "mineral tissue"). Most tissues are at least 70 percent water.

Macromolecules are giant **polymers** (*poly-*, "many"; *-mer*, "unit") constructed by the covalent linking of smaller molecules called **monomers** (Table 3.1). These monomers may or may not be identical, but they always have similar chemical structures. Molecules with molecular weights exceeding 1,000 are usually considered macromolecules, and the proteins, polysaccharides (large carbohydrates), and nucleic acids of living systems certainly fall into this category.

Each type of macromolecule performs some combination of functions: energy storage, structural support, protection, catalysis, transport, defense, regulation, movement, and information storage. These roles are not necessarily exclusive. For example, both carbohydrates and proteins can play structural roles, supporting and protecting tissues and organisms. However, only nucleic acids specialize in information storage and function as hereditary material, carrying both species and individual traits from generation to generation.

3.1 The Building Blocks of Organisms

MONOMER	SIMPLE POLYMER	COMPLEX POLYMER (MACROMOLECULE)
Amino acid	Peptide or oligopeptide	Polypeptide (protein)
Nucleotide	Oligonucleotide	Nucleic acid
Monosaccharide (sugar)	Oligosaccharide	Polysaccharide (carbohydrate)

The functions of macromolecules are directly related to their shapes and to the sequences and chemical properties of their monomers. Some macromolecules fold into compact spherical forms with surface features that make them water-soluble and capable of intimate interaction with other molecules. Other proteins and carbohydrates form long, fibrous systems that provide strength and rigidity to cells and organisms. Still other long, thin assemblies of proteins can contract and cause movement.

Because macromolecules are so large, they contain many different functional groups (see Figure 2.20). For example, a large protein may contain hydrophobic, polar, and charged functional groups that give specific properties to local sites on the macromolecule. As we will see, this diversity of properties determines the shapes of macromolecules and their interactions with both other macromolecules and smaller molecules.

Condensation and Hydrolysis Reactions

Polymers are constructed from monomers by a series of reactions called **condensation reactions** or dehydration reactions (both terms refer to the loss of water). Condensation reactions result in covalently bonded monomers (Figure 3.3*a*) and release a molecule of water for each bond formed. The condensation reactions that produce the different kinds of polymers differ in detail, but in all cases, polymers form only if energy is added to the system. In living systems, specific energy-rich molecules supply this energy.

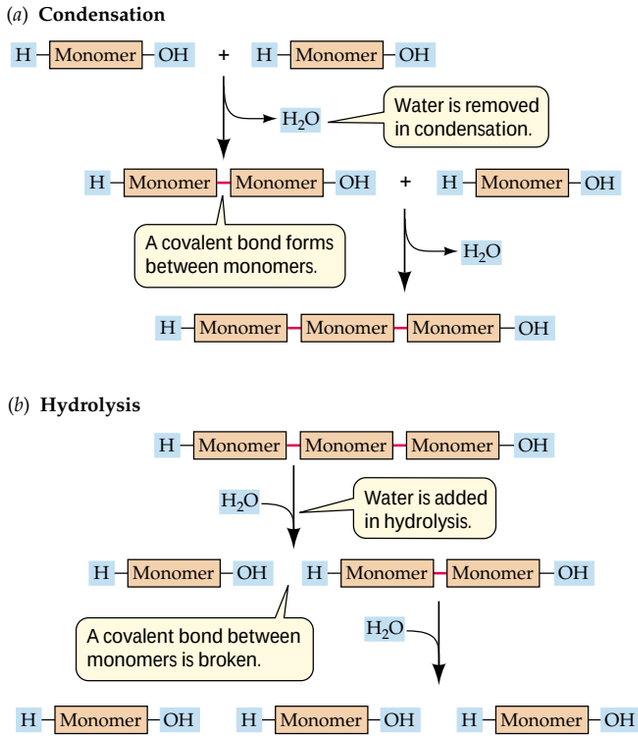
The reverse of a condensation reaction is a **hydrolysis reaction** (*hydro-*, "water"; *-lysis*, "break"). Hydrolysis reactions digest polymers and produce monomers. Water reacts with the bonds that link the polymer together, and the products are free monomers. The elements (H and O) of H_2O become part of the products (Figure 3.3*b*).

These two types of reactions are universal in living things, and as we have seen, were an important step in the origin of life in an aqueous environment. We begin our study of biological macromolecules with a very diverse group of polymers, the proteins.

Proteins: Polymers of Amino Acids

The functions of **proteins** include structural support, protection, transport, catalysis, defense, regulation, and movement. Among the functions of macromolecules listed earlier, only energy storage and information storage are not usually performed by proteins.

Proteins range in size from small ones such as the RNA-digesting enzyme ribonuclease A, which has a molecular weight of 5,733 and 51 amino acid residues, to huge molecules such as the cholesterol transport protein apolipoprotein B, which has a molecular weight of 513,000 and 4,636 amino



3.3 Condensation and Hydrolysis of Polymers (*a*) A condensation reaction links monomers into polymers. (*b*) A hydrolysis reaction digests polymers into individual monomers.

acid residues. (The word *residue* refers to a monomer when it is part of a polymer.) Each of these proteins consists of a single unbranched polymer of amino acids (a *polypeptide chain*), which is folded into a specific three-dimensional shape.

Many proteins require more than one polypeptide chain to make up the functional unit. For example, the oxygen-carrying protein hemoglobin has four chains that are folded separately and associate together to make the functional protein. As we will see later in this book, numerous functional proteins can associate, forming "multi-protein machines" to carry out complex roles such as DNA synthesis.

The *composition* of a protein refers to the relative amounts of the different amino acids it contains. Not every protein contains all kinds of amino acids, nor an equal number of different ones. The diversity in amino acid content and sequence is the source of the diversity in protein structures and functions.

The next several chapters will describe the many functions of proteins. To understand them, we must first explore protein structure. First we will examine the properties of amino acids and see how they link together to form proteins. Then we will systematically examine protein structure and look at how a linear chain of amino acids is consistently folded into a compact three-dimensional shape. Finally, we will see how this three-dimensional structure provides a specific physical and chemical environment that influences how other molecules can interact with the protein.

Proteins are composed of amino acids

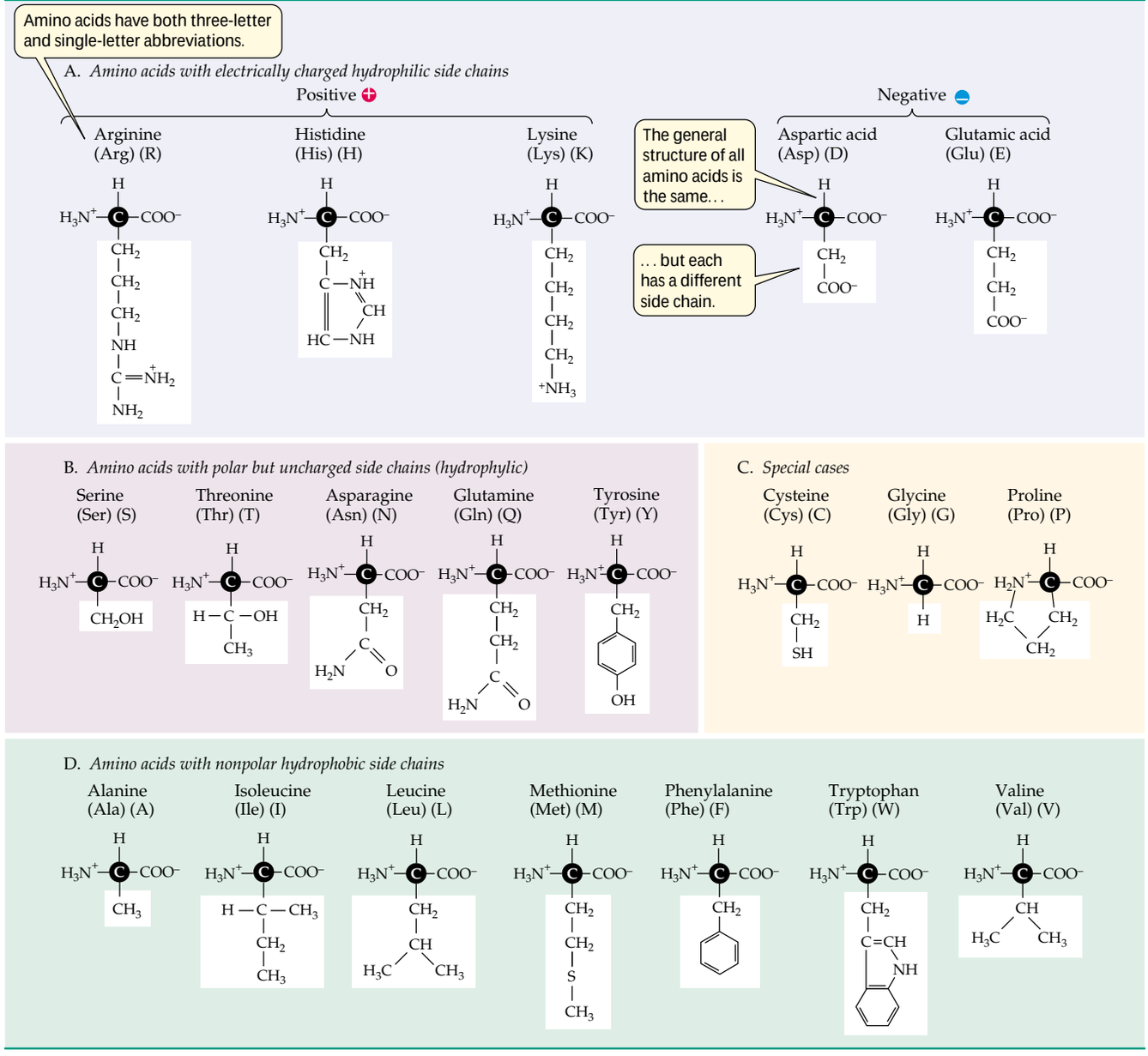
In Chapter 2, we looked at the structure of amino acids and identified four different groups attached to a central (α) carbon atom: a hydrogen atom, an amino group (NH_3^+), a carboxyl group (COO^-), and a unique **side chain**, or **R group**.

The R groups of amino acids are important in determining the three-dimensional structure and function of the protein macromolecule. They are highlighted in white in Table 3.2.

As Table 3.2 shows, amino acids are grouped and distinguished by their side chains. Some side chains are electrically charged (+1, –1), while others are polar (δ^+, δ^-), and still others are nonpolar and hydrophobic.

- The five amino acids that have electrically charged side chains attract water (are hydrophilic) and oppositely charged ions of all sorts.
- The five amino acids that have polar side chains tend to form weak hydrogen bonds with water and with other polar or charged substances. These amino acids are hydrophilic.

3.2 *The Twenty Amino Acids Found in Proteins*

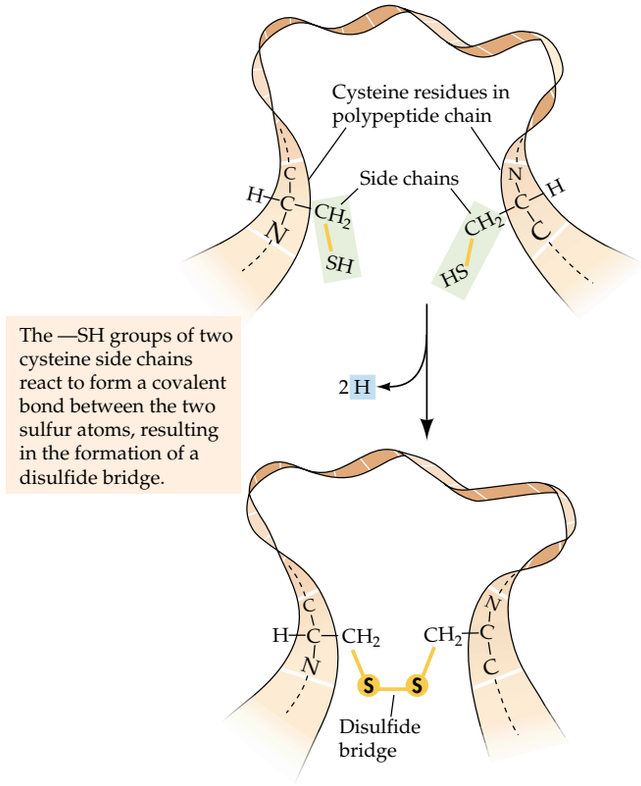


- Seven amino acids have side chains that are nonpolar hydrocarbons or very slightly modified hydrocarbons. In the watery environment of the cell, these hydrophobic side chains may cluster together in the interior of the protein. These amino acids are hydrophobic.
- Three amino acids—cysteine, glycine, and proline—are special cases, although their R groups are generally hydrophobic.

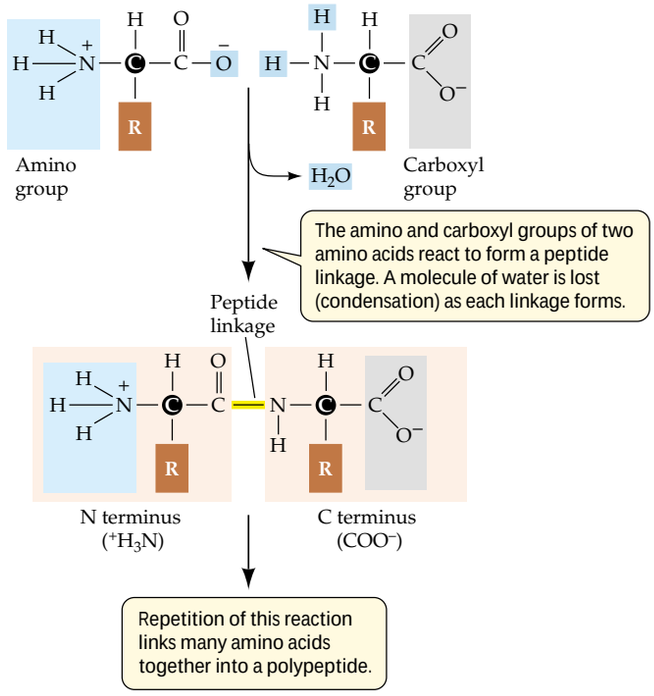
The *cysteine* side chain, which has a terminal —SH group, can react with another cysteine side chain to form a covalent bond called a **disulfide bridge** (—S—S—) (Figure 3.4). Disulfide bridges help determine how a polypeptide chain folds. When cysteine is not part of a disulfide bridge, its side chain is hydrophobic.

The *glycine* side chain consists of a single hydrogen atom and is small enough to fit into tight corners in the interior of a protein molecule, where a larger side chain could not fit.

Proline differs from other amino acids because it possesses a modified amino group lacking a hydrogen on its nitrogen, which limits its hydrogen-bonding ability. Also, the ring system of proline limits rotation about its α carbon, so proline is often found at bends or loops in a protein.



3.4 A Disulfide Bridge Disulfide bridges (—S—S—) are important in maintaining the proper three-dimensional shapes of some protein molecules.



3.5 Formation of Peptide Linkages In living things, the reaction leading to a peptide linkage has many intermediate steps, but the reactants and products are the same as those shown in this simplified diagram.

Peptide linkages covalently bond amino acids together

When amino acids polymerize, the carboxyl and amino groups attached to the α carbon are the reactive groups. The carboxyl group of one amino acid reacts with the amino group of another, undergoing a condensation reaction that forms a *peptide linkage*. Figure 3.5 gives a simplified description of this reaction. (In living systems, other molecules must activate the amino acids in order for this reaction to proceed, and there are intermediate steps in the process. We will examine these steps in Chapter 12.)

Just as a sentence begins with a capital letter and ends with a period, polypeptide chains have a linear order. The chemical "capital letter" marking the beginning of a polypeptide is the amino group of the first amino acid in the chain and is known as the *N terminus*. The "punctuation mark" for the end of the chain is the carboxyl group of the last amino acid—the *C terminus*. All the other amino and carboxyl groups in the chain (except those in side chains) are involved in peptide bond formation, so they do not exist in the chain

as "free," intact groups. Biochemists refer to the "N → C," or "amino-to-carboxyl" orientation of polypeptides.

The peptide linkage has two characteristics that are important in the three-dimensional structure of proteins:

- Unlike many single covalent bonds, in which the groups on either side of the bond are free to rotate in space, the C—N peptide linkage is relatively inflexible. The adjacent atoms (the α carbons of the two adjacent amino acids) are not free to rotate because of the partial double-bond character of the peptide bond. This characteristic limits the folding of the polypeptide chain.
- The oxygen bound to the carbon (C—O) in the carboxyl group carries a slight negative charge (δ^-), whereas the hydrogen bound to the nitrogen (N—H) in the amino group is slightly positive (δ^+). This asymmetry of charge favors hydrogen bonding within the protein molecule itself and with other molecules, contributing to both the structure and the function of many proteins. Before we explore the significance of such hydrogen bonds, we need to examine the importance of the order of amino acids.

The primary structure of a protein is its amino acid sequence

There are four levels of protein structure, called primary, secondary, tertiary, and quaternary. The precise sequence of amino acids in a polypeptide constitutes the **primary structure** of a protein (Figure 3.6*a*). The peptide backbone of this primary structure consists of a repeating sequence of three atoms (—N—C—C—): the N from the amino group, the α carbon, and the C from the carboxyl group of each amino acid.

Scientists have deduced the primary structure of many proteins. The single-letter abbreviations for amino acids (see Table 3.2) are used to record the amino acid sequence of a protein. Here, for example, are the first 20 amino acids (out of a total of 124) in the protein ribonuclease from a cow:

KETAAAKFERQHMDSSTSAA

The theoretical number of different proteins is enormous. Since there are 20 different amino acids, there could be $20 \times 20 = 400$ distinct dipeptides (two linked amino acids), and $20 \times 20 \times 20 = 8{,}000$ different tripeptides (three linked amino acids). Imagine this process of multiplying by 20 extended to a protein made up of 100 amino acids (which is considered a small protein). There could be 20^{100} such small proteins, each with its own distinctive primary structure. How large is the number 20^{100}? There aren't that many electrons in the entire universe!

At the higher levels of protein structure, local coiling and folding give the molecule its final functional shape, but all of these levels derive from the primary structure—that is, the precise location of specific amino acids in the polypeptide chain. The properties associated with a precise sequence of amino acids determine how the protein can twist and fold, thus adopting a specific stable structure that distinguishes it from every other protein.

Primary structure is determined by covalent bonds. But the next level of protein structure is determined by weaker hydrogen bonds.

The secondary structure of a protein requires hydrogen bonding

A protein's **secondary structure** consists of regular, repeated patterns in different regions of a polypeptide chain. There are two basic types of secondary structure, both of them determined by hydrogen bonding between the amino acid residues that make up the primary structure.

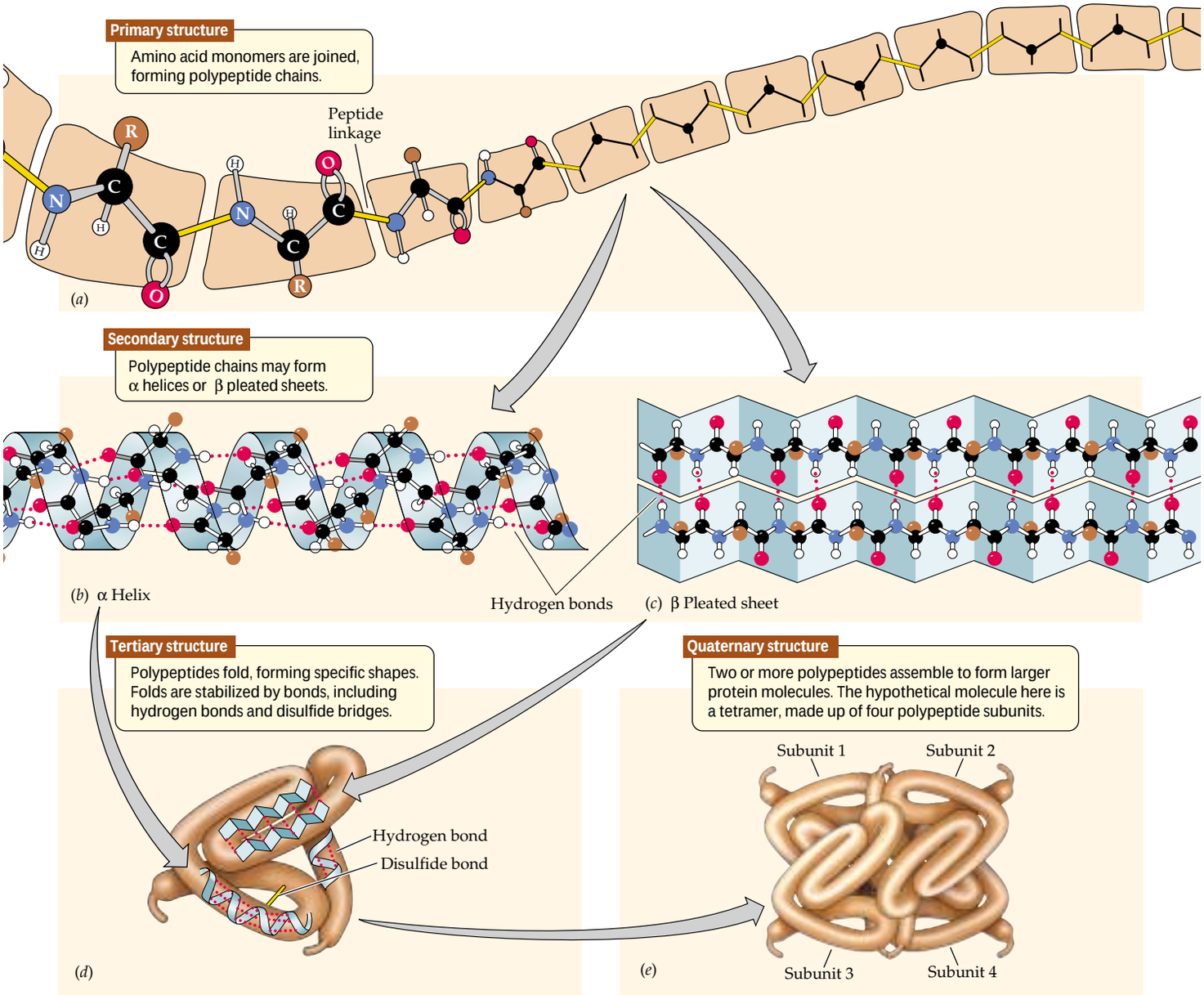
THE α HELIX. The **α (alpha) helix** is a right-handed coil that is "threaded" in the same direction as a standard wood screw (Figure 3.6*b*). The R groups extend outward from the peptide backbone of the helix. The coiling results from hydrogen bonds that form between the δ^+ hydrogen of the N—H of one amino acid residue and the δ^- oxygen of the C═O of another. When this pattern of hydrogen bonding is established repeatedly over a segment of the protein, it stabilizes the coil, resulting in an α helix. The presence of amino acids with large R groups that distort the coil or otherwise prevent the formation of the necessary hydrogen bonds will keep an α helix from forming.

The α-helical secondary structure is common in the fibrous structural proteins called keratins, which make up hair, hooves, and feathers. Hair can be stretched because stretching requires that only the hydrogen bonds of the α helix, not the covalent bonds, be broken; when the tension on the hair is released, both the hydrogen bonds and the helix re-form.

THE β PLEATED SHEET. A **β (beta) pleated sheet** is formed from two or more polypeptide chains that are almost completely extended and lying next to one another. The sheet is stabilized by hydrogen bonds between the N—H groups on one chain and the C═O groups on the other (Figure 3.6*c*). A β pleated sheet may form between separate polypeptide chains, as in spider silk, or between different regions of the same polypeptide chain that is bent back on itself. Many proteins contain regions of both α helix and β pleated sheet in the same polypeptide chain.

The tertiary structure of a protein is formed by bending and folding

In many proteins, the polypeptide chain is bent at specific sites and then folded back and forth, resulting in the **tertiary**



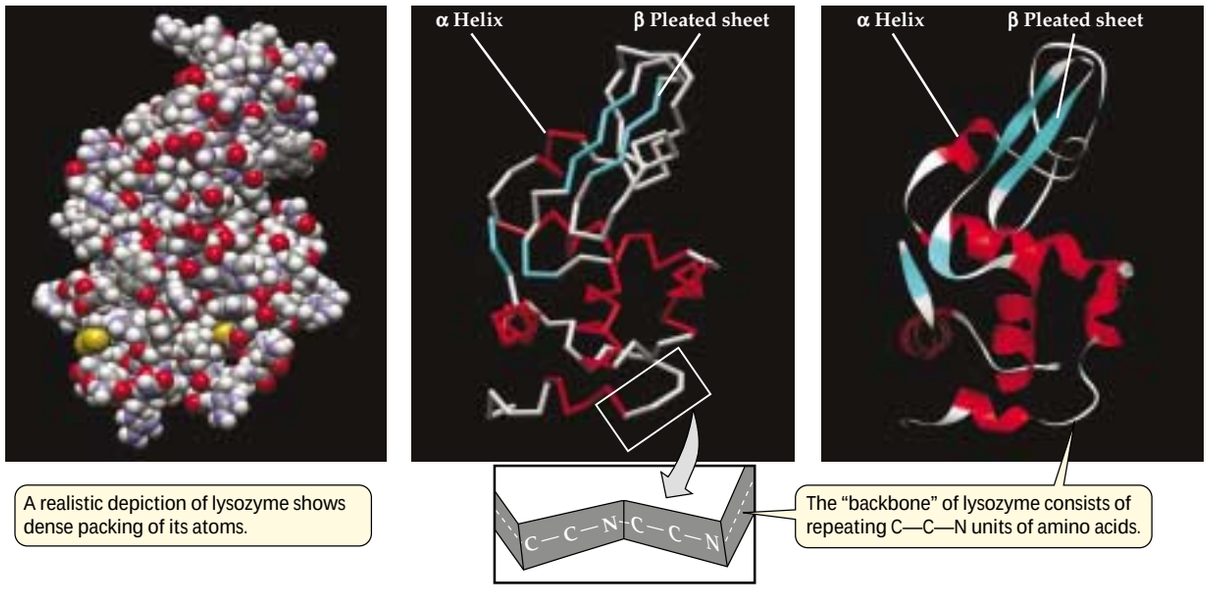
3.6 The Four Levels of Protein Structure Secondary, tertiary, and quaternary structure all arise from the primary structure of the protein.

structure of the protein (Figure 3.6*d*). Although the α helices and β pleated sheets contribute to the tertiary structure, only parts of the macromolecule usually have these secondary structures, and large regions consist of structures unique to a particular protein.

While hydrogen bonding between the N—H and C═O groups within and between chains is responsible for secondary structure, the interactions between R groups—the amino acid side chains—determine tertiary structure. We described the various strong and weak interactions between atoms in Chapter 2 (see Table 2.1). Many of these interactions are involved in determining tertiary structure.

- Covalent *disulfide bridges* can form between specific cysteine residues (see Figure 3.4), holding a folded polypeptide in place.
- *Hydrophobic* side chains can aggregate together in the interior of the protein, away from water, folding the polypeptide in the process.
- *van der Waals forces* can stabilize the close interactions between the hydrophobic residues.
- *Ionic bonds* can form between positively and negatively charged side chains buried deep within a protein, away from water, forming a *salt bridge*.

A complete description of a protein's tertiary structure specifies the location of every atom in the molecule in three-dimen-



3.7 Three Representations of Lysozyme Different molecular representations of a protein emphasize different aspects of its tertiary structure. These three representations of lysozyme are similarly oriented.

sional space in relation to all the other atoms. Such a description is available for the protein lysozyme (Figure 3.7). The first tertiary structures to be determined took years to figure out, but today, dozens of new structures are published every week. The major advances making this possible have been the ability to produce large quantities of specific proteins by biotechnology and the use of computers to analyze the atomic data.

Bear in mind that both tertiary structure and secondary structure derive from a protein's primary structure. If lysozyme is heated slowly, the heat energy will disrupt only the weak interactions and cause only the tertiary structure to break down. But the protein will return to its normal tertiary structure when it cools, demonstrating that all the information needed to specify the unique shape of a protein is contained in its primary structure.

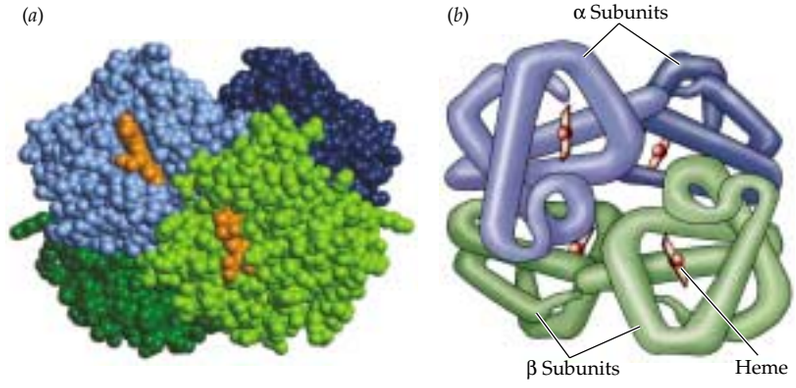
The quaternary structure of a protein consists of subunits

As mentioned earlier, many functional proteins contain two or more polypeptide chains, called *subunits*, each of them folded into its own unique tertiary structure. The protein's **quaternary structure** results from the ways in which these subunits bind together and interact (see Figure 3.6*e*).

Quaternary structure is illustrated by hemoglobin (Figure 3.8). Hydrophobic interactions, van der Waals forces, hydrogen bonds, and ionic bonds all help hold the four subunits together to form the hemoglobin molecule. The function of hemoglobin is to carry oxygen in red blood cells. As hemoglobin binds one O_2 molecule, the four subunits shift their relative positions slightly, changing the quaternary structure. Ionic bonds are broken, exposing buried side chains that enhance the binding of additional O_2 molecules. The structure changes again when hemoglobin releases its O_2 molecules to the cells of the body.

The surfaces of proteins have specific shapes

Small molecules in a solution are in constant motion. They vibrate, rotate, and move from place to place like corn in a



3.8 Quaternary Structure of a Protein Hemoglobin consists of four folded polypeptide subunits that assemble themselves into the quaternary structure shown here. In these two graphic representations, each type of subunit is a different color. The heme groups contain iron and are the oxygen-carrying sites.

popper. If two of them collide in the right circumstances, a chemical reaction can occur. The specific shapes of proteins allow them to bind *noncovalently* to other molecules, which in turn allows other important biological events to occur. Here are just a few examples:

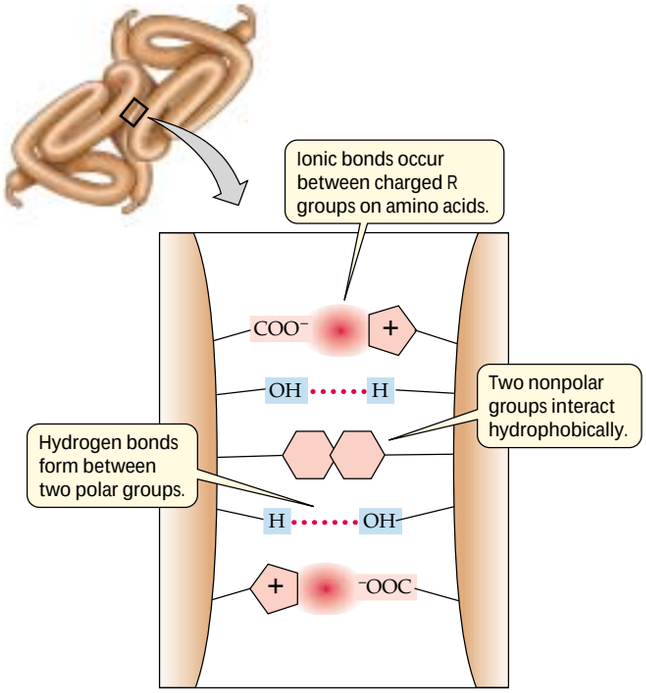
- Two adjacent cells can stick together because proteins protruding from each of the cells interact with each other (see Chapter 5).
- A substance can enter a cell by binding to a carrier protein in the cell surface membrane (see Chapter 5).
- A chemical reaction can be speeded up when an enzyme protein binds to one of the reactants (see Chapter 6).
- A "multi-protein machine," DNA polymerase, can bind to and copy DNA (see Chapter 11).
- Another "multi-protein machine," RNA polymerase, can synthesize RNA (see Chapter 12).
- Chemical signals such as hormones can bind to proteins on a cell's outer surface (see Chapter 15).
- Defensive proteins called antibodies can recognize the shape of a virus coat and bind to it (see Chapter 18).

The biological specificity of protein function depends on two general properties of the protein: its shape and the chemistry of its exposed surface groups.

- *Shape*. When a molecule collides with and binds to a much larger protein, it is like a baseball being caught by a catcher's mitt: The mitt has a shape that binds to the ball and fits around it. A hockey puck or a ping-pong ball would not fit a baseball catcher's mitt. The binding of a molecule to a protein involves a general "fit" between two three-dimensional objects that becomes even more specific after initial binding.
- *Chemistry*. The surface of a protein has certain chemical groups that it presents to a substance attempting to bind to it (Figure 3.9). These groups are the R groups of the exposed amino acids, and are therefore a property of the protein's primary structure.

Look again at the structures of the 20 amino acids in Table 3.2, noting the properties of the R groups. Exposed hydrophobic groups can bind to similarly nonpolar groups in the substance with which the protein interacts (often called the **ligand**). Charged R groups can bind to oppositely charged groups on the ligand. Polar R groups containing a hydroxyl (—OH) group can form a hydrogen bond with the ligand. These three types of interactions—hydrophobic, ionic, and hydrogen bonding—are weak by themselves, but strong when all of them act together. So the exposure of appropriate amino acid R groups on the protein surface allows the binding of a specific ligand to occur.

Knowing the exact shape of a protein and what can bind to it is important not only in understanding basic biology,



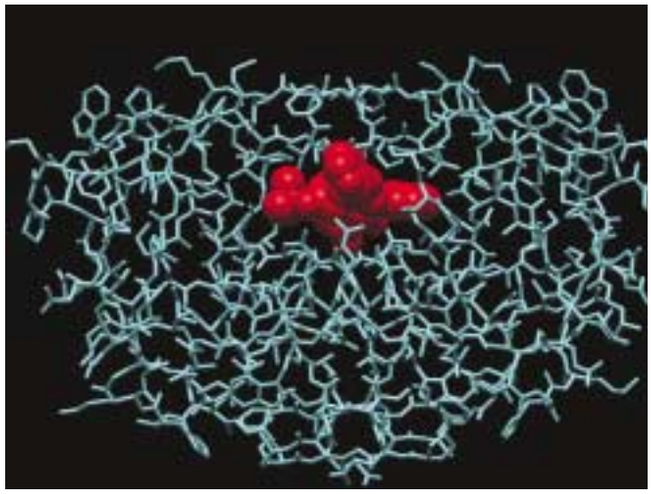
3.9 Noncovalent Interactions between Proteins and Other Molecules Noncovalent interactions allow a protein to bind tightly to another molecule with specific properties, or allow regions within a protein to interact with one another.

but also in applied fields such as medicine. For example, the three-dimensional structure of a protease, a protein essential for the replication of HIV—the virus that causes AIDS—was first determined, then specific proteins were designed to bind to it and block its action. These protease inhibitors have prolonged the lives of countless people living with HIV (Figure 3.10).

Protein shapes are sensitive to the environment

Because it is determined by weak forces, protein shape is sensitive to environmental conditions that would not break covalent bonds, but do upset the weaker noncovalent interactions that determine secondary and tertiary structure.

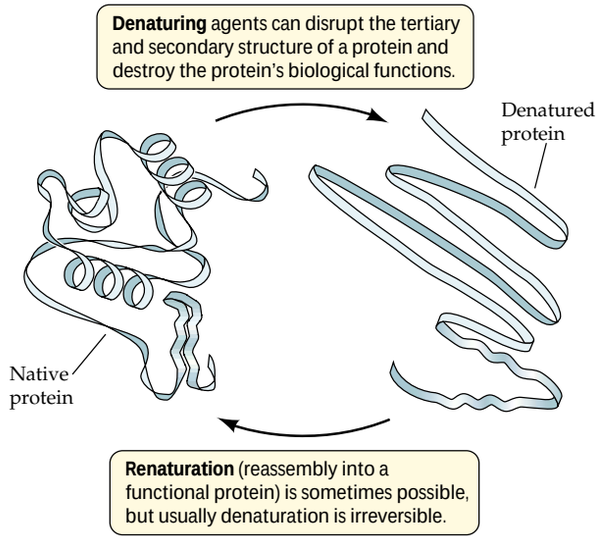
- Increases in *temperature* cause more rapid molecular movements and thus can break hydrogen bonds and hydrophobic interactions.
- Alterations in *pH* can change the pattern of ionization of carboxyl and amino groups in the R groups of amino acids, thus disrupting the pattern of ionic attractions and repulsions.
- High concentrations of polar substances such as urea can disrupt the hydrogen bonding that is crucial to protein



3.10 An HIV Protease Inhibitor After determining the structure of HIV protease (the blue polypeptide chain), a protein essential to the life cycle of HIV, biochemists designed a drug (the red space-filling model) to fit into the protease and block its function. Many people living with HIV and AIDS now take this drug.

structure. Nonpolar solvents may also disrupt normal protein structure.

The loss of a protein's normal three-dimensional structure is called **denaturation**, and it is always accompanied by a loss of the normal biological function of the protein (Figure 3.11).



3.11 Denaturation Is the Loss of Tertiary Protein Structure and Function Agents that can cause denaturation include high temperatures and certain chemicals.

Denaturation is often irreversible, because amino acids that were buried may now be exposed at the surface, and vice versa, causing a new structure to form or different molecules to bind to the protein. Boiling an egg denatures its proteins and is, as you know, not reversible. However, as we saw earlier in the case of lysozyme, denaturation may be reversible in the laboratory. If the protein is allowed to cool or the denaturing chemicals are removed, the protein may return to its "native" shape and normal function.

Chaperonins help shape proteins

There are two occasions when a polypeptide chain is in danger of binding the wrong ligand. First, following denaturation, hydrophobic R groups, previously on the inside of the protein away from water, become exposed on the surface. Since these groups can interact with similar groups on other molecules, the denatured proteins may aggregate and become insoluble, losing their function. Second, when a protein has just been made and has not yet folded completely, it can present a surface that binds the wrong molecule. In the cell, a protein can sometimes fold incorrectly as it is made. This can have serious consequences: In Alzheimer's disease, misfolded proteins accumulate in the brain and bind to one anther, forming fibers in the areas of the brain that control memory, mood, and spatial awareness.

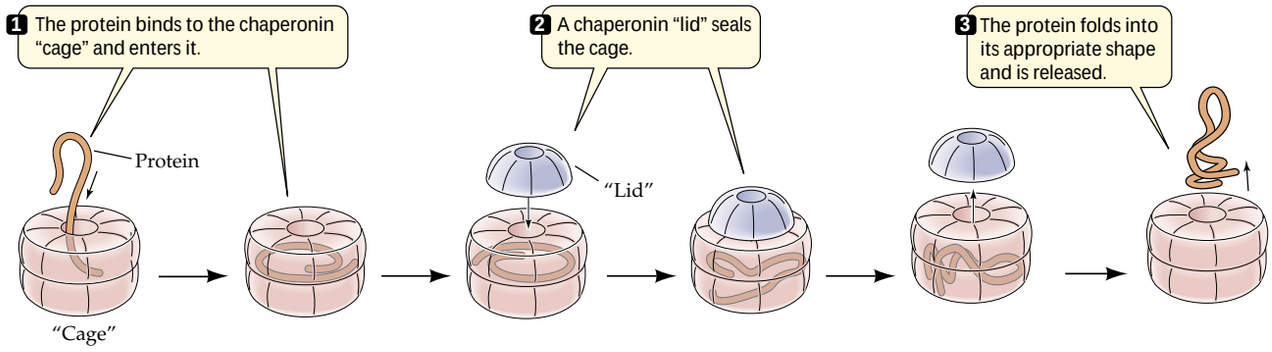
Living systems limit inappropriate protein interactions by making a class of proteins called, appropriately, **chaperonins** (recall the chaperones—usually teachers—at school dances who try to prevent "inappropriate interactions" among the students). Chaperonins were first identified in fruit flies as "heat shock" proteins, which prevented denaturing proteins from clumping together when the flies' temperatures were raised.

Some chaperonins work by trapping proteins in danger of inappropriate binding inside a molecular "cage" (Figure 3.12). This cage is composed of several identical subunits, and is itself a good example of quaternary protein structure. Inside the cage, the targeted protein folds into the right shape, and then is released at the appropriate time and place.

Carbohydrates: Sugars and Sugar Polymers

The second class of biological molecules, the **carbohydrates**, is a diverse group of compounds. Carbohydrates contain primarily carbon atoms flanked by hydrogen atoms and hydroxyl groups (H—C—OH). They have two major biochemical roles:

- They act as a source of energy that can be released in a form usable by body tissues.



3.12 Chaperonins Protect Proteins from Inappropriate Binding
Chaperonins surround new or denatured proteins and prevent them from binding to the wrong ligand.

- They serve as carbon skeletons that can be rearranged to form other molecules that are essential for biological structures and functions.

Some carbohydrates are relatively small, with molecular weights of less than 100. Others are true macromolecules, with molecular weights in the hundreds of thousands.

There are four categories of biologically important carbohydrates, which we will discuss in turn:

- *Monosaccharides* (*mono-*, "one"; *saccharide*, "sugar"), such as glucose, ribose, and fructose, are simple sugars. They are the monomers out of which the larger carbohydrates are constructed.
- *Disaccharides* (*di-*, "two") consist of two monosaccharides linked together by covalent bonds.
- *Oligosaccharides* (*oligo-*, "several") are made up of several (3 to 20) monosaccharides.
- *Polysaccharides* (*poly-*, "many"), such as starch, glycogen, and cellulose, are large polymers composed of hundreds or thousands of monosaccharides.

The general formula for carbohydrates, CH_2O, gives the relative proportions of carbon, hydrogen, and oxygen in a monosaccharide (i.e., the proportions of these atoms are 1:2:1). In disaccharides, oligosaccharides, and polysaccharides, these proportions differ slightly from the general formula because two hydrogens and an oxygen are lost during each of the condensation reactions that form them.

Monosaccharides are simple sugars

Green plants produce monosaccharides through photosynthesis, and animals acquire them directly or indirectly from plants. All living cells contain the monosaccharide **glucose**. Cells use glucose as an energy source, breaking it down through a series of reactions that release stored energy and produce water and carbon dioxide.

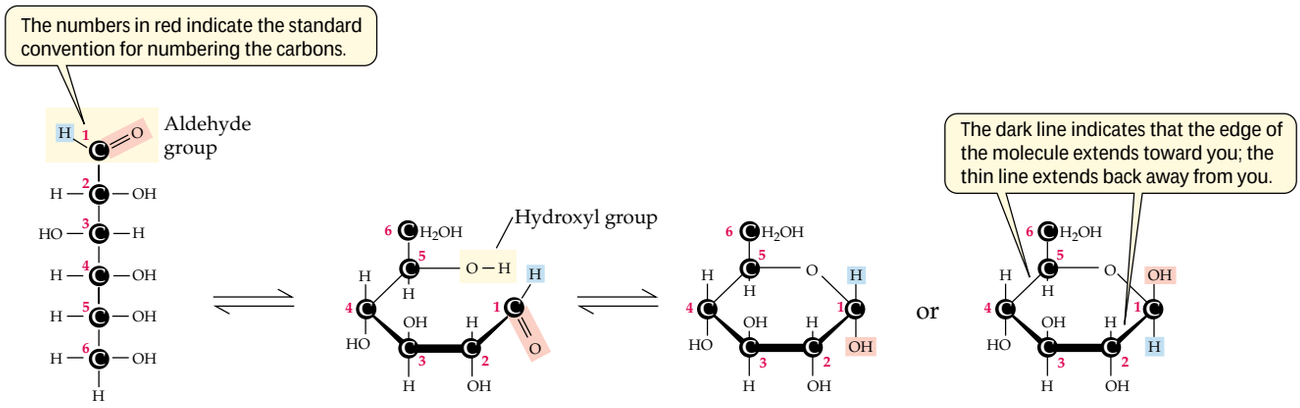
Glucose exists in two forms, the straight chain and the ring. The ring form predominates in more than 99 percent of circumstances because it is more stable under cellular conditions. There are two forms of the ring structure (α-glucose and β-glucose), which differ only in the placement of the —H and —OH attached to carbon 1 (Figure 3.13). The α and β forms interconvert and exist in equilibrium when dissolved in water.

Different monosaccharides contain different numbers of carbons. (The standard convention for numbering carbons in carbohydrates shown in Figure 3.13 is used throughout this book.) Most of the monosaccharides found in living systems belong to the D series of optical isomers (see Chapter 2). But some monosaccharides are structural isomers, which have the same kinds and numbers of atoms, but arranged differently. For example, the *hexoses* (*hex-*, "six"), a group of structural isomers, all have the formula $C_6H_{12}O_6$. Included among the hexoses are glucose, fructose (so named because it was first found in fruits), mannose, and galactose (Figure 3.14).

Pentoses (*pent-*, "five") are five-carbon sugars. Some pentoses are found primarily in the cell walls of plants. Two pentoses are of particular biological importance: Ribose and deoxyribose form part of the backbones of the nucleic acids RNA and DNA, respectively. These two pentoses are not isomers; rather, one oxygen atom is missing from carbon 2 in deoxyribose (*de-*, "absent") (see Figure 3.14). As we will see in Chapter 12, the absence of this oxygen atom has important consequences for the functional distinction of RNA and DNA.

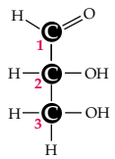
Glycosidic linkages bond monosaccharides together

The disaccharides and polysaccharides described above are all constructed from monosaccharides that are covalently bonded together by condensation reactions that form *glycosidic linkages*. One such linkage between two monosaccharides forms a disaccharide. For example, a molecule of su-



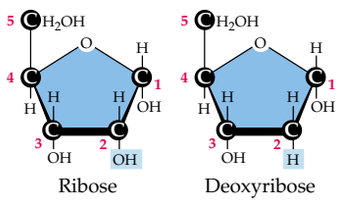
3.13 Glucose: From One Form to the Other All glucose molecules have the formula $C_6H_{12}O_6$, but their structures vary. When dissolved in water, the α and β "ring" forms of glucose interconvert.

Three-carbon sugar



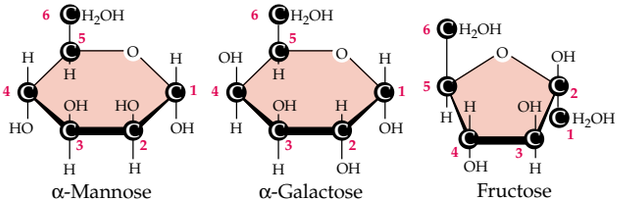
Glyceraldehyde is the smallest monosaccharide and exists only as the straight-chain form.

Five-carbon sugars (pentoses)



Ribose and deoxyribose each have five carbons, but very different chemical properties and biological roles.

Six-carbon sugars (hexoses)



These hexoses are isomers. All have the formula $C_6H_{12}O_6$, but each has distinct chemical properties and biological roles.

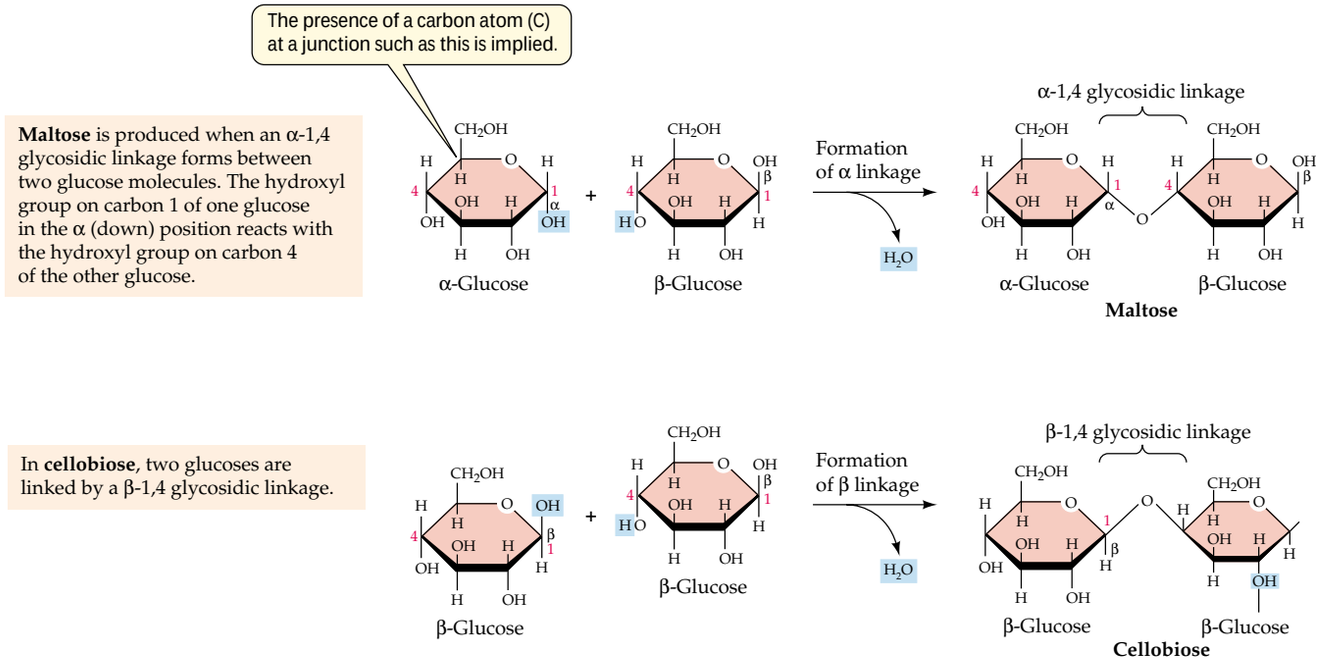
crose (table sugar) is formed from a glucose molecule and a fructose molecule, while lactose (milk sugar) contains glucose and galactose.

The disaccharide maltose contains two glucose molecules, but it is not the only disaccharide that can be made from two glucoses. When glucose molecules form a glycosidic linkage, the linkage will be one of two types, α or β, depending on whether the molecule that bonds its carbon 1 is α-glucose or β-glucose (see Figure 3.13). An α linkage with carbon 4 of a second glucose molecule gives maltose, whereas a β linkage gives cellobiose (Figure 3.15).

Maltose and cellobiose are disaccharide isomers, both having the formula $C_{12}H_{22}O_{11}$. However, they are different compounds with different properties. They undergo different chemical reactions and are recognized by different enzymes. For example, maltose can be hydrolyzed to its monosaccharides in the human body, whereas cellobiose cannot. Certain microorganisms have the chemistry needed to break down cellobiose.

Oligosaccharides contain several monosaccharides bound by glycosidic linkages at various sites. Many oligosaccharides have additional functional groups, which give them special properties. Oligosaccharides are often covalently bonded to proteins and lipids on the outer cell surface, where they serve as cell recognition signals. The human blood groups (such as ABO) get their specificity from oligosaccharide chains.

3.14 Monosaccharides Are Simple Sugars Monosaccharides are made up of varying numbers of carbons. Some are structural isomers, which have the same number of carbons, but arranged differently. Fructose, for example, is a hexose, but forms a five-sided ring like the pentoses.



3.15 Disaccharides Are Formed by Glycosidic Linkages
Glycosidic linkages between two monosaccharides create many different disaccharides. Which disaccharide is formed depends on which monosaccharides are linked, and on the site (which carbon atom is linked) and form (α or β) of the linkage.

Polysaccharides serve as energy stores or structural materials

Polysaccharides are giant polymers of monosaccharides connected by glycosidic linkages (Figure 3.16).

- **Starch** is a polysaccharide of glucose with α-glycosidic linkages.
- **Glycogen** is a highly branched polysaccharide of glucose.
- **Cellulose** is also a polysaccharide of glucose, but its individual monosaccharides are connected by β-glycosidic linkages.

Starch actually comprises a large family of giant molecules of broadly similar structure. While all starches are large polymers of glucose with α linkages (Figure 3.16*a*), the different starches can be distinguished by the amount of branching that occurs at carbons 1 and 6 (Figure 3.16*b*). Some plant starches are unbranched, as in plant amylose; others are moderately branched, as in plant amylopectin. Starch readily binds water, and when that water is removed, unbranched starch tends to form hydrogen bonds between the polysaccharide chains, which then aggregate. This is what causes bread to become hard and stale. Adding water and gentle heat separates the chains and the bread becomes softer.

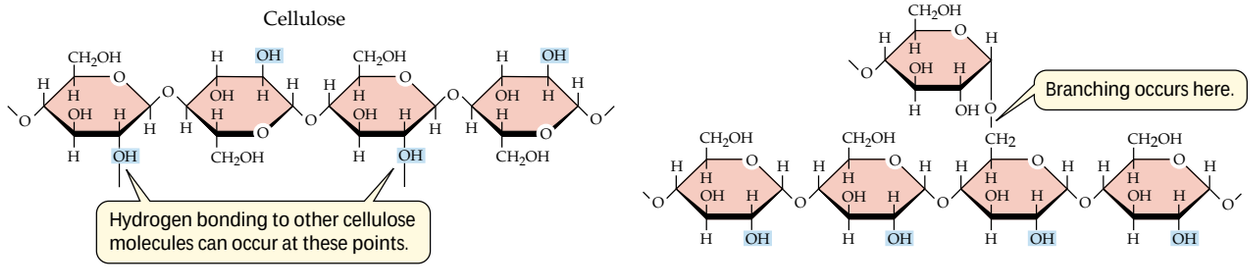
The polysaccharide glycogen stores glucose in animal livers and muscles. Starch and glycogen serve as energy storage compounds for plants and animals, respectively. These polysaccharides are readily hydrolyzed to glucose monomers, which in turn can be further degraded to liberate their stored energy and convert it to forms that can be used for cellular activities. If it is glucose that is actually needed for fuel, why must it be stored as a polymer? The reason is that 1,000 glucose molecules would exert 1,000 times the osmotic pressure (causing water to enter the cells; see Chapter 5) of a single glycogen molecule. If it were not for polysaccharides, many organisms would expend a lot of time and energy expelling excess water.

Cellulose is the predominant component of plant cell walls, and is by far the most abundant **organic** (carbon-containing) compound on Earth. Starch can be easily degraded by the actions of chemicals or enzymes. Cellulose, however, is chemically more stable because of its β-glycosidic linkages (Figure 3.16*a*). Thus starch is a good storage medium that can be easily broken down to supply glucose for energy-producing reactions, while cellulose is an excellent structural material that can withstand harsh environmental conditions without changing.

Chemically modified carbohydrates contain other groups

Some carbohydrates are chemically modified by the addition of functional groups, such as phosphate and amino groups

(*a*) **Molecular structure**

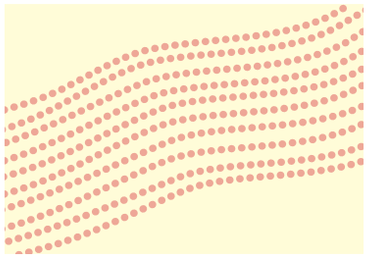


Cellulose is an unbranched polymer of glucose with β-1,4 glycosidic linkages that are chemically very stable.

Glycogen and starch are polymers of glucose with α-1,4 glycosidic linkages. α-1,6 glycosidic linkages produce branching at carbon 6.

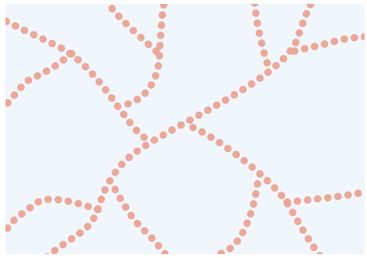
(*b*) **Macromolecular structure**

Linear (cellulose)



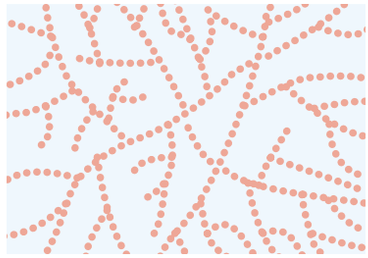
Parallel cellulose molecules from hydrogen-bonds, resulting in thin fibrils.

Branched (starch)



Branching limits the number of hydrogen bonds that can form in starch molecules, making starch less compact than cellulose.

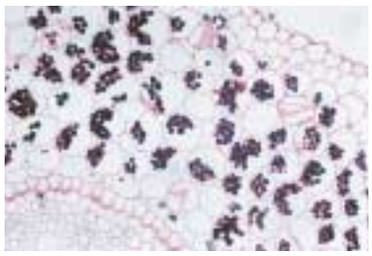
Highly branched (glycogen)



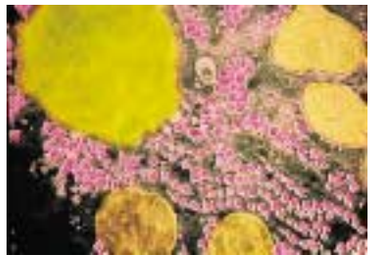
The high amount of branching in glycogen makes its solid deposits more compact than starch.

(*c*) **Polysaccharides in cells**

Layers of cellulose fibrils, as seen in this scanning electron micrograph, give plant cell walls great strength.



Dyed purple in this micrograph, starch deposits have a large granular shape within cells.



Colored pink in this electron micrograph of human liver cells, glycogen deposits have a small granular shape.

3.16 Representative Polysaccharides Cellulose, starch, and glycogen demonstrate different levels of branching and compaction in polysaccharides.

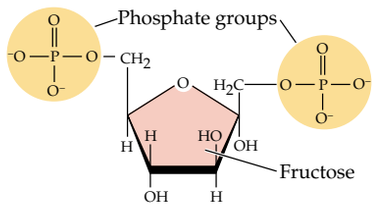
(Figure 3.17). For example, carbon 6 in glucose may be oxidized from —CH_2OH to a carboxyl group (—COOH), producing glucuronic acid. Or a phosphate group may be added to one or more of the —OH sites. Some of the resulting *sugar phosphates*, such as fructose 1,6-bisphosphate, are important intermediates in cellular energy reactions.

When an amino group is substituted for an —OH group, *amino sugars*, such as glucosamine and galactosamine, are produced. These compounds are important in the extracellular matrix, where they form parts of proteins involved in keeping tissues together. Galactosamine is a major component of cartilage, the material that forms caps on the ends of bones and stiffens the protruding parts of the ears and nose. A derivative of glucosamine produces the polymer *chitin*, which is the principal structural polysaccharide in the skele-

3.17 Chemically Modified Carbohydrates Added functional groups modify the form and properties of a carbohydrate.

(*a*) **Sugar phosphate**

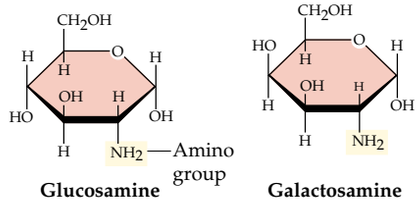
Fructose 1,6 bisphosphate is involved in the reactions that liberate energy from glucose. (The numbers in its name refer to the carbon sites of phosphate bonding; *bis-* indicates that two phosphates are present.)



Fructose 1,6 bisphosphate

(*b*) **Amino sugars**

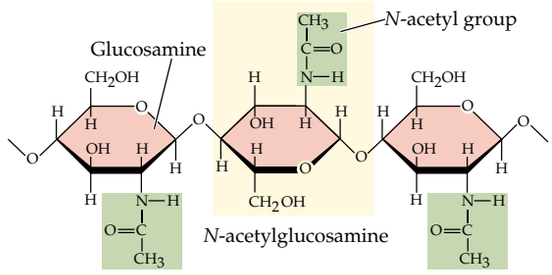
The monosaccharides glucosamine and galactosamine are amino sugars with an amino group in place of a hydroxyl group.



Glucosamine **Galactosamine**

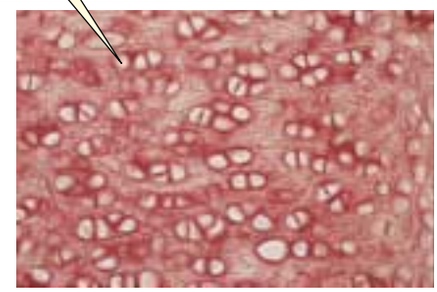
(*c*) **Chitin**

Chitin is a polymer of *N*-acetylglucosamine; *N*-acetyl groups provide additional sites for hydrogen bonding between the polymers.



Chitin

Galactosamine is an important component of cartilage, a connective tissue in vertebrates.



The external skeletons of insects are made up of chitin.

tons of insects, crabs, and lobsters, as well as in the cell walls of fungi. Fungi and insects (and their relatives) constitute more than 80 percent of the species ever described, and so chitin is one of the most abundant substances on Earth.

Lipids: Water-Insoluble Molecules

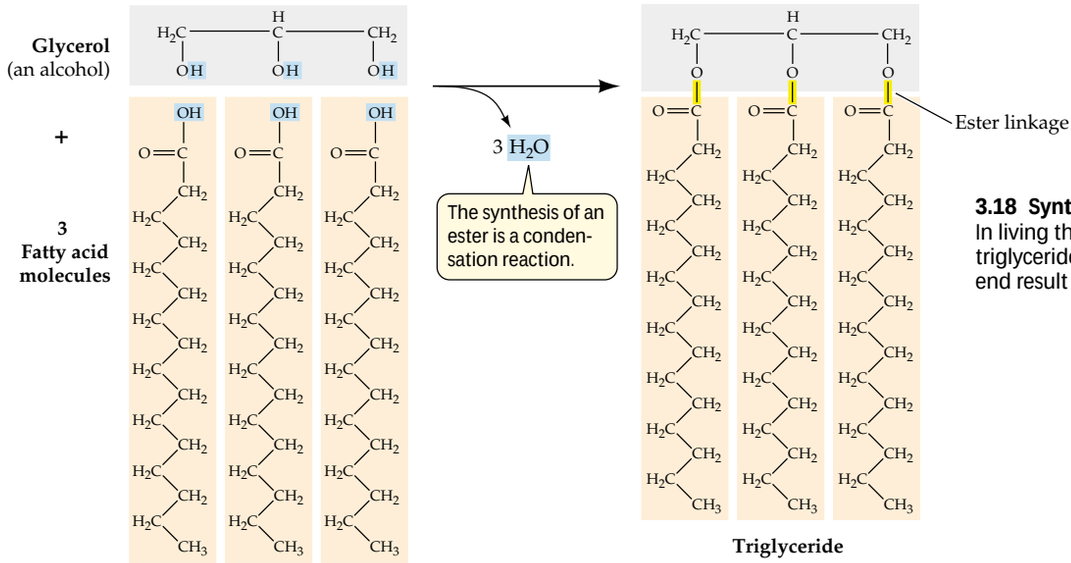
The **lipids** are a chemically diverse group of hydrocarbons. The property they all share is insolubility in water, which is due to the presence of many nonpolar covalent bonds. As we saw in Chapter 2, nonpolar hydrocarbon molecules are hydrophobic and preferentially aggregate among themselves, away from water, which is polar. When these nonpolar molecules are sufficiently close together, weak but additive van der Waals forces hold them together. These huge macromolecular aggregations are not polymers in a strict chemical sense, since their units (lipid molecules) are not held together by covalent bonds, as are, for example, the amino acids in proteins. But they can be considered polymers of individual lipid units.

In this section, we will describe the different types of lipids. Lipids have a number of roles in living organisms:

- Fats and oils store energy.
- Phospholipids play important structural roles in cell membranes.
- The carotenoids help plants capture light energy.
- Steroids and modified fatty acids play regulatory roles as hormones and vitamins.
- The fat in animal bodies serves as thermal insulation.
- A lipid coating around nerves acts as electrical insulation.
- Oil or wax on the surfaces of skin, fur, and feathers repels water.

Fats and oils store energy

Chemically, fats and oils are **triglycerides**, also known as *simple lipids*. Triglycerides that are solid at room temperature (20°C) are called *fats;* those that are liquid at room temperature are called *oils*. Triglycerides are composed of two types of building blocks: fatty acids and glycerol. *Glycerol* is a small molecule with three hydroxyl (—OH) groups (an alcohol). A *fatty acid* is made up of a long nonpolar hydrocarbon chain and a polar carboxyl group (—COOH). A triglyceride contains three fatty acid molecules and one molecule of glycerol.



3.18 Synthesis of a Triglyceride In living things, the reaction that forms triglycerides is more complex, but the end result is as shown here.

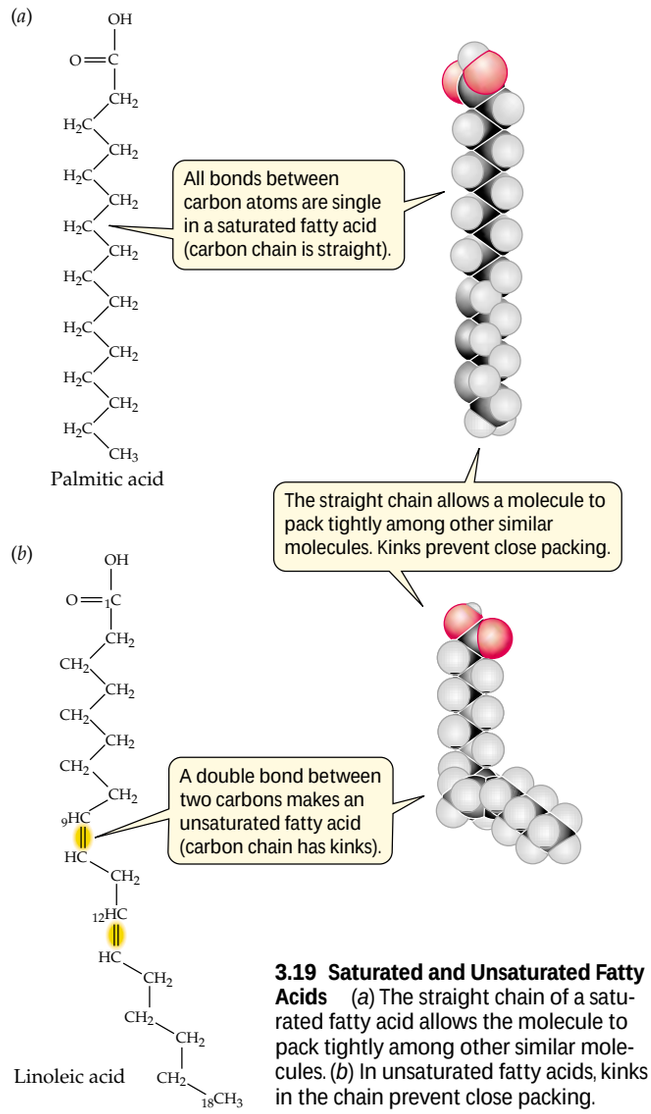
The carboxyl group of a fatty acid can form a covalent bond with the hydroxyl group of glycerol, resulting a functional group called an *ester* and water (Figure 3.18).

The three fatty acids in a triglyceride molecule need not all have the same hydrocarbon chain length or structure:

- In **saturated** fatty acids, all the bonds between the carbon atoms in the hydrocarbon chain are single bonds—there are no double bonds. That is, all the bonds are saturated with hydrogen atoms (Figure 3.19*a*). These fatty acid molecules are relatively rigid and straight, and they pack together tightly, like pencils in a box.
- In **unsaturated** fatty acids, the hydrocarbon chain contains one or more double bonds. Oleic acid, for example, is a *monounsaturated* fatty acid that has one double bond near the middle of the hydrocarbon chain, which causes a kink in the molecule (Figure 3.19*b*). Some fatty acids have more than one double bond—are *polyunsaturated*—and have multiple kinks. These kinks prevent the molecules from packing together tightly.

The kinks in fatty acid molecules are important in determining the fluidity and melting point of a lipid. The triglycerides of animal fats tend to have many long-chain saturated fatty acids, packed tightly together; these fats are usually solids at room temperature and have a high melting point. The triglycerides of plants, such as corn oil, tend to have short or unsaturated fatty acids. Because of their kinks, these fatty acids pack together poorly and have a low melting point, and these triglycerides are usually liquids at room temperature.

Fats and oils are marvelous storehouses for energy. When they take in excess food, many animal species deposit fat droplets in their cells as a means of storing energy. Some plant species, such as olives, avocados, sesame, castor beans, and all nuts, have substantial amounts of lipids in their seeds



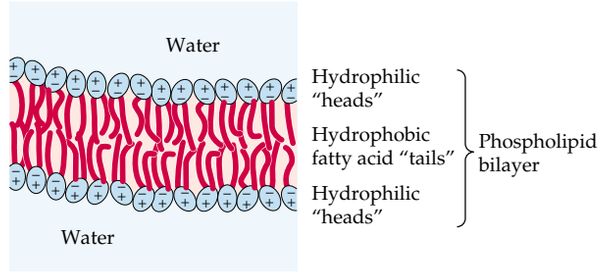
3.19 Saturated and Unsaturated Fatty Acids (*a*) The straight chain of a saturated fatty acid allows the molecule to pack tightly among other similar molecules. (*b*) In unsaturated fatty acids, kinks in the chain prevent close packing.

or fruits that serve as energy reserves for the next generation. This energy can be tapped by people who eat these plant oils or use them for fuel. Indeed, the famous German engineer Rudolf Diesel used peanut oil to power one of his early automobile engines in 1900.

Phospholipids form the core of biological membranes

Because lipids and water do not interact, a mixture of water and lipids forms two distinct layers. Many biologically important substances—such as ions, sugars, and free amino acids—that are soluble in water are insoluble in lipids.

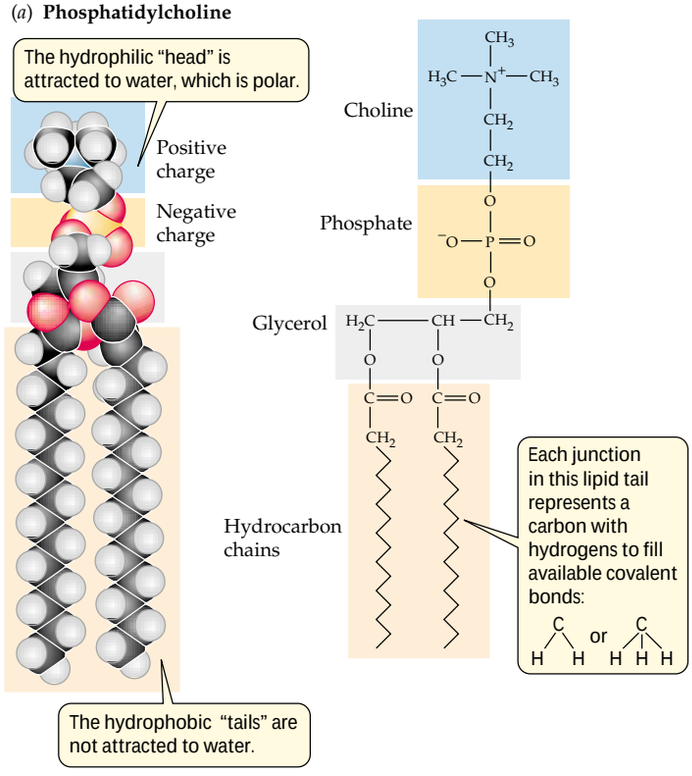
Like triglycerides, **phospholipids** contain fatty acids bound to glycerol by ester linkages. In phospholipids, however, any one of several phosphate-containing compounds replaces one of the fatty acids (Figure 3.20). The phosphate functional group has a negative electric charge, so this portion of the molecule is hydrophilic, attracting polar water molecules. But the two fatty acids are hydrophobic, so they tend to aggregate away from water.



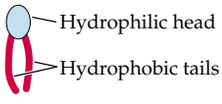
3.21 Phospholipids Form a Bilayer In an aqueous environment, hydrophobic interactions bring the "tails" of phospholipids together in the interior of a phospholipid bilayer. The hydrophilic "heads" face outward on both sides of the bilayer, where they interact with the surrounding water molecules.

In an aqueous environment, phospholipids line up in such a way that the nonpolar, hydrophobic "tails" pack tightly together and the phosphate-containing "heads" face outward, where they interact with water. The phospholipids thus form a *bilayer*, a sheet two molecules thick, with water excluded from the core (Figure 3.21). Biological membranes have this kind of phospholipid bilayer structure, and we will devote all of Chapter 5 to their biological functions.

(*a*) **Phosphatidylcholine**



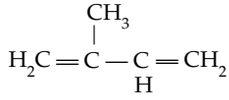
(*b*) **Membrane phospholipid,** generalized symbol



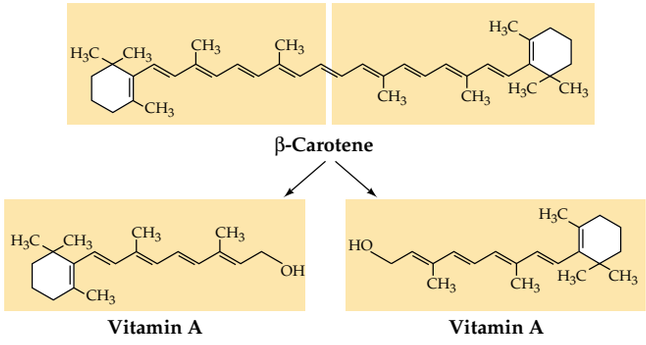
3.20 Phospholipid Structure (*a*) Phosphatidylcholine (lecithin) demonstrates the structure of a phospholipid molecule. In other phospholipids, the amino acid serine, the sugar alcohol inositol, or other compounds replace choline. (*b*) This generalized symbol is used throughout this book to represent a membrane phospholipid.

Carotenoids and steroids

The next two lipid classes we'll discuss—the carotenoids and the steroids—have chemical structures very different from those of triglycerides and phospholipids and from each other. Both carotenoids and steroids are synthesized by covalent linking and chemical modification of isoprene to form a series of isoprene units:



CAROTENOIDS TRAP LIGHT ENERGY. The **carotenoids** are a family of light-absorbing pigments found in plants and animals. Beta-carotene (β-carotene) is one of the pigments that traps light energy in leaves during photosynthesis. In humans, a molecule of β-carotene can be broken down into two vitamin A molecules (Figure 3.22), from which we make the pigment rhodopsin, which is required for vision. Carotenoids are responsible for the colors of carrots, tomatoes, pumpkins, egg yolks, and butter.



3.22 β-Carotene is the Source of Vitamin A The carotenoid β-carotene is symmetrical around its central double bond; when split, β-carotene becomes two vitamin A molecules. The simplified structural formula used here is standard chemical shorthand for large organic molecules with many carbon atoms. Structural formulas are simplified by omitting the C (indicating a carbon atom) at the intersections of the lines representing covalent bonds. Hydrogen atoms (H) to fill all the available bonding sites on each C are assumed.

STEROIDS ARE SIGNAL MOLECULES. The **steroids** are a family of organic compounds whose multiple rings share carbons (Figure 3.23). The steroid cholesterol is an important constituent of membranes. Other steroids function as hormones, chemical signals that carry messages from one part of the body to another. Testosterone and the estrogens are steroid hormones that regulate sexual development in vertebrates. Cortisol and related hormones play many regulatory roles in the digestion of carbohydrates and proteins, in the maintenance of salt balance and water balance, and in sexual development.

Cholesterol is synthesized in the liver and is the starting material for making testosterone and other steroid hormones, as well as the bile salts that help break down dietary fats so that they can be digested. Cholesterol is absorbed from foods such as milk, butter, and animal fats.

Some lipids are vitamins

Vitamins are small molecules that are not synthesized by the body, but are necessary for its normal functioning. Vitamins must be acquired from dietary sources.

- *Vitamin A* is formed from the β-carotene found in green and yellow vegetables (see Figure 3.22). In humans, a deficiency of vitamin A leads to dry skin, eyes, and internal body surfaces, retarded growth and development, and night blindness, which is a diagnostic symptom for the deficiency.
- *Vitamin D* regulates the absorption of calcium from the intestines. It is necessary for the proper deposition of calcium in bones; a deficiency of vitamin D can lead to rickets, a bone-softening disease.
- *Vitamin E* seems to protect cells from the damaging effects of oxidation–reduction reactions. For example, it has an important role in preventing unhealthy changes in the double bonds in the unsaturated fatty acids of membrane phospholipids. Commercially, vitamin E is added to some foods to slow spoilage.
- *Vitamin K* is found in green leafy plants and is also synthesized by bacteria normally present in the human intestine. This vitamin is essential to the formation of blood clots.

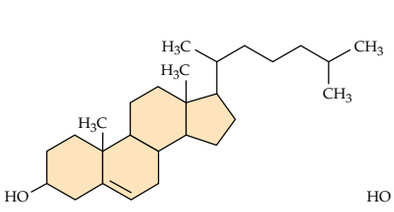
Inadequate vitamin intake can lead to deficiency diseases.

Wax coatings repel water

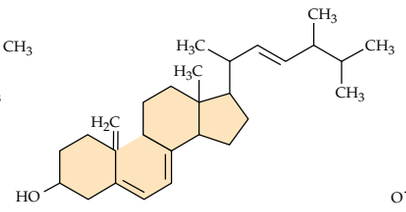
The sheen on human hair is not there only for cosmetic purposes. Glands in the skin secrete a waxy coating that repels water and keeps the hair pliable. Birds that live near water have a similar waxy coating on their feathers. The shiny leaves of holly plants, familiar during winter holidays, also have a waxy coating. Finally, bees make their honeycombs out of wax.

All waxes have the same basic structure: They are formed by an ester linkage between a saturated, long-chain fatty acid and a saturated, long-chain alcohol. The result is a very long

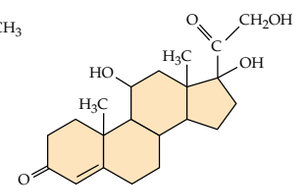
3.23 All Steroids Have the Same Ring Structure The steroids shown here, all important in vertebrates, are composed of carbon and hydrogen and are highly hydrophobic. However, small chemical variations, such as the presence or absence of a methyl or hydroxyl group, can produce enormous functional differences.



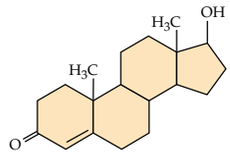
Cholesterol is a constituent of membranes and is the source of steroid hormones.



Vitamin D_2 can be produced in the skin by the action of light on a cholesterol derivative.

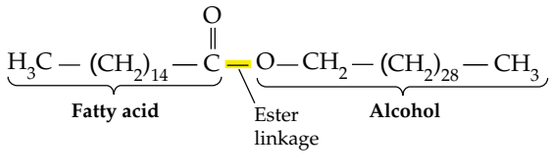


Cortisol is a hormone secreted by the adrenal glands.



Testosterone is a male sex hormone.

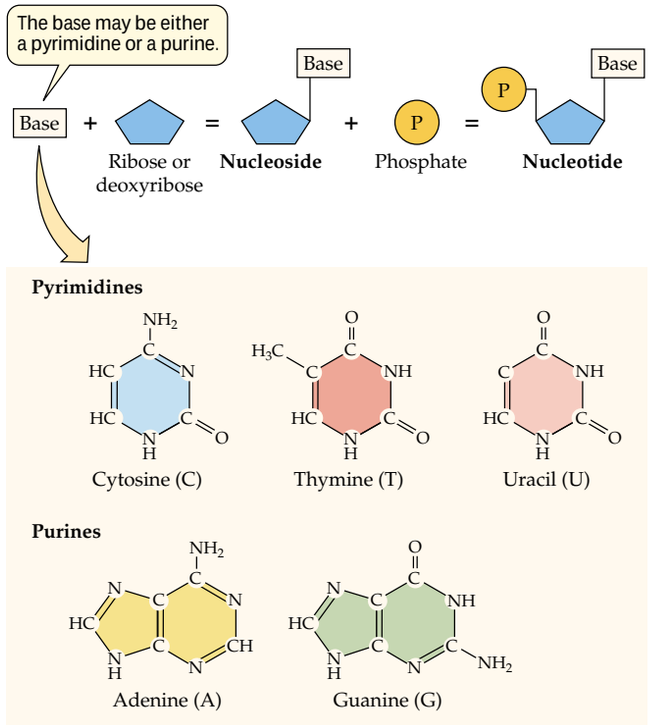
molecule, with 40–60 CH_2 groups. For example, here is the structure of beeswax:



This highly nonpolar structure accounts for the impermeability of wax to water.

Nucleic Acids: Informational Macromolecules That Can Be Catalytic

The **nucleic acids** are polymers specialized for the storage, transmission, and use of information. There are two types of nucleic acids: **DNA** (deoxyribonucleic acid) and **RNA** (ribonucleic acid). DNA molecules are giant polymers that encode hereditary information and pass it from generation to generation. Through an RNA intermediate, the information encoded in DNA is also used to specify the amino acid sequence of proteins. Information flows from DNA to DNA in reproduction, but in the nonreproductive activities of the cell, information flows from DNA to RNA to proteins, which ultimately carry out these functions. In addition, certain RNAs act as catalysts for important reactions in cells.



3.24 Nucleotides Have Three Components A nucleotide consists of a phosphate group, a pentose sugar (ribose or deoxyribose), and a nitrogen-containing base, all linked together by covalent bonds. The nitrogenous bases fall into two categories: Purines have two fused rings, and the smaller pyrimidines have a single ring.

The nucleic acids have characteristic chemical properties

Nucleic acids are composed of monomers called **nucleotides**, each of which consists of a pentose sugar, a phosphate group, and a nitrogen-containing **base**—either a pyrimidine or a purine (Figure 3.24). (Molecules consisting of a pentose sugar and a nitrogenous base, but no phosphate group, are called **nucleosides**.) In DNA, the pentose sugar is deoxyribose, which differs from the ribose found in RNA by one oxygen atom (see Figure 3.14).

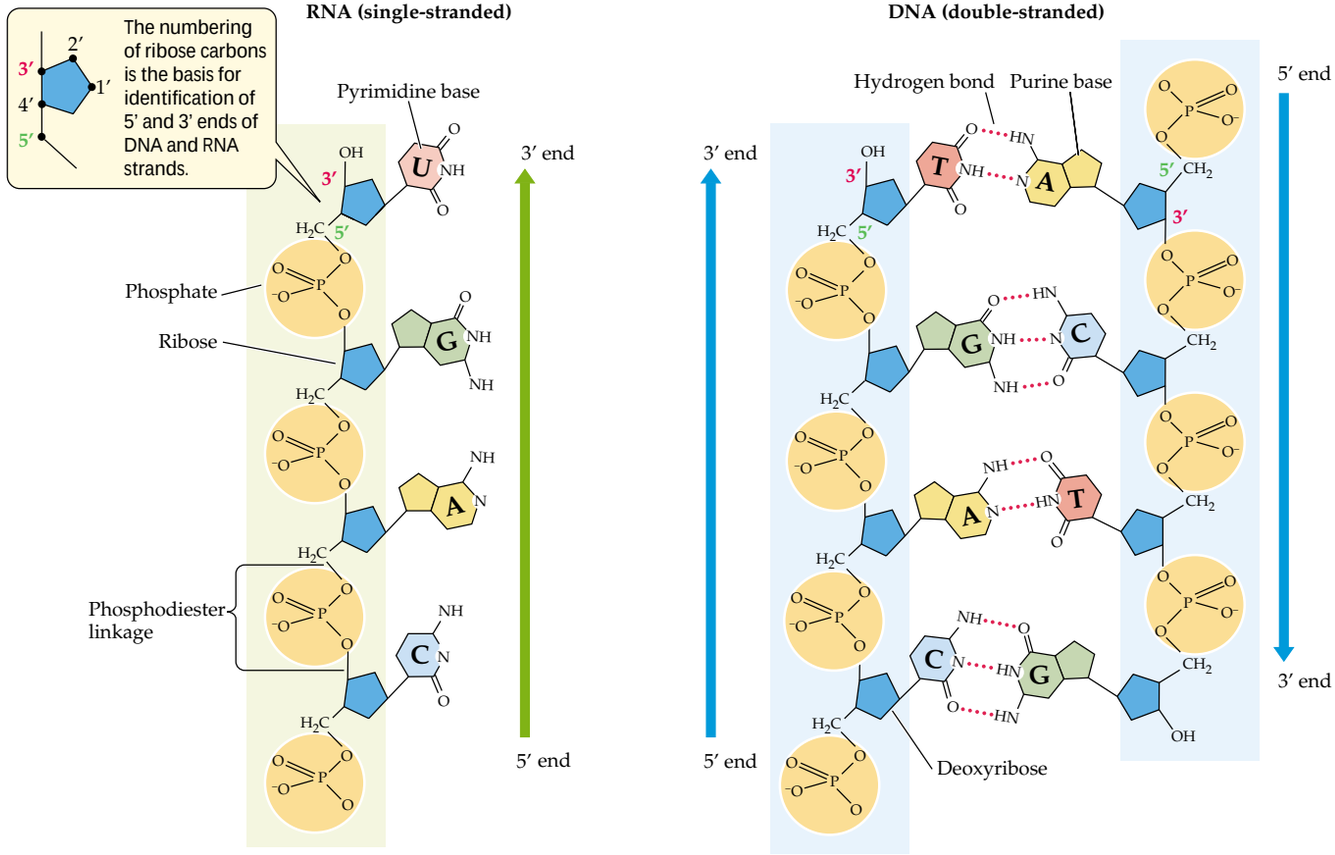
In both RNA and DNA, the backbone of the macromolecule consists of alternating pentose sugars and phosphates (sugar—phosphate—sugar—phosphate). The bases are attached to the sugars and project from the chain (Figure 3.25). The nucleotides are joined by *phosphodiester linkages* between the sugar of one nucleotide and the phosphate of the next (*-diester* refers to the two covalent bonds formed by —OH groups reacting with acidic phosphate groups). The phosphate groups link carbon 3 in one pentose sugar to carbon 5 in the adjacent sugar.

Most RNA molecules consist of only one polynucleotide chain. DNA, however, is usually double-stranded; it has two polynucleotide strands held together by hydrogen bonding between their nitrogenous bases. The two strands of DNA run in opposite directions. You can see what this means by drawing an arrow through the phosphate group from carbon 5′ to carbon 3′ in the next ribose. If you do this for both strands of the DNA in Figure 3.25, the arrows will point in opposite directions. This antiparallel orientation is necessary for the strands to fit together in three-dimensional space.

The uniqueness of a nucleic acid resides in its nucleotide sequence

Only four nitrogenous bases—and thus only four nucleotides—are found in DNA. The DNA bases and their abbreviations are adenine (A), cytosine (C), guanine (G), and thymine (T). A key to understanding the structure and function of nucleic acids is the principle of **complementary base pairing**. In double-stranded DNA, adenine and thymine always pair (A-T), and cytosine and guanine always pair (C-G).

Base pairing is complementary because of three factors: the sites for for hydrogen bonding on each base, the geometry of the sugar–phosphate backbone, which brings opposite bases near each other, and the molecular sizes of the paired bases. Adenine and guanine are both purines, consisting of two fused rings. Thymine and cytosine are both pyrimidines, consisting of only one ring. The pairing of a large purine with



In RNA, the bases are attached to ribose. The bases in RNA are the purines adenine (A) and guanine (G) and the pyrimidines cytosine (C) and uracil (U).

In DNA, the bases are attached to deoxyribose, and the base thymine (T) is found instead of uracil. Hydrogen bonds between purines and pyrimidines hold the two strands of DNA together.

3.25 Distinguishing Characteristics of DNA and RNA
RNA is usually a single strand. DNA usually consists of two strands running in opposite directions.

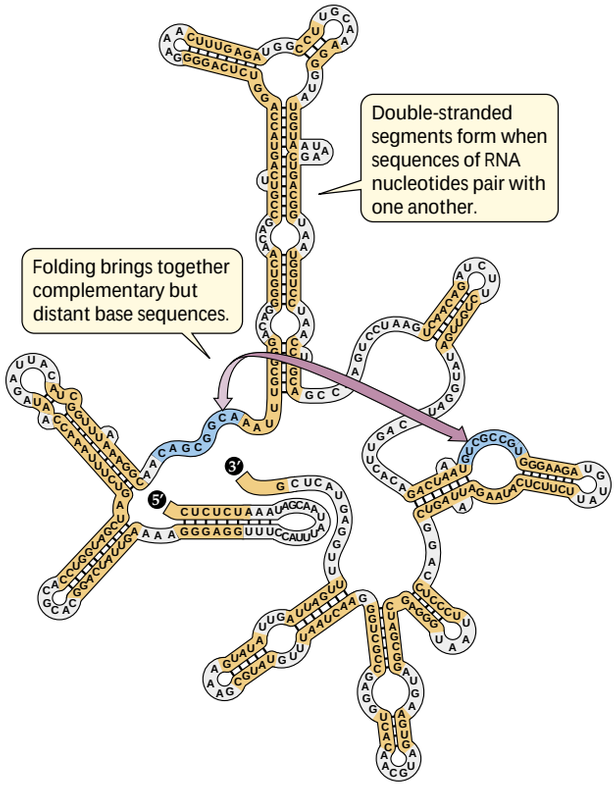
a small pyrimidine ensures stability and consistency in the double-stranded molecule of DNA.

Ribonucleic acids are also made up of four different monomers, but their nucleotides differ from those of DNA. In RNA the nucleotides are termed *ribonucleotides* (the ones in DNA are *deoxyribonucleotides*). They contain ribose rather than deoxyribose, and instead of the base thymine, RNA uses the base uracil (U) (Table 3.3). The other three bases are the same as in DNA.

Although RNA is generally single-stranded, complementary hydrogen bonding between ribonucleotides can take place. These bonds play important roles in determining the shapes of some RNA molecules and in associations between RNA molecules during protein synthesis (Figure 3.26). When the base sequence of DNA is copied in the synthesis of RNA, complementary base pairing also takes place between ribonucleotides and deoxyribonucleotides. In RNA, guanine and cytosine pair (G-C), as in DNA, but adenine pairs with uracil (A-U). Adenine in an RNA strand can pair either with uracil (in another RNA strand) or with thymine (in a DNA strand).

3.3 *Distinguishing RNA from DNA*

NUCLEIC ACID	SUGAR	BASES
RNA	Ribose	Adenine Cytosine Guanine Uracil
DNA	Deoxyribose	Adenine Cytosine Guanine Thymine



3.26 Hydrogen Bonding in RNA When a single-stranded RNA folds in on itself, hydrogen bonds can stabilize it into a three-dimensional shape.

DNA is a purely *informational* molecule. The information in DNA is encoded in the sequence of bases carried in its strands—the information encoded in the sequence TCAG is different from the information in the sequence CCAG. The information can be read easily and reliably, in a specific order.

The three-dimensional appearance of DNA is strikingly uniform. The segment shown in Figure 3.27 could be from any DNA molecule. The variations in DNA—the different sequences of bases—are strictly "internal." Through hydrogen bonding, the two complementary polynucleotide strands pair and twist to form a double helix. When compared with the complex and varied tertiary structures of different proteins, this uniformity is surprising. But this structural contrast makes sense in terms of the functions of these two classes of macromolecules.

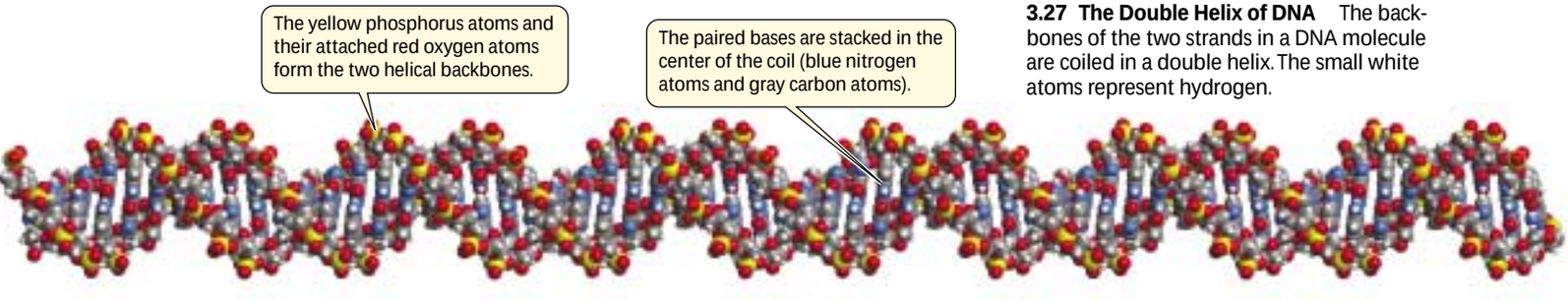
It is their different and unique shapes that permit proteins to recognize specific "target" molecules. The unique three-dimensional form of each protein matches at least a portion of the surface of the target molecule. In other words, structural diversity in the molecules to which proteins bind requires corresponding diversity in the structure of the proteins themselves.

In DNA, then, the information is in the sequence of the bases; in proteins, the information is in the shape of the molecule.

DNA is a guide to evolutionary relationships

Because DNA carries hereditary information between generations, a theoretical series of DNA molecules with changes in base sequences stretches back through evolutionary time. Of course, we cannot study all of these DNA molecules, because many of their organisms have become extinct. However, we can study the DNA of living organisms, which are judged to have changed little over millions of years. Comparisons and contrasts of these DNA molecules can be added to evidence from fossils and other sources to reveal the evolutionary record, as we will see in Chapter 24.

Closely related living species should have more similar base sequences than species judged by other criteria to be more distantly related. The examination of base sequences has confirmed many of the evolutionary relationships that have been inferred from the more traditional study of body structures, biochemistry, and physiology. For example, the closest living relative of humans (*Homo sapiens*) is the chimpanzee (genus *Pan*), which shares more than 98 percent of its DNA base sequence with human DNA. This confirmation of well-established evolutionary relationships gives credibility to the use of DNA to elucidate relationships when studies of structure are not possible or are not conclusive. For example,



3.27 The Double Helix of DNA The backbones of the two strands in a DNA molecule are coiled in a double helix. The small white atoms represent hydrogen.

DNA studies revealed a close evolutionary relationship between starlings and mockingbirds that was not expected on the basis of their anatomy or behavior.

DNA studies support the division of the prokaryotes into two domains, Bacteria and Archaea. Each of these two groups of prokaryotes is as distinct from the other as either is from the Eukarya, the third domain into which living things are classified (see Chapter 1). In addition, DNA comparisons support the hypothesis that certain subcellular compartments of eukaryotes (the organelles called mitochondria and chloroplasts) evolved from early bacteria that established a stable and mutually beneficial way of life inside larger cells.

RNA may have been the first biological catalyst

The three-dimensional structure of a folded RNA molecule presents a unique surface to the external environment (see Figure 3.26). These surfaces are every bit as specific as those of proteins. We noted above that an important role of proteins in biology is to act as catalysts, speeding up reactions that would ordinarily take place too slowly to be biologically useful, and that the spatial property of proteins is vital to this role.

As we will see, certain RNA molecules can also act as catalysts, using their three-dimensional shapes and other chemical properties. They can catalyze reactions on their own nucleotides as well as in other cellular substances. These catalytic RNAs are called **ribozymes**. Their discovery had implications for theories of the origin of life.

The Miller-Urey experiment and other such experiments in prebiotic chemistry yielded both amino acids and nucleotides. Organisms can synthesize both RNA and proteins from these monomers. As we noted above, in current organisms on Earth, protein synthesis requires DNA and RNA, and nucleic acid synthesis requires proteins (as enzymes). So the question is, when life originated, which came first, the proteins or the nucleic acids?

The discovery of catalytic RNAs provided a solution to this dilemma and led to the hypothesis that early life was part of an "RNA world." RNA can be informational (in its nucleotide sequence) as well as catalytic. So when RNA was first made, it could have acted as a catalyst for its own replication, as well as for the synthesis of proteins. Then DNA could have eventually evolved by being made from RNA. There is some laboratory evidence supporting this scenario:

- RNAs of different sequences have been put in a test tube and made to replicate on their own. Such self-replicating ribozymes speed up the synthesis of RNA 7 million-fold.
- In living organisms today, the formation of peptide linkages (see Figure 3.5) is catalyzed by a ribozyme.
- In certain viruses called retroviruses, there is an enzyme called reverse transcriptase that catalyzes the synthesis of DNA from RNA.

Nucleotides have other important roles

Nucleotides are more than just the building blocks of nucleic acids. As we will see in later chapters, there are several nucleotides with other functions:

- ATP (adenosine triphosphate) acts as an energy transducer in many biochemical reactions (see Chapter 6).
- GTP (guanosine triphosphate) serves as an energy source, especially in protein synthesis. It also has a role in the transfer of information from the environment to the body tissues (see Chapters 12 and 15).
- cAMP (cyclic adenosine monophosphate), a special nucleotide in which a bond forms between the sugar and phosphate groups within adenosine monophosphate, is essential in many processes, including the actions of hormones and the transmission of information by the nervous system (see Chapter 15).

All Life from Life

The concepts conveyed throughout this chapter—that large molecules obey the mechanistic laws of physics and chemistry, and that life could have arisen from inanimate, self-replicating macromolecules—have come to be generally accepted by the scientific community. So should we expect to see new life forms arise at any time from the biochemical environment?

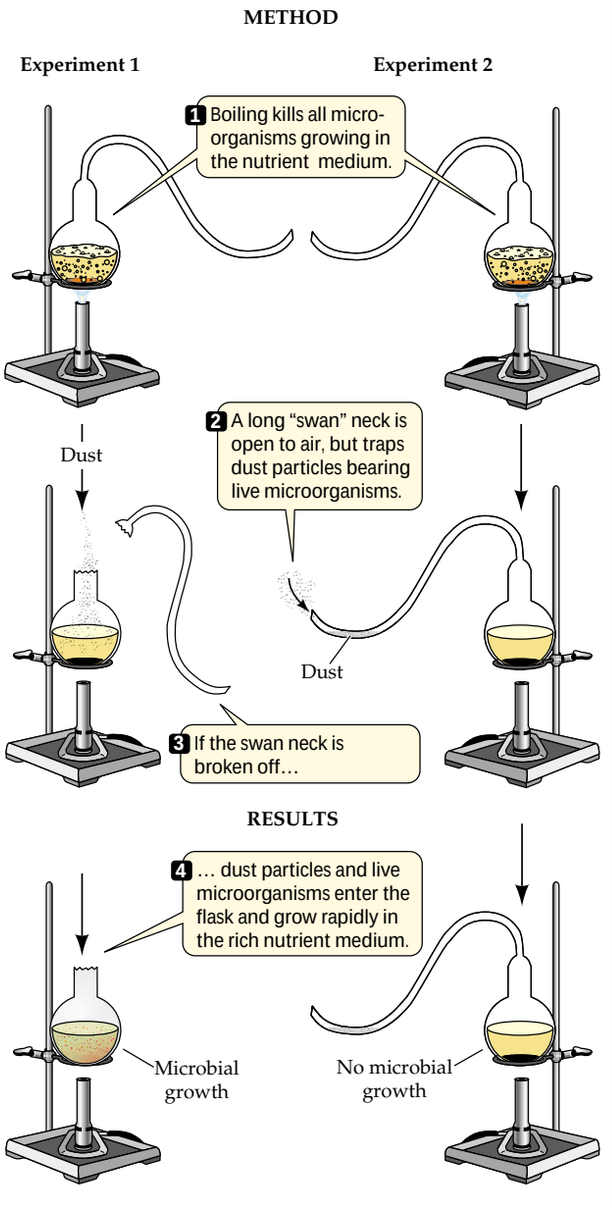
During the Renaissance (a period from about 1350 to 1700 A.D., marked by the birth of modern science), most people thought that at least some forms of life arose directly from inanimate or decaying matter by *spontaneous generation*. For instance, it was suggested that mice arose from sweaty clothes placed in dim light, frogs came from moist soil, and flies were produced from meat. These ideas were attacked by scientists such as the Italian doctor and poet Francisco Redi using the relatively new idea of using experiments to test an idea. In 1668, Redi proposed that flies arose not by some mysterious transformation of decaying meat, but from other flies, who laid their eggs on the meat. The eggs developed into wormlike maggots (the immature form of flies). Redi set out several jars containing chunks of meat.

- One jar contained meat exposed both to the air and to flies.
- A second jar contained meat in a container wrapped in a fine cloth so that the meat was exposed to the air, but not to flies.
- The meat in the third jar was in a sealed container and thus was not exposed to either air or flies.

As he had hypothesized, Redi found maggots, which then hatched into flies, only in the first container. The idea that a complex organism like a fly could come from a totally different substance was laid to rest.

EXPERIMENT

Question: Pasteur asked "Does life generate spontaneously or does it come only from already existing life"?



Conclusion: All life comes from existing life.

3.28 Disproving the Spontaneous Generation of Life Louis Pasteur's classic experiments showed that, under today's conditions, an inanimate solution remains lifeless unless a living organism contaminates it.

With the invention of the microscope in the 1660s, a vast new biological world was unveiled. Under microscopic observation, virtually every environment on Earth was found to be teeming with tiny organisms such as bacteria. Some scientists believed that these organisms arose spontaneously from their rich chemical environment.

The experiments that disproved this idea were done by the great French scientist Louis Pasteur. His experiments showed that microorganisms come only from other microorganisms, and that an environment without life remains lifeless unless contaminated by living creatures (Figure 3.28).

These experiments by Redi, Pasteur, and others provided solid evidence that neither small (bacteria) nor large (flies) organisms come from inanimate matter, but instead come from living parent organisms.

Indeed, life on Earth no longer arises from nonliving materials. This is because the atmospheric and planetary conditions that exist on Earth today are vastly different from those on the prebiotic, anaerobic planet. The oxygen present in today's atmosphere would break down the prebiotic molecules before they could accumulate. In addition, the necessary energy sources—including constant lightning strikes, immense volcanic eruptions, and bombardment by intense ultraviolet light—are no longer present with anything like their primeval force.

Chapter Summary

Theories of the Origin of Life

▶ Life may have come from outside Earth. The evidence for this proposal comes primarily from chemicals contained in meteorites that have landed on Earlth.

▶ The theory of chemical evolution proposes that life on Earth originated on Earth. Experiments using model systems that attempt to duplicate the ancient Earth have shown that chemical evolution could have produced the four types of macromolecules that distinguish living things. **Review Figure 3.1.** **See Web/CD Tutorial 3.1**

Macromolecules: Giant Polymers

▶ Macromolecules are polymers constructed by the formation of covalent bonds between smaller molecules called monomers. Macromolecules in living organisms include polysaccharides, proteins, and nucleic acids. **Review Figure 3.2 and Table 3.1**

▶ Macromolecules have specific, characteristic three-dimensional shapes that depend on the structure, properties, and sequence of their monomers.

▶ Different functional groups give local sites on macromolecules specific properties that are important for their biological functioning and their interactions with other macromolecules. **See Web/CD Tutorial 3.2**

Condensation and Hydrolysis Reactions

▶ Monomers are joined by condensation reactions, which release a molecule of water for each bond formed. Hydrolysis

reactions use water to break polymers into monomers. **Review Figure 3.3**

Proteins: Polymers of Amino Acids

▶ The functions of proteins include support, protection, catalysis, transport, defense, regulation, and movement. Protein function sometimes requires an attached prosthetic group.

▶ There are 20 amino acids found in proteins. Each amino acid consists of an amino group, a carboxyl group, a hydrogen, and a side chain bonded to the α carbon atom. **Review Table 3.2**

▶ The side chains, or R groups, of amino acids may be charged, polar, or hydrophobic; there are also special cases, such as the —SH groups of cysteine, which can form disulfide bridges. The side chains give different properties to each of the amino acids. **Review Table 3.2 and Figure 3.4**

▶ Amino acids are covalently bonded together into polypeptide chains by peptide linkages, which form by condensation reactions between the carboxyl and amino groups. **Review Figure 3.5**

▶ Polypeptide chains are folded into specific three-dimensional shapes to form functional proteins. Four levels of protein structure are possible: primary, secondary, tertiary, and quaternary.

▶ The primary structure of a protein is the sequence of amino acids bonded by peptide linkages. This primary structure determines both the higher levels of structure and protein function. **Review Figure 3.6*a***

▶ The two types of secondary structure—α helices and β pleated sheets—are maintained by hydrogen bonds between atoms of the amino acid residues. **Review Figure 3.6*b*,*c***

▶ The tertiary structure of a protein is generated by bending and folding of the polypeptide chain. **Review Figures 3.6*d*, 3.7**

▶ The quaternary structure of a protein is the arrangement of two or more polypeptides into a single functional protein consisting of two or more polypeptide subunits. **Review Figures 3.6*e*, 3.8**

▶ Weak chemical interactions are important in the three-dimensional structure of proteins and in their binding to other molecules. **Review Figure 3.9, 3.10**

▶ Proteins denatured by heat, alterations in pH, or certain chemicals lose their tertiary and secondary structure as well as their biological function. Renaturation is not often possible. **Review Figure 3.11**

▶ Chaperonins assist protein folding by preventing binding to inappropriate ligands. **Review Figure 3.12**

Carbohydrates: Sugars and Sugar Polymers

▶ All carbohydrates contain carbon bonded to hydrogen atoms and hydroxyl groups.

▶ Hexoses are monosaccharides that contain six carbon atoms. Examples of hexoses include glucose, galactose, and fructose, which can exist as chains or rings. **Review Figures 3.13, 3.14.** **See Web/CD Activity 3.1**

▶ The pentoses are five-carbon monosaccharides. Two pentoses, ribose and deoxyribose, are components of the nucleic acids RNA and DNA, respectively. **Review Figure 3.14**

▶ Glycosidic linkages may have either α or β orientation in space. They covalently link monosaccharides into larger units such as disaccharides, oligosaccharides, and polysaccharides. **Review Figure 3.15**

▶ Cellulose, a very stable glucose polymer, is the principal component of the cell walls of plants. It is formed by glucose units linked together by β-glycosidic linkages between carbons 1 and 4. Starches, less dense and less stable than cellulose, store energy in plants. Starches and glycogen are formed by α-glycosidic linkages between carbons 1 and 4 and are distinguished by the amount of branching they exhibit. **Review Figure 3.16**

▶ Chemically modified monosaccharides include the sugar phosphates and amino sugars. A derivative of the amino sugar glucosamine polymerizes to form the polysaccharide chitin, which is found in the cell walls of fungi and the exoskeletons of insects. **Review Figure 3.17**

Lipids: Water-Insoluble Molecules

▶ Although lipids can form gigantic structures, these aggregations are not chemically macromolecules because the individual units are not linked by covalent bonds.

▶ Fats and oils are triglycerides, composed of three fatty acids covalently bonded to a glycerol molecule by ester linkages. **Review Figure 3.18**

▶ Saturated fatty acids have a hydrocarbon chain with no double bonds. The hydrocarbon chains of unsaturated fatty acids have one or more double bonds that bend the chain, making close packing less possible. **Review Figure 3.29**

▶ Phospholipids have a hydrophobic hydrocarbon "tail" and a hydrophilic phosphate "head." **Review Figure 3.20**

▶ In water, the interactions of the hydrophobic tails and hydrophilic heads of phospholipids generate a phospholipid bilayer that is two molecules thick. The head groups are directed outward, where they interact with the surrounding water. The tails are packed together in the interior of the bilayer. **Review Figure 3.21**

▶ Carotenoids trap light energy in green plants. Carotene can be split to form vitamin A, a lipid vitamin. **Review Figure 3.22**

▶ Some steroids, such as testosterone, function as hormones. Cholesterol is synthesized by the liver and has a role in cell membranes, as well as in the digestion of fats. **Review Figure 3.23**

▶ Vitamins are substances that are required for normal functioning, but must be acquired from the diet.

Nucleic Acids: Informational Macromolecules

▶ DNA is the hereditary material. Both DNA and RNA play roles in the formation of proteins. Information flows from DNA to RNA to protein.

▶ Nucleic acids are polymers made up of nucleotides. A nucleotide consists of a phosphate group, a sugar (ribose in RNA and deoxyribose in DNA), and a nitrogen-containing base. In DNA the bases are adenine, guanine, cytosine, and thymine, but in RNA uracil substitutes for thymine. **Review Figure 3.24 and Table 3.3.** **See Web/CD Activity 3.2**

▶ In the nucleic acids, the bases extend from a sugar–phosphate backbone. The information content of DNA and RNA resides in their base sequences. RNA is single-stranded. DNA is a double-stranded helix in which there is complementary, hydrogen-bonded base pairing between adenine and thymine (A-T) and guanine and cytosine (G-C). The two strands of the DNA double helix run in opposite directions. **Review Figures 3.25, 3.27.** **See Web/CD Activity 3.3**

▶ Base pairing of single-stranded RNAs can lead to three-dimensional structures, which can be catalytic. This finding has led to the proposal that in the origin of life, RNA preceded protein. **Review Figure 3.26**

▶ Comparing the DNA base sequences of different living species provides information on their evolutionary relationships.

All Life from Life

▶ One of the earliest conclusions from biology as a modern experimental science was that even the tiniest microbe comes from others of the same type—that is, that life begets life. **Review Figure 3.28.** **See Web/CD Tutorial 3.3**

▶ The conditions on primeval Earth that may have enabled life to arise from inanimate self-replicating chemicals no longer exist. Today all life comes from pre-existing life.

Self-Quiz

1. The most abundant molecule in the cell is
   a. carbohydrate.
   b. lipid.
   c. nucleic acid.
   d. protein.
   e. water.
2. All lipids are
   a. triglycerides.
   b. polar.
   c. hydrophilic.
   d. polymers of fatty acids.
   e. more soluble in nonpolar solvents than in water.
3. All carbohydrates
   a. are polymers.
   b. are simple sugars.
   c. consist of one or more simple sugars.
   d. are found in biological membranes.
   e. are more soluble in nonpolar solvents than in water.
4. Which of the following is *not* a carbohydrate?
   a. Glucose
   b. Starch
   c. Cellulose
   d. Hemoglobin
   e. Deoxyribose
5. All proteins
   a. are enzymes.
   b. consist of one or more polypeptides.
   c. are amino acids.
   d. have quaternary structures.
   e. are more soluble in nonpolar solvents than in water.
6. Which of the following statements about the primary structure of a protein is *not* true?
   a. It may be branched.
   b. It is determined by the structure of the corresponding DNA.
   c. It is unique to that protein.
   d. It determines the tertiary structure of the protein.
   e. It is the sequence of amino acids in the protein.
7. The amino acid leucine (see Table 3.2)
   a. is found in all proteins.
   b. cannot form peptide linkages.
   c. is likely to appear in the part of a membrane protein that lies within the phospholipid bilayer.
   d. is likely to appear in the part of a membrane protein that lies outside the phospholipid bilayer.
   e. is identical to the amino acid lysine.
8. The quaternary structure of a protein
   a. consists of four subunits—hence the name *quaternary*.
   b. is unrelated to the function of the protein.
   c. may be either alpha or beta.
   d. depends on covalent bonding among the subunits.
   e. depends on the primary structures of the subunits.
9. All nucleic acids
   a. are polymers of nucleotides.
   b. are polymers of amino acids.
   c. are double-stranded.
   d. are double-helical.
   e. contain deoxyribose.
10. Which of the following statements about condensation reactions is *not* true?
    a. Protein synthesis results from them.
    b. Polysaccharide synthesis results from them.
    c. Nucleic acid synthesis results from them.
    d. They consume water as a reactant.
    e. Different condensation reactions produce different kinds of macromolecules.

For Discussion

1. Phospholipids make up a major part of most biological membranes; cellulose is the major constituent of the cell walls of plants. How do the chemical structures and physical properties of phospholipids and cellulose relate to their functions in cells?
2. Suppose that, in a given protein, one lysine is replaced by aspartic acid (see Table 3.2). Does this change occur in the primary structure or in the secondary structure? How might it result in a change in tertiary structure? In quaternary structure?
3. If there are 20 different amino acids commonly found in proteins, how many different dipeptides are there? How many different tripeptides? How many different trinucleotides? How many different single-stranded RNAs composed of 200 nucleotides?
4. Contrast the following three structures, emphasizing the surfaces they present to their environment: hemoglobin; a DNA molecule; a protein that spans a biological membrane.
5. Why might RNA have preceded proteins in the evolution of biological macromolecules?

3 Cells: The Basic Units of Life

Charles Darwin faced a dilemma. In his great book *On the Origin of Species*, published in 1859, he proposed the theory of natural selection to explain the gradual appearance and disappearance of different forms of animals and plants over long periods of time. But he realized that the fossil record, on which he based his theory, was incomplete, especially for the beginning of life. The oldest fossils that had been found in Darwin's time were complex organisms in rocks dated at about 550 million years ago (the Cambrian period). Where were the missing Precambrian fossils? These would surely provide a link to the origin of life.

As we saw in Chapter 3, conditions on Earth were probably suitable for the emergence of life by 4 billion years ago, about 600 million years after Earth began to form. But until recently there was no evidence for life older than the Cambrian. By the turn of the twentieth century, there was evidence for fossilized clumps of algae (simple aquatic photosynthetic organisms) in rocks at the base of the Grand Canyon that were close to 1 billion years old.

It took nearly another century to push the clock of life back nearer to its origins. In 1993, geologist J. William Schopf found fossilized chains of cylindrical objects, quite similar in size and shape to contemporary cyanobacteria ("blue-green algae"), in rocks in Western Australia that he dated at an astonishing 3.5 billion years old. He then used a chemical analysis method called laser Raman spectroscopy to show that these objects apparently contain carbon deposits that are chemical signatures of life.

Rounded or cylindrical objects in Earth's rocks or in a meteorite from Mars (see Chapter 3) get scientists excited because they realize that life is not just a bunch of macromolecules. Rather, life is macromolecules that can perform unique functions because they are enclosed in a structural compartment that is separate from the external environment. This separation allows living things to maintain a constant internal environment (homeostasis).

The "living compartment" is the cell, the subject of this chapter. The water-insoluble phospholipid structure (see Figure 3.21) that defines and contains cells is called the plasma membrane. It and its functions are so important that we will devote the entire next chapter to membranes. Subsequent chapters will be devoted to the chemical activities that take place inside all cells.

The Earliest Trace of Life? This fossil from Western Australia is 3.5 billion years old and shows carbon traces that indicate life. Its form is similar to that of modern filamentous cyanobacteria (inset).

The Cell: The Basic Unit of Life

Just as atoms are the units of chemistry, cells are the building blocks of life. Three statements constitute the **cell theory**:

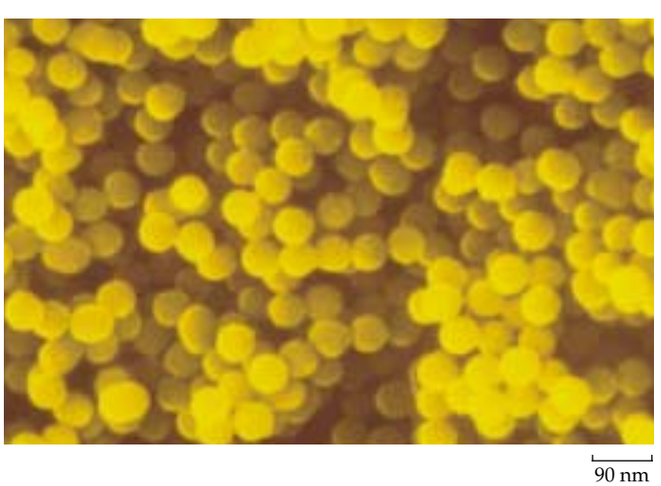
- Cells are the fundamental units of life.
- All organisms are composed of cells.
- All cells come from preexisting cells.

Cells are composed of water molecules and the small and large molecules we examined in the previous two chapters. Each cell contains at least 10,000 different types of molecules, most of them present in many copies. Cells use these molecules to transform matter and energy, to respond to their environment, and to reproduce themselves.

The cell theory has three important implications. First, it means that studying cell biology is in some sense the same as studying life. The principles that underlie the functions of the single cell in a bacterium are similar to those governing the 60 trillion cells of your body. Second, it means that life is continuous. All those cells in your body came from a single cell, the fertilized egg, which came from the fusion of two cells, a sperm and an egg from your parents, whose cells came from their fertilized eggs, and so on. Finally, it means that the origin of life on Earth was marked by the origin of the first cells.

Cells may have come from stable bubbles

Isolation from the general environment can be achieved in the laboratory within aggregates produced from molecules made in prebiotic synthesis experiments. Called **protobionts**, these aggregates cannot reproduce, but they can maintain internal chemical environments that differ from their surroundings. Under the microscope, they look a lot like tiny cells (Figure 4.1).



4.1 Protobionts These aggregates, made by agitating a solution of macromolecules, are chemical compartments, can perform some metabolic reactions, and can exchange materials with their environment. They are a model of how cells may have originated.

In the 1920s, Russian scientist Alexander Oparin mixed a large protein and a polysaccharide in solution. When he agitated this mixture, bubbles formed. He could also do this with other polymers. The interiors of these bubbles had much higher concentrations of the macromolecules than their surroundings. Moreover, they catalyzed chemical reactions, and they had some control over what left them and crossed the boundary into the environment. In other words, they were protobionts. Later, other researchers showed that if lipids are mixed in an aqueous environment, they spontaneously arrange themselves into droplets surrounded by a bilayer.

Taken together with the prebiotic chemistry models and RNA world hypothesis described in Chapter 3, these experiments suggest a *bubble theory* for the origin of cells.

Cell size is limited by the surface area-to-volume ratio

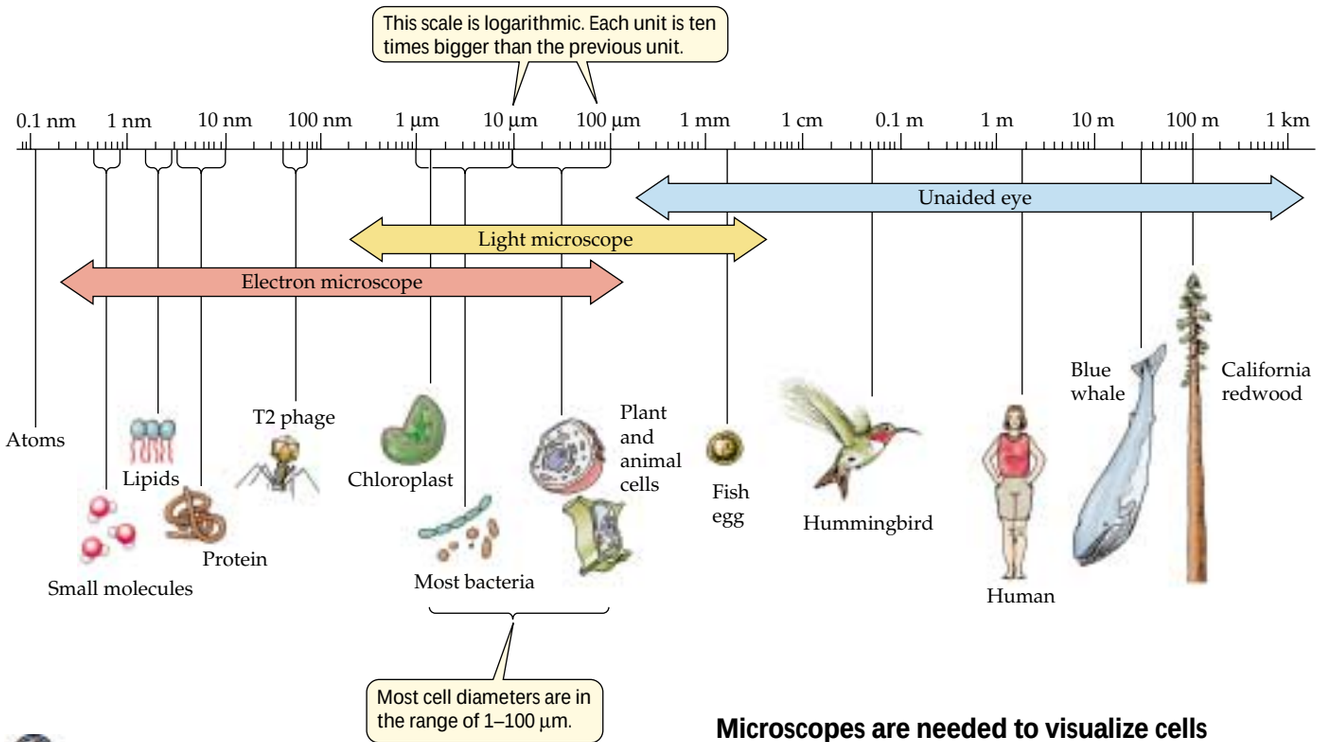
Most cells are tiny. The volume of cells ranges from 1 to 1,000 μm^3 (Figure 4.2). The eggs of some birds are enormous exceptions, to be sure, and individual cells of several types of algae and bacteria are large enough to be viewed with the unaided eye. And although neurons (nerve cells) have a volume that is within the "normal" cell range, they often have fine projections that may extend for meters, carrying signals from one part of a large animal to another. But by and large, cells are minuscule. The reason for this relates to the change in the **surface area-to-volume ratio** (SA/V) of any object as it increases in size.

As a cell increases in volume, its surface area also increases, but not to the same extent (Figure 4.3). This phenomenon has great biological significance for two reasons:

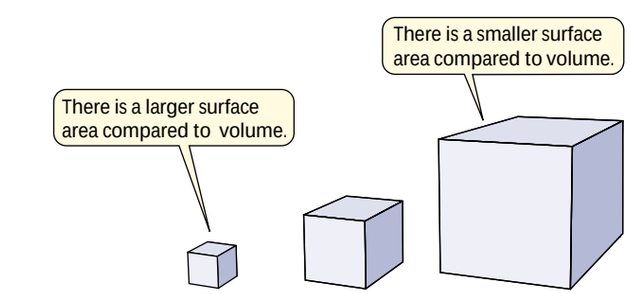
- The *volume* of a cell determines the amount of chemical activity it carries out per unit of time.
- The *surface area* of a cell determines the amount of substances the cell can take in from the outside environment and the amount of waste products it can release to the environment.

As a living cell grows larger, its rate of waste production and its need for resources increase faster than its surface area. This explains why large organisms must consist of many small cells: Cells are small in volume in order to maintain a large surface area-to-volume ratio.

In a multicellular organism, the large surface area represented by the multitude of small cells that make up the organism enables it to carry out the multitude of functions required for survival. Special structures transport food, oxygen, and waste materials to and from the small cells that are distant from the external surface of the organism.



4.2 The Scale of Life This scale shows the relative sizes of molecules, cells, and multicellular organisms.



	1-mm cube	2-mm cube	4-mm cube
Surface area	6 sides × 1^2 = 6 mm^2	6 sides × 2^2 = 24 mm^2	6 sides × 4^2 = 96 mm^2
Volume	1^3 = 1 mm^3	2^3 = 8 mm^3	4^3 = 64 mm^3
Surface area-to-volume ratio	6/1	3/1	1.5/1

4.3 Why Cells Are Small As an object grows larger, its volume increases more rapidly than its surface area. Cells must maintain a large surface area-to-volume ratio in order to function, which explains why large organisms must be composed of many small cells rather than a few huge ones.

Microscopes are needed to visualize cells

Most cells are invisible to the human eye. The smallest object a person can typically discern is about 0.2 mm (200 µm) in size. We refer to this measure as *resolution*, the distance apart two objects must be in order for them to be distinguished as separate; if they are closer together, they appear as a single blur. Many cells are much smaller than 200 µm. *Microscopes* are instruments used to improve resolution so that cells and their internal structures can be seen.

There are two basic types of microscopes: light microscopes and electron microscopes. The **light microscope** (LM) uses glass lenses and visible light to form a magnified image of an object. It has a resolving power of about 0.2 µm, which is 1,000 times that of the human eye. It allows visualization of cell sizes and shapes and some internal cell structures. The latter are hard to see under ordinary light, so cells are often killed and stained with various dyes to make certain structures stand out.

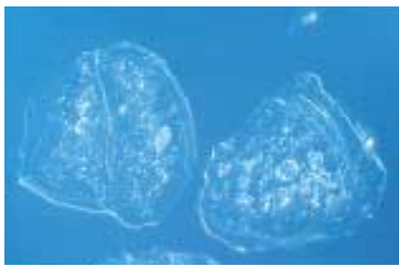
An **electron microscope** (EM) uses magnets to focus an electron beam, much as a light microscope uses glass lenses to focus a beam of light. Since we cannot see electrons, the electron microscope directs them at a fluorescent screen or photographic film to create a visible image. The resolving power of electron microscopes is about 0.5 nm, which is 400,000 times that of the human eye. This resolving power permits the details of many subcellular structures to be distinguished.

Many techniques have been developed to enhance the views of cells we see under the light and electron microscopes (Figure 4.4).

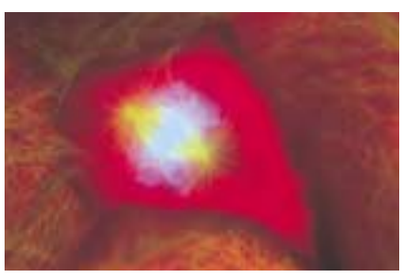
RESEARCH METHOD

In **bright-field microscopy**, light passes directly through the cells. Unless natural pigments are present, there is little contrast and details are not distinguished.

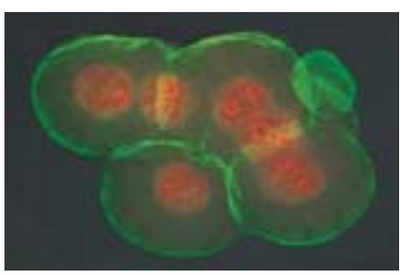
In **phase-contrast microscopy**, contrast in the image is increased by emphasizing differences in refractive index (the capacity to bend light), thereby enhancing light and dark regions in the cell.



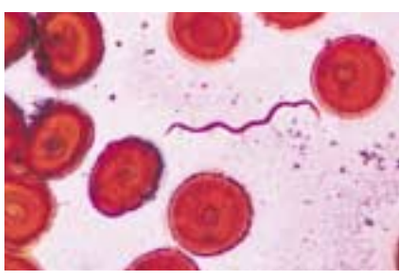
Differential interference-contrast microscopy uses two beams of polarized light. The combined images look as if the cell is casting a shadow on one side.



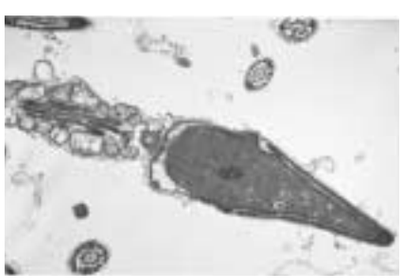
In **fluorescence microscopy**, a natural substance in the cell or a fluorescent dye that binds to a specific cell material is stimulated by a beam of light, and the longer-wavelength fluorescent light is observed coming directly from the dye.



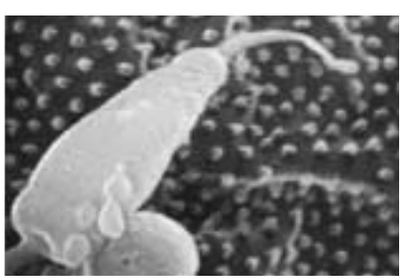
Confocal microscopy uses fluorescent materials but adds a system of focusing both the stimulating and emitted light so that a single plane through the cell is seen. The result is a sharper two-dimentional image than with standard fluorescence microscopy.



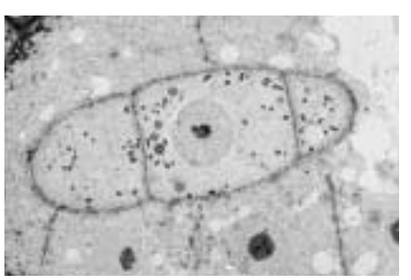
In **stained bright-field microscopy**, a stain added to preserve cells enhances contrast and reveals details not otherwise visible. Stains differ greatly in their chemistry and their capacity to bind to cell materials, so many choices are available.



In **transmission electron microscopy** (TEM), a beam of electrons is focused on the object by magnets. Objects appear darker if they absorb the electrons. If the electrons pass through they are detected on a fluorescent screen.



Scanning electron microscopy (SEM) directs electrons to the surface of the sample, where they cause other electrons to be emitted. These electrons are viewed on a screen. The three-dimentional surface of the object can be visualized.



Cryoelectron microscopy uses quickly frozen samples to reduce aberrations that are seen when samples are treated chemically. Computer analysis of thick sections can reconstruct a sample in three dimensions.

4.4 Looking at Cells The top six panels show some techniques used in light microscopy. The lower three images were created using electron microscopes.

Cells are surrounded by a plasma membrane

As we have noted, a **plasma membrane** separates each cell from its environment, creating a segregated (but not isolated) compartment. The plasma membrane is composed of a phospholipid bilayer, with the hydrophilic "heads" of the lipids facing the cell's aqueous interior on one side of the membrane and the extracellular environment on the other (see Figure 3.21). Proteins are embedded in the lipids. In many cases, these proteins protrude into the cytoplasm and into the extracellular environment. We will devote most of Chapter 5 to detailing the structure and functions of the plasma membrane, but summarize its roles here.

- The plasma membrane allows the cell to maintain a more or less constant internal environment. A self-maintaining, *constant internal environment* is a key characteristic of life that will be discussed in detail in Chapter 41.
- The plasma membrane acts as a *selectively permeable barrier*, preventing some substances from crossing while permitting other substances to enter and leave the cell.
- As the cell's boundary with the outside environment, the plasma membrane is important in *communicating with adjacent cells and receiving extracellular signals*. We will describe this function in Chapter 15.
- The plasma membrane often has molecules protruding from it that are responsible for *binding and adhering* to adjacent cells.

Cells show two organizational patterns

Prokaryotic cell organization is characteristic of the domains Bacteria and Archaea. Organisms in these domains are called *prokaryotes*. Their cells do not have membrane-enclosed internal compartments. The first cells ever to form were undoubtedly similar in organization to modern prokaryotes.

Eukaryotic cell organization is found in the domain Eukarya, which includes the protists, plants, fungi, and animals. The genetic material (DNA) of eukaryotic cells is contained in a special membrane-enclosed compartment called the nucleus. Eukaryotic cells also contain other membrane-enclosed compartments in which specific chemical reactions take place. Organisms with this type of cell organization are known as *eukaryotes*.

Both prokaryotes and eukaryotes have prospered for many hundreds of millions of years of evolution, and both are great success stories. Let's look first at prokaryotic cells.

Prokaryotic Cells

Prokaryotes can live off more different and diverse energy sources than any other living creatures, and they inhabit greater environmental extremes, such as very hot springs and very salty water. The vast diversity within the prokaryotic domains is the subject of Chapter 27.

Prokaryotic cells are generally smaller than eukaryotic cells, ranging from 0.25×1.2 μm to 1.5×4 μm. Each prokaryote is a single cell, but many types of prokaryotes are usually seen in chains, small clusters, or even clusters containing hundreds of individuals. In this section, we will first consider the features that cells in the domains Bacteria and Archaea have in common. Then we will examine structural features that are found in some, but not all, prokaryotes.

Prokaryotic cells share certain features

All prokaryotic cells have the same basic structure:

- The plasma membrane encloses the cell, regulating the traffic of materials into and out of the cell and separating it from its environment.
- A region called the **nucleoid** contains the hereditary material (DNA) of the cell.

The rest of the material enclosed in the plasma membrane is called the **cytoplasm**. The cytoplasm is composed of two parts: the liquid cytosol, and insoluble suspended particles, including ribosomes.

- The **cytosol** consists mostly of water that contains dissolved ions, small molecules, and soluble macromolecules such as proteins.
- **Ribosomes** are granules about 25 nm in diameter that are sites of protein synthesis.

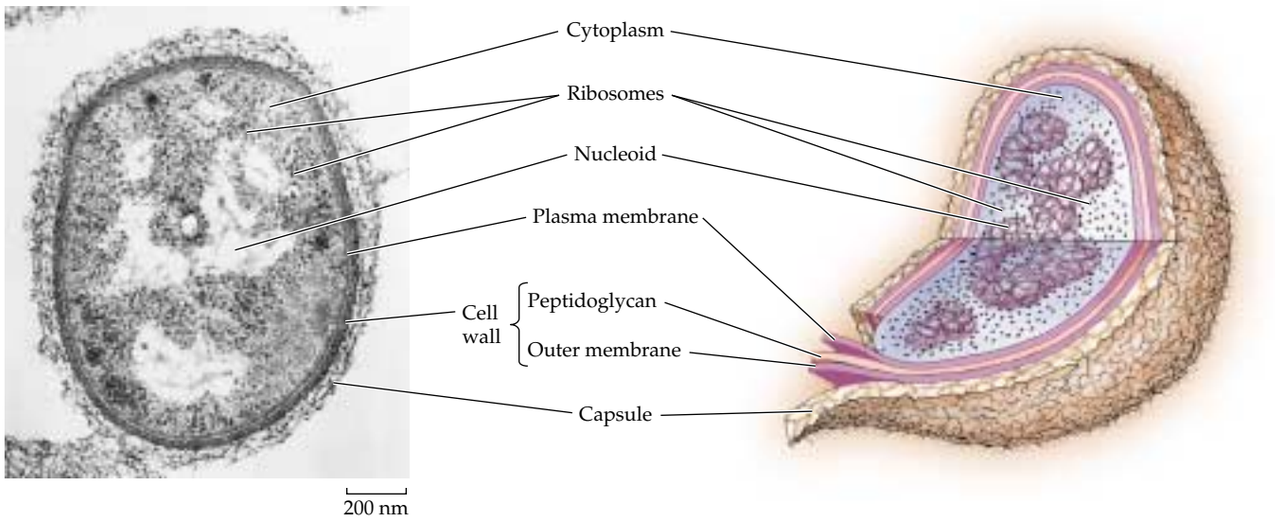
The cytoplasm is not a static region. Rather, the substances in this aqueous environment are in constant motion. For example, a typical protein moves around the entire cell within a minute, and encounters many molecules along the way.

Although structurally less complicated than eukaryotic cells, prokaryotic cells are functionally complex, carrying out thousands of biochemical transformations.

Some prokaryotic cells have specialized features

As they evolved, some prokaryotes developed specialized structures that gave a selective advantage to those cells that had them. These structures include a protective cell wall, an internal membrane for compartmentalization of chemical reactions, and flagella for cell movement through the watery environment. These features are shown in Figures 4.5 and 4.6.

CELL WALLS. Most prokaryotes have a **cell wall** located outside the plasma membrane. The rigidity of the cell wall



4.5 A Prokaryotic Cell The bacterium *Pseudomonas aeruginosa* illustrates typical prokaryotic cell structures. Note the existence of several protective structures external to the plasma membrane.

supports the cell and determines its shape. The cell walls of most bacteria, but not archaea, contain *peptidoglycan*, a polymer of amino sugars, cross-linked by covalent bonds to form a single giant molecule around the entire cell. In some bacteria, another layer—the outer membrane (a polysaccharide-rich phospholipid membrane)—encloses the peptidoglycan layer. Unlike the plasma membrane, this outer membrane is not a major permeability barrier, and some of its polysaccharides are disease-causing toxins.

Enclosing the cell wall in some bacteria is a layer of slime, composed mostly of polysaccharides and referred to as a *capsule*. The capsules of some bacteria may protect them from attack by white blood cells in the animals they infect. The capsule helps keep the cell from drying out, and sometimes it helps the bacterium attach to other cells. Many prokaryotes produce no capsule, and those that do have capsules can survive even if they lose them, so the capsule is not essential to cell life.

As you will see later in this chapter, eukaryotic plant cells also have a cell wall, but it differs in composition and structure from the cell walls of prokaryotes.

INTERNAL MEMBRANES. Some groups of bacteria—the cyanobacteria and some others—carry on photosynthesis. In these photosynthetic bacteria, the plasma membrane folds into the cytoplasm to form an internal membrane system that contains bacterial chlorophyll and other compounds needed for photosynthesis. The development of photosynthesis, probably by such internal membranes, was an important event in the early evolution of life on Earth. Other prokaryotes have internal membrane folds that remain attached to the plasma membrane. These *mesosomes* may function in cell division or in various energy-releasing reactions.

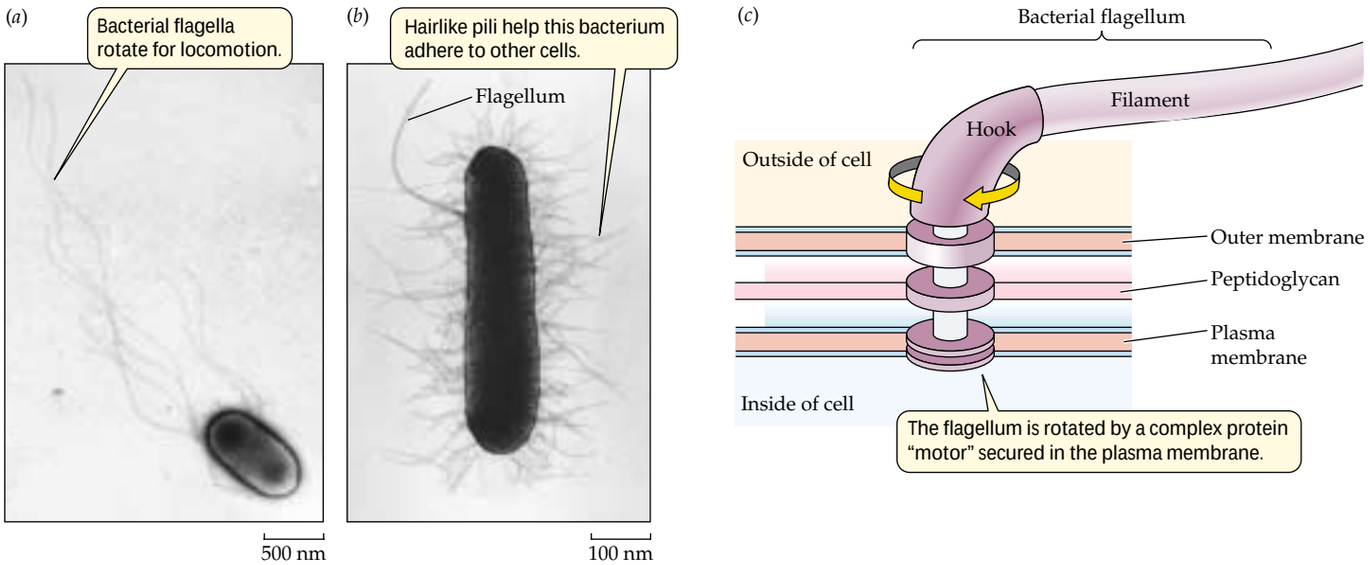
FLAGELLA AND PILI. Some prokaryotes swim by using appendages called *flagella* (Figure 4.6*a*,*c*). A single flagellum, made of a protein called *flagellin*, looks at times like a tiny corkscrew. It spins on its axis like a propeller, driving the cell along. Ring structures anchor the flagellum to the plasma membrane and, in some bacteria, to the outer membrane of the cell wall (Figure 4.6*c*). We know that the flagella cause the motion of the cell because if they are removed, the cell cannot move.

Pili project from the surfaces of some groups of bacteria (Figure 4.6*b*). Shorter than flagella, these threadlike structures help bacteria adhere to one another during mating, as well as to animal cells for protection and food.

CYTOSKELETON. Recent evidence suggests that some prokaryotes, especially rod-shaped bacteria, have an internal filamentous helical structure just below the plasma membrane. The proteins that make up this structure are similar in amino acid sequence to actin in eukaryotic cells, and since actin is part of the cytoskeleton in those cells (see below), it has been suggested that the helical filaments in prokaryotes play a role in cell shape.

Eukaryotic Cells

Animals, plants, fungi, and protists have cells that are usually larger and structurally more complex than those of the prokaryotes. To get a sense of the most prominent differ-



4.6 Prokaryotic Projections Surface projections such as bacterial flagella *(a)* and pili *(b)* contribute to the movement, adhesion, and complexity of prokaryotic cells. *(c)* Complex protein ring structures anchored in the bacterial cell membranes form a motor unit that rotates the flagellum and propels the cell.

ences, compare the eukaryotic plant and animal cells shown in Figure 4.7 with the prokaryotic cell in Figure 4.5.

Eukaryotic cells generally have dimensions ten times greater than those of prokaryotes; for example, the spherical yeast cell has a diameter of 8 µm. Like prokaryotic cells, eukaryotic cells have a plasma membrane, cytoplasm, and ribosomes. But added on to this basic organization are compartments in the cytoplasm whose interiors are separated from the cytosol by a membrane.

Compartmentalization is the key to eukaryotic cell function

Some of the compartments in eukaryotic cells are like little factories that make specific products. Others are like power plants that take in energy in one form and convert it to a more useful form. These membranous compartments, as well as other structures (such as ribosomes) that lack membranes but possess distinctive shapes and functions, are called **organelles** (see Figure 4.7). Each of these organelles has specific roles in its particular cell. These roles are defined by chemical reactions.

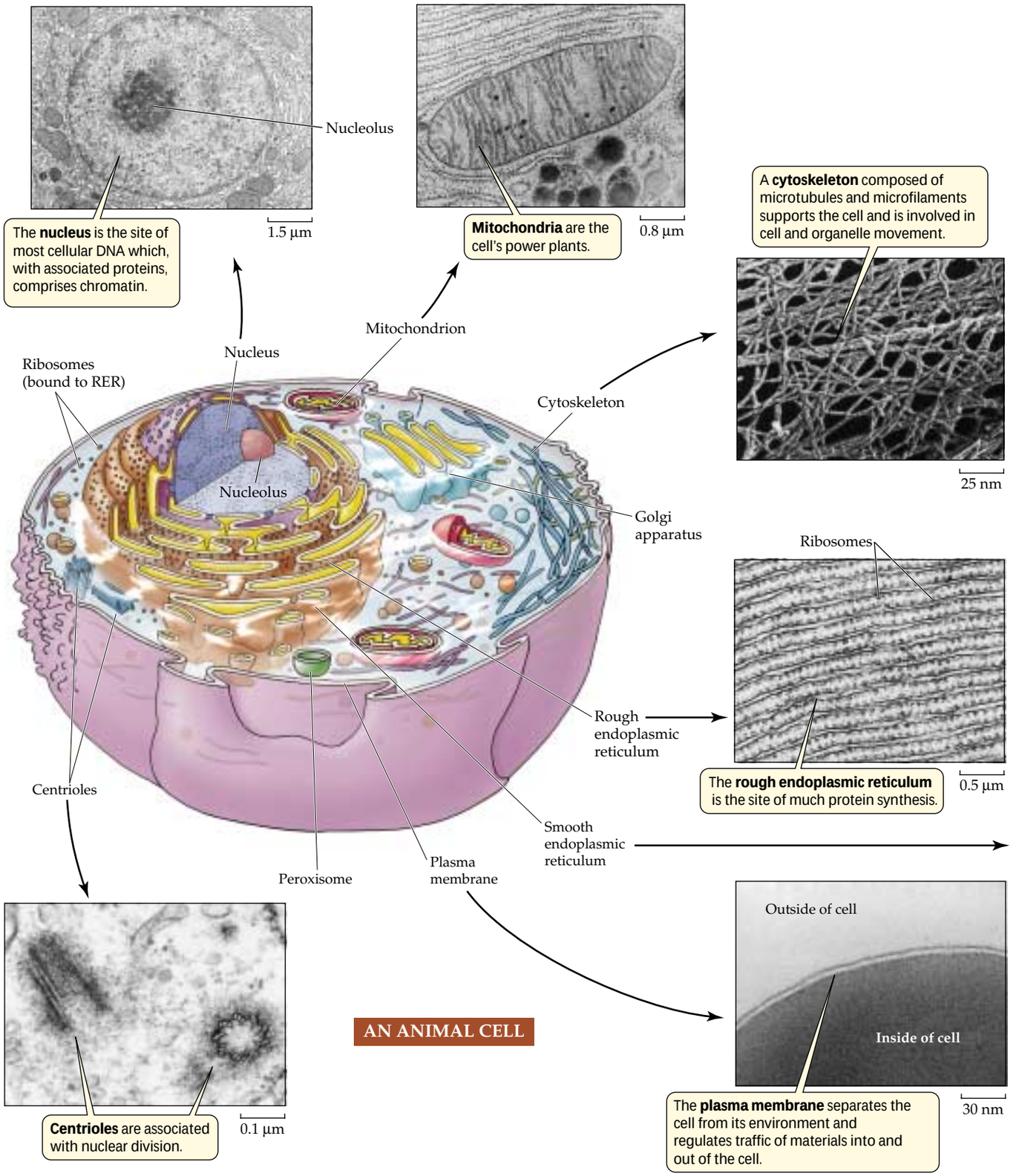
- The **nucleus** contains most of the cell's genetic material (DNA). The duplication of the genetic material and the first steps in decoding genetic information take place in the nucleus.
- The **mitochondrion** is a power plant and industrial park, where energy stored in the bonds of carbohydrates is converted to a form more useful to the cell (ATP) and certain essential biochemical conversions of amino acids and fatty acids occur.
- The **endoplasmic reticulum** and **Golgi apparatus** are compartments in which proteins are packaged and sent to appropriate locations in the cell.
- **Lysosomes** and **vacuoles** are cellular digestive systems in which large molecules are hydrolyzed into usable monomers.
- **Chloroplasts** perform photosynthesis.

The membrane surrounding each organelle does two essential things: First, it keeps the organelle's molecules away from other molecules in the cell with which they might react inappropriately. Second, it acts as a traffic regulator, letting important raw materials into the organelle and releasing its products to the cytoplasm. The evolution of compartmentalization was an important development in the ability of eukaryotic cells to specialize, forming the organs and tissues of a complex body.

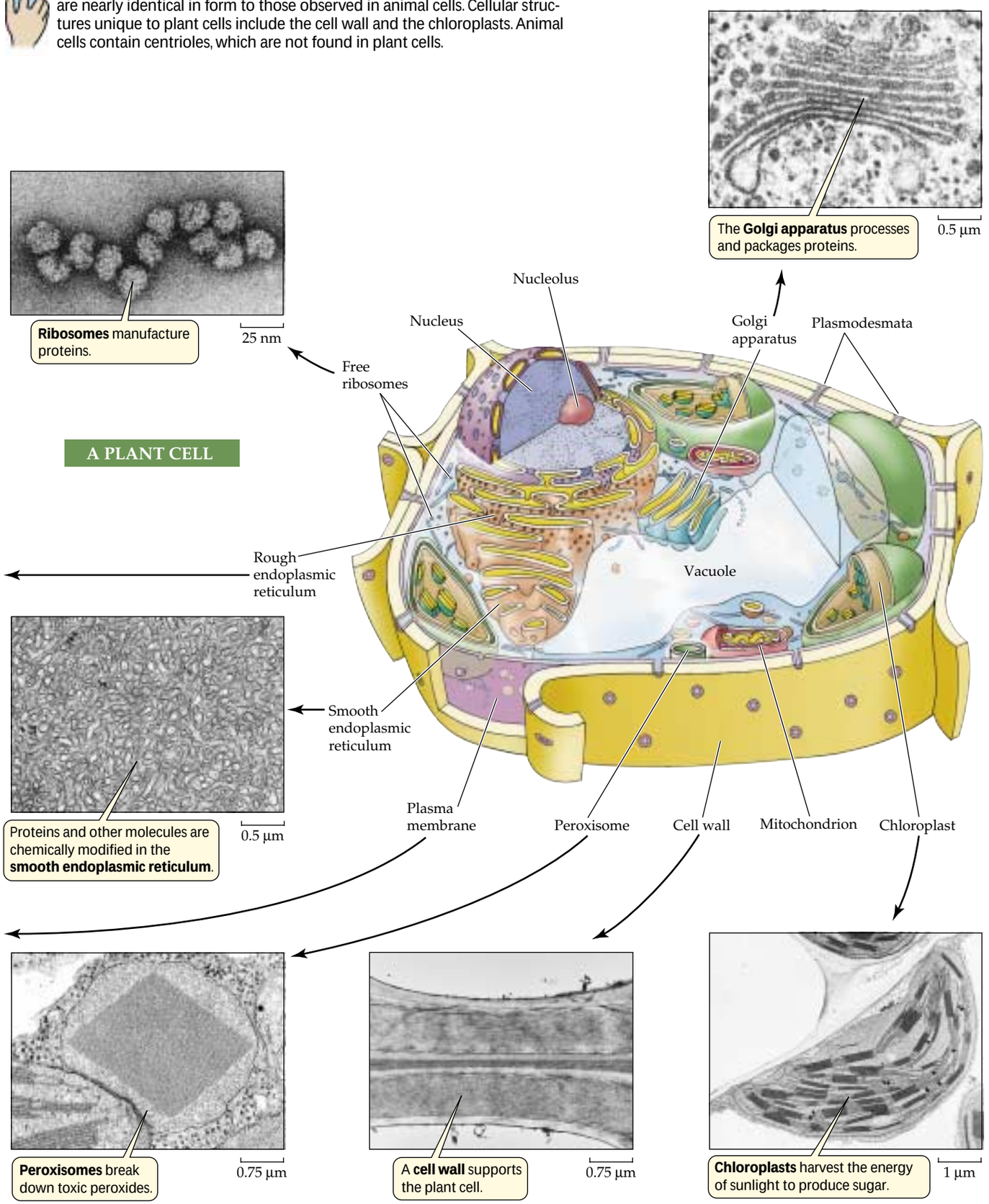
Organelles can be studied by microscopy or isolated for chemical analysis

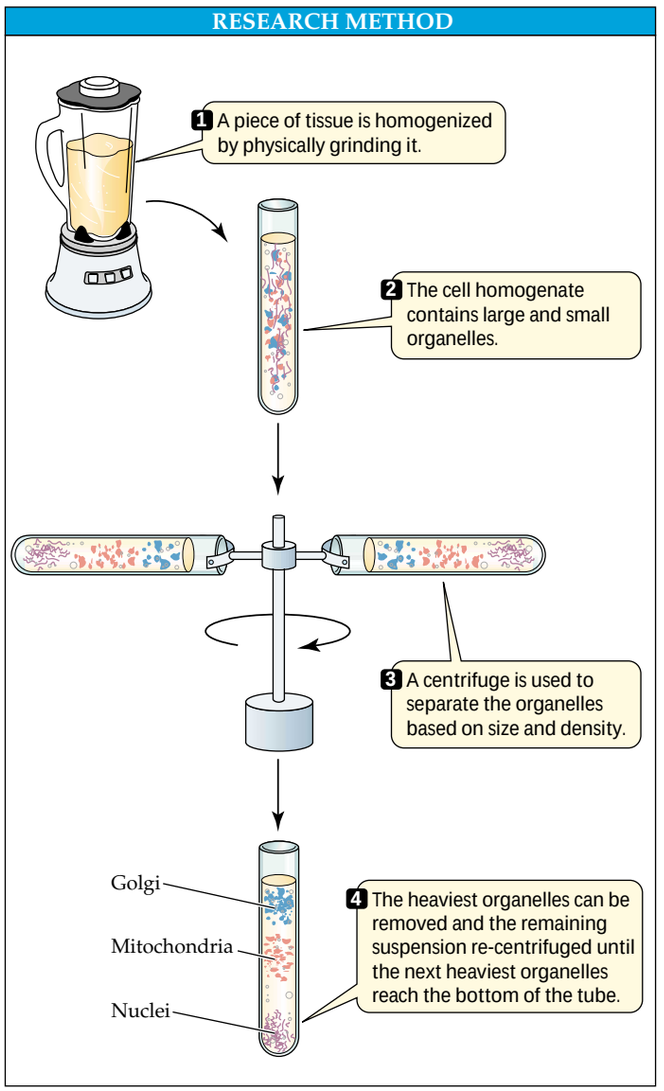
Cell organelles were first detected by light and electron microscopy. The use of stains targeted to specific macromolecules has allowed cell biologists to determine the chemical compositions of organelles. (See Figure 4.21, in which a single cell is stained for three different proteins.)

Besides microscopy, another way to look at cells is to take them apart. **Cell fractionation** begins with the destruction of the cell membrane. This allows the cytoplasmic components


Nucleolus
The **nucleus** is the site of most cellular DNA which, with associated proteins, comprises chromatin.
1.5 µm
Mitochondria are the cell's power plants.
0.8 µm
A **cytoskeleton** composed of microtubules and microfilaments supports the cell and is involved in cell and organelle movement.
25 nm
Mitochondrion
Nucleus
Ribosomes (bound to RER)
Cytoskeleton
Nucleolus
Golgi apparatus
Ribosomes
Rough endoplasmic reticulum
The **rough endoplasmic reticulum** is the site of much protein synthesis.
0.5 µm
Centrioles
Smooth endoplasmic reticulum
Peroxisome
Plasma membrane
Outside of cell
Inside of cell
AN ANIMAL CELL
Centrioles are associated with nuclear division.
0.1 µm
The **plasma membrane** separates the cell from its environment and regulates traffic of materials into and out of the cell.
30 nm

4.7 Eukaryotic Cells In electron micrographs, many plant cell organelles are nearly identical in form to those observed in animal cells. Cellular structures unique to plant cells include the cell wall and the chloroplasts. Animal cells contain centrioles, which are not found in plant cells.





4.8 Cell Fractionation The organelles of cells can be separated from one another after cells are broken open and centrifuged.

to flow out into a test tube. The various organelles can then be separated from one another on the basis of size or density (Figure 4.8). Biochemical analyses can then be done on the isolated organelles. Microscopy and cell fractionation have complemented each other, giving a complete picture of the structure and function of each organelle.

Organelles that Process Information

Living things depend on accurate, appropriate information—internal signals, environmental cues, and stored instructions—to respond appropriately to changing conditions and maintain a constant internal environment. In the cell, information is stored in the sequence of nucleotides in DNA molecules. Most of the DNA in eukaryotic cells resides in the nucleus. Information is translated from the language of DNA into the language of proteins at the ribosomes. This process is described in detail in Chapter 12.

The nucleus contains most of the cell's DNA

The single nucleus is usually the largest organelle in a cell (Figure 4.9; see also Figure 4.7). The nucleus of most animal cells is approximately 5 µm in diameter—substantially larger than most entire prokaryotic cells. The nucleus has several roles in the cell:

- The nucleus is the site of DNA duplication.
- The nucleus is the site of genetic control of the cell's activities.
- A region within the nucleus, the **nucleolus**, begins the assembly of ribosomes from specific proteins and RNA.

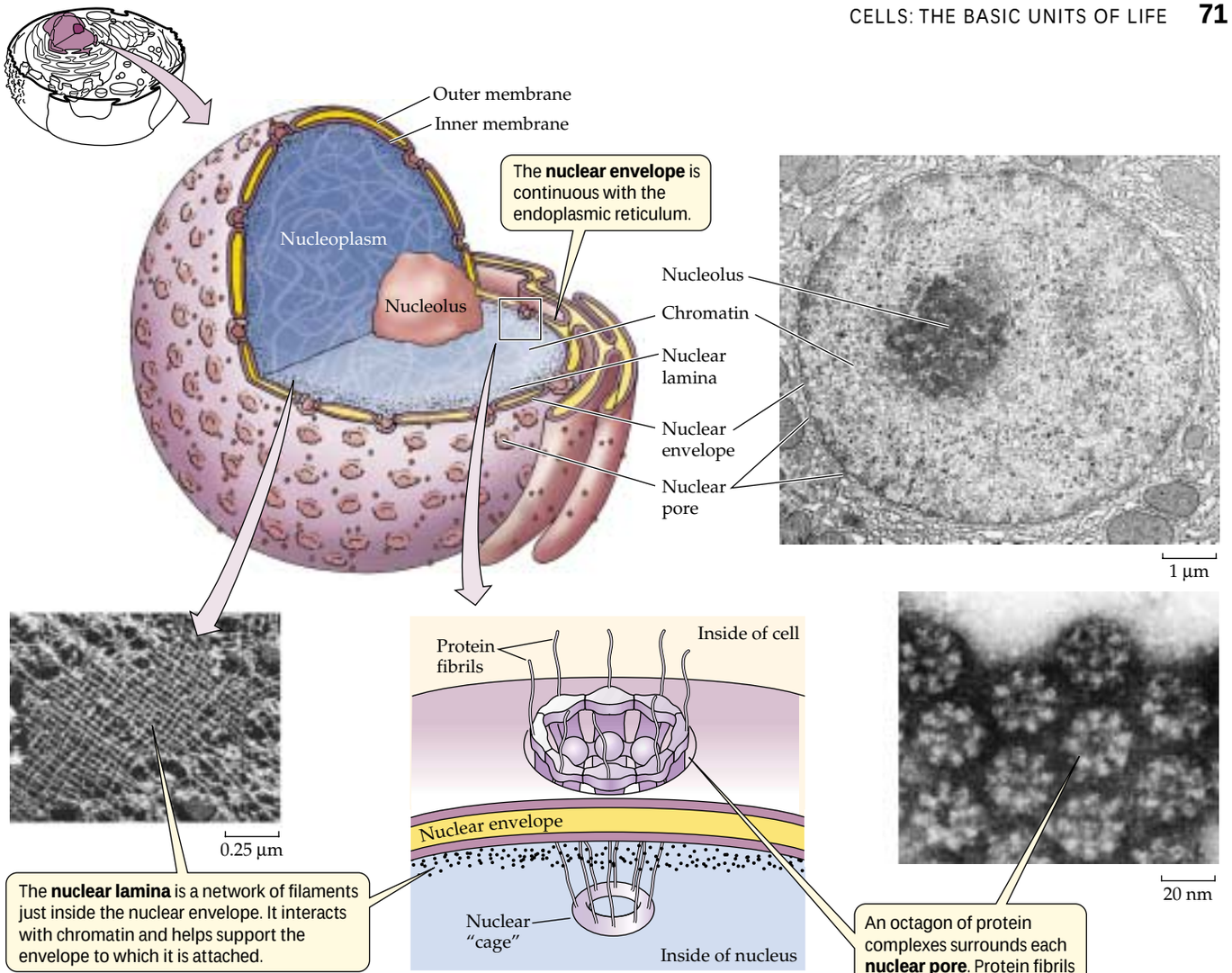
The nucleus is surrounded by two membranes, which together form the **nuclear envelope**. The two membranes of the nuclear envelope are separated by 10–20 nm and are perforated by **nuclear pores** approximately 9 nm in diameter, which connect the interior of the nucleus with the cytoplasm. At these pores, the outer membrane of the nuclear envelope is continuous with the inner membrane. Each pore is surrounded by a pore complex made up of eight large protein granules arranged in an octagon where the inner and outer membranes merge (see Figure 4.9). RNA and proteins pass through these pores to enter or leave the nucleus.

At certain sites, the outer membrane of the nuclear envelope folds outward into the cytoplasm and is continuous with the membrane of another organelle, the endoplasmic reticulum (discussed later in this chapter).

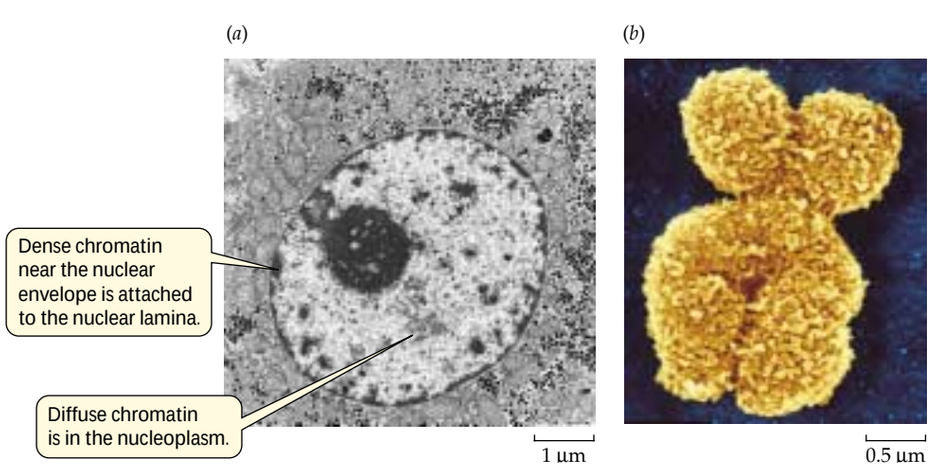
Inside the nucleus, DNA combines with proteins to form a fibrous complex called **chromatin**. Chromatin consists of exceedingly long, thin, entangled threads. Prior to cell division, the chromatin aggregates to form discrete, readily visible structures called **chromosomes** (Figure 4.10).

Surrounding the chromatin are water and dissolved substances collectively referred to as the **nucleoplasm**. Within the nucleoplasm, a network of apparently structural proteins called the *nuclear matrix* organizes the chromatin. At the periphery of the nucleus, the chromatin is attached to a protein meshwork, called the *nuclear lamina*, which is formed by the polymerization of proteins called *lamins* into filaments. The nuclear lamina maintains the shape of the nucleus by its attachment to both the chromatin and the nuclear envelope.

During most of a cell's life cycle, the nuclear envelope is a stable structure. When the cell divides, however, the nuclear envelope fragments into pieces of membrane with attached pore complexes. The envelope re-forms when distribution of the duplicated DNA to the daughter cells is completed.



4.9 The Nucleus Is Enclosed by a Double Membrane The double-membraned nuclear envelope, nucleolus, nuclear lamina, and nuclear pores are common features of all cell nuclei. The pores are the gateways through which proteins from the cytoplasm enter the nucleus and genetic material (mRNA) from the nucleus enters the cytoplasm.



4.10 Chromatin and Chromosomes (*a*) When a cell is not dividing, the nuclear DNA is aggregated with proteins to form chromatin, which is dispersed throughout the nucleus. (*b*) The chromatin in a dividing cell is packed into dense bodies called chromosomes.

Ribosomes are the sites of protein synthesis

In prokaryotic cells, ribosomes float freely in the cytoplasm. In eukaryotic cells they occur in two places: in the cytoplasm, where they may be free or attached to the surface of the endoplasmic reticulum (described in the next section); and inside the mitochondria and chloroplasts, where energy is processed. In each of these locations, the ribosomes are the sites where proteins are synthesized under the direction of nucleic acids. Although they seem small in comparison to the cell in which they are contained, ribosomes are huge machines made up of several dozen kinds of molecules.

The ribosomes of prokaryotes and eukaryotes are similar in that both consist of two different-sized subunits. Eukaryotic ribosomes are somewhat larger, but the structure of prokaryotic ribosomes is better understood. Chemically, ribosomes consist of a special type of RNA, called *ribosomal RNA* (*rRNA*), to which more than 50 different protein molecules are noncovalently bound.

The Endomembrane System

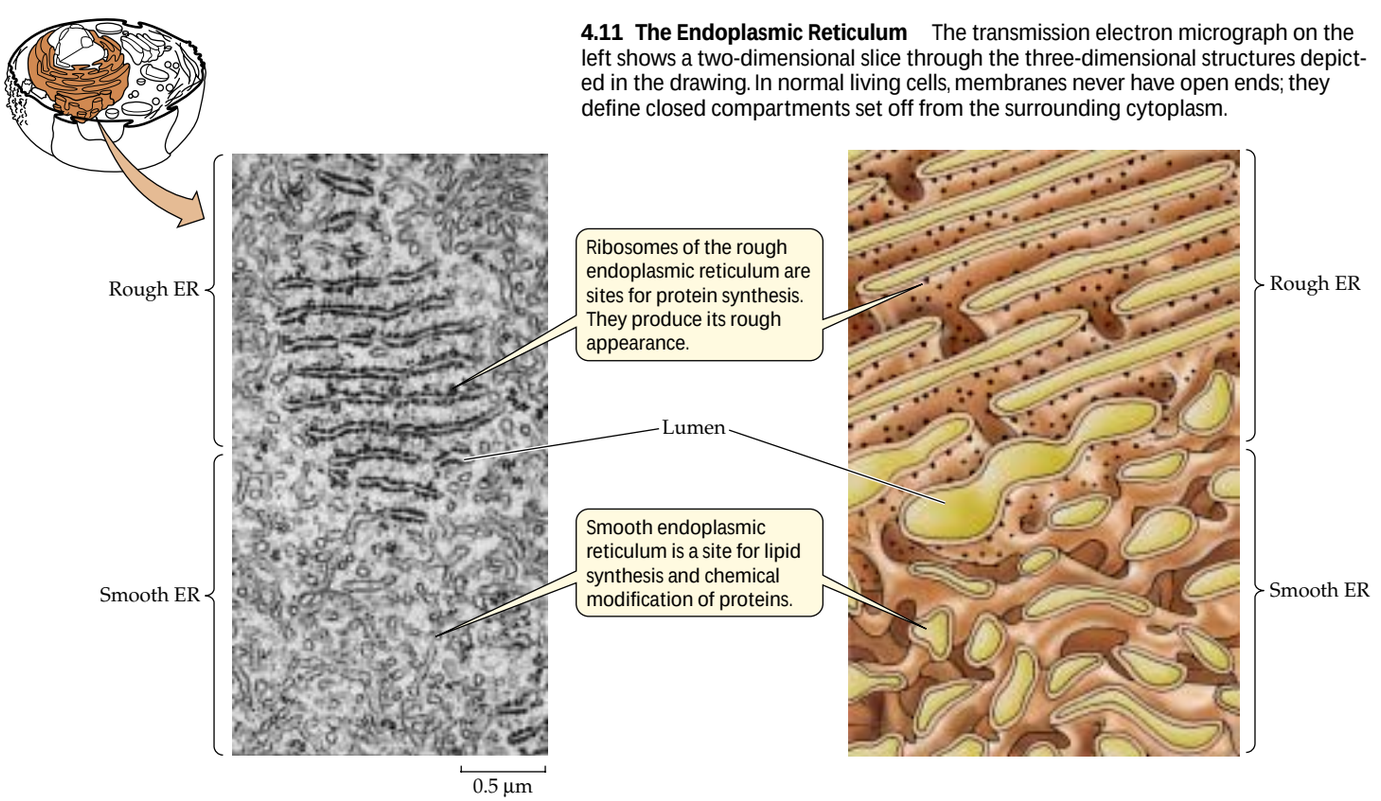
Much of the volume of some eukaryotic cells is taken up by an extensive **endomembrane system**. This system includes two main components, the endoplasmic reticulum and the Golgi apparatus. Continuities between the nuclear envelope and the endomembrane system are visible under the electron microscope. Tiny, membrane-surrounded droplets called **vesicles** appear to shuttle between the various components of the endomembrane system. This system has various structures, but all of them are essentially compartments, closed off by their membranes from the cytoplasm.

In this section, we will examine the functional significance of these compartments, and we will see how materials synthesized in one organelle, the endoplasmic reticulum, are transferred to another organelle, the Golgi apparatus, for further processing, storage, or transport. We will also describe the role of the lysosome in cellular digestion.

The endoplasmic reticulum is a complex factory

Electron micrographs reveal a network of interconnected membranes branching throughout the cytoplasm of a eukaryotic cell, forming tubes and flattened sacs. These membranes are collectively called the **endoplasmic reticulum**, or **ER**. The interior compartment of the ER, referred to as the *lumen*, is separate and distinct from the surrounding cytoplasm (Figure 4.11). The ER can enclose up to 10 percent of the interior volume of the cell, and its foldings result in a surface area many times greater than that of the plasma membrane.

Parts of the ER are studded with ribosomes, which are temporarily attached to the outer faces of its flattened sacs. Because of their appearance under the electron microscope,



4.11 The Endoplasmic Reticulum The transmission electron micrograph on the left shows a two-dimensional slice through the three-dimensional structures depicted in the drawing. In normal living cells, membranes never have open ends; they define closed compartments set off from the surrounding cytoplasm.

these regions are called **rough endoplasmic reticulum**, or **RER**. RER has two roles:

- As a compartment, it segregates certain newly synthesized proteins away from the cytoplasm and transports them to other locations in the cell.
- While inside the RER, proteins can be chemically modified so as to alter their function and eventual destination.

The attached ribosomes are sites for the synthesis of proteins that function outside the cytosol—that is, proteins that are to be exported from the cell, incorporated into membranes, or moved into the organelles of the endomembrane system. These proteins enter the lumen of the ER as they are synthesized. Once in the lumen of the ER, these proteins undergo several changes, including the formation of disulfide bridges and folding into their tertiary structures (see Figure 3.4).

Proteins gain carbohydrate groups in the RER, thus becoming glycoproteins. In the case of proteins directed to the lysosomes, the carbohydrate groups are part of an "addressing" system that ensures that the right proteins are directed to the organelle.

Smooth endoplasmic reticulum or **SER** is more tubular (less like flattened sacs) and lacks ribosomes (see Figure 4.11). Within the lumen of the SER, proteins that have been synthesized on the RER are chemically modified. In addition, the SER has three other important roles:

- It is responsible for chemically modifying small molecules taken in by the cell. This is especially true for drugs and pesticides.
- It is the site for the hydrolysis of glycogen in animal cells.
- It is the site for the synthesis of lipids and steroids.

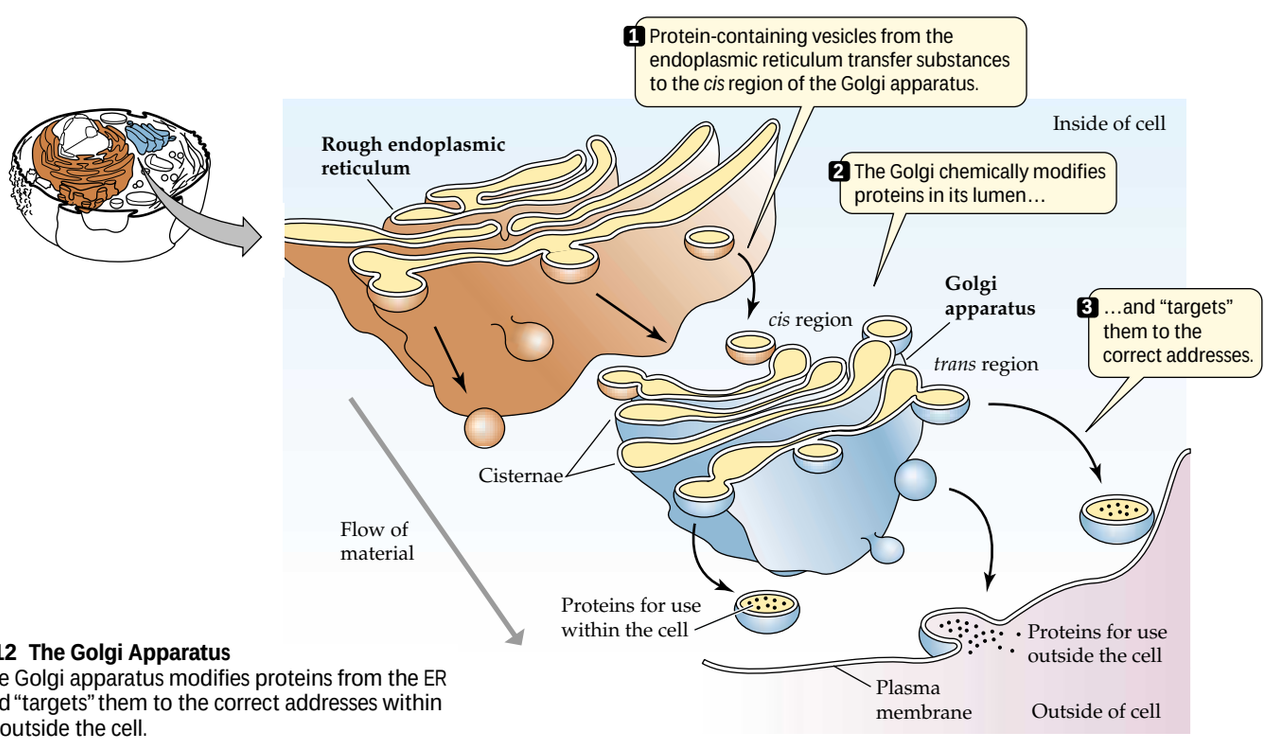
Cells that synthesize a lot of protein for export are usually packed with endoplasmic reticulum. Examples include glandular cells that secrete digestive enzymes and plasma cells that secrete antibodies. In contrast, cells that carry out less protein synthesis (such as storage cells) contain less ER. Liver cells, which modify molecules that enter the body from the digestive system, have abundant smooth ER.

The Golgi apparatus stores, modifies, and packages proteins

The exact appearance of the Golgi apparatus (named for its discoverer, Camillo Golgi) varies from species to species, but it always consists of flattened membranous sacs called *cisternae* and small membrane-enclosed vesicles. The cisternae appear to be lying together like a stack of saucers (Figure 4.12). The entire apparatus is about 1 µm long.

The Golgi apparatus has several roles:

- It receives proteins from the ER and may further modify them.
- It concentrates, packages, and sorts proteins before they are sent to their cellular or extracellular destinations.
- It is where some polysaccharides for the plant cell wall are synthesized.



4.12 The Golgi Apparatus
The Golgi apparatus modifies proteins from the ER and "targets" them to the correct addresses within or outside the cell.

In the cells of plants, protists, fungi, and many invertebrate animals, the stacks of cisternae are individual units scattered throughout the cytoplasm. In vertebrate cells, a few such stacks usually form a larger, single, more complex Golgi apparatus.

The Golgi apparatus appears to have three functionally distinct parts: a bottom, a middle, and a top. The bottom cisternae, constituting the *cis* region of the Golgi apparatus, lie nearest to the nucleus or a patch of RER (see Figure 4.12). The top cisternae, constituting the *trans* region, lie closest to the surface of the cell. The cisternae in the middle make up the *medial* region of the complex. These three parts of the Golgi apparatus contain different enzymes and perform different functions.

The Golgi apparatus receives proteins from the ER, packages them, and sends them on their way. Since there is often no direct membrane continuity between ER and Golgi apparatus, how does a protein get from one organelle to the other? The protein could simply leave the ER, travel across the cytoplasm, and enter the Golgi apparatus. But that would expose the protein to interactions with other molecules in the cytoplasm. On the other hand, segregation from the cytoplasm could be maintained if a piece of the ER could "bud off," forming a vesicle that contains the protein—and that is exactly what happens. The protein makes the passage from ER to Golgi apparatus safely enclosed in the vesicle. Once it arrives, the vesicle fuses with the membrane of the Golgi apparatus, releasing its cargo.

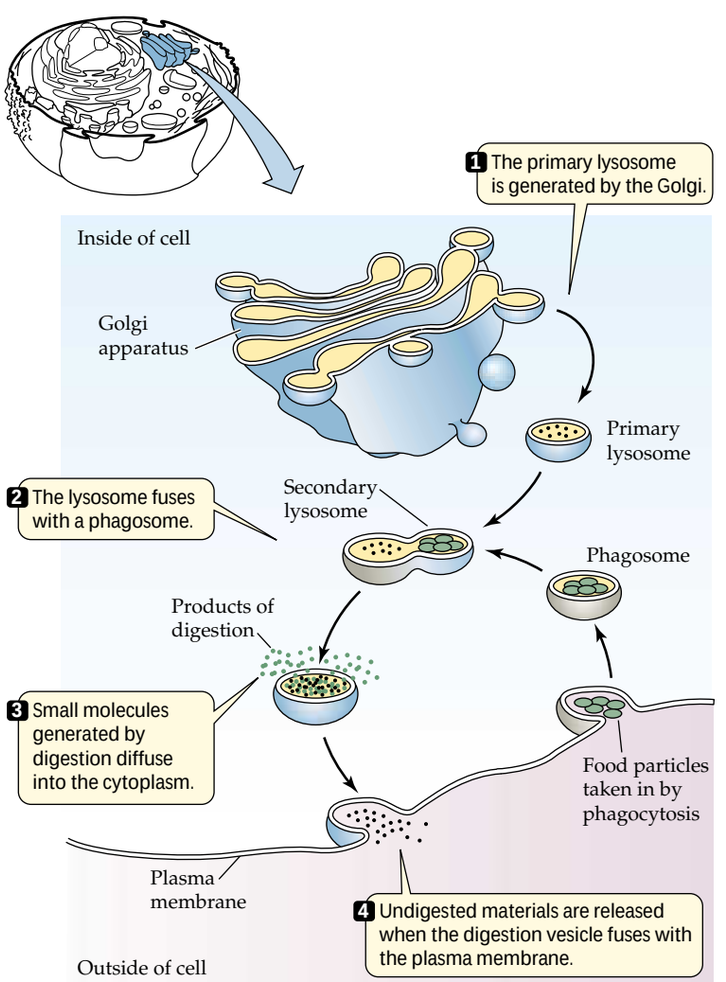
Vesicles form from the rough ER, move through the cytoplasm, and fuse with the *cis* region of the Golgi apparatus, releasing their contents into the lumen. The vesicles may not have far to go: If living cells are stained specifically for ER and Golgi apparatus, the Golgi apparatus can be seen moving rapidly along the ER, possibly picking up vesicles as they go. Other small vesicles may move between the cisternae, transporting proteins, and it appears that some proteins move from one cisterna to the next by tiny channels. Vesicles budding off from the *trans* region carry their contents away from the complex (see Figure 4.12).

Lysosomes contain digestive enzymes

Originating in part from the Golgi apparatus are organelles called **lysosomes**. They contain digestive enzymes, and they are the sites where macromolecules—proteins, polysaccharides, nucleic acids, and lipids—are hydrolyzed into their monomers (see Figure 3.3). Lysosomes are about 1 µm in diameter, are surrounded by a single membrane, and have a densely staining, featureless interior (Figure 4.13). There may be dozens of lysosomes in a cell, depending on its needs.

Lysosomes are sites for the breakdown of food and foreign objects taken up by the cell. These materials get into the cell by a process called **phagocytosis** (*phago-*, "eating"; *cytosis*, "cellular"), in which a pocket forms in the plasma membrane and eventually deepens and encloses material from outside the cell. This pocket becomes a small vesicle that breaks free of the plasma membrane to move into the cytoplasm as a *phagosome* containing food or other material (see Figure 4.13). The phagosome fuses with a *primary lysosome*, forming a *secondary lysosome* where digestion occurs.

The effect of this fusion is rather like releasing hungry foxes into a chicken coop: The enzymes in the secondary lysosome quickly hydrolyze the food particles. These reactions are enhanced by the mild acidity of the lysosome's interior, where the pH is lower than in the surrounding cytoplasm. The products of digestion exit through the membrane of the lysosome, providing fuel molecules and raw materials for other cell processes. The "used" secondary lysosome, now



4.13 Lysosomes Isolate Digestive Enzymes from the Cytoplasm
Lysosomes are sites for the hydrolysis of material taken into the cell by phagocytosis.

containing undigested particles, then moves to the plasma membrane, fuses with it, and releases the undigested contents to the environment.

Lysosomes are also where the cell digests its own material in a process called **autophagy**. Autophagy is an ongoing process in which organelles such as mitochondria are engulfed by lysosomes and hydrolyzed to monomers, which pass out of the lysosome through its membrane into the cytoplasm for reuse.

The importance of lysosome function is indicated by a group of human diseases called *lysosomal storage diseases*. If a cell lacks the ability to hydrolyze one or more macromolecules, these substances pile up in lysosomes, with harmful consequences. An example is Tay-Sachs disease, in which a lipid accumulates in the lysosomes of brain cells, resulting in death in early childhood.

Plant cells do not appear to contain lysosomes, but the central vacuole of a plant cell (which we will describe below) may function in an equivalent capacity because it, like lysosomes, contains many digestive enzymes.

Organelles that Process Energy

A cell uses energy to synthesize cell-specific materials that it can use for activities such as growth, reproduction, and movement. Energy is transformed from one form to another in mitochondria (found in all eukaryotic cells) and in chloroplasts (found in eukaryotic cells that harvest energy from sunlight). In contrast, energy transformations in prokaryotic cells are associated with enzymes attached to the inner surface of the plasma membrane or extensions of the plasma membrane that protrude into the cytoplasm.

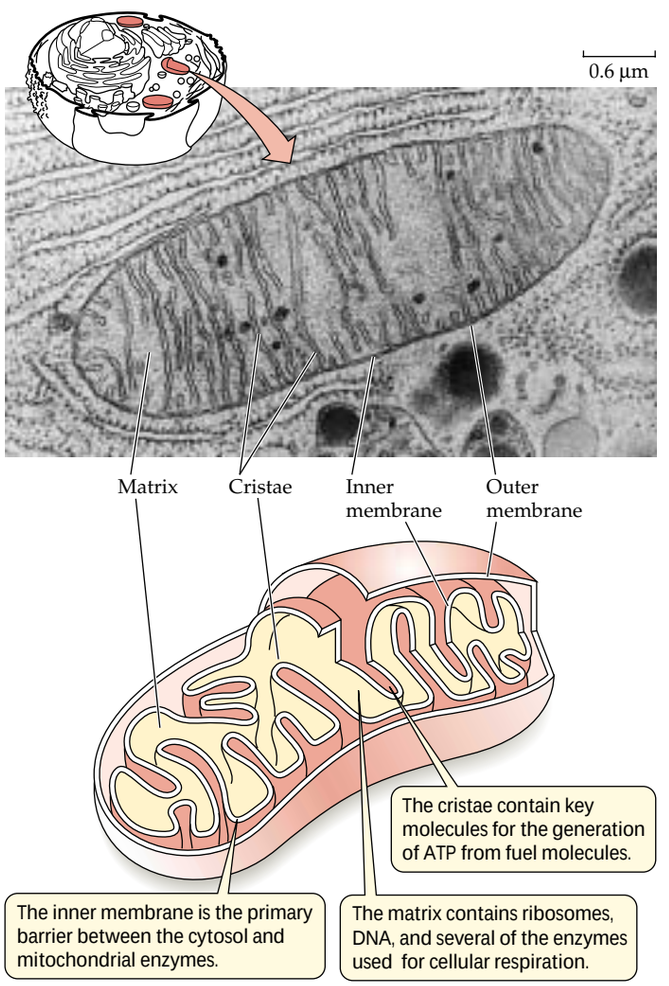
Mitochondria are energy transformers

In eukaryotic cells, the breakdown of fuel molecules such as glucose begins in the cytosol. The molecules that result from this partial degradation enter the mitochondria (singular, mitochondrion), whose primary function is to convert the potential chemical energy of those fuel molecules into a form that the cell can use: the energy-rich molecule ATP (adenosine triphosphate). The production of ATP in the mitochondria using fuel molecules and molecular oxygen (O_2) is called **cellular respiration**.

Typical mitochondria are small—somewhat less than 1.5 µm in diameter and 2–8 µm in length—about the size of many bacteria. The number of mitochondria per cell ranges from one contorted giant in some unicellular protists to a few hundred thousand in large egg cells. An average human liver cell contains more than a thousand mitochondria. Cells that require the most chemical energy tend to have the most mitochondria per unit of volume.

Mitochondria have two membranes. The *outer membrane* is smooth and protective, and it offers little resistance to the movement of substances into and out of the mitochondrion. Immediately inside the outer membrane is an *inner membrane*, which folds inward in many places, giving it a much greater surface area than that of the outer membrane (Figure 4.14). These folds tend to be quite regular, giving rise to shelflike structures called *cristae*.

The inner mitochondrial membrane contains many large protein molecules that participate in cellular respiration. The inner membrane exerts much more control over what enters and leaves the mitochondrion than does the outer membrane. The region enclosed by the inner membrane is referred to as the *mitochondrial matrix*. In addition to many proteins, the matrix contains some ribosomes and DNA that are used to make some of the proteins needed for cellular respiration.



4.14 A Mitochondrion Converts Energy from Fuel Molecules into ATP The electron micrograph is a two-dimensional slice through a three-dimensional organelle. As the drawing emphasizes, the cristae are extensions of the inner mitochondrial membrane.

In Chapter 7 we will see how the different parts of the mitochondrion work together in cellular respiration.

Plastids photosynthesize or store materials

One class of organelles—the **plastids**—is produced only in plants and certain protists. There are several types of plastids, with different functions.

CHLOROPLASTS. Chloroplasts contain the green pigment **chlorophyll** and are the sites of photosynthesis (Figure 4.15). In **photosynthesis**, light energy is converted into the chemical energy of bonds between atoms. The molecules formed in photosynthesis provide food for the photosynthetic organisms, as well as for other organisms that eat them. Directly or indirectly, photosynthesis is the energy source for most of the living world.

Chloroplasts are variable in size and shape (Figure 4.16). Like a mitochondrion, a chloroplast is surrounded by two membranes. In addition, there is a series of internal membranes whose structure and arrangement vary from one group of photosynthetic organisms to another. Here we concentrate on the chloroplasts of the flowering plants. Even these chloroplasts show some variation, but the pattern shown in Figure 4.15 is typical.

The internal membranes of chloroplasts look like stacks of flat, hollow pita bread. These stacks, called **grana** (singular, granum), consist of a series of flat, closely packed, circular compartments called **thylakoids**. In addition to phospholipids and proteins, the membranes of the thylakoids contain chlorophyll and other pigments that harvest light for photosynthesis. The thylakoids of one granum may be connected to those of other grana, making the interior of the chloroplast a highly developed network of membranes, much like the ER.

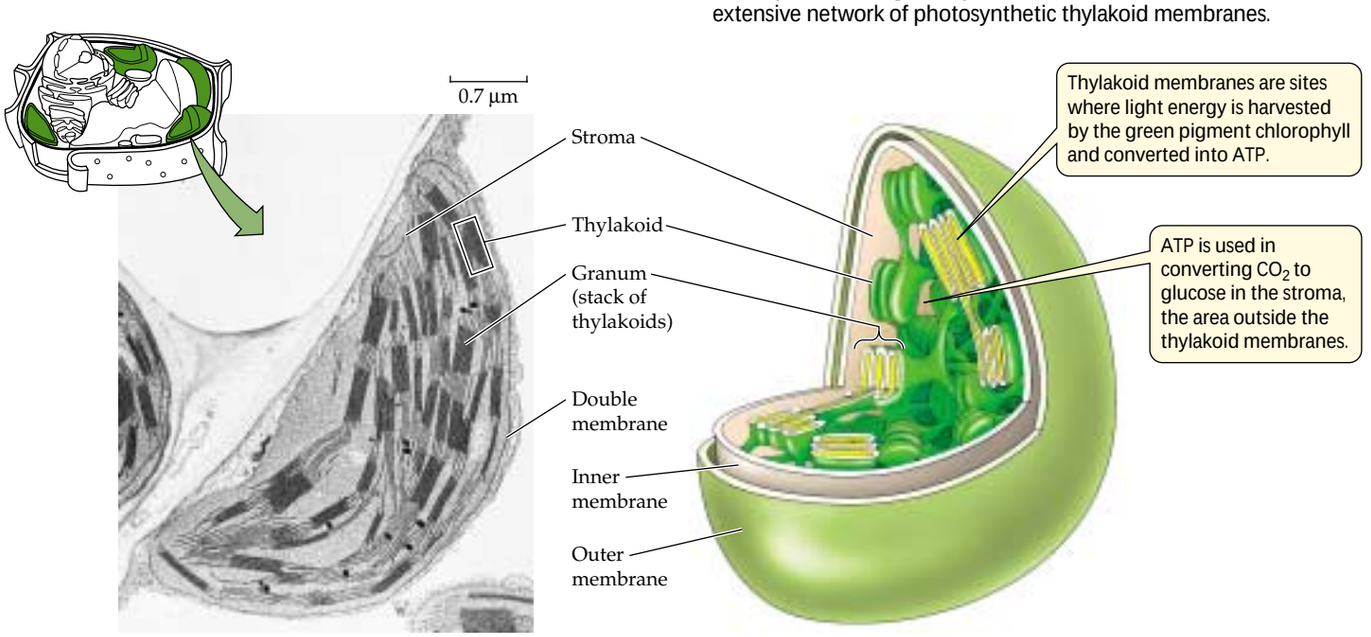
The fluid in which the grana are suspended is the **stroma**. Like the mitochondrial matrix, the chloroplast stroma contains ribosomes and DNA, which are used to synthesize some, but not all, of the proteins that make up the chloroplast.

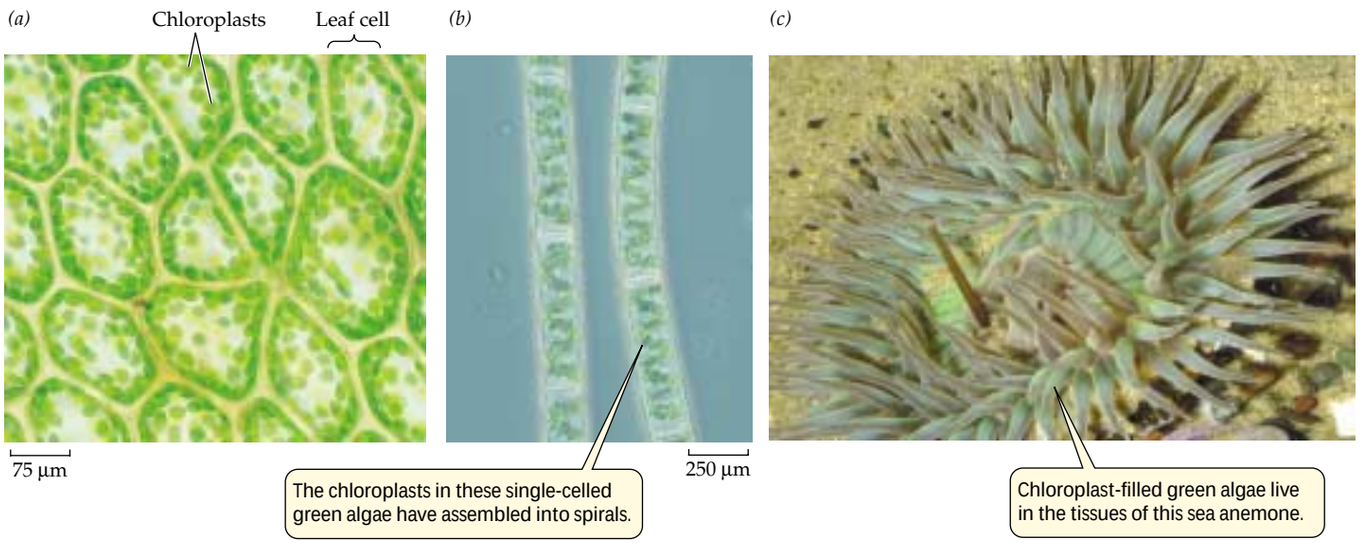
Animal cells do not produce chloroplasts, but some do contain functional chloroplasts. These are either taken up as free chloroplasts derived from the partial digestion of green plants or contained within unicellular algae that live within the animal's tissues. The green color of some corals and sea anemones results from the chloroplasts in algae that live within those animals (Figure 4.16*c*). The animals derive some of their nutrition from the photosynthesis that their chloroplast-containing "guests" carry out. Such an intimate relationship between two different organisms is called **symbiosis**.

OTHER TYPES OF PLASTIDS. Other types of plastids also store pigments or polysaccharides:

- *Chromoplasts* contain red, orange, and/or yellow pigments and give color to plant organs such as flowers (Figure 4.17*a*). The chromoplasts have no known chemical func-

4.15 The Chloroplast: The Organelle That Feeds the World The electron micrograph shows a chloroplast from a leaf of corn. Chloroplasts are large compared with mitochondria and contain an extensive network of photosynthetic thylakoid membranes.





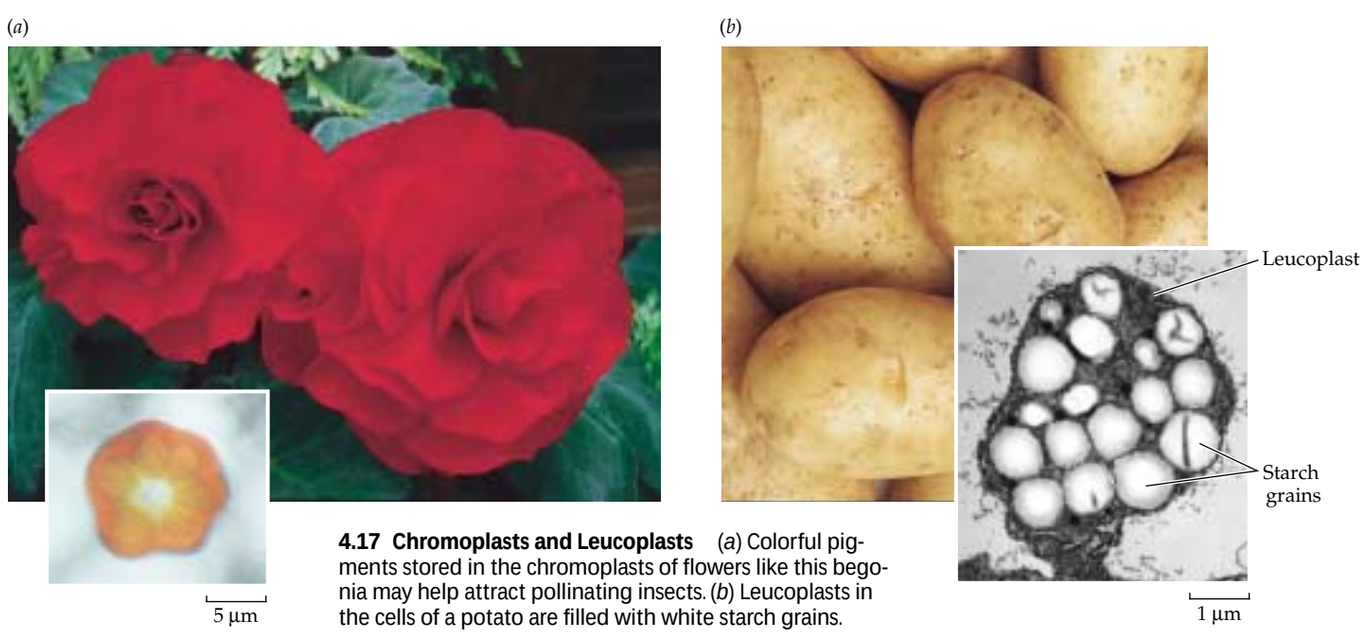
4.16 Being Green (*a*) In green plants, chloroplasts are concentrated in the leaf cells. (*b*) Green algae are photosynthetic and filled with chloroplasts. (*c*) No animal species produces its own chloroplasts, but in this symbiotic arrangement, unicellular green algae nourish a sea anemone.

tion in the cell, but the colors they give to some petals and fruits probably encourage animals to visit flowers and thus aid in pollination, or to eat fruits and thus aid in seed dispersal. (On the other hand, carrot roots gain no apparent advantage from being orange.)

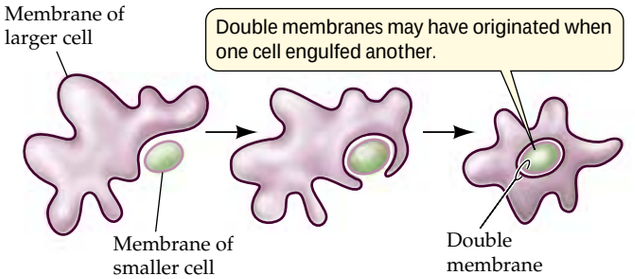
- *Leucoplasts* are storage depots for starch and fats (Figure 4.17*b*).

Endosymbiosis may explain the origin of mitochondria and chloroplasts

Although chloroplasts and mitochondria are about the size of prokaryotic cells and have the genetic material and protein synthesis machinery needed to make some of their own components, they are not independent of control by the nucleus. The vast majority of their proteins are encoded by nuclear DNA, made in the cytoplasm, and imported into the organelle. Observations of these organelles have led to the proposal that they originated by endosymbiosis—that is, that they were once independent prokaryotic organisms.



4.17 Chromoplasts and Leucoplasts (*a*) Colorful pigments stored in the chromoplasts of flowers like this begonia may help attract pollinating insects. (*b*) Leucoplasts in the cells of a potato are filled with white starch grains.



4.18 The Endosymbiosis Theory Chloroplasts and mitochondria may be descended from a small prokaryote that was engulfed by another, larger prokaryote.

About 2 billion years ago, only prokaryotes inhabited Earth. Some of them absorbed their food directly from the environment. Others were photosynthetic. Still others fed on smaller prokaryotes by engulfing them (Figure 4.18).

Suppose that a small, photosynthetic prokaryote was ingested by a larger one, but was not digested. Instead, it somehow survived, trapped within a vesicle in the cytoplasm of the larger cell. The smaller, ingested prokaryote divided at about the same rate as the larger one, so successive generations of the larger cell also contained the offspring of the smaller one. This phenomenon, called **endosymbiosis** (*endo-*, "within"; *symbiosis*, "living together"), exists today, as in the case of the algae that live within sea anemones (see Figure 4.16*c*).

According to this scenario, endosymbiosis provided benefits for both partners: The larger cell obtained the photosynthetic products from the smaller cell, and the smaller cell was protected by the larger one. Over evolutionary time, the smaller cell gradually lost much of its DNA to the nucleus of the larger cell, resulting in the modern chloroplast.

Much circumstantial evidence favors the endosymbiosis theory:

- On an evolutionary time scale of millions of years, there is evidence for DNA moving between organelles in the cell.
- There are many biochemical similarities between chloroplasts and modern photosynthetic bacteria.
- DNA sequencing shows strong similarities between modern chloroplast DNA and that of a photosynthetic prokaryote.
- The double membrane that encloses mitochondria and chloroplasts could have arisen through endosymbiosis. The outer membrane may have come from the engulfing cell's plasma membrane and the inner membrane from the engulfed cell's plasma membrane.

Similar evidence and arguments also support the proposition that mitochondria are the descendants of respiring prokaryotes engulfed by larger prokaryotes. The benefit of this endosymbiotic relationship might have been the capacity of the engulfed prokaryote to detoxify molecular oxygen (O_2), which was increasing in Earth's atmosphere as a result of photosynthesis.

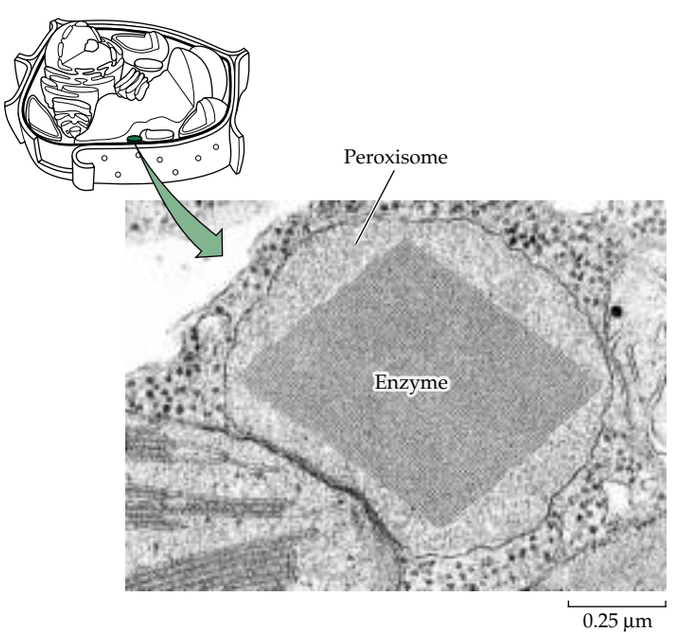
Other Organelles

Eukaryotic cells have several other organelles that are surrounded by a single membrane.

Peroxisomes house specialized chemical reactions

Peroxisomes are organelles that collect the toxic peroxides (such as hydrogen peroxide, H_2O_2) that are the unavoidable by-products of chemical reactions. These peroxides can be safely broken down inside the peroxisomes without mixing with other parts of the cell. Peroxisomes are small organelles, about 0.2 to 1.7 µm in diameter. They have a single membrane and a granular interior containing specialized enzymes (Figure 4.19). Peroxisomes are found at one time or another in at least some of the cells of almost every eukaryotic species.

A structurally similar organelle, the **glyoxysome**, is found only in plants. Glyoxysomes, which are most prominent in young plants, are the sites where stored lipids are converted into carbohydrates for transport to growing cells.

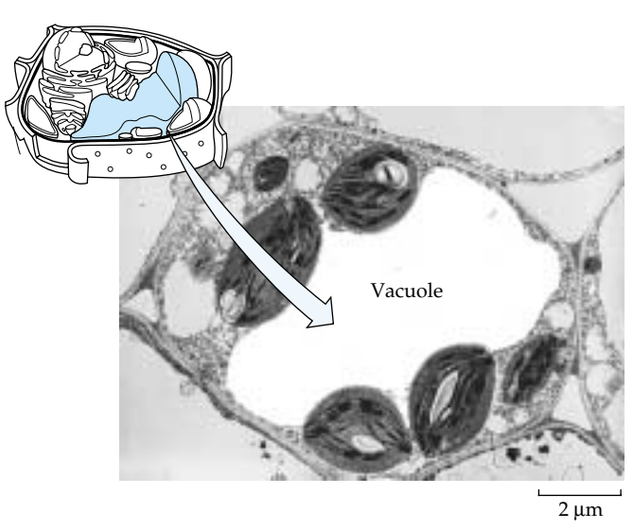


4.19 A Peroxisome A diamond-shaped crystal, composed of an enzyme, almost entirely fills this rounded peroxisome in a leaf cell. The enzyme catalyzes one of the reactions that breaks down toxic peroxides in the peroxisome.

Vacuoles are filled with water and soluble substances

Many eukaryotic cells, but particularly those of plants and protists, contain membrane-enclosed **vacuoles** filled with aqueous solutions containing many dissolved substances (Figure 4.20). Plant vacuoles have several functions:

- *Storage:* Plant cells produce a number of toxic by-products and waste materials, many of which are simply stored within vacuoles. And since they are poisonous or distasteful, these stored materials deter some animals from eating the plants. Thus these stored wastes may contribute to plant survival.
- *Structure:* In many plant cells, enormous vacuoles take up more than 90 percent of the cell volume and grow as the cell grows. The dissolved substances in the vacuole, working together with the vacuolar membrane, provide the *turgor*, or stiffness, of the cell, which in turn provides support for the structure of nonwoody plants. The presence of the dissolved substances causes water to enter the vacuole, making it swell like a balloon. Plant cells have a rigid cell wall, which resists the swelling of the vacuole, providing strength in the process.
- *Reproduction:* Some pigments (especially blue and pink ones) in petals and fruits are contained in vacuoles. These pigments—the *anthocyanins*—are visual cues that help attract the animals that assist in pollination or seed dispersal.
- *Digestion:* In some plants, the vacuoles contain enzymes that hydrolyze seed proteins into monomers that a developing plant embryo can use as food.



4.20 Vacuoles in Plant Cells Are Usually Large The large central vacuole in this cell is typical of mature plant cells. Smaller vacuoles are visible toward each end of the cell.

Food vacuoles are found in some simple and evolutionarily ancient groups of organisms—single-celled protists and simple multicellular organisms such as sponges, for example. In these organisms, the cells engulf food particles by phagocytosis, generating a food vacuole. Fusion of this vacuole with a lysosome results in digestion, and small molecules leave the vacuole and enter the cytoplasm for use or distribution to other organelles.

Contractile vacuoles are found in many freshwater protists. Their function is to get rid of the excess water that rushes into the cell because of the imbalance in salt concentration between the relatively salty interior of the cell and its freshwater environment. The contractile vacuole enlarges as water enters, then abruptly contracts, forcing the water out of the cell through a special pore structure.

The Cytoskeleton

In addition to its many membrane-enclosed organelles, the eukaryotic cytoplasm contains a set of long, thin fibers called the **cytoskeleton**. The cytoskeleton fills at least three important roles:

- It maintains cell shape and support.
- It provides for various types of cellular movement.
- Some of its fibers act as tracks or supports for motor proteins, which help move things within the cell.

In the discussion that follows, we'll look at three components of the cytoskeleton: microfilaments, intermediate filaments, and microtubules (Figure 4.21).

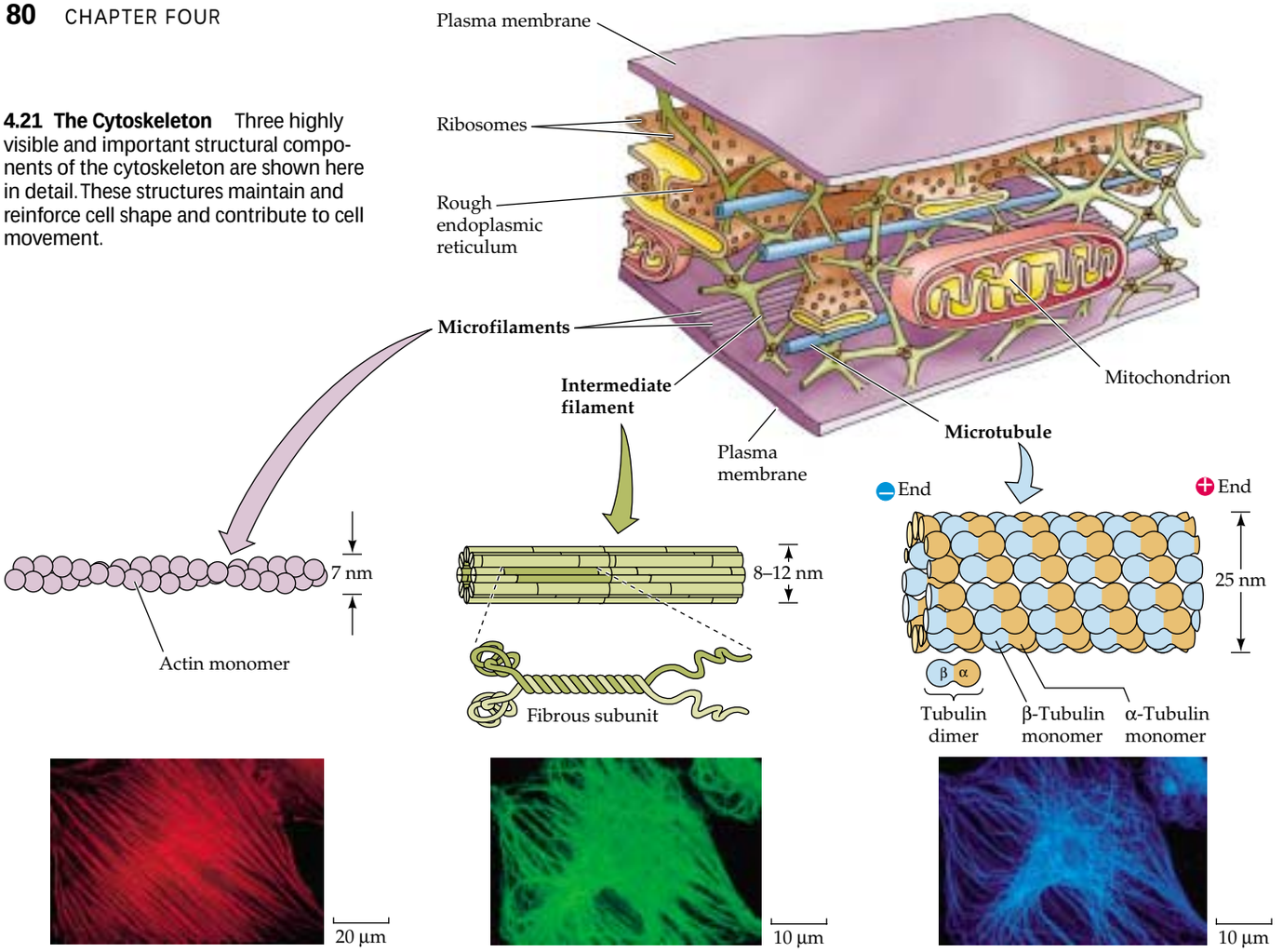
Microfilaments function in support and movement

Microfilaments can exist as single filaments, in bundles, or in networks. They are about 7 nm in diameter and several micrometers long. They are assembled from **actin**, a protein that exists in several forms and has many functions among members of the animal phyla. The actin found in microfilaments (which are also known as *actin filaments*) is extensively folded and has distinct "head" and "tail" sites. These sites interact with other actin molecules to form long, double helical chains (see Figure 4.21). The polymerization of actin into microfilaments is reversible, and they can disappear from cells, breaking down into units of free actin.

Microfilaments have two major roles:

- They help the entire cell or parts of the cell to move.
- They stabilize cell shape.

In muscle cells, actin fibers are associated with another protein, **myosin**, and the interactions of these two proteins account for the contraction of muscles. In non-muscle cells, actin fibers are associated with localized changes of shape in the cell.



4.21 The Cytoskeleton Three highly visible and important structural components of the cytoskeleton are shown here in detail. These structures maintain and reinforce cell shape and contribute to cell movement.

Microfilaments are made up of strands of the protein actin and often interact with strands of other proteins. They change cell shape and drive cellular motion, including contraction, cytoplasmic streaming, and the "pinched" shape changes that occur during cell division. Microfilaments and myosin strands together drive muscle action.

Intermediate filaments are made up of fibrous proteins organized into tough, ropelike assemblages that stabilize a cell's structure and help maintain its shape. Some intermediate filaments help to hold neighboring cells together. Others make up the nuclear lamina.

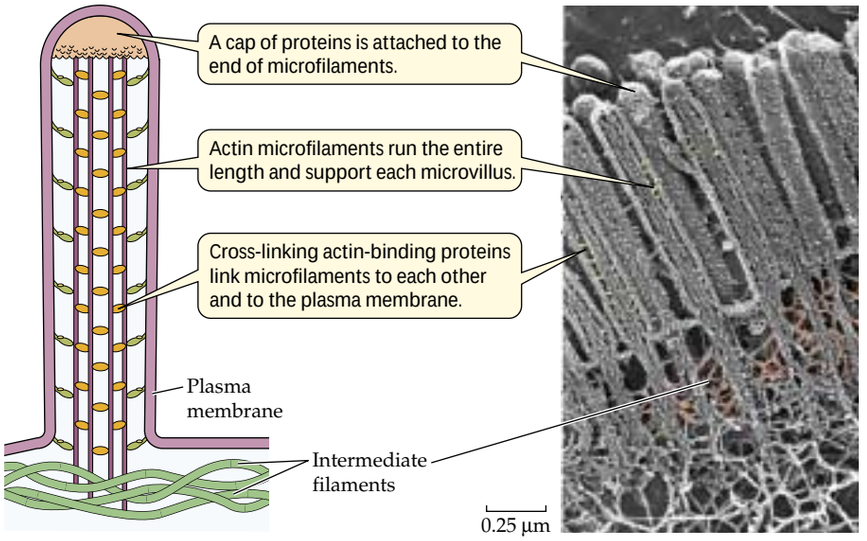
Microtubules are long, hollow cylinders made up of many molecules of the protein tubulin. Tubulin consists of two subunits, α-tubulin and β-tubulin. Microtubules lengthen or shorten by adding or subtracting tubulin dimers. Microtubule shortening moves chromosomes. Interactions between microtubules drive the movement of cells. Microtubules serve as "tracks" for the movement of vesicles.

For example, microfilaments are involved in a flowing movement of the cytoplasm called *cytoplasmic streaming* and in the "pinching" contractions that divide an animal cell into two daughter cells. Microfilaments are also involved in the formation of cellular extensions called *pseudopodia* (*pseudo-*, "false;" *podia*, "feet") that enable some cells to move.

In some cell types, microfilaments form a meshwork just inside the plasma membrane. Actin-binding proteins then cross-link the microtubules to form a rigid structure that supports the cell. For example, microfilaments support the tiny microvilli that line the intestine, giving it a larger surface area through which to absorb nutrients (Figure 4.22).

Intermediate filaments are tough supporting elements

Intermediate filaments (see Figure 4.21) are found only in multicellular organisms. In contrast to the other components of the cytoskeleton, there are at least 50 different kinds of intermediate filaments, often specific to a few cell types. They generally fall into six molecular classes, based on amino acid sequence, and share the same general structure, being composed of fibrous proteins of the keratin family, similar to the protein that makes up hair and fingernails. In cells, these proteins are organized into tough, ropelike assemblages 8 to 12 nm in diameter.



4.22 Microfilaments for Support Microfilaments form the backbone of the microvilli that increase the surface area of some cells, such as intestinal cells that absorb nutrients.

Intermediate filaments have two major structural functions:

- They stabilize cell structure.
- They resist tension.

In some cells, intermediate filaments radiate from the nuclear envelope and may maintain the positions of the nucleus and other organelles in the cell. The lamins of the nuclear lamina are intermediate filaments. Other kinds of intermediate filaments help hold a complex apparatus of microfilaments in place in muscle cells. Still other kinds stabilize and help maintain rigidity in surface tissues by connecting "spot welds" called *desmosomes* between adjacent cells (see Figure 5.6).

Microtubules are long and hollow

Microtubules are long, hollow, unbranched cylinders about 25 nm in diameter and up to several micrometers long. Microtubules have two roles in the cell:

- They form a rigid internal skeleton for some cells.
- They act as a framework along which motor proteins can move structures in the cell.

Microtubules are assembled from molecules of the protein **tubulin**. Tubulin is a *dimer*—a molecule made up of two monomers. The polypeptide monomers that make up this protein are known as *α-tubulin* and *β-tubulin*. Thirteen chains of tubulin dimers surround the central cavity of the microtubule (see Figure 4.21). The two ends of a microtubule are different. One end is designated the plus (+) end, the other the minus (–) end.

Tubulin dimers can be added or subtracted, mainly at the plus end, lengthening or shortening the microtubule. This capacity to change length rapidly makes microtubules *dynamic structures*. This dynamic property of microtubules is seen in animal cells, where they are often found in parts of the cell that are changing shape.

Many microtubules radiate from a region of the cell called the *microtubule organizing center*. Tubule polymerization results in rigidity, and tubule depolymerization leads to a collapse of this rigid structure.

In plants, microtubules help control the arrangement of the cellulose fibers of the cell wall. Electron micrographs of plants frequently show microtubules lying just inside the plasma membrane of cells that are forming or extending their cell walls. Experimental alteration of the orientation of these microtubules leads to a similar change in the cell wall and a new shape for the cell.

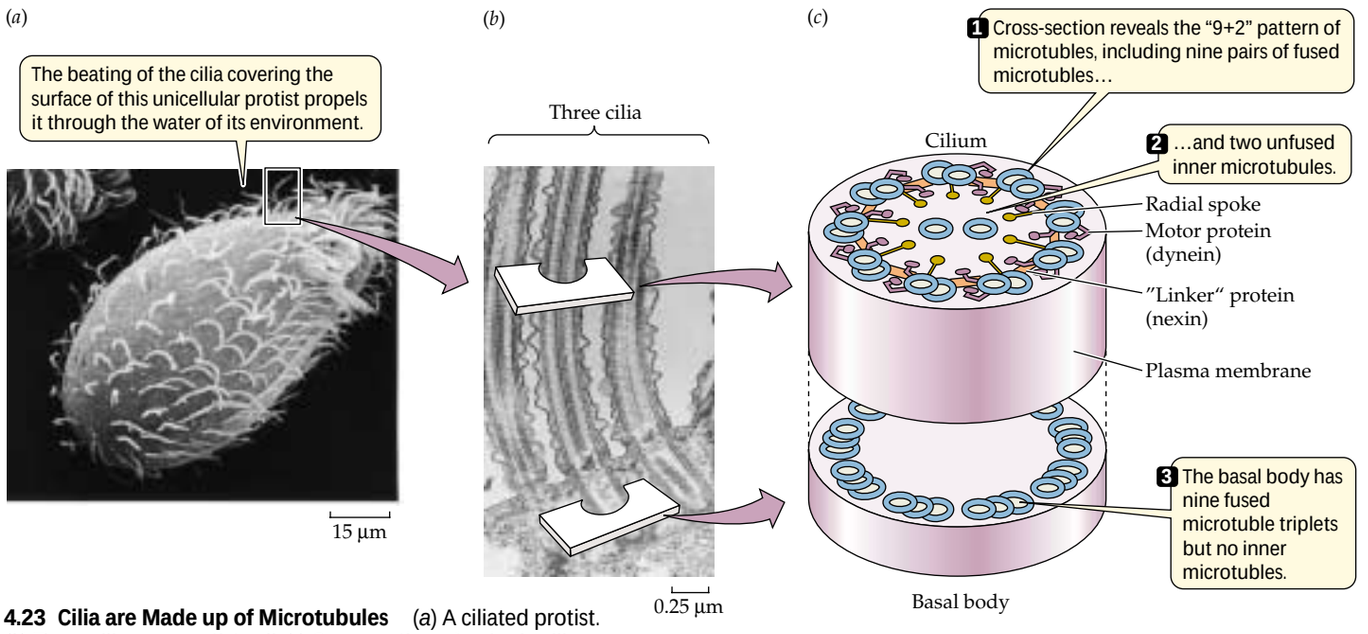
In many cells, microtubules serve as tracks for **motor proteins**, specialized molecules that use energy to change their shape and move. Motor proteins bond to and move along the microtubules, carrying materials from one part of the cell to another. Microtubules are also essential in distributing chromosomes to daughter cells during cell division. And they are intimately associated with movable cell appendages: the flagella and cilia.

Microtubules power cilia and flagella

Many eukaryotic cells possess flagella* or cilia, or both. These whiplike organelles may push or pull the cell through its aqueous environment, or they may move surrounding liquid over the surface of the cell (Figure 4.23*a*). Cilia and eukaryotic (but not prokaryotic) flagella are both assembled from specialized microtubules and have identical internal structures, but they differ in their relative lengths and their patterns of beating:

- **Flagella** are longer than cilia and are usually found singly or in pairs. Waves of bending propagate from one end of a flagellum to the other in snakelike undulation.
- **Cilia** are shorter than flagella and are usually present in great numbers. They beat stiffly in one direction and recover flexibly in the other direction (like a swimmer's arm), so that the recovery stroke does not undo the work of the power stroke.

*Some prokaryotes have flagella, as we saw earlier, but prokaryotic flagella lack microtubules and dynein. The flagella of prokaryotes are neither structurally nor evolutionarily related to those of eukaryotes. The prokaryotic flagellum is assembled from a protein called flagellin, and it has a much simpler structure and a smaller diameter than a single eukaryotic microtubule. And whereas eukaryotic flagella beat in a wavelike motion, prokaryotic flagella rotate (see Figure 4.6).



4.23 Cilia are Made up of Microtubules (*a*) A ciliated protist. (*b*) Three cilia on a protist cell. (*c*) Cross section of a single cilium.

In cross section, a typical cilium or eukaryotic flagellum is surrounded by the plasma membrane and contains a "9 + 2" array of microtubules. As Figure 4.23*c* shows, nine fused pairs of microtubules—called *doublets*—form an outer cylinder, and one pair of unfused microtubules runs up the center. A spoke radiates from one microtubule of each doublet and connects the doublet to the center of the structure.

In the cytoplasm at the base of every eukaryotic flagellum or cilium is an organelle called a **basal body**. The nine microtubule doublets extend into the basal body. In the basal body, each doublet is accompanied by another microtubule, making nine sets of three microtubules. The central, unfused microtubules do not extend into the basal body.

Centrioles are almost identical to the basal bodies of cilia and flagella. Centrioles are found in all eukaryotes except the flowering plants, pine trees and their relatives, and some protists. Under the light microscope, a centriole looks like a small, featureless particle, but the electron microscope reveals that it is made up of a precise bundle of microtubules arranged as nine sets of three fused microtubules each. Centrioles lie in the microtubule organizing center in cells that are about to divide. As you will see in Chapter 9, they are involved in the formation of the mitotic spindle, to which the chromosomes attach (see Figure 9.8).

Motor proteins move along microtubules

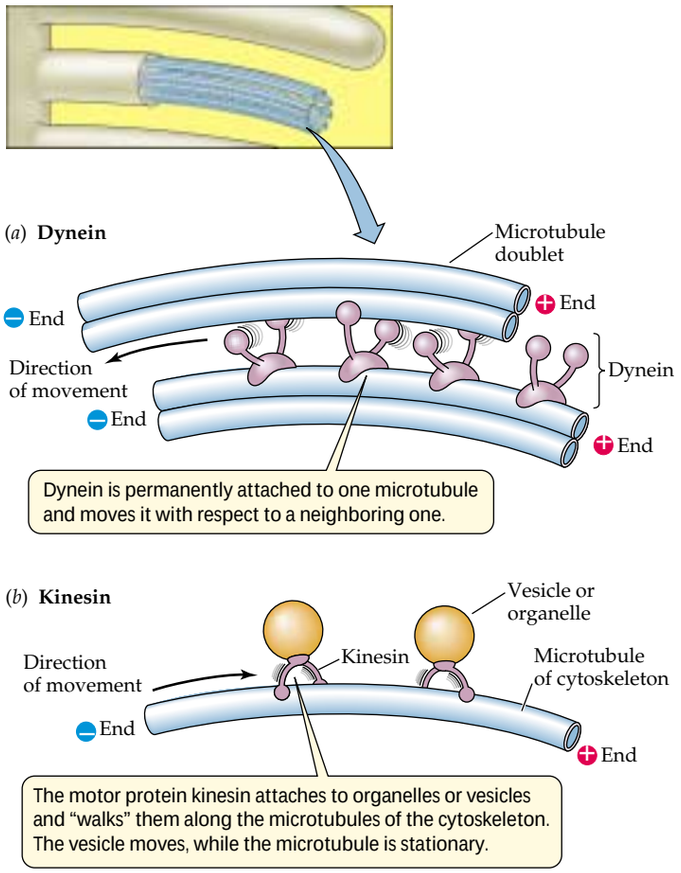
The nine microtubule doublets of cilia and flagella are linked by proteins. The motion of cilia and flagella results from the sliding of the microtubules past each other. This sliding is driven by a motor protein called **dynein**, which can undergo changes in its shape.

All motor proteins work by undergoing reversible shape changes powered by energy from ATP. Dynein molecules attached to one microtubule doublet bind to a neighboring doublet. As the dynein molecules change shape, they move the microtubule past its neighbor (Figure 4.24*a*).

Dynein and another motor protein, **kinesin**, are responsible for carrying protein-laden vesicles from one part of the cell to another. These motor proteins bind to a vesicle or other organelle, then "walk" it along a microtubule by changing their shape. Recall that microtubules have a plus end and a minus end. Dynein moves attached organelles toward the minus end, while kinesin moves them toward the plus end (Figure 4.24*b*).

Extracellular Structures

Although the plasma membrane is the functional barrier between the inside and the outside of a cell, many structures are produced by cells and secreted to the outside of the plasma membrane, where they play essential roles in protecting, supporting, or attaching cells. Because they are outside the plasma membrane, these structures are said to be *extracellular*. The peptidoglycan cell wall of bacteria is such an extracellular structure. In eukaryotes, other extracellular structures—the cell walls of plants and the extracellular matrices found between the cells of multicellular animals—play similar roles. Both of these structures are made up of a prominent fibrous macromolecule embedded in a jellylike medium.



4.24 Motor Proteins Use Energy from ATP to Move Things (*a*) Dynein is responsible for flagellar movement. (*b*) Kinesin, like dynein, delivers vesicles to various parts of the cell. (*c*) This SEM shows a vesicle attached to a microtubule by a motor protein.

The plant cell wall consists largely of cellulose

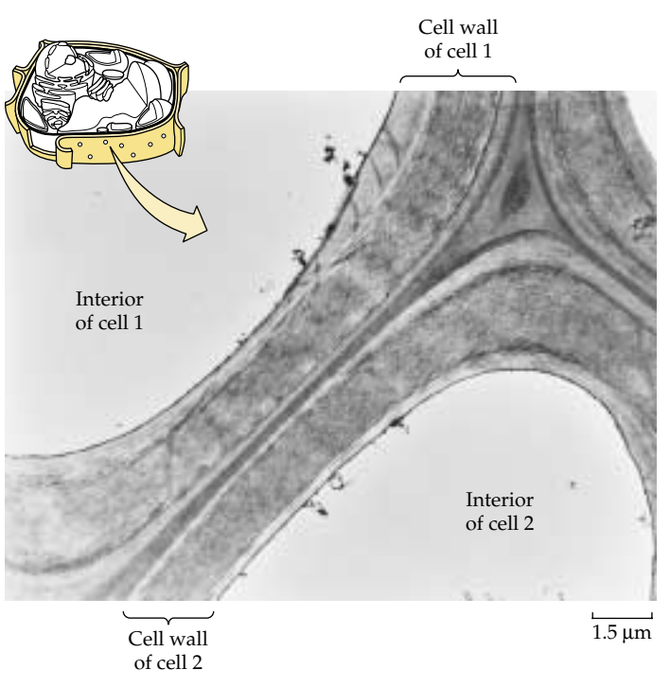
The cell wall of plant cells is a semirigid structure outside the plasma membrane (Figure 4.25). It consists of cellulose fibers embedded in other complex polysaccharides and proteins. The cell wall has three major roles in plants:

- It provides support for the cell and limits its volume by remaining rigid.
- It acts as a barrier to infections by fungi and other organisms that can cause plant diseases.
- It contributes to plant form by growing as plant cells expand.

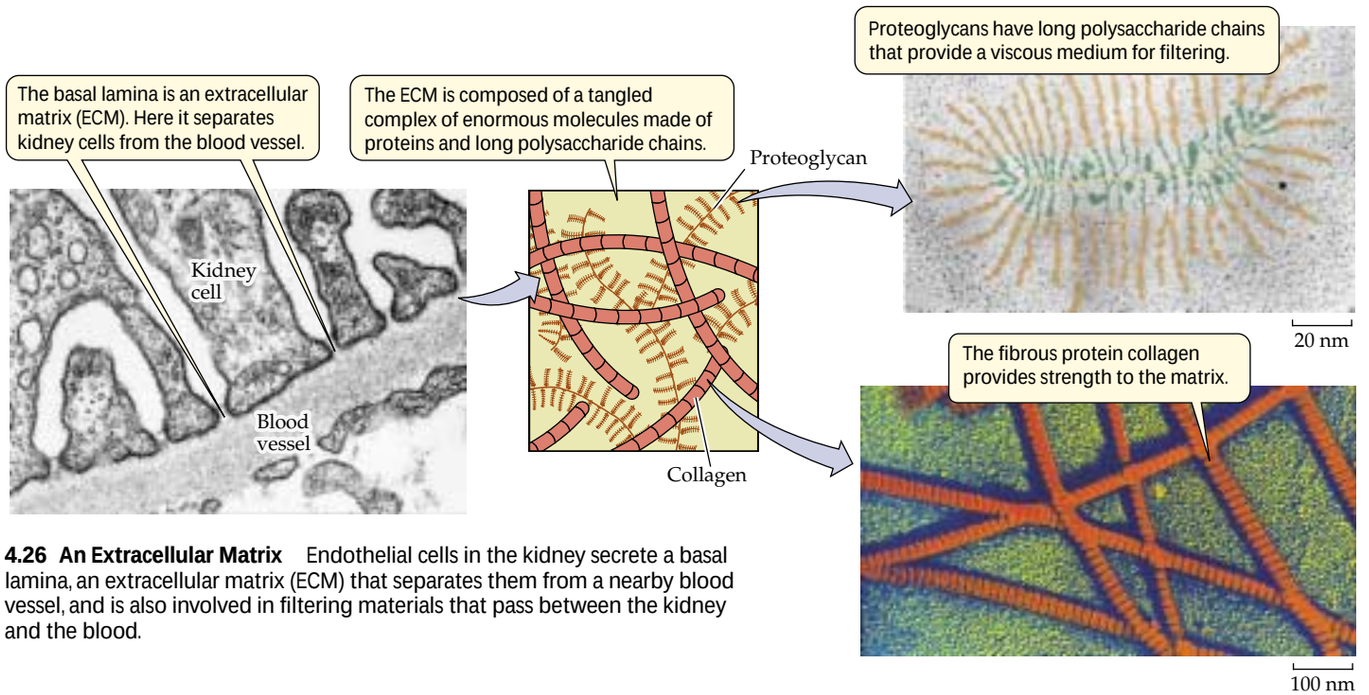
Because of their thick cell walls, plant cells viewed under a light microscope appear to be entirely isolated from each other. But electron microscopy reveals that this is not the case. The cytoplasm of adjacent plant cells is connected by numerous plasma membrane-lined channels, called *plasmodesmata*, that are about 20 to 40 nm in diameter and extend through the walls of adjoining cells (see Figure 15.19). Plasmodesmata permit the diffusion of water, ions, small molecules, and RNA and proteins between connected cells. Such diffusion ensures that the cells of a plant have uniform concentrations of these substances.

Animal cells have elaborate extracellular matrices

The cells of multicellular animals lack the semirigid cell wall that is characteristic of plant cells, but many animal cells are surrounded by, or are in contact with, an **extracellular matrix**. This matrix is composed of fibrous proteins such as *collagen* (the most abundant protein in mammals) and glyco-



4.25 The Plant Cell Wall The semirigid cell wall provides support for plant cells.



4.26 An Extracellular Matrix Endothelial cells in the kidney secrete a basal lamina, an extracellular matrix (ECM) that separates them from a nearby blood vessel, and is also involved in filtering materials that pass between the kidney and the blood.

proteins (Figure 4.26). These proteins, along with other substances that are specific to certain body tissues, are secreted by cells that are present in or near the matrix. The functions of the extracellular matrix are many:

- It holds cells together in tissues.
- It contributes to the physical properties of cartilage, skin, and other tissues.
- It helps filter materials passing between different tissues.
- It helps orient cell movements during embryonic development and during tissue repair.
- It plays a role in chemical signaling from one cell to another.

In the human body, some tissues, such as those in the brain, have very little extracellular matrix; other tissues, such as bone and cartilage, have large amounts. Bone cells are embedded in an extracellular matrix that consists primarily of collagen and calcium phosphate. This matrix gives bone its familiar rigidity. Epithelial cells, which line body cavities, lie together as a sheet spread over a *basal lamina*, or *basement membrane*, a form of extracellular matrix (see Figure 4.26).

Some extracellular matrices are made up, in part, of an enormous *proteoglycan*. A single molecule of this proteoglycan consists of many hundreds of polysaccharides covalently attached to about a hundred proteins, all of which are attached to one enormous polysaccharide. The molecular weight of this proteoglycan can exceed 100 million; the molecule takes up as much space as an entire prokaryotic cell.

Chapter Summary

The Cell: The Basic Unit of Life

- All cells come from preexisting cells and have certain processes, types of molecules, and structures in common.
- The first cells may have arisen from aggregates of macromolecules in bubbles. **Review Figure 4.1**
- To maintain adequate exchanges with its environment, a cell's surface area must be large compared with its volume. **Review Figures 4.2, 4.3.** **See Web/CD Activity 4.1**
- Cells can be visualized by various methods using microscopes. **Review Figure 4.4.** **See Web/CD Activity 4.2**
- All cells are surrounded by a plasma membrane.

Prokaryotic Cells

- All prokaryotic cells have a plasma membrane, a nucleoid region with DNA, and a cytoplasm that contains ribosomes, water, and dissolved proteins and small molecules. **Review Figure 4.5**
- Some prokaryotes have additional protective structures: cell wall, outer membrane, and capsule. Some prokaryotes contain photosynthetic membranes or mesosomes, and some have flagella or pili. **Review Figure 4.6**

Eukaryotic Cells

- Like prokaryotic cells, eukaryotic cells have a plasma membrane, cytoplasm, and ribosomes. However, eukaryotic cells are larger and contain many membrane-enclosed organelles. **Review Figure 4.7.** **See Web/CD Tutorial 4.1**
- The membranes that envelop organelles in the eukaryotic cell are partial barriers, ensuring that the chemical composition of the interior of the organelle differs from that of the surrounding cytoplasm.
- Organelles can be isolated by cell fractionation. **Review Figure 4.8**

Organelles that Process Information

▶ The nucleus is usually the largest organelle in a cell. It is surrounded by a double membrane, the nuclear envelope, which disassembles during cell division. Within the nucleus, the nucleolus is the source of the ribosomes found in the cytoplasm. Nuclear pores have a complex structure. **Review Figure 4.9**

▶ The nucleus contains most of the cell's DNA, which associates with protein to form chromatin. Chromatin is diffuse throughout the nucleus until just before cell division, when it condenses to form chromosomes. **Review Figure 4.10**

The Endomembrane System

▶ The endomembrane system is made up of a series of interrelated compartments enclosed by membranes.

▶ The rough endoplasmic reticulum has attached ribosomes that synthesize proteins. The smooth endoplasmic reticulum lacks ribosomes and is associated with the synthesis of lipids. **Review Figures 4.7, 4.11**

▶ The Golgi apparatus receives materials from the rough ER by means of vesicles that fuse with its *cis* region. Vesicles originating from the *trans* region of the Golgi contain proteins targeted to different cellular locations. Some of these vesicles fuse with the plasma membrane and release their contents outside the cell. **Review Figures 4.7, 4.12.** **See Web/CD Tutorial 4.2**

▶ Lysosomes contain many digestive enzymes. Lysosomes fuse with the phagosomes produced by phagocytosis to form secondary lysosomes, in which engulfed materials are digested. Undigested materials are secreted from the cell when the secondary lysosome fuses with the plasma membrane. **Review Figure 4.13.** **See Web/CD Activity 4.3**

Organelles that Process Energy

▶ Mitochondria are enclosed by an outer membrane and an inner membrane that folds inward to form cristae. Mitochondria contain the proteins needed for cellular respiration. **Review Figure 4.14**

▶ The cells of photosynthetic eukaryotes contain chloroplasts. These organelles are enclosed by double membranes and contain an internal system of thylakoids organized as grana. **Review Figures 4.7, 4.15**

▶ Thylakoids within chloroplasts contain the chlorophyll and proteins that harvest light energy for photosynthesis.

▶ Both mitochondria and chloroplasts contain their own DNA and ribosomes and are capable of making some of their own proteins.

▶ The endosymbiosis theory of the evolutionary origin of mitochondria and chloroplasts states that these organelles originated when larger prokaryotes engulfed, but did not digest, smaller prokaryotes. Mutual benefits permitted this symbiotic relationship to be maintained, allowing the smaller cells to evolve into the eukaryotic organelles observed today. **Review Figure 4.18**

Other Organelles

▶ Peroxisomes and glyoxysomes contain special enzymes and carry out specialized chemical reactions inside the cell. **Review Figure 4.19**

▶ Vacuoles are prominent in many plant cells and consist of a membrane-enclosed compartment full of water and dissolved substances. By taking in water, vacuoles enlarge and provide the pressure needed to stretch the cell wall and provide structural support for the plant. **Review Figure 4.20**

The Cytoskeleton

▶ The cytoskeleton within the cytoplasm of eukaryotic cells provides shape, strength, and movement. It consists of three interacting types of protein fibers. **Review Figure 4.21**

▶ Microfilaments consist of two chains of actin units that together form a double helix. Microfilaments strengthen cellular structures and provide the movement in animal cell division, cytoplasmic streaming, and pseudopod extension. Microfilaments may be found as individual fibers, bundles of fibers, or networks of fibers joined by linking proteins. **Review Figures 4.21, 4.22**

▶ Intermediate filaments are formed of keratins and are organized into tough, ropelike structures that hold organelles in place within the celland add strength to cell attachments in multicellular organisms. **Review Figure 4.21**

▶ Microtubules are composed of dimers of the protein tubulin. They can lengthen and shorten by adding and losing tubulin dimers. They are involved in the structure and function of cilia and flagella, both of which have a characteristic "9 + 2" pattern of microtubules. **Review Figures 4.21, 4.23**

▶ The movements of cilia and flagella result from the binding of the motor protein dynein to the microtubules. Dynein and another motor protein, kinesin, also bind to microtubules to move organelles through the cell. **Review Figure 4.24**

▶ Centrioles, made up of triplets of microtubules, are involved in the distribution of chromosomes during cell division.

Extracellular Structures

▶ Materials external to the plasma membrane provide protection, support, and attachment for cells in multicellular systems.

▶ The cell walls of plants consist principally of cellulose. They are pierced by plasmodesmata that join the cytoplasm of adjacent cells. **Review Figure 4.25**

▶ In multicellular animals, the extracellular matrix consists of different kinds of proteins, including proteoglycans. In bone and cartilage, the protein collagen predominates. **Review Figure 4.26**

Self-Quiz

1. Which is present in both prokaryotic cells and in eukaryotic plant cells?
   a. Chloroplasts
   b. Cell walls
   c. Nucleus
   d. Mitochondria
   e. Microtubules
2. The major factor limiting cell size is the
   a. concentration of water in the cytoplasm.
   b. need for energy.
   c. presence of membranous organelles.
   d. ratio of surface area to volume.
   e. composition of the plasma membrane.
3. Which statement about mitochondria is *not* true?
   a. Their inner membrane folds to form cristae.
   b. They are usually 1 μm or smaller in diameter.
   c. They are green because they contain chlorophyll.
   d. Energy-rich substances from the cytosol are oxidized in them.
   e. Much ATP is synthesized in them.
4. Which statement about plastids is *true*?
   a. They are found in prokaryotes.
   b. They are surrounded by a single membrane.
   c. They are the sites of cellular respiration.
   d. They are found in fungi.
   e. They are of several types, with different functions.

5. If all the lysosomes within a cell suddenly ruptured, what would be the most likely result?
   a. The macromolecules in the cytosol would begin to break down.
   b. More proteins would be made.
   c. The DNA within mitochondria would break down.
   d. The mitochondria and chloroplasts would divide.
   e. There would be no change in cell function.
6. The Golgi apparatus
   a. is found only in animals.
   b. is found in prokaryotes.
   c. is the appendage that moves a cell around in its environment.
   d. is a site of rapid ATP production.
   e. packages and modifies proteins.
7. Which organelle is *not* surrounded by one or more membranes?
   a. Ribosome
   b. Chloroplast
   c. Mitochondrion
   d. Peroxisome
   e. Vacuole
8. The cytoskeleton consists of
   a. cilia, flagella, and microfilaments.
   b. cilia, microtubules, and microfilaments.
   c. internal cell walls.
   d. microtubules, intermediate filaments, and microfilaments.
   e. calcified microtubules.
9. Microfilaments
   a. are composed of polysaccharides.
   b. are composed of actin.
   c. provide the motive force for cilia and flagella.
   d. make up the spindle that aids the movement of chromosomes.
   e. maintain the position of the chloroplast in the cell.
10. Which statement about the plant cell wall is *not* true?
    a. Its principal chemical components are polysaccharides.
    b. It lies outside the plasma membrane.
    c. It provides support for the cell.
    d. It completely isolates adjacent cells from one another.
    e. It is semirigid.

For Discussion

1. Which organelles and other structures are found in both plant and animal cells? Which are found in plant but not animal cells? In animal but not plant cells? Discuss these differences in relation to the activities of plants and animals.
2. Through how many membranes would a molecule have to pass in going from the interior of a chloroplast to the interior of a mitochondrion? From the interior of a lysosome to the outside of a cell? From one ribosome to another?
3. How does the possession of double membranes by chloroplasts and mitochondria relate to the endosymbiosis theory of the origins of these organelles? What other evidence supports the theory?
4. Compare the extracellular matrix of the animal cell with the plant cell wall with respect to composition of the fibrous and nonfibrous components, rigidity, and presence of cytoplasmic "bridges."
5. Plastids and mitochondria may have arisen via endosymbiosis. Propose a hypothesis for the origin of the cell nucleus.

4 Cellular Membranes

During his years as an undergraduate at Oxford, physics student Stephen Hawking took up rowing. Although he had never been much of an athlete, he was doing a passable job. But he noticed he was getting increasingly clumsy, and by the time he went to Cambridge as a graduate student, he was falling over for no particular reason. After weeks of tests, his physicians told him that he had motor neuron disease, an incurable condition in which the nerve cells that stimulate muscles gradually die and the patient loses all muscular control.

As the years progressed, Hawking made major contributions to theoretical physics, especially to the study of black holes and the origin of the universe. He ascended the ladder of academic success and now holds a professorship at Cambridge once held by Isaac Newton. But his disease has gotten worse, and he has lost almost all muscular control.

Stephen Hawking The effects of motor neuron disease have confined the famous physicist to a wheelchair. A major cellular manifestation of this disease is the lack of ability of the nerve cells to stimulate the opening of channels through muscle cell membranes that would result in normal muscle function.

A hallmark of living cells is their ability to regulate the substances that enter and leave them. This regulation is a function of the plasma membrane, which is composed of a hydrophobic lipid bilayer with associated proteins. Muscle cells respond to stimulation by nerve cells by opening protein-lined channels in their plasma membranes. Because his nerves cannot stimulate them, the channels of Hawking's muscle cells do not open, and his muscles do not contract. Channels in plasma membranes underlie the biological activities of many organisms, ranging from the beating of an animal's heart to the opening of tiny pores in leaves to let outside air in.

Membranes are dynamic structures whose components move and change. They perform their vital physiological roles by allowing cells to interact with other cells and with molecules in the environment. We describe the structural aspects of those interactions here. Membranes also regulate the traffic of chemicals into and out of the cell. The selective permeability of membranes, which we describe in this chapter, is an important characteristic of life. Later in this book, we will see it in action in such diverse situations as the transduction of light energy into chemical energy in the chloroplast and the retention of water and ions in the mammalian kidney.

Membrane Composition and Structure

The physical organization and functioning of all biological membranes depend on their constituents: lipids, proteins, and carbohydrates. The lipids establish the physical integrity of the membrane

and create an effective barrier to the rapid passage of hydrophilic materials such as water and ions. In addition, the phospholipid bilayer serves as a lipid "lake" in which a variety of proteins "float" (Figure 5.1). This general design is known as the **fluid mosaic model**.

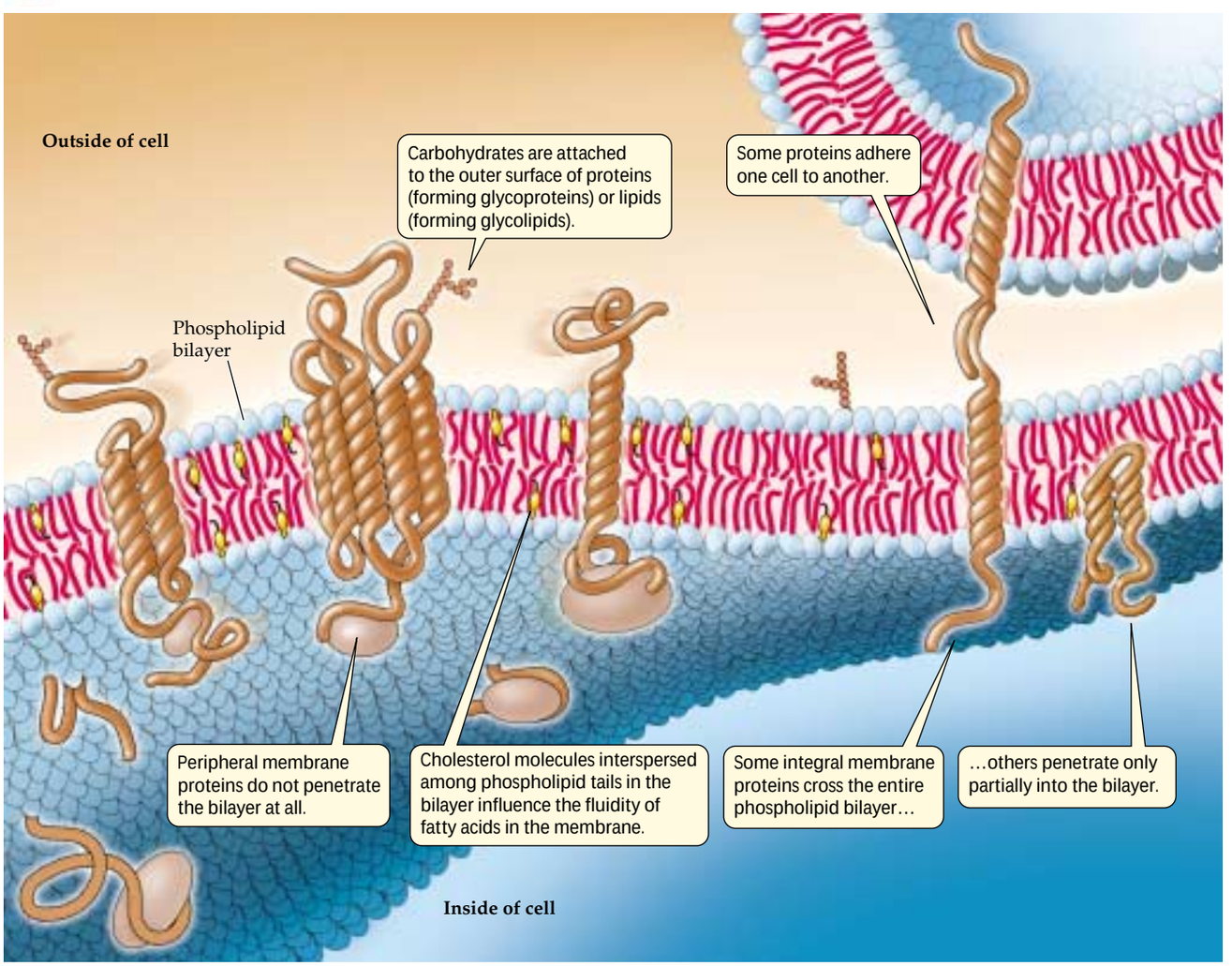
Proteins embedded in the phospholipid bilayer have a number of functions, including moving materials through the membrane and receiving chemical signals from the cell's external environment. Each membrane has a set of proteins suitable to the specialized function of the cell or organelle it surrounds.

The carbohydrates associated with membranes are attached either to the lipids or to protein molecules. They are located on the outside of the plasma membrane, where they protrude into the environment, away from the cell. Like some of the proteins, carbohydrates are crucial in recognizing specific molecules.

Lipids constitute the bulk of a membrane

Most of the lipids in biological membranes are phospholipids. Recall from Chapter 2 that some compounds are hydrophilic ("water-loving") and others are hydrophobic ("water-hating") and from Chapter 3 that phospholipids have both hydrophilic regions and hydrophobic regions:

- *Hydrophilic regions:* The phosphorus-containing "head" of the phospholipid is electrically charged and hence associates with polar water molecules.
- *Hydrophobic regions:* The long, nonpolar fatty acid "tails" of the phospholipid associate with other nonpolar materials, but they do not dissolve in water or associate with hydrophilic substances.



5.1 The Fluid Mosaic Model The general molecular structure of biological membranes is a continuous phospholipid bilayer in which proteins are embedded.

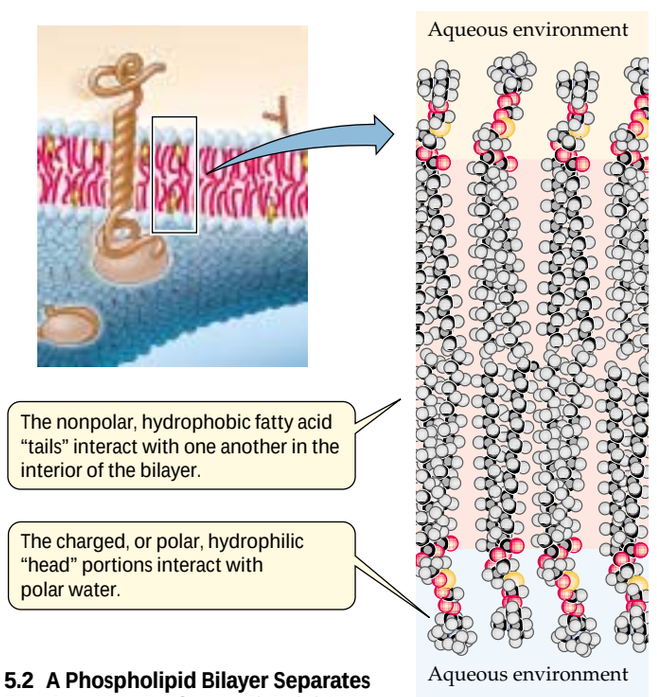
As a consequence of these properties, one way in which phospholipids can coexist with water is to form a *bilayer*, with the fatty acids of the two layers interacting with each other and the polar regions facing the outside aqueous environment (Figure 5.2).

In the laboratory, it is easy to make artificial bilayers with the same organization as natural membranes. In addition, small holes in a bilayer seal themselves spontaneously. This capacity of lipids to associate and maintain a bilayer organization helps biological membranes fuse during vesicle formation, phagocytosis, and related processes.

All biological membranes have a similar structure, but membranes from different cells or organelles may differ greatly in their lipid composition. Phospholipids differ in terms of fatty acid chain length, degree of unsaturation (double bonds) in the fatty acids, and the polar (phosphate-containing) groups present. In addition to phospholipids, membranes may contain cholesterol, a different type of lipid. In some membranes, 25 percent of the lipid is cholesterol (see Chapter 3), but other membranes have no cholesterol at all. When present, cholesterol is important to membrane integrity; most cholesterol in membranes is not hazardous to your health. A molecule of cholesterol is commonly situated next to an unsaturated fatty acid (see Figure 5.1).

The phospholipid bilayer stabilizes the entire membrane structure, but leaves it flexible, not rigid. At the same time, the fatty acids of the phospholipids make the hydrophobic interior of the membrane somewhat *fluid*—about as fluid as lightweight machine oil. This fluidity permits some molecules to move laterally within the plane of the membrane. A given phospholipid molecule in the plasma membrane may travel from one end of the cell to the other in a little more than a second. On the other hand, seldom does a phospholipid molecule in one half of the bilayer flip over to the other side and trade places with another phospholipid molecule. For such a swap to happen, the polar part of each molecule would have to move through the hydrophobic interior of the membrane. Since phospholipid flip-flops are rare, the inner and outer halves of the bilayer may be quite different in the kinds of phospholipids they contain.

The amount of cholesterol present in membranes, along with the degree of saturation of the fatty acids present, can increase or decrease membrane fluidity. Shorter fatty acid chains make for a more fluid membrane, as do unsaturated fatty acids. Adequate membrane fluidity is essential for many membrane functions. Since molecules move more slowly and fluidity decreases at reduced temperatures, membrane functions may decline in organisms that cannot keep their bodies warm. To address this problem, some organisms simply change the lipid composition of their membranes under cold conditions, replacing saturated with unsaturated fatty acids and using fatty acids with shorter tails. Such changes play a part in the survival of plants and hibernating animals and bacteria during the winter.



5.2 A Phospholipid Bilayer Separates Two Aqueous Regions The eight phospholipid molecules shown here represent a small cross section of a membrane bilayer.

Membrane proteins are asymmetrically distributed

All biological membranes contain proteins. Typically, plasma membranes have one protein molecule for every 25 phospholipid molecules. This ratio varies, however, depending on membrane function. In the inner membrane of the mitochondrion, which is specialized for energy processing, there is one protein for every 15 lipids. On the other hand, myelin, a membrane that encloses some nerve cells and uses the properties of lipids to act as an electrical insulator, has only one protein per 70 lipids.

Many membrane proteins are embedded in, or extend across, the lipid bilayer. Like phospholipids, these proteins have both hydrophilic and hydrophobic regions:

- *Hydrophilic regions:* Stretches of amino acids with hydrophilic R groups (side chains; see Table 3.2) give certain regions of the protein a polar character. Those regions, or *domains*, interact with water, sticking out into the aqueous extracellular environment or cytoplasm.
- *Hydrophobic regions:* Stretches of amino acids with hydrophobic R groups give other regions of the protein a nonpolar character. Those domains interact with the fatty acid chains in the interior of the lipid bilayer, away from water.

A special preparation method for electron microscopy, *freeze-fracturing*, reveals proteins embedded in the lipid bilayer of cellular membranes (Figure 5.3). The bumps that can be seen protruding from the interior of these membranes are not observed in pure lipid bilayers.

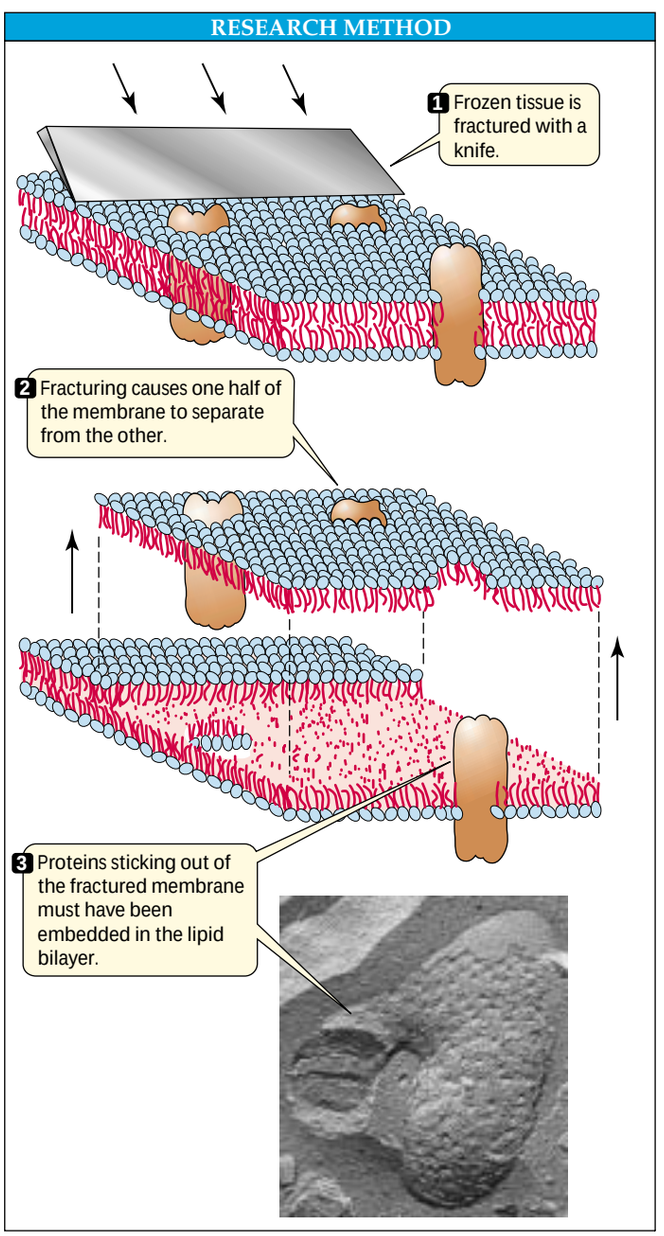
According to the fluid mosaic model, the proteins and lipids in a membrane are independent of each other and interact only noncovalently. The polar ends of proteins can interact with the polar ends of lipids, and the nonpolar regions of both molecules interact hydrophobically (see Figure 5.1).

There are two general types of membrane proteins:

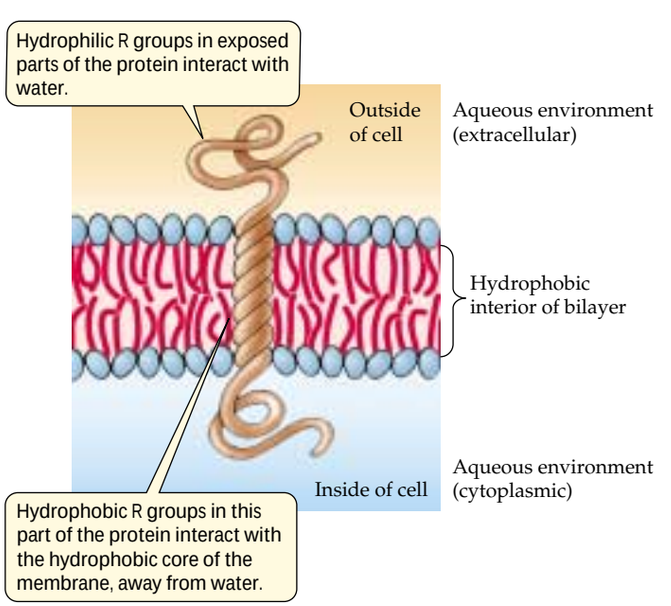
- **Integral membrane proteins** have hydrophobic regions and penetrate the phospholipid bilayer. Many of these proteins have long hydrophobic α-helical regions that span the hydrophobic core of the bilayer. Their hydrophilic ends protrude into the aqueous environments on either side of the membrane (Figure 5.4).
- **Peripheral membrane proteins** lack hydrophobic regions and are not embedded in the bilayer. Instead, they have polar or charged regions that interact with similar regions on exposed parts of integral membrane proteins or phospholipid molecules (see Figure 5.1).

Some membrane proteins are covalently attached to fatty acids or other lipid groups. These proteins can be classified as a special type of integral protein, as their hydrophobic lipid component allows them to insert themselves into the lipid bilayer.

Proteins are asymmetrically distributed on the inner and outer surfaces of a membrane. Integral membrane proteins that protrude on both sides of the membrane, known as **transmembrane proteins**, show different "faces" on the two membrane surfaces. Such proteins have certain specific domains on the outer side of the membrane, other domains within the membrane, and still other domains on the inner side of the membrane. Peripheral membrane proteins are localized on one side of the membrane or the other, but not



5.3 Membrane Proteins Revealed by the Freeze-Fracture Technique This membrane from a spinach chloroplast was first frozen and then separated so that the membrane bilayer was split open.



5.4 Interactions of Integral Membrane Proteins An integral membrane protein is held in the membrane by the distribution of the hydrophilic and hydrophobic R groups of its amino acids.

both. This arrangement gives the two surfaces of the membrane different properties. As we will soon see, these differences have great functional significance.

Like lipids, many membrane proteins move around relatively freely within the phospholipid bilayer. Experiments using the technique of *cell fusion* illustrate this migration dramatically. When two cells are fused, a single continuous membrane forms and surrounds both cells, and some proteins from each cell distribute themselves uniformly around this membrane.

Although many proteins are free to migrate in the membrane, others are not, but rather appear to be "anchored" to a specific region of the membrane. These membrane regions are like a corral of horses on a farm: The horses are free to move within the fenced area, but cannot get out of it. For instance, the protein in the membrane of a muscle cell that recognizes a molecular signal from nerve cells is normally found only at the site where a nerve cell meets the muscle cell. None of this protein is found elsewhere on the surface of the muscle cell. There are two ways in which the movement of proteins within a membrane can be restricted:

- The *cytoskeleton* may have components just below the inner face of the membrane that are attached to membrane proteins protruding into the cytoplasm.
- *Lipid rafts*, which are groups of lipids in a semisolid (not quite fluid) state, may trap proteins within a region. These lipids have a different composition from the surrounding phospholipids; for example, they may have very long fatty acid chains.

Membrane carbohydrates are recognition sites

In addition to lipids and proteins, many membranes contain significant amounts of carbohydrates. The carbohydrates are located on the outer surface of the membrane and serve as recognition sites for other cells and molecules (see Figure 5.1).

Membrane-associated carbohydrates may be covalently bound to lipids or to proteins:

- *Glycolipids* consist of a carbohydrate covalently bound to a lipid. The carbohydrate units often extend to the outside of the membrane, where they serve as recognition signals for interactions between cells. For example, the carbohydrate of some glycolipids changes when a cell becomes cancerous. This change may allow white blood cells to target cancer cells for destruction.
- *Glycoproteins* consist of a carbohydrate covalently bound to a protein. The bound carbohydrates are oligosaccharide chains, usually not exceeding 15 monosaccharide units in length. Glycoproteins enable a cell to be recognized by other cells and proteins.

An "alphabet" of monosaccharides on membranes can be used to generate a diversity of messages. Recall from Chapter 3 that sugar molecules can be formed from 3–7 carbons attached at different sites to one another, forming linear or branched oligosaccharides with many different three-dimensional shapes. An oligosaccharide of a specific shape from one cell can bind to a mirror-image shape on an adjacent cell. This binding forms the basis of cell-to-cell adhesion.

Cell Recognition and Adhesion

Some organisms, such as bacteria, are *unicellular*; that is, the entire organism is a single cell. Others, such as plants and animals, are *multicellular*—composed of many cells. Often these cells exist in specialized blocks of cells with similar functions, called *tissues*. Your body has about 60 trillion cells, arranged in different kinds of tissues such as muscle, nerve, skin, and so forth. Two processes allow cells to arrange themselves in groups:

- *Cell recognition*, in which one cell specifically binds to another cell of a certain type
- *Cell adhesion*, in which the relationship between the two cells is "cemented"

Both processes involve the plasma membrane. They are most easily studied if the cells in a tissue are separated into individual cells, then allowed to adhere to one another again. Simple organisms provide a good model for the complex tissues of larger species.

A living sponge is a multicellular marine animal with a simple body plan (see Chapter 32). The cells of the sponge are stuck together, but they can be disaggregated mechanically by passing the animal several times through a fine wire screen. What was an animal is now hundreds of individual of cells, suspended in seawater. Remarkably, if the cell suspension is shaken for a few hours, the cells bump into one another and stick together in the same shape as a sponge! The cells recognize and adhere to one another.

There are many different types (species) of sponges. If disaggregated cells from two different species of sponge are placed in the same container, the cells of the two species float around and bump into one another. But the cells of each species stick only to other cells of the same species. Two different sponges form, just like the ones at the start of the experiment.

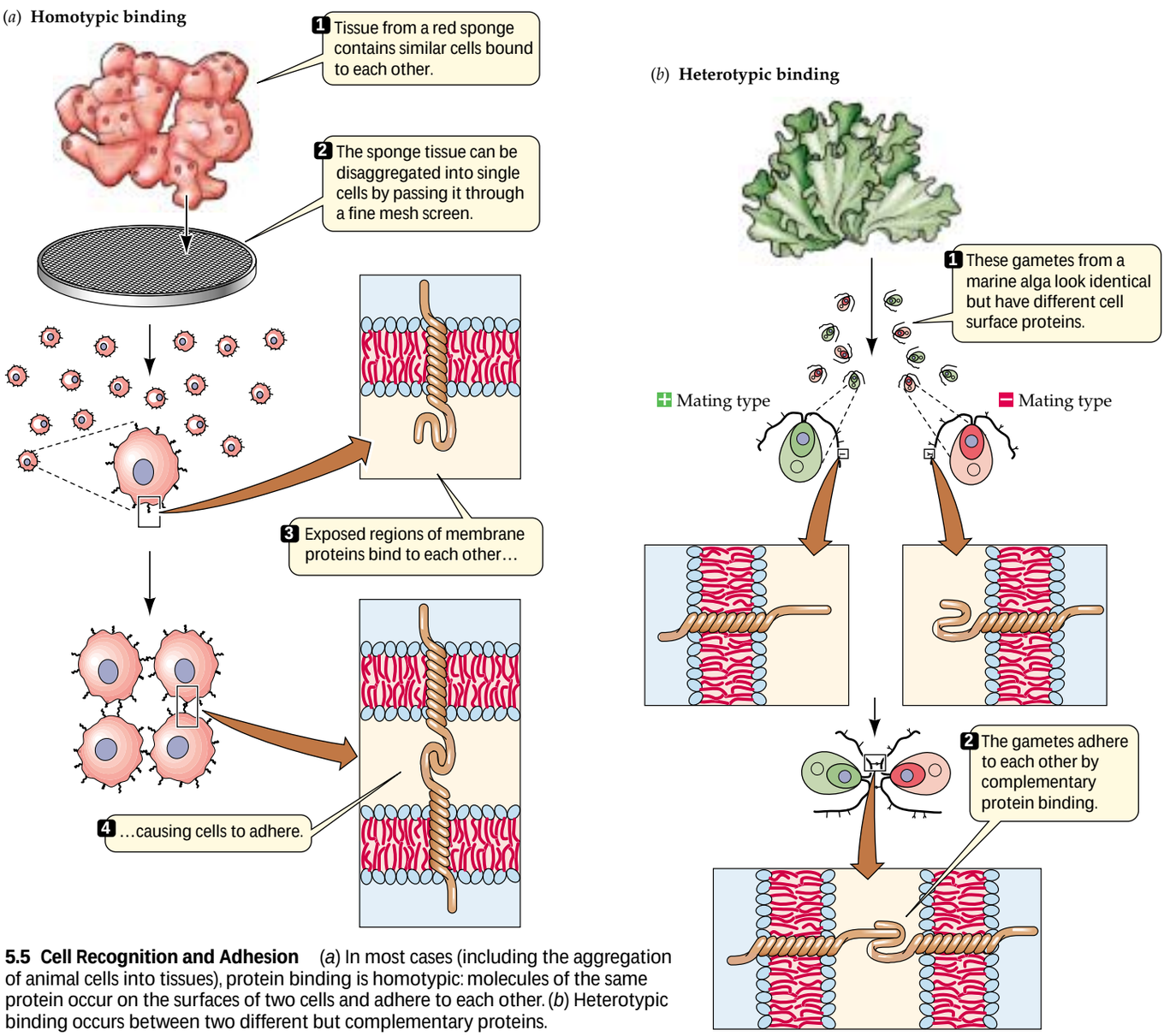
Such tissue-specific and species-specific cell adhesion is essential in the formation and maintenance of tissues and multicellular organisms. Think of your body. What keeps muscle cells bound to muscle cells and skin to skin? This is so obvious a characteristic of complex organisms that it is easy to overlook. You will see many examples of specific cell adhesion throughout this book; here, we describe its general

principles. As you will see, cell recognition and adhesion depend on membrane proteins.

Cell recognition and adhesion involve proteins at the cell surface

The molecule responsible for cell recognition and adhesion in sponges is a huge membrane glycoprotein (80% sugar) that is partly embedded in the plasma membrane, with the recognition part sticking out and exposed to the environment (and to other sponge cells). As we saw in Chapter 3, a macromolecule such as a protein not only has a specific shape, but also has specific chemical groups exposed on its surface where they can interact with other substances, including other proteins. Both of these features allow binding to other specific molecules (Figure 5.5*a*).

In most cases, the binding of cells in a tissue is **homotypic**; that is, the same molecule sticks out of both cells, and the exposed surfaces bind to each other. But **heterotypic** binding (of cells with different proteins) also can occur (Figure 5.5*b*). For example, when the mammalian sperm meets the egg, different proteins on the two types of cells have complementary binding surfaces. Similarly, some algae form similar-appearing male and female reproductive cells (analogous to sperm and eggs) that have flagella to propel them toward each other. Male and female cells can recognize each other by heterotypic proteins on their flagella. In the majority of plant cells, the plasma membrane is covered with a thick cell wall,



5.5 Cell Recognition and Adhesion (*a*) In most cases (including the aggregation of animal cells into tissues), protein binding is homotypic: molecules of the same protein occur on the surfaces of two cells and adhere to each other. (*b*) Heterotypic binding occurs between two different but complementary proteins.

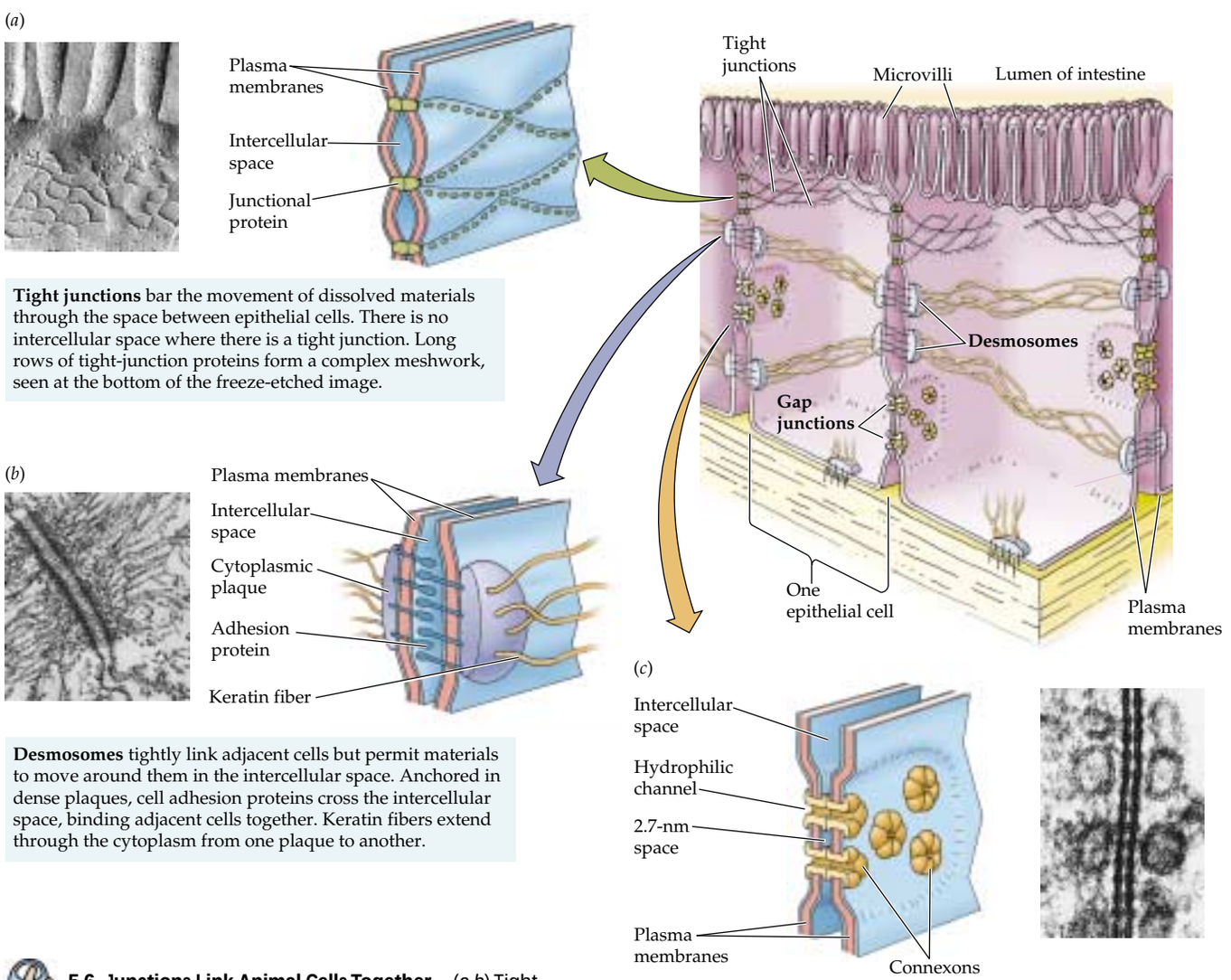
but this structure, too, has adhesion proteins that allow cells to bind to one another.

Cell adhesion proteins from many multicellular organisms have been characterized. Some of them do not just bind the two cells together, but initiate the formation of specialized cell junctions. In this case, the functions of cell recognition and cell adhesion reside in different molecules.

Specialized cell junctions

In a complex multicellular organism, cell recognition proteins allow specific kinds of cells to bind to each other. Often, both cells contribute material to additional membrane structures that "cement" their relationship. These specialized structures, called **cell junctions**, are most evident in electron micrographs of *epithelial tissues*, which are layers of cells that line body cavities or cover body surfaces. We will examine three types of cell junctions that enable cells to make direct physical contact and link with one another: tight junctions, desmosomes, and gap junctions.

TIGHT JUNCTIONS SEAL TISSUES AND PREVENT LEAKS. **Tight junctions** are specialized structures at the plasma membrane that link adjacent epithelial cells. They result from the mutual binding of strands of specific membrane proteins, which form a series of joints encircling each epithelial cell (Figure 5.6*a*). They are found in the region surrounding the



Tight junctions bar the movement of dissolved materials through the space between epithelial cells. There is no intercellular space where there is a tight junction. Long rows of tight-junction proteins form a complex meshwork, seen at the bottom of the freeze-etched image.

Desmosomes tightly link adjacent cells but permit materials to move around them in the intercellular space. Anchored in dense plaques, cell adhesion proteins cross the intercellular space, binding adjacent cells together. Keratin fibers extend through the cytoplasm from one plaque to another.

Gap junctions let adjacent cells communicate. Dissolved molecules and electric signals may pass from one cell to the other through the channel formed by two connexons extending from adjacent cells.

5.6 Junctions Link Animal Cells Together (*a*,*b*) Tight junctions and desmosomes are abundant in epithelial tissues. (*c*) Gap junctions are also found in some muscle and nerve tissues, in which rapid communication between cells is important.

lumen (cavity) of organs such as the intestine. Tight junctions have two functions:

- They prevent substances from moving through the spaces between cells. Thus, any substance entering the body from the lumen must pass through the epithelial cells.
- They restrict the migration of membrane proteins and phospholipids from one region of the cell to another.

Thus, the proteins and phospholipids in the *apical* (tip) region of the cell facing the lumen can be different from those in the *basolateral* regions facing the sides and bottom of the cell (basolateral: basal = bottom; lateral = side).

By forcing materials to enter certain cells, and by allowing different ends of cells to have different membrane proteins with different functions, tight junctions help ensure the directional movement of materials into the body.

DESMOSOMES HOLD CELLS TOGETHER. **Desmosomes** are also specialized structures associated with the plasma membrane. They hold adjacent cells firmly together, acting like spot welds or rivets (Figure 5.6*b*). Each desmosome has a dense structure called a plaque on the cytoplasmic surface of the plasma membrane. This plaque is attached to special cell adhesion proteins in the plasma membrane. These proteins stretch from the plaque through the plasma membrane of one cell, across the intercellular space, and through the plasma membrane of the adjacent cell, where they bind to the plaque proteins in that cell.

The plaque is also attached to fibers in the cytoplasm. These fibers, which are intermediate filaments of the cytoskeleton (see Figure 4.20), are made of a protein called *keratin*. They stretch from one cytoplasmic plaque across the cell to connect with another plaque on the other side of the cell. Anchored thus on both sides of the cell, these extremely strong keratin fibers provide great mechanical stability to epithelial tissues, which often receive rough wear in protecting the organism's body surface integrity.

GAP JUNCTIONS ARE A MEANS OF COMMUNICATION. Whereas tight junctions and desmosomes have mechanical roles, **gap junctions** facilitate communication between cells. Each gap junction is made up of specialized protein channels, called *connexons*, that span the plasma membranes of two adjacent cells and the intercellular space between them (Figure 5.6*c*). Dissolved molecules and electric signals can pass from cell to cell through these junctions. We will describe their role in more detail, as well as that of plasmodesmata, which perform a similar role in plants, when we discuss cell communication in later chapters, especially in Chapter 15.

Passive Processes of Membrane Transport

We have examined membrane structure and how it is used to perform one major membrane function: the binding of one cell to another. Now we turn to the second major membrane function: the ability to allow some substances, but not others, to pass through the membrane and enter or leave a cell or organelle. This characteristic of membranes is called **selective permeability**.

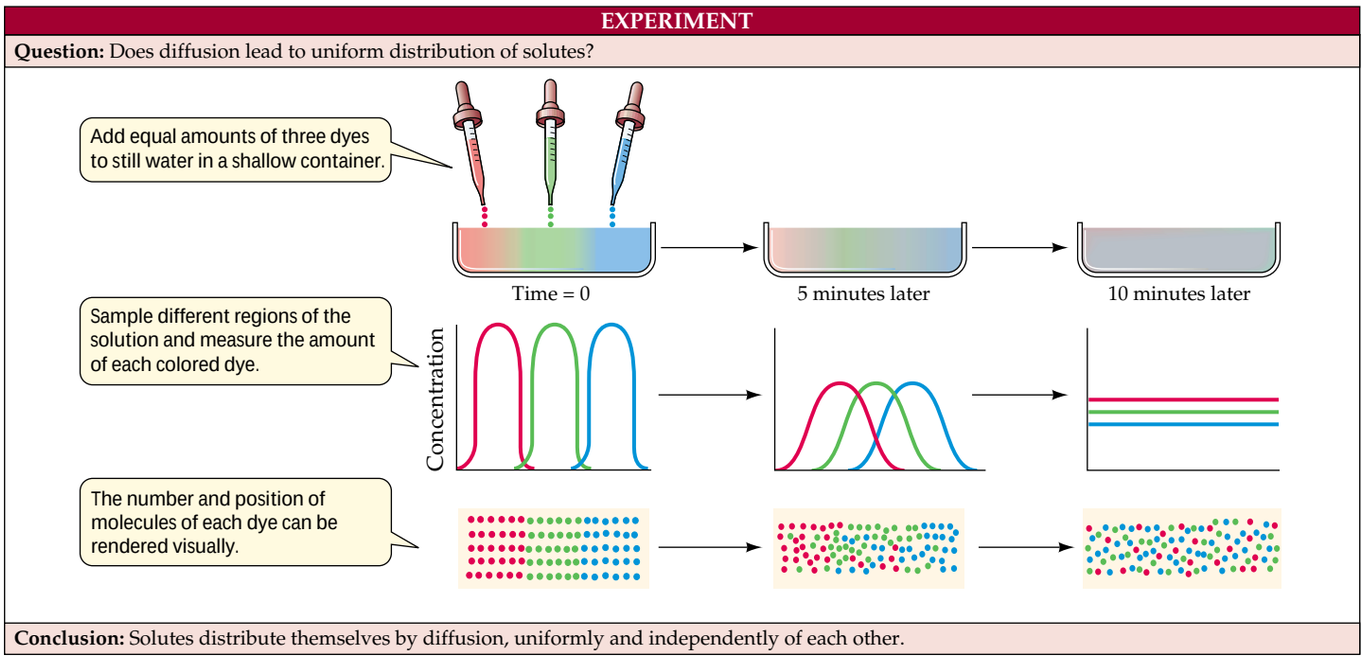
There are two fundamentally different kinds of processes by which substances cross biological membranes to enter and leave cells or organelles:

- **Passive transport** processes do not require any input of outside energy to drive them. The energy for these processes is in the substances themselves and the difference in their concentration on the two sides of the membrane. Passive transport processes include the different types of diffusion: simple diffusion through the phospholipid bilayer and facilitated diffusion through channel proteins or by means of carrier molecules.
- **Active transport** processes, on the other hand, require the input of chemical energy. They do not use the intrinsic property of a concentration gradient.

We'll discuss the active transport processes later in this chapter, after first focusing on the passive transport processes. Before considering diffusion as it works across a membrane, however, we must understand the basic principles of diffusion.

The physical nature of diffusion

Nothing in this world is ever absolutely at rest. Everything is in motion, although the motions may be very small. An important consequence of all this random jiggling is that all the components of a solution tend eventually to become evenly distributed throughout the system. For example, if a drop of ink is allowed to fall into a container of water, the pigment molecules of the ink are initially very concentrated. Without human intervention such as stirring, the pigment molecules of the ink move about at random, spreading slowly through the water until eventually the concentration of pigment—and thus the intensity of color—is exactly the same in every drop of liquid in the container. A solution in which the particles are uniformly distributed is said to be at *equilibrium*, because there will be no future net change in concentration. Equilibrium does not mean that the particles have stopped moving; it just means that they are moving in such a way that their overall distribution does not change.



5.7 Diffusion Leads to Uniform Distribution of Solutes Diffusion is the net movement of a solute from regions of greater concentration to regions of lesser concentration. The speed of diffusion varies with the substances involved, but the process continues until the solution reaches equilibrium.

Diffusion is the process of random movement toward a state of equilibrium. Although the motion of each individual particle is absolutely random, in diffusion the *net* movement of particles is directional until equilibrium is reached. Diffusion is thus net movement from regions of greater concentration to regions of lesser concentration (Figure 5.7).

In a complex solution (one with many different solutes), the diffusion of each substance is independent of that of the others. How fast a substance diffuses depends on four factors:

- the *diameter* of the molecules or ions: smaller molecules diffuse faster
- the *temperature* of the solution: higher temperatures lead to faster diffusion
- the *electric charge*, if any, of the diffusing material: electric charge has a variable effect on diffusion
- the *concentration gradient* in the system—that is, the change in concentration with distance in a given direction. The greater the concentration gradient, the more rapidly a substance diffuses.

DIFFUSION WITHIN CELLS AND TISSUES. Within cells, or wherever distances are very short, solutes distribute themselves rapidly by diffusion. Small molecules and ions may move from one end of an organelle to another in a millisecond (10^{-3} s). On the other hand, the usefulness of diffusion as a transport mechanism declines drastically as distances become greater. In the absence of mechanical stirring, diffusion across more than a centimeter may take an hour or more, and diffusion across meters may take years! Diffusion would not be adequate to distribute materials over the length of the human body, but within our cells or across layers of one or two cells, diffusion is rapid enough to distribute small molecules and ions almost instantaneously.

DIFFUSION ACROSS MEMBRANES. In a solution without barriers, all the solutes diffuse at rates determined by temperature, their physical properties, and the concentration gradient of each solute. If a biological membrane divides the solution into separate compartments, the movement of the different solutes can be affected by the properties of the membrane. The membrane is said to be *permeable* to solutes that can cross it more or less easily, but *impermeable* to substances that cannot move across it.

Molecules to which the membrane is impermeable remain in separate compartments, and their concentrations remain different on the two sides of the membrane. Molecules to which the membrane is permeable diffuse from one compartment to the other until their concentrations are equal on both sides of the membrane. When the concentrations of the diffusing substance are identical on both sides of the permeable membrane, equilibrium is reached. Individual molecules are still passing through the membrane after equilibrium is established, but equal numbers of molecules are moving in each direction, so there is *no net change* in concentration.

Simple diffusion takes place through the membrane bilayer

In **simple diffusion**, small molecules pass through the lipid bilayer of the membrane. The more lipid-soluble the molecule, the more rapidly it diffuses through the bilayer. This statement holds true over a wide range of molecular weights. Only water and the smallest of molecules seem to deviate from this rule, passing through bilayers much more rapidly than their lipid solubilities would predict.

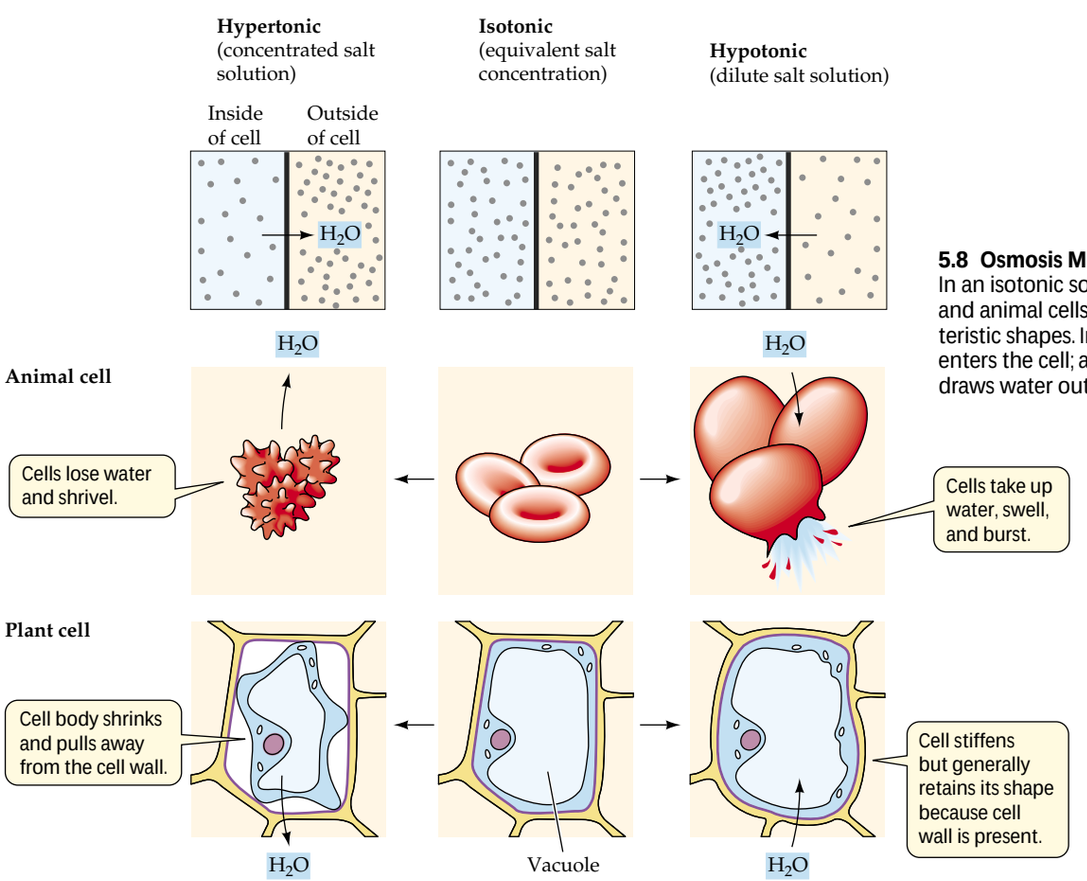
Charged or polar molecules such as amino acids, sugars, and ions do not pass readily through a membrane, for two reasons. First, cells are made up of, and exist in, water, and polar or charged substances form many hydrogen bonds with water, preventing their "escape" to the membrane. Second, the interior of the membrane is hydrophobic, and hydrophilic substances tend to be excluded from it. On the other hand, a molecule that is itself hydrophobic, and hence soluble in lipids, enters the membrane readily and is thus able to pass through it. For example, consider two types of molecules, a small protein of a few amino acids and a steroid of equivalent size. The protein, being polar, will diffuse slowly through the membrane, while the nonpolar steroid will diffuse through it readily.

Osmosis is the diffusion of water across membranes

Water molecules are abundant enough and small enough that they move through membranes by a diffusion process called **osmosis**. This completely passive process uses no metabolic energy and can be understood in terms of the concentrations of solutions. Osmosis depends on the *number* of solute particles present, not on the kinds of particles. We will describe osmosis using red blood cells and plant cells as examples.

Red blood cells are normally suspended in a fluid called plasma, which contains salts, proteins, and other solutes. If a drop of blood is examined under the light microscope, the red cells are seen to have a characteristic doughnut shape. If pure water is added to the drop of blood, the cells quickly swell and burst (Figure 5.8, top). Similarly, if slightly wilted lettuce is placed in pure water, it soon becomes crisp; by weighing it before and after, we can show that it has taken up water (Figure 5.8, bottom). If, on the other hand, red blood cells or crisp lettuce leaves are placed in a relatively concentrated solution of salt or sugar, the leaves become limp (wilt) and the red blood cells pucker and shrink.

From analyses of such observations, we know that the difference in solute concentration between a cell and its surrounding environment determines whether water will move from the environment into the cell or out of the cell into the



5.8 Osmosis Modifies the Shapes of Cells In an isotonic solution (center column), plant and animal cells maintain consistent, characteristic shapes. In a hypotonic solution, water enters the cell; a hypertonic environment draws water out of the cell.

environment. Other things being equal, if two different solutions are separated by a membrane that allows water, but not solutes, to pass through, water molecules will move across the membrane toward the solution with a higher solute concentration. In other words, *water will diffuse from a region of its higher concentration (lower concentration of solutes) to a region of its lower concentration (higher concentration of solutes).*

Three terms are used to compare the solute concentrations of two solutions separated by a membrane:

- **Isotonic** solutions have equal solute concentrations.
- A **hypertonic** solution has a higher solute concentration than the other solution with which it is being compared.
- A **hypotonic** solution has a lower solute concentration than the other solution with which it is being compared.

Water moves from a hypotonic solution across a membrane to a hypertonic solution (Figure 5.8). When we say that "water moves," bear in mind that we are referring to the net movement of water. Since it is so abundant, water is constantly moving across the plasma membrane into and out of cells. Whether the overall movement is greater in one direction or the other is what concerns us here.

The concentration of solutes in the environment determines the direction of osmosis in all animal cells. A red blood cell takes up water from a solution that is hypotonic to the cell's contents. The cell bursts because its plasma membrane cannot withstand the swelling of the cell (see Figure 5.8). The integrity of red blood cells (and other blood cells) is absolutely dependent on the maintenance of a constant solute concentration in the plasma in which they are suspended: The plasma must be isotonic to the cells if the cells are not to burst or shrink. Regulation of cell volume is an important process for cells without cell walls.

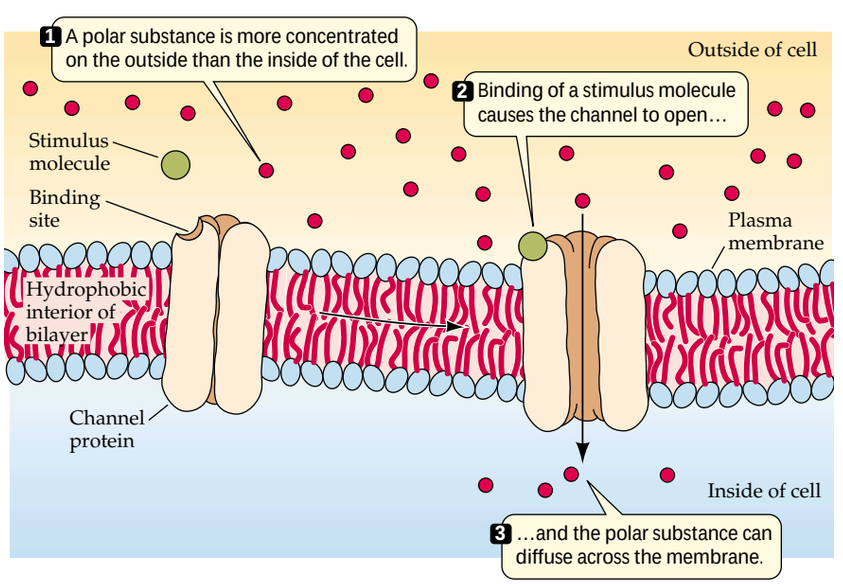
In contrast to animal cells, the cells of plants, archaea, bacteria, fungi, and some protists have cell walls, which limit the volume of the cells and keep them from bursting. Cells with sturdy cell walls take up a limited amount of water and, in so doing, build up internal pressure against the cell wall that prevents further water from entering. This pressure within the cell, called *turgor pressure*, keeps plants upright and is the driving force for the enlargement of plant cells—it is a normal and essential component of plant growth.

Diffusion may be aided by channel proteins

As we saw earlier, polar substances such as amino acids and sugars and charged substances such as ions do not diffuse across membranes. Instead, they cross the hydrophobic lipid barrier in two ways : One way is through integral membrane proteins that form channels through which the polar substance can pass, and the other is by binding to a membrane protein that speeds up its membrane crossing. Both of these processes are forms of **facilitated diffusion**.

Membrane **channel proteins** (Figure 5.9) have polar amino acids and water on the inside of the channel pore (to bind to the polar or charged substance and allow it to pass through) and nonpolar amino acids on the outside of the macromolecule (to allow the channel protein to insert itself into the lipid bilayer). The best-studied channel proteins are the *ion channels*. As you will see in later chapters, the movement of ions into and out of cells is important in many biological processes, ranging from the electrical activity of the nervous system to the opening of the pores in leaves that allow gas exchange with the environment. Hundreds of ion channels have been identified, each of them specific for a particular ion. All of them show the basic structure of a water-lined pore that allows a particular ion to move through it.

Just as the front gate on a fence can be open or closed, most ion channels are *gated:* they can be closed to ion passage or opened. A gated channel opens when something happens to change the shape of the protein. Depending on the channel, this stimulus can range from the binding of a chemical signal to an electrical charge caused by an imbalance of ions. Once the channel opens, millions of ions can rush through it per second. How fast this happens, and in which direction (into or out of the cell), depends on the concentration gradi-



5.9 A Gated Channel Protein Opens in Response to a Stimulus The membrane protein changes its three-dimensional shape when a stimulus molecule binds to it.

ent of the ion between the cytoplasm and the exterior environment of the cell. For example, if the concentration of potassium ion (K^+) is much higher outside of the cell than inside, K^+ will enter the cell through a potassium channel by diffusion; if the concentration is higher inside the cell, K^+ will diffuse out of the cell.

How does an ion channel exert its specificity for one ion and not another? It is not simply a matter of charge or size. For example, sodium ion (Na^+) is 0.095 nm in radius and K^+ is larger, at 0.130 nm; both have the same positive charge. Yet the potassium channel lets only K^+ pass through the membrane, and not the smaller Na^+. How this happens was recently discovered when Roderick MacKinnon determined the structure of the potassium channel from a bacterium (Figure 5.10). The explanation is elegant and provides a good review of cell chemistry.

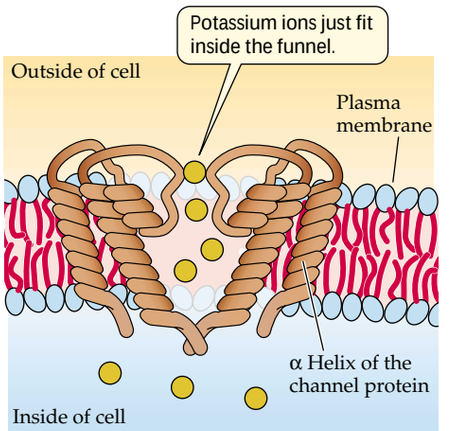
Being charged, both Na^+ and K^+ are attracted to water molecules. They have water "shells" in solution, held electrostatically by attraction of the positively charged ions to the negatively charged oxygen atom on the polar water molecules (see Figure 2.14). To get through a membrane channel, an ion must let go of its water. The "naked" ion is now attracted to the oxygen atoms on the channel pore protein.

In the potassium channel, oxygen atoms are located at the stem of a funnel-shaped protein region. The K^+ ion just fits the stem, and so can get into a position where it is more strongly attracted to the oxygen atoms there than to those of water. The smaller Na^+ ion, on the other hand, is kept a bit more distant from the oxygen atoms on the stem of the channel, and so prefers to be surrounded by water. So Na^+ does not enter the potassium channel.

As we mentioned, water crosses the plasma membrane at a rate far in excess of expectations, given its polarity. One way that water can do this is by hydrating ions as they pass through some ion channels. Up to 12 water molecules may coat an ion as it traverses a channel. Another way that water enters cells rapidly is through water channels, called **aquaporins**. Membrane proteins that allow water to pass through them have been characterized in many cells, from the plant vacuole, where they are important in maintaining turgor, to the mammalian kidney, where they act in retaining water that would otherwise be lost through urine.

Carrier proteins aid diffusion by binding substances

Another kind of facilitated diffusion involves not just the opening of a channel, but the actual binding of the transported substance to a membrane protein. These proteins are called **carrier proteins**, and like channel proteins, they allow diffusion both into and out of the cell. They are used to transport polar molecules such as sugars and amino acids.



5.10 The K^+ Channel This structure was first worked out for a channel from the bacterium *Streptomyces lividans*.

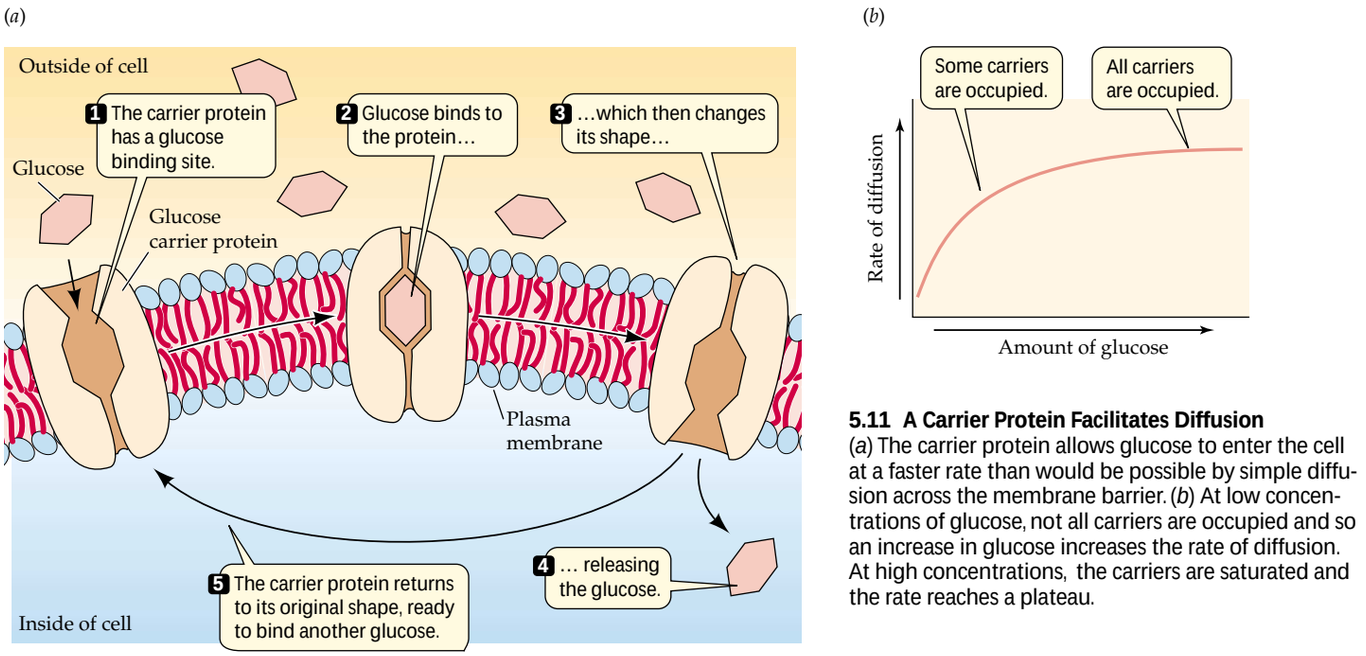
Glucose, for example, is the major energy source for most mammalian cells. The membranes of those cells contain a carrier protein called the *glucose transporter* that facilitates the uptake of glucose (Figure 5.11*a*). Since glucose is rapidly broken down as soon as it gets into a cell, there is almost always a strong concentration gradient favoring glucose entry, with a higher concentration outside the cell than inside.

Transport by carrier proteins is different from simple diffusion. In both processes, the rate of movement depends on the concentration gradient across the membrane. However, in facilitated diffusion, a point is reached at which further increases in the concentration gradient are not accompanied by an increased rate of diffusion. At this point, the facilitated diffusion system is said to be *saturated* (Figure 5.11*b*). Because there are only a limited number of carrier protein molecules per unit of membrane area, the rate of diffusion reaches a maximum when all the carrier molecules are fully loaded with solute molecules. In other words, when the difference in solute concentration across the membrane is sufficiently high, not enough carrier molecules are free at a given moment to handle all the solute molecules.

Passive transport allows substances to enter cells from the environment so that, after equilibrium is reached, the concentrations of a substance inside the cell and just outside the cell are equal. But a hallmark of living things is that they can have a composition quite different from that of their environment. One way that they achieve this is by not relying solely on concentration gradients, but instead by moving substances against their natural tendencies to diffuse. Because it requires an input of chemical energy, this process is called active transport.

Active Transport

In many biological situations, an ion or molecule must be moved across a membrane from a region of lower con-



5.11 A Carrier Protein Facilitates Diffusion (*a*) The carrier protein allows glucose to enter the cell at a faster rate than would be possible by simple diffusion across the membrane barrier. (*b*) At low concentrations of glucose, not all carriers are occupied and so an increase in glucose increases the rate of diffusion. At high concentrations, the carriers are saturated and the rate reaches a plateau.

centration to a region of higher concentration. In these cases, the substance cannot rush into or out of cells by diffusion. The movement of a substance across a biological membrane *against* a concentration gradient—active transport—requires the expenditure of chemical energy. The differences between diffusion and active transport are summarized in Table 5.1.

Active transport is directional

Three types of transporter proteins are involved in active transport (Figure 5.12):

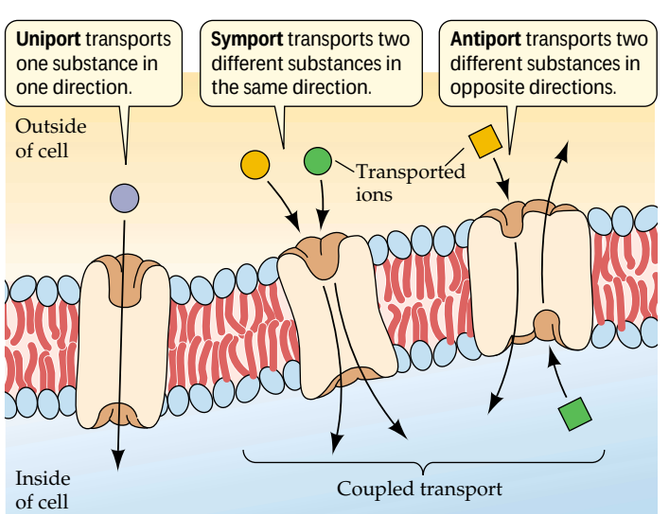
- **Uniports** move a single solute in one direction. For example, a calcium-binding protein found in the plasma membrane and endoplasmic reticulum of many cells actively transports Ca^{2+} to regions of higher concentration either outside the cell or inside the ER.
- **Symports** move two solutes in the same direction. For example, the uptake of amino acids from the intestine into the cells that line it requires the simultaneous binding of Na^+ and an amino acid to the same transporter protein.
- **Antiports** move two solutes in opposite directions, one into the cell and the other out of the cell. For example, many cells have a "sodium–potassium pump" that moves Na^+ out of the cell and K^+ into it.

Symports and antiports are known as *coupled transporters* because they move two solutes at once.

Primary and secondary active transport rely on different energy sources

There are two basic types of active transport processes:

- **Primary active transport** requires the direct participation of the energy-rich molecule ATP.
- **Secondary active transport** does not use ATP directly; rather, its energy is supplied by an ion concentration gradient established by primary active transport.



5.12 Three Types of Proteins for Active Transport Note that in each of the three cases, transport is directional.

5.1 Membrane Transport Mechanisms

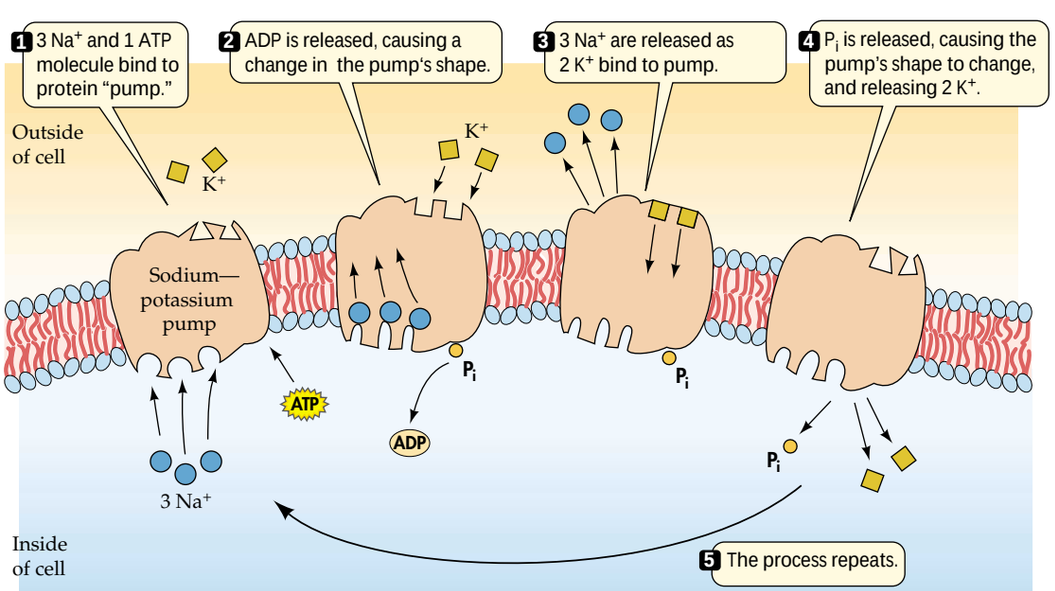
TRANSPORT MECHANISM	EXTERNAL ENERGY REQUIRED?	DRIVING FORCE	MEMBRANE PROTEIN REQUIRED?	SPECIFICITY
Simple diffusion	No	With concentration gradient	No	Not specific
Facilitated diffusion	No	With concentration gradient	Yes	Specific
Active transport	Yes	ATP hydrolysis (primary) (against concentration gradient)	Yes	Specific

In primary active transport, energy released by the hydrolysis of ATP drives the movement of specific ions against a concentration gradient. For example, if we compare the concentrations of potassium ions (K^+) and sodium ions (Na^+) inside a nerve cell and in the fluid bathing the nerve, the K^+ concentration is much higher inside the cell, whereas the Na^+ concentration is much higher outside. Nevertheless, a protein in the nerve cell continues to pump Na^+ out and K^+ in against these concentration gradients, ensuring that the gradients are maintained. This *sodium–potassium* (Na^+–K^+) *pump* is found in all animal cells and is an integral membrane glycoprotein. It breaks down a molecule of ATP to ADP and phosphate (P_i) and uses the energy released to bring two K^+ ions into the cell and export three Na^+ ions (Figure 5.13). The Na^+–K^+ pump is thus an antiport.

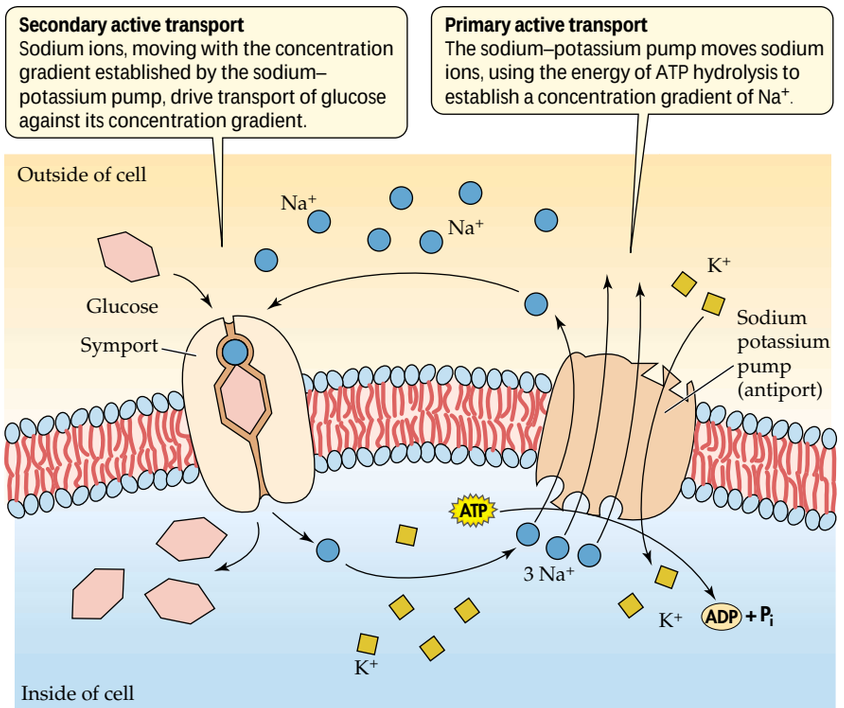
In secondary active transport, the movement of the solute against its concentration gradient is accomplished using energy "regained" by letting ions move across the membrane *with* their concentration gradient. For example, once the sodium–potassium pump establishes a concentration gradient of Na^+ ions, the passive diffusion of some Na^+ ions back into the cell can provide energy for the secondary active transport of glucose into the cell (Figure 5.14). Other secondary active transporters aid in the uptake of amino acids and other sugars, which are essential raw materials for cell maintenance and growth. Both types of coupled transport proteins—symports and antiports—are used for secondary active transport.

Endocytosis and Exocytosis

Macromolecules such as proteins, polysaccharides, and nucleic acids are simply too large and too charged or polar to pass through membranes. This is a fortunate property. Think of the consequences if these molecules could diffuse out of cells: A red blood cell would not retain its hemoglobin! On the other hand, cells must sometimes take up or secrete intact large molecules. As we saw in Chapter 4, this can be done by means of vesicles that either pinch off from the plasma membrane and enter the cell (endocytosis) or fuse with the plasma membrane and release their contents (exocytosis).



5.13 Primary Active Transport: The Sodium–Potassium Pump In active transport, energy is used to move a solute against its concentration gradient. Even though the Na^+ concentration is higher outside the cell and the K^+ concentration is higher inside the cell, for each molecule of ATP used, two K^+ are pumped into the cell and three Na^+ are pumped out of the cell.



5.14 Secondary Active Transport The Na^+ concentration gradient established by primary active transport (right) powers the secondary active transport of glucose (left). The movement of glucose across the membrane against its concentration gradient is coupled by a symport protein to the movement of Na^+ into the cell.

Macromolecules and particles enter the cell by endocytosis

Endocytosis is a general term for a group of processes that bring macromolecules, large particles, small molecules, and even small cells into the eukaryotic cell (Figure 5.15*a*). There are three types of endocytosis: phagocytosis, pinocytosis, and receptor-mediated endocytosis. In all three, the plasma membrane invaginates (folds inward) around materials from the environment, forming a small pocket. The pocket deepens, forming a vesicle. This vesicle separates from the plasma membrane and migrates with its contents to the cell's interior.

In **phagocytosis** ("cellular eating"), part of the plasma membrane engulfs large particles or even entire cells. Phagocytosis is used as a cellular feeding process by unicellular protists and by some white blood cells that defend the body by engulfing foreign cells and substances. The food vacuole or phagosome that forms usually fuses with a lysosome, where its contents are digested (see Figure 4.13).

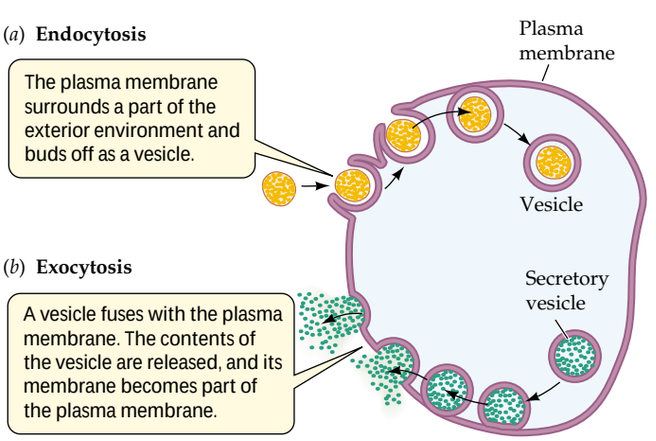
In **pinocytosis** ("cellular drinking"), vesicles also form. However, these vesicles are smaller, and the process operates to bring small dissolved substances or fluids into the cell. Like phagocytosis, pinocytosis is relatively nonspecific as to what it brings into the cell. For example, pinocytosis goes on constantly in the *endothelium*, the single layer of cells that separates a tiny blood capillary from its surrounding tissue, allowing the cells to rapidly acquire fluids from the blood.

In **receptor-mediated endocytosis**, specific reactions at the cell surface trigger the uptake of specific materials. Let's take a closer look at this process.

Receptor-mediated endocytosis is highly specific

Receptor-mediated endocytosis is used by animal cells to capture specific macromolecules from the cell's environment. This process depends on *receptor proteins*, integral membrane proteins that can bind to a specific molecule in the cell's environment. The uptake process is similar to nonspecific endocytosis, as already described. However, in receptor-mediated endocytosis, receptor proteins at particular sites on the extracellular surface of the plasma membrane bind to specific substances. These sites are called **coated pits** because they form a slight depression in the plasma membrane. The cytoplasmic surface of a coated pit is coated by proteins, such as *clathrin*.

When a receptor protein binds to its specific macromolecule outside the cell, its coated pit invaginates and forms a **coated vesicle** around the bound macromolecule. Strength-



5.15 Endocytosis and Exocytosis Endocytosis and exocytosis are used by all eukaryotic cells to take up substances from and release substances to the outside environment.

ened and stabilized by clathrin molecules, this vesicle carries the macromolecule into the cell (Figure 5.16). Once inside, the vesicle loses its clathrin coat and may fuse with a lysosome, where the engulfed material is processed and released into the cytoplasm. Because of its specificity for particular macromolecules, receptor-mediated endocytosis is a rapid and efficient method of taking up what may be minor constituents of the cell's environment.

Receptor-mediated endocytosis is the method by which cholesterol is taken up by most mammalian cells. Water-insoluble cholesterol and triglycerides are packaged by liver cells into lipoprotein particles, which are then secreted into the bloodstream to provide body tissues with lipids. One type of these particles, called *low-density lipoproteins*, or LDLs, must be taken up by the liver for recycling. This uptake also occurs via receptor-mediated endocytosis. This process begins with the binding of LDLs to specific receptor proteins on the cell surface. Once engulfed by endocytosis, the LDL particle is freed from the receptors. The receptors segregate to a region of the vesicle that buds off to form a new vesicle, which is recycled to the plasma membrane. The freed LDL particle remains in the original vesicle, which fuses with a lysosome in which the LDL is digested and the cholesterol made available for cell use. Persons with the inherited disease *familial hypercholesterolemia* (*-emia*, "blood") have dangerously high levels of cholesterol in their blood because of a deficient receptor for LDL.

Exocytosis moves materials out of the cell

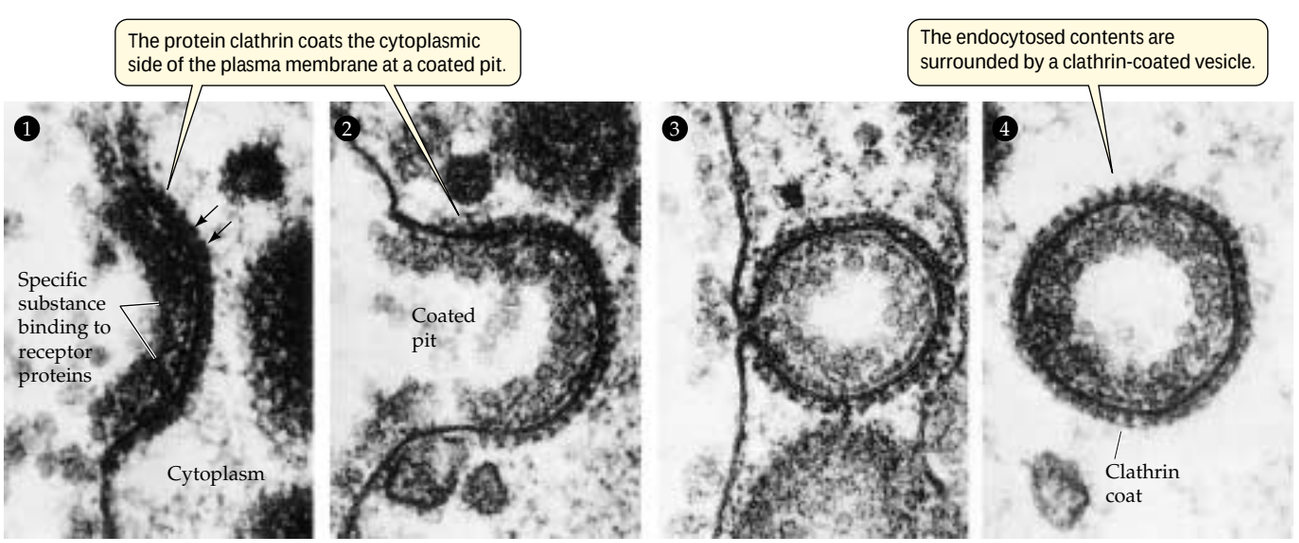
Exocytosis is the process by which materials packaged in vesicles are secreted from a cell when the vesicle membrane fuses with the plasma membrane (see Figure 5.15*b*). The initial event in this process is the binding of a membrane protein protruding from the cytoplasmic side of the vesicle with a membrane protein on the cytoplasmic side of the target site on the plasma membrane. The phospholipid regions of the two membranes merge, and an opening to the outside of the cell develops. The contents of the vesicle are released to the environment, and the vesicle membrane is smoothly incorporated into the plasma membrane.

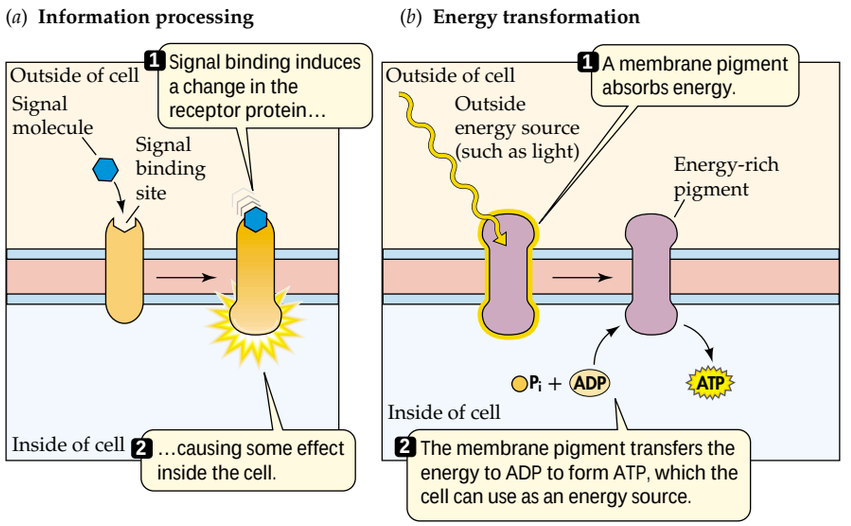
In Chapter 4, we encountered exocytosis as the last step in the processing of material engulfed by phagocytosis: the secretion of indigestible materials to the environment. Exocytosis is also important in the secretion of many different substances, including digestive enzymes from the pancreas, neurotransmitters from nerve cells, and materials for the construction of the plant cell wall.

Membranes Are Not Simply Barriers

We have discussed two major functions of membranes, cell adhesion and transport, but there are more. In Chapter 4, we described how the membrane of the rough endoplasmic reticulum serves as a site for ribosome attachment. Newly formed proteins are passed from the ribosomes through the membrane and into the interior of the ER for modification and delivery to other parts of the cell. This system sets up a separate compartment that segregates these proteins from the rest of the cell. On the other hand, the plasma membranes of nerve cells, muscle cells, and some eggs are electrically excitable. In nerve cells, the plasma membrane is the conductor of the nerve impulse from one end of the cell to the other.

5.16 Formation of a Coated Vesicle In receptor-mediated endocytosis, the receptor proteins in a coated pit bind specific macromolecules, which are then carried into the cell by a coated vesicle.





5.17 More Membrane Functions
(*a*) Membrane proteins conduct signals from outside the cell by triggering changes inside the cell. (*b*) The membranes of organelles such as mitochondria and chloroplasts are specialized for the transformation of energy. (*c*) When a series of biochemical reactions must take place in sequence, the membrane can sometimes arrange the needed enzymes in an "assembly line" to ensure that the reactions occur in proximity to each other.

Other biological activities and properties associated with membranes are discussed in the chapters that follow. These activities have been essential to the specialization of cells, tissues, and organisms throughout evolution. We review three of these activities here.

INFORMATION PROCESSING. As we have seen, biological membranes may have protruding integral membrane proteins or attached carbohydrates that can bind to specific substances in the environment. The binding of a specific substance can serve as a signal to initiate, modify, or turn off a cell function (Figure 5.17*a*).

In this type of information processing, specificity in binding is essential. We have already seen the role of a specific receptor protein in the endocytosis of LDL and its cargo of cholesterol. Another example is the binding of a hormone, such as insulin, to specific receptors on a target cell, such as a liver cell, to elicit a response in the cell—in this case, the uptake of glucose. There are many other examples, which we will discuss in Chapter 15.

ENERGY TRANSFORMATION. In a variety of cells, the membranes of organelles are specialized for processing energy (Figure 5.17*b*). For example, the inner mitochondrial membrane helps convert the energy of fuel molecules to the energy of ATP, and the thylakoid membranes of chloroplasts participate in the conversion of light energy to the energy of chemical bonds. These important processes, vital to the life of most eukaryotic organisms, are discussed in detail in Chapters 7 and 8.

ORGANIZING CHEMICAL REACTIONS. Many processes in cells depend on a series of enzyme-catalyzed reactions in which the products of one reaction serve as the reactants for the next. For such a reaction to occur, all the necessary molecules must come together. In a solution, the reactants and enzymes are all randomly distributed, and collisions among them are random. For this reason, a complete series of chemical reactions in solution may occur very slowly. However, if the different enzymes are bound to a membrane in sequential order, the product of one reaction can be released close to the enzyme for the next reaction. With such an "assembly line," reactions proceed more rapidly and efficiently (Figure 5.17*c*).

Membranes Are Dynamic

Membranes are constantly forming, transforming from one type to another, fusing with one another, and breaking down.

- *Phospholipids* in eukaryotes are synthesized on the surface of the smooth endoplasmic reticulum and rapidly dis-

tributed to membranes throughout the cell as vesicles form from the ER, move away, and fuse with other organelles.

- *Membrane proteins* are inserted into the rough endoplasmic reticulum as they form on ribosomes.
- *Functioning membranes* also move about within eukaryotic cells. Portions of the rough ER bud away as vesicles and join the *cis* face of the Golgi apparatus (see Figure 4.12). Rapidly—often in less than an hour—these segments of membrane find themselves in the *trans* regions of the Golgi, from which they bud away to join the plasma membrane.
- Membrane from *vesicles* is constantly merging with the plasma membrane by exocytosis, but this process is largely balanced by the removal of membrane in endocytosis, affording a recovery path by which internal membranes are replenished.

Because all membranes appear similar under the electron microscope, and because they interconvert readily, we might expect all subcellular membranes to be chemically identical. However, that is not the case, for there are major chemical differences among the membranes of even a single cell. Membranes are changed chemically when they form parts of certain organelles. In the Golgi apparatus, for example, the membranes of the *cis* face closely resemble those of the endoplasmic reticulum in chemical composition, but the *trans*-face membranes are more similar to the plasma membrane. As a vesicle is formed, the mix of proteins and lipids in its membrane is selected, just as its internal contents are selected, to correspond with the vesicle's target membrane.

In sum, there is a steady flux of membranes within the cell. Ceaselessly moving, functioning, changing their composition and roles, biological membranes are central to life.

Chapter Summary

Membrane Composition and Structure

▶ Biological membranes consist of lipids, proteins, and carbohydrates. The fluid mosaic model of membrane structure describes a phospholipid bilayer in which proteins can move about laterally within the membrane. **Review Figures 5.1, 5.2.** **See Web/CD Activity 5.1**

▶ Integral membrane proteins are at least partially inserted into the phospholipid bilayer. Peripheral membrane proteins are attached to the surface of the bilayer by ionic bonds. **Review Figure 5.1, 5.3, 5.4**

▶ The two surfaces of a membrane may have different properties because of their different phospholipid composition, exposed domains of integral membrane proteins, and peripheral membrane proteins. **Review Figures 5.1, 5.2**

▶ Carbohydrates attached to proteins or phospholipids project from the external surface of the plasma membrane and function as recognition signals for interactions between cells. **Review Figure 5.1**

Cell Recognition and Adhesion

▶ Some organisms consist of a single cell, but many are multicellular. The assembly of cells into tissues requires that they recognize and adhere to one another. Recognition and adhesion depend on membrane proteins that protrude from the cell surface. **Review Figure 5.5**

▶ Tight junctions prevent the passage of molecules through the spaces between cells, and they define functional regions of the plasma membrane by restricting the migration of membrane proteins uniformly over the cell surface. Desmosomes allow cells to adhere strongly to one another. Gap junctions provide channels for chemical and electrical communication between adjacent cells. **Review Figure 5.6.** **See Web/CD Activity 5.2**

Passive Processes of Membrane Transport

▶ Substances can diffuse passively across a membrane by three processes: unaided diffusion through the phospholipid bilayer, facilitated diffusion through protein channels, or facilitated diffusion by means of a carrier protein. **Review Table 5.1**

▶ A solute diffuses across a membrane from a region with a greater concentration of that solute to a region with a lesser concentration of that solute. Equilibrium is reached when the concentrations of the solute are identical on both sides of the membrane. **Review Figure 5.7**

▶ The rate of simple diffusion of a solute across a membrane is directly proportional to its concentration gradient across the membrane. An important factor in simple diffusion across a membrane is the lipid solubility of the solute.

▶ In osmosis, water diffuses from regions of higher water concentration to regions of lower water concentration.

▶ In hypotonic solutions, cells tend to take up water, whereas cells in hypertonic solutions tend to lose water. Animal cells must remain isotonic to the environment to prevent destructive loss or gain of water. **Review Figure 5.8*a*,*b***

▶ The cell walls of plants and some other organisms prevent the cells from bursting under hypotonic conditions. The turgor pressure that develops under these conditions keeps plants upright and stretches the cell wall during plant cell growth. **Review Figure 5.8*c***

▶ Channel proteins and carrier proteins function in facilitated diffusion. **Review Figures 5.9, 5.10, 5.11*a***

▶ The rate of carrier-mediated facilitated diffusion reaches a maximum when a solute concentration is reached that saturates the carrier proteins so that no increase in rate is observed with further increases in solute concentration. **Review Figure 5.11*b*.** **See Web/CD Tutorial 5.1**

Active Transport

▶ Active transport requires the use of chemical energy to move substances across a membrane against a concentration gradient. **Review Table 5.1**

▶ Active transport proteins may be uniports, symports, or antiports. **Review Figure 5.12**

▶ In primary active transport, energy from the hydrolysis of ATP is used to move ions into or out of cells against their concentration gradients. **Review Figure 5.13**

▶ Secondary active transport couples the passive movement of one solute with its concentration gradient to the movement of another solute against its concentration gradient. Energy from ATP is used indirectly to establish the concentration gradient that results in the movement of the first solute. **Review Figure 5.14.** **See Web/CD Tutorial 5.2**

Endocytosis and Exocytosis

▶ Endocytosis transports macromolecules, large particles, and small cells into eukaryotic cells by means of engulfment by and vesicle formation from the plasma membrane. Phagocytosis and pinocytosis are both nonspecific types of endocytosis. **Review Figure 5.15**

▶ In receptor-mediated endocytosis, a specific membrane receptor protein binds to a particular macromolecule. **Review Figure 5.16.** **See Web/CD Tutorial 5.3**

▶ In exocytosis, materials in vesicles are secreted from the cell when the vesicles fuse with the plasma membrane.

Membranes Are Not Simply Barriers

▶ Membranes function as sites for recognition and initial processing of extracellular signals, for energy transformations, and for organizing chemical reactions. **Review Figure 5.17**

Membranes Are Dynamic

▶ Modifications in membrane composition accompany the conversions of one type of membrane into another type.

Self-Quiz

1. Which statement about membrane phospholipids is *not* true?
   a. They associate to form bilayers.
   b. They have hydrophobic "tails."
   c. They have hydrophilic "heads."
   d. They give the membrane fluidity.
   e. They flop readily from one side of the membrane to the other.
2. Human growth hormone binds to a specific protein on the plasma membrane. This protein is called a
   a. ligand.
   b. clathrin.
   c. receptor.
   d. hydrophobic protein.
   e. cell adhesion molecule.
3. Which statement about membrane proteins is *not* true?
   a. They all extend from one side of the membrane to the other.
   b. Some serve as channels for ions to cross the membrane.
   c. Many are free to migrate laterally within the membrane.
   d. Their position in the membrane is determined by their tertiary structure.
   e. Some play roles in photosynthesis.
4. Which statement about membrane carbohydrates is *not* true?
   a. Most are bound to proteins.
   b. Some are bound to lipids.
   c. They are added to proteins in the Golgi apparatus.
   d. They show little diversity.
   e. They are important in recognition reactions at the cell surface.
5. Which statement about animal cell junctions is *not* true?
   a. Tight junctions are barriers to the passage of molecules between cells.
   b. Desmosomes allow cells to adhere strongly to one another.
   c. Gap junctions block communication between adjacent cells.
   d. Connexons are made of protein.
   e. The fibers associated with desmosomes are made of protein.
6. You are studying how the protein transferrin enters cells. When you examine cells that have taken up transferring, it is inside clathrin-coated vesicles. Therefore, the most likely mechanism for uptake of transferrin is
   a. facilitated diffusion.
   b. proton antiport.
   c. receptor-mediated endocytosis.
   d. gap junctions.
   e. ion channels.
7. Which statement about membrane channels is *not* true?
   a. They are pores in the membrane.
   b. They are proteins.
   c. All ions pass through the same type.
   d. Movement through them is from high concentration to low.
   e. Movement through them is by simple diffusion.
8. Facilitated diffusion and active transport both
   a. require ATP.
   b. require the use of proteins as carriers.
   c. carry solutes in only one direction.
   d. increase without limit as the solute concentration increases.
   e. depend on the solubility of the solute in lipid.
9. Primary and secondary active transport both
   a. generate ATP.
   b. are based on passive movement of sodium ions.
   c. include the passive movement of glucose molecules.
   d. use ATP directly.
   e. can move solutes against their concentration gradients.
10. Which statement about osmosis is *not* true?
    a. It obeys the laws of diffusion.
    b. In animal tissues, water moves into the cell which is hypertonic to the medium.
    c. Red blood cells must be kept in a plasma that is hypoosmotic to the cells.
    d. Two cells with identical osmotic potentials are isosmotic to each other.
    e. Solute concentration is the principal factor in osmosis.

For Discussion

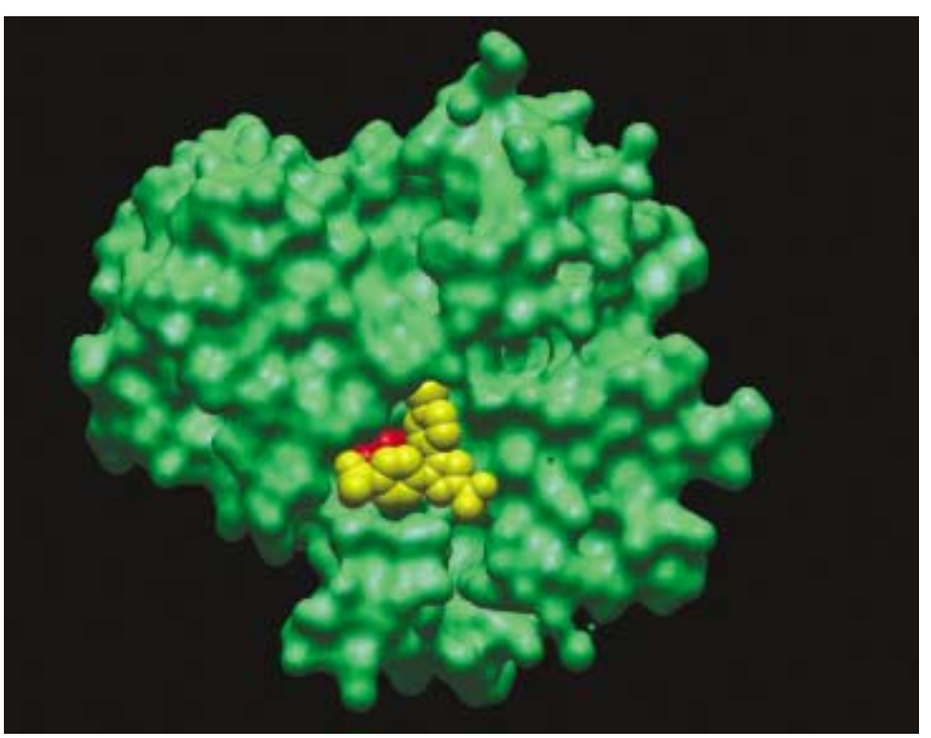
1. In Chapter 47, we will see that the functioning of muscles requires calcium ions to be pumped into a subcellular compartment against a calcium concentration gradient. What types of molecules are required for this to happen?
2. Some algae have complex glassy structures in their cell walls. These structures form within the Golgi apparatus. How do these structures reach the cell wall without having to pass through a membrane?
3. Organisms that live in fresh water are almost always hypertonic to their environment. In what way is this a serious problem? How do some organisms cope with this problem?
4. Contrast nonspecific endocytosis and receptor-mediated endocytosis with respect to mechanism and to performance.
5. The emergence of the phosopholipid membrane was important to the origin of cells. Describe the most important properties of membranes that allowed cells containing them to thrive in comparison with molecular aggregates without membranes.

5 Energy, Enzymes, and Metabolism

Millions of people, including famous athletes such as hockey star Wayne Gretzky, baseball great Hank Aaron, and Olympic decathlete Bruce Jenner, suffer from crippling arthritis. Until recently, physicians prescribed aspirin to calm the swelling that plagues arthritic joints. Aspirin has a long tradition in medicine. For thousands of years, healers in many cultures knew that the bark of a willow tree had anti-inflammatory properties that reduced swelling and pain. In 1829, German chemists isolated the active ingredient in willow bark, and later in that century others modified it chemically to make an even more effective drug—aspirin. While this drug was effective, it had several negative side effects, such as severe stomach irritation and reduced blood clotting.

It was only when biochemists discovered how aspirin works that its beneficial and undesirable effects could be explained. They discovered that aspirin binds to and adds an acetyl group to a particular amino acid (a serine) in a protein called cyclooxygenase, or COX. The normal role of COX in the body is to act as an enzyme: a catalyst to speed up the conversion of a linear fatty acid to a ring structure. The fatty acid enters a channel in the enzyme macromolecule, where it undergoes a specific chemical conversion, then departs from the enzyme. The ring form of the fatty acid stimulates inflammation in joints, repairs damage in the stomach wall, and helps blood clotting. When the serine residue in COX is acetylated by reactions with aspirin, the enzyme no longer speeds the production of the ring structure. Without COX, the formation of the ring structure still occurs, but at an exceedingly slow rate. So when COX is inhibited by aspirin, inflammation of joints is reduced, but stomach damage and a reduction in blood clotting also occur.

The search went on for a "better" anti-inflammatory drug: one that would block COX only in the joints. During the 1990s, biochemists hit paydirt. They found that there are actually *two* COX enzymes, one that acts in the stomach and blood cells (COX-1) and another that acts in the joints (COX-2). When they determined the primary structures of these two enzymes, they found that only one amino acid differs between them: COX-1 has a bulky



An Inflammatory Enzyme The COX-2 enzyme (green), represented here in a cutaway view, catalyzes the formation of a molecule that stimulates inflammation. Certain drugs (shown in yellow) block the substrate from binding to the active site. Such drugs inhibit the enzyme and depress inflammation. The valine residue that distinguishes COX-2 from the COX-1 enzyme is shown in red.

isoleucine where COX-2 has a smaller valine (see Table 3.2). The effect on tertiary structure was dramatic: The smaller amino acid in COX-2 exposed a side channel in the enzyme macromolecule that was blocked by the larger one in COX-1.

Using their knowledge of protein composition and structure, chemists designed molecules to plug up the COX-2 channel specifically, with no effect on COX-1. These new drugs (celecoxib and rofecoxib) relieve arthritis symptoms without the side effects in the stomach and blood. Their rapid development represents a case study in rational drug design: block a specific chemical transformation in cells by blocking the specific enzyme that catalyzes it.

Thousands of enzyme-catalyzed reactions go on all the time in every organism, each of them catalyzed by a specific protein with a particular three-dimensional structure. Taken together, these reactions make up **metabolism**, which is the total chemical activity of a living organism; at any instant, metabolism consists of thousands of individual chemical reactions. Many metabolic reactions can be classified as either the building up of complexity in the cell, using energy to do so, or the breaking down of complex substances into simpler ones, releasing energy in the process.

This chapter is concerned with energy and enzymes. Without them, neither we nor any other organism would be able to function. Before discovering how enzymes perform their molecular wizardry, we will consider the general principles of energy in biological systems.

Energy and Energy Conversions

Physicists define **energy** as the capacity to do work, which occurs when a force operates on an object over a distance. In biochemistry, it is more useful to consider energy as *the capacity for change*. No cell creates energy—all living things must obtain energy from the environment. Indeed, one of the fundamental physical laws is that energy can neither be created nor destroyed. However, energy can be *transformed* from one type into another, and living cells carry out many such energy transformations. Energy transformations are linked to the chemical transformations that occur in cells—the breaking of chemical bonds, the movement of substances across membranes, and so forth.

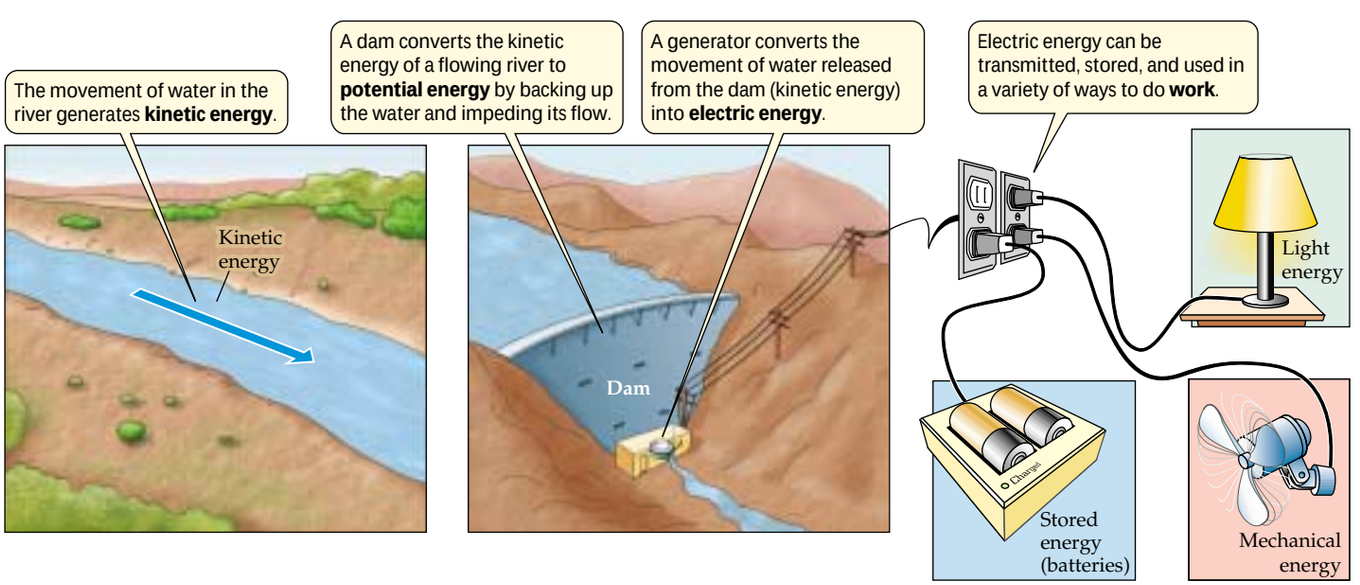
Energy changes are related to changes in matter

Energy comes in many forms, such as chemical energy, light energy, and mechanical energy. But all forms of energy can be considered as one of two basic types:

- **Kinetic energy** is the energy of movement. This type of energy does work that alters the state or motion of matter. It can exist in the form of heat, light, electric energy, and mechanical energy, among others.
- **Potential energy** is the energy of state or position—that is, stored energy. It can be stored in chemical bonds, as a concentration gradient, and as electric potential, among other ways.

Water stored behind a dam has potential energy. When the water is released from the dam, some of this potential energy is converted into kinetic energy, which can be harnessed to do work (Figure 6.1). Likewise, fatty acids store chemical energy in their C—H bonds and C—C bonds, and that energy can be released to do biochemical work.

6.1 Energy Conversions and Work The kinetic energy of a flowing river can be converted to potential energy by a dam. Release of water from the dam converts the potential energy back into kinetic energy, which a generator can convert into electric energy.



In all cells of all organisms, two types of metabolic reactions occur:

- **Anabolic reactions** (anabolism) link together simple molecules to form more complex molecules. The synthesis of a protein from amino acids is an anabolic reaction. Anabolic reactions require an input of energy and capture it in the chemical bonds that are formed.
- **Catabolic reactions** (catabolism) break down complex molecules into simpler ones and release the energy stored in chemical bonds.

Catabolic and anabolic reactions are often linked. The energy released in catabolic reactions is used to drive anabolic reactions—that is, to do biological work.

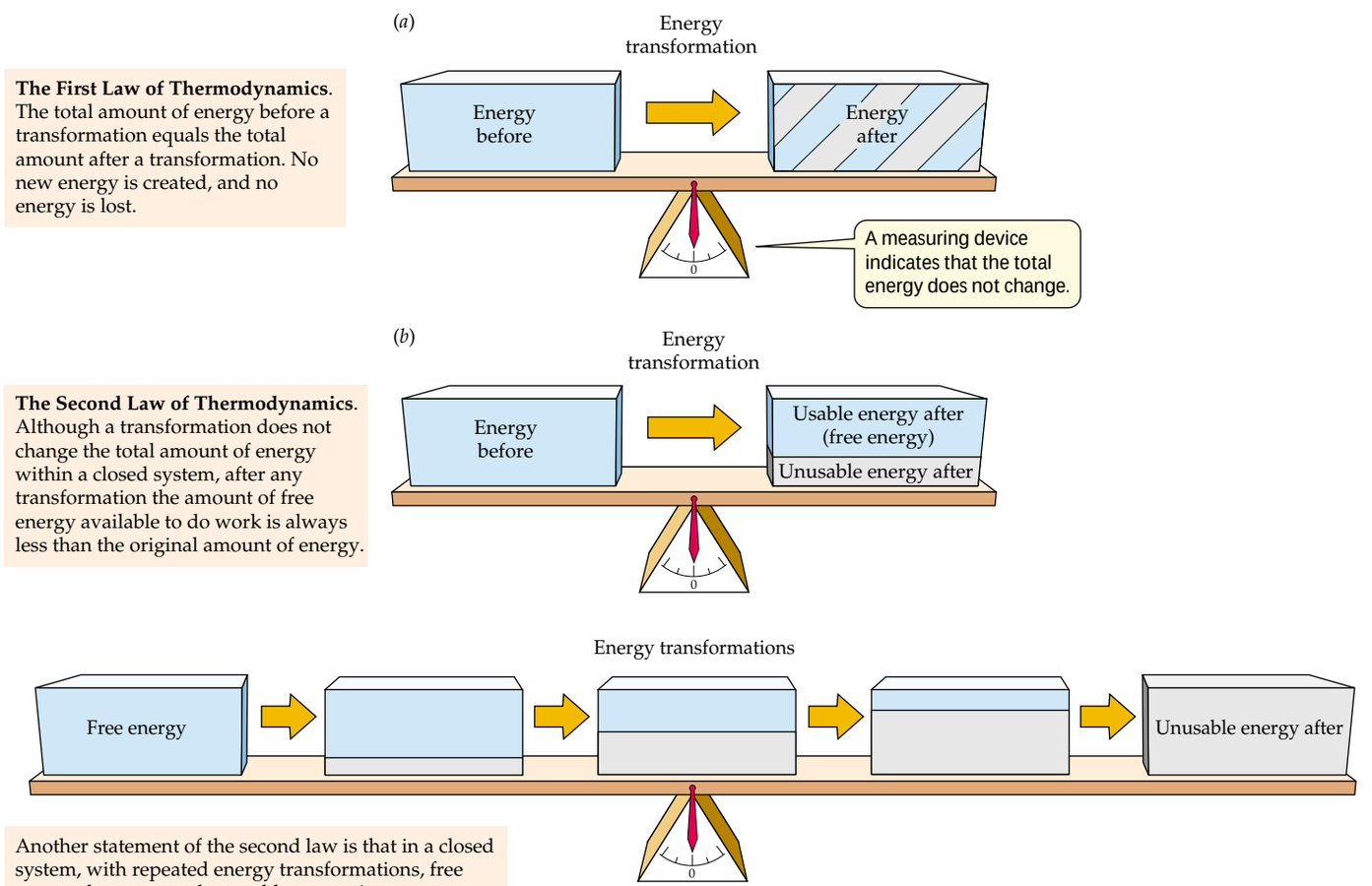
Cellular activities such as growth, movement, and active transport of ions across a membrane all require energy, and none of them would proceed without a source of energy. In the discussion that follows, we will discover the physical laws that govern all energy transformations, identify the energy available to do biological work, and consider the direction of energy flow.

The first law: Energy is neither created nor destroyed

Energy can be converted from one form to another. For example, by striking a match, you convert potential chemical energy to light and heat. The **first law of thermodynamics** states that in any such conversion of energy, energy is neither created nor destroyed.

The first law tells us that *in any conversion of energy from one form to another, the total energy before and after the conversion is the same* (Figure 6.2*a*). As you will see in the next two chapters, potential energy in the chemical bonds of carbohydrates and lipids can be converted to potential energy in the form of ATP. This energy can then be used to produce potential energy in the form of concentration gradients established by active transport, and can be converted to kinetic energy and used to do mechanical work, such as muscle contraction.

6.2 The Laws of Thermodynamics (*a*) The first law states that energy cannot be created or destroyed. (*b*) The second law states that during energy transformations, some of the free energy is lost.



The First Law of Thermodynamics. The total amount of energy before a transformation equals the total amount after a transformation. No new energy is created, and no energy is lost.

The Second Law of Thermodynamics. Although a transformation does not change the total amount of energy within a closed system, after any transformation the amount of free energy available to do work is always less than the original amount of energy.

Another statement of the second law is that in a closed system, with repeated energy transformations, free energy decreases and unusable energy increases—a phenomenon known as creation of **entropy**.

The second law: Not all energy can be used, and disorder tends to increase

The **second law of thermodynamics** states that, although energy cannot be created or destroyed, *when energy is converted from one form to another, some of that energy becomes unavailable to do work* (Figure 6.2*b*). In other words, no physical process or chemical reaction is 100 percent efficient, and not all the energy released can be converted to work. Some energy is lost to a form associated with disorder. The second law applies to all energy transformations, but we will focus here on chemical reactions in living systems.

NOT ALL ENERGY CAN BE USED. In any system, the total energy includes the usable energy that can do work *and* the unusable energy that is lost to disorder:

total energy = usable energy + unusable energy

In biological systems, the total energy is called **enthalpy** (*H*). The usable energy that can do work is called **free energy** (*G*). Free energy is what cells require for all the chemical reactions of cell growth, cell division, and the maintenance of cell health. The unusable energy is represented by **entropy** (*S*), which is a measure of the disorder of the system, multiplied by the absolute temperature (*T*). Thus we can rewrite the word equation above more precisely as

$$H = G + TS$$

Because we are interested in usable energy, we rearrange this expression:

$$G = H - TS$$

Although we cannot measure *G*, *H*, or *S* absolutely, we can determine the *change* in each at a constant temperature. Such energy changes are measured in calories (cal) or joules (J) (see Chapter 2). A change in energy is represented by the Greek letter delta (Δ). For example, the change in free energy (ΔG) of any chemical reaction is equal to the difference in free energy between the products and the reactants,

$$\Delta G_{\text{reaction}} = G_{\text{products}} - G_{\text{reactants}}$$

Such a change can be either positive or negative.

At a constant temperature, Δ*G* is defined in terms of the change in total energy (Δ*H*) and the change in entropy (Δ*S*):

$$\Delta G = \Delta H - T\Delta S$$

This equation tells us whether free energy is released or consumed by a chemical reaction:

- If Δ*G* is negative (Δ*G* < 0), free energy is released.
- If Δ*G* is positive (Δ*G* > 0), free energy is required (consumed).

IF THE NECESSARY FREE ENERGY IS NOT AVAILABLE, THE REACTION DOES NOT OCCUR. The sign and magnitude of Δ*G* depend on the two factors on the right of the equation:

- Δ*H:* In a chemical reaction, Δ*H* is the total amount of energy added to the system (Δ*H* > 0) or released (Δ*H* < 0).
- Δ*S:* Depending on the sign and magnitude of Δ*S*, the entire term, *T*Δ*S*, may be negative or positive, large or small. In other words, in living systems at a constant temperature (no change in *T*), the magnitude and sign of Δ*G* can depend a lot on changes in entropy. Large changes in entropy make Δ*G* more negative in value, as shown by the negative sign in front of the *T*Δ*S* term.

If a chemical reaction increases entropy, its products are more disordered or random than its reactants. If there are more products than reactants, as in the hydrolysis of a protein to its amino acids, the products have considerable freedom to move around. The disorder in a solution of amino acids will be large compared with that in the protein, in which peptide bonds and other forces prevent free movement. So in hydrolysis, the change in entropy (Δ*S*) will be positive.

If there are fewer products, and they are more restrained in their movements than the reactants, Δ*S* will be negative. For example, a large protein linked by peptide bonds is less free in its movements than a solution of the hundreds or thousands of amino acids from which it was synthesized.

DISORDER TENDS TO INCREASE. The second law of thermodynamics also predicts that, *as a result of energy conversions, disorder tends to increase*. Chemical changes, physical changes, and biological processes all tend to increase entropy and therefore tend toward disorder or randomness (Figure 6.2*b*). This tendency for disorder to increase gives a directionality to physical processes and chemical reactions. It explains why some reactions proceed in one direction rather than another.

How does the second law apply to organisms? Consider the human body, with its highly complex structures constructed of simpler molecules. This increase in complexity is in apparent disagreement with the second law. But this is not the case! Constructing 1 kg of a human body requires that about 10 kg of biological materials be metabolized and in the process converted to CO_2, H_2O, and other simple molecules, and these conversions require a lot of energy. This metabolism creates far more disorder than the order in 1 kg of flesh. *Life requires a constant input of energy to maintain order*. There is no disagreement with the second law of thermodynamics.

Having seen that the physical laws of energy apply to living things, we'll now turn to a consideration of how these laws apply to biochemical reactions.

Chemical reactions release or take up energy

In cells, anabolic reactions may make a single product, such as a protein (a highly ordered substance), out of many smaller reactants, such as amino acids (less ordered). Such reactions require or consume energy. Catabolic reactions may break down an ordered reactant, such as a glucose molecule, into smaller, more randomly distributed products, such as carbon dioxide and water. Such reactions give off energy. In other words, some reactions release free energy, and others take it up.

The amount of energy released ($-\Delta G$) or taken up ($+\Delta G$) by a reaction is related directly to the tendency of the reaction to run to *completion* (the point at which all the reactants are converted to products):

- Some reactions tend to run toward completion without any input of energy. These reactions, which release free energy, are said to be **exergonic** and have a negative ΔG (Figure 6.3*a*).
- Reactions that proceed toward completion only with the addition of free energy from the environment are **endergonic** and have a positive ΔG (Figure 6.3*b*).

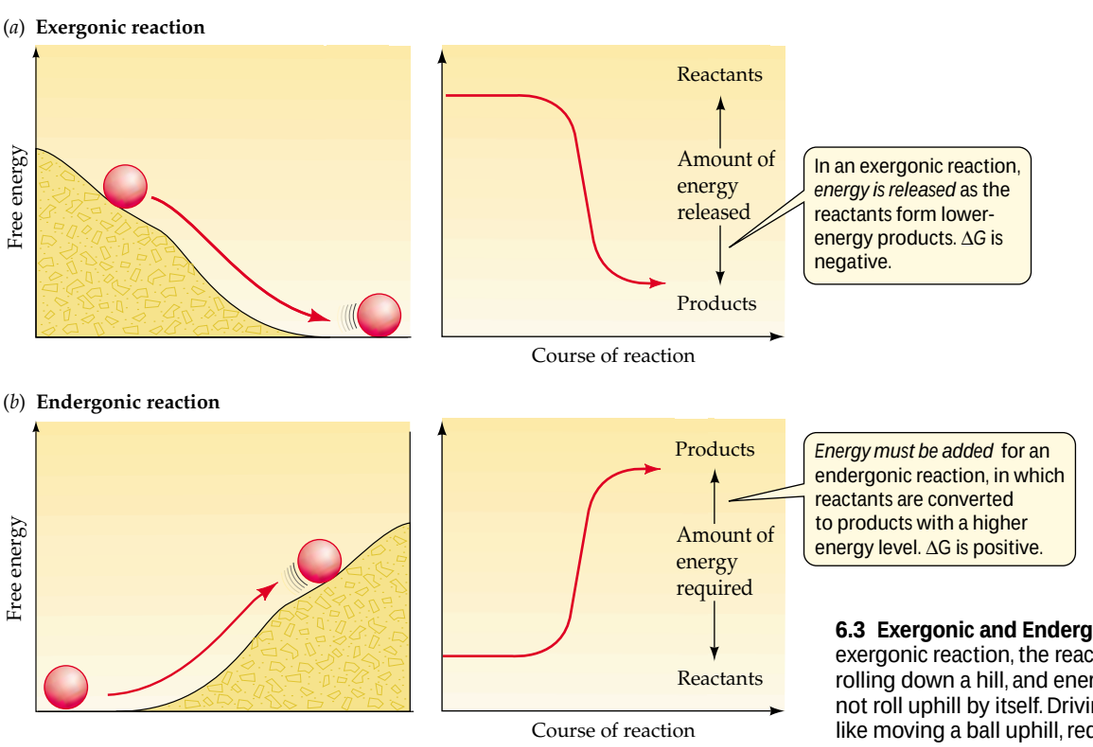
If a reaction runs exergonically in one direction (from reactant A to product B, for example), then the reverse reaction (B to A) requires a steady supply of energy to drive it. If $A \rightarrow B$ is exergonic ($\Delta G < 0$) , then $B \rightarrow A$ is endergonic ($\Delta G > 0$) .

In principle, chemical reactions can run both forward and backward. For example, if compound A can be converted into compound B ($A \rightarrow B$), then B, in principle, can be converted into A ($B \rightarrow A$), although at given concentrations of A and B, only one of these directions will be favored. Think of the overall reaction as resulting from competition between forward and reverse reactions ($A \rightleftharpoons B$). Increasing the concentration of the reactants (A) speeds up the forward reaction, and increasing the concentration of the products (B) favors the reverse reaction. At some concentration of A and B, the forward and reverse reactions take place at the same rate. At this concentration, no further net change in the system is observable, although individual molecules are still forming and breaking apart. This balance between forward and reverse reactions is known as **chemical equilibrium**. Chemical equilibrium is a static state, a state of no net change, and a state in which $\Delta G = 0$.

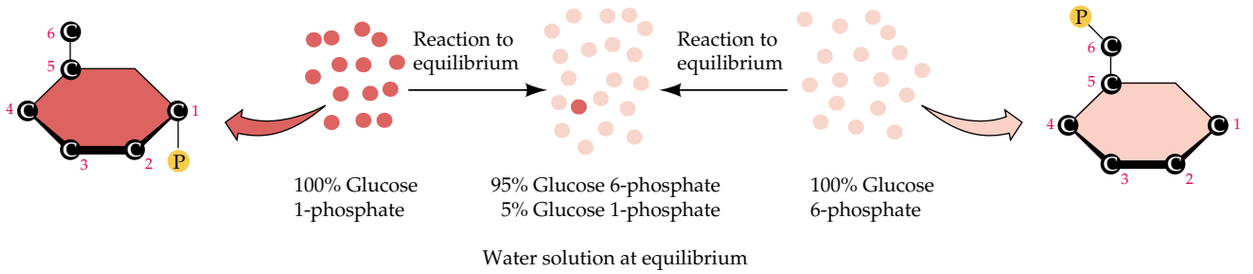
Chemical equilibrium and free energy are related

Every chemical reaction proceeds to a certain extent, but not necessarily to completion. In other words, all the reactants present are not necessarily converted to products. Each reaction has a specific equilibrium point, and that equilibrium point is related to the free energy released by the reaction under specified conditions. To understand the principle of equilibrium, consider the following example.

Most cells contain glucose 1-phosphate, which is converted in the cell to glucose 6-phosphate. Imagine that we



6.3 Exergonic and Endergonic Reactions (*a*) In an exergonic reaction, the reactants behave like a ball rolling down a hill, and energy is released. (*b*) A ball will not roll uphill by itself. Driving an endergonic reaction, like moving a ball uphill, requires adding free energy.



6.4 Concentration at Equilibrium No matter what quantities of glucose 1-phosphate and glucose 6-phosphate are dissolved in water, when equilibrium is attained, there will always be 95% glucose 6-phosphate and 5% glucose 1-phosphate.

start out with an aqueous solution of glucose 1-phosphate that has a concentration of 0.02 M. (M stands for molar concentration; see Chapter 2.) The solution is maintained under constant environmental conditions (25°C and pH 7). As the reaction proceeds slowly to equilibrium, the concentration of the product, glucose 6-phosphate, rises from 0 to 0.019 M, while the concentration of the reactant, glucose 1-phosphate, falls to 0.001 M. At this point, equilibrium is reached (Figure 6.4). From then on, the reverse reaction, from glucose 6-phosphate to glucose 1-phosphate, progresses at the same rate as the forward reaction.

At equilibrium, then, this reaction has a product-to-reactant ratio of 19:1 (0.019/0.001), so the forward reaction has gone 95 percent of the way to completion ("to the right," as written). Therefore, the forward reaction is an exergonic reaction. This result is obtained every time the experiment is run under the same conditions. The reaction is described by the equation

$$\text{glucose 1-phosphate} \rightleftharpoons \text{glucose 6-phosphate}$$

The change in free energy (ΔG) for any reaction is related directly to its point of equilibrium. The further toward completion the point of equilibrium lies, the more free energy is given off. In an exergonic reaction, such as the conversion of glucose 1-phosphate to glucose 6-phosphate, ΔG is a negative number (in this example, $\Delta G = -1.7$ kcal/mol, or –7.1 kJ/mol).

A large, positive ΔG for a reaction means that it proceeds hardly at all to the right (A → B). But if the product is present, such a reaction runs backward, or "to the left" (A ← B), (nearly all B is converted to A). A ΔG value near zero is characteristic of a readily reversible reaction: reactants and products have almost the same free energies.

The principles of thermodynamics we have been discussing apply to all energy exchanges in the universe, so they are very powerful and useful. Next, we'll apply them to reactions in cells that involve the biological energy currency, ATP.

ATP: Transferring Energy in Cells

All living cells rely on **adenosine triphosphate**, or **ATP**, for the capture and transfer of the free energy needed to do chemical work and maintain the cells. ATP operates as a kind of energy currency. That is, just as you may earn money from a job and then spend it on a meal, some of the free energy released by certain exergonic reactions is captured in ATP, which can then release free energy to drive endergonic reactions.

ATP is produced by cells in a number of ways (which we will describe in the next two chapters), and it is used in many ways. ATP is not an unusual molecule. In fact, it has another important use in the cell: it can be converted into a building block for DNA and RNA. But two things about ATP make it especially useful to cells: it releases a relatively large amount of energy when hydrolyzed, and it can phosphorylate (donate a phosphate group to) many different molecules. We will examine these two properties in the discussion that follows.

ATP hydrolysis releases energy

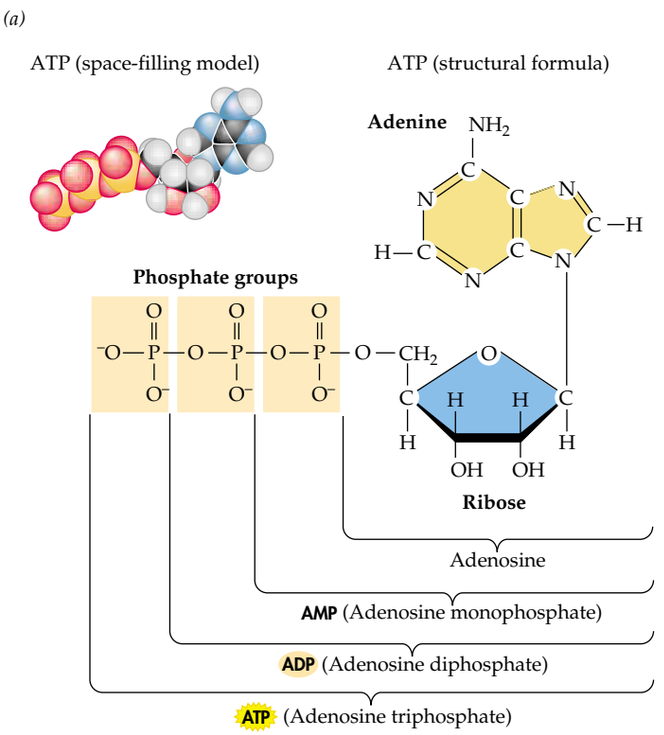
An ATP molecule consists of the nitrogenous base adenine bonded to ribose (a sugar), which is attached to a sequence of three phosphate groups (Figure 6.5). The hydrolysis of ATP yields **ADP** (**adenosine diphosphate**) and an inorganic phosphate ion (abbreviated P_i, short for HPO_4^{2-}), as well as free energy:

$$ATP + H_2O \rightarrow ADP + P_i + \text{free energy}$$

The important property of this reaction is that it is exergonic, releasing free energy. The change in free energy (ΔG) is about –12 kcal/mol (–50 kJ/mol) at the temperature, pH, and substrate concentrations typical of living cells.*

What characteristics of ATP account for the free energy released by the loss of one of its phosphate groups? First and foremost, the free energy of the P—O bond between phos-

*The "standard" ΔG for ATP hydrolysis is –7.3 kcal/mol or –30 kJ/mol, but that value is valid only at pH 7 and with ATP, ADP, and phosphate present at concentrations of 1 M—concentrations that differ greatly from those found in cells.



6.5 ATP (*a*) ATP is richer in energy than its relatives ADP and AMP. The hydrolysis of ATP releases this energy. (*b*) Fireflies use ATP to initiate the oxidation of luciferin. This converts chemical energy into light energy, emitting rhythmic flashes that signal the insect's readiness to mate. Very little of the energy in this conversion is lost as heat.

phate groups is much higher than the energy of the H—O bond that forms after hydrolysis. So some usable energy is released upon hydrolysis. Second, because phosphates are negatively charged and so repel each other, it takes energy to get phosphates near enough to each other to make the covalent bond that links them together (e.g., to add a phosphate to ADP to make ATP).

ATP couples exergonic and endergonic reactions

As we have just seen, the hydrolysis of ATP is exergonic and yields ADP, P_i, and free energy. The reverse reaction, the formation of ATP from ADP and P_i, is endergonic and consumes as much free energy as is released by the breakdown of ATP:

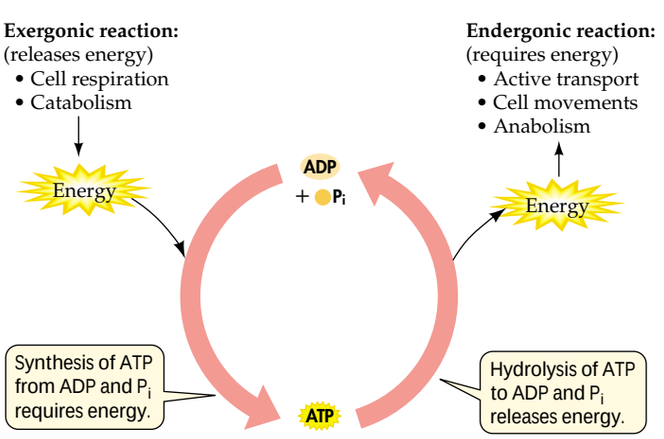
$$ADP + P_i + \text{free energy} \rightarrow ATP + H_2O$$

Many different exergonic reactions in the cell can provide the energy to convert ADP to ATP. In eukaryotes, the most important of these reactions is cellular respiration, in which some of the energy released from fuel molecules is captured in ATP. The formation and hydrolysis of ATP constitute what might be called an "energy-coupling cycle," in which ADP picks up energy from exergonic reactions to become ATP, which donates energy to endergonic reactions.

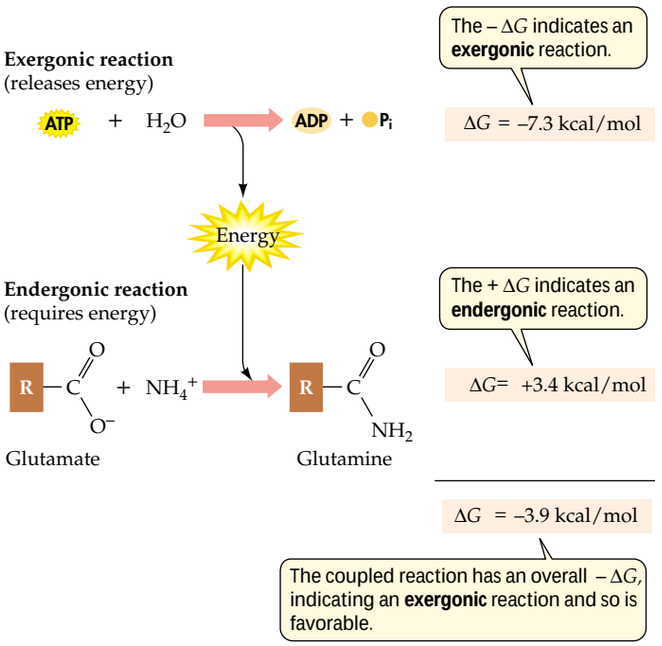
How does this ATP cycle trap and release energy? An exergonic reaction is coupled to the endergonic reaction that forms ATP from ADP and P_i (Figure 6.6). Coupling of exergonic and endergonic reactions is very common in metabolism. When it forms, ATP captures free energy and retains it like a compressed spring. ATP then diffuses to another site in the cell, where its hydrolysis releases free energy to drive an endergonic reaction.

A specific example of this energy-coupling cycle is shown in Figure 6.7. The formation of the amino acid glutamine has a positive ΔG (is endergonic) and will not proceed without the input of free energy from ATP hydrolysis, which has a negative ΔG (is exergonic). The total ΔG for the coupled reactions is negative (the two ΔGs are added together). Hence the reactions proceed exergonically when they are coupled, and glutamine is synthesized.

An active cell requires millions of molecules of ATP per second to drive its biochemical machinery. An ATP molecule



6.6 The Energy-Coupling Cycle of ATP Exergonic cellular processes release the energy needed to make ATP from ADP. The energy released from the conversion of ATP back to ADP can be used to fuel endergonic processes.



6.7 Coupling ATP Hydrolysis to an Endergonic Reaction The synthesis of the amino acid glutamine from glutamate and an ammonium ion is endergonic and must be coupled with the exergonic hydrolysis of ATP.

is consumed within a second following its formation, on average. At rest, an average person hydrolyzes and produces about 40 kg of ATP per day—as much as some people weigh! This means that each ATP molecule undergoes about 10,000 cycles of synthesis and hydrolysis every day.

Enzymes: Biological Catalysts

When we know the change in free energy (ΔG) of a reaction, we know where the equilibrium point of the reaction lies: The more negative ΔG is, the further the reaction proceeds toward completion. However, ΔG tells us nothing about the *rate* of a reaction—the speed at which it moves toward equilibrium. The reactions that occur in cells are so slow that they could not contribute to life unless the cells did something to speed them up. That is the role of **catalysts**: substances that speed up a reaction without being permanently altered by that reaction. A catalyst does not cause a reaction that would not take place eventually without it, but merely speeds up the rates of both forward and backward reactions, allowing equilibrium to be approached faster.

Most biological catalysts are proteins called **enzymes**. Although we will focus here on proteins, some catalysts—perhaps the earliest ones in the origin of life—are RNA molecules called **ribozymes** (see Chapter 3). A biological catalyst, whether protein or RNA, is a framework or scaffold in which chemical catalysis takes place. It does not matter whether the framework is RNA or protein—indeed, artificial catalysts can be made from DNA. Evolution has selected proteins as catalysts, probably because of their great diversity in three-dimensional structure and variety of chemical functions.

In the discussion that follows, we will identify the energy barrier that controls the rate of reactions. Then we'll focus on the role of enzymes: how they interact with reactants, how they lower the energy barrier, and how they permit reactions to proceed faster. After exploring the nature and significance of enzyme specificity, we'll look at how enzymes contribute to the coupling of reactions.

For a reaction to proceed, an energy barrier must be overcome

An exergonic reaction may release a great deal of free energy, but the reaction may take place very slowly. Some reactions are slow because there is an *energy barrier* between reactants and products. Think about a gas stove. The burning of the natural gas (methane + $O_2 \rightarrow CO_2 + H_2O$) is obviously an exergonic reaction—heat and light are released. Once started, the reaction goes to completion: all of the methane reacts with oxygen to form carbon dioxide and water vapor.

Because burning methane liberates so much energy, you might expect this reaction to proceed rapidly whenever methane is exposed to oxygen. But this does not happen. Simply mixing methane with air produces no reaction. Methane will start burning only if a spark—an input of energy—is provided. (On the stove, this energy is supplied by electricity.) The need for this spark to start the reaction shows that there is an energy barrier between the reactants and the products.

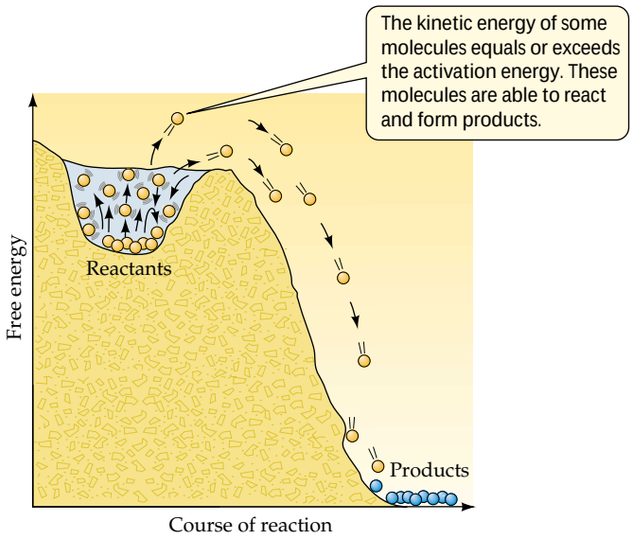
In general, exergonic reactions proceed only after the reactants are pushed over the energy barrier by a small amount of added energy. The energy barrier thus represents the amount of energy needed to start the reaction, known as the **activation energy** (E_a) (Figure 6.8*a*). Recall the ball rolling down the hill in Figure 6.3. The ball has a lot of potential energy at the top of the hill. However, if the ball is stuck in a small depression, it won't roll down the hill, even though that action is exergonic (Figure 6.8*b*). To start the ball rolling, a small amount of energy (activation energy) is needed to get the ball out of the depression (Figure 6.8*c*).

In a chemical reaction, the activation energy is the energy needed to change the reactants into unstable molecular forms called *transition-state species*. Transition-state species have higher free energies than either the reactants or the products. Their bonds may be stretched and hence unstable. Although the amount of activation energy needed for different reactions varies, it is often small compared with the change in free energy of the reaction. The activation energy that starts

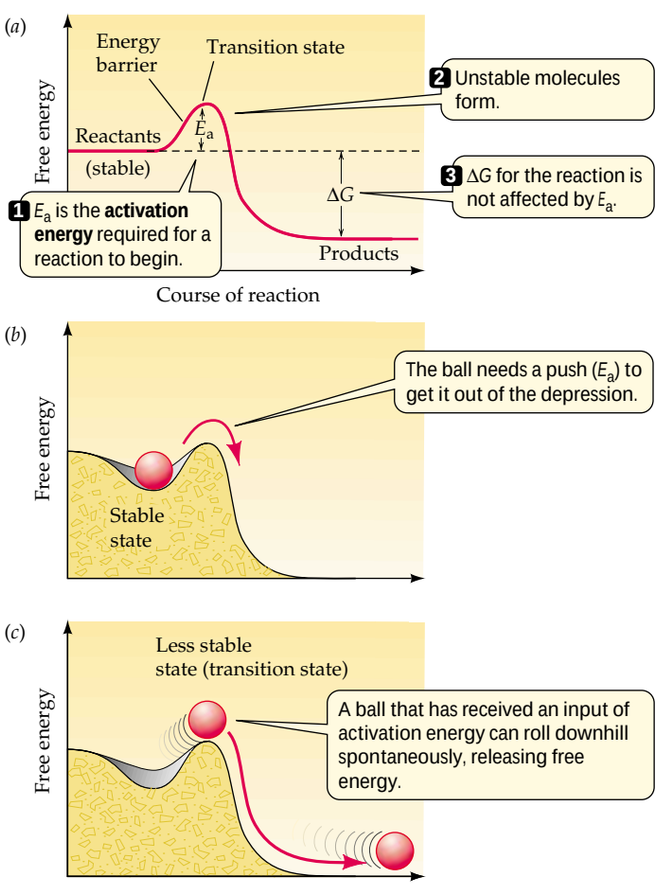
a reaction is recovered during the ensuing "downhill" phase of the reaction, so it is not a part of the net free energy change, ΔG (see Figure 6.8*a*).

Where does the activation energy come from? In any collection of reactants at room or body temperature, molecules are moving around and could use their kinetic energy of motion to overcome the energy barrier, enter the transition state, and react (Figure 6.9). However, at normal temperatures, only a few molecules have enough energy to do this; most have insufficient kinetic energy for activation, so the reaction takes place slowly. If the system were heated, all the reactant molecules would move faster and have more kinetic energy. Since more of them would have energy exceeding the required activation energy, the reaction would speed up.

However, adding enough heat to increase the average kinetic energy of the molecules won't work in living systems. Such a nonspecific approach would accelerate all reactions, including destructive ones, such as the denaturation of proteins (see Figure 3.11). A more effective way to speed up a reaction in a living system is to lower the energy barrier. In living cells, enzymes accomplish this task.



6.9 Over the Energy Barrier Some molecules have enough kinetic energy to surmount the energy barrier and react, forming products. At the temperatures of most organisms, however, only a small proportion of the molecules have that much kinetic energy.



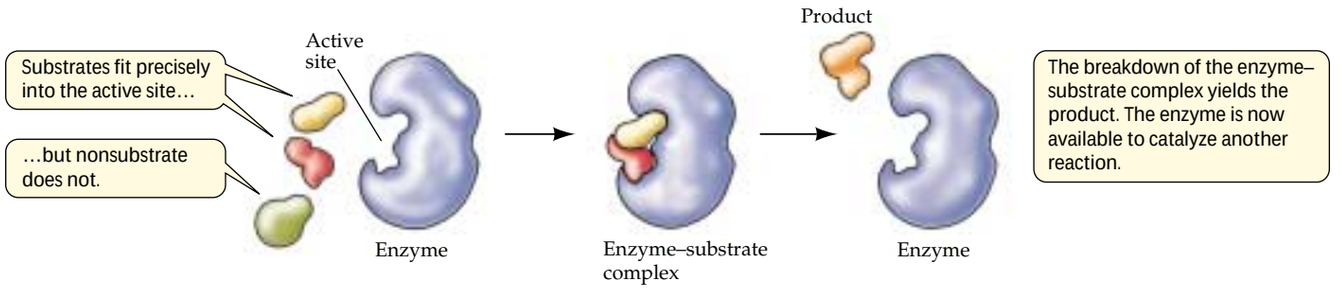
6.8 Activation Energy Initiates Reactions (*a*) In any chemical reaction, an initial stable state must become less stable before change is possible. (*b*,*c*) A ball on a hillside provides a physical analogy to the biochemical principle graphed in (*a*).

Enzymes bind specific reactant molecules

Catalysts increase the rate of chemical reactions. Most *nonbiological* catalysts are *nonspecific*. For example, powdered platinum catalyzes virtually any reaction in which molecular hydrogen (H_2) is a reactant. In contrast, most *biological* catalysts are *highly specific*. These complex molecules of protein (enzymes) or RNA (ribozymes) catalyze relatively simple chemical reactions. An enzyme or ribozyme usually recognizes and binds to only one or a few closely related reactants, and it catalyzes only a single chemical reaction. In the discussion that follows, we focus on enzymes, but you should note that similar rules of chemical behavior apply to ribozymes as well.

In an enzyme-catalyzed reaction, the reactants are called **substrates**. Substrate molecules bind to a particular site on the enzyme, called the **active site**, where catalysis takes place (Figure 6.10). The specificity of an enzyme results from the exact three-dimensional shape and structure of its active site, into which only a narrow range of substrates can fit. Other molecules—with different shapes, different functional groups, and different properties—cannot properly fit and bind to the active site.

The names of enzymes reflect the specificity of their functions and often end with the suffix "-ase." For example, the



6.10 Enzyme and Substrate An enzyme is a protein catalyst with an active site capable of binding one or more substrate molecules. The enzyme–substrate complex yields product and free enzyme.

enzyme RNA polymerase catalyzes the formation of RNA, but not DNA, and the enzyme hexokinase accelerates the phosphorylation of hexose sugars, but not pentose sugars.

The binding of a substrate to the active site of an enzyme produces an **enzyme–substrate complex** (ES) held together by one or more means, such as hydrogen bonding, ionic attraction, or covalent bonding. The enzyme–substrate complex gives rise to product and free enzyme:

$$E + S \rightarrow ES \rightarrow E + P$$

where E is the enzyme, S is the substrate, P is the product, and ES is the enzyme–substrate complex. The free enzyme (E) is in the same chemical form at the end of the reaction as at the beginning. While bound to the substrate, it may change chemically, but by the end of the reaction it has been restored to its initial form.

Enzymes lower the energy barrier but do not affect equilibrium

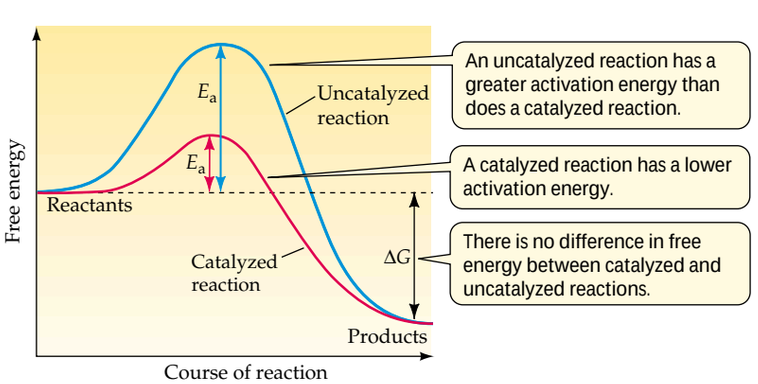
When reactants are part of an enzyme–substrate complex, they require less activation energy than the transition-state species of the corresponding uncatalyzed reaction (Figure 6.11). Thus the enzyme lowers the energy barrier for the reaction—it offers the reaction an easier path. When an enzyme lowers the energy barrier, both the forward and the reverse reactions speed up, so the enzyme-catalyzed overall reaction proceeds toward equilibrium more rapidly than the uncatalyzed reaction. The final equilibrium (and ΔG) is the same with or without the enzyme.

Adding an enzyme to a reaction does not change the difference in free energy (ΔG) between the reactants and the products (see Figure 6.11). It does change the activation energy and, consequently, the rate of reaction. For example, if 600 molecules of a protein with arginine as its terminal amino acid just sit in solution, the proteins tend toward disorder, and the terminal peptide bonds break, releasing the arginines (ΔS increases). After 7 years, about half (300) of the proteins will have undergone this reaction. With the enzyme carboxypeptidase A catalyzing the reaction, however, the 300 arginines are released in half a second!

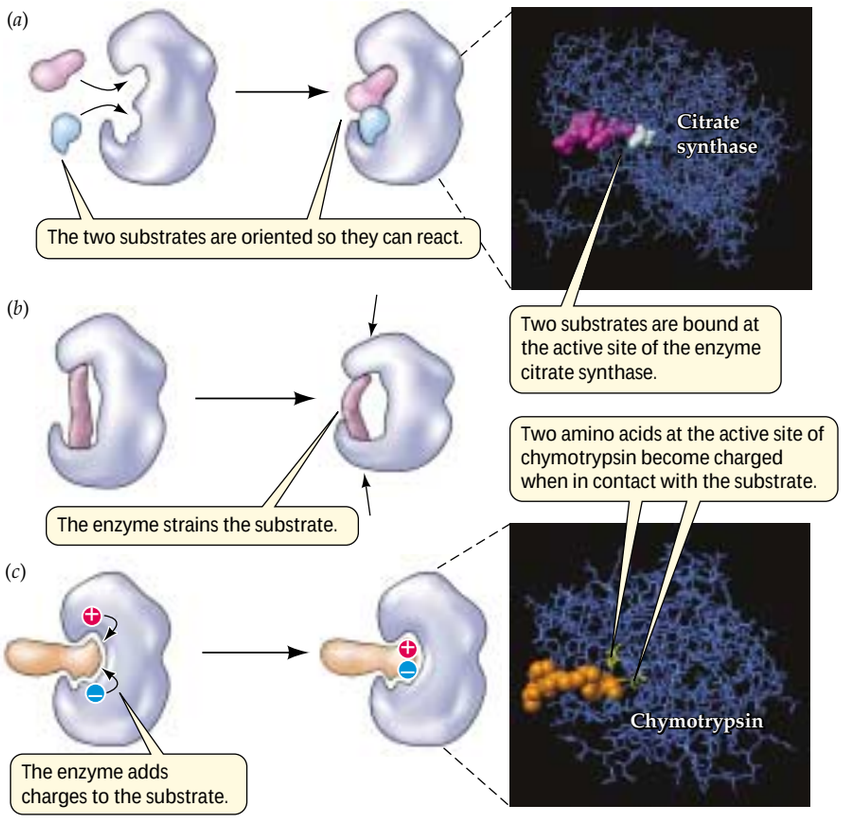
What are the chemical events at active sites of enzymes?

After formation of the enzyme–substrate complex, chemical interactions occur. These interactions contribute directly to the breaking of old bonds and the formation of new ones (Figure 6.12). In catalyzing a reaction, an enzyme may use one or more of the following mechanisms:

ENZYMES ORIENT SUBSTRATES. While free in solution, substrates are rotating and tumbling around and may not have the proper orientation to interact when they collide. Part of the activation energy needed to start a reaction is used to make the substrates collide with the right atoms for bond formation next to each other. When proteins are synthe-



6.11 Enzymes Lower the Energy Barrier Although the activation energy is lower in an enzyme-catalyzed reaction than in an uncatalyzed reaction, the energy released is the same with or without catalysis. In other words, E_a is lower, but ΔG is unchanged.

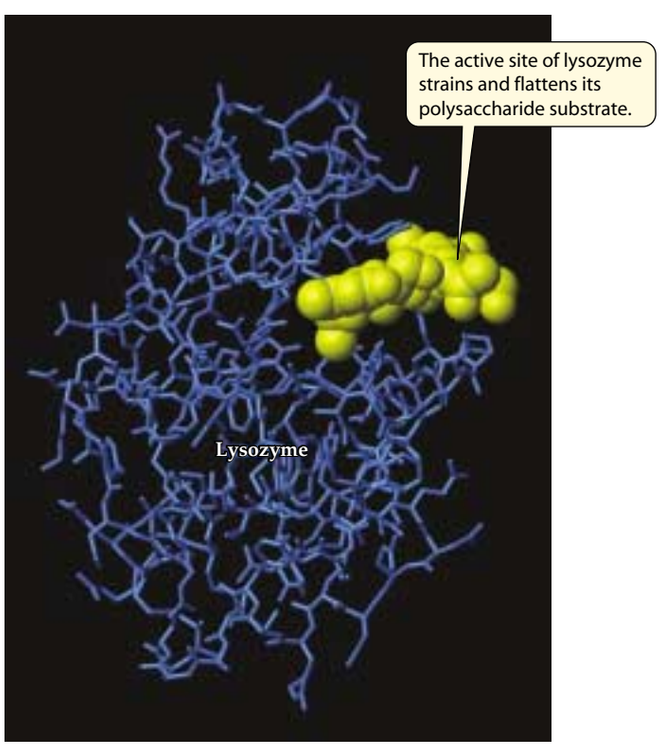


6.12 Life at the Active Site Enzymes have several ways of causing their substrates to enter the transition state: (*a*) orientation, (*b*) chemical change, and (*c*) physical strain.

sized, for example, a peptide bond is formed between the carboxyl group of one amino acid and the amino group of the next (see Figure 3.5). If two amino acids are to form a peptide bond, the carboxyl group of one and amino group of the other must be the sites of collision. When the active site of an enzyme binds to one amino acid, however, it is held in the right orientation to react with a second amino acid when that substrate binds to the enzyme.

ENZYMES INDUCE STRAIN IN THE SUBSTRATE. Once a substrate has bound to the active site, the enzyme can cause bonds in the substrate to stretch, putting it in an unstable transition state. For example, the polysaccharide substrate for the enzyme lysozyme enters the active site in a flat-ringed "chair" shape, but the active site quickly causes it to flatten out into a "sofa" (Figure 6.13). The resulting stretching of its bonds causes them to be less stable and more reactive to the enzyme's other substrate, water.

ENZYMES TEMPORARILY ADD CHEMICAL GROUPS TO SUBSTRATES. The side chains (R groups) of an enzyme's amino acids may be direct participants in making its substrates more chemically reactive. For example, in *acid-base catalysis*, the acidic or basic side chains of the amino acids forming the active site may transfer H^+ to or from the substrate, destabilizing a covalent bond in the substrate and permitting it to break. In *covalent catalysis*, a functional group in a side chain forms a temporary covalent bond with a portion of the substrate. In *metal ion catalysis*, metal ions such as copper, zinc, iron, and manganese, which are firmly bound to side chains of the protein, can lose or gain electrons without detaching from the protein (Figure 6.12*c*). This ability makes them important participants in oxidation–reduction reactions, which involve loss or gain of electrons.



6.13 Tertiary Structure of Lysozyme Lysozyme is an enzyme that protects the animals that produce it by destroying invading bacteria. To destroy the bacteria, it cleaves certain polysaccharide chains in their cell walls.

Molecular Structure Determines Enzyme Function

Most enzymes (and ribozymes) are much larger than their substrates. An enzyme is typically a protein containing hundreds of amino acids, and may consist of a single folded polypeptide chain or several subunits. Its substrate is generally a small molecule. The active site of the enzyme is usually quite small, not more than 6–12 amino acids. Two questions arise from this observation:

- What is the nature of the active site that allows it to recognize and bind the substrate?
- What is the role of the rest of the huge protein?

The active site is specific to the substrate

The remarkable ability of an enzyme to select exactly the right substrate depends on a precise interlocking of molecular shapes and interactions of chemical groups at the binding site. The binding of the substrate to the active site depends on the same kinds of forces that maintain the tertiary structure of the enzyme: hydrogen bonds, the attraction and repulsion of electrically charged groups, and hydrophobic interactions.

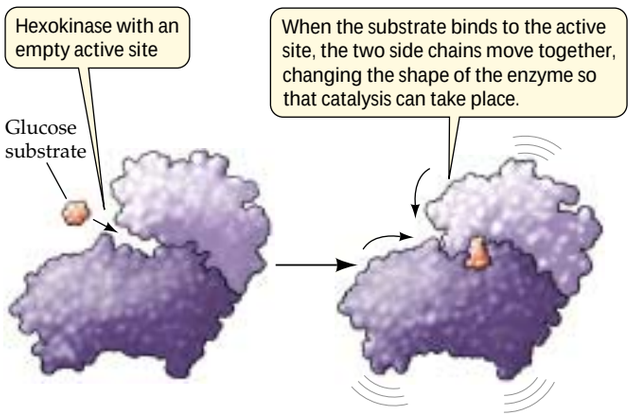
In 1894, the German chemist Emil Fischer compared the fit between an enzyme and its substrate to that of a lock and key. Fischer's model persisted for more than half a century with only indirect evidence to support it. The first direct evidence came in 1965, when David Phillips and his colleagues at the Royal Institution in London succeeded in crystallizing the enzyme lysozyme and determined its tertiary structure using the techniques of X-ray crystallography (described in Chapter 11). They observed a pocket in lysozyme that neatly fits its substrate (see Figure 6.13).

An enzyme changes shape when it binds a substrate

As proteins, enzymes are not immutable structures. Just as the structure of egg white protein changes when the egg is heated, many enzymes change their structure (albeit less dramatically) when they bind to their substrates. These shape changes expose those regions of the enzyme—the active sites—that actually react with the substrate. Such a change in enzyme shape caused by substrate binding is called **induced fit**.

Induced fit can be observed in the enzyme hexokinase (Figure 6.14) when it is studied with and without one of its substrates, glucose (its other substrate is ATP). It catalyzes the reaction

glucose + ATP → glucose 6-phosphate + ADP



6.14 Some Enzymes Change Shape When Substrate Binds to Them Shape changes result in an induced fit between enzyme and substrate, improving the catalytic ability of the enzyme.

Induced fit brings reactive side chains from the enzyme's active site into alignment with the substrates, facilitating the catalytic mechanisms described earlier (see Figure 6.12).

Equally important, the folding of hexokinase to fit around the glucose substrate excludes water from the active site. This is essential, because the two molecules binding to the active site are glucose and ATP. If water were present, ATP could be hydrolyzed to ADP and phosphate. But since water is absent, the transfer of a phosphate from ATP to glucose is favored.

Induced fit at least partly explains why enzymes are so large. The rest of the macromolecule may have two roles:

- It provides a framework so that the amino acids of the active site are properly positioned in relation to the substrate.
- It participates in the small but significant changes in protein shape and structure that result in induced fit.

Some enzymes require other molecules in order to operate

As large and complex as enzymes are, many of them require the presence of other, nonprotein molecules in order to function (Table 6.1). Some of these molecular "partners" include:

- **Cofactors**. These are inorganic ions such as copper, zinc, or iron that bind to certain enzymes and are essential to their function.
- **Coenzymes**. These carbon-containing molecules are required for the action of one or more enzymes. Coenzymes are usually relatively small compared with the enzyme to which they temporarily bind (Figure 6.15).

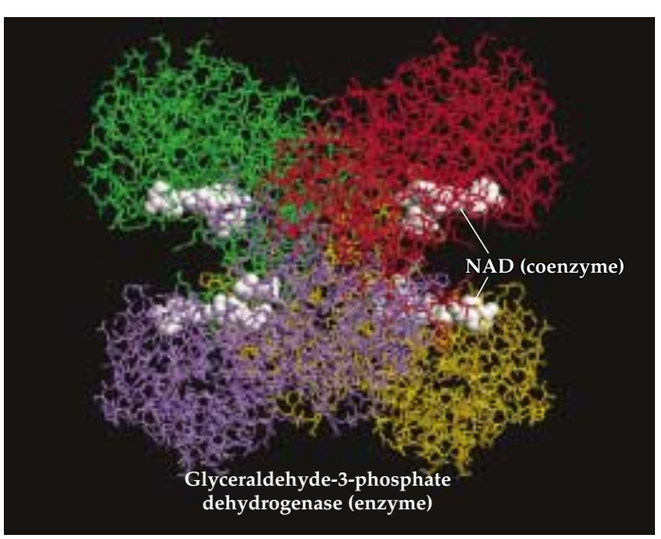
6.1 A Few Examples of Nonprotein "Partners" of Enzymes

TYPE OF MOLECULE	ROLE IN CATALYZED REACTIONS
Cofactors	
Iron (Fe^{2+} or Fe^{3+})	Oxidation/reduction
Copper (Cu^{+} or Cu^{2+})	Oxidation/reduction
Zinc (Zn^{2+})	Helps bind NAD
Coenzymes	
Biotin	Carries $—COO^-$
Coenzyme A	Carries $—CH_2—CH_3$
NAD	Carries electrons
FAD	Carries electrons
ATP	Provides/extracts energy
Prosthetic groups	
Heme	Binds ions, O_2, and electrons; contains iron cofactor
Flavin	Binds electrons
Retinal	Converts light energy

▸ **Prosthetic groups**. These distinctive molecular groupings are permanently bound to their enzymes. They include the heme groups that are attached to the oxygen-carrying protein hemoglobin (shown in Figure 3.8).

Coenzymes are like substrates in that they are not permanently bound to the enzyme, and must collide with the enzyme and bind to its active site. A coenzyme can be considered a substrate because it changes chemically during the reaction and then separates from the enzyme to participate in other reactions. Coenzymes move from enzyme molecule to enzyme molecule, adding or removing chemical groups from the substrate.

ATP and ADP can be considered coenzymes because they are necessary for some reactions, are changed by reactions, and bind to and detach from the enzyme. In the next chapter, we will encounter coenzymes that function in energy processing by accepting or donating electrons or hydrogen atoms. In animals, some coenzymes are produced from *vitamins* that must be obtained from food—they cannot be synthesized by the body. For example, the B vitamin niacin is used to make the coenzyme NAD.

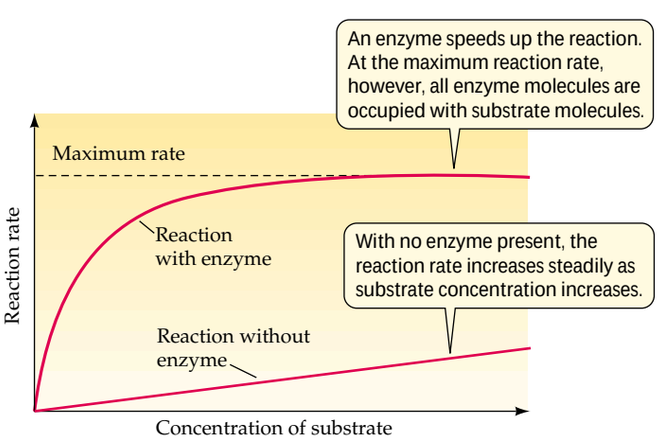


6.15 An Enzyme with a Coenzyme Some enzymes require coenzymes in order to function. This illustration shows the relative sizes of the four subunits (red, orange, green, and purple) of the enzyme glyceraldehyde 3-phosphate dehydrogenase and its coenzyme, NAD (white).

Substrate concentration affects reaction rate

For a reaction of the type A → B, the rate of the uncatalyzed reaction is directly proportional to the concentration of A (Figure 6.16). The higher the concentration of substrate, the more reactions per unit of time. Addition of the appropriate enzyme speeds up the reaction, of course, but it also changes the shape of the plot of rate versus substrate concentration. At first, the rate of the enzyme-catalyzed reaction increases as the substrate concentration increases, but then it levels off. When further increases in the substrate concentration do not significantly increase the reaction rate, the maximum rate is attained.

Since the concentration of an enzyme is usually much lower than that of its substrate, what we are seeing is a *saturation* phenomenon like the one that occurs in facilitated diffusion (see Chapter 5). When all the enzyme molecules are bound to substrate molecules, the enzyme is working as fast as it can—at its maximum rate. Nothing is gained by adding more substrate, because no free enzyme molecules are left to act as catalysts.



6.16 Catalyzed Reactions Reach a Maximum Rate Because there is usually less enzyme than substrate present, the reaction rate levels off when the enzyme becomes saturated.

The maximum rate of an enzyme reaction can be used to measure how efficient the enzyme can be—that is, how many molecules of substrate are converted to product per unit of time when there is an excess of substrate present. This turnover number ranges from 1 molecule every 2 seconds for lysozyme (see Figure 6.13) to an amazing 40 million molecules per second for the liver enzyme catalase.

Metabolism and the Regulation of Enzymes

A major characteristic of life is *homeostasis*, the maintenance of stable internal conditions. Regulation of the rates at which our thousands of different enzymes operate contributes to metabolic homeostasis. In the remainder of this chapter, we will investigate the role of enzymes in organizing and regulating metabolism. In living cells, the activity of enzymes can be activated or inhibited in various ways, so the presence of an enzyme does not necessarily ensure that it is functioning. There are mechanisms that alter the rate at which some enzymes catalyze reactions, making enzymes the target points at which entire sequences of chemical reactions can be regulated. Finally, we examine how the environment—namely, temperature and pH—affects enzyme activity.

Metabolism is organized into pathways

An organism's **metabolism** is the totality of the biochemical reactions that take place within it. Metabolism transforms raw materials and stored potential energy into forms that can be used by living cells. Metabolism consists of sequences of enzyme-catalyzed chemical reactions called **pathways**. In these sequences, the product of one reaction is the substrate for the next:

$$A \xrightarrow{\text{enzyme 1}} B \xrightarrow{\text{enzyme 2}} C \xrightarrow{\text{enzyme 3}} D$$

Some metabolic pathways are anabolic, synthesizing the important chemical building blocks from which macromolecules are built. Others are catabolic, breaking down molecules for usable free energy, recycling monomers, or inactivating toxic substances. The balance among these anabolic and catabolic pathways may change depending on the cell's (and the organism's) needs. So a cell must regulate all its metabolic pathways constantly.

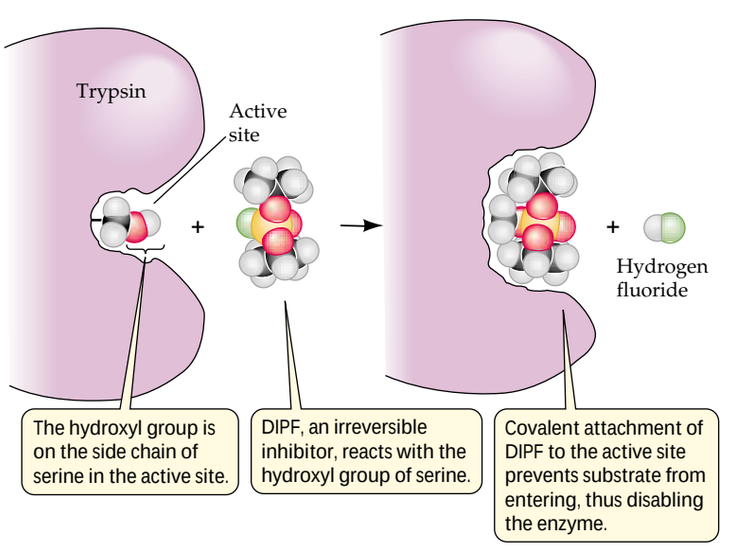
Enzyme activity is subject to regulation by inhibitors

Various *inhibitors* can bind to enzymes, slowing down the rates of enzyme-catalyzed reactions. Some inhibitors occur naturally in cells; others are artificial. Naturally occurring inhibitors regulate metabolism; artificial ones can be used to treat disease, to kill pests, or in the laboratory to study how enzymes work. Some inhibitors irreversibly inhibit the enzyme by permanently binding to it. Others have reversible effects; that is, they can become unbound from the enzyme. The removal of a natural reversible inhibitor increases an enzyme's rate of catalysis.

IRREVERSIBLE INHIBITION. Some inhibitors covalently bond to certain side chains at the active sites of an enzyme, thereby permanently inactivating the enzyme by destroying its capacity to interact with its normal substrate. At the beginning of this chapter we described aspirin, which adds an acetyl group to a serine residue at the active site of cyclooxygenase, preventing this serine from taking part in chemical catalysis.

Another example of an **irreversible inhibitor** is DIPF (diisopropylphosphorofluoridate), which also reacts with serine (Figure 6.17). DIPF is an irreversible inhibitor of acetylcholinesterase, an enzyme that is essential for the orderly propagation of impulses from one nerve cell to another. Because of their effect on acetylcholinesterase, DIPF and other similar compounds are classified as *nerve gases*. One of them, Sarin, was used in an attack on the Tokyo subway in 1995, resulting in a dozen deaths and hundreds hospitalized. The widely used insecticide malathion is a derivative of DIPF that inhibits only insect acetylcholinesterase, not the mammalian enzyme.

REVERSIBLE INHIBITION. Not all inhibition is irreversible. Some inhibitors are similar enough to a particular enzyme's natural substrate to bind noncovalently to its active site, yet different enough that the enzyme catalyzes no chemical reaction. While such a molecule is bound to the enzyme,



6.17 Irreversible Inhibition DIPF forms a stable covalent bond with the side chain of the amino acid serine at the active site of the enzyme trypsin.

the natural substrate cannot enter the active site; thus, the inhibitor effectively wastes the enzyme's time, preventing its catalytic action. Such molecules are called **competitive inhibitors** because they compete with the natural substrate for the active site (Figure 6.18*a*). In these cases, the inhibition is reversible. When the concentration of the competitive inhibitor is reduced, it detaches from the active site, and the enzyme is again active.

The enzyme succinate dehydrogenase is subject to competitive inhibition. This enzyme, found in all mitochondria, catalyzes the conversion of the compound succinate to fumarate. A third molecule, oxaloacetate, is similar to succinate and can act as a competitive inhibitor of succinate dehydrogenase by binding to its active site. Once bound to oxaloacetate, the enzyme can do nothing more with it—no reaction occurs. An enzyme molecule cannot bind a succinate molecule until the oxaloacetate molecule has moved out of its active site—which can occur if more substrate (succinate) molecules are added.

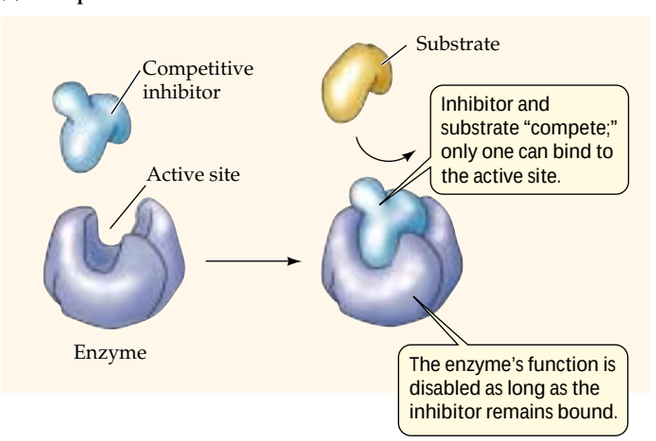
Some inhibitors that do not react with the active site are called **noncompetitive inhibitors**. Noncompetitive inhibitors bind to the enzyme at a site distinct from the active site. Their binding can cause a conformational change in the enzyme that alters the active site (Figure 6.18*b*). In this case, the active site may still bind substrate molecules, but the rate of product formation may be reduced. Noncompetitive inhibitors, like competitive inhibitors, can become unbound, so their effects are reversible.

Allosteric enzymes control their activity by changing their shape

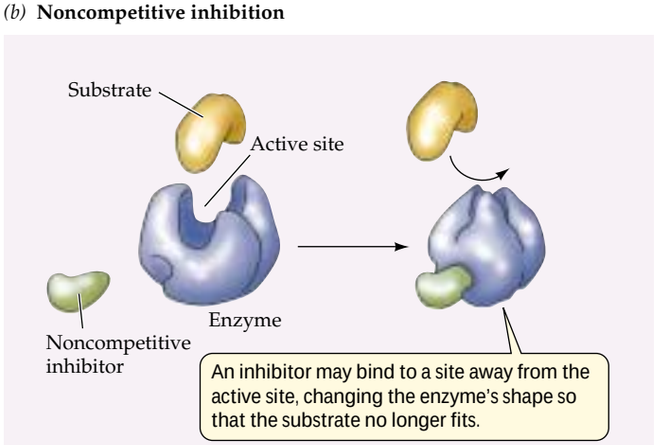
The change in enzyme shape due to noncompetitive inhibitor binding is an example of **allostery** (*allo-*, "different"; *-stery*, "shape"). In that case, the binding of the inhibitor *induces* the protein to change its shape. More common are enzymes that

6.18 Reversible Inhibition (*a*) A competitive inhibitor binds temporarily to the active site of an enzyme. Succinate dehydrogenase, for example, is subject to competitive inhibition by oxaloacetate. (*b*) A noncompetitive inhibitor binds temporarily to the enzyme at a site away from the active site, but still blocks enzyme function.

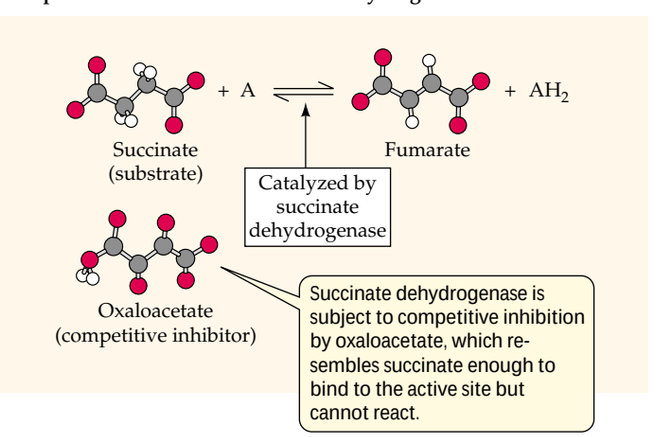
(a) **Competitive inhibition**



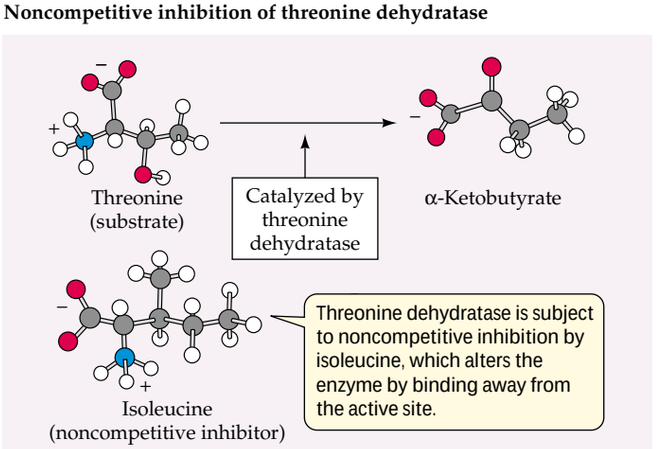
(b) **Noncompetitive inhibition**

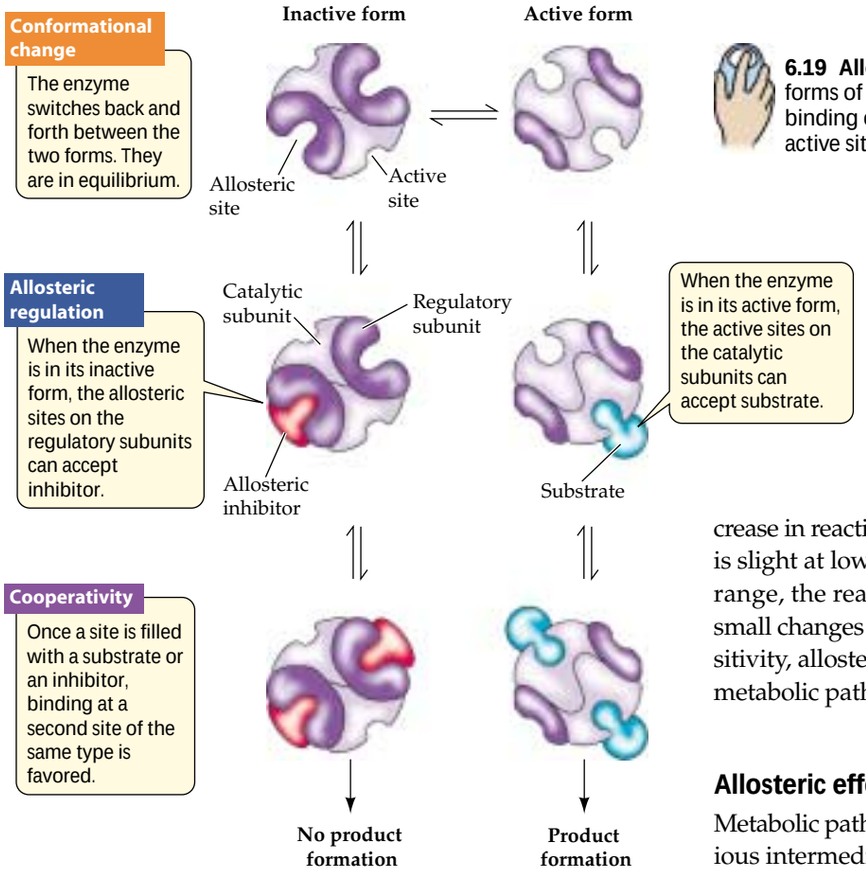


Competitive inhibition of succinate dehydrogenase



Noncompetitive inhibition of threonine dehydratase





6.19 Allosteric Regulation of Enzymes Active and inactive forms of an enzyme are interconverted, depending on the binding of regulatory molecules at a location distant from the active site.

already exist in the cell in more than one possible shape. The inactive form of the enzyme has a shape that cannot bind the substrate, while the active form has the proper shape at the active site to bind the substrate. These two forms can interconvert, and this process is regulated by the binding of an **allosteric regulator** to a site on the enzyme away from the active site. Regulator binding is just like substrate binding: it is highly specific. So an enzyme may have several sites for binding: one for the substrate(s) and others for regulators.

Allosteric regulators work in two ways:

- *Positive regulators* stabilize the active form of the enzyme.
- *Negative regulators* stabilize the inactive form of the enzyme (Figure 6.19).

Most (but not all) enzymes that are allosterically regulated are proteins with quaternary structure; that is, they are made up of multiple polypeptide subunits. The active site is present on one subunit, called the *catalytic subunit*, while the regulatory site(s) are present on different subunit(s), the *regulatory subunit(s)*.

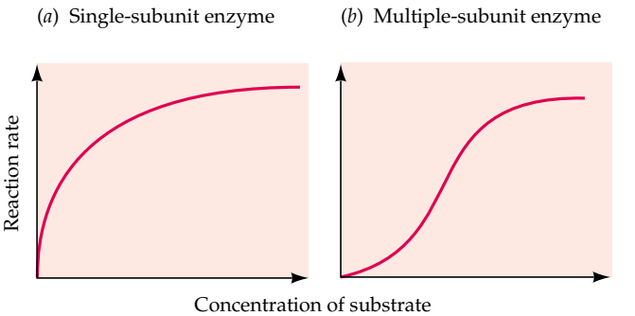
Allosteric enzymes and nonallosteric enzymes differ greatly in their reaction rates when the substrate concentration is low. Graphs of reaction rate plotted against substrate concentration show this relationship. For an enzyme with a single subunit, the plot looks like that in Figure 6.20*a*. The reaction rate first increases very sharply with increasing substrate concentration, then tapers off to a constant maximum rate as the supply of enzyme becomes saturated with substrate.

The plot for many allosteric enzymes is radically different, having a *sigmoid* (S-shaped) appearance (Figure 6.20*b*). The increase in reaction rate with increasing substrate concentration is slight at low substrate concentrations, but within a certain range, the reaction rate is extremely sensitive to relatively small changes in substrate concentration. Because of this sensitivity, allosteric enzymes are important in regulating entire metabolic pathways.

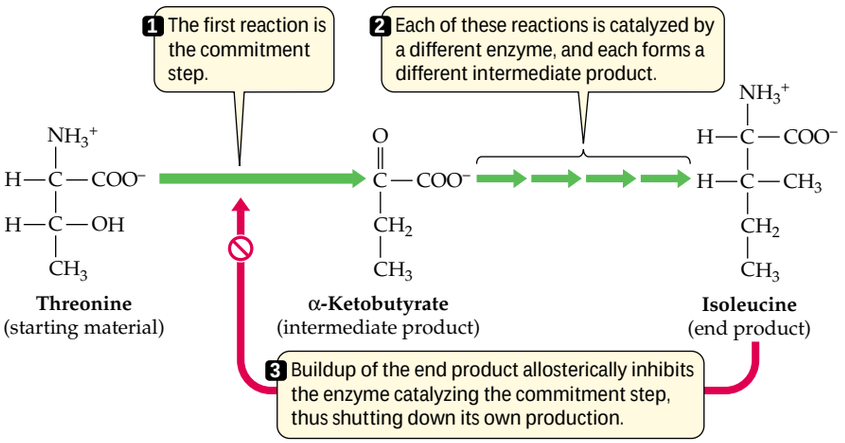
Allosteric effects regulate metabolism

Metabolic pathways typically involve a starting material, various intermediate products, and an end product that is used for some purpose by the cell. In each pathway, there are a number of reactions, each forming an intermediate product and each catalyzed by a different enzyme. The first step in a pathway is called the **commitment step**, meaning that once this enzyme-catalyzed reaction occurs, the "ball is rolling," and the other reactions happen in sequence, leading to the end product. But what if the cell has no need for that product—for example, if that product is available from its environment in adequate amounts? It would be energetically wasteful for the cell to continue making something it does not need.

One way that cells solve this problem is to shut down the metabolic pathway by having the final product allosterically



6.20 Allostery and Reaction Rate How the rate of an enzyme-catalyzed reaction changes with increasing substrate concentration depends on whether the enzyme is allosterically regulated.



6.21 Inhibition of Metabolic Pathways The commitment step is catalyzed by an allosteric enzyme that can be inhibited by the end product of the pathway. The specific pathway shown here is the synthesis of isoleucine, an amino acid, from threonine in bacteria. This pathway is typical of many enzyme-catalyzed biological reactions.

inhibit the enzyme that catalyzes the commitment step (Figure 6.21). This mechanism is known as **end-product inhibition** or **feedback inhibition**. When the end product is present in a high concentration, some of it binds to an allosteric site on the commitment step enzyme, thereby causing it to become inactive. We will describe many other examples of allosteric interactions in later chapters.

Enzymes are affected by their environment

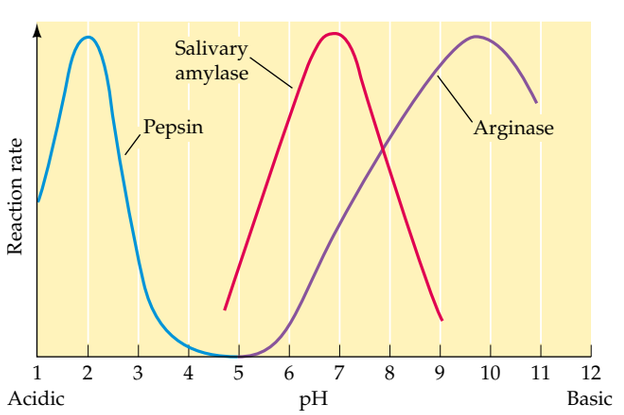
Enzymes enable cells to perform chemical reactions and carry out complex processes rapidly without using the extremes of temperature and pH employed by chemists in the laboratory. However, because of their three-dimensional structures and the chemistry of the side chains in their active sites, enzymes are highly sensitive to temperature and pH. We described the general effects of these environmental factors on proteins in Chapter 3. Here, we will examine their effects on enzyme function, which, of course, depends on enzyme structure and chemistry.

pH AFFECTS ENZYME ACTIVITY. The rates of most enzyme-catalyzed reactions depend on the pH of the medium in which they occur. Each enzyme is most active at a particular pH; its activity decreases as the solution is made more acidic or more basic than its "ideal" (optimal) pH (Figure 6.22).

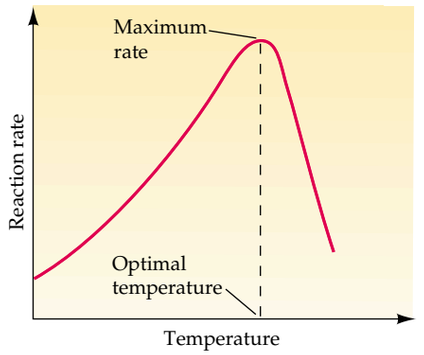
Several factors contribute to this effect. One is the ionization of carboxyl, amino, and other groups on either the substrate or the enzyme. In neutral or basic solutions, carboxyl groups (—COOH) release H^+ to become negatively charged carboxylate groups ($—COO^-$). Similarly, amino groups ($—NH_2$) accept H^+ ions in neutral or acidic solutions, becoming positively charged $—NH_3^+$ groups (see Chapter 2). Thus, in a neutral solution, a molecule with an amino group is attracted electrically to another molecule that has a carboxyl group, because both groups are ionized and the two groups have opposite charges.

If the pH changes, however, the ionization of these groups may change. For example, at a low pH (high H^+ concentration), the excess H^+ may react with the $—COO^-$ to form COOH. If this happens, the group is no longer charged and cannot interact with other charged groups in the protein, so the folding of the protein may be altered. If such a change occurs at the active site of an enzyme, the enzyme may no longer have the correct shape to bind to its substrate.

TEMPERATURE AFFECTS ENZYME ACTIVITY. In general, warming increases the rate of an enzyme-catalyzed reaction because at higher temperatures, a greater fraction of the reactant molecules have enough energy to provide the activation energy for the reaction (Figure 6.23). Temperatures that are too high, however, inactivate enzymes, because at high temperatures enzyme molecules vibrate and twist so rapidly that some of their noncovalent bonds break. When heat changes their tertiary structure, enzymes become inactivated, or thermally denatured. Some enzymes denature at temperatures only slightly above that of the human body, but a few are stable even at the boiling or freezing points of water. All enzymes, however, show an optimal temperature for activity.



6.22 pH Affects Enzyme Activity Each enzyme catalyzes its reaction at a maximum rate at a particular pH. The activity curves peak at the pH where each enzyme is most effective.



6.23 Temperature Affects Enzyme Activity Each enzyme is most active at a particular optimal temperature. At higher temperatures, denaturation reduces the enzyme's activity.

Individual organisms adapt to changes in the environment in many ways, one of which is based on groups of enzymes, called **isozymes**, that catalyze the same reaction but have different chemical compositions and physical properties. Different isozymes within a given group may have different optimal temperatures. The rainbow trout, for example, has several isozymes of the enzyme acetylcholinesterase, whose operation is essential to the normal transmission of nerve impulses. If a rainbow trout is transferred from warm water to near-freezing water (2°C), the fish produces an isozyme of acetylcholinesterase that is different from the one it produces at the higher temperature. The new isozyme has a lower optimal temperature, allowing the fish to perform normally in the colder water.

In general, enzymes adapted to warm temperatures fail to denature at those temperatures because their tertiary structures are held together largely by covalent bonds, such as disulfide bridges, instead of the more heat-sensitive weak chemical interactions. Most enzymes in humans are more stable at high temperatures than those of the bacteria that infect us, so that a moderate fever tends to denature bacterial enzymes, but not our own.

Chapter Summary

Energy and Energy Conversions

▶ Energy is the capacity to do work. Potential energy is the energy of state or position; it includes the energy stored in chemical bonds. Kinetic energy is the energy of motion (and related forms such as electric energy, light, and heat).

▶ Potential energy can be converted to kinetic energy, which can do work. **Review Figure 6.1**

▶ Living things, like everything else, obey the laws of thermodynamics. The first law of thermodynamics tells us that energy cannot be created or destroyed. The second law of thermodynamics tells us that the quantity of energy available to do work (free energy) decreases and unusable energy (associated with entropy) increases. **Review Figure 6.2**

▶ Changes in free energy, total energy, temperature, and entropy are related by the equation $\Delta G = \Delta H - T\Delta S$.

▶ Exergonic reactions release free energy and have a negative ΔG. Endergonic reactions take up free energy and have a positive ΔG. Endergonic reactions proceed only if free energy is provided. **Review Figure 6.3**

▶ The change in free energy (ΔG) of a reaction determines its point of chemical equilibrium, at which the forward and reverse reactions proceed at the same rate. For exergonic reactions, the equilibrium point lies toward completion (the conversion of all reactants into products). **Review Figure 6.4**

ATP: Transferring Energy in Cells

▶ ATP (adenosine triphosphate) serves as an energy currency in cells. Hydrolysis of ATP releases a relatively large amount of free energy. **Review Figure 6.5**

▶ The ATP cycle couples exergonic and endergonic reactions, transferring free energy from the exergonic to the endergonic reaction. **Review Figures 6.6, 6.7.** **See web/CD Activity 6.1**

Enzymes: Biological Catalysts

▶ The rate of a chemical reaction is independent of ΔG, but is determined by the size of the energy barrier. Catalysts speed reactions by lowering the energy barrier. **Review Figures 6.8, 6.9**

▶ Enzymes are biological catalysts, proteins that are highly specific for their substrates. Substrates bind to the active site, where catalysis takes place, forming an enzyme–substrate complex. **Review Figure 6.10**

▶ At the active site, a substrate can be oriented correctly, chemically modified, or strained. As a result, the substrate readily forms its transition state, and the reaction proceeds. **Review Figures 6.11, 6.12.** **See web/CD Activity 6.2**

Molecular Structure Determines Enzyme Function

▶ The active site where substrate binds determines the specificity of an enzyme. Upon binding to substrate, some enzymes change shape, facilitating catalysis. **Review Figures 6.13, 6.14**

▶ Some enzymes require cofactors to carry out catalysis. Prosthetic groups are permanently bound to the enzyme. Coenzymes are not usually bound to the enzyme. They can be considered substrates, as they are changed by the reaction and then released from the enzyme. **Review Table 6.1 and Figure 6.15**

▶ Substrate concentration affects the rate of an enzyme-catalyzed reaction. **Review Figure 6.16**

Metabolism and the Regulation of Enzymes

▶ Metabolism is organized into pathways in which the product of one reaction is a reactant for the next reaction. Each reaction in the pathway is catalyzed by an enzyme.

▶ Enzyme activity is subject to regulation. Some inhibitors react irreversibly with enzymes and block their catalytic activity. Others react reversibly with enzymes, inhibiting their action only temporarily. A compound closely similar in structure to an enzyme's normal substrate may competitively inhibit the action of the enzyme. **Review Figures 6.17, 6.18.** **See web/CD Tutorial 6.1**

▶ Allosteric regulators bind to a site different from the active site and stabilize the active or inactive form of an enzyme. Many such enzymes have multiple subunits. **Review Figure 6.19.** **See web/CD Tutorial 6.2**

▶ For allosteric enzymes, plots of reaction rate versus substrate concentration are sigmoid, in contrast to plots of the same variables for nonallosteric enzymes. **Review Figure 6.20**

▶ The end product of a metabolic pathway may inhibit the allosteric enzyme that catalyzes the commitment step of that pathway. **Review Figure 6.21**

▶ Enzymes are sensitive to their environment. Both pH and temperature affect enzyme activity. **Review Figures 6.22, 6.23**

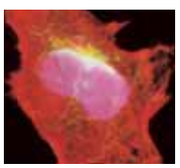
Self-Quiz

1. Coenzymes differ from enzymes in that coenzymes are
   a. only active outside the cell.
   b. polymers of amino acids.
   c. smaller, such as vitamins.
   d. specific for one reaction.
   e. always carriers of high-energy phosphate.
2. Which statement about thermodynamics is true?
   a. Free energy is used up in an exergonic reaction.
   b. Free energy cannot be used to do work, such as chemical transformations.
   c. The total amount of energy can change after a chemical transformation.
   d. Free energy can be kinetic but not potential energy
   e. Entropy tends always to a maximum.
3. In a chemical reaction,
   a. the rate depends on the value of ΔG.
   b. the rate depends on the activation energy.
   c. the entropy change depends on the activation energy.
   d. the activation energy depends on the value of ΔG.
   e. the change in free energy depends on the activation energy.
4. Which statement about enzymes is *not* true?
   a. They consist of proteins, with or without a nonprotein part.
   b. They change the rate of the catalyzed reaction.
   c. They change the value of ΔG of the reaction.
   d. They are sensitive to heat.
   e. They are sensitive to pH.
5. The active site of an enzyme
   a. never changes shape.
   b. forms no chemical bonds with substrates.
   c. determines, by its structure, the specificity of the enzyme.
   d. looks like a lump projecting from the surface of the enzyme.
   e. changes ΔG of the reaction.
6. The molecule ATP is
   a. a component of most proteins.
   b. high in energy because of the presence of adenine (A).
   c. required for many energy-producing biochemical reactions.
   d. a catalyst.
   e. used in some endergonic reactions to provide energy.
7. In an enzyme-catalyzed reaction,
   a. a substrate does not change.
   b. the rate decreases as substrate concentration increases.
   c. the enzyme can be permanently changed.
   d. strain may be added to a substrate.
   e. the rate is not affected by substrate concentration.
8. Which statement about enzyme inhibitors is *not* true?
   a. A competitive inhibitor binds the active site of the enzyme.
   b. An allosteric inhibitor binds a site on the active form of the enzyme.
   c. A noncompetitive inhibitor binds a site other than the active site.
   d. Noncompetitive inhibition cannot be completely overcome by the addition of more substrate.
   e. Competitive inhibition can be completely overcome by the addition of more substrate.
9. Which statement about feedback inhibition of enzymes is *not* true?
   a. It is exerted through allosteric effects.
   b. It is directed at the enzyme that catalyzes the first committed step in a branch of a pathway.
   c. It affects the rate of reaction, not the concentration of enzyme.
   d. It acts very slowly.
   e. It is an example of negative feedback.
10. Which statement about temperature effects is *not* true?
    a. Raising the temperature may reduce the activity of an enzyme.
    b. Raising the temperature may increase the activity of an enzyme.
    c. Raising the temperature may denature an enzyme.
    d. Some enzymes are stable at the boiling point of water.
    e. All enzymes have the same optimal temperature.

For Discussion

1. How is it possible for endergonic reactions to proceed in organisms?
2. Consider two proteins: One is an enzyme dissolved in the cytosol, the other is an ion channel in a membrane. Contrast the structures of the two proteins, indicating at least two important differences.
3. Plot free energy versus the course of an endergonic reaction and that of an exergonic reaction. Include the activation energy in both plots. Label E_a and ΔG on both graphs.
4. Consider an enzyme that is subject to allosteric regulation. If a competitive inhibitor (not an allosteric inhibitor) is added to a solution of such an enzyme, the ratio of enzyme molecules in the active form to those in the inactive form increases. Explain this observation.
5. If you were presented with a radioactively labeled substance, what experiments would you perform to determine whether it enters cells by simple diffusion or active transport?

6 Chromosomes, the Cell Cycle, and Cell Division



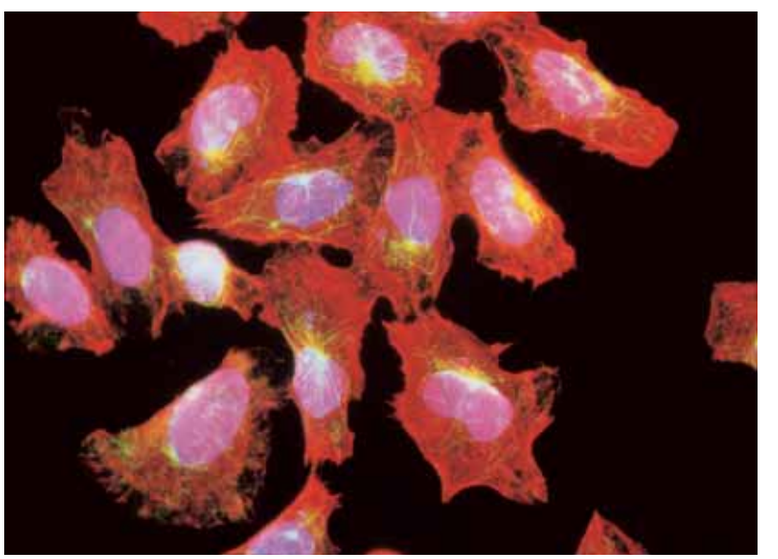
In 1951, 31-year-old Henrietta Lacks entered Johns Hopkins Hospital to be treated for a cancerous tumor. Although she died a few months later, her tumor cells are still alive today. Scientists found that, given adequate nourishment, cancerous cells from the tumor could reproduce themselves indefinitely in a laboratory dish, where they grew as a formless mass. These "HeLa cells" became a test-tube model for many studies of human biology. Over the past half-century, tens of thousands of research articles have been published using information obtained from Henrietta's cells. But are these "immortal" cells really a good model for human biology?

In one sense, they are indeed a good model. Most multicellular organisms come from a single cell: the fertilized egg. This cell reproduces itself to make two cells, which in turn divide to become four cells, and so on until all the cells of a new organism have been produced. An organism is not just a mass of cells like the HeLa culture, however; its cells must form specialized tissues and organs, each with specific roles to perform.

In normal tissues, cell reproduction (cell "births") is offset by cell loss (cell "deaths"). We know that cell death is important from careful studies of a tiny worm, *Caenorhabditis elegans*, in which 1,090 cells are produced from the fertilized egg and exactly 131 of them die before the worm is born. If the cells that are programmed to die do not do so, the worm's organs are severely malformed. Another example occurs in the mammalian brain. Young mice, for instance, lose hundreds of thousands of brain cells each day; if these cells do not die, the mouse's overcrowded brain simply does not work.

A cell's death is often programmed into its genetic instructions: normal cells "sacrifice" themselves for the greater good of the organism. Once an organism reaches its adult size, it stays that way through a combination of cell division and programmed cell death. Unlike most normal cells, but like most cancerous cells, HeLa cells keep growing because they have a genetic imbalance that heavily favors cell reproduction over cell death.

In this chapter, we will first describe how prokaryotic cells produce two new organisms from the original single-celled organism. Then we will describe two types of cell and nuclear division—mitosis and meiosis—and relate them to asexual and sexual reproduction in eukaryotic organisms. Finally, to



HeLa Cells: More Births Than Deaths
These tumor cells grow and reproduce as an unspecialized mass on the surface of a solid medium. They have been cultured in a laboratory since 1951. They are the source of much data relating to the reproduction of human cells.

balance our discussion of cell proliferation through division, we will describe the important process of programmed cell death, also known as apoptosis.

Systems of Cell Reproduction

Unicellular organisms use cell division primarily to reproduce themselves, whereas in multicellular organisms cell division also plays important roles in growth and in the repair of tissues (Figure 9.1). In order for any cell to divide, four events must occur:

- There must be a *reproductive signal*. This signal, which may come either from inside or outside the cell, initiates the cellular reproductive events.
- **Replication** of DNA (the genetic material) and other vital cell components must occur so that each of the two new cells will be identical and have complete cell functions.
- The cell must distribute the replicated DNA to each of the two new cells. This process is called **segregation**.
- New material must be added to the cell membrane (and the cell wall, in organisms that have one) in order to separate the two new cells in a process called **cytokinesis**.

These four events occur in somewhat different ways in prokaryotes and eukaryotes.

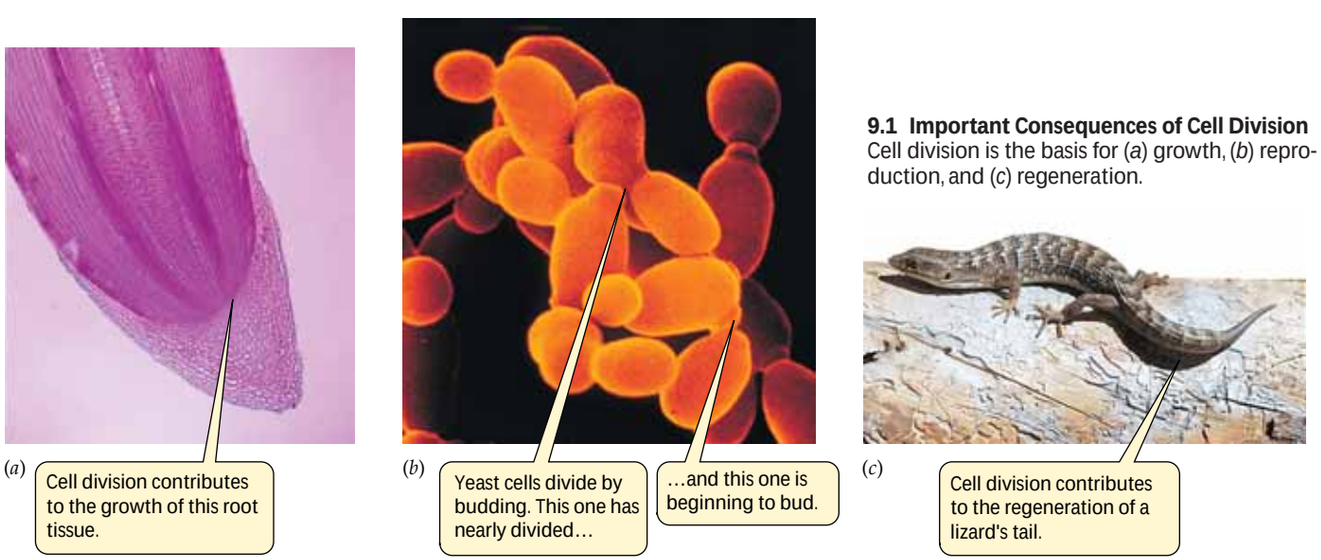
Prokaryotes divide by fission

In prokaryotes, cell division results in the reproduction of the entire single-celled organism. The cell grows in size, replicates its DNA, and then essentially divides into two new cells, a process called **fission**.

REPRODUCTIVE SIGNALS. The reproductive rates of many prokaryotes respond to conditions in the environment. The bacterium *Escherichia coli*, a species that is commonly used in genetic studies, is a "cell division machine" that essentially divides continuously. Typically, cell division takes 40 minutes at 37°C. But if there are abundant sources of carbohydrates and salts available, the division cycle speeds up so that cells may divide in 20 minutes. Another bacterium, *Bacillus subtilis*, stops dividing when food supplies are low, then resumes dividing when conditions improve. These observations suggest that external factors, such as materials in the environment, control the initiation of cell division in prokaryotes.

REPLICATION OF DNA. A **chromosome**, as we saw in Chapter 4, is a DNA molecule containing genetic information. When a cell divides, all of its chromosomes must be replicated, and each of the two resulting copies must find its way into one of the two new cells.

Most prokaryotes have only one chromosome, a single long DNA molecule with proteins bound to it. In the bacterium *E. coli*, the DNA is a continuous molecule often referred to as a *circular chromosome*. If the bacterial DNA were actually arranged in a circle, it would be about 1.6 million nm (1.6 mm) in circumference. The bacterium itself is only about 1 µm (1,000 nm) in diameter and about 4 µm long. Thus the bacterial DNA, fully extended, would form a circle over 100 times larger than the cell! To fit it into the cell, the DNA must be packaged. The DNA molecule accomplishes some packaging by folding in on itself, and positively charged (basic) proteins bound to negatively charged (acidic) DNA contribute to this folding. Circular chromosomes appear to be characteristic of all prokaryotes, as well as some viruses, and



9.1 Important Consequences of Cell Division
Cell division is the basis for (*a*) growth, (*b*) reproduction, and (*c*) regeneration.

are also found in the chloroplasts and mitochondria of eukaryotic cells.

Functionally, the prokaryotic chromosome has two regions that are important for cell reproduction:

- The site where replication of the circle starts: the origin of replication, designated *ori*
- The site where replication ends: the terminus of replication, *ter*

The process of chromosome replication occurs as the DNA is threaded through a "replication complex" of proteins at the center of the cell. These proteins include the enzyme DNA polymerase, and their operation will be discussed further in Chapter 11. During the process of prokaryotic DNA replication, the cell grows and provides a mechanism for the ordered distribution of the DNA into the newly formed daughter cells.

SEGREGATION OF DNA. DNA replication actively drives the segregation of the replicated DNA molecules to the two new cells. The first region to be replicated is *ori*, which is attached to the plasma membrane. The two resulting *ori* regions separate as the new chromosome forms and new plasma membrane forms between them as the cell grows longer (Figure 9.2). By the end of replication, there are two chromosomes, one at either end of the lengthened bacterial cell.

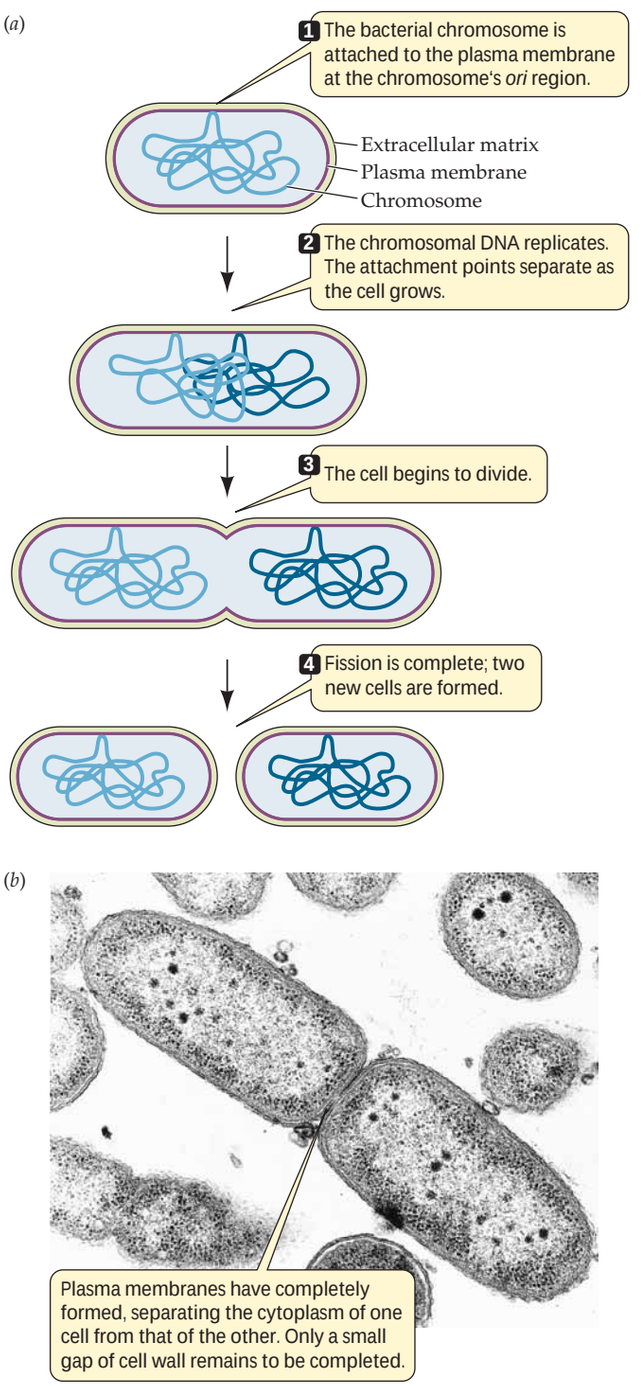
CYTOKINESIS. Cell separation, or cytokinesis, begins 20 minutes after chromosome replication is finished. The first event of cytokinesis is a pinching in of the plasma membrane to form a ring similar to a purse string. Fibers composed of a protein similar to eukaryotic tubulin (which makes up microtubules) are major components of this ring. As the membrane pinches in, new cell wall materials are synthesized, which finally separate the two cells.

Eukaryotic cells divide by mitosis or meiosis

Complex eukaryotes, such as humans and flowering plants, originate from a single cell, the fertilized egg. This cell derives from the union of two sex cells, called **gametes**, from the organism's parents—that is, a sperm and egg—and so contains genetic material from both of these parental cells. This means that the fertilized egg contains one set of chromosomes from the male parent and one set from the female parent.

The formation of a multicellular organism from a fertilized egg is called *development*. It involves both cell reproduction and cell specialization. For example, an adult human has several trillion cells, all ultimately deriving from the fertilized egg, and many of them have specialized roles. We will discuss how cells specialize later in this book, in Part Three. For now, we will focus on cell reproduction.

Cell reproduction in eukaryotes, like that in prokaryotes, involves reproductive signals, DNA replication, segregation, and cytokinesis. But, as you might expect, events in eukary-



9.2 Prokaryotic Cell Division (*a*) The steps of cell division in prokaryotes. (*b*) These two cells of the bacterium *Pseudomonas aeruginosa* have almost completed fission. Each cell contains a complete chromosome, visible as the nucleoid in the center of the cell.

otes are somewhat more complex. First, unlike prokaryotes, eukaryotic cells do not constantly divide whenever environmental conditions are adequate. In fact, eukaryotic cells that are part of a multicellular organism and have become specialized seldom divide. So the signals for cell division are related not to the environment of a single cell, but to the needs of the entire organism. Second, instead of a single chromosome, eukaryotes usually have many (humans have 46), so the processes of replication and segregation, while basically the same as in prokaryotes, are more intricate (see Table 9.1). Third, eukaryotic cells have a distinct nucleus, which has to be replicated and then divided into two new nuclei. Thus, in eukaryotes, cytokinesis is distinct from division of the genetic material. Finally, cytokinesis is different in plant cells (which have a cell wall) than in animal cells (which do not).

The key difference between prokaryotic and eukaryotic cell reproduction is that in the eukaryotes, newly replicated chromosomes remain associated with each other as **sister chromatids**, and a new mechanism, **mitosis**, is used to segregate them into the two new nuclei.

The reproduction of a eukaryotic cell typically consists of three steps:

- The replication of DNA within the nucleus
- The packaging and segregation of the replicated DNA into two new nuclei (nuclear division)
- The division of the cytoplasm (cytokinesis)

A second mechanism of nuclear division, **meiosis**, occurs in germ cells that produce gametes that contribute to the reproduction of a new organism. While the two products of mitosis are genetically identical to the cell that produced them—they both have the same DNA—the products of meiosis are not. As we will see later in the chapter, meiosis generates diversity by shuffling the genetic material, resulting in new gene combinations. It plays a key role in sexual life cycles.

What determines whether a cell will divide? How does mitosis lead to identical cells, and meiosis to diversity? Why do we need both identical copies and diverse cells? Why do most eukaryotic organisms reproduce sexually? In the pages that follow, we will describe the details of interphase, mitosis, and meiosis, as well as their consequences for heredity, development, and evolution.

Interphase and the Control of Cell Division

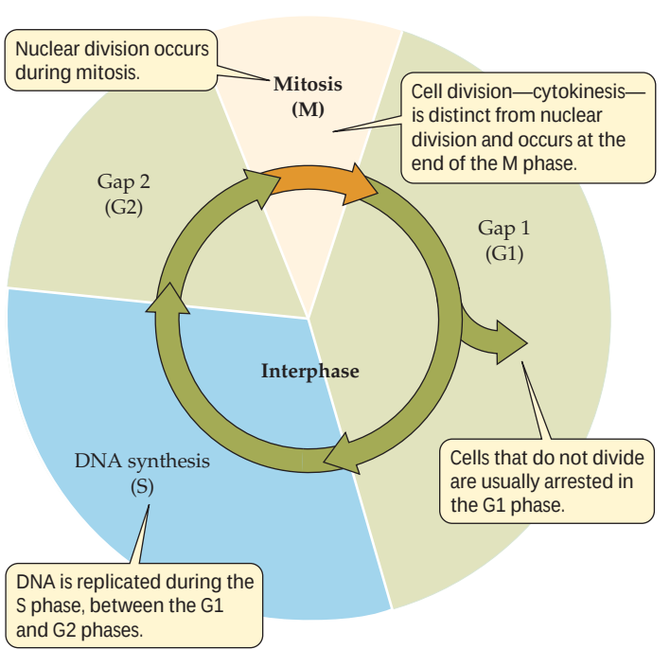
A cell lives and functions until it divides or dies. Or, if it is a gamete, it lives until it fuses with another gamete. Some types of cells, such as red blood cells, muscle cells, and nerve cells, lose the capacity to divide as they mature. Other cell types, such as cortical cells in plant stems, divide only rarely. Some cells, like the cells in a developing embryo, are specialized for rapid division.

Between divisions—that is, for most of its life—a eukaryotic cell is in a condition called **interphase**. For most types of cells, we may speak of a **cell cycle** that has two phases: mitosis and interphase. In this section, we will describe the cell cycle events that occur during interphase, especially the "decision" to enter mitosis.

A given cell lives for one turn of the cell cycle and then becomes two cells. The cell cycle, when repeated again and again, is a constant source of new cells. However, even in tissues engaged in rapid growth, cells spend most of their time in interphase. Examination of any collection of dividing cells, such as the tip of a root or a slice of liver, will reveal that most of the cells are in interphase most of the time; only a small percentage of the cells will be in mitosis at any given moment.

Interphase consists of three subphases, identified as G1, S, and G2. The cell's DNA replicates during the **S phase** (the S stands for synthesis). The period between the end of mitosis and the onset of the S phase is called **G1**, or Gap 1. Another gap phase—**G2**—separates the end of the S phase and the beginning of mitosis, when nuclear and cytoplasmic division take place and two new cells are formed. Mitosis and cytokinesis are referred to as the **M phase** of the cell cycle (Figure 9.3).

The process of DNA replication, which we will describe in Chapter 11, is completed by the end of S phase. Where there was formerly one chromosome, there are now two, joined



9.3 The Eukaryotic Cell Cycle The cell cycle consists of a mitotic (M) phase, during which first nuclear division (mitosis) and then cell division (cytokinesis) take place. The M phase is followed by a long period of growth known as interphase. Interphase has three subphases (G1, S, and G2) in cells that divide.

together and awaiting segregation into two new cells by mitosis or meiosis.

Although one key event—DNA replication—dominates and defines the S phase, important cell cycle processes take place in the gap phases as well. G1 is quite variable in length in different cell types. Some rapidly dividing embryonic cells dispense with it entirely, while other cells may remain in G1 for weeks or even years. In many cases, these cells enter a resting phase called **G0**. Special internal and external signals are needed to prompt a cell to leave G0 and re-enter the cell cycle at G1.

The biochemical hallmark of a G1 cell is that it is preparing for the S phase, so at this stage each chromosome is a single, unreplicated structure. It is at the G1-to-S transition that the commitment to enter another cell cycle is made.

During G2, the cell makes preparations for mitosis—for example, by synthesizing components of the microtubules that will move the chromosomes to opposite ends of the dividing cell. Because the chromosomes were replicated during the S phase, each chromosome now consists of two identical sister chromatids.

Cyclins and other proteins signal events in the cell cycle

How are appropriate decisions to enter the S or M phases made? These transitions—from G1 to S and from G2 to M—depend on the activation of a type of protein called **cyclin-dependent kinase**, or **Cdk**. Remember that a *kinase* is an enzyme that catalyzes the transfer of a phosphate group from ATP to another molecule; this phosphate transfer is called *phosphorylation*:

$$\text{protein} + \text{ATP} \xrightarrow{\text{kinase}} \text{protein}\text{—P} + \text{ADP}$$

What does phosphorylation do to a protein? Recall from Chapter 3 that proteins have both hydrophilic regions (which tend to interact with water on the outside of the macromolecule) and hydrophobic regions (which tend to interact with one another on the inside of the macromolecule). These regions are important in giving a protein its three-dimensional shape. Phosphate groups are charged, so an amino acid with such a group tends to be on the outside of the protein. In this way, phosphorylation changes the shape and function of a protein. By catalyzing the phosphorylation of certain target proteins, Cdk's play important roles in initiating the steps of the cell cycle.

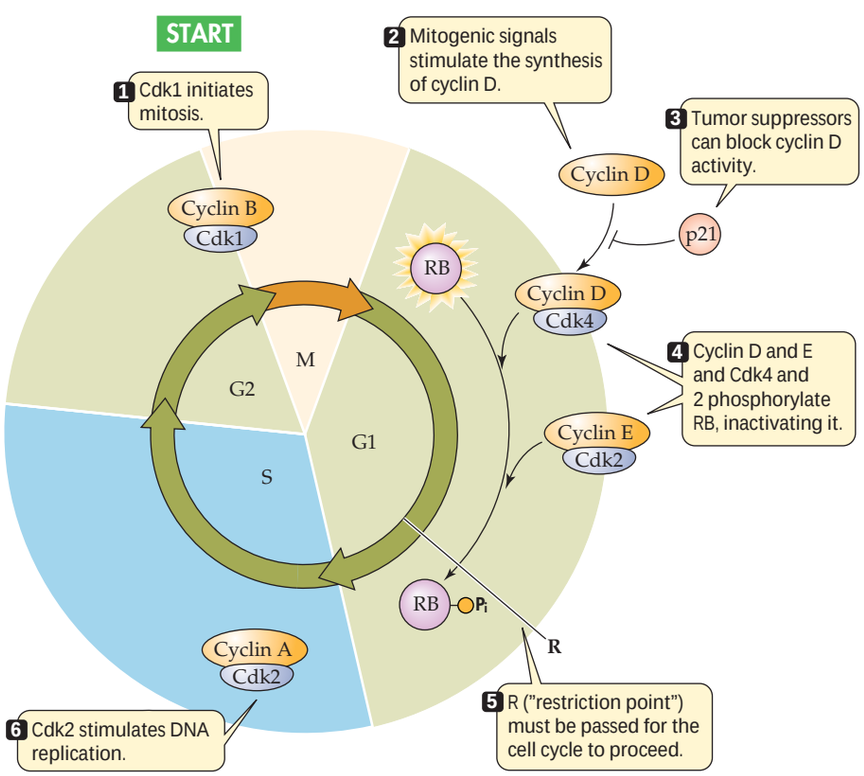
The discovery that Cdk's induce cell division is a beautiful example of research on different organisms and different cell types converging on a single mechanism:

- One group of scientists was studying immature sea urchin eggs, trying to find out how they are stimulated to divide and form mature eggs. A protein called *maturation promoting factor* was purified from maturing eggs, which by itself prodded immature eggs into division.
- Other scientists studying the cell cycle in yeast, a single-celled eukaryote, found a strain that was stalled at the G1–S boundary because it lacked a Cdk. This yeast Cdk was discovered to be very similar to the sea urchin's maturation promoting factor.

Similar Cdk's were soon found to control the G1-to-S transition in many other organisms, including humans.

But Cdk's are not active by themselves. Rather, they must bind to a second type of protein, called **cyclin**. This binding—an example of allosteric regulation—activates the Cdk by altering its shape and exposing its active site (see Figure 6.19). It is the cyclin-Cdk complex that acts as a protein kinase and triggers the transition from G1 to S phase. Then cyclin breaks down, and the Cdk becomes inactive.

Several different cyclin-Cdk combinations act at various stages of the mammalian cell cycle (Figure 9.4):



9.4 Cyclin-Dependent Kinases and Cyclins Trigger Transitions in the Cell Cycle There are four cyclin-Cdk controls during the typical cell cycle in humans.

- Cyclin D-Cdk4 acts during the middle of G1. This is the *restriction point* (*R*), a key decision point beyond which the rest of the cell cycle is normally inevitable.
- Cyclin E-Cdk2 also acts in the middle of G1.
- Cyclin A-Cdk2 acts during S, and also stimulates DNA replication.
- Cyclin B-Cdk1 acts at the G2–M boundary, initiating the transition to mitosis.

The key to progress past the restriction point is a protein called *RB* (retinoblastoma protein, named for a childhood cancer in which it was first discovered). RB normally inhibits the cell cycle. But when RB is phosphorylated by a protein kinase, it becomes inactive and no longer blocks the restriction point, and the cell progresses past G1 into S phase (Note the double negative here—a cell function happens because an inhibitor is inhibited! This phenomenon is rather common in the control of cellular metabolism.) The enzymes that catalyze RB phosphorylation are Cdk4 and Cdk2. So what is needed for a cell to pass the restriction point is the synthesis of cyclins D and E, which activate Cdk 4 and 2, which phosphorylate RB, which becomes inactivated.

The cyclin-Cdk complexes act as *checkpoints*, points at which a cell cycle's progress can be monitored to determine whether the next step can be taken. For example, if DNA is damaged by radiation during G1, a protein called p21 is made. (The p stands for "protein" and the 21 stands for its molecular weight—about 21,000 daltons.) The p21 protein then binds to the two G1 Cdk's, preventing their activation by cyclins. So the cell cycle stops while repairs are made to DNA. The p21 protein itself breaks down after the DNA is repaired, allowing cyclins to bind to the Cdk's and the cell cycle to proceed.

Because cancer results from inappropriate cell division, it is not surprising that these cyclin-Cdk controls are disrupted in cancer cells. For example, some fast-growing breast cancers have too much cyclin D, which overstimulates Cdk4 and thus cell division. As we will describe in Chapter 17, a major protein in normal cells that prevents them from dividing is p53, which leads to the synthesis of p21 and therefore inhibition of Cdk's. More than half of all human cancers contain defective p53, resulting in the absence of cell cycle controls.

Growth factors can stimulate cells to divide

Cyclin-Cdk complexes provide an internal control on progress through the cell cycle. But there are tissues in the body in which cells no longer go through the cell cycle, or go through it slowly and divide infrequently. If such cells are to divide, they must be stimulated by external signals (chemical messengers) called **growth factors**. For example, when you cut yourself and bleed, specialized cell fragments called platelets gather at the wound and help to initiate blood clotting. The platelets also produce and release a protein, called *platelet-derived growth factor*, that diffuses to the adjacent cells in the skin and stimulates them to divide and heal the wound.

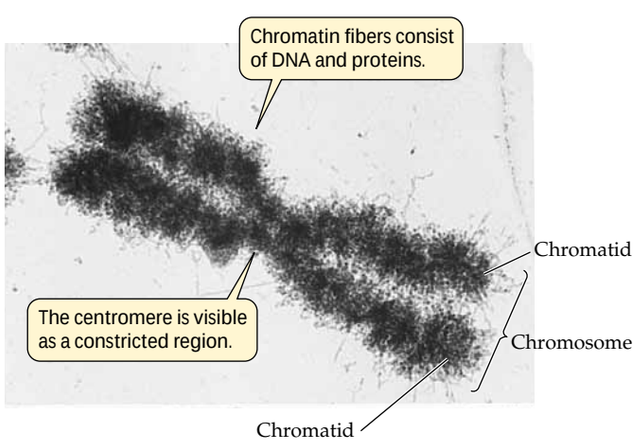
Other growth factors include *interleukins*, which are made by one type of white blood cell and promote cell division in other cells that are essential for the body's immune system defenses. *Erythropoietin*, made by the kidney, stimulates the division of bone marrow cells and the production of red blood cells. In addition, many hormones promote division in specific cell types.

We will describe the physiological roles of growth factors in later chapters, but all of them act in a similar way. They bind to their target cells via specialized receptor proteins on the target cell surface. This specific binding triggers events within the target cell that initiate the cell cycle. Cancer cells often divide inappropriately because they make their own growth factors, or because they no longer require growth factors to start cycling.

Eukaryotic Chromosomes

Most human cells other than gametes contain two full sets of genetic information, one from the mother and the other from the father. As in prokaryotes, this genetic information consists of DNA molecules. However, unlike prokaryotes, humans and other eukaryotes have more than one chromosome, and during interphase those chromosomes reside within a membrane-enclosed organelle, the nucleus.

The basic unit of the eukaryotic chromosome is a gigantic, linear, double-stranded molecule of DNA complexed with many proteins to form a dense material called **chromatin** (Figure 9.5). Before the S phase, each chromosome contains only one such double-stranded DNA molecule. However,



9.5 Chromosomes, Chromatids, and Chromatin A human chromosome, shown as the cell prepares to divide.

after the DNA molecule replicates during the S phase, the two resulting DNA molecules, now called chromatids, are held together along most of their length by a protein called **cohesin**. They stay this way until mitosis, when most of the cohesin is removed, except in a region called the **centromere** at which the chromatids are still held together (see Figure 9.9). A second group of proteins called **condensins** coats the DNA molecules at this time and makes them more compact.

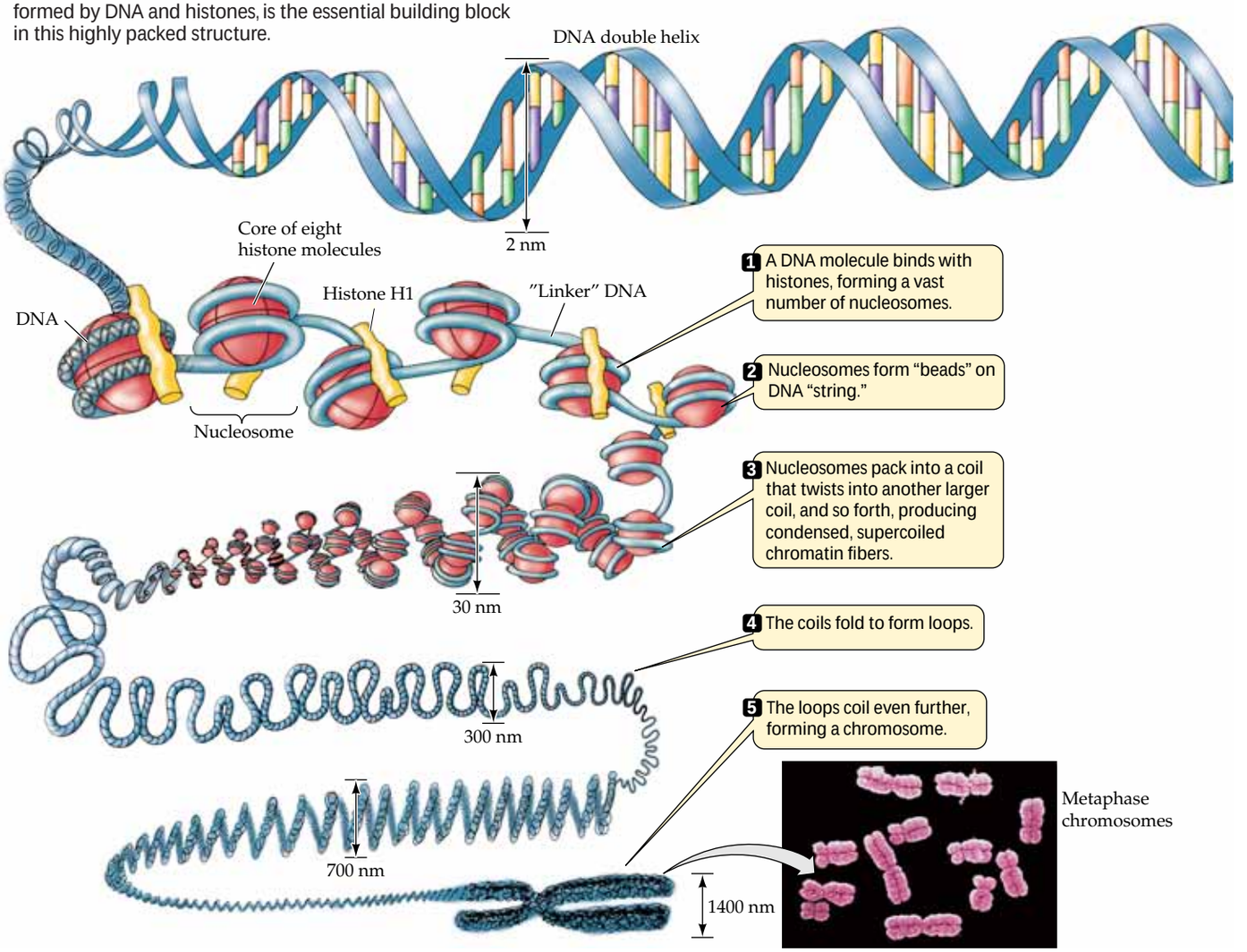
The DNA in a typical human cell has a total length of 2 meters. Yet the nucleus is only 5 µm (0.000005 meters) in diameter. So, although the DNA in an interphase nucleus is "unwound," it is still impressively packed! This packing is achieved largely by proteins associated closely with the chromosomal DNA (Figure 9.6).

Chromosomes contain large quantities of proteins called **histones** (from the Greek, "web"). There are five classes of histones. All of them have a positive charge at cellular pH levels because of their high content of the basic amino acids lysine and arginine. These positive charges electrostatically attract the negative phosphate groups on DNA. These interactions, as well as interactions among the histones themselves, form beadlike units called **nucleosomes**. Each nucleosome contains the following components:

- Eight histone molecules, two each of four of the histone classes, united to form a core or spool.
- 146 base pairs of DNA, 1.65 turns of it wound around the histone core.
- Histone H1 (the remaining histone class) on the outside of the DNA, which may clamp it to the histone core.

During interphase, a chromosome is made up of a single DNA molecule running around vast numbers of nucleosomes like beads on a string. Between the nucleosomes stretches a variable amount of non-nucleosomal "linker"

9.6 DNA Packs into a Mitotic Chromosome The nucleosome, formed by DNA and histones, is the essential building block in this highly packed structure.



DNA. Since this DNA is exposed to the nuclear environment, it is accessible to proteins involved in its duplication and the regulation of its expression, as we will see in Chapter 14.

During both mitosis and meiosis, the chromatin becomes ever more coiled and condensed as its nucleosomes pack together and coil, with further folding of the chromatin continuing up to the time at which the chromosomes begin to move apart.

Mitosis: Distributing Exact Copies of Genetic Information

In mitosis, a single nucleus gives rise to two nuclei that are genetically identical to each other and to the parent nucleus. This process ensures the accurate distribution of the eukaryotic cell's multiple chromosomes to the daughter nuclei. In reality, mitosis is a continuous process in which each event flows smoothly into the next. For discussion, however, it is convenient to look at mitosis—the M phase of the cell cycle—as a series of separate events: prophase, prometaphase, metaphase, anaphase, and telophase.

The centrosomes determine the plane of cell division

Once the commitment to enter mitosis has been made, the cell enters S phase, and DNA is replicated. At the same time, in the cytoplasm, the **centrosome** ("central body"), an organelle that lies near the nucleus, doubles, forming a pair of centrosomes. In many organisms, each centrosome consists of a pair of **centrioles**, each one a hollow tube lined with nine microtubules. The two tubes are at right angles to each other.

At the G2-to-M transition, the two centrosomes separate from each other, moving to opposite ends of the nuclear envelope. The orientation of the centrosomes determines the plane at which the cell will divide, and therefore the spatial relationship of the two new cells to the parent cell. This relationship may be of little consequence to single free-living cells such as yeasts, but it is important for cells that make up part of a body tissue.

The material around the centrioles initiates the formation of microtubules, which will orchestrate chromosomal movement. Plant cells lack centrosomes, but distinct microtubule organizing centers at either end of the cell serve the same role. The formation of microtubules leads to the formation of the spindle structure that is required for the orderly segregation of the chromosomes in cell division.

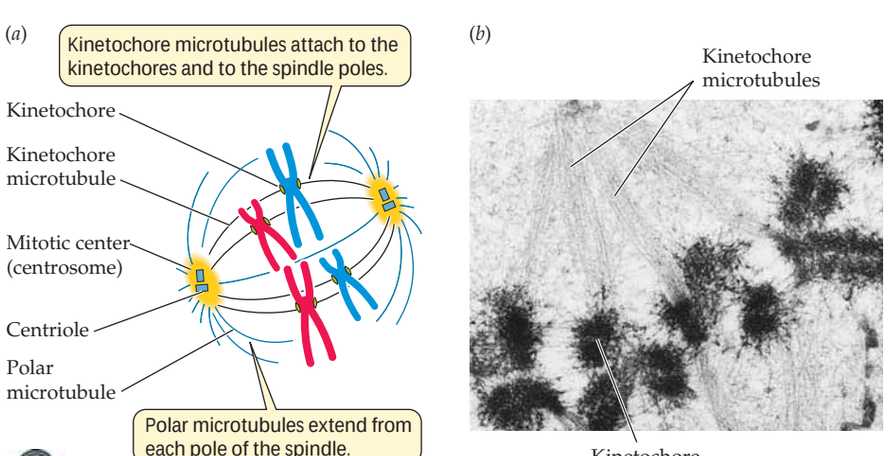
Chromatids become visible and the spindle forms during prophase

During interphase, only the nuclear envelope, the nucleoli, and a barely discernible tangle of chromatin are visible under the light microscope. The appearance of the nucleus changes as the cell enters **prophase**—the beginning of mitosis. Most of the cohesin that held the two products of DNA replication together since S phase has been destroyed, so the individual chromatids become visible. They are still held together by a small amount of cohesin at the centromere (see Figure 9.9). Late in prophase, specialized three-layered structures called **kinetochores** develop in the centromere region, one on each chromatid. These structures will be important in chromosome movements.

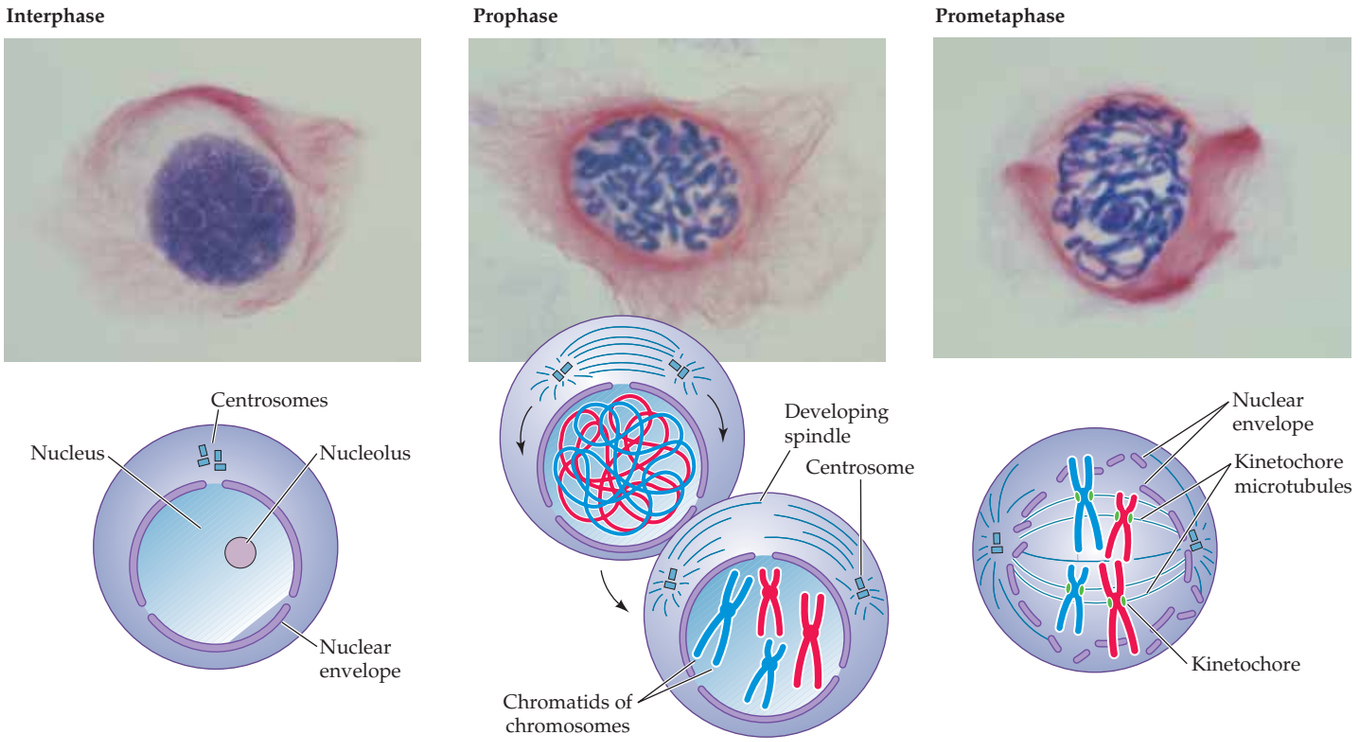
Each of the two centrosomes serves as a *mitotic center*, or *pole*, toward which the chromosomes will move. Microtubules form between each pole and the chromosomes to make up a **spindle**, which serves both as a structure to which the chromosomes will attach and as a framework keeping the two poles apart. The spindle is actually two half-spindles: Each polar microtubule runs from one pole to the middle of the spindle, where it overlaps with polar microtubules extending from the other half-spindle (Figure 9.7). The polar microtubules are initially unstable, constantly forming and falling apart, until they contact polar microtubules from the other half-spindle and become more stable.

There are two types of microtubules in the spindle:

- *Polar microtubules* have abundant tubulin around the centrioles. Tubulin subunits can aggregate to form long fibers that extend beyond the equatorial plate.
- *Kinetochore microtubules* attach to the kinetochores on the chromosomes.



9.7 The Mitotic Spindle Consists of Microtubules (*a*) Diagram of the spindle apparatus in a cell at metaphase. (*b*) An electron micrograph of the stage shown in (*a*).



9.8 Mitosis Mitosis results in two new nuclei that are genetically identical to each other and to the nucleus from which they were formed. The photomicrographs are of plant nuclei, which lack centrioles. The diagrams are of corresponding phases in animal cells and introduce the structures not found in plants. In the micrographs, the red dye stains microtubules (and thus the spindle); the blue dye stains the chromosomes. In the diagrams, the chromosomes are stylized to emphasize the fates of the individual chromatids.

The two sister chromatids in each chromosome pair become attached by their kinetochore to microtubules in opposite halves of the spindle. This ensures that one chromatid of the pair will eventually move to one pole and the other chromatid will move to the other pole. Movement of the chromatids is the central feature and accomplishment of mitosis.

Chromosome movements are highly organized

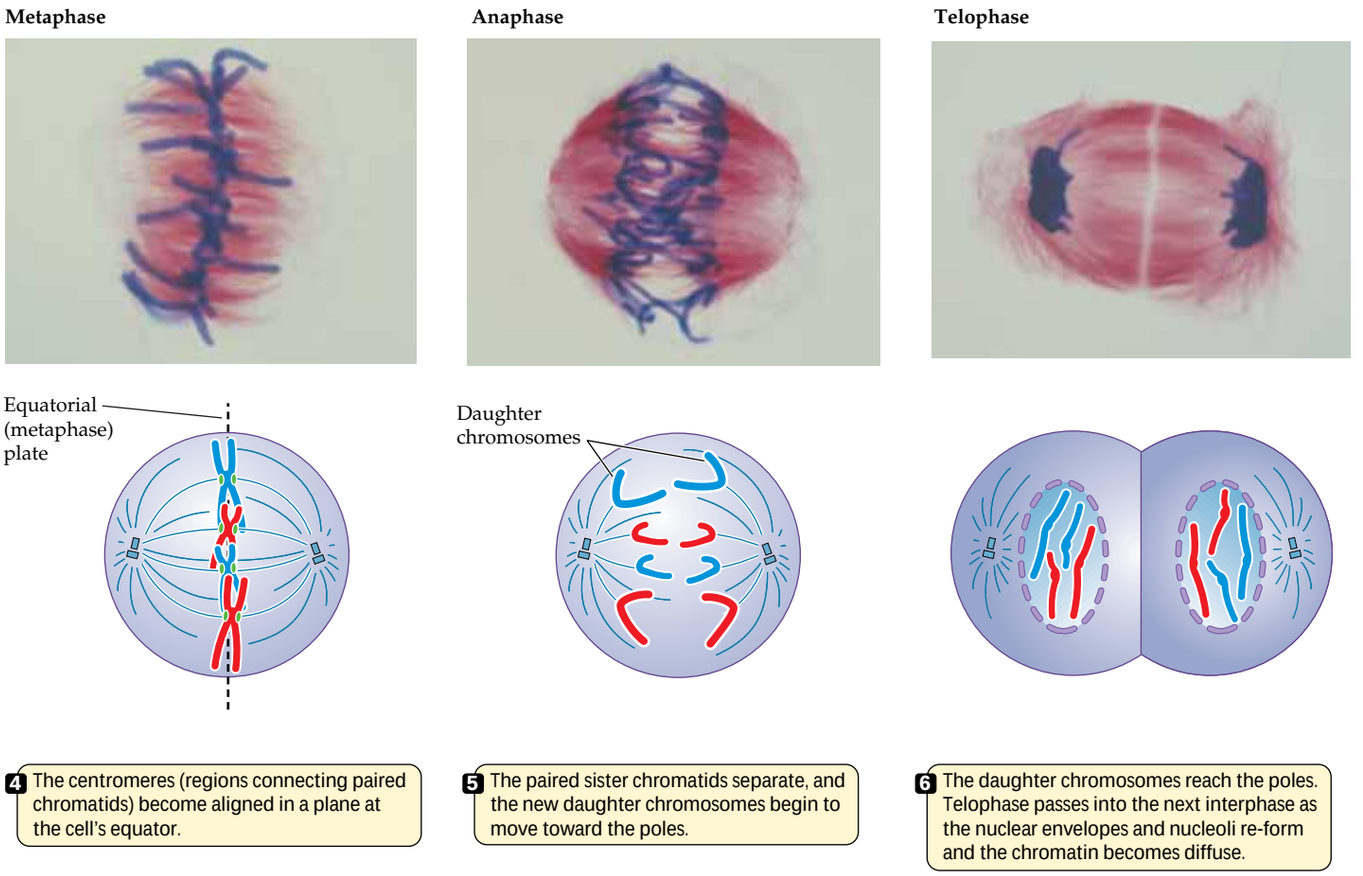
The next three phases of mitosis—prometaphase, metaphase, and anaphase—are the phases during which chromosomes actually move (Figure 9.8). During these phases, the centromeres holding the two chromatids together separate, and the former sister chromatids move away from each other in opposite directions.

PROMETAPHASE. **Prometaphase** is marked by the disappearance of the nuclear envelope. The material of the envelope remains in the cytoplasm, however, to be reassembled when the daughter nuclei re-form. In prometaphase, the chromosomes begin to move toward the poles, but this movement is counteracted by two factors:

- A repulsive force from the poles pushes the chromosomes toward the middle region, or **equatorial plate** (metaphase plate), of the cell.
- The two chromatids are still held together at the centromere by cohesin.

Thus, during prometaphase chromosomes appear to move aimlessly back and forth between the poles and the middle of the spindle. Gradually, the centromeres approach the equatorial plate.

METAPHASE. The cell is said to be in **metaphase** when all the centromeres arrive at the equatorial plate. Metaphase is the best time to see the sizes and shapes of chromosomes

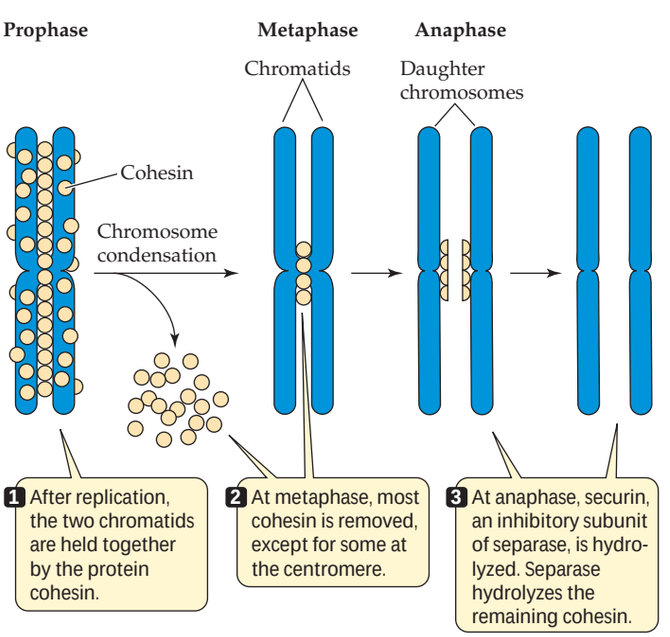


because they are maximally condensed. The chromatids are now clearly connected to one pole or the other by microtubules. At the end of metaphase, all of the chromatid pairs separate simultaneously.

This separation occurs because the cohesin holding the sister chromatids together is hydrolyzed by a specific protease, appropriately called *separase*. Until this point, separase has been present but inactive, because it has been bound to an inhibitory subunit called *securin*. Once all the chromatids are connected to the spindle, securin is hydrolyzed, allowing separase to catalyze cohesin breakdown (Figure 9.9). In this way, chromosome alignment is connected to chromatid separation. This process, called the *spindle checkpoint*, apparently senses whether there are any kinetochores that are unattached to the spindle. If there are, securin breakdown is blocked, and the sister chromatids stay together.

ANAPHASE. Separation of the chromatids marks the beginning of **anaphase**, during which the two sister chromatids move to opposite ends of the spindle. Each chromatid contains one double-stranded DNA molecule and is now referred to as a **daughter chromosome**.

What propels this highly organized mass migration is not clear. Two things seem to move the chromosomes along.



9.9 Molecular Biology of Chromatid Attachment and Separation
Cohesin holds sister chromatids together. Separase hydrolyzes cohesin at the onset of anaphase.

First, at the kinetochores are proteins that act as "molecular motors." These proteins, called *cytoplasmic dynein*, have the ability to hydrolyze ATP to ADP and phosphate, thus releasing energy to move the chromosomes along the microtubules toward the poles. These motor proteins account for about 75 percent of the force of motion. Second, the kinetochore microtubules shorten from the poles, drawing the chromosomes toward them. This shortening accounts for about 25 percent of the motion.

During anaphase the poles of the spindle are pushed farther apart, doubling the distance between them. The distance between poles increases because the overlapping polar microtubules extending from opposite ends of the spindle contain motor proteins that cause them to slide past each other, in much the same way that microtubules slide in cilia and flagella (see Figure 4.24*a*). This polar separation further separates one set of daughter chromosomes from the other.

The movements of chromosomes are slow, even in cellular terms. At about 1 µm per minute, it takes about 10–60 minutes for them to complete their journey to the poles. This speed is equivalent to a person taking 9 million years to travel across the United States! This slow speed may ensure that the chromosomes segregate accurately.

Nuclei re-form during telophase

When the chromosomes stop moving at the end of anaphase, the cell enters **telophase**. Two sets of chromosomes (formerly referred to as daughter chromosomes), containing identical DNA and carrying identical sets of hereditary instructions, are now at the opposite ends of the spindle, which begins to break down. The chromosomes begin to uncoil, continuing until they become the diffuse tangle of chromatin that is characteristic of interphase. The nuclear envelopes and nucleoli, which were disaggregated during prophase, coalesce and reform their respective structures. When these and other changes are complete, telophase—and mitosis—is at an end, and each of the daughter nuclei enters another interphase.

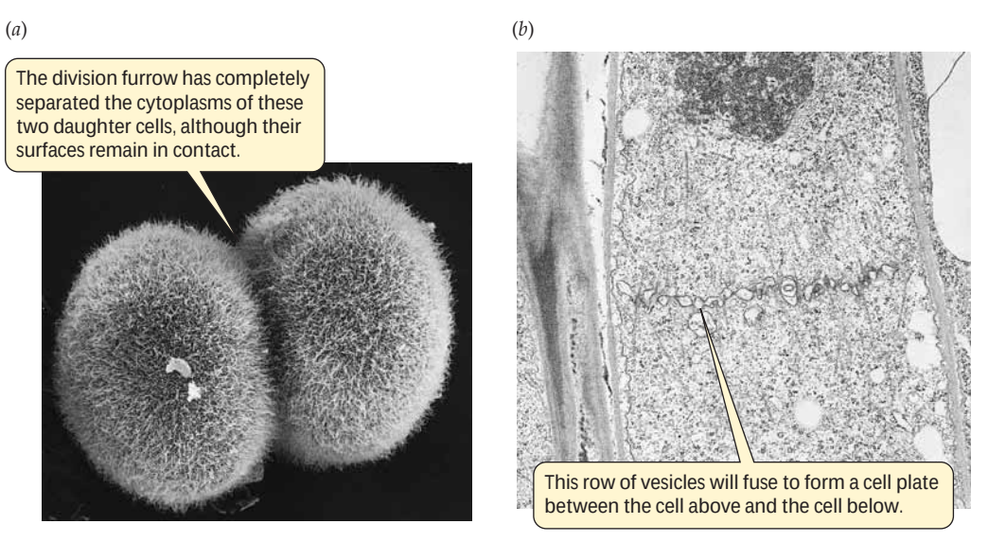
Mitosis is beautifully precise. Its result is two nuclei that are identical to each other and to the parent nucleus in chromosomal makeup, and hence in genetic constitution. Next, the two nuclei must be isolated in separate cells, which requires the division of the cytoplasm.

Cytokinesis: The Division of the Cytoplasm

Mitosis refers only to the division of the nucleus. The division of the cell's cytoplasm, which follows mitosis, is accomplished by cytokinesis, which may actually begin before telophase ends. In different organisms, cytokinesis may be accomplished in different ways. The differences between the process in plants and in animals are substantial.

Animal cells usually divide by a furrowing of the plasma membrane, as if an invisible thread were tightening between the two poles (Figure 9.10*a*). The invisible thread is actually microfilaments of actin and myosin (see Figure 4.21) located in a ring just beneath the plasma membrane. These two proteins interact to produce a contraction, just as they do in muscles, thus pinching the cell in two. These microfilaments assemble rapidly from actin monomers that are present in the interphase cytoskeleton. Their assembly appears to be under the control of Ca^{2+} released from storage sites in the center of the cell.

Plant cell cytoplasm divides differently because plants have cell walls. As the spindle breaks down after mitosis, membranous vesicles derived from the Golgi apparatus appear in the equatorial region roughly midway between the two daughter nuclei. Propelled along microtubules by the



9.10 Cytokinesis Differs in Animal and Plant Cells Plant cells must divide differently from animal cells because they have cell walls. (*a*) A sea urchin egg that has just completed cytokinesis at the end of the first cell division of its development into an embryo. (*b*) A dividing plant cell in late telophase.

motor protein kinesin, these vesicles fuse to form new plasma membrane and contribute their contents to a *cell plate*, which is the beginning of a new cell wall (Figure 9.10*b*).

Following cytokinesis, both daughter cells contain all the components of a complete cell. A precise distribution of chromosomes is ensured by mitosis. Organelles such as ribosomes, mitochondria, and chloroplasts need not be distributed equally between daughter cells as long as some of each are present in both cells; accordingly, there is no mechanism with a precision comparable to that of mitosis to provide for their equal allocation to daughter cells. As we will see in Part Three of this book, during development the unequal distribution of cytoplasmic components can have functional significance for the two new cells.

Reproduction: Asexual and Sexual

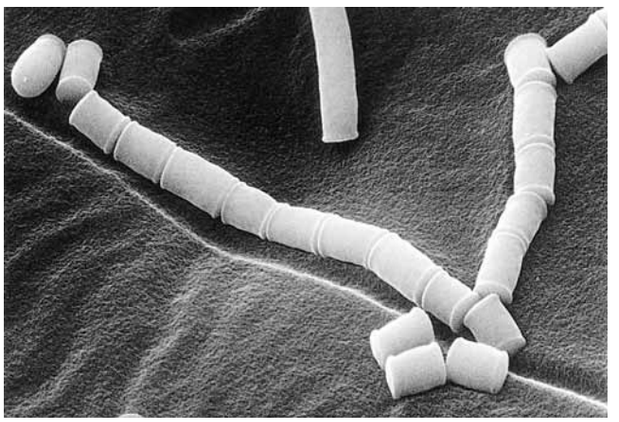
The mitotic cell cycle repeats itself over and over. By this process, a single cell can give rise to a vast number of other cells. Meiosis, on the other hand, results in only four progeny cells, which usually do not undergo further duplications. These two methods of nuclear and cell division are both involved in reproduction, but they have different reproductive roles.

Reproduction by mitosis results in genetic constancy

A cell undergoing mitosis may be an entire single-celled organism reproducing itself with each cell cycle. Alternatively, it may be a cell produced by a multicellular organism that divides further to produce a new multicellular organism. Some multicellular organisms can reproduce themselves by releasing cells derived from mitosis and cytokinesis or by having a multicellular piece break away and grow on its own (Figure 9.11).

Asexual reproduction, sometimes called *vegetative reproduction*, is based on mitotic division of the nucleus. Accordingly, it produces a *clone* of offspring that are genetically identical to the parent. If there is any variation among the offspring, it is likely to be due to *mutations*, or changes, in the genetic material. Asexual reproduction is a rapid and effective means of making new individuals, and it is common in nature.

Sexual reproduction, which involves meiosis, is very different. In sexual reproduction, two parents each contribute one cell, a gamete, to their offspring. This method produces offspring that differ genetically from each parent as well as from one another. This genetic variation among the offspring means that some of them may be better adapted than others to survive and reproduce in a particular environment. Thus this genetic diversity provides the raw material for natural selection and evolution.



9.11 Asexual Reproduction These spool-shaped cells are asexual spores formed by a fungus. Each spore contains a nucleus produced by a mitotic division. A spore is the same genetically as the parent that fragmented to produce it.

Reproduction by meiosis results in genetic diversity

Sexual reproduction, which combines genetic information from two different cells, generates genetic diversity. All sexual life cycles have certain hallmarks:

- There are two parents, each of which provides chromosomes to the offspring in the form of a gamete produced by meiosis.
- Each gamete contains a single set of chromosomes.
- The two gametes—often identifiable as a female egg and a male sperm—fuse to produce a single cell, the **zygote**, or fertilized egg. The zygote thus contains two sets of chromosomes.

In multicellular organisms, **somatic cells**—those body cells that are *not* specialized for reproduction—each contain two sets of chromosomes, which are found in pairs (see Figure 9.13). One chromosome of each pair comes from each of the organism's two parents. The members of such a **homologous pair** are similar in size and appearance (except for the sex chromosomes found in some species, as we will see in Chapter 10). The two chromosomes (the homologs) of a homologous pair bear corresponding, though generally not identical, genetic information.

Gametes, on the other hand, contain only a single set of chromosomes—that is, one homolog from each pair. The number of chromosomes in such a cell is denoted by n, and the cell is said to be **haploid**. Two haploid gametes fuse to form a new organism in a process called **fertilization**. The resulting zygote thus has two sets of chromosomes, just as somatic cells do. Its chromosome number is denoted by $2n$, and the zygote is said to be **diploid**.

As you can see in Figure 9.12, different kinds of sexual life cycles exhibit different patterns of development after zygote formation:

- In *haplontic* organisms, such as protists and many fungi, the tiny zygote is the only diploid cell in the life cycle; the mature organism is haploid. The zygote undergoes meiosis to produce haploid cells, or **spores**. These spores form the new organism, which may be single-celled or multicellular, by mitosis. The mature haploid organism produces gametes by mitosis, which fuse to form the diploid zygote.
- Most plants and some protists have *alternation of generations*. Here, meiosis does not give rise to gametes but to haploid spores. The spores divide by mitosis to form an alternate, haploid life stage (the *gametophyte*). It is this haploid stage that forms gametes by mitosis. The gametes fuse to form a diploid zygote, which divides by mitosis to become the diploid *sporophyte*.
- In *diplontic* organisms, including animals and some plants, the gametes are the only haploid cells in the life cycle, and the mature organism is diploid. Gametes are formed by meiosis, which fuse to form a diploid zygote. The organism is formed by mitosis of diploid cells.

We will describe all these life cycles in greater detail in Part Five. In this chapter, our focus is on the role of sexual reproduction in generating diversity among individual organisms.

The essence of sexual reproduction is the random selection of half of a parent's diploid chromosome set to make a haploid gamete, followed by the fusion of two such haploid gametes to produce a diploid cell that contains genetic information from both gametes. Both of these steps contribute to a shuffling of genetic information in the population, so that no two individuals have exactly the same genetic constitution. The diversity provided by sexual reproduction opens up enormous opportunities for evolution.

The number, shapes, and sizes of the metaphase chromosomes constitute the karyotype

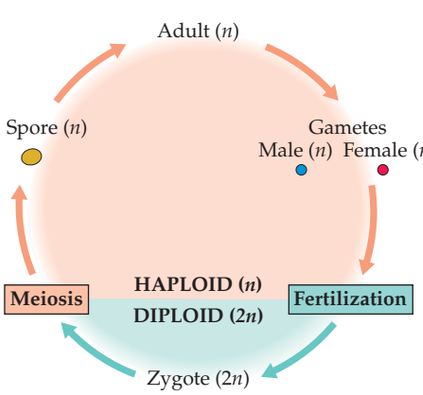
When nuclei are in metaphase of mitosis, it is often possible to count and characterize the individual chromosomes. This is a relatively simple process in some organisms, thanks to

9.12 Fertilization and Meiosis Alternate in Sexual Reproduction In sexual reproduction, haploid (*n*; peach) cells or organisms alternate with diploid (2*n*; green) cells or organisms.

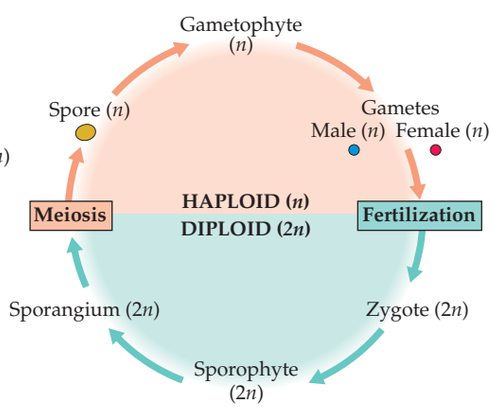
Fungus (*Rhizopus oligosporus*)
(Haploid)

Fern (*Osmunda cinnamomcea*)
(Diploid sporophyte)

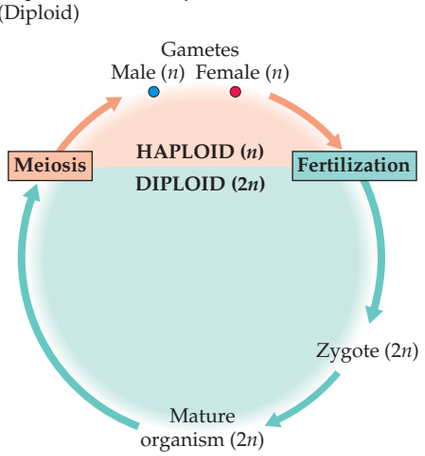
Elephant (*Loxodonta africana*)
(Diploid)



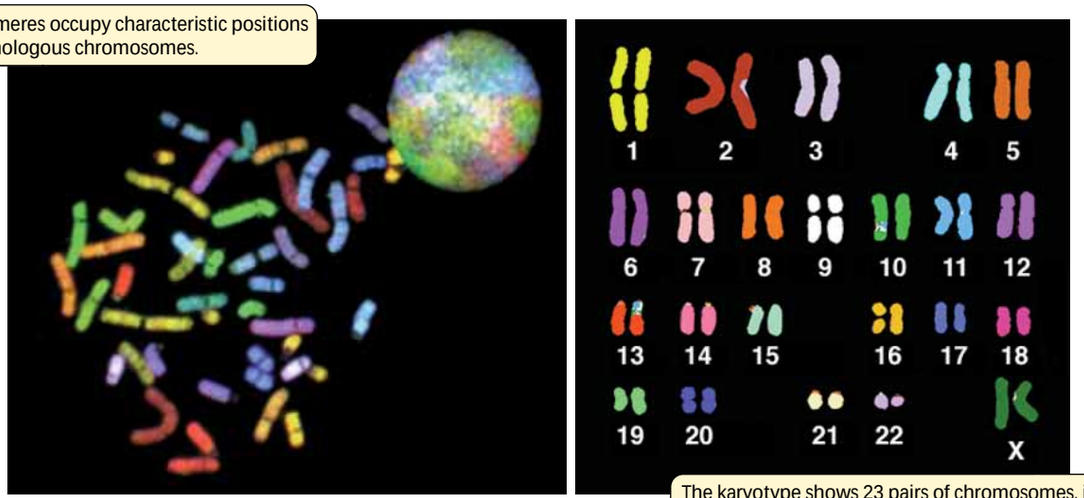
In the **haplontic life cycle**, the organism is haploid and the zygote is the only diploid stage.



In **alternation of generations**, the organism passes through both haploid and diploid stages.



In the **diplontic life cycle**, the organism is diploid and the gametes are the only haploid stage.



9.13 Human Cells Have 46 Chromosomes Chromosomes from a human cell are shown in metaphase of mitosis. The DNA of each chromosome has a specific nucleotide sequence that is stained by a specific colored dye, so that each homologous pair shares a distinctive color. Each chromosome at this stage is composed of two chromatids, but these cannot be distinguished by this "chromosome painting" technique. The multicolored globe is an interphase nucleus. The karyotype on the right is produced by computerized analysis of the image on the left.

techniques that can capture cells in metaphase and spread out the chromosomes. A photograph of the entire set of chromosomes can then be made, and the images of the individual chromosomes can be placed in an orderly arrangement. Such a rearranged photograph reveals the number, shapes, and sizes of chromosomes in a cell, which together constitute its **karyotype** (Figure 9.13).

Individual chromosomes can be recognized by their lengths, the positions of their centromeres, and characteristic banding when they are stained and observed at high magnification. When the cell is diploid, the karyotype consists of homologous pairs of chromosomes—23 pairs for a total of 46 chromosomes in humans, and greater or smaller numbers of pairs in other diploid species. There is no simple relationship between the size of an organism and its chromosome number (Table 9.1).

Meiosis: A Pair of Nuclear Divisions

Meiosis consists of two nuclear divisions that reduce the number of chromosomes to the haploid number in preparation for sexual reproduction. Although the *nucleus divides twice* during meiosis, the *DNA is replicated only once*. Unlike the products of mitosis, the products of meiosis are different from one an other and from the parent cell. To understand the process of meiosis and its specific details, it is useful to keep in mind the overall functions of meiosis:

- To reduce the chromosome number from diploid to haploid
- To ensure that each of the haploid products has a complete set of chromosomes
- To promote genetic diversity among the products

The first meiotic division reduces the chromosome number

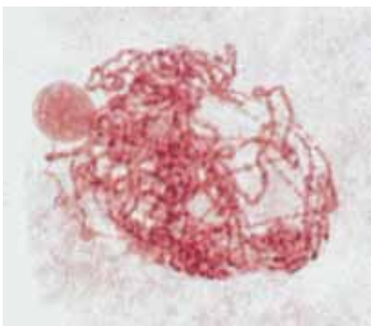
Two unique features characterize the first of the two meiotic divisions, **meiosis I**. The first is that homologous chromosomes pair along their entire lengths. No such pairing occurs in mitosis. The second is that after metaphase I, the homologous chromosomes separate. The individual chromosomes, each consisting of two sister chromatids, remain intact until the end of metaphase II in the second meiotic division. Throughout this discussion, refer to Figure 9.14 on the two following pages to help you visualize each step.

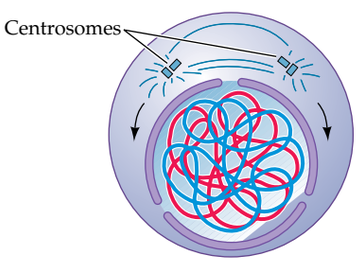
9.1 *Numbers of Pairs of Chromosomes in Some Plant and Animal Species*

COMMON NAME	SPECIES	NUMBER OF CHROMOSOME PAIRS
Mosquito	*Culex pipiens*	3
Housefly	*Musca domestica*	6
Toad	*Bufo americanus*	11
Rice	*Oryza sativa*	12
Frog	*Rana pipiens*	13
Alligator	*Alligator mississippiensis*	16
Rhesus monkey	*Macaca mulatta*	21
Wheat	*Triticum aestivum*	21
Human	*Homo sapiens*	23
Potato	*Solanum tuberosum*	24
Donkey	*Equus asinus*	31
Horse	*Equus caballus*	32
Dog	*Canis familiaris*	39
Carp	*Cyprinus carpio*	52

MEIOSIS I

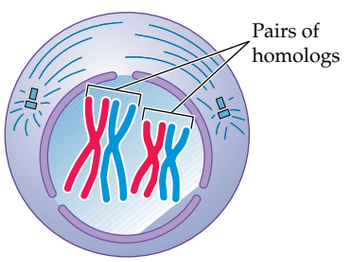
Early Prophase I





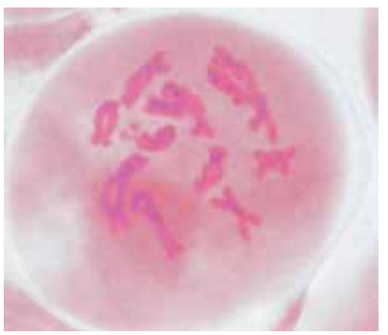
1 The chromatin begins to condense following interphase.

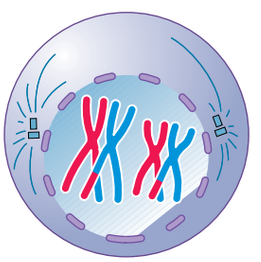
Mid-Prophase I



2 Synapsis aligns homologs, and chromosomes condense. Homologs are shown in different colors indicating those coming from each parent. In reality, their differences are very small, usually comprising different alleles of some genes.

Late Prophase I–Prometaphase

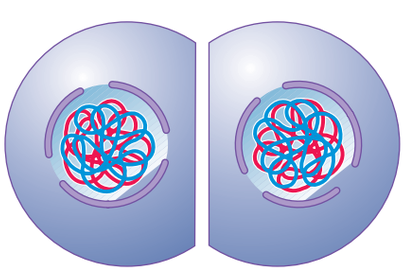




3 The chromosomes continue to coil and shorten. Crossing over results in an exchange of genetic material. In prometaphase the nuclear envelope breaks down.

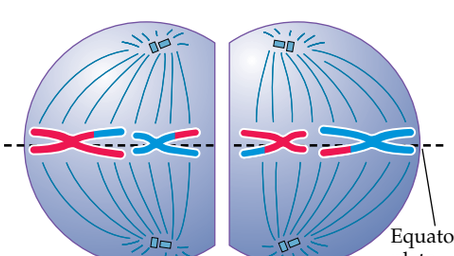
MEIOSIS II

Prophase II



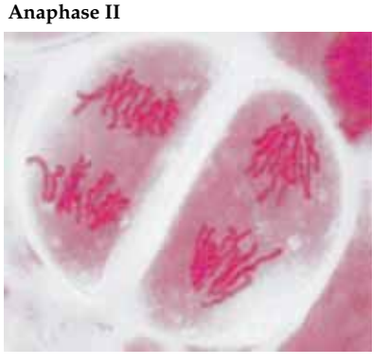
7 The chromosomes condense again, following a brief interphase (interkinesis) in which DNA does not replicate.

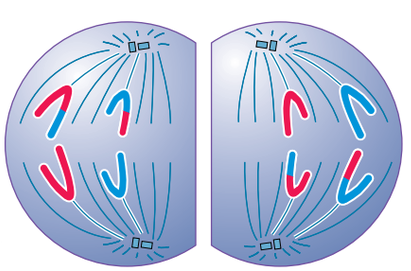
Metaphase II



8 Kinetochores of the paired chromatids line up across the equatorial plates of each cell.

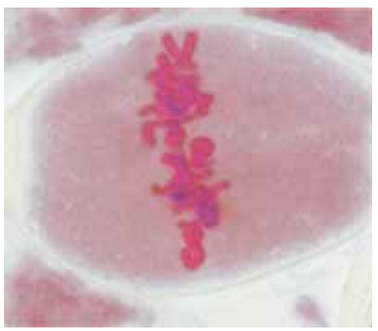
Anaphase II

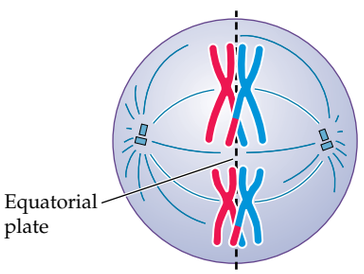




9 The chromatids finally separate, becoming chromosomes in their own right, and are pulled to opposite poles. Because of crossing over in prophase I, each new cell will have a different genetic makeup.

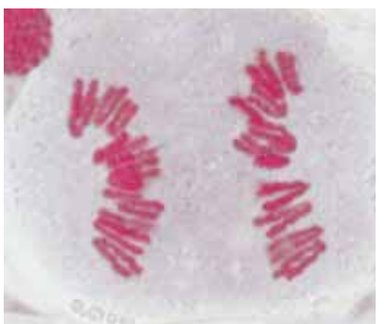
Metaphase I

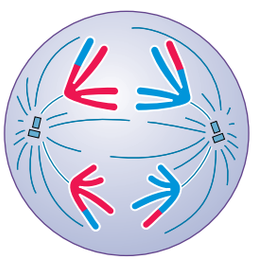




4 The homologous chromosomes line up on the equatorial (metaphase) plate.

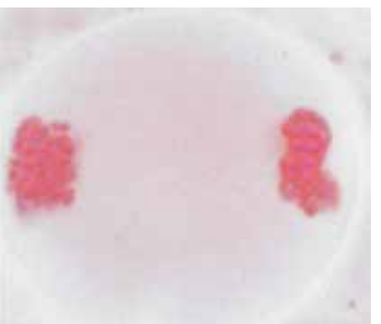
Anaphase I

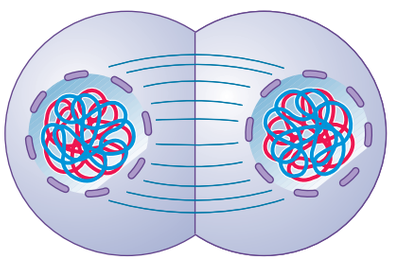




5 The homologous chromosomes (each with two chromatids) move to opposite poles of the cell.

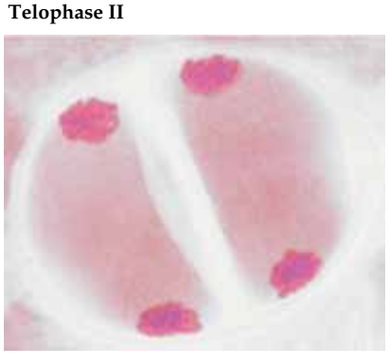
Telophase I

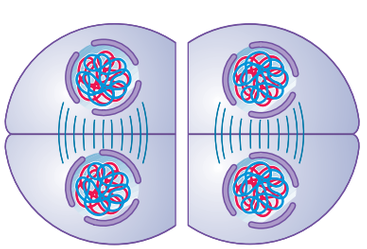




6 The chromosomes gather into nuclei, and the original cell divides.

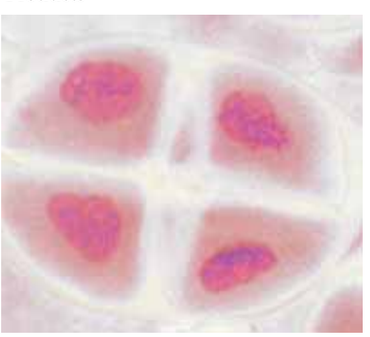
Telophase II

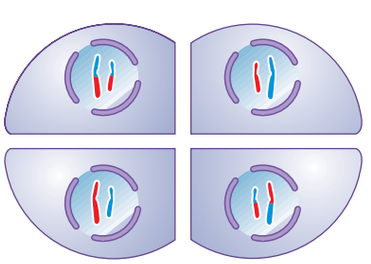




10 The chromosomes gather into nuclei, and the cells divide.

Products





11 Each of the four cells has a nucleus with a haploid number of chromosomes.

9.14 Meiosis In meiosis, two sets of chromosomes are divided among four nuclei, each of which then has half as many chromosomes as the original cell. These four haploid cells are the result of two successive nuclear divisions. The photomicrographs show meiosis in the male reproductive organ of a lily. As in Figure 9.8, the diagrams show corresponding phases in an animal.

Like mitosis, meiosis I is preceded by an interphase with an S phase during which each chromosome is replicated. As a result, each chromosome consists of two sister chromatids, held together by cohesin proteins.

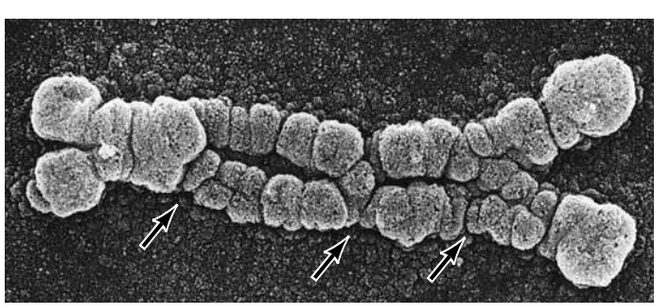
Meiosis I begins with a long prophase I (the first three frames of Figure 9.14), during which the chromosomes change markedly. The homologous chromosomes pair by adhering along their lengths, a process called **synapsis**. This process lasts from prophase I to the end of metaphase I.

By the time chromosomes can be clearly seen under light microscope, the two homologs are already tightly joined. This joining begins at the centromeres and is mediated by a recognition of homologous DNA sequences on homologous chromosomes. In addition, a special group of proteins may form a scaffold called the *synaptonemal complex*, which runs lengthwise along the homologous chromosomes and appears to join them together.

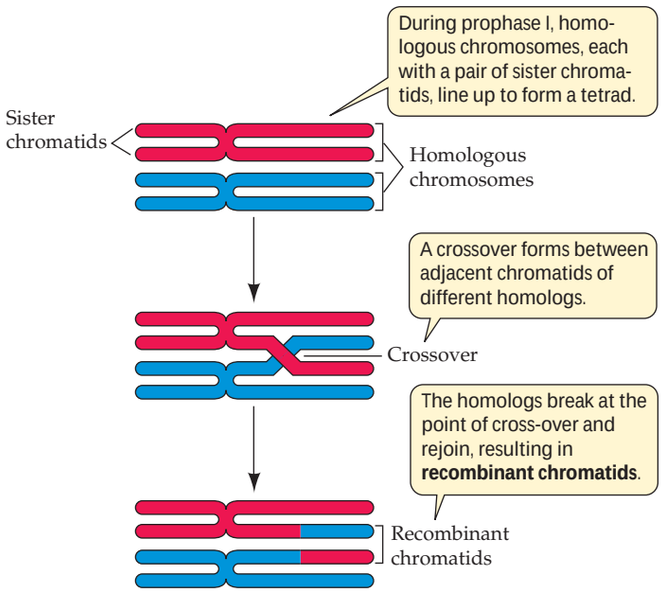
The four chromatids of each pair of homologous chromosomes form what is called a **tetrad**, or *bivalent*. In other words, a tetrad consists of four chromatids, two each from two homologous chromosomes. For example, there are 46 chromosomes in a human diploid cell at the beginning of meiosis, so there are 23 homologous pairs of chromosomes, each with two chromatids (that is, 23 tetrads), for a total of 92 chromatids during prophase I.

Throughout prophase I and metaphase I, the chromatin continues to coil and compact, so that the chromosomes appear ever thicker. At a certain point, the homologous chromosomes seem to repel each other, especially near the centromeres, but they are held together by physical attachments mediated by cohesins. These cohesins are different from the ones holding the two sister chromatids together. Regions having these attachments take on an X-shaped appearance and are called **chiasmata** (from the Greek *chiasma*, "cross"; Figure 9.15).

A chiasma reflects an exchange of genetic material between nonsister chromatids on homologous chromosomes—



9.15 Chiasmata: Evidence of Exchange between Chromatids
Chiasmata are visible near the middle of this scanning electron micrograph of some chromatids from a desert locust, and near the ends of others. Three chiasmata are indicated with arrows.



9.16 Crossing Over Forms Genetically Diverse Chromosomes
The exchange of genetic material by crossing over may result in new combinations of genetic information on the recombinant chromosomes.

what geneticists call **crossing over** (Figure 9.16). The chromosomes begin exchanging material shortly after synapsis begins, but chiasmata do not become visible until later, when the homologs are repelling each other. Crossing over increases genetic variation among the products of meiosis by reshuffling genetic information among the homologous pairs. We will have a great deal to say about crossing over and its genetic consequences in the coming chapters.

There seems to be plenty of time for the complicated events of prophase I to occur. Whereas mitotic prophase is usually measured in minutes, and all of mitosis seldom takes more than an hour or two, meiosis can take much longer. In human males, the cells in the testis that undergo meiosis take about a week for prophase I and about a month for the entire meiotic cycle. In the cells that will become eggs, prophase I begins long before a woman's birth, during her early fetal development, and ends as much as decades later, during the monthly ovarian cycle.

Prophase I is followed by prometaphase I (not pictured in Figure 9.14), during which the nuclear envelope and the nucleoli disaggregate. A spindle forms, and microtubules become attached to the kinetochores of the chromosomes. In meiosis I, the kinetochores of both chromatids in each chromosome become attached to the same half-spindle. Thus the entire chromosome, consisting of two chromatids, will migrate to one pole. Which member of a homologous chromosome pair becomes attached to each half-spindle,

and thus which member will go to which pole, is random. By metaphase I, all the chromosomes have moved to the equatorial plate. Up to this point, homologous pairs are held together by chiasmata.

The homologous chromosomes separate in anaphase I, when the individual chromosomes, each still consisting of two chromatids, are pulled to the poles, with one homolog of a pair going to one pole and the other homolog going to the opposite pole. (Note that this process differs from the separation of chromatids during mitotic anaphase.) Each of the two daughter nuclei from this division thus contains only one set of chromosomes, not the two sets that were present in the original diploid nucleus. However, because they consist of two chromatids rather than just one, each of these chromosomes has twice the mass that a chromosome at the end of a mitotic division has.

In some organisms, there is a telophase I, with the reappearance of the nuclear envelopes. When there is a telophase I, it is followed by an interphase, called **interkinesis**, similar to the mitotic interphase. During interkinesis the chromatin is partially uncoiled; however, there is no replication of the genetic material, because each chromosome already consists of two chromatids. Furthermore, the sister chromatids in interkinesis are generally not genetically identical, because crossing over in prophase I has reshuffled genetic material between the maternal and paternal chromosomes. In other organisms, the chromosomes move directly into the second meiotic division.

The second meiotic division separates the chromatids

Meiosis II is similar to mitosis in many ways. In each nucleus produced by meiosis I, the chromosomes line up at equatorial plate at metaphase II. The centromeres of the sister chromatids separate because of cohesin breakdown, and the daughter chromosomes move to the poles in anaphase II.

The three major differences between meiosis II and mitosis are:

- DNA replicates before mitosis, but not before meiosis II.
- In mitosis, the sister chromatids that make up a given chromosome are identical. In meiosis II, they may differ over part of their length if they participated in crossing over during prophase I.
- The number of chromosomes on the equatorial plate in meiosis II is half the number in the mitotic nucleus.

The result of meiosis is four nuclei; each nucleus is haploid and has a single set of unreplicated chromosomes that differs from other such sets in its exact genetic composition. The differences among the haploid nuclei result from crossing over during prophase I and from the random segregation of homologous chromosomes during anaphase I.

Meiosis leads to genetic diversity

What are the consequences of the synapsis and segregation of homologous chromosomes during meiosis? In mitosis, each chromosome behaves independently of its homolog; its two chromatids are sent to opposite poles at anaphase. If a mitotic division begins with x chromosomes, we end up with x chromosomes in each daughter nucleus, and each chromosome consists of one chromatid. Each of the two sets of chromosomes (one of paternal and one of maternal origin) is divided equally and distributed equally to each daughter cell. In meiosis, things are very different.

In meiosis, chromosomes of maternal origin pair with their paternal homologs during synapsis. Separation of the homologs during meiotic anaphase I ensures that each pole receives one member of each homologous pair. For example, at the end of meiosis I in humans, each daughter nucleus contains 23 of the original 46 chromosomes. In this way, the chromosome number is decreased from diploid to haploid. Furthermore, meiosis I guarantees that each daughter nucleus gets one full set of chromosomes.

The products of meiosis I are genetically diverse for two reasons:

- Synapsis during prophase I allows the maternal chromosome in each homologous pair to exchange segments with the paternal one by crossing over. The resulting *recombinant* chromatids contain some genetic material from each parent.
- It is a matter of chance which member of a homologous pair goes to which daughter cell at anaphase I. For example, if there are two homologous pairs of chromosomes in the diploid parent nucleus, a particular daughter nucleus could get paternal chromosome 1 and maternal chromosome 2, or paternal 2 and maternal 1, or both maternal, or both paternal. It all depends on the way in which the homologous pairs line up at metaphase I. This phenomenon is termed **independent assortment**.

Note that of the four possible chromosome combinations just described, two produce daughter nuclei that are the same as one of the parental types (except for any material exchanged by crossing over). The greater the number of chromosomes, the less probable that the original parental combinations will be reestablished, and the greater the potential for genetic diversity. Most species of diploid organisms do indeed have more than two pairs of chromosomes. In humans, with 23 chromosome pairs, 2^{23} (8,388,608) different combinations can be produced, just by the mechanism of independent assort-

ment. Taking the extra genetic shuffling afforded by crossing over into account, the number of possible combinations is virtually infinite.

Figure 9.17 compares meiosis with mitosis.

Meiotic Errors

In the complex process of cell division, things occasionally go wrong. A pair of homologous chromosomes may fail to separate during meiosis I, or sister chromatids may fail to separate during meiosis II or during mitosis. This phenomenon is called **nondisjunction**. Conversely, homologous chromosomes may fail to remain together. These problems can result in the production of aneuploid cells. **Aneuploidy** is a condition in which one or more chromosomes are either lacking or present in excess.

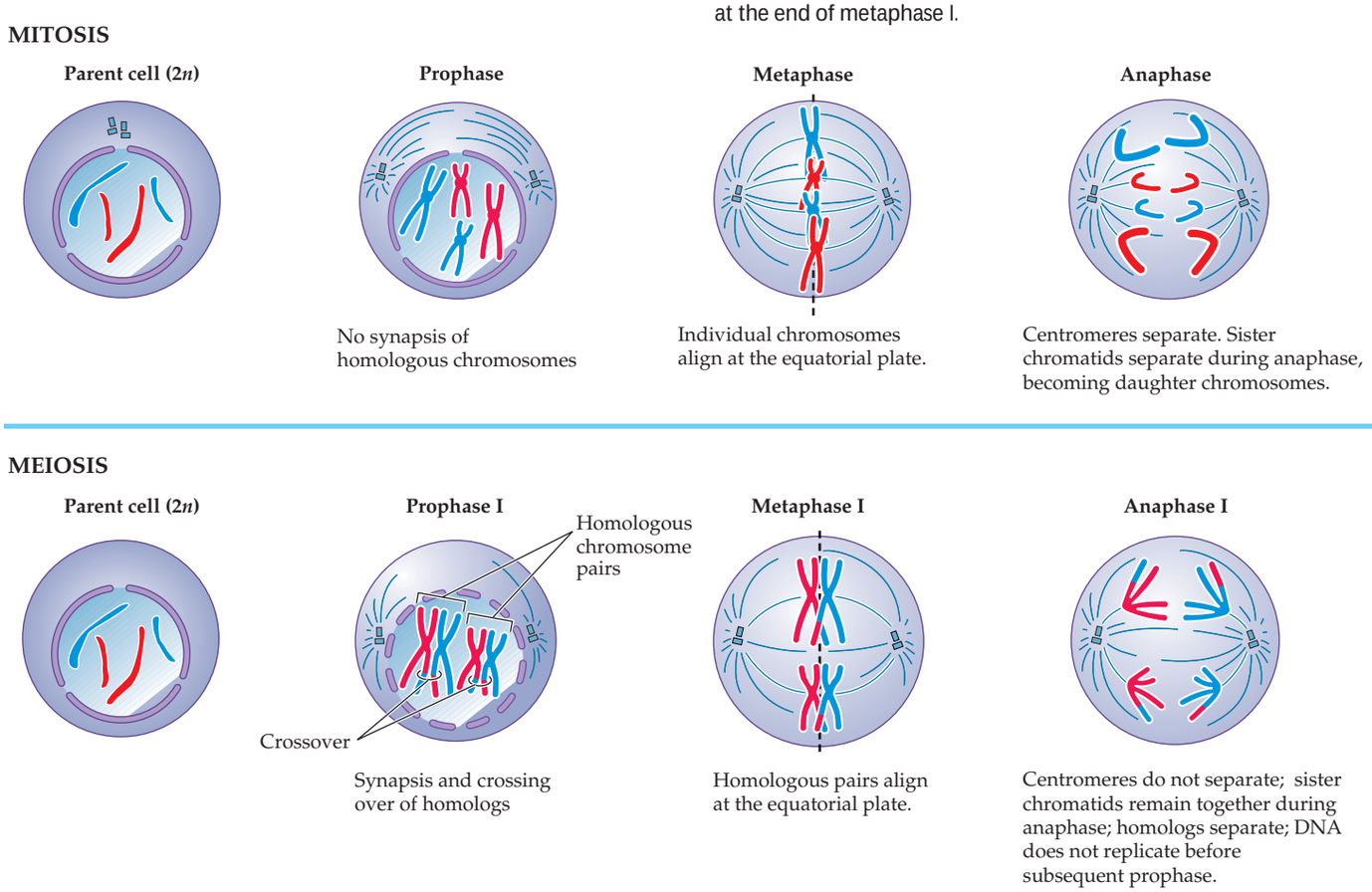
Aneuploidy can give rise to genetic abnormalities

One reason for aneuploidy may be a lack of cohesins. Recall that these molecules, formed during prophase I, hold the two homologous chromosomes together into metaphase I. They ensure that when the chromosomes line up at the equatorial plate, one homolog will face one pole and the other homolog will face the other pole. Without this "glue," the two homologs may line up randomly at metaphase I, just like chromosomes during mitosis, and there is a 50 percent chance that both will go to the same pole.

If, for example, during the formation of a human egg, both members of the chromosome 21 pair go to the same pole during anaphase I, the resulting eggs will contain either two of chromosome 21 or none at all. If an egg with two of these chromosomes is fertilized by a normal sperm, the resulting zygote will have three copies of the chromosome: it will be **trisomic** for chromosome 21. A child with an extra chromosome 21 demonstrates the symptoms of Down syndrome: impaired intelligence; characteristic abnormalities of the hands, tongue, and eyelids; and an increased susceptibility to cardiac abnormalities and diseases such as leukemia. If an egg that did not receive chromosome 21 is fertilized by a normal sperm, the zygote will have only one copy: it will be **monosomic** for chromosome 21 (Figure 9.18).

Other abnormal chromosomal events can also occur. In a process called **translocation**, a piece of a chromosome may

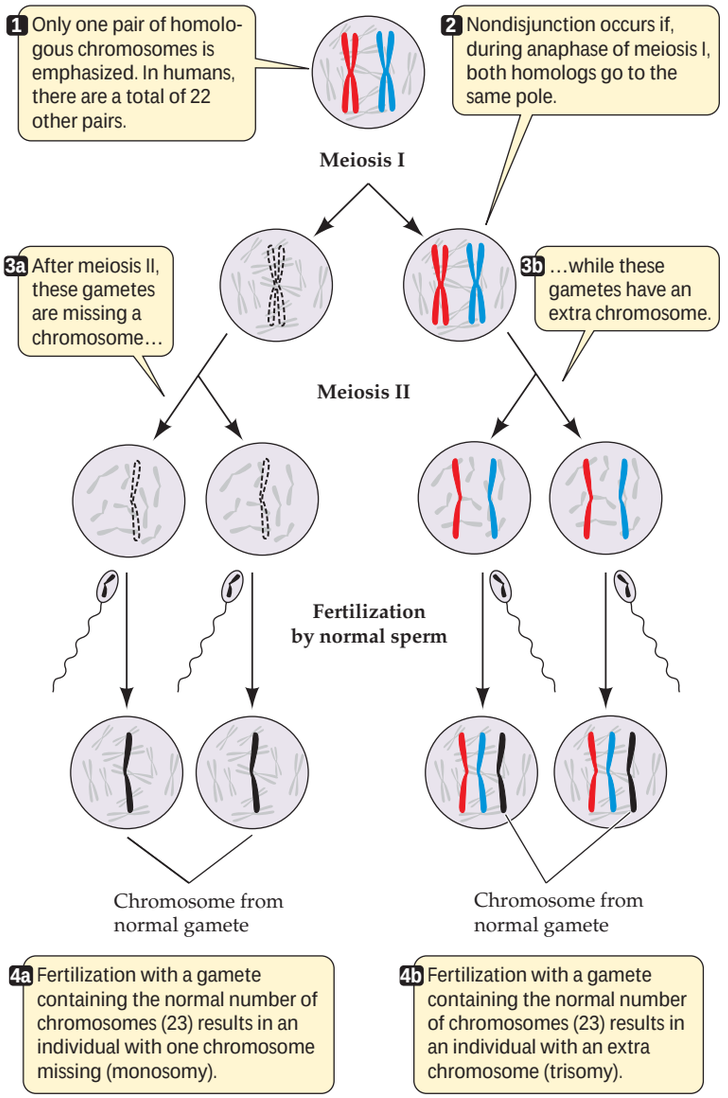
9.17 Mitosis and Meiosis: A Comparison Meiosis differs from mitosis by synapsis and by the failure of the centromeres to separate at the end of metaphase I.

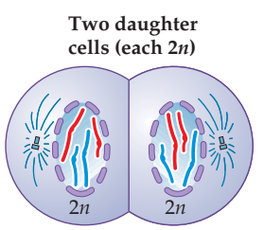


9.18 Nondisjunction Leads to Aneuploidy Nondisjunction occurs if homologous chromosomes fail to separate during meiosis I. The result is aneuploidy: One or more chromosomes are either lacking or present in excess.

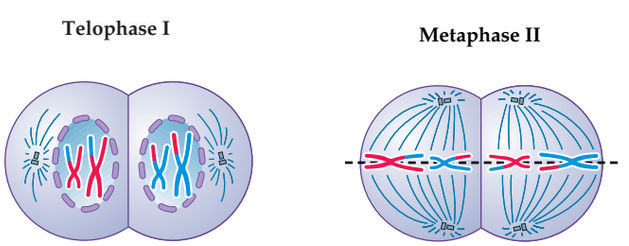
break away and become attached to another chromosome. For example, a particular large part of one chromosome 21 may be translocated to another chromosome. Individuals who inherit this translocated piece along with two normal chromosomes 21 will have Down syndrome.

Trisomies (and the corresponding monosomies) are surprisingly common in human zygotes, with 10–30 percent of all conceptions showing aneuploidy. But most of the embryos that develop from such zygotes do not survive to birth, and those that do often die before the age of 1 year. Trisomies and monosomies for most chromosomes other than chromosome 21 are lethal to the embryo. At least one-fifth of all recognized pregnancies are spontaneously terminated during the first 2 months, largely because of such trisomies and monosomies. (The actual proportion of spontaneously terminated pregnancies is certainly higher, because the earliest ones often go unrecognized.)





Mitosis is a mechanism for constancy: The parent nucleus produces two identical daughter nuclei.



Chromatids separate.

Meiosis is a mechanism for diversity: The parent nucleus produces four different haploid daughter nuclei.

Polyploids can have difficulty in cell division

As we saw earlier in our discussion of sexual life cycles, both diploid and haploid nuclei can divide by mitosis. Multicellular diploid and multicellular haploid individuals both develop from single-celled beginnings by mitotic divisions. Likewise, mitosis may proceed in diploid organisms even when a chromosome is missing from one of the haploid sets or when there is an extra copy of one of the chromosomes (as in people with Down syndrome).

Organisms with complete extra sets of chromosomes may sometimes be produced by artificial breeding or by natural accidents. Under some circumstances, triploid ($3n$), tetraploid ($4n$), and higher-order **polyploid** nuclei may form. Each of these *ploidy levels* represents an increase in the number of complete sets of chromosomes present.

If a nucleus has one or more extra full sets of chromosomes, its abnormally high ploidy in itself does not prevent mitosis. In mitosis, each chromosome behaves independently of the others. In meiosis, by contrast, homologous chromosomes must synapse to begin division. If even one chromosome has no homolog, anaphase I cannot send representatives of that chromosome to both poles. A diploid nucleus can undergo normal meiosis; a haploid one cannot. Similarly, a tetraploid nucleus has an even number of each kind of chromosome, so each chromosome can pair with its homolog. But a triploid nucleus cannot undergo normal meiosis, because one-third of the chromosomes would lack partners.

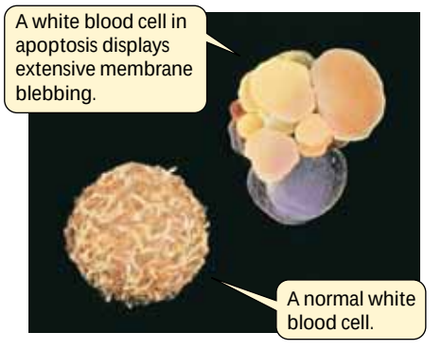
This limitation has important consequences for the fertility of triploid, tetraploid, and other chromosomally unusual organisms. Modern bread wheat plants are hexaploids, the result of naturally occurring crosses between three different grasses, each having its own diploid set of 14 chromosomes. Over a period of 10,000 years, humans have selected favorable varieties of these hybrids to produce modern wheat strains.

Cell Death

As we mentioned at the start of this chapter, an essential role of cell division in complex eukaryotes is to replace cells that die. In humans, billions of cells die each day, mainly in the blood and in the epithelia lining organs such as the intestine. Cells die in one of two ways. The first, **necrosis**, occurs when cells either are damaged by poisons or are starved of essential nutrients. These cells usually swell up and burst, releasing their contents into the extracellular environment. This process often results in inflammation (see Chapter 18). The scab that forms around a wound is a familiar example of necrotic tissue. More typically, cell death in an organism is due to **apoptosis** (from the Greek, "falling off"). Apoptosis is a genetically programmed series of events that result in cell death. These two modes of cell death are compared in Table 9.2.

Why would a cell initiate apoptosis, which is essentially "cell suicide"? There are two possible reasons:

- The cell is no longer needed by the organism. For example, before birth, a human fetus has weblike hands, with connective tissue between the fingers. As development proceeds, this unneeded tissue disappears as its cells undergo apoptosis (see Figure 19.11).
- The longer cells live, the more prone they are to genetic damage that could lead to cancer. This is especially true of cells in the blood and intestine, which are exposed to high levels of toxic substances. Such cells normally die after only days or weeks.



9.19 Apoptosis: Programmed Cell Death Many cells are genetically programmed to "self-destruct" when they are no longer needed, or when they have lived long enough to accumulate a burden of DNA damage that might harm the organism.

9.2 Two Different Ways for Cells to Die

	NECROSIS	APOPTOSIS
Stimuli	Low O_2, toxins, ATP depletion, damage	Specific, genetically programmed physiological signals
ATP required	No	Yes
Cellular pattern	Swelling, organelle disruption, tissue death	Chromatin condensation, membrane blebbing, single-cell death
DNA breakdown	Random fragments	Nucleosome-sized fragments
Plasma membrane	Burst	Blebbed (see Figure 9.19)
Fate of dead cells	Ingested by phagocytes	Ingested by neighboring cells
Reaction in tissue	Inflammation	No inflammation

Like the cell division cycle, the cell death cycle is controlled by signals, which may come either from inside or outside the cell. These signals include the lack of a mitotic signal (such as a growth factor), and the recognition of DNA damage. As we will see in Chapter 17, many of the drugs used to treat diseases of excess cell proliferation, such as cancer, work through these signals.

The events of apoptosis are very similar in most organisms. The cell becomes isolated from its neighbors, chops up its chromatin into nucleosome-sized pieces, and then fragments itself (Figure 9.19). In a remarkable example of the economy of nature, the surrounding living cells usually ingest the remains of the dead cell. The genetic signals that lead to apoptosis are also common to many organisms.

Chapter Summary

Systems of Cell Reproduction

▶ Cell division is necessary for the reproduction, growth, and repair of an organism. **Review Figure 9.1**

▶ Cell division must be initiated by a reproductive signal. Cell division consists of three steps: replication of the genetic material (DNA), segregation of the two DNA molecules to separate portions of the cell, and cytokinesis, or division of the cytoplasm.

▶ In prokaryotes, cellular DNA is a single molecule, or chromosome. Prokaryotes reproduce by cell fission. **Review Figure 9.2**

▶ In eukaryotes, cells divide by either mitosis or meiosis.

Interphase and the Control of Cell Division

▶ The mitotic cell cycle has two main phases: interphase (during which cells are not dividing) and mitosis (when cells are dividing).

▶ During most of the cell cycle, the cell is in interphase, which is divided into three subphases: S, G1, and G2. DNA is replicated during the S phase. **Review Figure 9.3**

▶ Cyclin-Cdk complexes regulate the passage of cells through checkpoints in the cell cycle. The most important one is the R point in G1, which determines whether the rest of the cycle will proceed. **Review Figure 9.4**

▶ In addition to the internal cyclin-Cdk complexes, controls external to the cell, such as growth factors and hormones, can also stimulate the cell to begin a division cycle.

Eukaryotic Chromosomes

▶ A eukaryotic chromosome contains a DNA molecule bound to proteins in a complex called chromatin. At mitosis, the replicated chromatids are held together at the centromere. Each chromatid consists of one double-stranded DNA molecule. **Review Figure 9.5**

▶ During interphase, the DNA in chromatin is wound around cores of histones to form nucleosomes. DNA folds over and over again, packing itself within the nucleus. During mitosis or meiosis, it folds even more. **Review Figure 9.6**

Mitosis: Distributing Exact Copies of Genetic Information

▶ After DNA is replicated during the S phase, the first sign of mitosis is the separation of the replicated centrosomes, which initiate microtubule formation for the spindle.

▶ Mitosis can be divided into several phases, called prophase, prometaphase, metaphase, anaphase, and telophase.

▶ During prophase, the chromosomes condense and appear as paired chromatids, and the spindle forms. **Review Figure 9.7.** **See Web/CD Activity 9.1**

▶ During prometaphase, the chromosomes move toward the middle of the spindle. In metaphase, they gather at the middle of the cell with their centromeres on the equatorial plate. At the end of metaphase, the centromeres holding the sister chromatids together separate, and during anaphase, each chromatid, now called the daughter chromosome, migrates to its pole along the microtubule track. **Review Figure 9.8.** **See Web/CD Activity 9.2**

▶ Cohesin holds sister chromatids together from the time they are formed in DNA replication until the onset of anaphase. Separin hydrolyzes cohesin when an inhibitory subunit, securin, is hydrolyzed. **Review Figure 9.9**

▶ During telophase, the chromosomes become less condensed. The nuclear envelopes and nucleoli re-form, thus producing two nuclei whose chromosomes are identical to each other and to those of the cell that began the cycle. **See Web/CD Tutorial 9.1**

Cytokinesis: The Division of the Cytoplasm

▶ Nuclear division is usually followed by cytokinesis. Animal cell cytoplasm usually divides by a furrowing of the plasma membrane, caused by the contraction of cytoplasmic microfilaments. In plant cells, cytokinesis is accomplished by vesicle fusion and the synthesis of new cell wall material. **Review Figure 9.10**

Reproduction: Asexual and Sexual

▶ The cell cycle can repeat itself many times, forming a clone of genetically identical cells.

▶ Asexual reproduction produces a new organism that is genetically identical to the parent. Any genetic variety is the result of mutations.

▶ In sexual reproduction, two haploid gametes—one from each parent—unite in fertilization to form a genetically unique, diploid zygote. **Review Figure 9.12.** **See Web/CD Activity 9.3**

▶ In sexually reproducing organisms, certain cells in the adult undergo meiosis, a process by which a diploid cell produces haploid gametes. Each gamete contains a random selection of one of each pair of homologous chromosomes from the parent.

▶ The number, shapes, and sizes of the chromosomes constitute the karyotype of an organism. **Review Figure 9.13**

Meiosis: A Pair of Nuclear Divisions

▶ Meiosis reduces the chromosome number from diploid to haploid, ensures that each haploid cell contains one member of each chromosome pair, and results in genetically diverse products. It consists of two nuclear divisions. **Review Figure 9.14.** **See Web/CD Activity 9.4**

▶ During prophase I of the first meiotic division, homologous chromosomes pair up with each other, and material may be exchanged between the two homologs by crossing over. In metaphase I, the paired homologs line up at the equatorial plate. **Review Figures 9.14, 9.16**

▶ In anaphase I, entire chromosomes, each with two chromatids, migrate to the poles. By the end of meiosis I, there are two nuclei, each with the haploid number of chromosomes. **Review Figures 9.14, 9.17**

▶ In meiosis II, the sister chromatids separate. No DNA replication precedes this division, which in other aspects is similar to mitosis. The result of meiosis is four cells, each with a haploid chromosome content. **Review Figures 9.14, 9.17**

▶ Both crossing over during prophase I and the random selection of which homolog of a pair migrates to which pole during anaphase I ensure that the genetic composition of each haploid

gamete is different from that of the parent cell and from that of the other gametes. The more chromosome pairs there are in a diploid cell, the greater the diversity of chromosome combinations generated by meiosis. **Review Figure 9.16, Table 9.1.** **See Web/CD Tutorial 9.2**

Meiotic Errors

▶ In nondisjunction, one member of a homologous pair of chromosomes fails to separate from the other, and both go to the same pole. Pairs of homologous chromosomes may also fail to stick together when they should. These events may lead to one gamete with an extra chromosome and another lacking that chromosome.

▶ The union of a gamete with an abnormal chromosome number with a normal haploid gamete at fertilization results in aneuploidy and genetic abnormalities that are invariably harmful or lethal to the organism. **Review Figure 9.18**

▶ Polyploid organisms can have difficulty in cell division. Natural and artificially produced polyploids underlie modern agriculture.

Cell Death

▶ Cells may die by necrosis, or they may self-destruct by apoptosis, a genetically programmed series of events that includes the detachment of the cell from its neighbors and the fragmentation of its nuclear DNA. **Review Figure 9.19, Table 9.2**

Self-Quiz

1. Which statement about eukaryotic chromosomes is *not* true?
   a. They sometimes consist of two chromatids.
   b. They sometimes consist of a single chromatid.
   c. They normally possess a single centromere.
   d. They consist of proteins.
   e. They are clearly visible as defined bodies under the light microscope.
2. Nucleosomes
   a. are made of chromosomes.
   b. consist entirely of DNA.
   c. consist of DNA wound around a histone core.
   d. are present only during mitosis.
   e. are present only during prophase.
3. Which statement about the cell cycle is *not* true?
   a. It consists of mitosis and interphase.
   b. The cell's DNA replicates during G1.
   c. A cell can remain in G1 for weeks or much longer.
   d. Proteins are formed throughout all subphases of interphase.
   e. Histones are synthesized primarily during S phase.
4. Which statement about mitosis is *not* true?
   a. A single nucleus gives rise to two identical daughter nuclei.
   b. The daughter nuclei are genetically identical to the parent nucleus.
   c. The centromeres separate at the onset of anaphase.
   d. Homologous chromosomes synapse in prophase.
   e. Mitotic centers organize the microtubules of the spindle fibers.
5. Which statement about cytokinesis is true?
   a. In animals, a cell plate forms.
   b. In plants, it is initiated by furrowing of the membrane.
   c. It generally immediately follows mitosis.
   d. In plant cells, actin and myosin play an important part.
   e. It is the division of the nucleus.
6. Apoptosis
   a. occurs in all cells.
   b. involves the cell membrane dissolving.
   c. does not occur in an embryo.
   d. involves a series of programmed events for cell death.
   e. is not involved with cancer.
7. In meiosis,
   a. meiosis II reduces the chromosome number from diploid to haploid.
   b. DNA replicates between meiosis I and II.
   c. the chromatids that make up a chromosome in meiosis II are identical.
   d. each chromosome in prophase I consists of four chromatids.
   e. homologous chromosomes separate from one another in anaphase I.
8. In meiosis,
   a. a single nucleus gives rise to two daughter nuclei.
   b. the daughter nuclei are genetically identical to the parent nucleus.
   c. the centromeres separate at the onset of anaphase I.
   d. homologous chromosomes synapse in prophase I.
   e. no spindle forms.
9. A plant has a diploid chromosome number of 12. An egg cell of the plant has 5 chromosomes. The most probable explanation of this is
   a. normal mitosis.
   b. normal meiosis.
   c. nondisjunction in meiosis I.
   d. nondisjunction in meiosis I and II.
   e. nondisjunction in mitosis.
10. The number of daughter chromosomes in a human cell in anaphase II of meiosis is
    a. 2.
    b. 23.
    c. 46.
    d. 69.
    e. 92.

For Discussion

1. How does a nucleus in the G2 phase of the cell cycle differ from one in the G1 phase?
2. Compare the roles of cohesins and condensin in mitosis, meiosis I, and meiosis II.
3. Compare and contrast mitosis (and subsequent cytokinesis) in animals and plants.
4. Suggest two ways in which, with the help of a microscope, one might determine the relative duration of the various phases of mitosis.
5. Contrast mitotic prophase and prophase I of meiosis. Contrast mitotic anaphase and anaphase I of meiosis.
6. Compare the sequence of events in the mitotic cell cycle with the sequence in programmed cell death.

7 Genetics: Mendel and Beyond

In the Middle Eastern desert 1,800 years ago, the rabbi faced a dilemma. A Jewish woman had given birth to a son. As required by the laws set down by God's commandment to Abraham almost 2,000 years previously and later reiterated by Moses, the mother brought her 8-day-old son to the rabbi for ritual penile circumcision. The rabbi knew that the woman's two previous sons had bled to death when their foreskins were cut. Yet the biblical commandment remained: Unless he was circumcised, the boy could not be counted among those with whom God had made His solemn covenant. After consultation with other rabbis, it was decided to exempt this, the third son.

Almost a thousand years later, in the twelfth century, the physician and biblical commentator Moses Maimonides reviewed this and numerous other cases in the rabbinical literature and stated that in such instances the third son should not be circumcised. Furthermore, the exemption should apply whether the mother's son was "from her first husband or from her second husband." The bleeding disorder, he reasoned, was clearly carried by the mother and passed on to her sons.

Knowing nothing of our modern vision of genetics, these rabbis linked a human disease (which we now know as hemophilia A) to a pattern of inheritance (which we know as sex linkage). Only in the past few decades have the precise biochemical nature of hemophilia A and its genetic determination been worked out.

How do we account for, and predict, such patterns of inheritance? In this chapter, we will discuss how the units of inheritance, called genes, are transmitted from generation to generation, and we will show how many of the rules that govern genetics can be explained by the behavior of chromosomes during meiosis. We will also describe the interactions of genes with one another and with the environment, and we will examine the consequences of the fact that genes occupy specific positions on chromosomes.

An Ancient Ritual
A male infant undergoes ritual circumcision in accordance with Jewish laws. Sons of Jewish mothers who carry the gene for hemophilia may be exempted from the ritual.

The Foundations of Genetics

Much of the early study of biological inheritance was done with plants and animals of economic importance. Records show that people were deliberately cross-breeding date palm trees and horses as early as 5,000 years ago. By the early nineteenth century, plant breed-

ing was widespread, especially with ornamental flowers such as tulips. Half a century later, Gregor Mendel used the existing knowledge of plant reproduction to design and conduct experiments on inheritance. Although his published results were neglected by scientists for more than 30 years, they ultimately became the foundation for the science of genetics.

Plant breeders showed that both parents contribute equally to inheritance

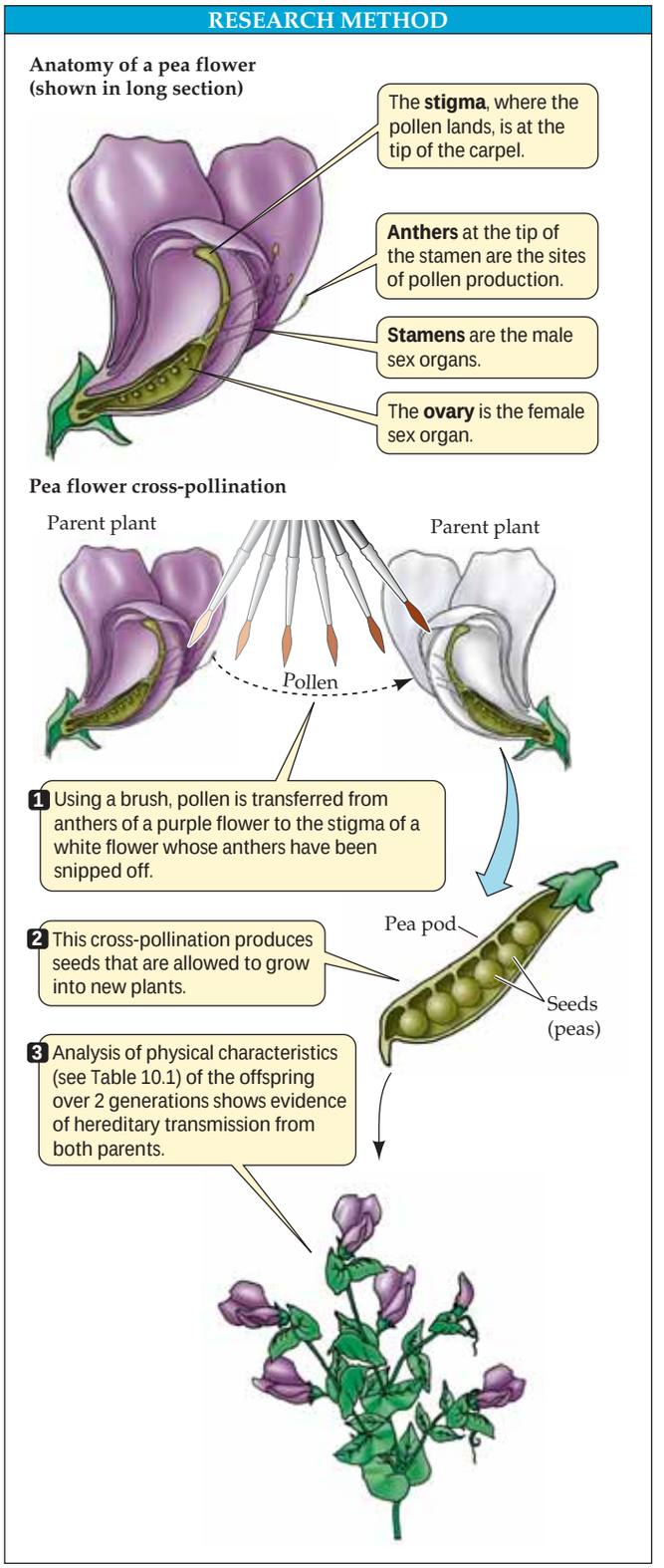
Plants are good experimental subjects for the study of genetics. Many plants are easily grown in large quantities, produce large numbers of offspring (in the form of seeds), and have relatively short generation times. In most plant species, the same individuals have both male and female reproductive organs, permitting each plant to reproduce as a male, as a female, or as both. Best of all, it is often easy to control which individuals mate (Figure 10.1).

Some discoveries that Mendel found useful in his studies had been made in the late eighteenth century by a German botanist, Josef Gottlieb Kölreuter. He had studied the offspring of **reciprocal crosses**, in which plants are crossed (mated with each other) in opposite directions. For example, in one cross, males that have white flowers are mated with females that have red flowers, while in a complementary cross, red-flowered males and white-flowered females are mated. In Kölreuter's studies, such reciprocal crosses always gave identical results, showing that both parents contributed equally to the offspring.

Although the concept of equal parental contributions was an important discovery, the nature of what exactly the parents were contributing to their offspring—the units of inheritance—remained unknown. Laws of inheritance proposed at the time favored the concept of *blending*. If a plant that had one form of a characteristic (say, red flowers) was crossed with one that had a different form of that characteristic (blue flowers), the offspring would be a blended combination of the two parents (purple flowers). According to the blending theory, it was thought that once heritable elements were combined, they could not be separated again (like inks of different colors mixed together). The red and blue genetic determinants were thought to be forever blended into the new purple one. Then, about a century after Kölreuter completed his work, Mendel began his.

Mendel brought new methods to experiments on inheritance

Gregor Mendel was an Austrian monk, not an academic scientist, but he was qualified to undertake scientific investigations. Although in 1850 he had failed an examination for a teaching certificate in natural science, he later undertook in-



10.1 A Controlled Cross between Two Plants Plants were widely used in early genetic studies because it is easy to control which individuals mate with which. Mendel used the garden pea (*Pisum sativum*) in many of his experiments.

tensive studies in physics, chemistry, mathematics, and various aspects of biology at the University of Vienna. His work in physics and mathematics probably led him to apply experimental and quantitative methods to the study of heredity, and these methods were the key ingredients in his success.

Mendel worked out the basic principles of inheritance in plants over a period of about 9 years. His work culminated in a public lecture in 1865 and a detailed written publication in 1866. Mendel's paper appeared in a journal that was received by 120 libraries, and he sent reprinted copies (of which he had obtained 40) to several distinguished scholars. However, his theory was not accepted. In fact, it was ignored.

The chief difficulty was that the most prominent biologists of Mendel's time were not in the habit of thinking in mathematical terms, even the simple terms used by Mendel. Even Charles Darwin, whose theory of evolution by natural selection depended on genetic variation among individuals, failed to understand the significance of Mendel's findings. In fact, Darwin performed breeding experiments on snapdragons similar to Mendel's on peas and got data similar to Mendel's, but he missed the point, still relying on the concept of blending. In addition, Mendel had little credibility as a biologist; indeed, his lowest grades were in biology! Whatever the reasons, Mendel's pioneering paper had no discernible influence on the scientific world for more than 30 years.

Then, in 1900, after meiosis had been observed and described, Mendel's discoveries burst into prominence as a result of independent experiments by three plant geneticists, Hugo DeVries, Carl Correns, and Erich von Tschermak. Each carried out crossing experiments and obtained quantitative data about the progeny; each published his principal findings in 1900; each cited Mendel's 1866 paper. They immediately realized that chromosomes and meiosis provided a physical explanation for the theory that Mendel had proposed to explain the data from his crosses. As we go through Mendel's work, we will describe first his experiments and conclusions, and then the chromosomal explanation of his theories.

Mendel's Experiments and the Laws of Inheritance

That Mendel was able to make his discoveries before the discovery of meiosis was due in part to the methods of experimentation he used. Mendel's work is a fine example of preparation, choice of experimental material, execution, and interpretation. Let's see how he approached each of these steps.

Mendel devised a careful research plan

Mendel chose the garden pea for his studies because of its ease of cultivation, the feasibility of controlled pollination (see Figure 10.1), and the availability of varieties with differing traits. He controlled pollination, and thus fertilization, of his parent plants by manually moving pollen from one plant to another. Thus he knew the parentage of the offspring in his experiments. The pea plants Mendel studied produce male and female sex organs and gametes in the same flower. If untouched, they naturally *self-pollinate*—that is, the female organ of each flower receives pollen from the male organs of the same flower. Mendel made use of this natural phenomenon in some of his experiments.

Mendel began by examining different varieties of peas in a search for heritable characters and traits suitable for study:

- A **character** is an observable feature, such as flower color.
- A **trait** is a particular form of a character, such as white flowers.
- A **heritable** character trait is one that is passed from parent to offspring.

Mendel looked for characters that had well-defined, contrasting alternative traits, such as purple flowers versus white flowers. Furthermore, these traits had to be **true-breeding**, meaning that the observed trait was the only form present for many generations. In other words, peas with white flowers, when crossed with one another, would have to give rise only to progeny with white flowers for many generations; tall plants bred to tall plants would have to produce only tall progeny.

Mendel isolated each of his true-breeding strains by repeated inbreeding (done by crossing of sibling plants that were seemingly identical or by allowing individuals to self-pollinate) and selection. In most of his work, Mendel concentrated on the seven pairs of contrasting traits shown in Table 10.1. Before performing any experimental cross, he made sure that each potential parent was from a true-breeding strain—an essential point in his analysis of his experimental results.

Mendel then collected pollen from one parental strain and placed it onto the stigma (female organ) of flowers of the other strain whose anthers were removed. The plants providing and receiving the pollen were the **parental generation**, designated **P**. In due course, seeds formed and were planted. The seeds and the resulting new plants constituted the **first filial generation**, or $\mathbf{F_1}$. Mendel and his assistants examined each F_1 plant to see which traits it bore and then recorded the number of F_1 plants expressing each trait. In some experiments the F_1 plants were allowed to self-pollinate and produce a **second filial generation**, $\mathbf{F_2}$. Again, each F_2 plant was characterized and counted.

In summary, Mendel devised a well-organized plan of research, pursued it faithfully and carefully, recorded great amounts of quantitative data, and analyzed the numbers he recorded to explain the relative proportions of the different kinds of progeny. His results and the conclusions to which they led are the subject of the next several sections.

10.1 Mendel's Results from Monohybrid Crosses

PARENTAL GENERATION PHENOTYPES		F_2 GENERATION PHENOTYPES			
DOMINANT	RECESSIVE	DOMINANT	RECESSIVE	TOTAL	RATIO
Spherical seeds ×	Wrinkled seeds	5,474	1,850	7,324	2.96:1
Yellow seeds ×	Green seeds	6,022	2,001	8,023	3.01:1
Purple flowers ×	White flowers	705	224	929	3.15:1
Inflated pods ×	Constricted pods	882	299	1,181	2.95:1
Green pods ×	Yellow pods	428	152	580	2.82:1
Axial flowers ×	Terminal flowers	651	207	858	3.14:1
Tall stems (1 m) ×	Dwarf stems (0.3 m)	787	277	1,064	2.84:1

Mendel's experiment 1 examined a monohybrid cross

"Experiment 1" in Mendel's paper involved a **monohybrid cross**—one involving offspring of a cross in which each member of the P generation is true-breeding for a different trait. He took pollen from pea plants of a true-breeding strain with wrinkled seeds and placed it on the stigmas of flowers of a true-breeding strain with spherical seeds (Figure 10.2). He also performed the reciprocal cross by placing pollen from the spherical-seeded strain on the stigmas of flowers of the wrinkled-seeded strain.

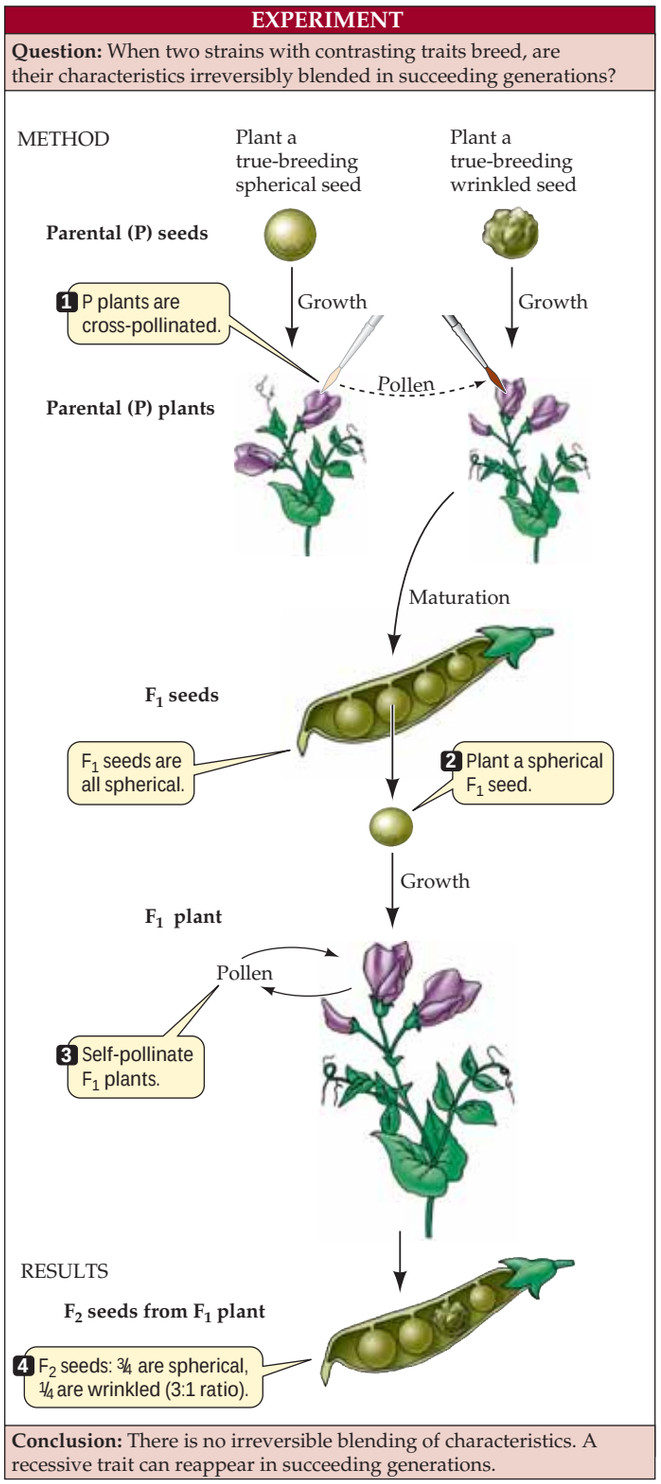
10.2 Contrasting Traits In experiment 1, Mendel studied the inheritance of seed shape. We know today that wrinkled seeds possess an abnormal form of starch.

In both cases, all the F_1 seeds produced were spherical—it was as if the wrinkled seed trait had disappeared completely. The following spring, Mendel grew 253 F_1 plants from these spherical seeds. Each of these plants was allowed to self-pollinate to produce F_2 seeds. In all, there were 7,324 F_2 seeds, of which 5,474 were spherical and 1,850 wrinkled (Figure 10.3 and Table 10.1).

Mendel observed that the wrinkled seed trait was never expressed in the F_1 generation, even though it reappeared in the F_2 generation. He concluded that the spherical seed trait was **dominant** to the wrinkled seed trait, which he called **recessive**. In each of the other six pairs of traits Mendel studied, one proved to be dominant over the other.

Of most importance, the ratio of the two traits in the F_2 generation was always the same—approximately 3:1. That is, three-fourths of the F_2 generation showed the dominant trait and one-fourth showed the recessive trait (see Table 10.1). In Mendel's experiment 1, the ratio was 5,474:1,850 = 2.96:1. The reciprocal crosses in the parental generation both gave similar outcomes in the F_2; it did not matter which parent contributed the pollen.

By themselves, the results from experiment 1 disproved the widely held belief that inheritance is always a blending



10.3 Mendel's Experiment 1 The pattern Mendel observed in the F_2 generation—¼ of the seeds wrinkled, ¾ spherical—was the same no matter which strain contributed the pollen in the parental generation.

phenomenon. According to the blending theory, Mendel's F_1 seeds should have had an appearance intermediate between those of the two parents—in other words, they should have been slightly wrinkled. Furthermore, the blending theory offered no explanation for the reappearance of the wrinkled trait in the F_2 seeds after its apparent absence in the F_1 seeds.

Mendel proposed that the units responsible for the inheritance of specific traits are present as discrete particles that occur in pairs and segregate (separate) from one another during the formation of gametes. According to this theory, the units of inheritance retain their integrity in the presence of other units. This **particulate theory** is in sharp contrast to the blending theory, in which the units of inheritance were believed to lose their identities when mixed together.

As he worked mathematically with his data, Mendel reached the tentative conclusion that each pea plant has two units of inheritance for each character, one from each parent. During the production of gametes, only one of these paired units is given to a gamete. Hence each gamete contains one unit, and the resulting zygote contains two, because it is produced by the fusion of two gametes. This conclusion is the core of Mendel's model of inheritance. Mendel's unit of inheritance is now called a **gene**.

Mendel reasoned that in experiment 1, the two true-breeding parent plants had different forms of the gene affecting seed shape. The spherical-seeded parent had two genes of the same form, which we will call *S*, and the parent with wrinkled seeds had two *s* genes. The *SS* parent produced gametes that each contained a single *S* gene, and the *ss* parent produced gametes each with a single *s* gene. Each member of the F_1 generation had an *S* from one parent and an *s* from the other; an F_1 could thus be described as *Ss*. We say that *S* is dominant over *s* because the trait specified by the *s* allele is not evident when both forms of the gene are present.

The different forms of a gene (*S* and *s* in this case) are called **alleles**. Individuals that are true-breeding for a trait contain two copies of the same allele. For example, all the individuals in a population of a strain of true-breeding peas with wrinkled seeds must have the allele pair *ss;* if *S* were present, the plants would produce spherical seeds.

We say that the individuals that produce wrinkled seeds are **homozygous** for the allele *s*, meaning that they have two copies of the same allele (*ss*). Some peas with spherical seeds—the ones with the genotype *SS*—are also homozygous. However, not all plants with spherical seeds have the *SS* genotype. Some spherical-seeded plants, like Mendel's F_1, are **heterozygous**: They have two different alleles of the gene in question (in this case, *Ss*).

To illustrate these terms with a more complex example, one in which there are three gene pairs, an individual with the genotype *AABbcc* is homozygous for the *A* and *C* genes, because it has two *A* alleles and two *c* alleles, but heterozygous for the *B* gene, because it contains the *B* and *b* alleles. An individual that is homozygous for a character is sometimes called a *homozygote;* a *heterozygote* is heterozygous for the character in question.

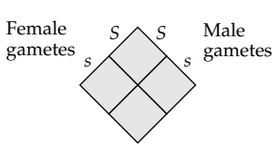
The physical appearance of an organism is its **phenotype**. Mendel correctly supposed the phenotype to be the result of the **genotype**, or genetic constitution, of the organism showing the phenotype. In experiment 1 we are dealing with two phenotypes (spherical seeds and wrinkled seeds). The F_2 generation contains these two phenotypes, but they are produced by three genotypes. The wrinkled seed phenotype is produced only by the genotype *ss*, whereas the spherical seed phenotype may be produced by the genotypes *SS* or *Ss*.

Mendel's first law says that alleles segregate

How does Mendel's model of inheritance explain the composition of the F_2 generation in experiment 1? Consider first the F_1, which has the spherical seed phenotype and the *Ss* genotype. According to Mendel's model, when any individual produces gametes, the two alleles separate, so that each gamete receives only one member of the pair of alleles. This is Mendel's first law, the **law of segregation**.

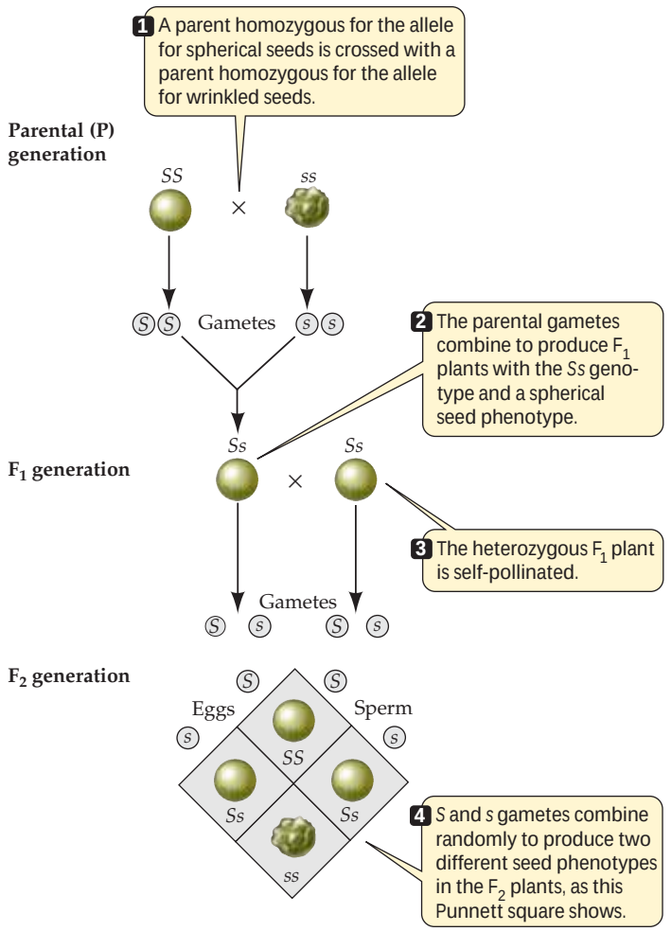
In experiment 1, half the gametes produced by the F_1 generation contained the *S* allele and half the *s* allele. In the F_2 generation, since both *SS* and *Ss* plants produce spherical seeds while *ss* produces wrinkled seeds, there are three ways to get a spherical-seeded plant, but only one way to get a wrinkled-seeded plant (*s* from both parents)—predicting a 3:1 ratio remarkably close to the values Mendel found experimentally for all six of the traits he compared (see Table 10.1).

While this simple example is easy to work out in your head, determination of expected allelic combinations for more complicated inheritance patterns can be aided by use of a **Punnett square**, devised in 1905 by the British geneticist Reginald Crundall Punnett. This device reminds us to consider all possible combinations of gametes when calculating expected genotype frequencies. A Punnett square looks like this:



It is a simple grid with all possible male gamete genotypes shown along one side and all possible female gamete genotypes along another side. To complete the grid, we fill in each square with the corresponding pollen genotype and egg genotype, giving the diploid genotype of a member of the F_2 generation. For example, to fill the rightmost square, we put in the *S* from the egg (female gamete) and the *s* from the pollen (male gamete), yielding *Ss* (Figure 10.4).

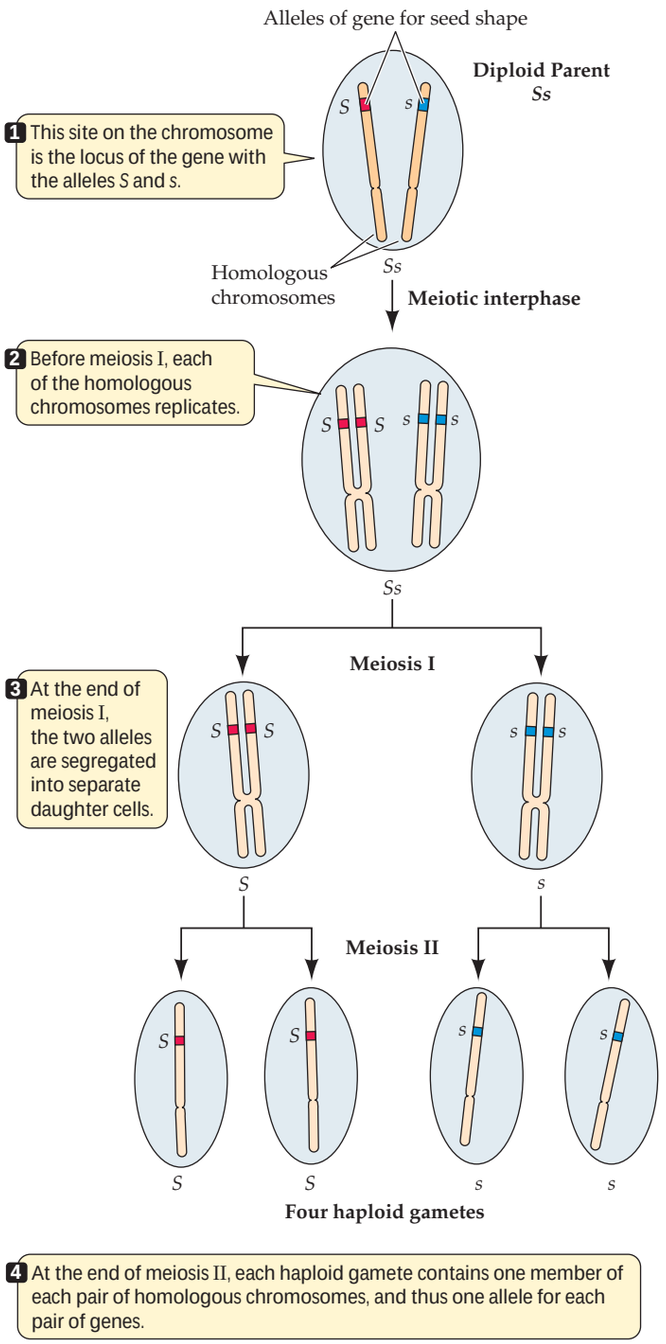
Mendel did not live to see his theory placed on a sound physical footing based on chromosomes and DNA. Genes are now known to be regions of the DNA molecules in chromosomes. More specifically, a gene is a portion of the DNA that resides at a particular site on a chromosome, called a **locus** (plural, **loci**), and encodes a particular character. Genes are expressed in the phenotype mostly as proteins with particular functions, such as enzymes. So a dominant gene can be thought of as a region of DNA that is expressed as a functional enzyme, while a recessive gene typically expresses a nonfunctional enzyme. Mendel arrived at his law of segregation with no knowledge of chromosomes or meiosis, but today we can picture the different alleles of a gene segregating as chromosomes separate in meiosis I (Figure 10.5).



10.4 Mendel's Explanation of Experiment 1 Mendel concluded that inheritance depends on factors from each parent, and that these factors are discrete units that do not blend in the offspring.

Mendel verified his hypothesis by performing a test cross

Mendel set out to test his hypothesis that there were two possible allelic combinations (*SS* and *Ss*) in the spherical-seeded F_1 generation. He did so by performing a **test cross**, which is a way of finding out whether an individual showing a dom-

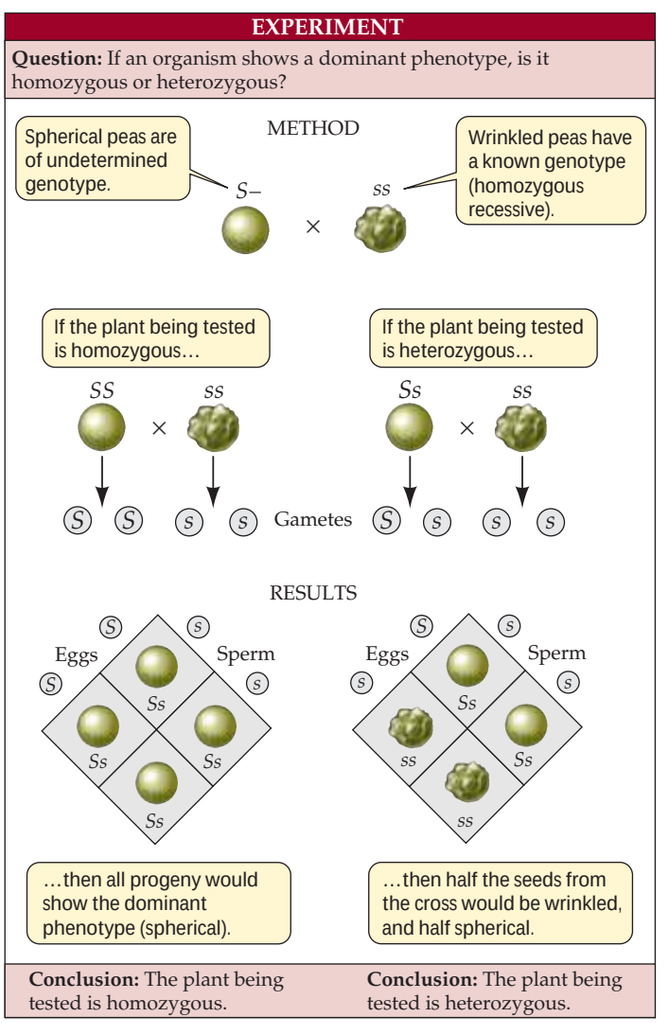


10.5 Meiosis Accounts for the Segregation of Alleles Although Mendel had no knowledge of chromosomes or meiosis, we now know that a pair of alleles resides on homologous chromosomes, and that meiosis segregates those alleles.

inant trait is homozygous or heterozygous. In a test cross, the individual in question is crossed with an individual known to be homozygous for the recessive trait—an easy individual to identify, because in order to have the recessive phenotype, it must be homozygous for the recessive trait.

For the seed shape gene that we have been considering, the recessive homozygote used for the test cross is *ss*. The individual being tested may be described initially as *S*–because we do not yet know the identity of the second allele. We can predict two possible results:

- If the individual being tested is homozygous dominant (*SS*), all offspring of the test cross will be *Ss* and show the dominant trait (spherical seeds).
- If the individual being tested is heterozygous (*Ss*), then approximately half of the offspring of the test cross will be heterozygous and show the dominant trait (*Ss*), but the other half will be homozygous for, and will show, the recessive trait (*ss*) (Figure 10.6).



10.6 Homozygous or Heterozygous? An individual with a dominant phenotype may be homozygous or heterozygous. Its genotype can be determined by crossing it with a homozygous recessive plant and observing the phenotypes of the progeny produced. This procedure is known as a test cross.

The second prediction closely matches the results that Mendel obtained; thus Mendel's hypothesis accurately predicted the results of his test cross.

With his first hypothesis confirmed, Mendel went on to ask another question: How do different pairs of genes behave in crosses when considered together?

Mendel's second law says that alleles of different genes assort independently

Consider an organism that is heterozygous for two genes (*SsYy*), in which the *S* and *Y* alleles came from its mother and *s* and *y* came from its father. When this organism produces gametes, do the alleles of maternal origin (*S* and *Y*) go together to one gamete and those of paternal origin (*s* and *y*) to another gamete? Or can a single gamete receive one maternal and one paternal allele, *S* and *y* (or *s* and *Y*)?

To answer these questions, Mendel performed another series of experiments. He began with peas that differed in two seed characters: seed shape and seed color. One true-breeding parental strain produced only spherical, yellow seeds (*SSYY*), and the other produced only wrinkled, green ones (*ssyy*). A cross between these two strains produced an F_1 generation in which all the plants were *SsYy*. Because the *S* and *Y* alleles are dominant, the F_1 seeds were all spherical and yellow.

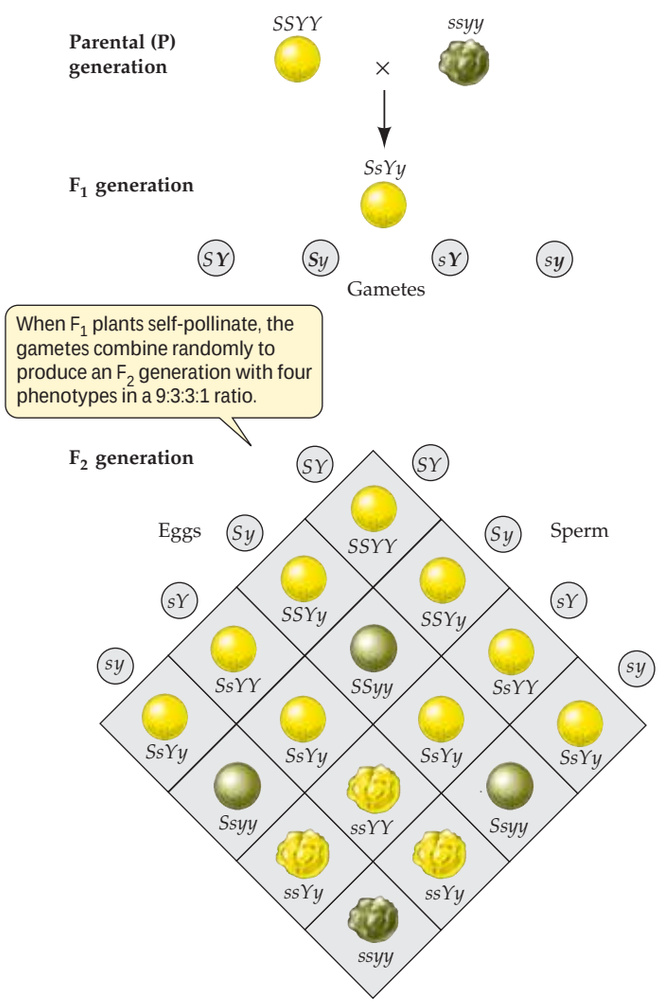
Mendel continued this experiment to the F_2 generation by performing a **dihybrid cross**, which is a cross made between individuals that are identical double heterozygotes. There are two possible ways in which such doubly heterozygous plants might produce gametes, as Mendel saw it. (Remember that he had never heard of chromosomes or meiosis.)

First, if the alleles maintain the associations they had in the parental generation (that is, if they are **linked**), then the F_1 plants should produce two types of gametes (*SY* and *sy*), and the F_2 progeny resulting from self-pollination of the F_1 plants should consist of three times as many plants bearing spherical, yellow seeds as ones with wrinkled, green seeds. Were such results to be obtained, there might be no reason to suppose that seed shape and seed color were regulated by two different genes, because spherical seeds would always be yellow and wrinkled ones always green.

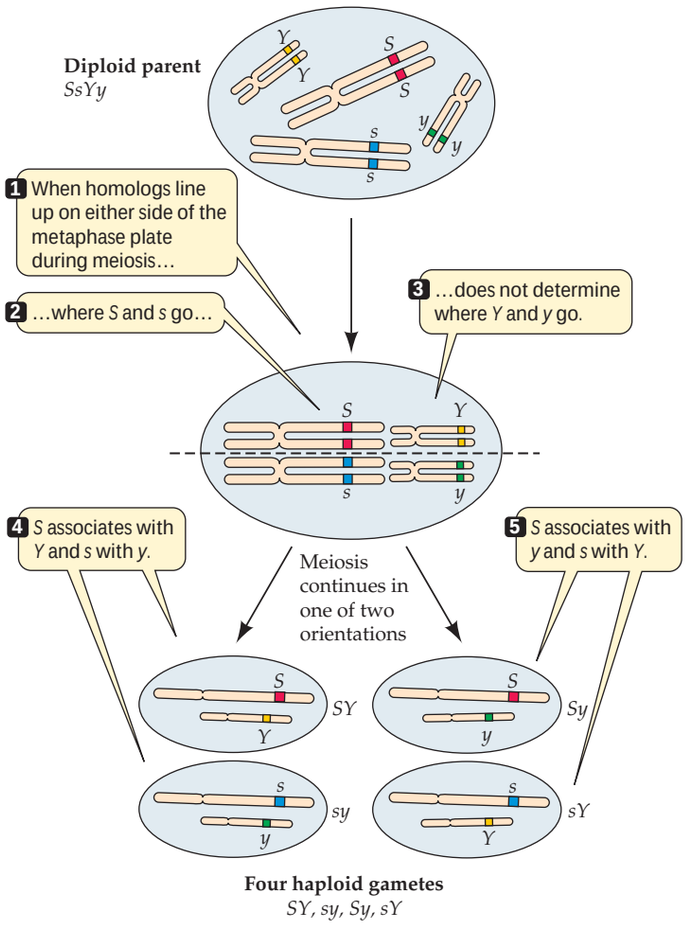
The second possibility is that the segregation of *S* from *s* is independent of the segregation of *Y* from *y* (that is, that the two genes are not linked). In this case, four kinds of gametes should be produced by the F_1 in equal numbers: *SY*, *Sy*, *sY*, and *sy*. When these gametes combine at random, they should produce an F_2 of nine different genotypes. The F_2 progeny could have any of three possible genotypes for shape (*SS*, *Ss*, or *ss*) and any of three possible genotypes for color (*YY*, *Yy*, or *yy*). The combined nine genotypes should produce just four phenotypes (spherical yellow, spherical green, wrinkled yellow, wrinkled green). By using a Punnett square, we can show that these four phenotypes would be expected to occur in a ratio of 9:3:3:1 (Figure 10.7).

The results of Mendel's dihybrid crosses matched the second prediction: Four different phenotypes appeared in the F_2 in a ratio of about 9:3:3:1. The parental traits appeared in new combinations (spherical green and wrinkled yellow). Such new combinations are called **recombinant** phenotypes.

These results led Mendel to the formulation of what is now known as Mendel's second law: Alleles of different genes assort independently of one another during gamete formation. That is, the segregation of the alleles of gene A is independent of the segregation of the alleles of gene B. We now know that this **law of independent assortment** is not as universal as the law of segregation, because it applies to genes located on separate chromosomes, but not necessarily to those located



10.7 Independent Assortment The 16 possible combinations of gametes in this dihybrid cross result in 9 different genotypes. Because *S* and *Y* are dominant over *s* and *y*, respectively, the 9 genotypes result in 4 phenotypes in a ratio of 9:3:3:1. These results show that the two genes segregate independently.



10.8 Meiosis Accounts for Independent Assortment of Alleles We now know that alleles of different genes are segregated independently during metaphase I of meiosis. Thus a parent of genotype SsYy can form gametes with four different genotypes.

on the same chromosome, as we will see below. However, it is correct to say that chromosomes segregate independently during the formation of gametes, and so do any two genes on separate homologous chromosome pairs (Figure 10.8).

One of Mendel's major contributions to the science of genetics was his use of the rules of statistics and probability to analyze his masses of data from hundreds of crosses producing thousands of plants. His mathematical analyses led to clear patterns in the data, and then to his hypotheses. Ever since Mendel, geneticists have used simple mathematics in the same ways that Mendel did.

Punnett squares or probability calculations: A choice of methods

Punnett squares provide one way of solving problems in genetics, and probability calculations provide another. Many people find it easiest to use the principles of probability, perhaps because they are so familiar. When we flip a coin, the law of probability states that it has an equal probability of landing "heads" or "tails." For any given toss of a fair coin, the probability of heads is independent of what happened in all the previous tosses. A run of ten straight heads implies nothing about the next toss. No "law of averages" increases the likelihood that the next toss will come up tails, and no "momentum" makes an eleventh occurrence of heads any more likely. On the eleventh toss, the odds of getting heads are still 50/50.

The basic conventions of probability are simple:

- If an event is absolutely certain to happen, its probability is 1.
- If it cannot possibly happen, its probability is 0.
- Otherwise, its probability lies between 0 and 1.

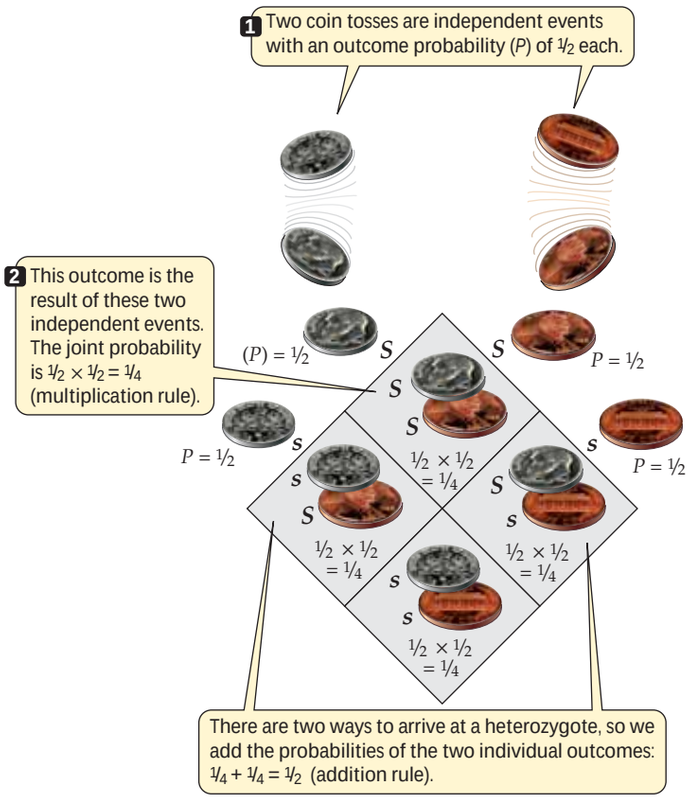
A coin toss results in heads approximately half the time, so the probability of heads is ½—as is the probability of tails.

MULTIPLYING PROBABILITIES. How can we determine the probability of two independent events happening together? If two coins (a penny and a dime, say) are tossed, each acts independently of the other. What, then, is the probability of both coins coming up heads? Half the time, the penny comes up heads; of that fraction, half the time the dime also comes up heads. Therefore, the joint probability of both coins coming up heads is half of one-half, or ½ × ½ = ¼. To find the joint probability of independent events, then, we multiply the probabilities of the individual events (Figure 10.9). How does this method apply to genetics?

THE MONOHYBRID CROSS. To apply the principles of probability to genetics problems, we need only deal with gamete formation and random fertilization instead of coin tosses. A homozygote can produce only one type of gamete, so, for example, the probability of an *SS* individual producing gametes with the genotype *S* is 1. The heterozygote *Ss* produces *S* gametes with a probability of ½, and *s* gametes with a probability of ½.

Consider the F_2 progeny of the cross in Figure 10.4. They are obtained by self-pollination of F_1 plants of genotype *Ss*. The probability that an F_2 plant will have the genotype *SS* must be ½ × ½ = ¼, because there is a 50:50 chance that the sperm will have the genotype *S*, and that chance is independent of the 50:50 chance that the egg will have the genotype *S*. Similarly, the probability of *ss* offspring is ½ × ½ = ¼.

ADDING PROBABILITIES. How are probabilities calculated when an event can happen in different ways? The probability of an F_2 plant getting an *S* allele from the sperm and an *s* allele from the egg is ¼, but remember that the same



10.9 Using Probability Calculations in Genetics The probability of any given combination of alleles from a sperm and an egg appearing in the offspring of a cross can be obtained by multiplying the probabilities of each event. Since a heterozygote can be formed in two ways, these two probabilities are added together.

genotype can also result from an *s* from the sperm and an *S* from the egg, also with a probability of ¼. The probability of an event that can occur in two or more different ways is the sum of the individual probabilities of those ways. Thus the probability that an F_2 plant will be a heterozygote is equal to the sum of the probabilities of the two ways of forming a heterozygote: ¼ + ¼ = ½ (see Figure 10.9). The three genotypes are therefore expected in the ratio ¼ *SS* : ½ *Ss* : ¼ *ss*—hence the 1:2:1 ratio of genotypes and the 3:1 ratio of phenotypes seen in Figure 10.4.

THE DIHYBRID CROSS. If F_1 plants heterozygous for two independent characters self-pollinate, the resulting F_2 plants express four different phenotypes. The proportions of these phenotypes are easily determined by probability calculations. Let's see how this works for the experiment shown in Figure 10.7.

Using the principle described above, we can calculate that the probability that an F_2 seed will be spherical is ¾: the probability of an *Ss* heterozygote (½) plus the probability of an *SS* homozygote (¼) = ¾. By the same reasoning, the probability that a seed will be yellow is also ¾. The two characters are determined by separate genes and are independent of each other, so the joint probability that a seed will be both spherical and yellow is ¾ × ¾ = 9/16. What is the probability of F_2 seeds being both wrinkled and yellow? The probability of being yellow is again ¾; the probability of being wrinkled is ½ × ½ = ¼. The joint probability that a seed will be both wrinkled and yellow, then, is ¼ × ¾ = 3/16. The same probability applies, for similar reasons, to spherical, green F_2 seeds. Finally, the probability that F_2 seeds will be both wrinkled and green is ¼ × ¼ = 1/16. Looking at all four phenotypes, we see they are expected in the ratio of 9:3:3:1.

Probability calculations and Punnett squares give the same results. Learn to do genetics problems both ways, and then decide which method you prefer.

Mendel's laws can be observed in human pedigrees

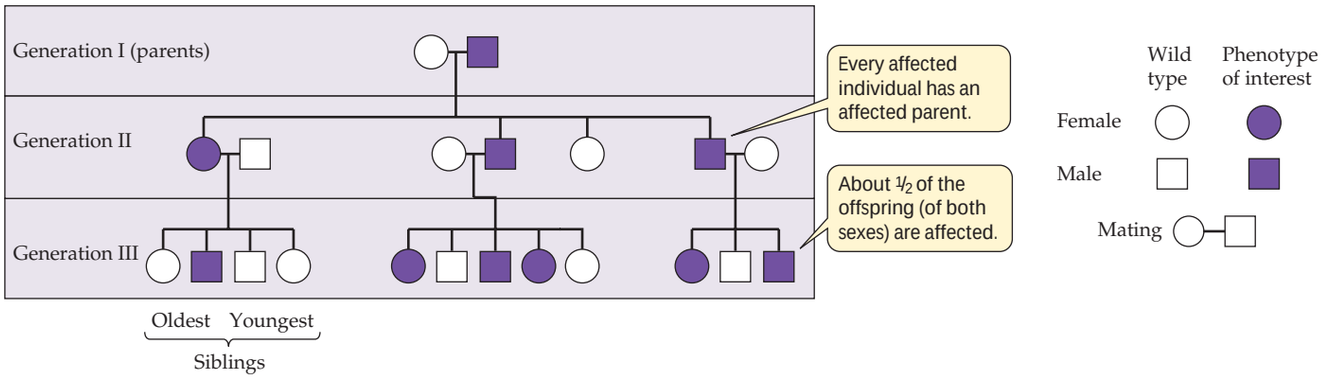
After Mendel's work was uncovered by plant breeders, Mendelian inheritance was observed in humans. Currently, the patterns of over 2,500 inherited human characteristics have been described.

How can Mendel's laws of inheritance be applied to humans? Mendel worked out his laws by performing many planned crosses and counting many offspring. Neither of these approaches is possible with humans. So human geneticists rely on **pedigrees**, family trees that show the occurrence of phenotypes (and alleles) in several generations of related individuals.

Because humans have such small numbers of offspring, human pedigrees do not show the clear proportions of offspring phenotypes that Mendel saw in his pea plants (see Table 10.1). For example, when two people who are both heterozygous for a recessive allele (say, *Aa*) marry, there will be, for each of their children, a 25 percent probability that the child will be a recessive homozygote (*aa*). Thus, over many such marriages, one-fourth of all the children will be recessive homozygotes (*aa*). But the offspring of a single marriage are likely to be too few to show the exact one-fourth proportion. In a family with only two children, for example, both could easily be *aa* (or *Aa*, or *AA*).

To deal with this ambiguity, human geneticists assume that any allele that causes an abnormal phenotype is rare in the human population. This means that if some members of a given family have a rare allele, it is highly unlikely that an outsider marrying into that family will have that same rare allele.

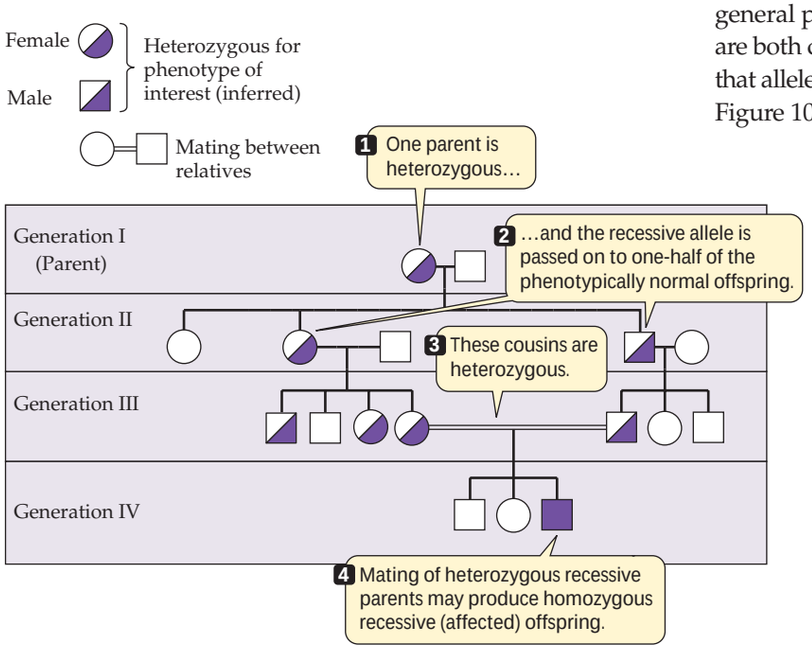
Human geneticists may wish to know whether a particular rare allele is dominant or recessive. Figure 10.10 depicts a pedigree showing the pattern of inheritance of a rare domi-



10.10 Pedigree Analysis and Dominant Inheritance This pedigree represents a family affected by Huntington's disease, which results from a rare dominant allele. Everyone who inherits this allele is affected.

nant allele. The following are the key features to look for in such a pedigree:

- Every affected person has an affected parent.
- About half of the offspring of an affected parent are also affected.
- The phenotype occurs equally in both sexes.



10.11 Recessive Inheritance This pedigree represents a family that carries the allele for albinism, a recessive trait. Because the trait is recessive, heterozygotes do not have the albino phenotype, but they can pass the allele on to their offspring. Affected persons must inherit the allele from two heterozygous parents or (rarely) from one homozygous and one heterozygous parent. In this family, the heterozygous parents are cousins, but the same result could occur if the parents were unrelated but heterozygous.

Compare this pattern with Figure 10.11, which shows the pattern of inheritance of a rare recessive allele:

- Affected people usually have two parents who are not affected.
- In affected families, about one-fourth of the children of unaffected parents can be affected.
- The phenotype occurs equally in both sexes.

In pedigrees showing inheritance of a recessive phenotype, it is not uncommon to find a marriage of two relatives. This pattern is a result of the rarity of recessive alleles that give rise to abnormal phenotypes. For two phenotypically normal parents to have an affected child (*aa*), the parents must both be heterozygous (*Aa*). If a particular recessive allele is rare in the general population, the chance of two people marrying who are both carrying that allele is quite low. On the other hand, if that allele is present in a family, two cousins might share it (see Figure 10.10). This is why studies on populations isolated either culturally (by religion, as with the Amish in the United States) or geographically (as on islands) have been so valuable to human geneticists. People in these groups tend to have large families, or to marry among themselves, or both.

Because the major use of pedigree analysis is in the clinical evaluation and counseling of patients with inherited abnormalities, a single pair of alleles is usually followed. However, just as pedigree analysis shows the segregation of alleles, it also can show independent assortment if two different allele pairs are considered.

Alleles and Their Interactions

In many cases, alleles do not show the simple relationships between dominance and recessiveness that we have described. In others, a single allele may have multiple phenotypic effects. Existing alleles can give rise to new alleles by mutation, so there can be many alleles for a single character.

New alleles arise by mutation

Different alleles of a gene exist because genes are subject to **mutations**, which are rare, stable, and inherited changes in the genetic material. In other words, an allele can mutate to become a different allele. Mutation, which will be discussed in detail in Chapter 12, is a random process; different copies of the same allele may be changed in different ways.

One particular allele of a gene may be defined as the **wild type**, because it is present in most individuals in nature ("the wild") and gives rise to an expected trait or phenotype. Other alleles of that gene, often called *mutant alleles*, may produce a different phenotype. The wild-type and mutant alleles reside at the same locus and are inherited according to the rules set forth by Mendel. A genetic locus with a wild-type allele that is present less than 99 percent of the time (the rest of the alleles being mutant) is said to be **polymorphic** (from the Greek *poly*, "many," and *morph*, "form").

Many genes have multiple alleles

Because of random mutations, a group of individuals may have more than two alleles of a given gene. (Any one individual has only two alleles, of course—one from its mother and one from its father.) In fact, there are many examples of such multiple alleles.

Coat color in rabbits is determined by one gene with four alleles. There is a dominance hierarchy among these alleles:

$$C > c^{ch} > c^{h} > c$$

Any rabbit with the *C* allele (paired with any of the four) is dark gray, and a rabbit with *cc* is albino. The intermediate colors result from the different allelic combinations shown in Figure 10.12.

Multiple alleles increase the number of possible phenotypes. In Mendel's monohybrid cross, there was just one pair of alleles (*Ss*) and two possible phenotypes (resulting from *SS* or *Ss* and *ss*). The four alleles of the rabbit coat color gene produce five phenotypes.

Dominance is not always complete

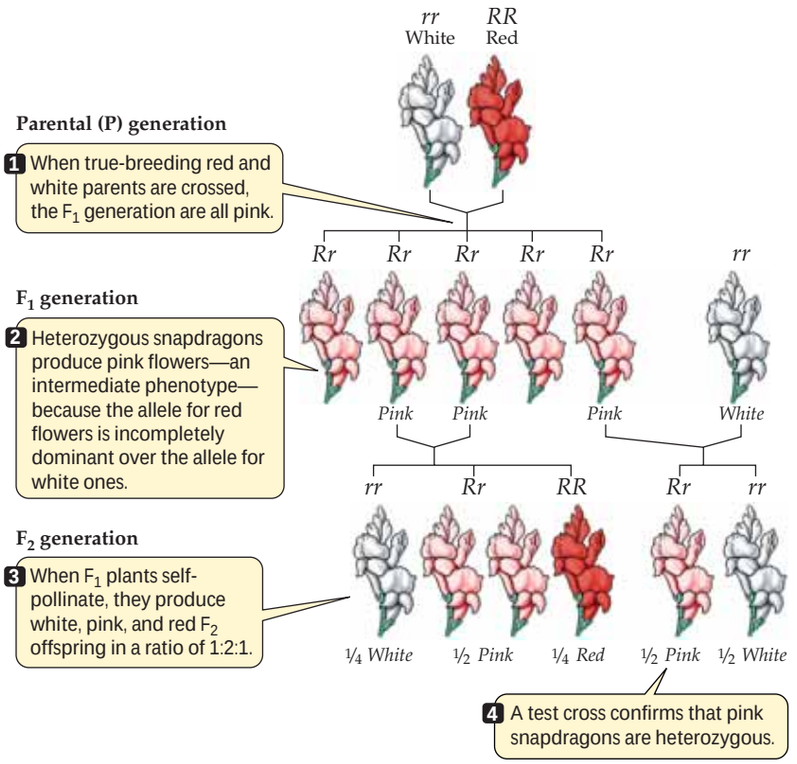
In the single pairs of alleles studied by Mendel, dominance is complete when an individual is heterozygous. That is, an *Ss* individual expresses the *S* phenotype. However, many genes have alleles that are not dominant or recessive to one another. Instead, the heterozygotes show an intermediate phenotype—at first glance, like that predicted by the old blending theory of inheritance. For example, if a true-breeding red snapdragon is crossed with a true-breeding white one, all the F_1 flowers are pink. That this phenomenon can still be explained in terms of Mendelian genetics, rather than blending, is readily demonstrated by a further cross.

The blending theory predicts that if one of the pink F_1 snapdragons is crossed with a true-breeding white one, all the offspring should be a still lighter pink. In fact, approximately ½ of the offspring are white, and ½ are the same shade of pink as the F_1 parent. When the F_1 pink snapdragons are allowed to self-pollinate, the resulting F_2 plants are distributed in a ratio of 1 red:2 pink:1 white (Figure 10.13). Clearly the hereditary particles—the genes—have not blended; they are readily sorted out in the F_2.

We can understand these results in terms of the Mendelian laws of inheritance. All we need to do is recognize that the heterozygotes show a phenotype intermediate between those of the two homozygotes. In such cases, the gene is said to be governed by **incomplete dominance**. Incomplete dominance is common in nature. In fact, Mendel's paper was unusual in that all seven of the examples he described (see Table 10.1) are characterized by complete dominance.

10.12 Inheritance of Coat Color in Rabbits There are four alleles of the gene for coat color in rabbits. Different combinations of two alleles give different coat colors.

Possible genotypes	CC, Cc^{ch}, Cc^{h}, Cc	$c^{ch}c^{ch}$	$c^{ch}c^{h}$, $c^{ch}c$	$c^{h}c^{h}$, $c^{h}c$	cc
Phenotype	Dark gray	Chinchilla	Light gray	Himalayan	Albino



10.13 Incomplete Dominance Follows Mendel's Laws An intermediate phenotype can occur in heterozygotes when neither allele is dominant. The heterozygous phenotype (here, pink flowers) may give the appearance of a blended trait, but the traits of the parental generation reappear in their original forms in succeeding generations, as predicted by Mendel's laws of inheritance.

Blood type of cells	Genotype	Antibodies made by body	Reaction to added antibodies	
			Anti-A	Anti-B
A	I^AI^A or I^Ai^O	Anti-B		
B	I^BI^B or I^Bi^O	Anti-A		
AB	I^AI^B	Neither anti-A nor anti-B		
O	i^Oi^O	Both anti-A and anti-B		

Red blood cells that do not react with antibody remain evenly dispersed.

Red blood cells that react with antibody clump together (speckled appearance).

10.14 ABO Blood Reactions Are Important in Transfusions Red blood cells of types A, B, AB, and O were mixed with serum containing anti-A or anti-B antibodies. As you look down the columns, note that each of the types, when mixed separately with anti-A and with anti-B, gives a unique pair of results; this is the basic method by which blood is typed. People with type O blood are good blood donors because O cells do not react with either anti-A or anti-B antibodies. People with type AB blood are good recipients, since they make neither type of antibody.

In codominance, both alleles are expressed

Sometimes the two alleles at a locus produce two different phenotypes that both appear in heterozygotes. An example of this phenomenon, called **codominance**, is seen in the ABO blood group system in humans.

Early attempts at blood transfusion frequently killed the patient. Around 1900, the Austrian scientist Karl Landsteiner mixed blood cells and serum (blood from which cells have been removed) from different individuals. He found that only certain combinations of blood are compatible. In other combinations, the red blood cells from one individual form clumps in the presence of serum from the other individual. This discovery led to our ability to administer compatible blood transfusions that do not kill the recipient.

Clumps form in incompatible transfusions because specific proteins in the serum, called *antibodies*, react with foreign, or "nonself," cells. The antibodies react with proteins on the surface of nonself cells, called *antigens*. Blood compatibility is determined by a set of three alleles (I^A, I^B, and i^O) at one locus, which determine the antigens on the surface of red blood cells. Different combinations of these alleles in different people produce four different blood types, or phenotypes: A, B, AB, and O (Figure 10.14). The AB phenotype found in individuals of I^AI^B genotype is an example of codominance—these individuals produce cell surface antigens of both the A and B types.

Some alleles have multiple phenotypic effects

Mendel's principles were further extended when it was discovered that a single allele can result in more than one phenotype. When a single allele has more than one distinguishable phenotypic effect, we say that the allele is **pleiotropic**. A familiar example of pleiotropy involves the allele responsible for the coloration pattern (light body, darker extremities) of Siamese cats, discussed later in this chapter. The same allele is also responsible for the characteristic crossed eyes of Siamese cats. Although these effects appear to be unrelated, both result from the same protein produced under the influence of the allele.

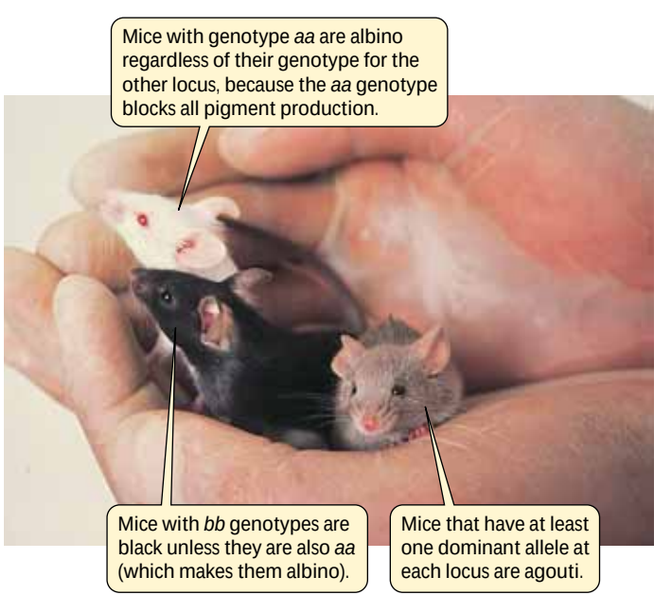
Gene Interactions

Thus far we have treated the phenotype of an organism, with respect to a given character, as a simple result of the alleles of a single gene. In many cases, however, several genes interact to determine a phenotype. To complicate things further, the physical environment may interact with the genetic constitution of an individual in determining the phenotype.

Some genes alter the effects of other genes

Epistasis occurs when the phenotypic expression of one gene is affected by another gene. For example, several genes determine coat color in mice. The wild-type color is agouti, a grayish pattern resulting from bands on the individual hairs. The dominant allele *B* determines that the hairs will have bands and thus that the color will be agouti, whereas the homozygous recessive genotype *bb* results in unbanded hairs, and the mouse appears black. A second locus, on another chromosome, affects an early step in the formation of hair pigments. The dominant allele *A* at this locus allows normal color development, but *aa* blocks all pigment production. Thus, *aa* mice are all-white albinos, irrespective of their genotype at the *B* locus (Figure 10.15).

If a mouse with genotype *AABB* (and thus the agouti phenotype) is crossed with an albino of genotype *aabb*, the F_1 mice are *AaBb* and have the agouti phenotype. If the F_1 mice are crossed with each other to produce an F_2 generation, then epistasis will result in an expected phenotypic ratio of 9 agouti:3 black:4 albino. (Can you show why? The underlying ratio is the usual 9:3:3:1 for a dihybrid cross with unlinked genes, but look closely at each genotype, and watch out for epistasis.)



10.15 Genes May Interact Epistatically Epistasis occurs when one gene alters the phenotypic effect of another gene. In these mice, the presence of the recessive genotype (*aa*) at one locus blocks pigment production, producing an albino mouse no matter what the genotype is at the second locus.

In another form of epistasis, two genes are mutually dependent: The expression of each depends on the alleles of the other. The epistatic action of such **complementary genes** may be explained as follows: Suppose gene *A* codes for enzyme A in the metabolic pathway for purple pigment in flowers, and gene *B* codes for enzyme B:

colorless precursor $\xrightarrow{\text{enzyme A}}$ colorless intermediate $\xrightarrow{\text{enzyme B}}$ purple pigment

In order for the pigment to be produced, both reactions must take place. The recessive alleles *a* and *b* code for nonfunctional enzymes. If a plant is homozygous for either *a* or *b*, the corresponding reaction will not occur, no purple pigment will form, and the flowers will be white.

Hybrid vigor results from new gene combinations and interactions

If Mendel's paper was the most important event in genetics in the nineteenth century, perhaps an equally important paper in applied genetics was published early in the twentieth century by G. H. Shull, titled "The composition of a field of maize." Farmers growing crops have known for centuries that mating among close relatives (known as *inbreeding*) can result in offspring of lower quality than those from matings between unrelated individuals. The reason for this is that close relatives tend to have the same recessive alleles, some of which may be harmful, as we saw in our discussion of human pedigrees above. In fact, it has long been known that if one crosses two true-breeding, homozygous genetic strains of a plant or animal, the result is offspring that are phenotypically much stronger, larger, and in general more "vigorous" than either of the parents (Figure 10.16).

Shull began his experiment with two of the thousands of existing varieties of corn (maize). Both varieties produced about 20 bushels of corn per acre. But when he crossed them, the yield of their offspring was an astonishing 80 bushels per acre. This phenomenon is known as **heterosis** (short for heterozygosis), or **hybrid vigor**. The cultivation of hybrid corn spread rapidly in the United States and all over the world, quadrupling grain production. The practice of hybridization has spread to many other crops and animals used in agriculture.

The actual mechanism by which heterosis works is not known. A widely accepted hypothesis is *overdominance*, in

10.16 Hybrid Vigor in Corn The heterozygous F_1 offspring is larger and more vigorous than either homozygous parent.

which the heterozygous condition in certain important genes is superior to either homozygote.

The environment affects gene action

The phenotype of an individual does not result from its genotype alone. Genotype and environment interact to determine the phenotype of an organism. Environmental variables such as light, temperature, and nutrition can affect the translation of a genotype into a phenotype.

A familiar example of this phenomenon involves the Siamese cat. This handsome animal normally has darker fur on its ears, nose, paws, and tail than on the rest of its body. These darkened extremities normally have a lower temperature than the rest of the body. A few simple experiments show that the Siamese cat has a genotype that results in dark fur, but only at temperatures below the general body temperature. If some dark fur is removed from the tail and the cat is kept at higher than usual temperatures, the new fur that grows in is light. Conversely, removal of light fur from the back, followed by local chilling of the area, causes the spot to fill in with dark fur.

Two parameters describe the effects of genes and environment on the phenotype:

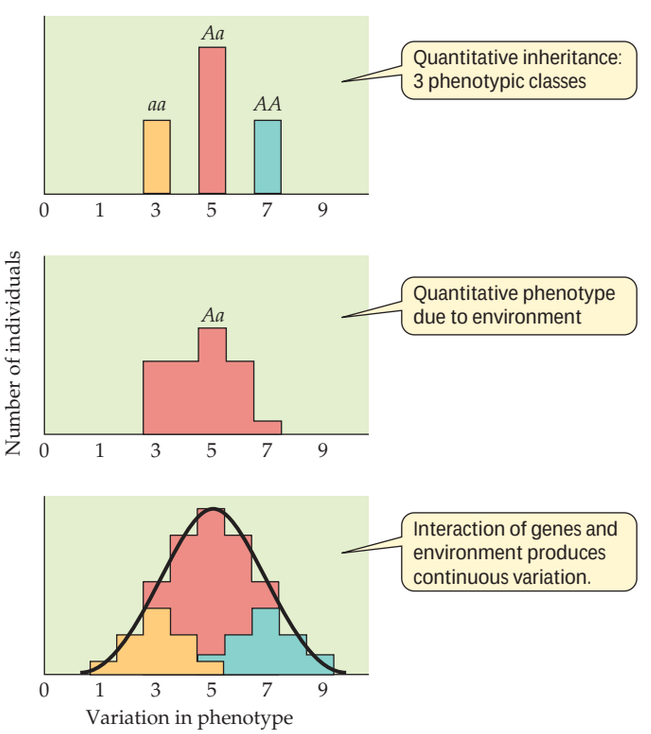
- **Penetrance** is the proportion of individuals in a group with a given genotype that actually show the expected phenotype.
- **Expressivity** is the degree to which a genotype is expressed in an individual.

For an example of environmental effects on expressivity, consider how Siamese cats kept indoors or outdoors in different climates might look.

Most complex phenotypes are determined by multiple genes and environment

The differences between individual organisms in simple characters, such as those that Mendel studied in peas, are discrete and **qualitative**. For example, the individuals in a population of peas are either short or tall. For most complex characters, however, such as height in humans, the phenotype varies more or less continuously over a range. Some people are short, others are tall, and many are in between the two extremes. Such variation within a population is called **quantitative**, or **continuous**, variation. In most cases, quantitative variation is due to two factors (Figure 10.17): multiple genes, each with multiple alleles, and environmental influences on the expression of these genes.

Geneticists call the genes that together determine a complex character **quantitative trait loci**. Identifying these loci is a major challenge, and an important one. For example, the amount of grain that a variety of rice produces in a growing season is determined by many interacting genetic factors. Crop plant breeders have worked hard to decipher these fac-



10.17 Quantitative Variation Quantitative variation is produced by the interaction of genes and environment. In this illustration, only a single gene with three alleles is considered. Most complex characters are determined by many genes and alleles, with the environment exerting an influence on each.

tors in order to breed higher-yielding rice strains. In a similar way, human characteristics such as disease susceptibility and behavior are caused in part by quantitative trait loci.

Genes and Chromosomes

The recognition that genes occupy characteristic positions on chromosomes and are segregated by meiosis enabled Mendel's successors to provide a physical explanation for his model of inheritance. It soon became apparent that the association of genes with chromosomes has other genetic consequences as well. We mentioned above that genes located on the same chromosome may not follow Mendel's law of independent assortment. What is the pattern of inheritance of such genes? How do we determine where genes are located on a chromosome, and the distances between them?

The answers to these and many other genetic questions were worked out in studies of the fruit fly *Drosophila melanogaster*. Its small size, its ease of cultivation, and its short generation time made this animal an attractive experimental subject. Beginning in 1909, Thomas Hunt Morgan and his students pioneered the study of *Drosophila* in Columbia University's famous "fly room," where they discovered the phenomena described in this section. *Drosophila* remains extremely important in studies of chromosome structure, population genetics, the genetics of development, and the genetics of behavior.

Genes on the same chromosome are linked

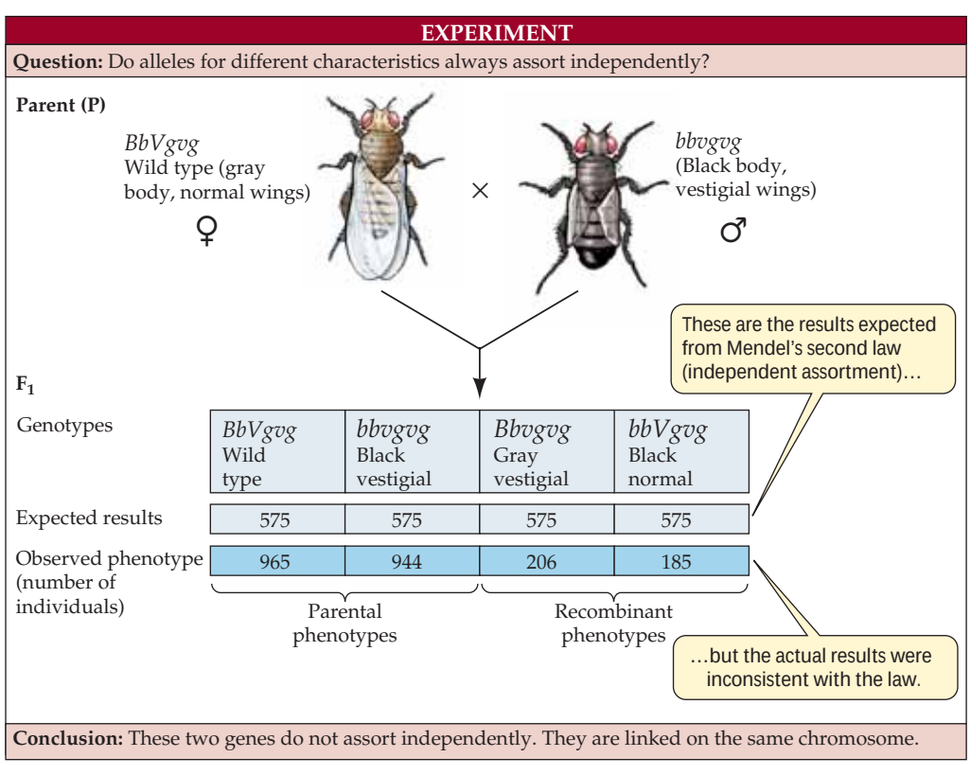
Some of the crosses Morgan performed with fruit flies resulted in phenotypic ratios that were not in accord with those predicted by Mendel's law of independent assortment. Morgan crossed *Drosophila* of two known genotypes, *BbVgvg* × *bbvgvg*, for two different characters, body color and wing shape:

- *B* (wild-type gray body), is dominant over *b* (black body)
- *Vg* (wild-type wing) is dominant over *vg* (vestigial, a very small wing)

(Do you recognize this type of cross? It is a test cross for the two gene pairs; see Figure 10.6.)

Morgan expected to see four phenotypes in a ratio of 1:1:1:1, but that is not what he observed. The body color gene and the wing size gene were not assorting independently; rather, they were for the most part inherited together (Figure 10.18).

These results became understandable to Morgan when he assumed that the two loci are on the same chromosome—that is, that they are linked. After all, since the number of genes in a cell far exceeds the number of chromosomes, each chromosome must contain many genes. The full set of loci on a given chromosome constitutes a *linkage group*. The number of linkage groups in a species equals the number of homologous chromosome pairs.



10.18 Some Alleles Do Not Assort Independently Morgan's studies showed that the genes for body color and wing size in *Drosophila* are linked, so their alleles do not assort independently. Linkage accounts for the departure of the phenotype ratios observed from the results predicted by Mendel's law of independent assortment.

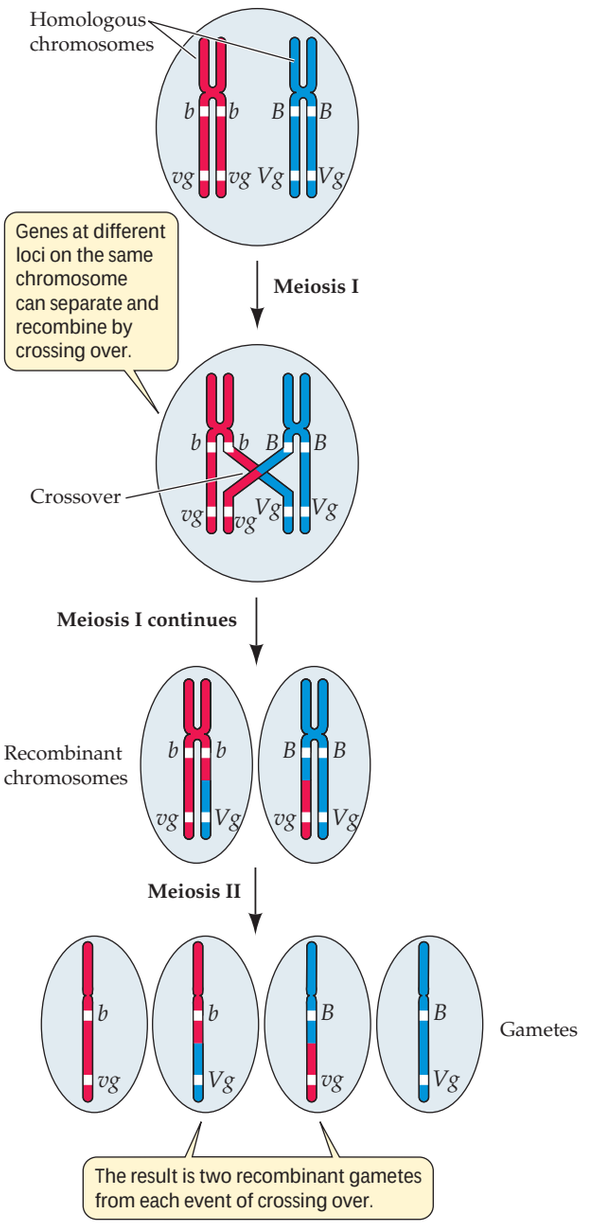
Suppose, now, that the *Bb* and *Vgvg* loci are indeed located on the same chromosome. Why, then, didn't all of Morgan's F_1 flies have the parental phenotypes—that is, why did his cross result in anything other than gray flies with normal wings (wild-type) and black flies with vestigial wings? If we assumed that linkage is *absolute*—that is, that chromosomes always remain intact and unchanged—we would expect to see just those two types of progeny. However, this is not always what happens.

Genes can be exchanged between chromatids

Absolute linkage is extremely rare. If linkage were absolute, Mendel's law of independent assortment would apply only to loci on different chromosomes. What actually happens is more complex, and therefore more interesting. Chromosomes are not unbreakable, so recombination of genes can occur. That is, genes at different loci on the same chromosome do sometimes separate from one another during meiosis.

Genes may recombine when two homologous chromosomes physically exchange corresponding segments during prophase I of meiosis—that is, by crossing over (Figure 10.19; see also Figure 9.16). Recall from Chapter 9 that the DNA is replicated during the S phase, so that by prophase I, when homologous chromosome pairs come together to form tetrads, each chromosome consists of two chromatids. The exchange event involves only two of the four chromatids in a tetrad, one from each member of the homologous pair, and can occur at any point along the length of the chromosome. The chromosome segments involved are exchanged reciprocally, so both chromatids involved in crossing over become recombinant (that is, each chromatid ends up with genes from both of the organism's parents). Usually several exchange events occur along the length of each homologous pair.

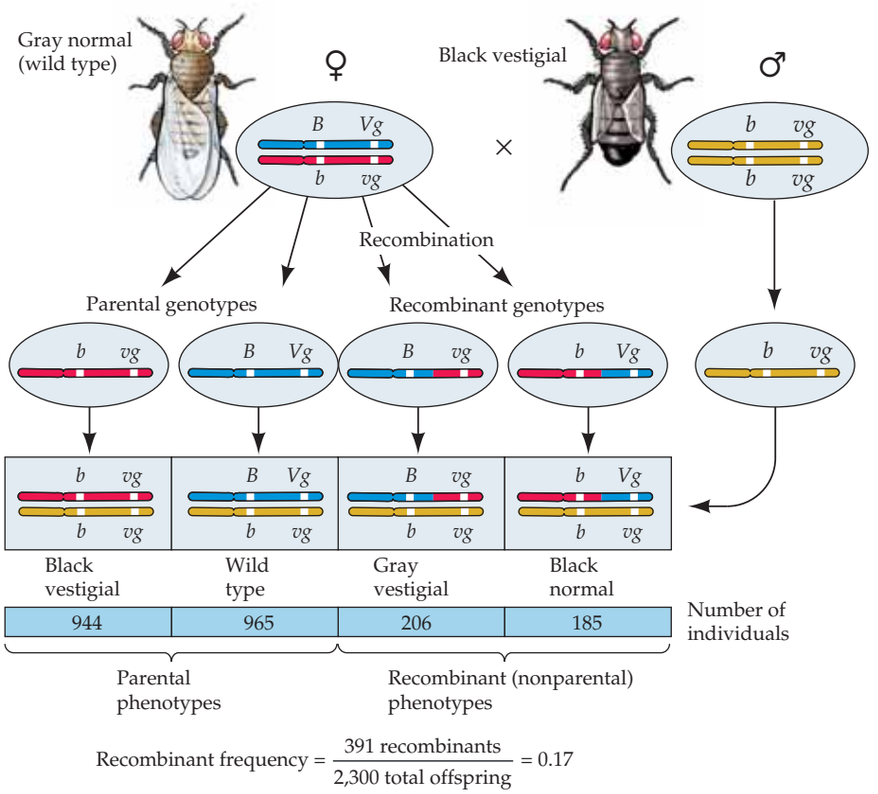
When crossing over takes place between two linked genes, not all progeny of a cross will have the parental phenotypes. Instead, recombinant offspring appear as well, as they did in Morgan's cross. They appear in proportions called **recombinant frequencies**, which are calculated by dividing the number of recombinant progeny by the total number of progeny (Figure 10.20). Recombinant frequencies will be greater for loci that are farther apart on the chromosome than for loci that are closer together, because an exchange event is more likely to occur between genes that are far apart than between genes that are close together.



10.19 Crossing Over Results in Genetic Recombination Genes at different loci on the same chromosome can be separated from one another and recombined by crossing over. Such recombination occurs during prophase I of meiosis.

Geneticists can make maps of chromosomes

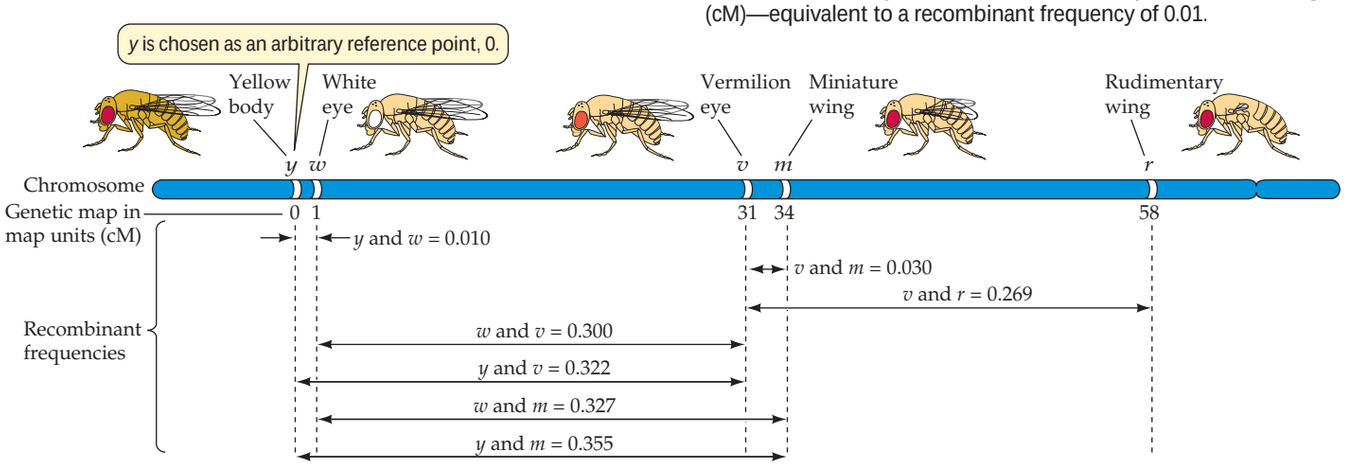
If two loci are very close together on a chromosome, the odds of crossing over between them are small. In contrast, if two loci are far apart, crossing over could occur between them at many points. In a population of cells undergoing meiosis, a greater proportion of the cells will undergo recombination between two loci that are far apart than between two loci that are close together. In 1911, Alfred Sturtevant, then an undergraduate student in T. H. Morgan's fly room, realized how that simple insight could be used to show where different genes lie on a chromosome in relation to one another.



10.20 Recombinant Frequencies The frequency of recombinant offspring (those with a phenotype different from either parent) can be calculated. Recombinant frequencies will be larger for loci that are far apart than for those that are close together on the chromosome.

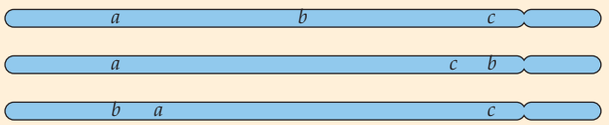
The Morgan group had determined recombinant frequencies for many pairs of linked genes. Sturtevant used these recombinant frequencies to create **genetic maps** that showed the arrangement of genes along the chromosome (Figure 10.21). Ever since Sturtevant demonstrated this method, geneticists have mapped the chromosomes of eukaryotes, prokaryotes, and viruses, assigning distances between genes in **map units**. A map unit corresponds to a recombinant frequency of 0.01; it is also referred to as a *centimorgan* (cM), in honor of the founder of the fly room. You, too, can work out a genetic map (Figure 10.22).

10.21 Steps toward a Genetic Map Because the chance of a recombinant genotype occurring increases with the distance between two loci on a chromosome, Sturtevant was able to derive this partial map of a *Drosophila* chromosome from the Morgan group's data on the recombinant frequencies of five recessive traits. He used an arbitrary unit of distance—the map unit, or centimorgan (cM)—equivalent to a recombinant frequency of 0.01.



10.22 Map These Genes The object of this exercise is to determine the order of three loci (*a*, *b*, and *c*) on a chromosome, as well as the map distances (in cM) between them.

1 At the outset, we have no idea of the individual distances, and there are several possible sequences (*a-b-c*, *a-c-b*, *b-a-c*).

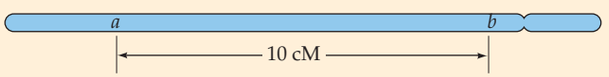


We make a cross *AABB* × *aabb*, and obtain an F_1 generation with a genotype *AaBb*. We test cross these *AaBb* individuals with *aabb*. Here are the genotypes of the first 1,000 progeny:

450 *AaBb*, 450 *aabb*, 50 *Aabb*, and 50 *aaBb*.

2 **How far apart are the *a* and *b* genes?** Well, what is the recombinant frequency? Which are the recombinant types, and which are the parental types?

Recombinant frequency (*a* to *b*) = (50 + 50)/1,000 = 0.1
So the map distance is
Map distance = 100 × recombinant frequency =
100 × 0.1 = 10 cM

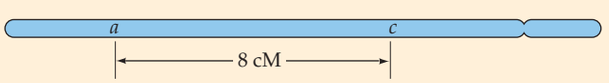


3 Now we make a cross *AACC* × *aacc*, obtain an F_1 generation, and test cross it, obtaining:

460 *AaCc*, 460 *aacc*, 40 *Aacc*, and 40 *aaCc*.

How far apart are the *a* and *c* genes?

Recombinant frequency (*a* to *c*) = (40 + 40)/1,000 = 0.08
Map distance = 100 × recombinant frequency =
100 × 0.08 = 8 cM



4 **How far apart are the *b* and *c* genes?**

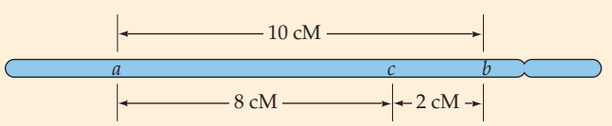
We make a cross *BBCC* × *bbcc*, obtain an F_1 generation, and test cross it, obtaining:

490 *BbCc*, 490 *bbcc*, 10 *Bbcc*, and 10 *bbCc*.

Determine the map distance between *b* and *c*.

Recombinant frequency (*b* to *c*) = (10 + 10)/1,000 = 0.02
Map distance = 100 × recombinant frequency =
100 × 0.02 = 2 cM

5 **Which of the three genes is between the other two?**
Because *a* and *b* are the farthest apart, *c* must be between them.



These numbers add up perfectly, but in most real cases they don't add up perfectly because of multiple crossovers.

Sex Determination and Sex-Linked Inheritance

In Mendel's work, reciprocal crosses always gave identical results; it did not matter, in general, whether a dominant allele was contributed by the mother or by the father. But in some cases, the parental origin of a chromosome does matter. For example, as we saw at the beginning of this chapter, human males inherit hemophilia A from their mother, not from their father. To understand the types of inheritance in which the parental origin of an allele is important, we must consider the ways in which sex is determined in different species.

Sex is determined in different ways in different species

In corn, a plant much studied by geneticists, every diploid adult has both male and female reproductive structures. The tissues in these two types of structures are genetically identical, just as roots and leaves are genetically identical. Plants such as corn, in which the same individual produces both male and female gametes, are said to be *monoecious* (from the Greek, "one house"). Other plants, such as date palms and oak trees, and most animals are *dioecious* ("two houses"), meaning that some individuals can produce only male gametes and the others can produce only female gametes. In other words, dioecious organisms have two sexes.

In most dioecious organisms, sex is determined by differences in the chromosomes, but such determination operates in different ways in different groups of organisms. For example, the sex of a honeybee depends on whether it develops from a fertilized or an unfertilized egg. A fertilized egg is diploid and gives rise to a female bee—either a worker or a queen, depending on the diet during larval life (again, note how the environment affects the phenotype). An unfertilized egg is haploid and gives rise to a male drone:

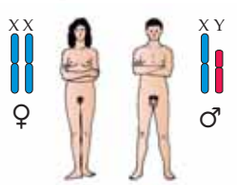
Diploid worker | Diploid queen | Haploid drone

In many other animals, including humans, sex is determined by a single **sex chromosome**, or by a pair of them. Both males and females have two copies of each of the rest of the chromosomes, which are called **autosomes**.

Female grasshoppers, for example, have two X chromosomes, whereas males have only one. Female grasshoppers are described as being XX (ignoring the autosomes) and males as XO (pronounced "ex-oh"):

Females form eggs that contain one copy of each autosome and one X chromosome. Males form approximately equal amounts of two types of sperm: One type contains one copy of each autosome and one X chromosome; the other type contains only autosomes. When an X-bearing sperm fertilizes an egg, the zygote is XX, and develops into a female. When a sperm without an X fertilizes an egg, the zygote is XO, and develops into a male. This chromosomal mechanism ensures that the two sexes are produced in approximately equal numbers.

As in grasshoppers, female mammals have two X chromosomes and males have one. However, male mammals also have a sex chromosome that is not found in females: the Y chromosome. Females may be represented as XX and males as XY:



Males produce two kinds of gametes. Each gamete has a complete set of autosomes, but half the gametes carry an X chromosome and the other half carry a Y. When an X-bearing sperm fertilizes an egg, the resulting XX zygote is female; when a Y-bearing sperm fertilizes an egg, the resulting XY zygote is male.

The X and Y chromosomes have different functions

Some subtle but important phenotypic differences show up clearly in mammals with abnormal sex chromosome constitutions. These conditions, which result from nondisjunctions, as described in Chapter 9, tell us something about the functions of the X and Y chromosomes. In humans, XO individuals sometimes appear. Human XO individuals are females who are physically moderately abnormal but mentally normal; usually they are also sterile. The XO condition in humans is called *Turner syndrome*. It is the only known case in which a human can survive with only one member of a chromosome pair (here, the XY pair), although most XO conceptions terminate spontaneously early in development. XXY individuals also occur; this condition is known as *Klinefelter syndrome*. People with this genotype are sometimes taller than average, always sterile, and always male.

These observations suggested that the gene that determines maleness is located on the Y chromosome. Observations of people with other types of chromosomal abnormalities helped researchers to pinpoint the location of that gene:

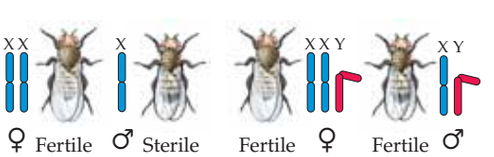
- Some XY individuals are phenotypically women and lack a small portion of the Y chromosome.
- Some men are genetically XX and have a small piece of the Y chromosome present but attached to another chromosome.

The Y fragment that is missing and present in these two examples, respectively, contains the maleness-determining gene, which was named *SRY* (*s*ex-determining *r*egion on the *Y* chromosome).

The *SRY* gene encodes a protein involved in *primary sex determination*—that is, the determination of the kinds of gametes that will be produced and the organs that will make them. In the presence of functional SRY protein, the embryo develops sperm-producing testes. (Notice that italic type is used for the name of a gene, but roman type is used for the name of a protein.) If the embryo has no Y chromosome, the *SRY* gene is absent, and thus the SRY protein is not made. In the absence of the SRY protein, the embryo develops egg-producing ovaries. In this case, a gene on the X chromosome called *DAX1* produces an anti-testis factor. So the role of *SRY* in a male is to inhibit the maleness inhibitor encoded by *DAX1*. The SRY protein does this in male cells, but since it is not present in females, *DAX1* can act to inhibit maleness.

Primary sex determination is not the same as *secondary sex determination*, which results in the outward manifestations of maleness and femaleness (body type, breast development, body hair, and voice). These outward characteristics are not determined directly by the presence or absence of the Y chromosome. Rather, they are determined by genes scattered on the autosomes and X chromosome that control the actions of hormones, such as testosterone and estrogen.

The Y chromosome functions differently in *Drosophila melanogaster*. Superficially, *Drosophila* follows the same pattern of sex determination as mammals—females are XX and males are XY. However, XO individuals are males (rather than females as in mammals) and almost always are indistinguishable from normal XY males except that they are sterile. XXY *Drosophila* are normal, fertile females:



Thus, in *Drosophila*, sex is determined by the ratio of X chromosomes to autosome sets. If there is one X chromosome for each set of autosomes, the individual is a female; if there is only one X chromosome for the two sets of autosomes, the individual is a male. The Y chromosome plays no sex-determining role in *Drosophila*, but it is needed for male fertility.

Caenorhabditis elegans is a favorite model organism for studies of development (see Chapter 19). This tiny worm has two sexes: male and hermaphrodite (self-fertilizing). As in fruit flies, sex is determined by the X:autosome ratio-individuals with a ratio below 0.67 are male.

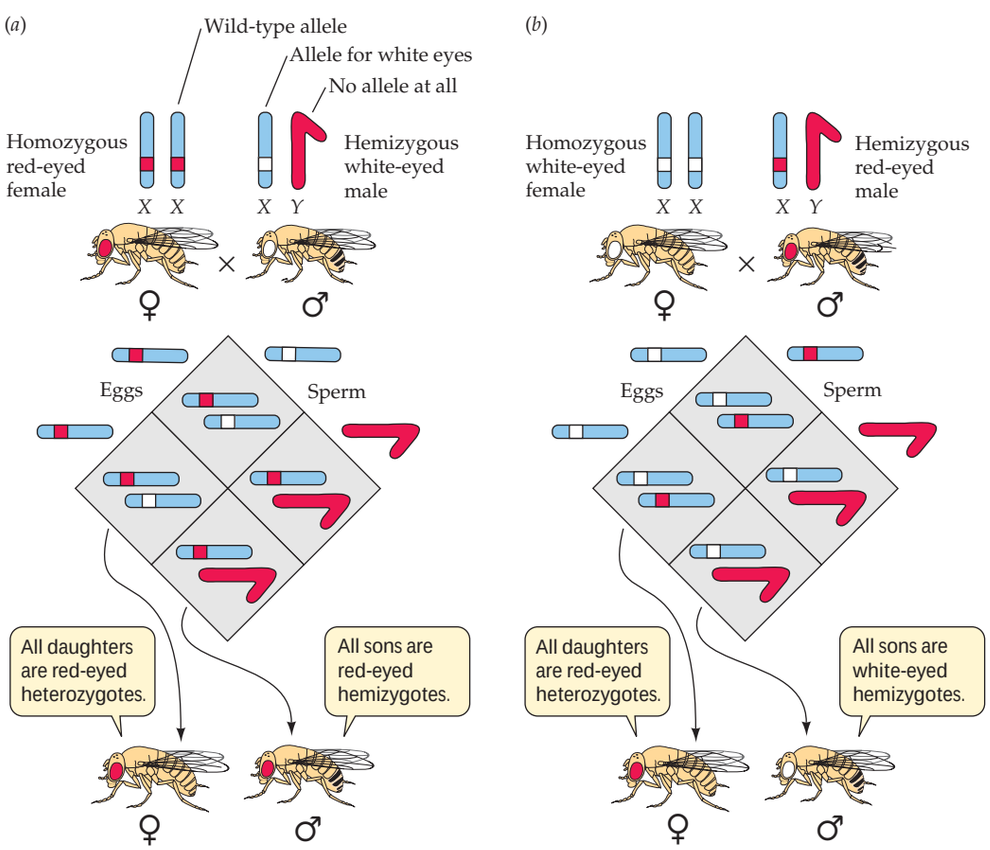
In birds, moths, and butterflies, males are XX and females are XY. To avoid confusion, these forms are usually expressed as ZZ (male) and ZW (female):

In these organisms, the female produces two types of gametes, carrying Z or W. Whether the egg is Z or W determines the sex of the offspring, in contrast to humans and fruit flies, in which the sperm, carrying either X or Y, determines the sex.

Genes on sex chromosomes are inherited in special ways

Genes on sex chromosomes do not show the Mendelian patterns of inheritance we have described above. In *Drosophila* and in humans, the Y chromosome carries few known genes, but a substantial number of genes affecting a great variety of characters are carried on the X chromosome. Any such gene is present in two copies in females, but in only one copy in males. Therefore, females may be heterozygous for genes that are on the X chromosome, but males will always be **hemizygous** for genes on the X chromosome—they will have only one copy of each, and it will be expressed. Thus, reciprocal crosses do not give identical results for characters whose genes are carried on the sex chromosomes, and these characters do not show the usual Mendelian ratios for the inheritance of genes located on autosomes.

The first and still one of the best examples of inheritance of characters governed by loci on the sex chromosomes (**sex-linked** inheritance) is that of eye color in *Drosophila*. The wild-type eye color of these flies is red. In 1910, Morgan discovered a mutation that causes white eyes. He experimented by crossing flies of the wild-type and mutant phenotypes. His results demonstrated that the eye color locus is on the X chromosome. Study Figure 10.23 as you follow the crosses and results:



10.23 Eye Color Is a Sex-Linked Trait in *Drosophila* Thomas Hunt Morgan demonstrated that a mutant allele that causes white eyes in *Drosophila* is carried on the X chromosome. Note that in this case, the reciprocal crosses do not have the same results.

- When a homozygous red-eyed female was crossed with a (hemizygous) white-eyed male, all the sons and daughters had red eyes, because red is dominant over white and all the progeny had inherited a wild-type X chromosome from their mothers (Figure 10.23*a*).
- However, in the reciprocal cross, in which a white-eyed female was mated with a red-eyed male, all the sons were white-eyed and all the daughters were red-eyed (Figure 10.23*b*).
- The sons from the reciprocal cross inherited their only X chromosome from their white-eyed mother; the Y chromosome they inherited from their father does not carry the eye color locus (Figure 10.23*b*).
- The daughters, on the other hand, got an X chromosome bearing the white allele from their mother and an X chromosome bearing the red allele from their father; they were therefore red-eyed heterozygotes (Figure 10.23*b*).
- When heterozygous females were mated with red-eyed males, half their sons had white eyes, but all their daughters had red eyes.

Together, these results showed that eye color was carried on the X chromosome and not on the Y.

Humans display many sex-linked characters

The human X chromosome carries about two thousand genes. The alleles at these loci follow the same pattern of inheritance as those for white eyes in *Drosophila*. One human X chromosome gene, for example, has a mutant recessive allele that leads to red-green color blindness, a hereditary disorder. Red-green color blindness appears in individuals who are homozygous or hemizygous for the mutant allele.

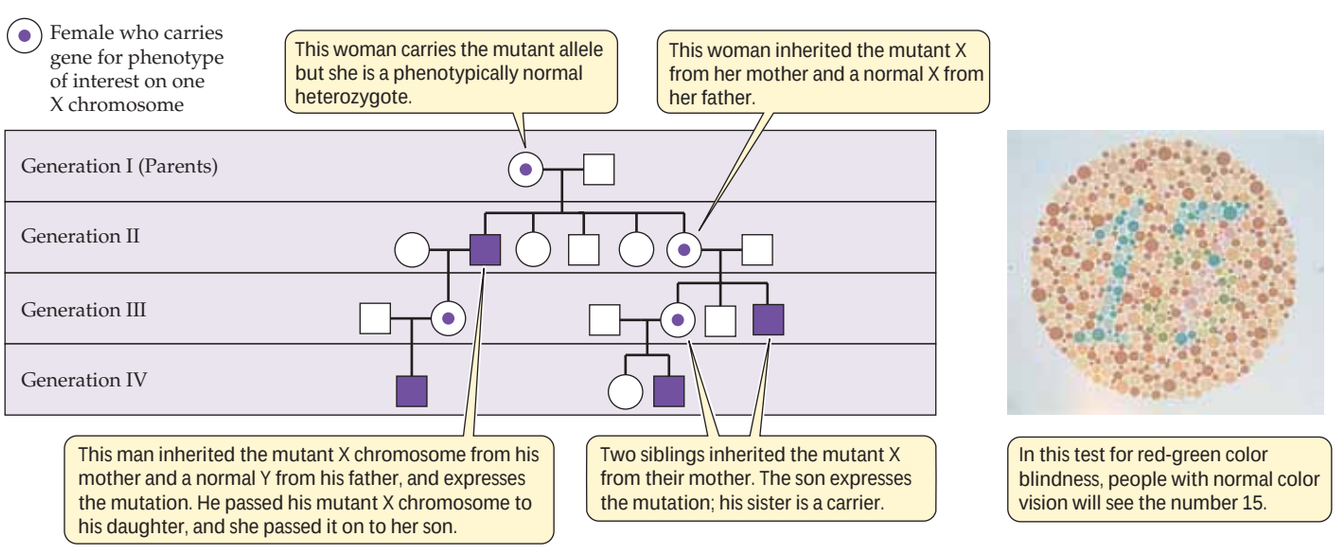
Pedigree analysis of X-linked recessive phenotypes (Figure 10.24) reveals the following patterns:

- The phenotype appears much more often in males than in females, because only one copy of the rare allele is needed for its expression in males, while two copies must be present in females.
- A male with the mutation can pass it on only to his daughters; all his sons get his Y chromosome.
- Daughters who receive one mutant X chromosome are heterozygous *carriers*. They are phenotypically normal, but they can pass the mutant X to both sons and daughters (but do so only half of the time, on average, since half of their X chromosomes carry the normal allele).
- The mutant phenotype can skip a generation if the mutation passes from a male to his daughter (who will be phenotypically normal) and thus to her son.

Hemophilia A, which affected the family described at the beginning of this chapter, is an X-linked recessive phenotype, as are several other important human diseases, as we will see in later chapters. Human mutations inherited as X-linked dominant phenotypes are rarer than X-linked recessives because dominant phenotypes appear in every generation, and because people carrying the harmful mutation, even as heterozygotes, often fail to survive and reproduce. (Look at the four points above and try to determine what would happen if the mutation were dominant.)

The small human Y chromosome carries several dozen genes. Among them is the maleness determinant, *SRY*. Interestingly, for some genes on the Y, there are similar, but not identical, genes on the X. For example, one of the proteins

10.24 Red-Green Color Blindness is a Sex-Linked Trait in Humans
The mutant allele for red-green color blindness is inherited as an X-linked recessive.



that make up ribosomes has a gene on the Y that is expressed only in male cells, while the X-linked counterpart is expressed in both sexes. This means that there are "male" and "female" ribosomes; the significance of this phenomenon is unknown. Y-linked alleles are passed only from father to son. (You can verify this with a Punnett square.)

Non-Nuclear Inheritance

The nucleus is not the only organelle in a eukaryotic cell that carries genetic material. As we described in Chapter 4, mitochondria and plastids, which may have arisen from prokaryotes that colonized other cells, contain small numbers of genes. For example, in humans, there are about 30,000 genes in the nuclear genome and 37 in the mitochondrial genome. Plastid genomes are about five times larger than those of mitochondria. In any case, several of the genes of cytoplasmic organelles are important for organelle assembly and function, so it is not surprising that mutations of these genes have profound effects on the organism.

The inheritance of organelle genes differs from that of nuclear genes for several reasons:

- In most organisms, mitochondria and plastids are inherited from the mother only. As you will see in later chapters, eggs contain abundant cytoplasm and organelles, but the only part of the sperm that survives to take part in the union of haploid gametes is the nucleus. So you have inherited your mother's mitochondria (with their genes), but not your father's.
- There may be hundreds of mitochondria or plastids in a cell. So a cell is not diploid for organelle genes; rather, it is highly polyploid.
- Organelle genes tend to mutate at much faster rates than nuclear genes, so there are multiple alleles of organelle genes.

The phenotypes of mutations in the DNA of organelles reflect the organelles' roles. For example, in plants and some eukaryotic algae, certain plastid mutations affect the proteins that assemble chlorophyll molecules into photosystems (see Figure 8.9) and result in a phenotype that is essentially white instead of green. Mitochondrial mutations that affect one of the complexes in the electron transport chain result in less ATP production. They have especially noticeable effects in tissues with a high energy requirement, such as the nervous system, muscles, and kidneys. In 1995, Greg Lemond, a professional cyclist who had won the famous Tour de France three times, was forced to retire because of muscle weakness suspected to be caused by a mitochondrial mutation.

Chapter Summary

The Foundations of Genetics

▶ Although it had long been known that both parents contribute to the character traits of their offspring, before Mendel's time it was believed that, once they were brought together, the units of inheritance blended and could never be separated.

▶ Although Gregor Mendel's work was meticulous and well documented, his discoveries, reported in the 1860s, were ignored until decades later.

Mendel's Experiments and the Laws of Inheritance

▶ Mendel used the garden pea for his studies because the plants were easily cultivated and crossed and because they showed numerous characters (such as seed shape) with clearly different traits (spherical or wrinkled). **Review Figure 10.1, Table 10.1**

▶ In a monohybrid cross, the offspring of the first generation (F_1) showed only one of the two parental traits. Mendel proposed that the trait observed in the F_1 was dominant and the other was recessive. **Review Table 10.1**

▶ When the F_1 offspring were self-pollinated, the resulting F_2 generation showed a 3:1 phenotypic ratio, with the recessive phenotype present in one-fourth of the offspring. This reappearance of the recessive phenotype refuted the blending theory. **Review Figure 10.3**

▶ Because some alleles are dominant and some are recessive, the same phenotype can result from different genotypes. Homozygous genotypes have two copies of the same allele; heterozygous genotypes have two different alleles. Heterozygous genotypes yield phenotypes that show the dominant trait.

▶ On the basis of many crosses using different characters, Mendel proposed his first law: that the units of inheritance (now known as genes) are particulate, that there are two alleles of each gene in each parent, and that during gamete formation the two alleles segregate from each other. **Review Figure 10.4**

▶ Geneticists who followed Mendel showed that genes are carried on chromosomes and that alleles are segregated during meiosis I. **Review Figure 10.5**

▶ Using a test cross, Mendel was able to determine whether a plant showing the dominant phenotype was homozygous or heterozygous. The appearance of the recessive phenotype in half of the offspring of such a cross indicates that the parent is heterozygous. **Review Figure 10.6. See Web/CD Activity 10.1**

▶ From studies of the inheritance of two characters using dihybrid crosses, Mendel concluded that alleles of different genes assort independently. **Review Figures 10.7, 10.8. See Web/CD Tutorial 10.1**

▶ We can predict the results of hybrid crosses either by using a Punnett square or by calculating probabilities. To determine the joint probability of independent events, we multiply the individual probabilities. To determine the probability of an event that can occur in two or more different ways, we add the individual probabilities. **Review Figure 10.9**

▶ The analysis of pedigrees can trace Mendelian inheritance patterns in humans. **Review Figures 10.10, 10.11**

Alleles and Their Interactions

▶ New alleles arise by mutation, and many genes have multiple alleles. **Review Figure 10.12**

▶ Dominance is sometimes not complete, since both alleles in a heterozygous organism may be expressed in the phenotype. **Review Figures 10.13, 10.14**

Gene Interactions

▶ In epistasis, the products of different genes interact to produce a phenotype. **Review Figure 10.15**

▶ Environmental variables such as temperature, nutrition, and light affect gene action.

▶ In some cases, the phenotype is the result of the effects of several genes and the environment, and inheritance is quantitative. **Review Figure 10.17**

Genes and Chromosomes

▶ Each chromosome carries many genes. Genes located on the same chromosome are said to be linked, and they are often inherited together. **Review Figure 10.18**

▶ Linked genes can recombine by crossing over in prophase I of meiosis. The result is recombinant gametes, which have new combinations of linked genes because of the exchange. **Review Figures 10.19, 10.20**

▶ The distance between two genes on a chromosome is proportional to the frequency of crossing over between them. Genetic maps are based on recombinant frequencies. **Review Figures 10.21, 10.22.**

See Web/CD Tutorial 10.2

Sex Determination and Sex-Linked Inheritance

▶ Sex chromosomes carry genes that determine whether the organism will produce male or female gametes. The specific functions of X and Y chromosomes differ among species.

▶ In fruit flies and mammals, the X chromosome carries many genes, but the Y chromosome has only a few. Males have only one allele for X-linked genes, so rare alleles show up phenotypically more often in males than in females. **Review Figures 10.23, 10.24**

Non-Nuclear Inheritance

▶ Cytoplasmic organelles such as plastids and mitochondria contain some heritable genes.

▶ Cytoplasmic organelle genes are generally inherited only from the mother because male gametes contribute only their nucleus to the zygote at fertilization.

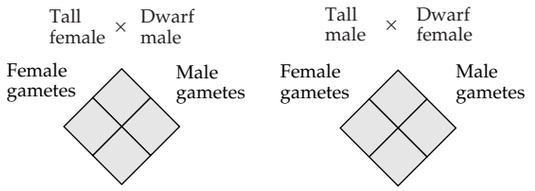
See Web/CD Activities 10.2 and 10.3 for a concept review of this chapter.

Self-Quiz

1. In a simple Mendelian monohybrid cross, tall plants were crossed with short plants and the F_1 were crossed among themselves. What fraction of the F_2 generation are both tall *and* heterozygous?
   a. $^1/_8$
   b. $^1/_4$
   c. $^1/_3$
   d. $^2/_3$
   e. $^1/_2$
2. The phenotype of an individual
   a. depends at least in part on the genotype.
   b. is either homozygous or heterozygous.
   c. determines the genotype.
   d. is the genetic constitution of the organism.
   e. is either monohybrid or dihybrid.
3. The ABO blood groups in humans are determined by a multiple allelic system where I^A and I^B are codominant and dominant to I^O. A newborn infant is type A. The mother is type O. Possible genotypes of the father are:
   a. A, B or AB
   b. A, B or O
   c. O only
   d. A or AB
   e. A or O
4. Which statement about an individual that is homozygous for an allele is *not* true?
   a. Each of its cells possesses two copies of that allele.
   b. Each of its gametes contains one copy of that allele.
   c. It is true-breeding with respect to that allele.
   d. Its parents were necessarily homozygous for that allele.
   e. It can pass that allele to its offspring.
5. Which statement about a test cross is *not* true?
   a. It tests whether an unknown individual is homozygous or heterozygous.
   b. The test individual is crossed with a homozygous recessive individual.
   c. If the test individual is heterozygous, the progeny will have a 1:1 ratio.
   d. If the test individual is homozygous, the progeny will have a 3:1 ratio.
   e. Test cross results are consistent with Mendel's model of inheritance.
6. Linked genes
   a. must be immediately adjacent to one another on a chromosome.
   b. have alleles that assort independently of one another.
   c. never show crossing over.
   d. are on the same chromosome.
   e. always have multiple alleles.
7. In the F_2 generation of a dihybrid cross
   a. 4 phenotypes appear in the ratio 9:3:3:1 if the loci are linked.
   b. 4 phenotypes appear in the ratio 9:3:3:1 if the loci are unlinked.
   c. 2 phenotypes appear in the ratio 3:1 if the loci are unlinked.
   d. 3 phenotypes appear in the ratio 1:2:1 if the loci are unlinked.
   e. 2 phenotypes appear in the ratio 1:1 whether or not the loci are linked.
8. The sex of a human is determined by
   a. ploidy, the male being haploid.
   b. the Y chromosome.
   c. X and Y chromosomes, the male being XY.
   d. the number of X chromosomes, the male being XO.
   e. Z and W chromosomes, the male being ZZ.
9. In epistasis
   a. nothing changes from generation to generation.
   b. one gene alters the effect of another.
   c. a portion of a chromosome is deleted.
   d. a portion of a chromosome is inverted.
   e. the behavior of two genes is entirely independent.
10. In humans, spotted teeth is caused by a dominant sex-linked gene. A man with spotted teeth whose mother had normal teeth marries a woman with normal teeth. Therefore,
    a. all of their daughters will have normal teeth.
    b. all of their daughters will have spotted teeth.
    c. all of their children will have spotted teeth.
    d. half of their sons will have spotted teeth.
    e. none of their sons will have spotted teeth.

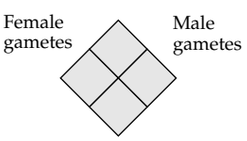
Genetics Problems

1. Using the Punnett squares below, show that for typical dominant and recessive autosomal traits, it does not matter which parent contributes the dominant allele and which the recessive allele. Cross true-breeding tall plants (*TT*) with true-breeding dwarf plants (*tt*).



2. The photograph shows the shells of 15 bay scallops, *Argopecten irradians*. These scallops are hermaphroditic; that is, a single individual can reproduce sexually, as did the pea plants of the F_1 generation in Mendel's experiments. Three color schemes are evident: yellow, orange, and black and white. The color-determining gene has three alleles. The top row shows a yellow scallop and a representative sample of its offspring, the middle row shows a black-and-white scallop and its offspring, and the bottom row shows an orange scallop and its offspring. Assign a suitable symbol to each of the three alleles participating in color control; then determine the genotype of each of the three parent individuals and tell what you can about the genotypes of the different offspring. Explain your results carefully.

3. Show diagrammatically what occurs when the F_1 offspring of the cross in Question 1 self-pollinate.



4. A new student of genetics suspects that a particular recessive trait in fruit flies (dumpy wings, which are somewhat smaller and more bell-shaped than the wild-type) is sex-linked. A single mating between a fly having dumpy wings (*dp*; female) and a fly with wild-type wings (*Dp*; male) produces 3 dumpy-winged females and 2 wild-type males. On the basis of these data, is the trait sex-linked or autosomal? What were the genotypes of the parents? Explain how these conclusions can be reached on the basis of so few data.

5. The sex of fishes is determined by the same X-Y system as in humans. An allele of one locus on the Y chromosome of the fish *Lebistes* causes a pigmented spot to appear on the dorsal fin. A male fish that has a spotted dorsal fin is mated with a female fish that has an unspotted fin. Describe the phenotypes of the F_1 and the F_2 generations from this cross.

6. In *Drosophila melanogaster*, the recessive allele *p*, when homozygous, determines pink eyes. *Pp* or *PP* results in wild-type eye color. Another gene, on another chromosome, has a recessive allele, *sw*, that produces short wings when homozygous. Consider a cross between females of genotype *PPSwSw* and males of genotype *ppswsw*. Describe the phenotypes and genotypes of the F_1 generation and of the F_2 generation produced by allowing the F_1 progeny to mate with one another.

7. On the same chromosome of *Drosophila melanogaster* that carries the *p* (pink eyes) locus, there is another locus that affects the wings. Homozygous recessives, *byby*, have blistery wings, while the dominant allele *By* produces wild-type wings. The *P* and *By* loci are very close together on the chromosome; that is, the two loci are tightly linked. In answering these questions, assume that no crossing over occurs.
   a. For the cross *PPByBy* × *ppbyby*, give the phenotypes and genotypes of the F_1 and of the F_2 generations produced by interbreeding of the F_1 progeny.
   b. For the cross *PPbyby* × *ppByBy*, give the phenotypes and genotypes of the F_1 and of the F_2 generations.
   c. For the cross of Question 7*b*, what further phenotype(s) would appear in the F_2 generation if crossing over occurred?
   d. Draw a nucleus undergoing meiosis, at the stage in which the crossing over (Question 7*c*) occurred. In which generation (P, F_1, or F_2) did this crossing over take place?

8. Consider the following cross of *Drosophila melanogaster* (alleles as described in Question 6): Males with genotype *Ppswsw* are crossed with females of genotype *ppSwsw*. Describe the phenotypes and genotypes of the F_1 generation.

9. In the Andalusian fowl, a single pair of alleles controls the color of the feathers. Three colors are observed: blue, black, and splashed white. Crosses among these three types yield the following results:

PARENTS	PROGENY
Black × blue	Blue and black (1:1)
Black × splashed white	Blue
Blue × splashed white	Blue and splashed white (1:1)
Black × black	Black
Splashed white × splashed white	Splashed white

   a. What progeny would result from the cross blue × blue?
   b. If you want to sell eggs, all of which would yield blue fowl, how should you proceed?

10. In *Drosophila melanogaster*, white (*w*), eosin (w^e), and wild-type red (w^+) are multiple alleles of a single locus for eye color. This locus is on the X chromosome. A female that has eosin (pale orange) eyes is crossed with a male that has wild-type eyes. All the female progeny are red-eyed; half the male offspring have eosin eyes, and half have white eyes.

    a. What is the order of dominance of these alleles?

    b. What are the genotypes of the parents and progeny?

11. Color blindness is a recessive trait. Two people with normal vision have two sons, one color-blind and one with normal vision. If the couple also has daughters, what proportion of them will have normal vision? Explain.

12. A mouse with an agouti coat is mated with an albino mouse of genotype *aabb*. Half of the offspring are albino, one-fourth are black, and one-fourth are agouti. What are the genotypes of the agouti parents and of the various kinds of offspring? (*Hint*: See the section on epistasis.)

13. The disease Leber's optic neuropathy is caused by a mutation in a gene carried on mitochondrial DNA. What would be the result in their first child if a man with this disease married a woman who did not have the disease? What would be the result if the wife had the disease and the husband did not?

8 DNA and Its Role in Heredity

In the novel and film *Jurassic Park*, fictional scientists were depicted using biotechnology to produce dinosaurs. In the story, the scientists isolated the DNA of dinosaurs from fossilized insects that had sucked the reptiles' blood. The insects, preserved intact in amber (fossilized tree resin), yielded DNA that could be used to produce living individuals of long-extinct organisms such as *Tyrannosaurus rex*. This premise of Michael Crichton's novel was based on an actual scientific paper that claimed to show reptilian DNA sequences in a fossil insect. Unfortunately, the scientific report was not correct; the "preserved" DNA turned out to be a contaminant from modern organisms.

Despite the fact that the preservation of intact DNA over millions of years is highly improbable, the popular success of *Jurassic Park* did bring the idea of DNA as the genetic material to the attention of millions of readers and viewers. Indeed, the double helix of DNA has become a familiar secular icon in the 50 years since that structure was first proposed.

In this and the next several chapters, we will focus on the structure, replication, and function of DNA. As you will see, the structure of DNA determines its function. This chapter will first describe the key experiments that led to the determination that the genetic material is DNA. Then the structure and replication of the molecule will be described. Finally, we will present two practical applications that have arisen from our knowledge of DNA replication: the polymerase chain reaction and DNA sequencing.

Jurassic Park In Michael Crichton's novel, DNA isolated from a fossil is used to produce living dinosaurs. Although such a procedure is strictly fiction, the movie based on the book did bring the role of DNA as the genetic material to the attention of millions of people.

DNA: The Genetic Material

As we saw in Chapter 10, by the early twentieth century geneticists had associated the presence of genes with chromosomes. Research began to focus on exactly which chemical component of chromosomes comprised this genetic material.

By the 1920s, scientists knew that chromosomes were made up of DNA and proteins. At this time a new dye was developed that could bind specifically to DNA and turned red in direct proportion to the amount of DNA present in a cell. This technique provided circumstantial evidence that DNA was the genetic material:

- It was in the right place, since it was an important component of the nucleus and the chromosomes, which were known to carry genes.

- It varied among species. When cells from different species were stained with the dye and their color intensity measured, each species appeared to have its own specific nuclear DNA content.
- It was present in the right amounts. The amount of DNA in somatic cells (body cells not specialized for reproduction) was twice that in the reproductive cells (eggs or sperm)—as might be expected for diploid and haploid cells, respectively.

But circumstantial evidence is *not* a scientific demonstration of cause and effect. After all, proteins are also present in nuclei. The convincing demonstration that DNA is the genetic material came from two lines of experiments, one on bacteria and the other on viruses.

DNA from one type of bacterium genetically transforms another type

The history of biology is filled with incidents in which research on one specific topic has—with or without answering the question originally asked—contributed richly to another, apparently unrelated area. Such a case of *serendipity* is the work of Frederick Griffith, an English physician.

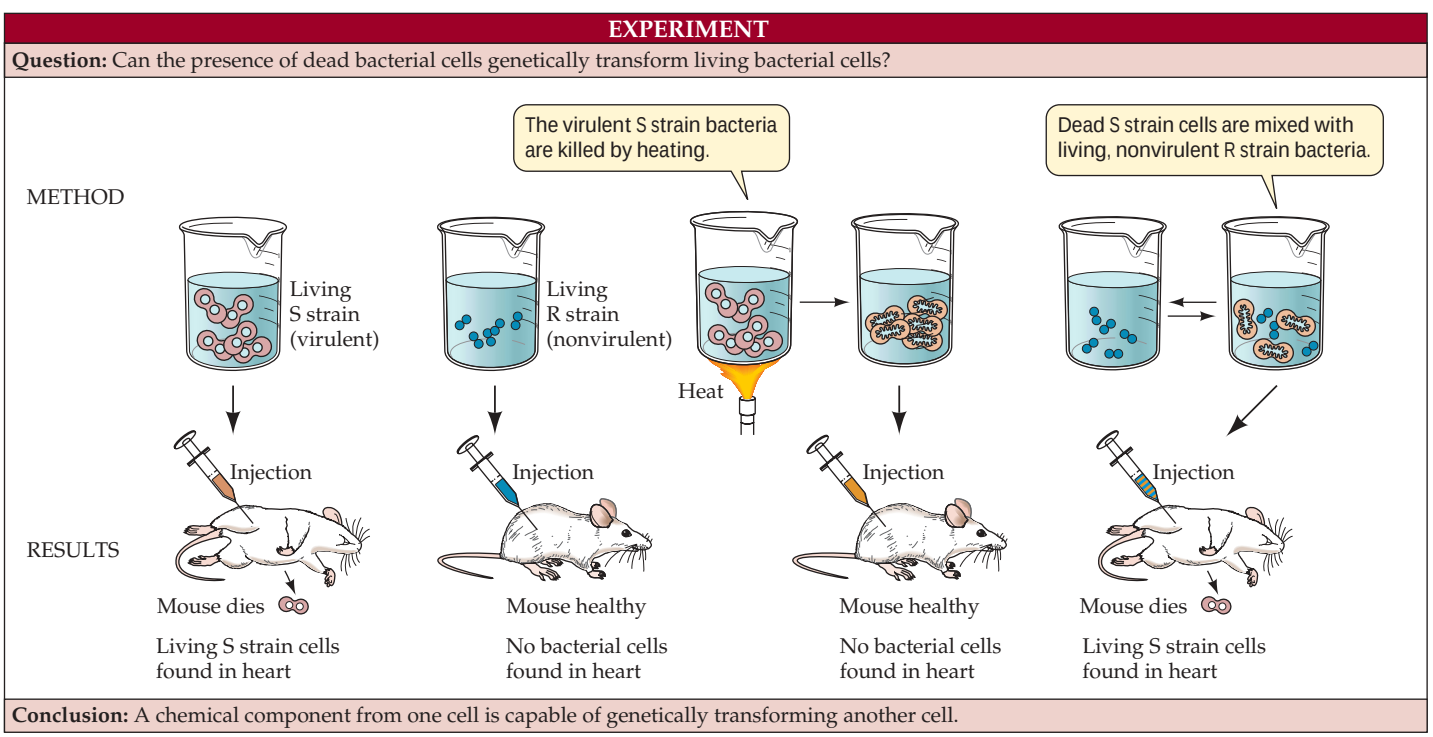
In the 1920s, Griffith was studying the bacterium *Streptococcus pneumoniae*, or pneumococcus, one of the agents that causes pneumonia in humans. He was trying to develop a vaccine against this devastating illness (antibiotics had not yet been discovered). Griffith was working with two strains* of pneumococcus:

- Cells of the S strain produced colonies that looked smooth (S). Covered by a polysaccharide capsule, these cells were protected from attack by a host's immune system. When S cells were injected into mice, they reproduced and caused pneumonia (the strain was *virulent*).
- Cells of the R strain produced colonies that looked rough (R), lacked the protective capsule, and were not virulent.

Griffith inoculated some mice with heat-killed S pneumococci. These heat-killed bacteria did not produce infection. However, when Griffith inoculated other mice with a mixture of living R bacteria and heat-killed S bacteria, to his astonishment, the mice died of pneumonia (Figure 11.1). When he examined blood from the hearts of these mice, he found it full of living bacteria—many of them with characteristics of the virulent S strain! Griffith concluded that, in the presence of the dead S pneumococci, some of the living R pneumococci had been transformed into virulent S-strain organisms.

*A bacterial *strain* is a population of bacterial cells descended from a single parent cell; strains may differ in one or more inherited characteristics.

11.1 Genetic Transformation of Nonvirulent Pneumococci
Frederick Griffith's experiments demonstrated that something in the virulent S strain could transform nonvirulent R strain bacteria into a lethal form, even when the S strain bacteria had been killed by high temperatures.



Did this transformation of the bacteria depend on something that happened in the mouse's body? No. It was shown that simply incubating living R and heat-killed S bacteria together in a test tube yielded the same transformation. Years later, another group of scientists discovered that a cell-free extract of heat-killed S cells also transformed R cells. (A *cell-free extract* contains all the contents of ruptured cells, but no intact cells.) This result demonstrated that some substance—called at the time a chemical **transforming principle**—from the dead S pneumococci could cause a heritable change in the affected R cells. This was an extraordinary discovery: Treatment with a substance permanently changed an inherited characteristic. Now it remained to identify the chemical structure of this substance.

The transforming principle is DNA

Identifying the transforming principle was a crucial step in the history of biology. It was accomplished over a period of years by Oswald Avery and his colleagues at what is now Rockefeller University. They treated samples known to contain the pneumococcal transforming principle in a variety of ways to destroy different types of molecules—proteins, nucleic acids, carbohydrates, and lipids—and tested the treated samples to see if they had retained transforming activity.

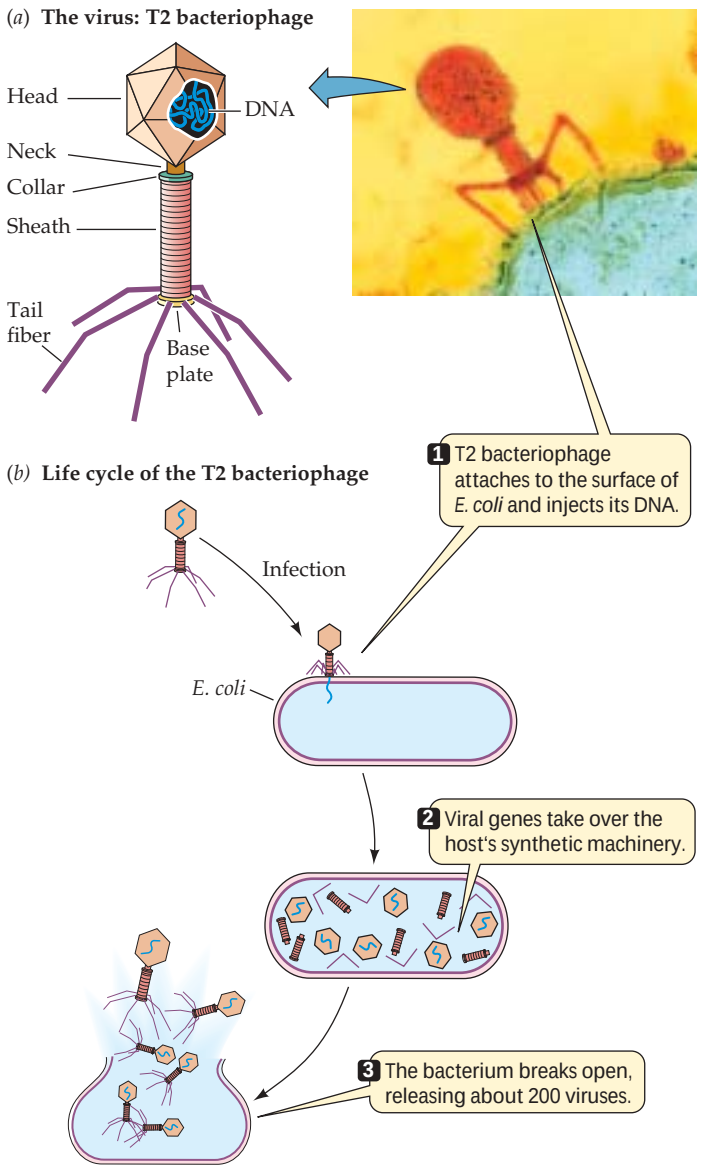
The answer was always the same: If the DNA in the sample was destroyed, transforming activity was lost, but there was no loss of activity when proteins, carbohydrates, or lipids were destroyed. As a final step, Avery, with Colin MacLeod and Maclyn McCarty, isolated virtually pure DNA from a sample containing pneumococcal transforming principle and showed that it caused bacterial transformation. We now know that the gene encoding the enzyme that catalyzes the synthesis of the pathogenic polysaccharide capsule was transferred during transformation.

The work of Avery, MacLeod, and McCarty, published in 1944, was a milestone in establishing that DNA is the genetic material in cells. However, it had little impact at the time, for two reasons. First, most scientists did not believe that DNA was chemically complex enough to be the genetic material, especially given the much greater chemical complexity of proteins. Second, and perhaps more important, bacterial genetics was a new field of study—it was not yet clear that bacteria even *had* genes.

Viral replication experiments confirm that DNA is the genetic material

The questions about bacteria were soon resolved as researchers identified genes and mutations in these organisms. Bacteria and viruses seemed to undergo genetic processes similar to those in fruit flies and pea plants. Experiments were designed with these relatively simple systems to discover the nature of the genetic material.

In 1952, Alfred Hershey and Martha Chase of the Carnegie Laboratory of Genetics published a paper that had a much greater immediate impact than Avery's 1944 paper. The Hershey–Chase experiment, which sought to determine whether DNA or protein was the hereditary material, was carried out with a virus that infects bacteria. This virus, called the T2 bacteriophage, consists of little more than a DNA core packed inside a protein coat (Figure 11.2*a*). The virus is thus made of the two materials that were, at the time, the leading candidates for the genetic material.



11.2 T2 and the Bacteriophage Reproduction Cycle (*a*) The external structures of T2 bacteriophage consist entirely of protein. This cutaway view shows a strand of DNA within the head. (*b*) T2 is parasitic on *E. coli*, depending on the bacterium to produce new viruses.

When a T2 bacteriophage attacks a bacterium, part (but not all) of the virus enters the bacterial cell. About 20 minutes later, the cell bursts, releasing dozens of viruses. The entry of a viral component changes the genetic program of the host bacterial cell: it is converted from a bacterium into a bacteriophage factory. Hershey and Chase set out to determine which part of the virus—protein or DNA—enters the bacterial cell. To trace the two components of the virus over its life cycle (Figure 11.2*b*), Hershey and Chase labeled each with a specific radioactive tracer:

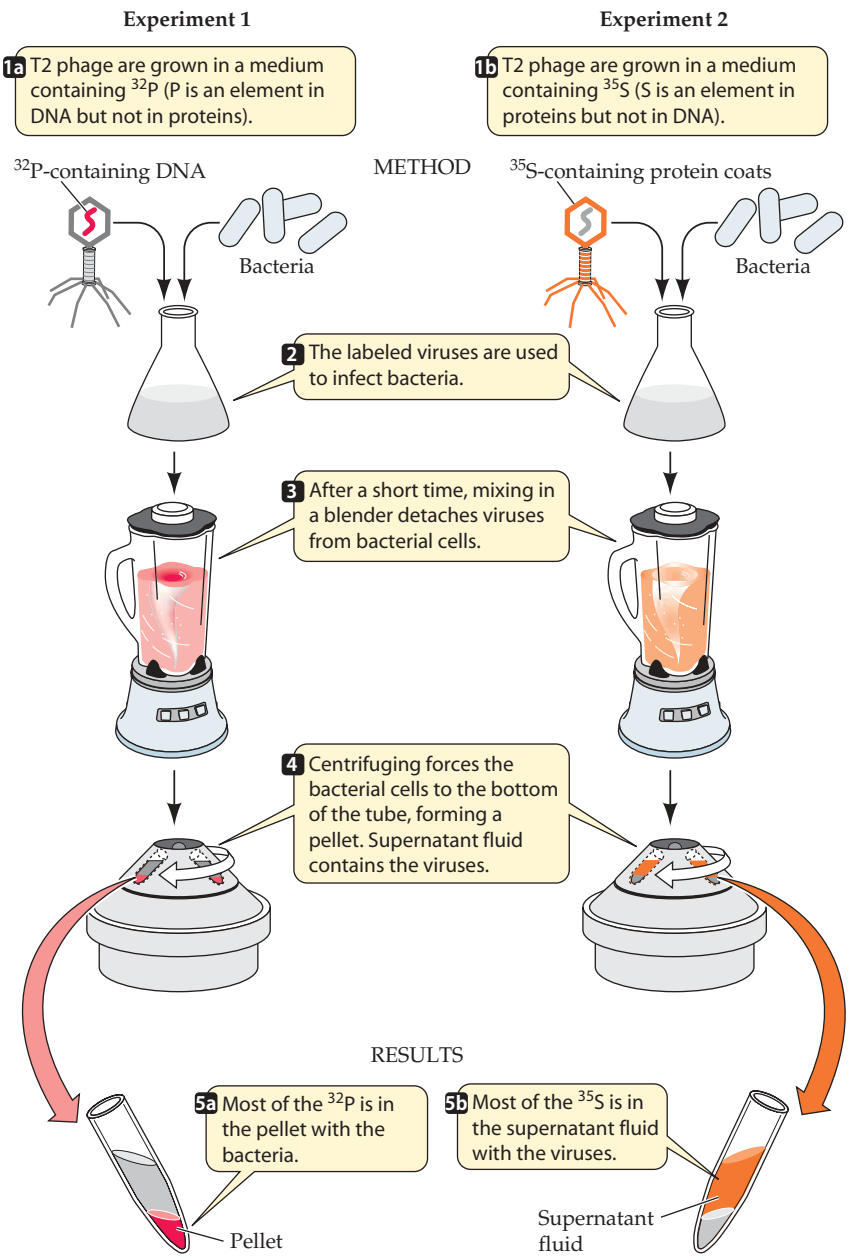
- Viral proteins contain some sulfur (in the amino acids cysteine and methionine), an element not present in DNA, and sulfur has a radioactive isotope, ^{35}S. So Hershey and Chase grew a batch of T2 bacteriophage in a bacterial culture in the presence of ^{35}S; the resulting viruses had their proteins labeled with this isotope.
- The deoxyribose–phosphate "backbone" of DNA, on the other hand, is rich in phosphorus (see Chapter 3), an element that is not present in most proteins, and phosphorus also has a radioactive isotope, ^{32}P. The researchers grew another batch of T2 in a bacterial culture in the presence of ^{32}P, so that all the viral DNA was labeled with ^{32}P.

With these radioactively labeled viruses, Hershey and Chase performed their revealing experiments. They allowed bacteriophage containing either ^{32}P or ^{35}S to attach to bacteria. After a few minutes, they agitated the mixtures vigorously in a kitchen blender, which (without bursting the bacteria) stripped away the parts of the virus that had not penetrated the bacteria. Then Hershey and Chase separated the bacteria from the rest of the material. They found that most of the ^{35}S (and thus the protein) had separated from the bacteria, and that most of the ^{32}P (and thus the DNA) had stayed with the bacteria. These results suggested that the DNA was transferred to the bacteria, whereas the protein remained outside, and thus that it was DNA that redirected the genetic program of the bacterial cell (Figure 11.3).

Hershey and Chase performed other similar but more long-range experiments, allowing a progeny (offspring) generation of viruses to be collected. The resulting viruses contained almost none of the original ^{35}S and none of the parental viral protein. However, they contained about one-

EXPERIMENT

Question: Which component of a bacteriophage—DNA or protein—is the hereditary material that enters a bacterial cell to direct the assembly of new virus particles?



Conclusion: DNA, not protein, enters bacterial cells and directs the assembly of new viruses.

11.3 The Hershey–Chase Experiment Because only DNA entered the bacterial cell during infection by labeled bacteriophage, the experiment demonstrated that DNA, not protein, is the hereditary material.

third of the original ^{32}P—and thus, presumably, one-third of the original DNA. Because DNA was carried over in the virus from generation to generation but protein was not, a logical conclusion was that the hereditary information of the virus is contained in the DNA.

The Hershey–Chase experiment convinced most scientists that DNA is the carrier of hereditary information.

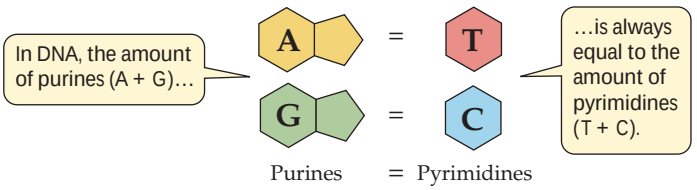
The Structure of DNA

As soon as scientists were convinced that the genetic material was DNA, they began efforts to learn its precise, three-dimensional chemical structure. In determining the structure of DNA, scientists hoped to find the answers to two questions: how DNA is replicated between nuclear divisions, and how it causes the synthesis of specific proteins. Both expectations were fulfilled.

X-ray crystallography provided clues to DNA structure

The structure of DNA was deciphered only after many types of experimental evidence and theoretical considerations were combined. The crucial evidence was obtained by *X-ray crystallography* (Figure 11.4). Some chemical substances, when they are isolated and purified, can be made to form crystals. The positions of atoms in a crystalline substance can be inferred from the pattern of diffraction of X-rays passed through it. Even today, however, this is not an easy task when the substance is of enormous molecular weight.

In the early 1950s, even highly talented X-ray crystallographers could (and did) look at the best available images from DNA preparations and fail to see what they meant. Nonetheless, the attempt to characterize DNA would have been impossible without the crystallographs prepared by the English chemist Rosalind Franklin. Franklin's work, in turn, depended on the success of the English biophysicist Maurice Wilkins, who prepared a sample containing very uniformly oriented DNA fibers. These DNA preparations provided samples for diffraction that were far better than previous ones.



11.5 Chargaff's Rule In DNA, the total abundance of purines is equal to the total abundance of pyrimidines.

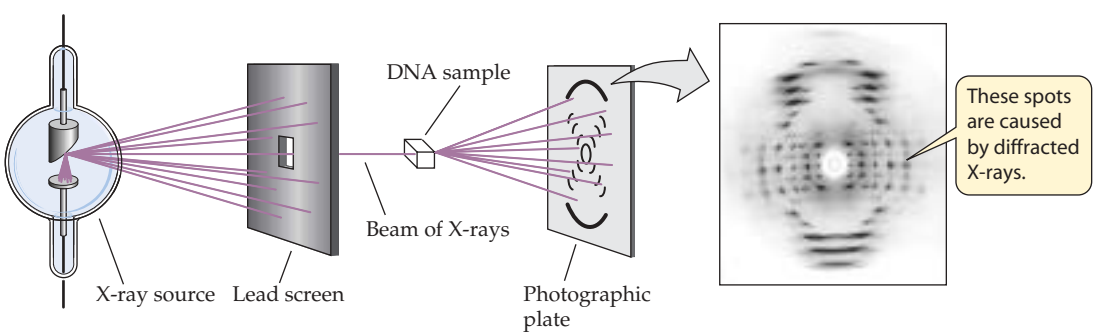
The chemical composition of DNA was known

The chemical composition of DNA also provided important clues about its structure. Biochemists knew that DNA was a polymer of nucleotides. Each nucleotide of DNA consists of a molecule of the sugar deoxyribose, a phosphate group, and a nitrogen-containing base (see Figures 3.24 and 3.25). The only differences among the four nucleotides of DNA are their nitrogenous bases: the purines **adenine** (**A**) and **guanine** (**G**), and the pyrimidines **cytosine** (**C**) and **thymine** (**T**).

In 1950, Erwin Chargaff at Columbia University reported some observations of major importance. He and his colleagues found that DNA from many different species—and from different sources within a single organism—exhibits certain regularities. In almost all DNA, the following rule holds: The amount of adenine equals the amount of thymine (A = T), and the amount of guanine equals the amount of cytosine (G = C) (Figure 11.5). As a result, the total abundance of purines (A + G) equals the total abundance of pyrimidines (T + C). The structure of DNA could not have been worked out without this information, now known as *Chargaff's rule*, yet its significance was overlooked for at least three years.

Watson and Crick described the double helix

The solution to the puzzle of the structure of DNA was accelerated by *model building:* the assembly of three-dimensional representations of possible molecular structures using known relative molecular dimensions and known bond angles. This technique, originally exploited in structural stud-



11.4 X-Ray Crystallography Revealed the Basic Helical Structure of the DNA Molecule The positions of atoms in a purified chemical substance can be inferred by the pattern of diffraction of X-rays passed through it, although the task requires tremendous skill.

ies by the American chemist Linus Pauling, was used by the English physicist Francis Crick and the American geneticist James D. Watson (Figure 11.6*a*), then both at the Cavendish Laboratory of Cambridge University.

Watson and Crick attempted to combine all that had been learned so far about DNA structure into a single coherent model. The crystallographers' results (see Figure 11.4) convinced Watson and Crick that the DNA molecule is **helical** (cylindrically spiral) and provided the values of certain distances within the helix. The results of density measurements and previous model building suggested that there are two polynucleotide chains in the molecule. Modeling studies had also led to the conclusion that the two chains in DNA run in opposite directions—that is, that they are **antiparallel**. (We'll clarify this point on the next page.)

Crick and Watson built several large models. Late in February of 1953, they built a model out of tin that established the general structure of DNA. This structure explained all the known chemical properties of DNA, and it opened the door to understanding its biological functions. There have been minor amendments to that first published structure, but its principal features remain unchanged.

Four key features define DNA structure

Four features summarize the molecular architecture of the DNA molecule:

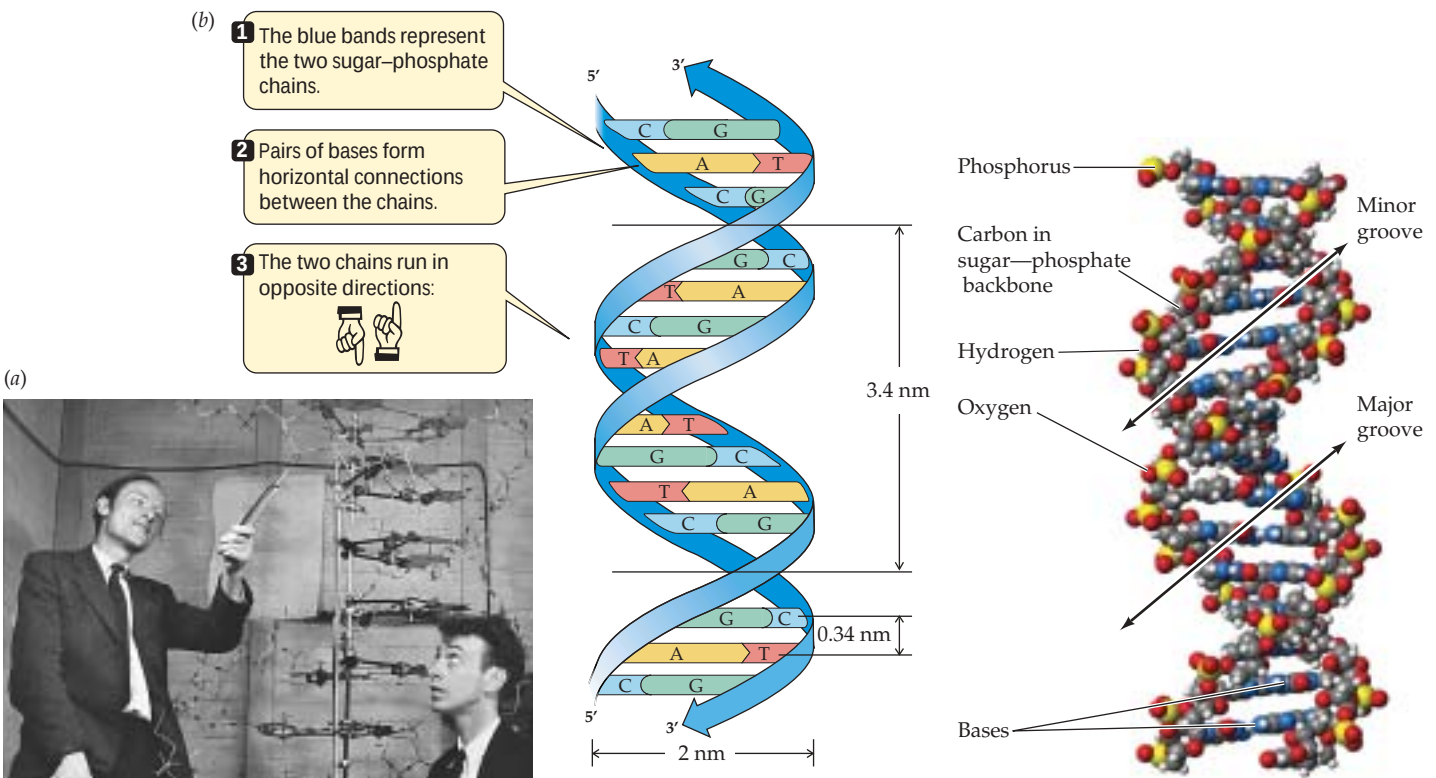
- It is a double-stranded helix.
- It has a uniform diameter.
- It is right-handed (that is, it twists to the right, as do the threads on most screws).
- It is antiparallel (the two strands run in opposite directions).

The sugar–phosphate "backbones" of the polynucleotide chains coil around the outside of the helix, and the nitrogenous bases point toward the center (Figure 11.6*b*).

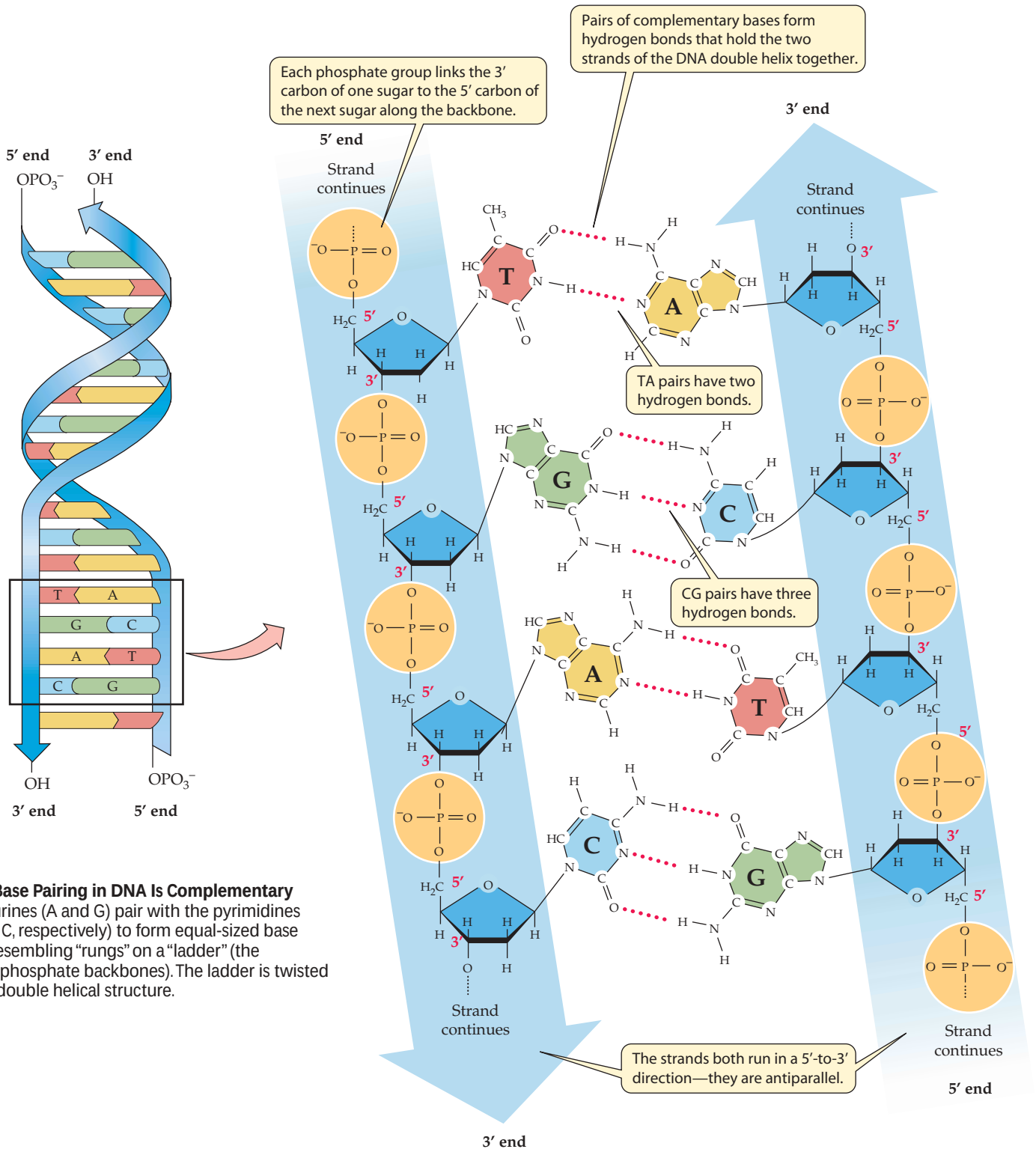
The two chains are held together by hydrogen bonding between specifically paired bases. Consistent with Chargaff's rule,

- adenine (A) pairs with thymine (T) by forming two hydrogen bonds; and
- guanine (G) pairs with cytosine (C) by forming three hydrogen bonds.

Every base pair consists of one purine (A or G) and one pyrimidine (T or C). This pattern is known as **complementary base pairing** (Figure 11.7).



11.6 DNA Is a Double Helix (*a*) Francis Crick and James Watson proposed that the DNA molecule has a double helical structure. (*b*) Biochemists can now pinpoint the position of every atom in a DNA macromolecule. To see that the essential features of the original Watson–Crick model have been verified, follow with your eyes the double helical chains of sugar–phosphate groups and note the horizontal rungs of the bases (see also Figure 3.27).



11.7 Base Pairing in DNA Is Complementary
The purines (A and G) pair with the pyrimidines (T and C, respectively) to form equal-sized base pairs resembling "rungs" on a "ladder" (the sugar–phosphate backbones). The ladder is twisted into a double helical structure.

Because the AT and GC pairs are of equal length and fit identically into the double helix (like rungs on a ladder), the diameter of the helix is uniform. The base pairs are flat (see Figure 11.6), and their stacking in the center of the molecule is stabilized by hydrophobic interactions (see Chapter 2), contributing to the overall stability of the double helix.

What does it mean to say that the two DNA strands are *antiparallel*? The direction of a polynucleotide can be defined by looking at the phosphodiester linkages between adjacent nucleotides (*-diester* refers to the two bonds formed by —OH groups reacting with acidic *phosphate* groups). In the sugar–phosphate backbone of DNA, the phosphate groups connect to the 3′ carbon of one deoxyribose molecule and the 5′ carbon of the next, linking successive sugars together (see Figure 11.7). The number followed by a prime (′) designates the position of a carbon atom in the five-carbon sugar deoxyribose.

Thus the two ends of a polynucleotide chain differ. At one end of a strand is a free (not connected to another nucleotide) 5′ phosphate group (—OPO_3^-); this end is called the 5′ end. At the other end is a free 3′ hydroxyl group (—OH); this end is called the 3′ end. In a DNA double helix, the 5′ end of one strand is paired with the 3′ end of the other strand, and vice versa. In other words, were you to draw an arrow for each strand running from 5′ to 3′, the arrows would point in different directions; thus it is said that the strands are antiparallel.

The double helical structure of DNA is essential to its function

The genetic material performs four important functions, and the DNA structure proposed by Watson and Crick was elegantly suited to three of them.

- *The genetic material stores an organism's genetic information*. With its millions of nucleotides, the base sequence of a DNA molecule could encode and store an enormous amount of information and could account for species and individual differences. DNA fits this role nicely.
- *The genetic material is susceptible to mutation*, or permanent changes in the information it encodes. For DNA, mutations might be simple changes in the linear sequence of base pairs.
- *The genetic material is precisely replicated* in the cell division cycle. Replication could be accomplished by complementary base pairing, A with T and G with C. In the original publication of their findings in the journal *Nature* in 1953, Watson and Crick coyly pointed out, "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material."
- *The genetic material is expressed as the phenotype*. This function is not obvious in the structure of DNA. However, as we will see in the next chapter, the nucleotide sequence of DNA is copied into RNA, which is in turn converted into a linear sequence of amino acids—a protein. The folded forms of proteins provide much of the phenotype of an organism.

Determining the DNA Replication Mechanism

The mechanism of DNA replication that had suggested itself to Watson and Crick was soon confirmed. First, experiments showed that single strands of DNA could be replicated in a test tube containing simple substrates and an enzyme. Then an elegant experiment showed that each of the two strands of the double helix serves as a template for a new strand of DNA.

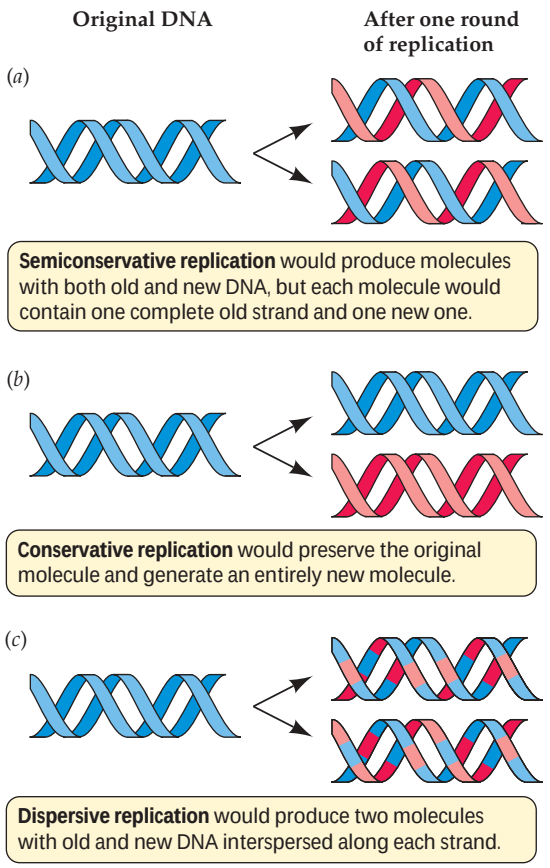
Three modes of DNA replication appeared possible

The prediction that the DNA molecule contains the information needed for its own replication was demonstrated by the work of Arthur Kornberg, then at Washington University in St. Louis. He showed that DNA can be synthesized in a test tube containing just three substances:

- The substrates, deoxyribonucleoside triphosphates dATP, dCTP, dGTP, and dTTP
- The enzyme **DNA polymerase**
- DNA, which serves as a **template** to guide the incoming nucleotides

There were three possible patterns that could result in complementary base pairing during DNA replication:

- *Semiconservative replication*, in which each parent strand serves as a template for a new strand, and the two new DNAs each have one old and one new strand (Figure 11.8*a*)
- *Conservative replication*, in which the original double helix serves as a template for, but does not contribute to, a new double helix (Figure 11.8*b*)
- *Dispersive replication*, in which fragments of the original DNA molecule serve as templates for assembling two



11.8 Three Models for DNA Replication In each model, original DNA is shown in blue and newly synthesized DNA in red.

new molecules, each containing old and new parts, perhaps at random (Figure 11.8*c*)

Watson and Crick's original paper suggested that DNA replication was semiconservative, but Kornberg's experiment did not provide a basis for choosing among these three models.

Meselson and Stahl demonstrated that DNA replication is semiconservative

A clever experiment conducted by Matthew Meselson and Franklin Stahl convinced the scientific community that **semiconservative replication** is the correct model. Working at the California Institute of Technology in 1957, they devised a simple way to distinguish old strands of DNA from new ones: *density labeling*.

The key to their experiment was the use of a "heavy" isotope of nitrogen. Heavy nitrogen (^{15}N) is a rare, nonradioactive isotope that makes molecules containing it more dense than chemically identical molecules containing the common isotope, ^{14}N. To distinguish DNA of different densities (that is, DNA containing ^{15}N versus DNA containing ^{14}N), Meselson, Stahl, and Jerome Vinograd developed a new procedure using a *centrifuge*. Spinning solutions or suspensions at high speed in a centrifuge causes the solutes or particles to separate and form a gradient according to their density.

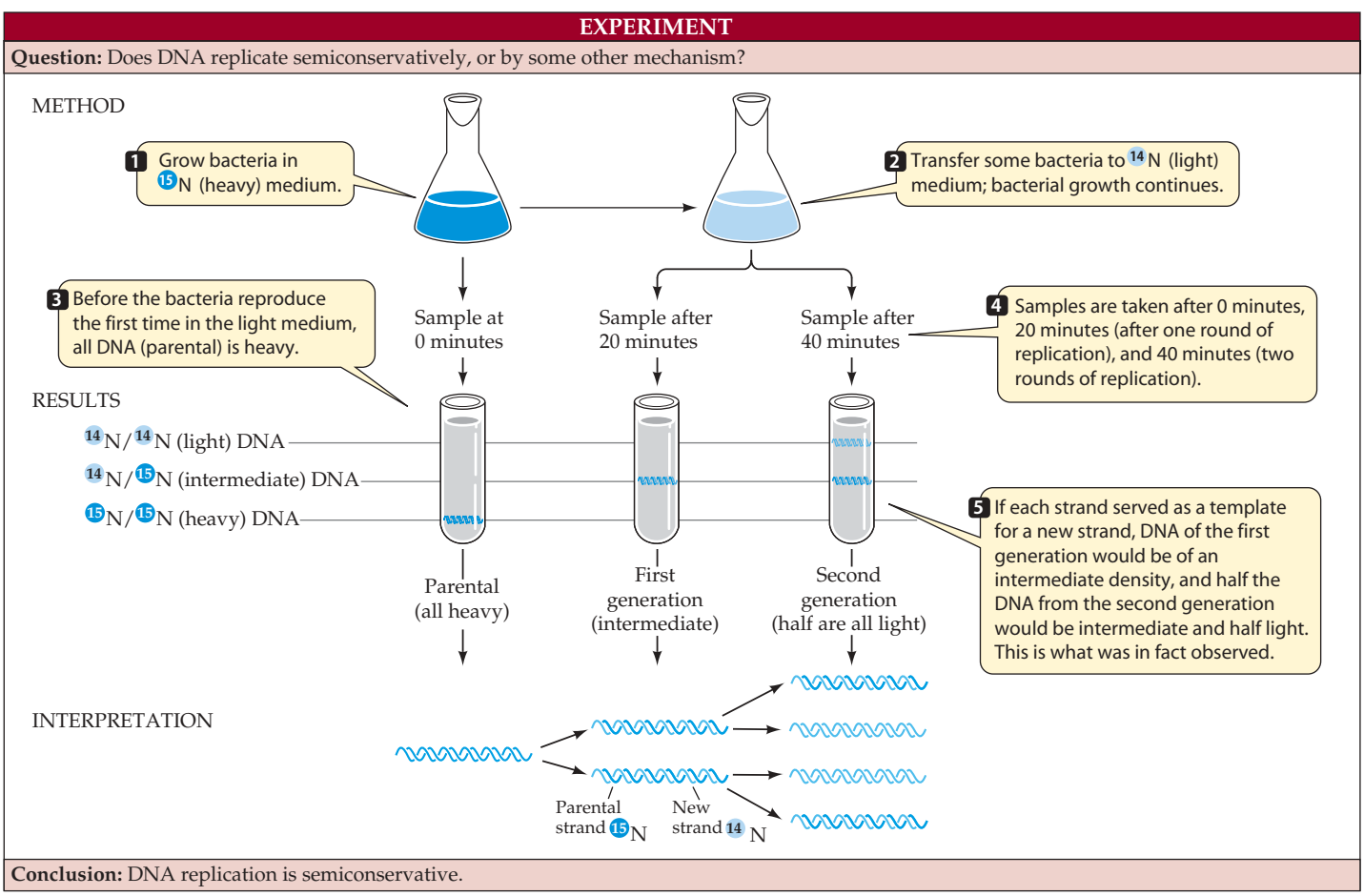
Meselson and Stahl grew two cultures of the bacterium *Escherichia coli* for many generations:

- One culture was grown in a medium whose nitrogen source (ammonium chloride, NH_4Cl) was made with ^{15}N instead of ^{14}N. As a result, all the DNA in the bacteria was "heavy."
- Another culture was grown in a medium with ^{14}N, and all the DNA in these bacteria was "light."

When extracts from the two cultures were combined and centrifuged, two separate DNA bands formed, showing that this method could distinguish DNA samples of slightly different densities.

Next, the researchers grew another *E. coli* culture on ^{15}N medium, then transferred it to normal ^{14}N medium and allowed the bacteria to continue growing (Figure 11.9). Under

11.9 The Meselson–Stahl Experiment A centrifuge was used to separate DNAs labeled with isotopes of different densities. This experiment revealed a pattern that supports the semiconservative model of DNA replication.



the conditions they used, *E. coli* replicates its DNA every 20 minutes. Meselson and Stahl collected some of the bacteria after each division and extracted DNA from the samples. They found that the DNA banding pattern in the density gradient was different in each bacterial generation:

- At the time of the transfer to the ^{14}N medium, the DNA was uniformly labeled with ^{15}N, and hence was relatively dense.
- After one generation, when the DNA had been duplicated once, all the DNA was of an intermediate density.
- After two generations, there were two equally large DNA bands: one of low density and one of intermediate density.
- In samples from subsequent generations, the proportion of low-density DNA increased steadily.

The results of this experiment can be explained only by the semiconservative model of DNA replication. In the first round of DNA replication, the strands of the double helix—both heavy with ^{15}N—separated. Each strand then acted as the template for a second strand, which contained only ^{14}N and hence was less dense. Each double helix then consisted of one ^{15}N strand and one ^{14}N strand, and was of intermediate density. In the second replication, the ^{14}N-containing strands directed the synthesis of partners with ^{14}N, creating low-density DNA, and the ^{15}N strands formed new ^{14}N partners (see Figure 11.9).

The crucial observation demonstrating the semiconservative model was that intermediate-density DNA (^{15}N–^{14}N) appeared in the first generation and continued to appear in subsequent generations. With the other models, the results would have been quite different (see Figure 11.8):

- In conservative replication, the first generation would have had both high-density DNA (^{15}N–^{15}N) and low-density DNA (^{14}N–^{14}N), but no intermediate-density DNA.
- In dispersive replication, the density of the new DNA would have been half that of parental DNA.

The Meselson-Stahl experiment, called by some scientists among the most elegant ever done by biologists, was an excellent example of the scientific method. It began with three hypotheses—the three models of DNA replication—and was designed so that the results could differentiate between them.

The Molecular Mechanisms of DNA Replication

Semiconservative DNA replication in the cell involves a number of different enzymes and other proteins. It takes place in two steps:

- The DNA is unwound to separate the two template strands and make them available for base pairing.
- New nucleotides are linked by covalent bonding to each growing new strand in a sequence determined by complementary base pairing with the bases on the template strand.

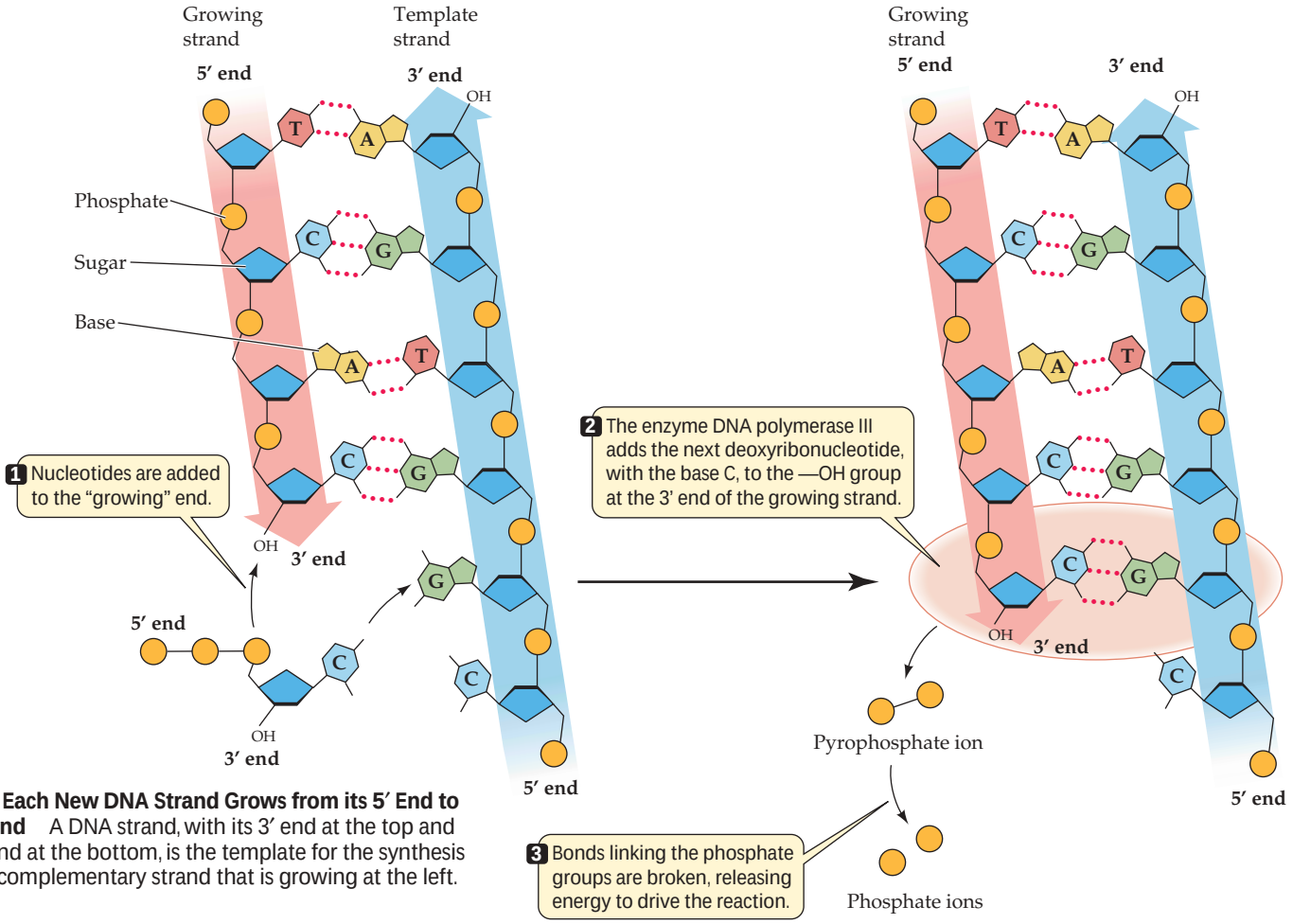
A key observation of virtually all DNA replication is that *nucleotides are always added to the growing strand at the 3′ end*—the end at which the DNA strand has a free hydroxyl (—OH) group on the 3′ carbon of its terminal deoxyribose (Figure 11.10). The three phosphate groups in a deoxyribonucleoside triphosphate are attached to the 5′ position of the sugar (see Figure 11.7). So when a new nucleotide is added to DNA, it can attach only to the 3′ end.

When DNA polymerase brings a deoxyribonucleoside triphosphate with the appropriate base to the 3′ end of a growing chain, the free hydroxyl group on the chain reacts with one of the substrate's phosphate groups. As this happens, the bond linking the terminal two phosphate groups to the rest of the deoxyribonucleoside triphosphate breaks, and stored energy is released as the phosphate groups separate from the molecule. The resulting *pyrophosphate ion*, consisting of the two terminal phosphate groups, also hydrolyzes, forming two separate phosphate ions and in the process releasing additional free energy. The phosphate group still on the nucleotide becomes part of the sugar–phosphate backbone of the growing DNA molecule.

DNA is threaded through a replication complex

DNA is replicated through the interaction of the template DNA with a huge protein complex called the **replication complex**, which catalyzes the reactions involved. All chromosomes have at least one base sequence, called the **origin of replication**, to which this replication complex initially binds. DNA replicates *in both directions* from the origin of replication, forming two **replication forks** (Figure 11.11). Both of the separated strands of the parent molecule act as templates simultaneously, and the formation of the new strands is guided by complementary base pairing.

Until recently, DNA replication was depicted as a locomotive (the replication complex) moving along a railroad track (the DNA) (Figure 11.11*a*). The current view is that this model may not be correct. Instead, the replication complex seems to be stationary, attached to nuclear structures, and it is the DNA that moves, essentially threading through the complex as single strands and emerging as double strands (Figure 11.11*b*). During S phase in eukaryotes, there are about 100 replication complexes, and each of them contains as many as 300 individual replication forks. All replication complexes contain several proteins with different roles in DNA replication; we will describe these proteins as we examine the steps of the process.

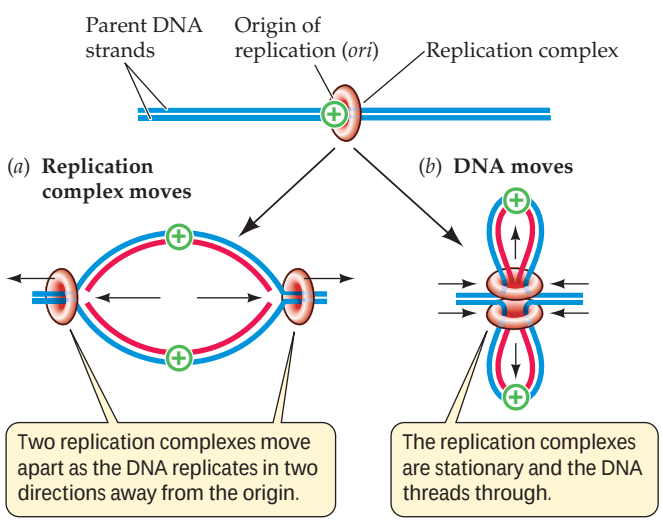


11.10 Each New DNA Strand Grows from its 5′ End to its 3′ End A DNA strand, with its 3′ end at the top and its 5′ end at the bottom, is the template for the synthesis of the complementary strand that is growing at the left.

Small, circular DNAs replicate from a single origin

The first event at the origin of replication is the localized unwinding (denaturation) of DNA. There are several forces that hold the two strands together, including hydrogen bonding and the hydrophobic interactions of bases. An enzyme called **DNA helicase** uses energy from ATP hydrolysis to unwind the DNA, and special proteins called **single-strand binding proteins** bind to the unwound strands to keep them from reassociating into a double helix. This process makes the two template strands available for complementary base pairing.

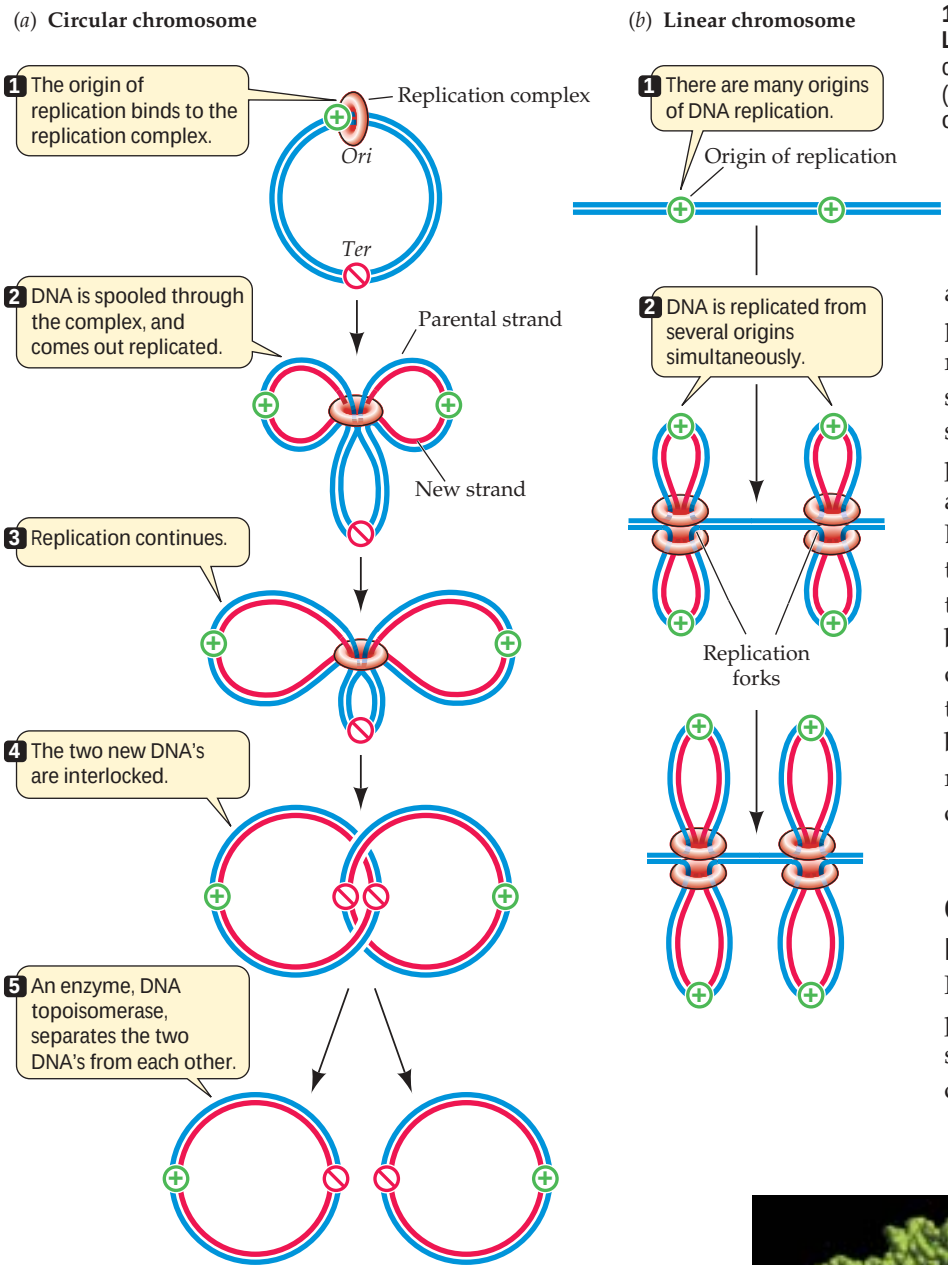
Small circular chromosomes, such as the 3-million-base-pair DNA of bacteria, have a single origin of replication. As the DNA moves through the replication complex, the replication forks grow around the circle (Figure 11.12*a*). Two interlocking circular DNAs are formed, and they are separated by an enzyme called **DNA topoisomerase**.



11.11 Two Views of DNA Replication (*a*) It was once thought that the replication complex moved along DNA. (*b*) Newer evidence suggests that the DNA is threaded through the stationary complex.

Large, linear DNAs have many origins

In large linear chromosomes, such as a human chromosome with 80 million base pairs, there are hundreds of origins of replication. Origins of replication that are adjacent to one another along the linear chromosome can be bound by replication complexes at the same time and replicated simultaneously. So there are many replication forks in eukaryotic DNA (Figure 11.12*b*).



11.12 Replication in Small Circular and Large Linear Chromosomes (*a*) Small circular chromosomes have a single origin of replication. (*b*) Larger linear chromosomes have many origins of replication.

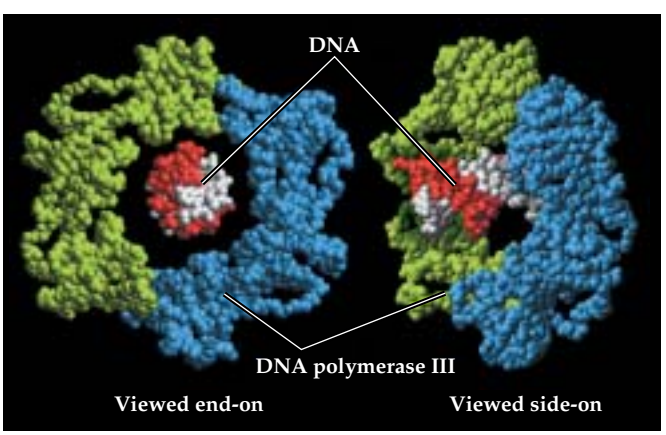
DNA polymerases need a primer

DNA polymerases are much larger than their substrates, the deoxyribonucleoside triphosphates, and the template DNA, which is very thin. Molecular models of the enzyme-substrate-template complex from bacteria (Figure 11.13) show that the enzyme is shaped like an open hand with a palm, a thumb, and fingers. The palm holds the active site of the enzyme and brings together the substrate and the template. The finger regions rotate inward and have precise shapes that can recognize the different shapes of the four nucleotide bases.

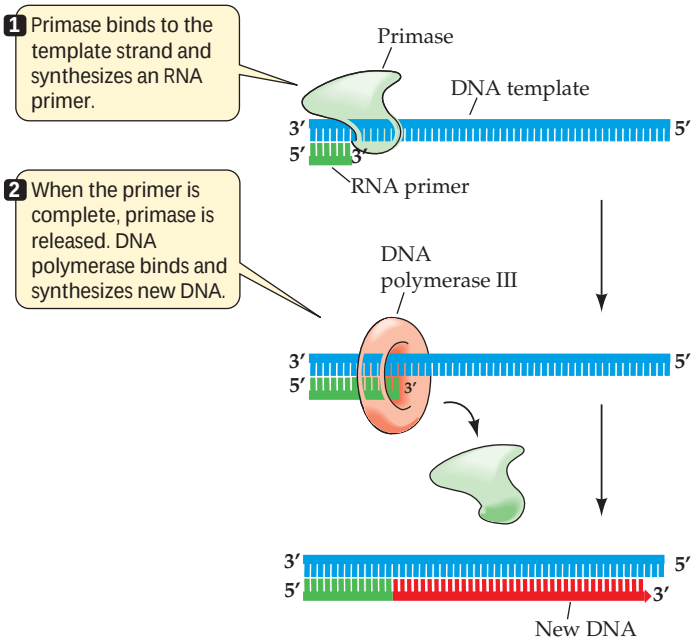
DNA polymerases can elongate a polynucleotide strand by covalently linking new nucleotides to a previously existing strand, but they cannot start a strand from scratch. Therefore, a "starter" strand of DNA or RNA, called a **primer**, is required for replication. In DNA replication, the primer is a short single strand of RNA (Figure 11.14). This RNA strand, complementary to the DNA template strand, is synthesized one nucleotide at a time by an enzyme called **primase**. DNA polymerase then adds nucleotides to the 3′ end of the primer and continues until the replication of that section of DNA has been completed. Then the RNA primer is degraded, DNA is added in its place, and the resulting DNA fragments are connected by the action of other enzymes. When DNA replication is complete, each new strand consists only of DNA.

Cells contain several different DNA polymerases

Most cells contain more than one DNA polymerase, but only one of them is responsible for chromosomal DNA replication. The others are involved in primer



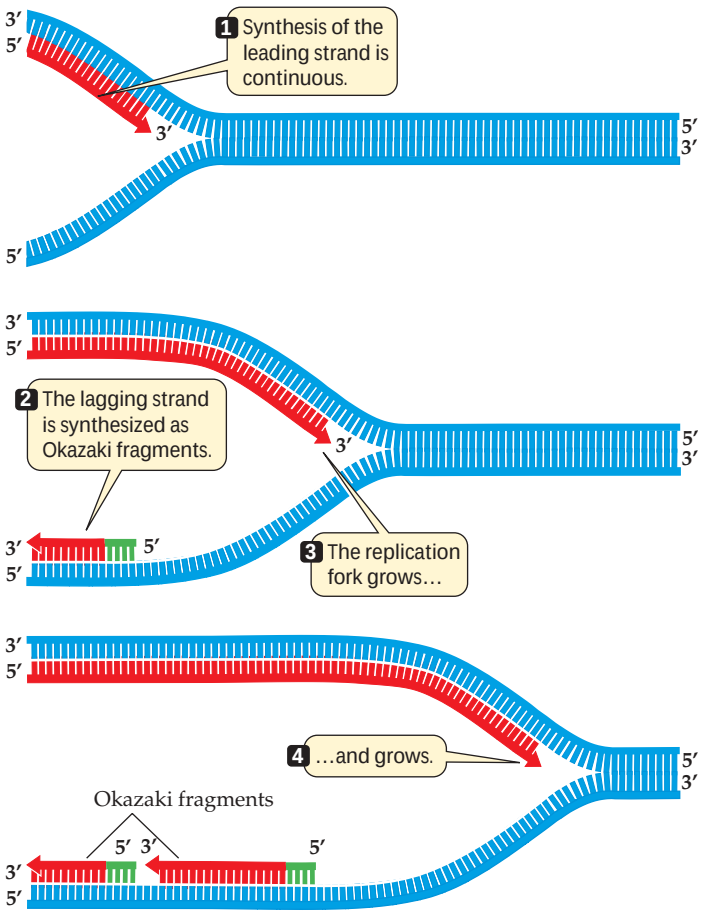
11.13 DNA Polymerase Binds to the Template Strand The DNA polymerase enzyme (blue and green) is much larger than the DNA molecule. DNA polymerase III is shaped like a hand, and in the side-on view, its "fingers" can be seen curling around the DNA. These "fingers" can recognize the different shapes of the four bases (white; the DNA "backbone" is shown in red).



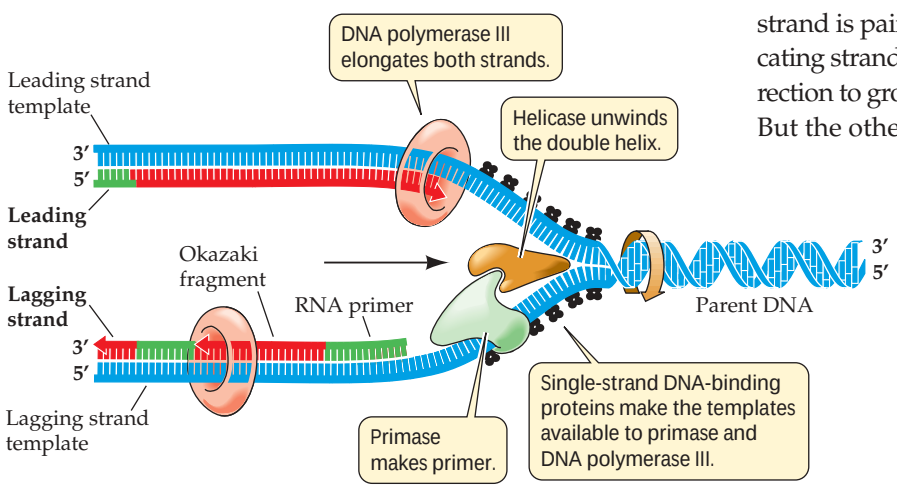
11.14 No DNA Forms without a Primer DNA polymerases require a primer—a "starter" strand of DNA or RNA to which they can add new nucleotides.

removal and DNA repair. Fourteen DNA polymerases have been identified in humans; the one catalyzing most replication is DNA polymerase α. In the bacterium *E. coli*, there are three DNA polymerases; the one responsible for replication is DNA polymerase III. Various other proteins play roles in replacing the RNA primer and in other replication tasks; some of these are shown in Figure 11.15.

The DNA at the replication fork opens up like a zipper in one direction. Study Figure 11.16 and try to imagine what is happening over a short period of time. Remember that in DNA the two strands are antiparallel; that is, the 3′ end of one



11.16 The Two New Strands Form in Different Ways As the template DNA unwinds, both new strands are synthesized in the 5′-to-3′ direction, although their template strands are antiparallel. The leading strand grows continuously forward, but the lagging strand grows in short discontinuous stretches called Okazaki fragments. Eukaryotic Okazaki fragments are hundreds of nucleotides long, with gaps between them.

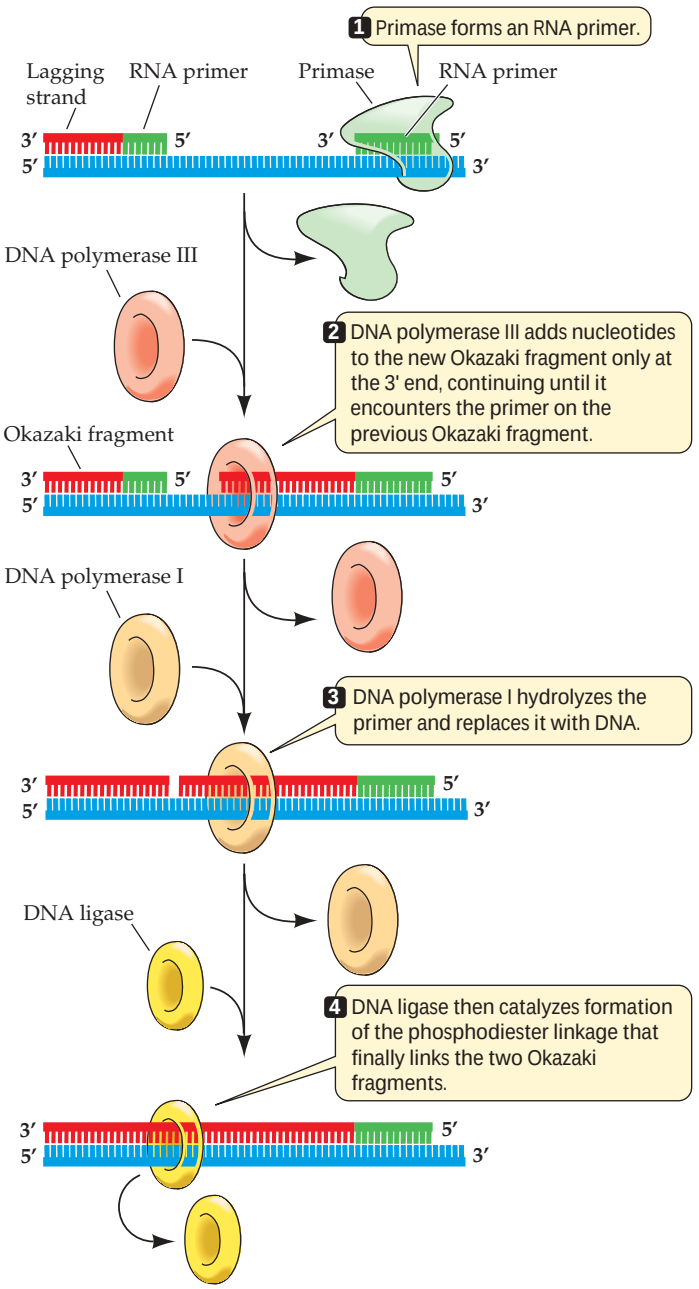


11.15 Many Proteins Collaborate at the Replication Fork Several proteins in addition to DNA polymerase III are involved in DNA replication. The two molecules of DNA polymerase (red) are actually part of the same complex.

strand is paired with the 5′ end of the other. One newly replicating strand (the **leading strand**) is pointing in the "right" direction to grow continuously at its 3′ end as the fork opens up. But the other strand (the **lagging strand**) is pointing in the "wrong" direction: As the fork opens up further, its exposed 3′ end gets farther and farther away from the fork, and an unreplicated gap is formed, which would get bigger and bigger if there were not a special mechanism to overcome this problem.

The lagging strand is synthesized from Okazaki fragments

Synthesis of the lagging strand requires working in relatively small, discontinuous stretches (100 to 200 nucleotides at a time in eukary-



11.17 The Lagging Strand Story In bacteria, DNA polymerase I and DNA ligase cooperate with DNA polymerase III to complete the complex task of synthesizing the lagging strand.

otes; 1,000 to 2,000 at a time in prokaryotes). These discontinuous stretches are synthesized just as the leading strand is, by the addition of new nucleotides one at a time to the 3′ end of the new strand, but the synthesis of this new strand moves in the direction opposite to that in which the replication fork is moving. These stretches of new DNA for the lagging strand are called **Okazaki fragments**, after their discoverer, the Japanese biochemist Reiji Okazaki. While the leading strand grows continuously "forward," the lagging strand grows in shorter, "backward" stretches with gaps between them.

A single primer suffices for synthesis of the leading strand, but each Okazaki fragment requires its own primer. In bacteria, DNA polymerase III synthesizes Okazaki fragments by adding nucleotides to a primer until it reaches the primer of the previous fragment. At this point, DNA polymerase I (the one discovered by Kornberg) removes the old primer and replaces it with DNA. Left behind is a tiny nick—the final phosphodiester linkage between the adjacent Okazaki fragments is missing. The enzyme **DNA ligase** catalyzes the formation of that bond, linking the fragments and making the lagging strand whole (Figure 11.17).

Working together, DNA helicase, the two DNA polymerases, primase, DNA ligase, and the other proteins of the replication complex do the job of DNA synthesis with a speed and accuracy that are almost unimaginable. In *E. coli*, the replication complex makes new DNA at a rate in excess of 1,000 base pairs per second, committing errors in fewer than one base in 10^6, or one in a million.

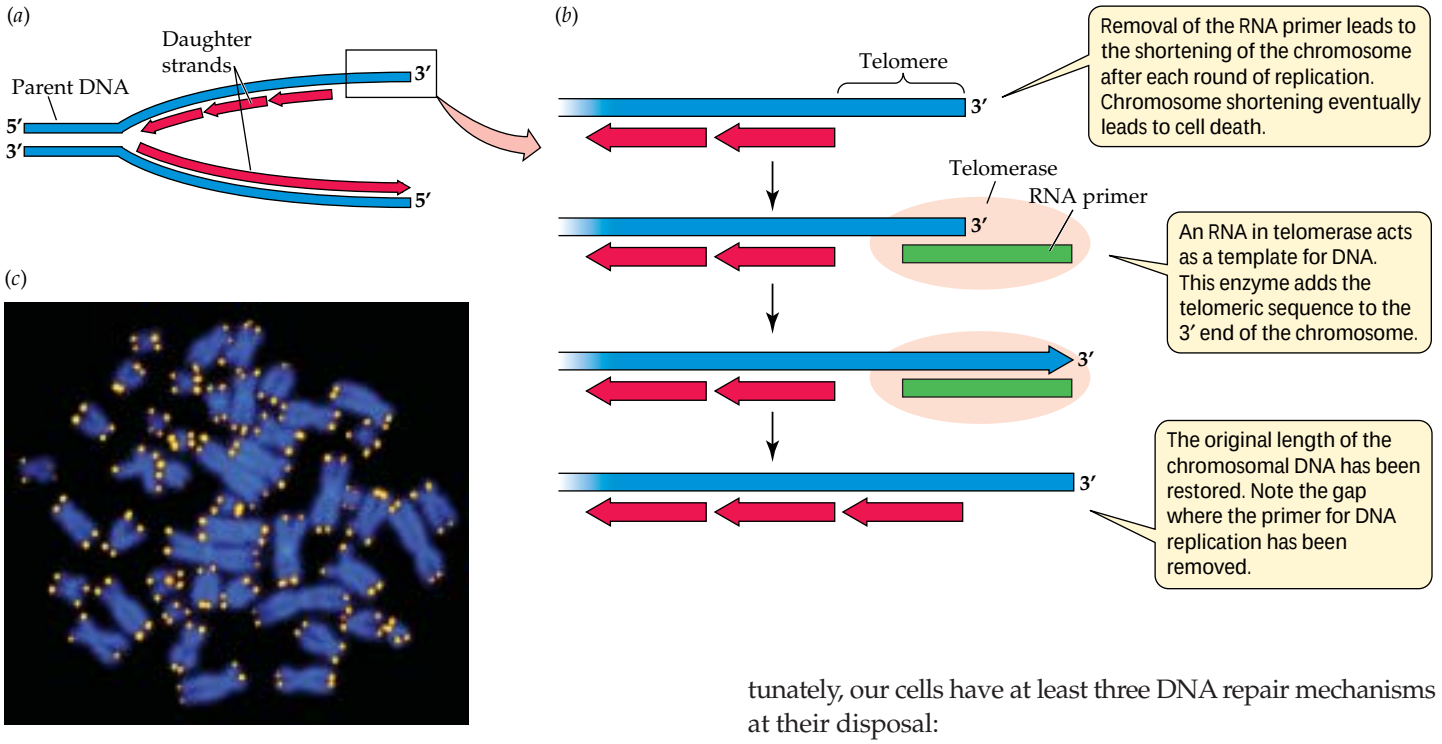
Telomeres are not fully replicated

As we have just seen, replication of the lagging strand occurs by the addition of Okazaki fragments to RNA primers. Beyond the very end of a linear DNA molecule, however, there is no place for a primer to bind (i.e., there is no complementary DNA strand). So the new chromosome formed after DNA replication has a bit of single-stranded DNA at each end (Figure 11.18*a*). This situation activates mechanisms that cut off the single-stranded region, along with some of the intact double-stranded end. Thus, the chromosome becomes slightly shorter with each cell division.

In many eukaryotes, there are repetitive sequences at the ends of chromosomes called **telomeres**. In humans, the telomere sequence is TTAGGG, and it is repeated about 2,500 times. These repeats bind special proteins that maintain the stability of chromosome ends. Each human chromosome can lose 50–200 base pairs of telomeric DNA after each round of DNA replication and cell division. After 20–30 divisions, the chromosomes are unable to take part in cell division, and the cell dies. This phenomenon explains in part why cells do not last the entire lifetime of the organism: Their telomeres shorten.

Yet constantly dividing cells, such as bone marrow and germ line cells, maintain their telomeric DNA. An enzyme, appropriately called **telomerase**, catalyzes the addition of any lost telomeric sequences (Figure 11.18*b*). Telomerase contains an RNA sequence that acts as a template for the telomeric repeat sequence.

Telomerase is expressed in more than 90 percent of human cancers and may be an important factor in the ability of can-



11.18 Telomeres and Telomerase (*a*) Removal of RNA primer at the 3′ end of the lagging strand leaves a region of DNA unreplicated. *(b)* The enzyme telomerase binds to the 3′ end and extends the lagging strand of DNA. An RNA sequence embedded in telomerase provides a template so that, overall, the DNA does not get shorter. *(c)* Bright fluorescent staining marks the telomeric regions on these blue-stained human chromosomes.

cer cells to divide continuously. Since most normal cells do not have this ability, telomerase is an attractive target for drugs designed to attack tumors specifically.

There is also interest in telomerase and aging. When a gene expressing high levels of telomerase is added to human cells in culture, their telomeres do not shorten. Instead of dying after 20–30 cell generations, the cells become immortal. It remains to be seen how this finding relates to the aging of a large organism.

DNA Proofreading and Repair

DNA is accurately replicated and faithfully maintained. The price of failure can be great: the transmission of genetic information is at stake, as is the functioning and even the life of a cell or multicellular organism. Yet the replication of DNA is not perfectly accurate, and the DNA of nondividing cells is subject to damage by environmental agents. In the face of these threats, how has life gone on so long?

The preservers of life are DNA repair mechanisms. DNA polymerases initially make a significant number of mistakes in assembling polynucleotide strands. The observed error rate of one for every 10^6 bases replicated would result in about 1,000 mutations every time a human cell divided. Fortunately, our cells have at least three DNA repair mechanisms at their disposal:

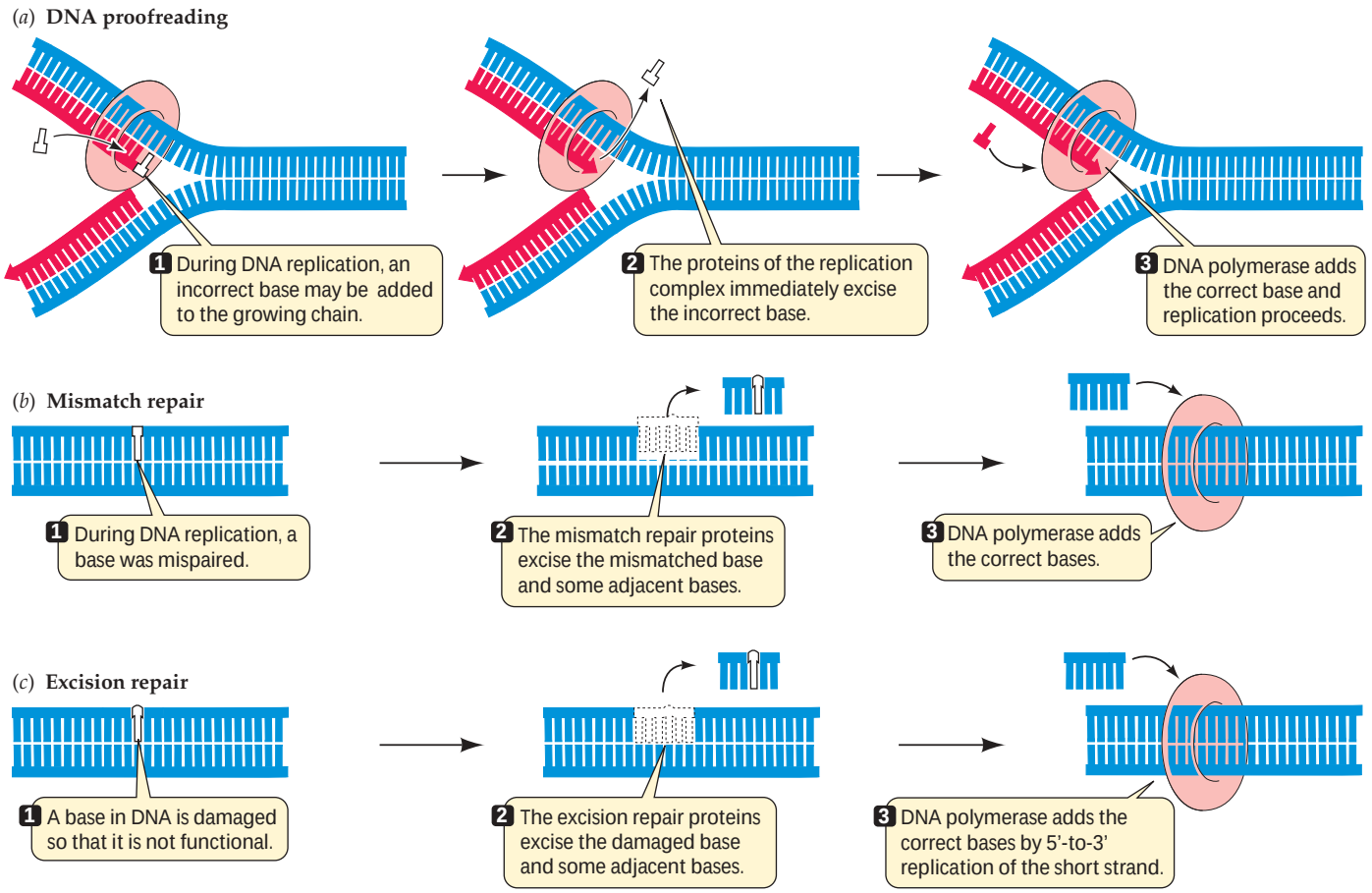
- A **proofreading** mechanism corrects errors in replication as DNA polymerase makes them.
- A **mismatch repair** mechanism scans DNA immediately after it has been replicated and corrects any base-pairing mismatches.
- An **excision repair** mechanism removes abnormal bases that have formed because of chemical damage and replaces them with functional bases.

Proofreading mechanisms ensure that DNA replication is accurate

After introducing a new nucleotide into a growing polynucleotide strand, DNA polymerase performs a proofreading function (Figure 11.19*a*). When a DNA polymerase recognizes a mispairing of bases, it removes the improperly introduced nucleotide and tries again. (Other proteins of the replication complex also play roles in proofreading.) The error rate for this process is only about 1 in 10,000 base pairs, and lowers the overall error rate for replication to about one base in every 10^{10} bases replicated.

Mismatch repair mechanisms correct base-pairing errors

After DNA has been replicated, a second set of proteins surveys the newly replicated molecule and looks for remaining mismatched base pairs (Figure 11.19*b*). For example, this mismatch repair system might detect an AC base pair instead of an AT pair. But how does the repair system "know" whether the AC pair should be repaired by removing the C and replacing it with T or by removing the A and replacing it with G?



11.19 DNA Repair Mechanisms The proteins of the replication complex also play roles in the life-preserving DNA repair mechanisms, helping to ensure the exact replication of template DNA and repair any damage that occurs.

The repair mechanism can detect the "wrong" base because a DNA strand is chemically modified some time after replication. In prokaryotes, methyl groups (—CH_3) are added to some guanines. Immediately after replication, methylation has not yet occurred, so the newly replicated strand is "marked," by being unmethylated, as the one in which errors should be corrected.

When mismatch repair fails, DNA sequences are altered. One form of colon cancer arises in part from a failure of mismatch repair.

Excision repair mechanisms repair chemical damage

DNA molecules can also be damaged during the life of a cell (e.g., when it is in G1). High-energy radiation, chemicals from the environment, and random spontaneous chemical reactions can all damage DNA.

Certain enzymes constantly "inspect" the cell's DNA (Figure 11.19*c*). When they find mispaired bases, chemically modified bases, or points at which one strand has more bases than the other (with the result that one or more bases of one strand form an unpaired loop), these enzymes cut the defective strand. Another enzyme cuts away the bases adjacent to and including the offending base, and DNA polymerase and DNA ligase synthesize and seal up a new (usually correct) base sequence to replace the excised one.

Our dependence on excision repair is underscored by our susceptibility to diseases that arise from excision repair defects. One example is the skin disease xeroderma pigmentosum. Persons with this disease lack a mechanism that normally repairs the damage caused by ultraviolet radiation in sunlight; they develop skin cancers after even minute exposure to sunlight.

Practical Applications of DNA Replication

The principles underlying DNA replication in cells have been used to develop two laboratory techniques that have been vital in analyzing genes and genomes. The first technique allows researchers to make multiple copies of short DNA sequences, and the second allows them to determine the base sequence of a DNA molecule.

The polymerase chain reaction makes multiple copies of DNA

Since DNA can be replicated in the laboratory, it is possible to make multiple copies of a DNA sequence. The **polymerase chain reaction** (**PCR**) technique essentially automates this process by copying a short region of DNA many times in a test tube.

PCR is a cyclic process in which a sequence of steps is repeated over and over again (Figure 11.20):

- Double-stranded fragments of DNA are separated into single strands by mild heating (denatured).
- A short, artificially synthesized primer is added to the mixture, along with the four deoxyribonucleotide triphosphates (dATP, dGTP, dCTP, and dTTP) and DNA polymerase.
- DNA polymerase catalyzes the production of complementary new strands.

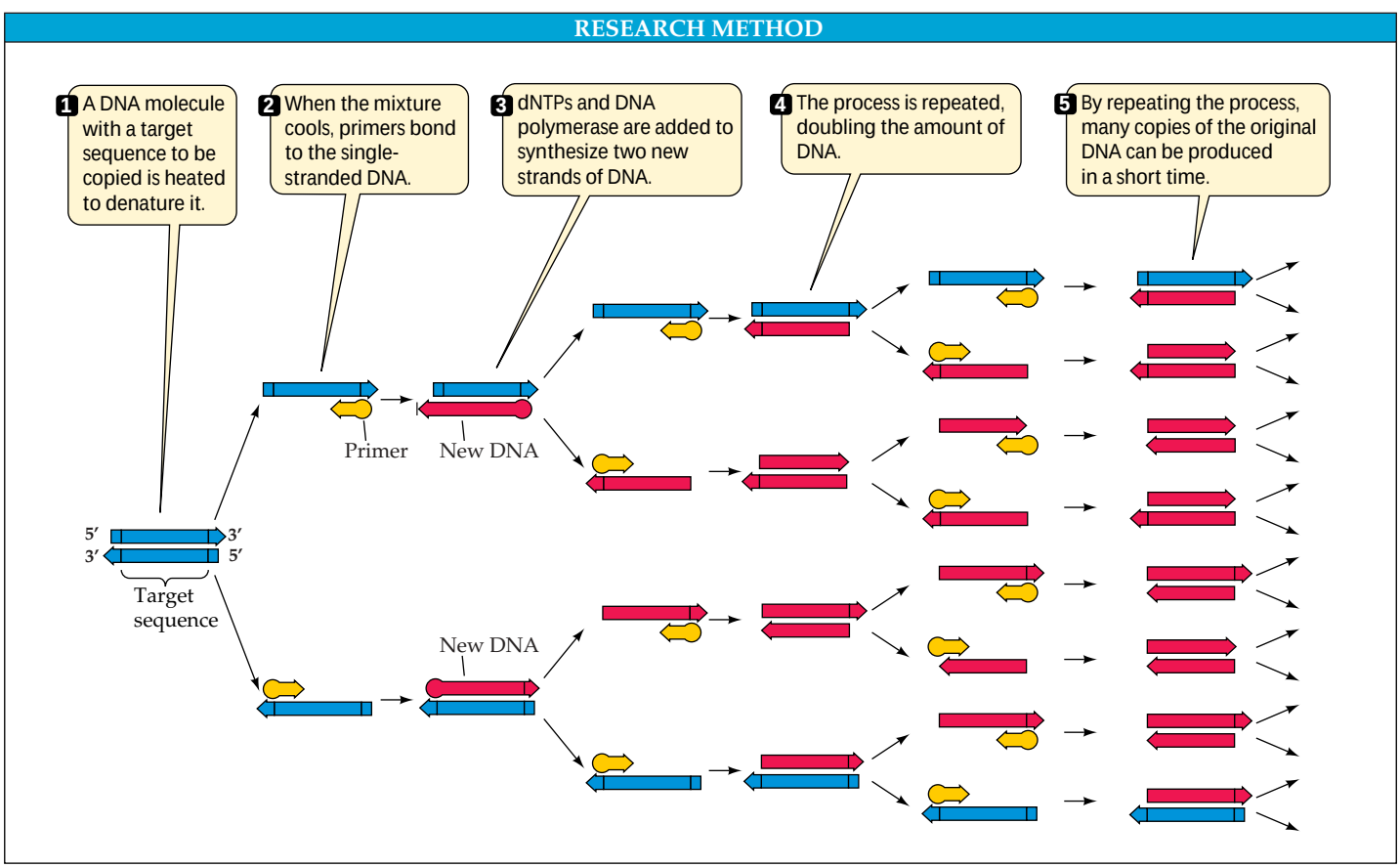
A single cycle takes a few minutes to double the amount of DNA, leaving the new DNA in the double-stranded state. Theoretically, repeating the cycle many times leads to an exponential increase in the number of copies of the DNA sequence.

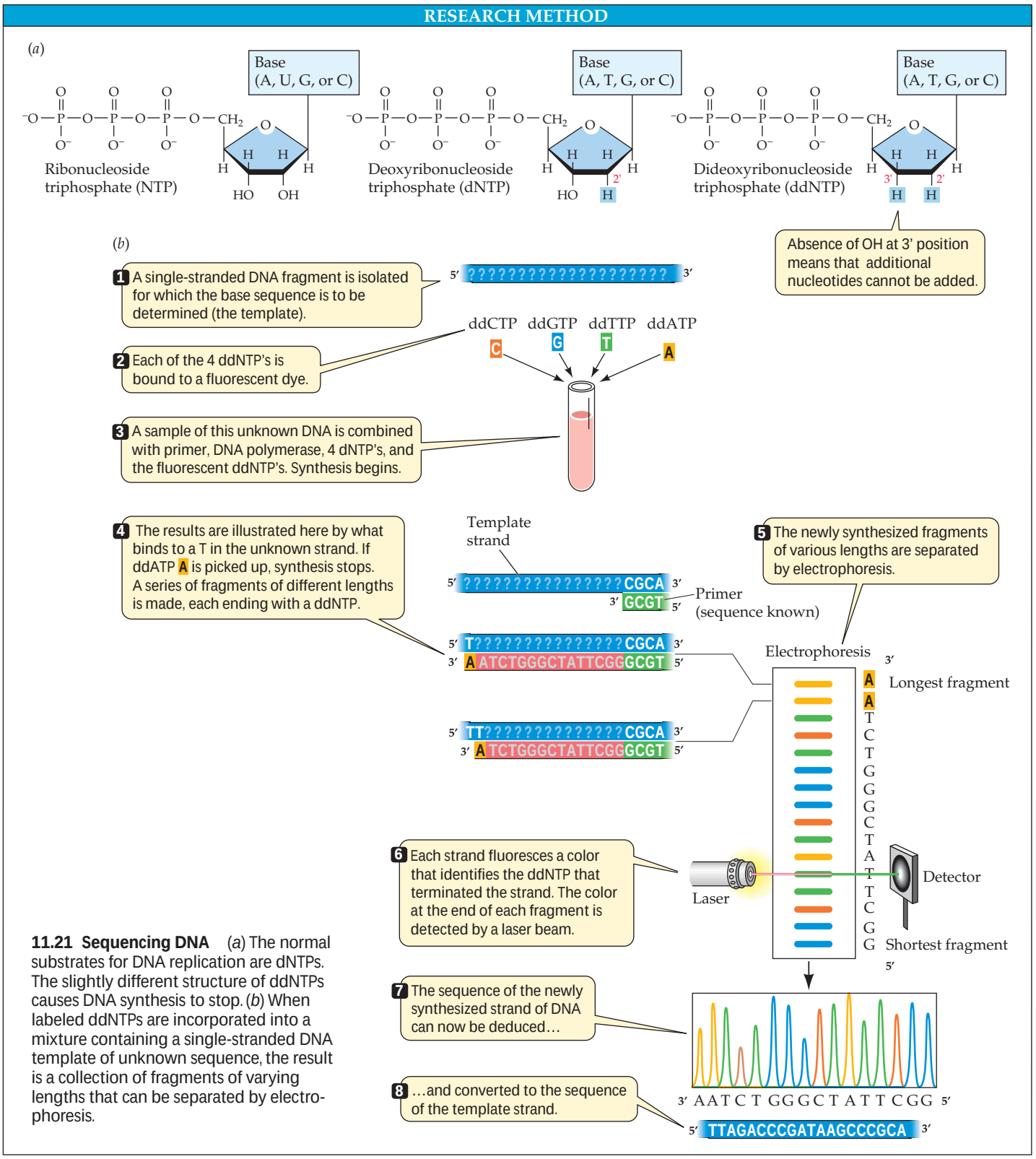
The PCR technique requires that the base sequences at the 3′ end of each strand of the target DNA sequence be known so that a complementary primer, usually 15–20 bases long, can be made in the laboratory. Because of the uniqueness of DNA sequences, usually only two primers of this length will bind to only one region of DNA in an organism's genome. This specificity in the face of the incredible diversity of target DNA is a key to the power of PCR.

One initial problem with PCR was its temperature requirements. To denature the DNA, it must be heated to more than 90°C—a temperature that destroys most DNA polymerases. The PCR method would not be practical if new polymerase had to be added after denaturation in each cycle.

This problem was solved by nature: In the hot springs at Yellowstone National Park, as well as other locations, lives a bacterium called, appropriately, *Thermus aquaticus*. The means by which this organism survives temperatures up to 95°C was investigated by Thomas Brock and his colleagues. They discovered that *T. aquaticus* has an entire metabolic machinery that is heat-resistant, including DNA polymerase that does not denature at these high temperatures.

11.20 The Polymerase Chain Reaction The steps in this cyclic process are repeated many times to produce multiple copies of a DNA fragment.





11.21 Sequencing DNA (*a*) The normal substrates for DNA replication are dNTPs. The slightly different structure of ddNTPs causes DNA synthesis to stop. (*b*) When labeled ddNTPs are incorporated into a mixture containing a single-stranded DNA template of unknown sequence, the result is a collection of fragments of varying lengths that can be separated by electrophoresis.

Scientists pondering the problem of copying DNA by PCR read Brock's basic research articles and got a clever idea: Why not use *T. aquaticus* DNA polymerase in the PCR reaction? It could withstand the 90°C temperature and would not have to be added during each cycle. The idea worked, and it earned biochemist Kerry Mullis a Nobel prize. PCR has had an enormous impact on genetic research. Some of its most striking applications will be described in Chapters 13 through 17.

The nucleotide sequence of DNA can be determined

Another important technique allows researchers to determine the base sequence of a DNA molecule. This **DNA sequencing** technique relies on the use of artificially altered nucleosides. As we saw earlier in this chapter, the deoxyribonucleoside triphosphates (dNTPs) that are the normal substrates for DNA replication contain the sugar deoxyribose. If that sugar is replaced with 2,3-dideoxyribose, the resulting dideoxyribonucleoside triphosphate (ddNTP) will still be added by DNA polymerase to a growing DNA chain. However, because ddNTPs lack a hydroxyl group at the 3′ position, the next nucleotide cannot be added (Figure 11.21*a*). Thus, synthesis stops at the position where ddNTP has been incorporated into the growing end of a DNA strand.

To determine the sequence of DNA, a fragment of DNA (usually no more than 700 base pairs long) is denatured. The resulting single strands of DNA are placed in a test tube and mixed with

- DNA polymerase, to synthesize the complementary strand;
- Short, artificially synthesized primers appropriate for that sequence;
- The four dNTPs (dATP, dGTP, dCTP, and dTTP); and
- Small amounts of the four ddNTPs, each bonded to a fluorescent "tag" that emits a different color of light.

DNA replication proceeds and the test tube soon contains a mixture of the template DNA strands and shorter, new complementary strands. The new strands, each ending with a fluorescent ddNTP, are of varying lengths. For example, each time a T is reached on the template strand, DNA polymerase adds either a dATP or ddATP to the growing complementary strand. If dATP is added, the strand continues to grow. If ddATP is added, chain growth stops.

After DNA replication has been allowed to proceed for a while the new DNA fragments are denatured from their templates. The fragments are then subjected to *electrophoresis* (see Figure 16.2). This technique sorts the DNA fragments by length, and can detect differences in fragment length as short as one base. During the electrophoresis run, the fragments pass in order of increasing length through a laser beam that excites the fluorescent tags. The light emitted is then detected, and the resulting information—that is, which color of fluorescence, and therefore which ddNTP, is at the end of a strand of which length—is fed into a computer. The computer processes this information and prints out the DNA sequence of the fragment (Figure 11.21*b*). DNA sequencing has formed the basis of the new science of genomics, as we will describe in Chapters 13, 14, and 17.

Chapter Summary

DNA: The Genetic Material

- Circumstantial evidence (its location and quantity in the cell) suggested that DNA might be the genetic material. Two experiments provided a convincing demonstration that this was the case. **Review Figures 11.1, 11.2, 11.3**

The Structure of DNA

- X-ray crystallography showed that the DNA molecule is a helix. **Review Figure 11.4**
- DNA is composed of nucleotides, each containing one of four bases: adenine, cytosine, thymine, or guanine. Biochemical analysis revealed that the amount of adenine equals the amount of thymine and the amount of guanine equals the amount of cytosine. **Review Figure 11.5**
- Putting the accumulated data together, Watson and Crick proposed that DNA is a double-stranded helix in which the strands are antiparallel and the bases are held together by hydrogen bonding. This model accounts for the genetic information, mutation, and replication functions of DNA. **Review Figures 11.6, 11.7**

Determining the DNA Replication Mechanism

- An experiment by Meselson and Stahl proved the replication of DNA to be semiconservative. Each parent strand acts as a template for the synthesis of a new strand; thus the two replicated DNA helices each contain one parent strand and one newly synthesized strand. **Review Figures 11.8, 11.9.** **See Web/CD Tutorial 11.1**

The Mechanisms of DNA Replication

- In DNA replication, the enzyme DNA polymerase catalyzes the addition of nucleotides to the 3′ end of each strand. Nucleotides are added by complementary base pairing with the template strand of DNA. The substrates are deoxyribonucleoside triphosphates, which are hydrolyzed as they are added to the growing chain, releasing energy that fuels the synthesis of DNA. **Review Figure 11.10**
- The DNA replication complex is attached to nuclear structures, and DNA is threaded through it for replication. **Review Figure 11.11**
- Many proteins assist in DNA replication. DNA helicase unwinds the double helix, and the template strands are stabilized by single-strand binding proteins.
- Prokaryotes have a single origin of replication; eukaryotes have many. Replication in both cases proceeds in both directions from an origin of replication. **Review Figure 11.12**
- An RNA primase catalyzes the synthesis a short RNA primer, to which nucleotides are added. **Review Figure 11.14**
- Through the action of DNA polymerase, the leading strand grows continuously in the 5′-to-3′ direction until the replication of that section of DNA has been completed. Then the RNA primer is degraded and DNA is added in its place.
- On the lagging strand, DNA is still made in the 5′-to-3′ direction. But synthesis of the lagging strand is discontinuous: The DNA is added as short fragments to primers, then the polymerase skips past the 5′ end to make the next fragment. **Review Figures 11.15, 11.16, 11.17.** **See Web/CD Tutorial 11.3**
- The very ends of linear chromosomes are usually not fully replicated because there is no place for a primer to bind on the lagging strand. This leads to a shortening of the DNA after each

round of replication, and ultimately cell death. Some cells have an enzyme, telomerase, that maintains chromosome length so that the cell can continue to divide. **Review Figure 11.18**
See Web/CD Tutorial 11.2

DNA Proofreading and Repair

▶ The machinery of DNA replication makes about one error in 10^6 nucleotides bases added. DNA is also subject to chemical damage. DNA is repaired by three different mechanisms: proofreading, mismatch repair, and excision repair. **Review Figure 11.19**

Practical Applications of DNA Replication

▶ The polymerase chain reaction technique uses DNA polymerase to repeatedly replicate DNA in the laboratory. **Review Figure 11.20**

▶ The principles of DNA replication can be used to determine the nucleotide sequence of DNA. **Review Figure 11.21**

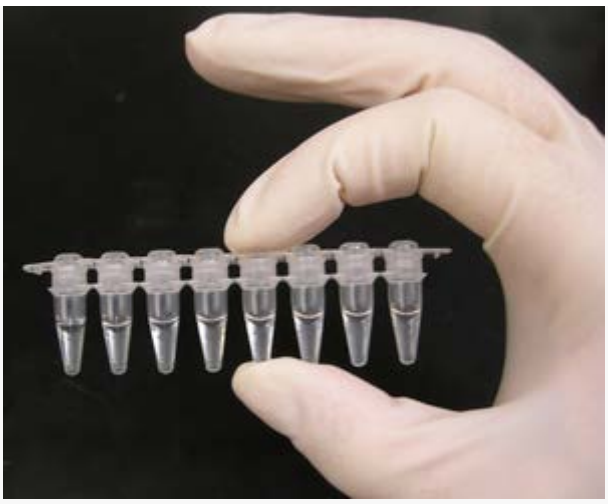
Self-Quiz

1. Griffith's studies of *Streptococcus pneumoniae*
   a. showed that DNA is the genetic material of bacteria.
   b. showed that DNA is the genetic material of bacteriophages.
   c. demonstrated the phenomenon of bacterial transformation.
   d. proved that prokaryotes reproduce sexually.
   e. proved that protein is not the genetic material.
2. In the Hershey–Chase experiment,
   a. DNA from parent bacteriophages appeared in progeny bacteriophages.
   b. most of the phage DNA never entered the bacteria.
   c. more than three-fourths of the phage protein appeared in progeny phages.
   d. DNA was labeled with radioactive sulfur.
   e. DNA formed the coat of the bacteriophages.
3. Which statement about complementary base pairing is *not* true?
   a. It plays a role in DNA replication.
   b. In DNA, T pairs with A.
   c. Purines pair with purines, and pyrimidines pair with pyrimidines.
   d. In DNA, C pairs with G.
   e. The base pairs are of equal length.
4. In semiconservative replication of DNA,
   a. the original double helix remains intact and a new double helix forms.
   b. the strands of the double helix separate and act as templates for new strands.
   c. polymerization is catalyzed by RNA polymerase.
   d. polymerization is catalyzed by a double helical enzyme.
   e. DNA is synthesized from amino acids.
5. Which of the following does not occur during DNA replication?
   a. Unwinding of the parent double helix
   b. Formation of short pieces that are connected by DNA ligase
   c. Complementary base pairing
   d. Use of a primer
   e. Polymerization in the 3′-to-5′ direction
6. The primer used for DNA replication
   a. is a short strand of RNA added to the 3′ end.
   b. is present only once on the leading strand.
   c. remains on the DNA after replication.
   d. ensures that there will be a free 5′ end to which nucleotides can be added.
   e. is added to only one of the two template strands.
7. One strand of DNA has the sequence 5′–ATTCCG–3′. The complementary strand for this is
   a. 5′–TAAGGC–3′
   b. 5′–ATTCCG–3′
   c. 5′–ACCTTA–3′
   d. 5′–CGGAAT–3′
   e. 5′–GCCTTA–3′
8. The role of DNA ligase in DNA replication is to
   a. add more nucleotides to the growing strand one at a time.
   b. open up the two DNA strands to expose template strands.
   c. ligate base to sugar to phosphate in a nucleotide.
   d. bond Okazaki fragments to one another.
   e. remove incorrectly paired bases.
9. The polymerase chain reaction
   a. is a method for sequencing DNA.
   b. is used to transcribe specific genes.
   c. amplifies specific DNA sequences.
   d. does not require DNA replication primers.
   e. uses a DNA polymerase that denatures at 55°C.
10. The following events occur in excision repair of DNA. What is their proper order?
    (1) Base-paired DNA is made complementary to the template.
    (2) Damaged bases are recognized.
    (3) DNA ligase seals the new strand to existing DNA.
    (4) Part of a single strand is excised.
    a. 1234
    b. 2134
    c. 2413
    d. 3421
    e. 4231

For Discussion

1. Outline a series of experiments using radioactive isotopes to show that bacterial DNA and not protein enters the host cell and is responsible for bacterial transformation.
2. Suppose that Meselson and Stahl had continued their experiment on DNA replication for another ten bacterial generations. Would there still have been any ^{14}N–^{15}N hybrid DNA present? Would it still have appeared in the centrifuge tube? Explain.
3. If DNA replication were conservative rather than semiconservative, what results would Meselson and Stahl have observed? Diagram the results using the conventions of Figure 11.9.
4. Using the following information, calculate the number of origins of DNA replication on a human chromosome: DNA polymerase adds nucleotides at 3,000 base pairs per minute in one direction; replication is bidirectional; S phase lasts 300 minutes; there are 120 million base pairs per chromosome. With a typical chromosome 3 μm long, how many origins are there per μm?
5. The drug dideoxycytidine (used to treat certain viral infections) is a nucleotide made with 2′,3′-dideoxyribose. This sugar lacks —OH groups at both the 2′ and the 3′ positions. Explain why this drug stops the growth of a DNA chain when added to DNA.

Polymerase chain reaction



A strip of eight PCR tubes, each containing a 100 µl reaction mixture

The **polymerase chain reaction** (**PCR**) is a technology in molecular biology used to amplify a single copy or a few copies of a piece of DNA across several orders of magnitude, generating thousands to millions of copies of a particular DNA sequence.

Developed in 1983 by Kary Mullis, PCR is now a common and often indispensable technique used in medical and biological research labs for a variety of applications. These include DNA cloning for sequencing, DNA-based phylogeny, or functional analysis of genes; the diagnosis of hereditary diseases; the identification of genetic fingerprints (used in forensic sciences and paternity testing); and the detection and diagnosis of infectious diseases. In 1993, Mullis was awarded theNobel Prize in Chemistry along with Michael Smith for his work on PCR.

The method relies on thermal cycling, consisting of cycles of repeated heating and cooling of the reaction for DNA meltingand enzymatic replication of the DNA. Primers (short DNA fragments) containing sequences complementary to the target region along with a DNA polymerase, which the method is named after, are key components to enable selective and repeated amplification. As PCR progresses, the DNA generated is itself used as a template for replication, setting in motion a chain reaction in which the DNA template is exponentially amplified. PCR can be extensively modified to perform a wide array ofgenetic manipulations.

Almost all PCR applications employ a heat-stable DNA polymerase, such as Taq polymerase (an enzyme originally isolated from the bacterium *Thermus aquaticus*). This DNA polymerase enzymatically assembles a new DNA strand from DNA building-blocks, the nucleotides, by using single-stranded DNA as a template and DNA oligonucleotides (also called DNA primers), which are required for initiation of DNA synthesis. The vast majority of PCR methods use thermal cycling, i.e., alternately heating and cooling the PCR sample through a defined series of temperature steps.

In the first step, the two strands of the DNA double helix are physically separated at a high temperature in a process called DNA melting. In the second step, the temperature is lowered and the two DNA strands become templates for DNA polymerase to selectively amplify the target DNA. The selectivity of PCR results from the use of primers that arecomplementary to the DNA region targeted for amplification under specific thermal cycling conditions.

Placing a strip of eight PCR tubes, each containing a 100 µl reaction mixture, into the thermal cycler

PCR principles and procedure

Figure 1a: A thermal cycler for PCR

Figure 1b: An older model three-temperature thermal cycler for PCR

PCR amplifies a specific region of a DNA strand (the DNA target). Most PCR methods typically amplify DNA fragments of between 0.1 and 10kilo base pairs (kbp), although some techniques allow for amplification of fragments up to 40 kbp in size. The amount of amplified product is determined by the available substrates in the reaction, which become limiting as the reaction progresses. A basic PCR set up requires several components and reagents. These components include:

- *DNA template* that contains the DNA region (target) to amplify
- Two *primers* that are complementary to the 3' (three prime) ends of each of the sense and anti-sense strand of the DNA target

- *Taq polymerase* or another DNA polymerase with a temperature optimum at around 70 °C
- *Deoxynucleoside triphosphates* (dNTPs, sometimes called "deoxynucleotide triphosphates"; nucleotides containing triphosphate groups), the building-blocks from which the DNA polymerase synthesizes a new DNA strand
- *Buffer solution*, providing a suitable chemical environment for optimum activity and stability of the DNA polymerase
- *Bivalent cations*, magnesium or manganese ions; generally Mg^{2+} is used, but Mn^{2+} can be used for PCR-mediated DNA mutagenesis, as higher Mn^{2+} concentration increases the error rate during DNA synthesis
- *Monovalent cation* potassium ions

The PCR is commonly carried out in a reaction volume of 10–200 µl in small reaction tubes (0.2–0.5 ml volumes) in a thermal cycler. The thermal cycler heats and cools the reaction tubes to achieve the temperatures required at each step of the reaction (see below). Many modern thermal cyclers make use of the Peltier effect, which permits both heating and cooling of the block holding the PCR tubes simply by reversing the electric current. Thin-walled reaction tubes permit favorable thermal conductivity to allow for rapid thermal equilibration. Most thermal cyclers have heated lids to prevent condensation at the top of the reaction tube. Older thermocyclers lacking a heated lid require a layer of oil on top of the reaction mixture or a ball of wax inside the tube.

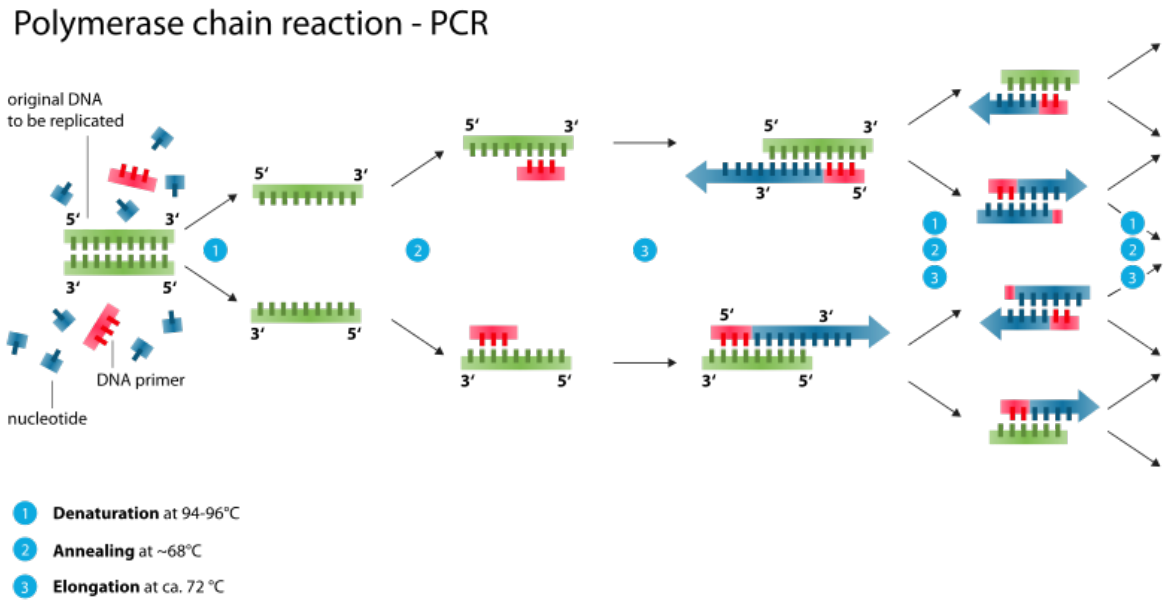
Procedure

Typically, PCR consists of a series of 20-40 repeated temperature changes, called cycles, with each cycle commonly consisting of 2-3 discrete temperature steps, usually three (Figure below). The cycling is often preceded by a single temperature step at a high temperature (>90 °C), and followed by one hold at the end for final product extension or brief storage. The temperatures used and the length of time they are applied in each cycle depend on a variety of parameters. These include the enzyme used for DNA synthesis, the concentration of divalent ions and dNTPs in the reaction, and the melting temperature (Tm) of the primers.[10]

- *Initialization step*(Only required for DNA polymerases that require heat activation by hot-start PCR.[11]): This step consists of heating the reaction to a temperature of 94–96 °C (or 98 °C if extremely thermostable polymerases are used), which is held for 1–9 minutes.
- *Denaturation step*: This step is the first regular cycling event and consists of heating the reaction to 94–98 °C for 20–30 seconds. It causes DNA melting of the DNA template by disrupting the hydrogen bonds between complementary bases, yielding single-stranded DNA molecules.
- *Annealing step*: The reaction temperature is lowered to 50–65 °C for 20–40 seconds allowing annealing of the primers to the single-stranded DNA template. This temperature must be low enough to allow for hybridization of the primer to the strand, but high enough for the hybridization to be specific, i.e., the primer should only bind to a perfectly complementary part of the template. If the temperature is too low, the primer could bind imperfectly. If it is too high, the primer might not bind. Typically the annealing temperature is about 3–5 °C below the Tm of the primers used. Stable DNA–DNA hydrogen bonds are only formed when the

primer sequence very closely matches the template sequence. The polymerase binds to the primer-template hybrid and begins DNA formation.

- *Extension/elongation step*: The temperature at this step depends on the DNA polymerase used; Taq polymerase has its optimum activity temperature at 75–80 °C, and commonly a temperature of 72 °C is used with this enzyme. At this step the DNA polymerase synthesizes a new DNA strand complementary to the DNA template strand by adding dNTPs that are complementary to the template in 5' to 3' direction, condensing the 5'-phosphate group of the dNTPs with the 3'-hydroxyl group at the end of the nascent (extending) DNA strand. The extension time depends both on the DNA polymerase used and on the length of the DNA fragment to amplify. As a rule-of-thumb, at its optimum temperature, the DNA polymerase polymerizes a thousand bases per minute. Under optimum conditions, i.e., if there are no limitations due to limiting substrates or reagents, at each extension step, the amount of DNA target is doubled, leading to exponential (geometric) amplification of the specific DNA fragment.
- *Final elongation*: This single step is occasionally performed at a temperature of 70–74 °C (this is the temperature needed for optimal activity for most polymerases used in PCR) for 5–15 minutes after the last PCR cycle to ensure that any remaining single-stranded DNA is fully extended.
- *Final hold*: This step at 4–15 °C for an indefinite time may be employed for short-term storage of the reaction.



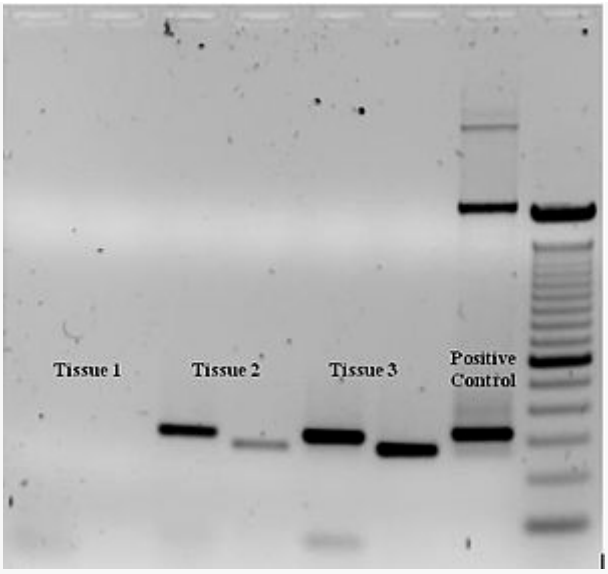


Figure 3: Ethidium bromide-stained PCR products after gel electrophoresis. Two sets of primers were used to amplify a target sequence from three different tissue samples. No amplification is present in sample #1; DNA bands in sample #2 and #3 indicate successful amplification of the target sequence. The gel also shows a positive control, and a DNA ladder containing DNA fragments of defined length for sizing the bands in the experimental PCRs.

To check whether the PCR generated the anticipated DNA fragment (also sometimes referred to as the amplimer oramplicon), agarose gel electrophoresis is employed for size separation of the PCR products. The size(s) of PCR products is determined by comparison with a DNA ladder (a molecular weight marker), which contains DNA fragments of known size, run on the gel alongside the PCR products (see Fig. 3).

PCR stages

The PCR process can be divided into three stages:

Exponential amplification: At every cycle, the amount of product is doubled (assuming 100% reaction efficiency). The reaction is very sensitive: only minute quantities of DNA must be present.[14]

Leveling off stage: The reaction slows as the DNA polymerase loses activity and as consumption of reagents such as dNTPs and primers causes them to become limiting.

Plateau: No more product accumulates due to exhaustion of reagents and enzyme.

PCR optimization

Main article: PCR optimization

In practice, PCR can fail for various reasons, in part due to its sensitivity to contamination causing amplification of spurious DNA products. Because of this, a number of techniques and procedures have been developed for optimizing PCR conditions.[15][16] Contamination with extraneous DNA is addressed with lab protocols and procedures that separate pre-PCR mixtures from potential DNA contaminants.[8] This usually involves spatial separation of PCR-setup areas from areas for analysis or purification of PCR products, use of disposable plasticware, and thoroughly cleaning the work surface between reaction setups. Primer-design techniques are important in improving PCR product yield and in avoiding the formation of spurious products, and the usage of alternate buffer components or polymerase enzymes can help with amplification of long or otherwise problematic regions of DNA. Addition of reagents, such as formamide, in buffer systems may increase the specificity and yield of PCR. Computer simulations of theoretical PCR results (Electronic PCR) may be performed to assist in primer design.

Application of PCR

Selective DNA isolation

PCR allows isolation of DNA fragments from genomic DNA by selective amplification of a specific region of DNA. This use of PCR augments many methods, such as generatinghybridization probes for Southern or northern hybridization and DNA cloning, which require larger amounts of DNA, representing a specific DNA region. PCR supplies these techniques with high amounts of pure DNA, enabling analysis of DNA samples even from very small amounts of starting material.

Other applications of PCR include DNA sequencing to determine unknown PCR-amplified sequences in which one of the amplification primers may be used in Sanger sequencing, isolation of a DNA sequence to expedite recombinant DNA technologies involving the insertion of a DNA sequence into a plasmid, phage, or cosmid (depending on size) or the genetic material of another organism. Bacterial colonies (E. coli) can be rapidly screened by PCR for correct DNA vector constructs.[19] PCR may also be used forgenetic fingerprinting; a forensic technique used to identify a person or organism by comparing experimental DNAs through different PCR-based methods.

Some PCR 'fingerprints' methods have high discriminative power and can be used to identify genetic relationships between individuals, such as parent-child or between siblings, and are used in paternity testing (Fig. 4). This technique may also be used to determine evolutionary relationships among organisms when certain molecular clocks are used (i.e., the 16S rRNA and recA genes of microorganisms).[*citation needed*]

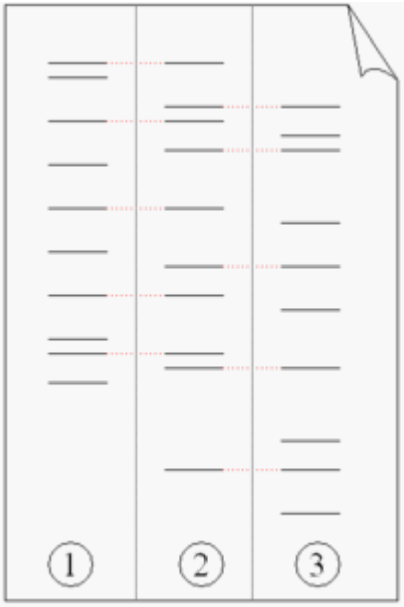


Figure 4: Electrophoresis of PCR-amplified DNA fragments. (1) Father. (2) Child. (3) Mother. The child has inherited some, but not all of the fingerprint of each of its parents, giving it a new, unique fingerprint.

Amplification and quantification of DNA

Because PCR amplifies the regions of DNA that it targets, PCR can be used to analyze extremely small amounts of sample. This is often critical for forensic analysis, when only a trace amount of DNA is available as evidence. PCR may also be used in the analysis of ADNA that is tens of thousands of years old. These PCR-based techniques have been successfully used on animals, such as a forty-thousand-year-oldmammoth, and also on human DNA, in applications ranging from the analysis of Egyptian mummies to the identification of a Russian tsar and the body of English king Richard III.

Quantitative PCR methods allow the estimation of the amount of a given sequence present in a sample—a technique often applied to quantitatively determine levels of gene

expression. Quantitative PCR is an established tool for DNA quantification that measures the accumulation of DNA product after each round of PCR amplification.

PCR in diagnosis of diseases

PCR permits early diagnosis of malignant diseases such as leukemia and lymphomas, which is currently the highest-developed in cancer research and is already being used routinely. PCR assays can be performed directly on genomic DNA samples to detect translocation-specific malignant cells at a sensitivity that is at least 10,000 fold higher than that of other methods.

PCR allows for rapid and highly specific diagnosis of infectious diseases, including those caused by bacteria or viruses. PCR also permits identification of non-cultivatable or slow-growing microorganisms such as mycobacteria, anaerobic bacteria, or viruses from tissue cultureassays and animal models. The basis for PCR diagnostic applications in microbiology is the detection of infectious agents and the discrimination of non-pathogenic from pathogenic strains by virtue of specific genes.

Viral DNA can likewise be detected by PCR. The primers used must be specific to the targeted sequences in the DNA of a virus, and the PCR can be used for diagnostic analyses or DNA sequencing of the viral genome. The high sensitivity of PCR permits virus detection soon after infection and even before the onset of disease. Such early detection may give physicians a significant lead time in treatment. The amount of virus ("viral load") in a patient can also be quantified by PCR-based DNA quantitation techniques (see below).

Limitations

DNA polymerase is prone to error, which in turn causes mutations in the PCR fragments that are made. Additionally, the specificity of the PCR fragments can mutate to the template DNA, due to nonspecific binding of primers. Furthermore prior information on the sequence is necessary in order to generate the primers. [23]

Variations on the basic PCR technique

Allele-specific PCR: a diagnostic or cloning technique based on single-nucleotide variations (SNVs not to be confused with SNPs) (single-base differences in a patient). It requires prior knowledge of a DNA sequence, including differences between alleles, and uses primers whose 3' ends encompass the SNV (base pair buffer around SNV usually incorporated). PCR amplification under stringent conditions is much less efficient in the presence of a mismatch between template and primer, so successful amplification with an SNP-specific primer signals presence of the specific SNP in a sequence. See SNP genotyping for more information.

- *Assembly PCR* or *Polymerase Cycling Assembly (PCA)*: artificial synthesis of long DNA sequences by performing PCR on a pool of long oligonucleotides with short overlapping segments. The oligonucleotides alternate between sense and antisense directions, and the overlapping segments determine the order of the PCR fragments, thereby selectively producing the final long DNA product.
- *Asymmetric PCR*: preferentially amplifies one DNA strand in a double-stranded DNA template. It is used in sequencing and hybridization probing where amplification of only one of the two complementary strands is required. PCR is carried out as usual, but with a great excess of the primer for the strand targeted for amplification. Because of the slow (arithmetic) amplification later in the reaction after the limiting primer has been used up, extra cycles of PCR are required. A recent modification on this process, known as *L*inear-*A*fter-*T*he-*E*xponential-PCR (LATE-PCR), uses a limiting primer with a higher melting temperature

(Tm) than the excess primer to maintain reaction efficiency as the limiting primer concentration decreases mid-reaction.

- *Dial-out PCR*: a highly parallel method for retrieving accurate DNA molecules for gene synthesis. A complex library of DNA molecules is modified with unique flanking tags before massively parallel sequencing. Tag-directed primers then enable the retrieval of molecules with desired sequences by PCR.[28]
- *Digital PCR (dPCR)*: used to measure the quantity of a target DNA sequence in a DNA sample. The DNA sample is highly diluted so that after running many PCRs in parallel, some of them do not receive a single molecule of the target DNA. The target DNA concentration is calculated using the proportion of negative outcomes. Hence the name 'digital PCR'.
- *Helicase-dependent amplification*: similar to traditional PCR, but uses a constant temperature rather than cycling through denaturation and annealing/extension cycles. DNA helicase, an enzyme that unwinds DNA, is used in place of thermal denaturation.[29]
- *Hot start PCR*: a technique that reduces non-specific amplification during the initial set up stages of the PCR. It may be performed manually by heating the reaction components to the denaturation temperature (e.g., 95 °C) before adding the polymerase.[30] Specialized enzyme systems have been developed that inhibit the polymerase's activity at ambient temperature, either by the binding of an antibody or by the presence of covalently bound inhibitors that dissociate only after a high-temperature activation step. Hot-start/cold-finish PCR is achieved with new hybrid polymerases that are inactive at ambient temperature and are instantly activated at elongation temperature.
- *In silico PCR* (digital PCR, virtual PCR, electronic PCR, e-PCR) refers to computational tools used to calculate theoretical polymerase chain reaction results using a given set of primers (probes) to amplify DNA sequences from a sequenced genome or transcriptome. In silico PCR was proposed as an educational tool for molecular biology.
- *Intersequence-specific PCR* (ISSR): a PCR method for DNA fingerprinting that amplifies regions between simple sequence repeats to produce a unique fingerprint of amplified fragment lengths.
- *Inverse PCR*: is commonly used to identify the flanking sequences around genomic inserts. It involves a series of DNA digestions and self ligation, resulting in known sequences at either end of the unknown sequence.
- *Ligation-mediated PCR*: uses small DNA linkers ligated to the DNA of interest and multiple primers annealing to the DNA linkers; it has been used for DNA sequencing,genome walking, and DNA footprinting.
- *Methylation-specific PCR* (MSP): developed by Stephen Baylin and Jim Herman at the Johns Hopkins School of Medicine, and is used to detect methylation of CpG islands in genomic DNA. DNA is first treated with sodium bisulfite, which converts unmethylated cytosine bases to uracil, which is recognized by PCR primers as thymine. Two PCRs are then carried out on

the modified DNA, using primer sets identical except at any CpG islands within the primer sequences. At these points, one primer set recognizes DNA with cytosines to amplify methylated DNA, and one set recognizes DNA with uracil or thymine to amplify unmethylated DNA. MSP using qPCR can also be performed to obtain quantitative rather than qualitative information about methylation.

- *Miniprimer PCR*: uses a thermostable polymerase (S-Tbr) that can extend from short primers ("smalligos") as short as 9 or 10 nucleotides. This method permits PCR targeting to smaller primer binding regions, and is used to amplify conserved DNA sequences, such as the 16S (or eukaryotic 18S) rRNA gene.[37]
- *Multiplex Ligation-dependent Probe Amplification* (*MLPA*): permits amplifying multiple targets with a single primer pair, thus avoiding the resolution limitations of multiplex PCR (see below).
- *Multiplex-PCR*: consists of multiple primer sets within a single PCR mixture to produce amplicons of varying sizes that are specific to different DNA sequences. By targeting multiple genes at once, additional information may be gained from a single test-run that otherwise would require several times the reagents and more time to perform. Annealing temperatures for each of the primer sets must be optimized to work correctly within a single reaction, and amplicon sizes. That is, their base pair length should be different enough to form distinct bands when visualized by gel electrophoresis.
- *Nanoparticle-Assisted PCR (nanoPCR)*: In recent years, it has been reported that some nanoparticles (NPs) can enhance the efficiency of PCR (thus being called nanoPCR), and some even perform better than the original PCR enhancers. It was also found that quantum dots (QDs) can improve PCR specificity and efficiency. Single-walled carbon nanotubes (SWCNTs) and multi-walled carbon nanotubes (MWCNTs) are efficient in enhancing the amplification of long PCR. Carbon nanopowder (CNP) was reported be able to improve the efficiency of repeated PCR and long PCR. ZnO, TiO_2, and Ag NPs were also found to increase PCR yield. Importantly, already known data has indicated that non-metallic NPs retained acceptable amplification fidelity. Given that many NPs are capable of enhancing PCR efficiency, it is clear that there is likely to be great potential for nanoPCR technology improvements and product development.
- *Nested PCR*: increases the specificity of DNA amplification, by reducing background due to non-specific amplification of DNA. Two sets of primers are used in two successive PCRs. In the first reaction, one pair of primers is used to generate DNA products, which besides the intended target, may still consist of non-specifically amplified DNA fragments. The product(s) are then used in a second PCR with a set of primers whose binding sites are completely or partially different from and located 3' of each of the primers used in the first reaction. Nested PCR is often more successful in specifically amplifying long DNA fragments than conventional PCR, but it requires more detailed knowledge of the target sequences.

- *Overlap-extension PCR* or *Splicing by overlap extension (SOEing)* : a genetic engineering technique that is used to splice together two or more DNA fragments that contain complementary sequences. It is used to join DNA pieces containing genes, regulatory sequences, or mutations; the technique enables creation of specific and long DNA constructs. It can also introduce deletions, insertions or point mutations into a DNA sequence.
- *PAN-AC*: uses isothermal conditions for amplification, and may be used in living cells.
- *quantitative PCR* (qPCR): used to measure the quantity of a target sequence (commonly in real-time). It quantitatively measures starting amounts of DNA, cDNA, or RNA.quantitative PCR is commonly used to determine whether a DNA sequence is present in a sample and the number of its copies in the sample. *Quantitative PCR* has a very high degree of precision. Quantitative PCR methods use fluorescent dyes, such as Sybr Green, EvaGreen or fluorophore-containing DNA probes, such as TaqMan, to measure the amount of amplified product in real time. It is also sometimes abbreviated to RT-PCR (*real-time* PCR) but this abbreviation should be used only for reverse transcription PCR. qPCR is the appropriate contractions for quantitative PCR (real-time PCR).
- *Reverse Transcription PCR (RT-PCR)*: for amplifying DNA from RNA. Reverse transcriptase reverse transcribes RNA into cDNA, which is then amplified by PCR. RT-PCR is widely used in expression profiling, to determine the expression of a gene or to identify the sequence of an RNA transcript, including transcription start and termination sites. If the genomic DNA sequence of a gene is known, RT-PCR can be used to map the location of exons and introns in the gene. The 5' end of a gene (corresponding to the transcription start site) is typically identified by RACE-PCR (*Rapid Amplification of cDNA Ends*).
- *Solid Phase PCR*: encompasses multiple meanings, including Polony Amplification (where PCR colonies are derived in a gel matrix, for example), Bridge PCR[44] (primers are covalently linked to a solid-support surface), conventional Solid Phase PCR (where Asymmetric PCR is applied in the presence of solid support bearing primer with sequence matching one of the aqueous primers) and Enhanced Solid Phase PCR (where conventional Solid Phase PCR can be improved by employing high Tm and nested solid support primer with optional application of a thermal 'step' to favour solid support priming).

- *Suicide PCR*: typically used in paleogenetics or other studies where avoiding false positives and ensuring the specificity of the amplified fragment is the highest priority. It was originally described in a study to verify the presence of the microbe Yersinia pestis in dental samples obtained from 14th Century graves of people supposedly killed by plague during the medieval Black Death epidemic. The method prescribes the use of any primer combination only once in a PCR (hence the term "suicide"), which should never have been used in any positive control PCR reaction, and the primers should always target a genomic region never

amplified before in the lab using this or any other set of primers. This ensures that no contaminating DNA from previous PCR reactions is present in the lab, which could otherwise generate false positives.

- *Thermal asymmetric interlaced PCR (TAIL-PCR)*: for isolation of an unknown sequence flanking a known sequence. Within the known sequence, TAIL-PCR uses a nested pair of primers with differing annealing temperatures; a degenerate primer is used to amplify in the other direction from the unknown sequence.
- *Touchdown PCR* (*Step-down PCR*): a variant of PCR that aims to reduce nonspecific background by gradually lowering the annealing temperature as PCR cycling progresses. The annealing temperature at the initial cycles is usually a few degrees (3-5 °C) above the T_m of the primers used, while at the later cycles, it is a few degrees (3-5 °C) below the primer T_m. The higher temperatures give greater specificity for primer binding, and the lower temperatures permit more efficient amplification from the specific products formed during the initial cycles.
- *Universal Fast Walking*: for genome walking and genetic fingerprinting using a more specific 'two-sided' PCR than conventional 'one-sided' approaches (using only one gene-specific primer and one general primer—which can lead to artefactual 'noise') by virtue of a mechanism involving lariat structure formation. Streamlined derivatives of UFW are LaNe RAGE (lariat-dependent nested PCR for rapid amplification of genomic DNA ends), 5'RACE LaNe and 3'RACE LaNe.

9 From DNA to Protein: Genotype to Phenotype

In 1978, Georgi Markov, a journalist who was living in London because he had written articles critical of the then-Communist government of Bulgaria, was standing at a bus stop near Waterloo Station. A man, possibly a Bulgarian secret agent, brushed up against him and, seemingly by accident, poked him with an umbrella. Markov felt a sharp pain, and within a few hours, he started to feel weak. High temperature, vomiting, and more severe symptoms soon followed. Two days later he was dead.

Police investigators found a tiny perforated pellet embedded in Markov's leg, and in that pellet was a small amount of ricin, a highly toxic molecule isolated from the seeds of the tropical castor bean plant, *Ricinus communis*. These seeds have been used for centuries as a source of castor oil, a natural product that used to be commonly given to children to "clean out" the digestive tract and is now used in the plastics industry. The toxin is a protein that is not present in the oil, and people found out the hard way that it is one of the most poisonous substances made by any organism.

Ricin kills cells by blocking protein synthesis. More specifically, it catalyzes the modification and cleavage of one of the large RNA molecules that make up the eukaryotic ribosome, the "workbench" of protein synthesis. Proteins are the major phenotypic expression of the genotype—the genetic information encoded in a cell's DNA. Ricin inhibits the cell's ability to express the genotype as phenotype through protein synthesis, and therefore ricin-poisoned cells cannot survive.

This chapter deals with the mechanisms by which genes are expressed as proteins. We will begin with evidence for the relationship between genes and proteins, and then fill in some of the details of the processes of transcription—the copying of the gene sequence of DNA into a sequence of RNA—and translation—the use of the sequence of RNA to make a polypeptide with a defined order of amino acids. Finally, we will define mutations and their phenotypes in specific molecular terms.

***Ricinus communis*, the Castor Bean Plant**
This brightly colored plant, grown in the Tropics as an ornamental, produces ricin, a lethal toxin that inhibits protein synthesis at the ribosome.

One Gene, One Polypeptide

There are many steps between genotype and phenotype. Genes cannot, all by themselves, directly produce a phenotypic result, such as a particular eye color, a specific seed shape, or a cleft chin, any more than a compact disk can play a symphony without the help of a CD player.

The first historical step in relating genes to phenotypes was to define phenotypes in molecular terms. The molecular basis of phenotypes was actually discovered before the discovery that DNA was the genetic material. Scientists had studied the chemical differences between individuals carrying wild-type and mutant alleles in organisms as diverse as humans and bread molds. They found that the major phenotypic differences were the result of differences in specific proteins.

In the 1940s, a series of experiments by George W. Beadle and Edward L. Tatum at Stanford University showed that when an altered gene resulted in an altered phenotype, that altered phenotype was always associated with an altered enzyme. This finding was critically important in defining the phenotype in chemical terms.

The roles of enzymes in biochemistry were being described at this time, and it occurred to Beadle and Tatum that the expression of a gene as phenotype could occur through an enzyme. They experimented with the bread mold *Neurospora crassa*. The nuclei in the body of this mold are haploid (*n*), as are its reproductive spores. (This fact is important because it means that even recessive mutant alleles are easy to detect in experiments.) Beadle and Tatum grew *Neurospora* on a minimal nutritional medium containing sucrose, minerals, and a vitamin. Using this medium, the enzymes of wild-type *Neurospora* could catalyze the metabolic reactions needed to make all the chemical constituents of their cells, including proteins. These wild-type strains are called *prototrophs* ("original eaters").

Beadle and Tatum treated wild-type *Neurospora* with X rays, which act as a *mutagen* (something known to cause mutations). When they examined the treated molds, they found some mutant strains could no longer grow on the minimal medium, but needed to be supplied with additional nutrients. The scientists hypothesized that these *auxotrophs* ("increased eaters") must have suffered mutations in genes that code for the enzymes used to synthesize the nutrients they now needed to obtain from their environment. For each auxotrophic strain, Beadle and Tatum were able to find a single compound that, when added to the minimal medium, supported the growth of that strain. This result suggested that mutations have simple effects, and that each mutation causes a defect in only one enzyme in a metabolic pathway described as *the one-gene, one-enzyme* hypothesis (Figure 12.1).

One group of auxotrophs, for example, could grow only if the minimal medium was supplemented with the amino acid arginine. (Wild-type *Neurospora* makes its own arginine.) These mutant strains were designated *arg* mutants. Beadle and Tatum found several different *arg* mutant strains. They proposed two alternative hypotheses to explain why these different genetic strains had the same phenotype:

- The different *arg* mutants could have mutations in *the same gene*, as in the case of the different eye color alleles of fruit flies. In this case, the gene might code for an enzyme involved in arginine synthesis.
- The different *arg* mutants could have mutations in *different genes*, each coding for a separate function that leads to arginine production. These independent functions might be different enzymes along the same biochemical pathway.

Some of the *arg* mutant strains fell into each of the two categories. Genetic crosses showed that some of the mutations were at the same chromosomal locus, and so were different alleles of the same gene. Other mutations were at different loci, or on different chromosomes, and so were not alleles of the same gene. Beadle and Tatum concluded that these different genes participated in governing a single biosynthetic pathway—in this case, the pathway leading to arginine synthesis (see the Conclusion in Figure 12.1).

By growing different *arg* mutants in the presence of various compounds suspected to be intermediates in the synthetic metabolic pathway for arginine, Beadle and Tatum were able to classify each mutation as affecting one enzyme or another, and to order the compounds along the pathway. Then they broke open the wild-type and mutant cells and examined them for enzyme activities. The results confirmed their hypothesis: Each mutant strain was indeed missing a single active enzyme in the pathway.

The gene–enzyme connection had been proposed 40 years earlier in 1908 by the Scottish physician Archibald Garrod, who studied the inherited human disease alkaptonuria. He linked the biochemical phenotype of the disease to an abnormal gene and a missing enzyme. Today we know of hundreds of examples of such hereditary diseases, which we will return to in Chapter 17.

The gene–enzyme relationship has undergone several modifications in light of our current knowledge of molecular biology. Many enzymes are composed of more than one polypeptide chain, or subunit (that is, they have a quaternary structure). In this case, each polypeptide chain is specified by its own separate gene. Thus, it is more correct to speak of a *one-gene, one-polypeptide* relationship: The function of a gene is to control the production of a single, specific polypeptide.

Much later, it was discovered that some genes code for forms of RNA that do not become translated into polypeptides, and that still other genes are involved in controlling which other DNA sequences are expressed. While these discoveries have supplanted the idea that all genes code for proteins, they did not invalidate the relationship between genes and polypeptides. But how does this relationship work—that is, how is the information encoded in DNA used to specify a particular polypeptide?

EXPERIMENT

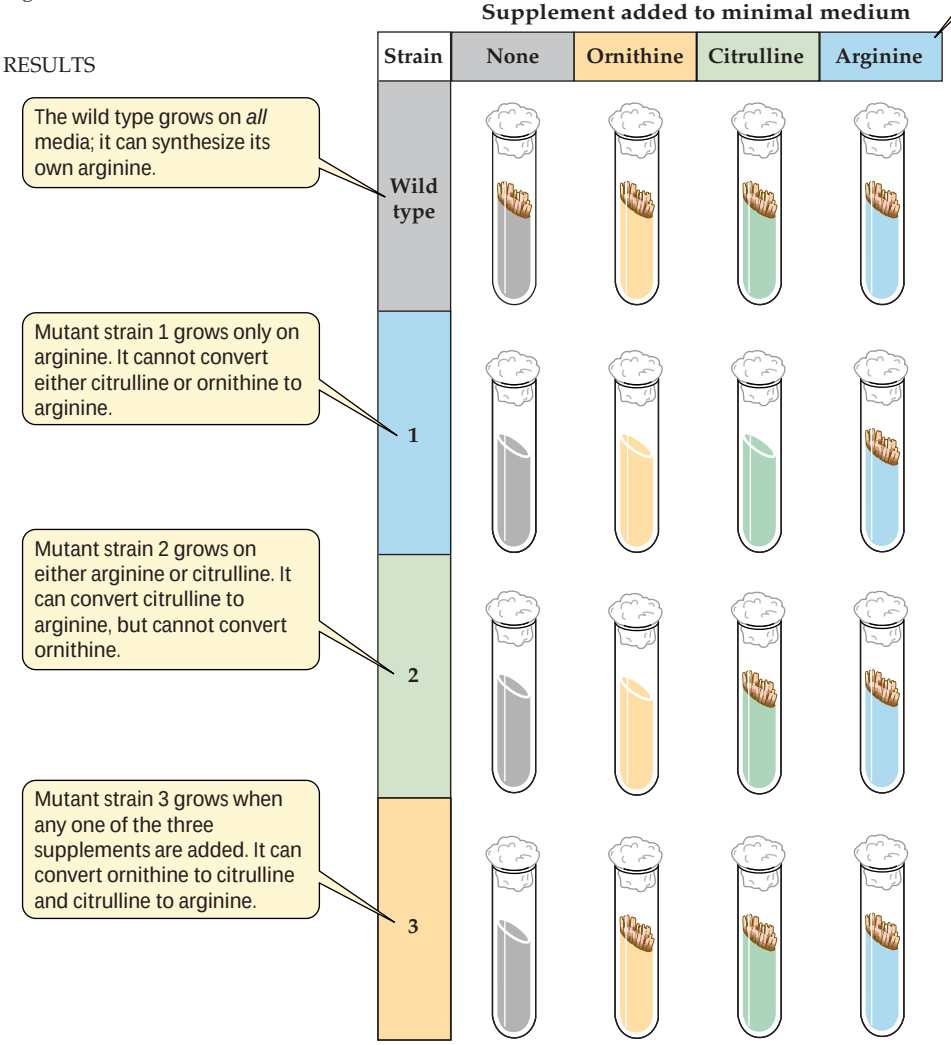
Question: What is the relationship between genes and enzymes in a biochemical pathway?

METHOD

Put spores of each mutant strain on a minimal medium (mm) with no supplements; mm + arginine; mm + citrulline; and mm + ornithine.

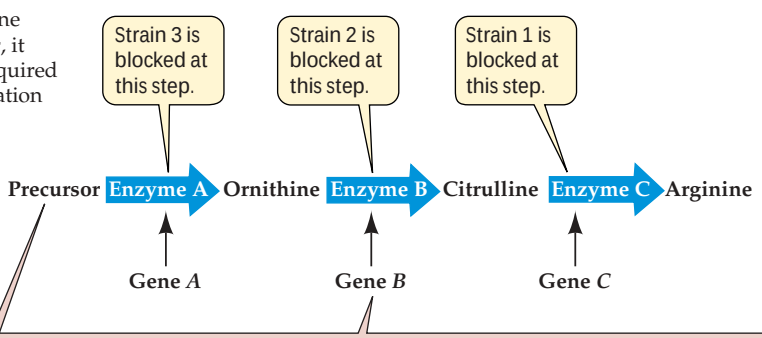
All the mutant strains grow if the amino acid arginine is added (the strains were selected because they require arginine).

RESULTS



INTERPRETATION

If an organism cannot convert one particular compound to another, it presumably lacks an enzyme required for the conversion, and the mutation is in the gene that codes for that enzyme.



Conclusion: The synthesis of arginine proceeds like this. Each gene specifies a particular enzyme.

12.1 One Gene, One Enzyme Beadle and Tatum studied several *arg* mutants of *Neurospora*. The different *arg* mutant strains required the addition of different compounds in order to synthesize the arginine required for their growth. Step through the figure to follow the reasoning that upheld the "one-gene, one-enzyme" hypothesis.

DNA, RNA, and the Flow of Information

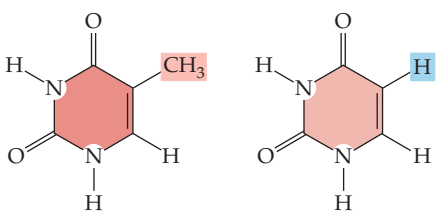
The expression of a gene to form a polypeptide occurs in two major steps:

- *Transcription* copies the information of a DNA sequence (the gene) into corresponding information in an RNA sequence.
- *Translation* converts this RNA sequence into the amino acid sequence of a polypeptide.

RNA differs from DNA

RNA is a key intermediary between DNA and polypeptide. **RNA** (**ribonucleic acid**) is a polynucleotide similar to DNA (see Figure 3.25), but it differs from DNA in three ways:

- RNA generally consists of only one polynucleotide strand.
- The sugar molecule found in RNA is ribose, rather than the deoxyribose found in DNA.
- Although three of the nitrogenous bases (adenine, guanine, and cytosine) in RNA are identical to those in DNA, the fourth base in RNA is **uracil** (**U**), which is similar to thymine but lacks the methyl (—CH_3) group.

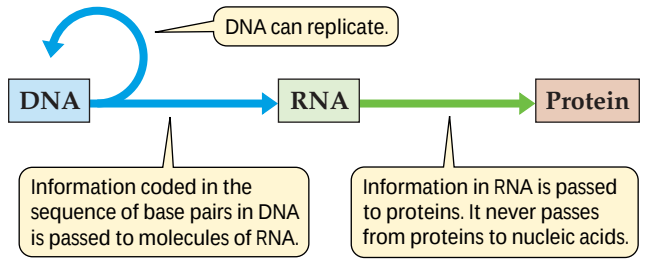


Thymine Uracil

RNA can base-pair with single-stranded DNA. This pairing obeys the same complementary base-pairing rules as in DNA, except that *adenine pairs with uracil* instead of thymine. Single-stranded RNA can fold into complex shapes by internal base pairing, as we will see later in this chapter.

Information flows in one direction when genes are expressed

Soon after he and Watson proposed their three-dimensional structure for DNA, Francis Crick pondered the problem of how DNA is functionally related to proteins. This led him to propose what he called the **central dogma** of molecular biology. The central dogma, simply stated, is that DNA codes for the production of RNA, RNA codes for the production of protein, and protein does not code for the production of protein, RNA, or DNA (Figure 12.2). In Crick's words, "once 'information' has passed into protein it cannot get out again."



12.2 The Central Dogma Information flows from DNA to RNA to proteins, as indicated by the arrows.

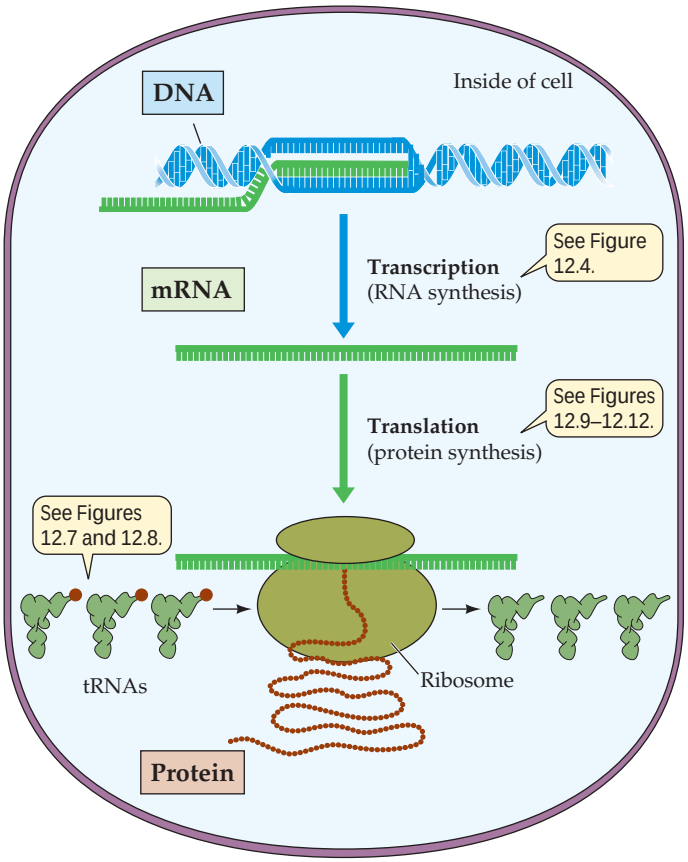
The central dogma raised two questions:

- How does genetic information get from the nucleus to the cytoplasm? (As we saw in Chapter 4, most of the DNA of a eukaryotic cell is confined to the nucleus, but proteins are synthesized in the cytoplasm.)
- What is the relationship between a specific nucleotide sequence in DNA and a specific amino acid sequence in a protein?

To answer these questions, Crick proposed two hypotheses.

THE MESSENGER HYPOTHESIS AND TRANSCRIPTION. To answer the first question, Crick and his colleagues proposed the *messenger hypothesis*. They proposed that an RNA molecule forms as a complementary copy of one DNA strand of a particular gene. The process by which this RNA forms is called **transcription** (Figure 12.3). This **messenger RNA**, or **mRNA**, then travels from the nucleus to the cytoplasm, where it serves as a template for the synthesis of proteins. Crick's hypothesis has been tested repeatedly for genes that code for proteins, and the answer is always the same: Each gene sequence in DNA that codes for a protein is expressed as a sequence in mRNA.

THE ADAPTER HYPOTHESIS AND TRANSLATION. To answer the second question, Crick proposed the *adapter hypothesis:* there must be an adapter molecule that can bind a specific amino acid with one region and recognize a sequence of nucleotides with another region. In due course, these adapters, called **transfer RNA**, or **tRNA**, were identified. Because they recognize the genetic message of mRNA and simultaneously carry specific amino acids, tRNAs can translate the language of DNA into the language of proteins. The tRNA adapters line up on the mRNA so that the amino acids are in the proper sequence for a growing polypeptide chain—a process called **translation** (see Figure 12.3). Once again, actual observations of the expression of thousands of genes have confirmed the hypothesis that tRNA acts as the intermediary between the nucleotide sequence information in mRNA and the amino acid sequence in a protein.



12.3 From Gene to Protein This diagram summarizes the processes of gene expression in prokaryotes. In eukaryotes, the processes are somewhat more complex.

Summarizing the main features of the central dogma, the messenger hypothesis, and the adapter hypothesis, we may say that a given gene is transcribed to produce a messenger RNA (mRNA) complementary to one of the DNA strands, and that transfer RNA (tRNA) molecules translate the sequence of bases in the mRNA into the appropriate sequence of linked amino acids during protein synthesis.

RNA viruses modify the central dogma

Certain viruses are rare exceptions to the central dogma. *Viruses* are infectious particles that reproduce inside cells. Many viruses, such as the tobacco mosaic virus, influenza virus, and poliovirus, have RNA rather than DNA as their genetic material. With its nucleotide sequence, RNA could potentially act as an information carrier and be expressed as proteins. But since RNA is usually single-stranded, its replication is a problem. The viruses generally solve this problem by transcribing from RNA to RNA, making an RNA strand that is complementary to their genome. This "opposite" strand is then used to make multiple copies of the viral genome by transcription:

The human immunodeficiency virus (HIV) and certain rare tumor viruses also have RNA as their genome, but do not replicate it as RNA-to-RNA. Instead, after infecting a host cell, they make a DNA copy of their genome and use it to make more RNA. This RNA is then used both as genomes for more copies of the virus and as mRNA to produce viral proteins.

Synthesis of DNA from RNA is called *reverse transcription*, and not surprisingly, such viruses are called **retroviruses**.

Transcription: DNA-Directed RNA Synthesis

Although the RNA viruses present a modification of the central dogma, the fact remains that in normal prokaryotic and eukaryotic cells, RNA synthesis is directed by DNA. Transcription—the formation of a specific RNA from a specific DNA—requires several components:

- A DNA template for complementary base pairing
- The appropriate ribonucleoside triphosphates (ATP, GTP, CTP, and UTP) to act as substrates
- An enzyme, **RNA polymerase**

Within each gene, only one of the two strands of DNA—the *template strand*—is transcribed. The other, complementary DNA strand, referred to as the *non-template strand*, remains untranscribed. For different genes in the same DNA molecule, different strands may be transcribed. That is, the strand that is the non-template strand in one gene may be the template strand in another.

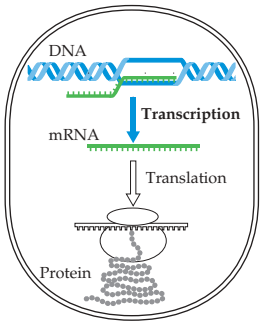
Not only mRNA is produced by transcription. The same process is responsible for the synthesis of tRNA and ribosomal RNA (rRNA), whose important roles in protein synthesis will be described below. Like polypeptides, these RNAs are encoded by specific genes.

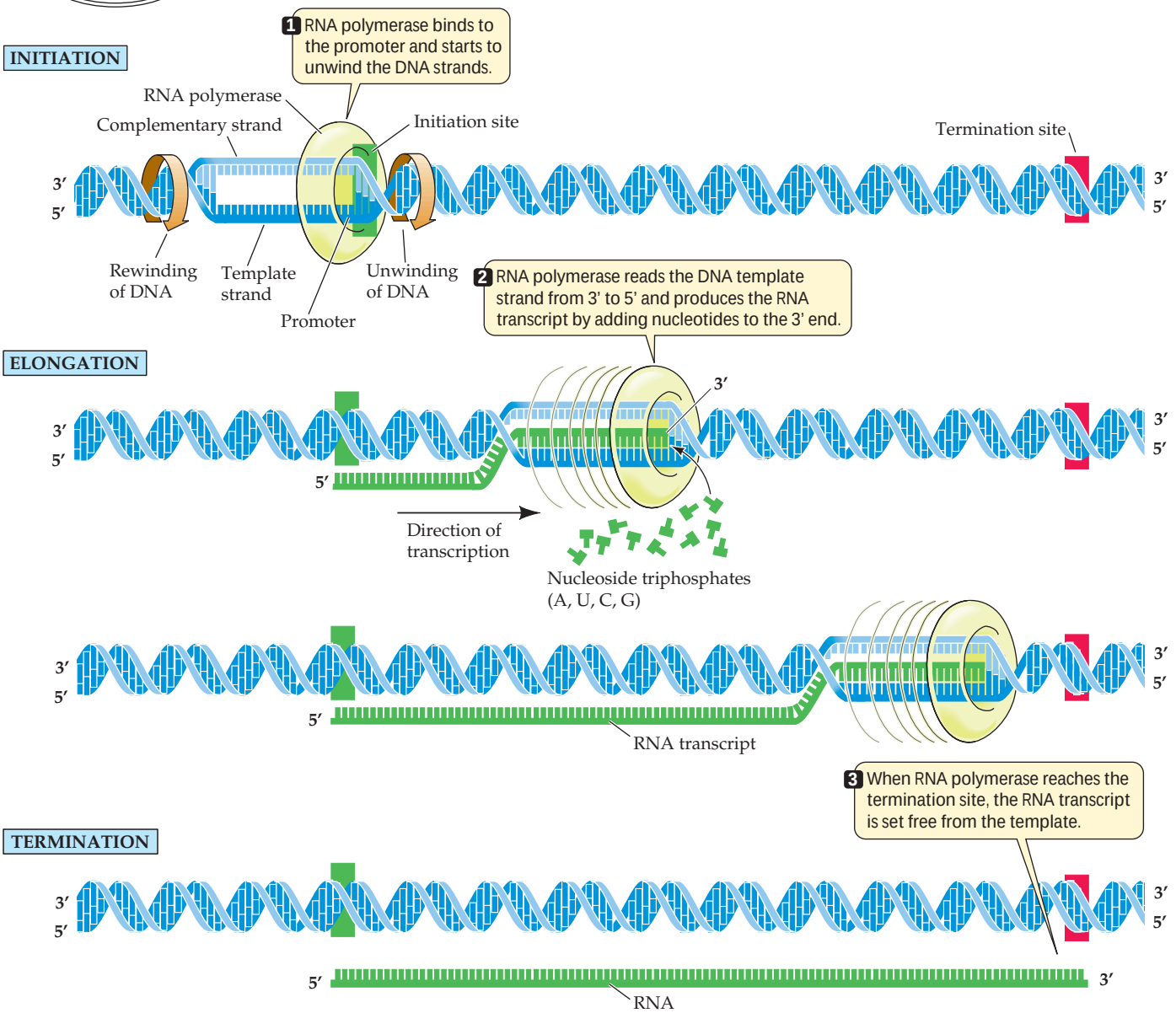
In DNA replication, as we know, the two strands of the parent molecule unwind, and each strand serves as the template for a new strand. In transcription, DNA partly unwinds so that it can serve as a template for RNA synthesis. As the RNA transcript is formed, it peels away, allowing the DNA to be rewound into the double helix (Figure 12.4).

Transcription can be divided into three distinct processes: initiation, elongation, and termination. Let's consider each of these in turn.

Initiation of transcription requires a promoter and RNA polymerase

Initiation begins transcription, and requires a **promoter**, a special sequence of DNA to which RNA polymerase binds very tightly. There is at least one promoter for each gene (or,





12.4 DNA Is Transcribed into RNA DNA is partially unwound to serve as a template for RNA synthesis. The RNA transcript is formed and then peels away, allowing the DNA that has already been transcribed to rewind into a double helix. Three distinct processes—initiation, elongation, and termination—constitute DNA transcription. RNA polymerase is much larger in reality than indicated here, covering about 50 base pairs.

in prokaryotes, each set of genes). Promoters are important control sequences that "tell" the RNA polymerase three things:

- where to start transcription
- which strand of DNA to read
- the direction to take from the start

A promoter, which is a specific sequence in the DNA that reads in a particular direction, orients the RNA polymerase and thus "aims" it at the correct strand to use as a template. Promoters function somewhat like the punctuation marks that determine how a sequence of words is to be read as a sentence. Part of each promoter is the *initiation site*, where transcription begins. Farther toward the 3′ end of the pro-

moter lie groups of nucleotides that help the RNA polymerase bind. RNA polymerase moves in a 3′-to-5′ direction along the template strand (see Figure 12.4).

Although every gene has a promoter, not all promoters are identical. Some promoters are more effective at transcription initiation than others. Furthermore, there are differences between transcription initiation in prokaryotes and in eukaryotes. These differences will be explored in Chapters 13 and 14.

RNA polymerase elongates the transcript

Once RNA polymerase has bound to the promoter, it begins the process of **elongation**. It unwinds the DNA about 20 base pairs at a time and reads the template strand in the 3′-to-5′ direction (see Figure 12.4). Like DNA polymerase, RNA polymerase adds new nucleotides to the 3′ end of the growing strand, but does not require a primer to get started. The new RNA elongates from the first base that forms its 5′ end to its 3′ end. The RNA transcript is thus antiparallel to the DNA template strand.

Unlike DNA polymerases, RNA polymerases do not inspect and correct their work. Transcription errors occur at a rate of one mistake for every 10^4 to 10^5 bases. Because many copies of RNA are made, and because they often have only a relatively short existence, these errors are not as potentially harmful as mutations in DNA.

Transcription terminates at particular base sequences

What tells RNA polymerase to stop adding nucleotides to a growing RNA transcript? Just as initiation sites specify the start of transcription, particular base sequences in the DNA specify its **termination**. The mechanisms of termination are complex and of more than one kind. For some genes, the newly formed transcript simply falls away from the DNA template and the RNA polymerase. For others, a helper protein pulls the transcript away.

In prokaryotes, in which there is no nuclear envelope and ribosomes can be near the chromosome, the translation of mRNA often begins near the 5′ end of the mRNA before transcription of the mRNA molecule is complete. In eukaryotes, the situation is more complicated. First, there is a spatial separation of transcription (in the nucleus) and translation (in the cytoplasm). Second, the first product of transcription is a pre-mRNA that is longer than the final mRNA and must undergo considerable processing before it can be translated. The advantages of this processing, and its mechanisms, will be discussed in Chapter 14.

The Genetic Code

How do transcription and translation produce specific and functional protein products? These processes require a *genetic code* that relates genes (DNA) to mRNA and mRNA to the amino acids of proteins. The genetic code specifies which amino acids will be used to build a protein. You can think of the genetic information in an mRNA molecule as a series of sequential, nonoverlapping three-letter "words." Each sequence of three nucleotide bases (the three "letters") along the chain specifies a particular amino acid. Each three-letter "word" is called a **codon**. Each codon is complementary to the corresponding triplet in the DNA molecule from which it was transcribed. Thus, the genetic code is the means of relating codons to their specific amino acids.

The complete genetic code is shown in Figure 12.5. Notice that there are many more codons than there are different amino acids in proteins. Combinations of the four available "letters" (the bases) give 64 (4^3) different three-letter codons, yet these codons determine only 20 amino acids. AUG, which codes for methionine, is also the **start codon**, the initiation signal for translation. Three of the codons (UAA, UAG, UGA) are **stop codons**, or termination signals for translation;

Second letter

First letter	U	C	A	G	Third letter
U	UUU, UUC: Phenylalanine UUA, UUG: Leucine	UCU, UCC, UCA, UCG: Serine	UAU, UAC: Tyrosine UAA: Stop codon UAG: Stop codon	UGU, UGC: Cysteine UGA: Stop codon UGG: Tryptophan	U C A G
C	CUU, CUC, CUA, CUG: Leucine	CCU, CCC, CCA, CCG: Proline	CAU, CAC: Histidine CAA, CAG: Glutamine	CGU, CGC, CGA, CGG: Arginine	U C A G
A	AUU, AUC, AUA: Isoleucine AUG: Methionine; start codon	ACU, ACC, ACA, ACG: Threonine	AAU, AAC: Asparagine AAA, AAG: Lysine	AGU, AGC: Serine AGA, AGG: Arginine	U C A G
G	GUU, GUC, GUA, GUG: Valine	GCU, GCC, GCA, GCG: Alanine	GAU, GAC: Aspartic acid GAA, GAG: Glutamic acid	GGU, GGC, GGA, GGG: Glycine	U C A G

12.5 The Universal Genetic Code Genetic information is encoded in mRNA in three-letter units—codons—made up of the bases uracil (U), cytosine (C), adenine (A), and guanine (G). To decode a codon, find its first letter in the left column, then read across the top to its second letter, then read down the right column to its third letter. The amino acid the codon specifies is given in the corresponding row. For example, AUG codes for methionine, and GUA codes for valine.

when the translation machinery reaches one of these codons, translation stops, and the polypeptide is released from the translation complex.

After describing the properties of the genetic code, we will examine some of the scientific thinking and experimentation that went into deciphering it.

The genetic code is redundant but not ambiguous

After the start and stop codons, the remaining 60 codons are far more than enough to code for the other 19 amino acids—and indeed there are repeats. Thus we say that the genetic code is *redundant;* that is, an amino acid may be represented by more than one codon. The redundancy is not evenly divided among the amino acids. For example, methionine and tryptophan are represented by only one codon each, whereas leucine is represented by six different codons (see Figure 12.5).

The term "redundancy" should not be confused with "ambiguity." To say that the code was *ambiguous* would mean that a single codon could specify either of two (or more) different amino acids; there would then be doubt whether to put in, say, leucine or something else. The genetic code is not ambiguous. Redundancy in the code means that there is more than one clear way to say, "Put leucine here." In other words, a given amino acid may be encoded by more than one codon, but a codon can code for only one amino acid. But just as people in different places prefer different ways of saying the same thing—"Good-bye!" "See you!" "Ciao!" and "So long!" have the same meaning—different organisms prefer one or another of the redundant codons.

The genetic code is (nearly) universal

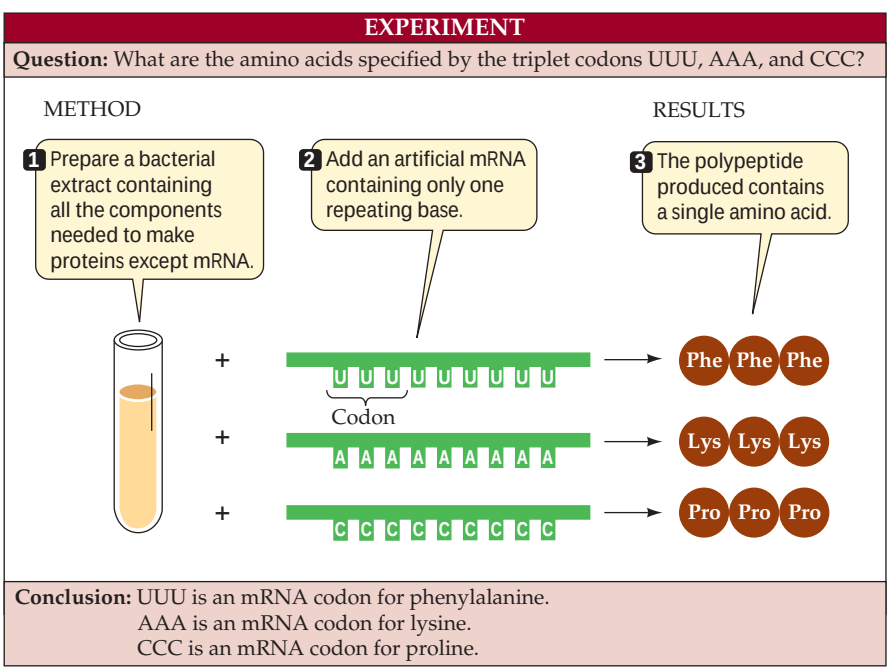
Over 40 years of experiments on thousands of organisms from all the living domains and kingdoms reveal that the genetic code appears to be nearly *universal*, applying to all the species on our planet. Thus the code must be an ancient one that has been maintained intact throughout the evolution of living organisms. Exceptions are known: within mitochondria and chloroplasts, the code differs slightly from that in prokaryotes and in the nuclei of eukaryotic cells; in one group of protists, UAA and UAG code for glutamine rather than functioning as stop codons. The significance of these differences is not yet clear. What is clear is that the exceptions are few and slight.

The common genetic code means that there is also a common language for evolution. As natural selection resulted in one species replacing another, the raw material of genetic variation has remained the same. The common code also has profound implications for genetic engineering, as we will see in Chapter 16, since it means that a human gene is in the same language as a bacterial gene. The differences are more like dialects of a single language than entirely different languages. So the transcription and translation machinery of a bacterium could theoretically utilize genes from a human as well as its own genes.

The codons in Figure 12.5 are *mRNA codons*. The base sequence on the DNA strand that was transcribed to produce the mRNA is complementary and antiparallel to these codons. Thus, for example, 3′-AAA-5′ in the template DNA strand corresponds to phenylalanine (which is encoded by the mRNA codon 5′-UUU-3′), and 3′-ACC-5′ in the template DNA corresponds to tryptophan (which is encoded by the mRNA codon 5′-UGG-3′). How did biologists assign these codons to specific amino acids?

Biologists deciphered the genetic code by using artificial messengers

Molecular biologists broke the genetic code in the early 1960s. The problem was perplexing: How could more than 20 "code words" be written with an "alphabet" consisting of only four "letters"? How, in other words, could four bases (A, U, G, and C) code for 20 different amino acids?



12.6 Deciphering the Genetic Code Nirenberg and Matthaei used a test-tube protein synthesis system to determine the amino acids specified by synthetic mRNAs of known codon composition.

That the code was a triplet code, based on three-letter codons, was considered likely. Since there are only four letters (A, G, C, U), a one-letter code clearly could not unambiguously encode 20 amino acids; it could encode only four of them. A two-letter code could contain only $4 \times 4 = 16$ codons—still not enough. But a triplet code could contain up to $4 \times 4 \times 4 = 64$ codons. This was more than enough to encode the 20 amino acids.

Marshall W. Nirenberg and J. H. Matthaei, at the National Institutes of Health, made the first decoding breakthrough in 1961 when they realized that they could use a simple artificial polynucleotide instead of a complex natural mRNA as a messenger. They could then identify the polypeptide that the artificial messenger encoded.

Scientists prepared an artificial mRNA in which all the bases were uracil (poly U). When poly U was added to a test tube containing all the ingredients necessary for protein synthesis (ribosomes, all the amino acids, activating enzymes, tRNAs, and other factors), a polypeptide formed. This polypeptide contained only one kind of amino acid: phenylalanine (Phe). Poly U coded for poly Phe! Accordingly, UUU appeared to be the mRNA code word—the codon—for phenylalanine. Following up on this success, Nirenberg and Matthaei soon showed that CCC codes for proline and AAA for lysine (Figure 12.6). (Poly G presented some chemical problems and was not tested initially.) UUU, CCC, and AAA were three of the easiest codons; different approaches were required to work out the rest.

Other scientists later found that simple artificial mRNAs only three nucleotides long—each amounting to a codon—could bind to a ribosome, and that the resulting complex could then cause the binding of the corresponding tRNA with its specific amino acid. Thus, for example, simple UUU caused the tRNA carrying phenylalanine to bind to the ribosome. After this discovery, complete deciphering of the genetic code was relatively simple. To find the "translation" of a codon, Nirenberg could use a sample of that codon as an artificial mRNA and see which amino acid became bound to it.

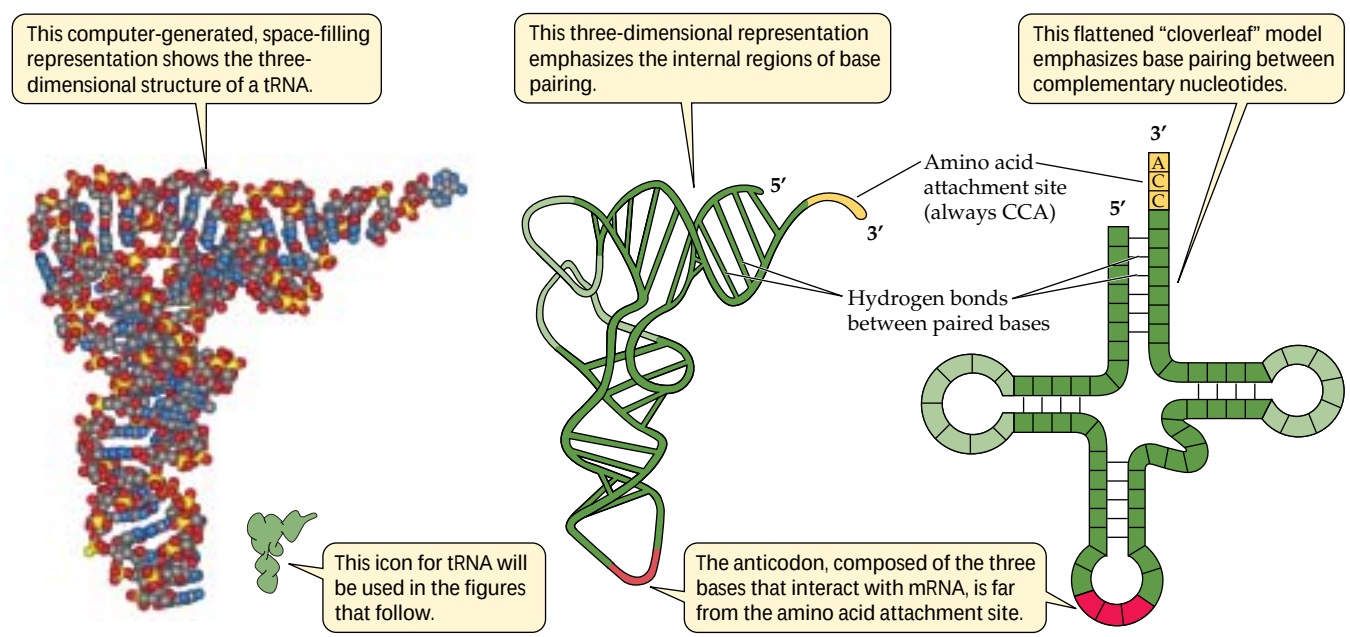
Preparation for Translation: Linking RNAs, Amino Acids, and Ribosomes

As Crick's adapter hypothesis proposed, the translation of mRNA into proteins requires a molecule that links the information contained in mRNA codons with specific amino acids in proteins. That function is performed by tRNA. Two key events must take place to ensure that the protein made is the one specified by mRNA:

- tRNA must read mRNA correctly.
- tRNA must carry the amino acid that is correct for its reading of the mRNA.

Transfer RNAs carry specific amino acids and bind to specific codons

The codon in mRNA and the amino acid in a protein are related by way of an adapter—a specific tRNA with an



12.7 Transfer RNA The tRNA molecule binds to amino acids, associates with mRNA molecules, and interacts with ribosomes. There is at least one specific tRNA molecule for each of the amino acids. When the tRNA is bonded to an amino acid, it is designated as a charged tRNA.

tached amino acid. For each of the 20 amino acids, there is at least one specific type ("species") of tRNA molecule.

The tRNA molecule has three functions: It carries (is "charged with") an amino acid, it associates with mRNA molecules, and it interacts with ribosomes. Its molecular structure relates clearly to all of these functions. A tRNA molecule has about 75 to 80 nucleotides. It has a *conformation* (a three-dimensional shape) that is maintained by complementary base pairing (hydrogen bonding) within its own sequence (Figure 12.7).

The conformation of a tRNA molecule allows it to combine specifically with binding sites on ribosomes. At the 3′ end of every tRNA molecule is a site to which its specific amino acid binds covalently. At about the midpoint of tRNA is a group of three bases, called the **anticodon**, that constitutes the site of complementary base pairing (hydrogen bonding) with mRNA. Each tRNA species has a unique anticodon, which is complementary to the mRNA codon for that tRNAs amino acid. At contact, the codon and the anticodon are antiparallel to each other. As an example of this process, consider the amino acid arginine:

- The DNA coding region for arginine is 3′-GCC-5′, which is transcribed, by complementary base pairing, to the mRNA codon 5′-CGG-3′.
- That mRNA codon binds by complementary base pairing to a tRNA with the anticodon 3′-GCC-5′, which is charged with arginine.

Recall that 61 different codons encode the 20 amino acids in proteins (see Figure 12.5). Does this mean that the cell must produce 61 different tRNA species, each with a different anticodon? No. The cell gets by with about two-thirds that number of tRNA species, because the specificity for the base at the 3′ end of the codon (and the 5′ end of the anticodon) is not always strictly observed. This phenomenon, called *wobble*, allows the alanine codons GCA, GCC, and GCU, for example, all to be recognized by the same tRNA. Wobble is allowed in some matches, but not in others; of most importance, it does not allow the genetic code to be ambiguous!

Activating enzymes link the right tRNAs and amino acids

The charging of each tRNA with its correct amino acid is achieved by a family of activating enzymes, known more formally as *aminoacyl-tRNA synthetases* (Figure 12.8). Each activating enzyme is specific for one amino acid and for its corresponding tRNA. The enzyme has a three-part active site that recognizes three smaller molecules: a specific amino acid, ATP, and a specific tRNA.

The activating enzyme reacts with tRNA and an amino acid (AA) in two steps:

$$\text{enzyme} + \text{ATP} + \text{AA} \rightarrow \text{enzyme—AMP—AA} + \text{PP}_i$$

$$\text{enzyme—AMP—AA} + \text{tRNA} \rightarrow \text{enzyme} + \text{AMP} + \text{tRNA—AA}$$

The amino acid is attached to the 3′ end of the tRNA (to a free OH group on the ribose) with an energy-rich bond, forming charged tRNA. This bond will provide the energy for the synthesis of the peptide bond that will join adjacent amino acids.

A clever experiment by Seymour Benzer and his colleagues at the California Institute of Technology demonstrated the importance of the specificity of the attachment of tRNA to its amino acid. In their laboratory, the amino acid cysteine, already properly attached to its tRNA, was chemically modified to become a different amino acid, alanine. Which component—the amino acid or the tRNA—would be recognized when this hybrid charged tRNA was put into a protein-synthesizing system? The answer was: the tRNA. Everywhere in the synthesized protein where cysteine was supposed to be, alanine appeared instead. The cysteine-specific tRNA had delivered its cargo (alanine) to every mRNA "address" where cysteine was called for. This experiment showed that the protein synthesis machinery recognizes the anticodon of the charged tRNA, not the amino acid attached to it.

If activating enzymes in nature did what Benzer did in the laboratory and charged tRNAs with the wrong amino acids, those amino acids would be inserted into proteins at inappropriate places, leading to alterations in protein shape and function. The fact that the activating enzymes are highly specific has led to the process of tRNA charging being called the "second genetic code."

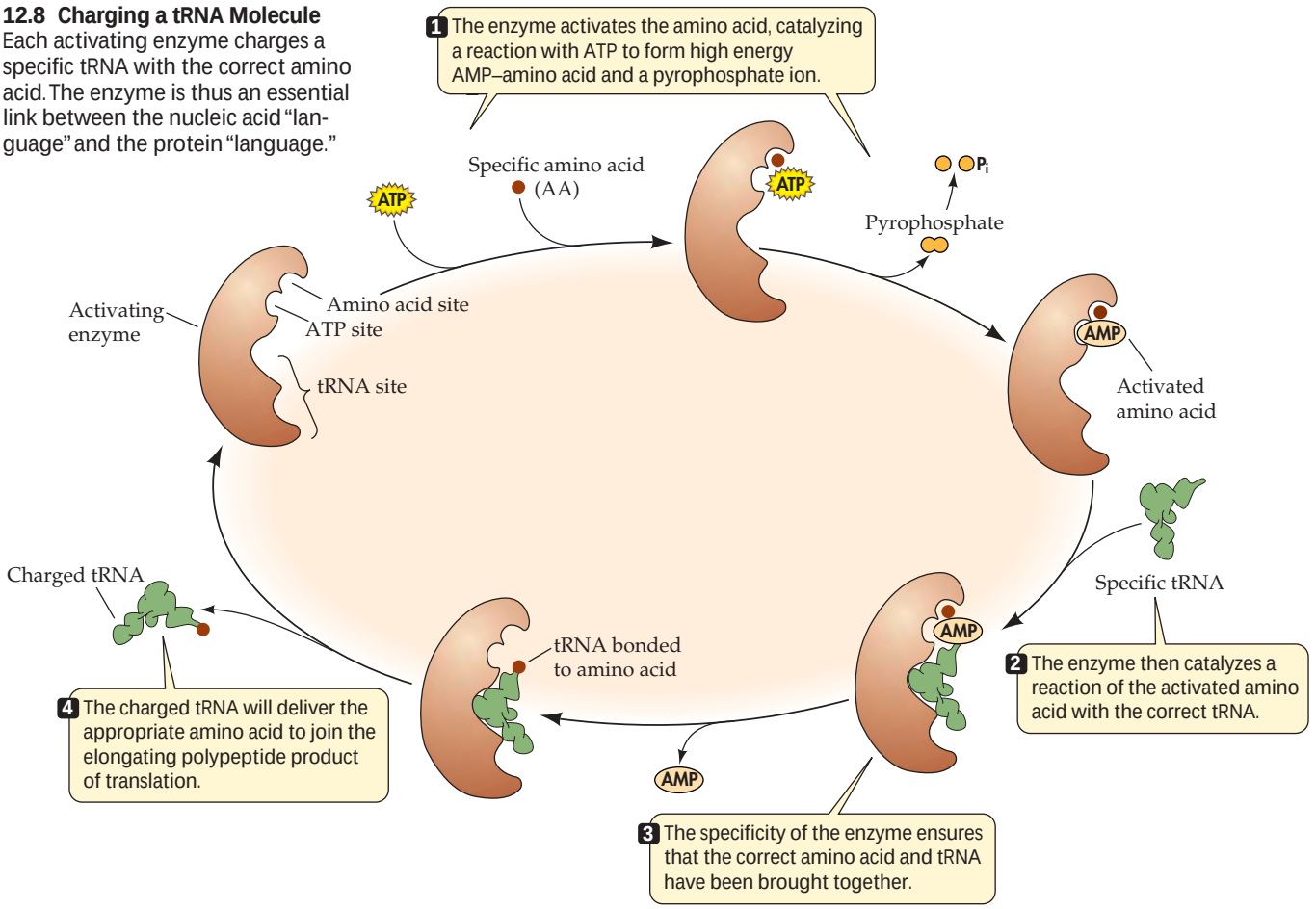
The ribosome is the workbench for translation

Ribosomes are required for the translation of the genetic information in mRNA into a polypeptide chain. Although ribosomes are small in contrast to other cellular organelles, their mass of several million daltons makes them large in comparison with charged tRNAs.

Each ribosome consists of two subunits, a large one and a small one (Figure 12.9). In eukaryotes, the large subunit consists of three different molecules of rRNA and about 45 different protein molecules, arranged in a precise pattern. The small subunit consists of one rRNA molecule and 33 different protein molecules. When not active in the translation of mRNA, the ribosomes exist as separated subunits.

The ribosomes of prokaryotes are somewhat smaller than those of eukaryotes, and their ribosomal proteins and RNAs are different. Mitochondria and chloroplasts also contain ribosomes, some of which are similar to those of prokaryotes.

The different proteins and rRNAs in a ribosomal subunit are held together by ionic and hydrophobic forces, not covalent bonds. If these forces are disrupted by detergents, for example, the proteins and rRNAs separate from one another.



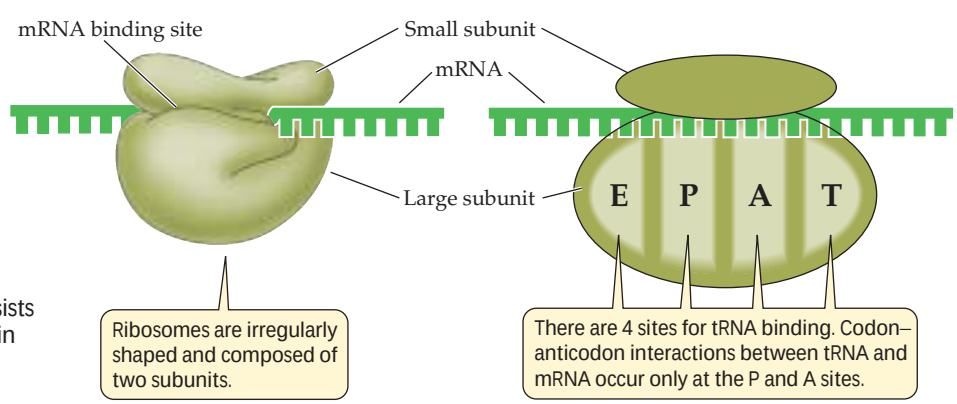
12.8 Charging a tRNA Molecule Each activating enzyme charges a specific tRNA with the correct amino acid. The enzyme is thus an essential link between the nucleic acid "language" and the protein "language."

When the detergent is removed, the entire complex structure self-assembles. This is like separating the pieces of a jigsaw puzzle and having them fit together again without human hands to guide them!

A given ribosome does not specifically produce just one kind of protein. A ribosome can use any mRNA and all species of charged tRNAs, and thus can be used to make many different polypeptide products. The mRNA, as a linear sequence of codons, specifies the polypeptide sequence to be made; the ribosome is simply the molecular workbench where the task is accomplished. Its structure enables it to hold the mRNA and charged tRNAs in the right positions, thus allowing the growing polypeptide to be assembled efficiently.

On the large subunit of the ribosome are four sites to which tRNA binds (see Figure 12.9). A charged tRNA traverses these four sites in order:

- The *T* (transfer) *site* is where a charged tRNA first lands on the ribosome, accompanied by a special protein "escort" called the *T*, or *transfer*, *factor*.



12.9 Ribosome Structure Each ribosome consists of a large and a small subunit. The subunits remain separate when they are not in use for protein synthesis.

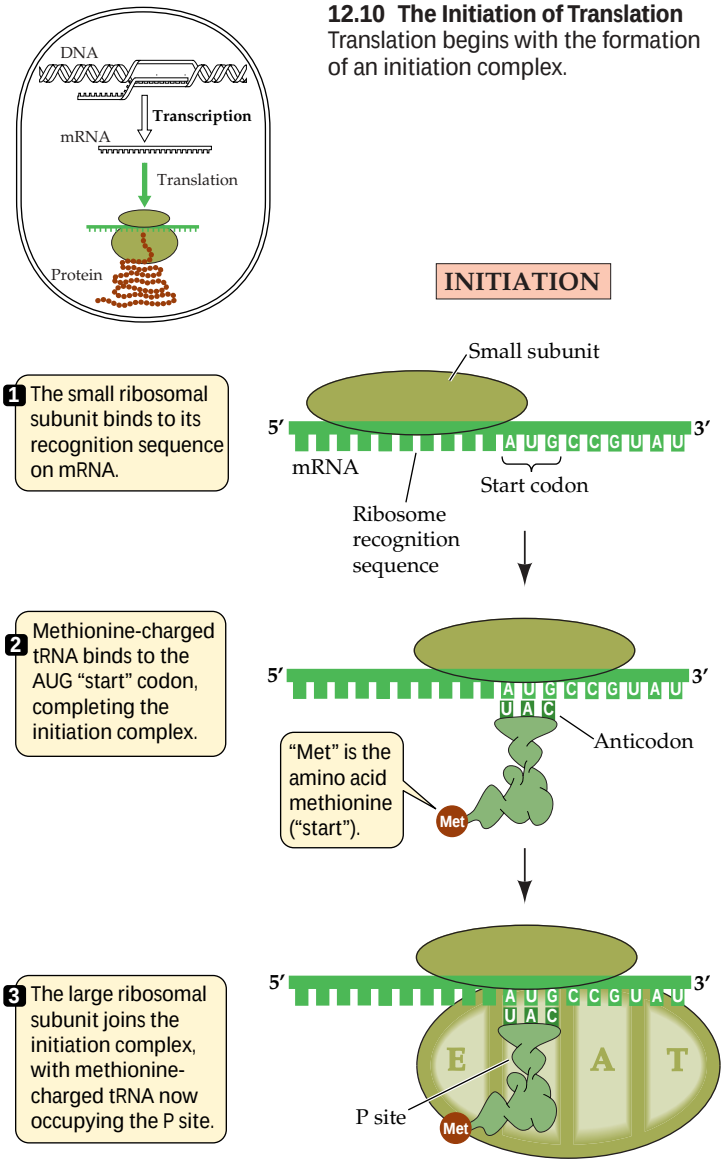
- The *A* (amino acid) *site* is where the tRNA anticodon binds to the mRNA codon, thus lining up the correct amino acid to be added to the growing polypeptide chain.
- The *P* (polypeptide) *site* is where the tRNA adds its amino acid to the growing polypeptide chain.
- The *E* (exit) *site* is where the tRNA, having given up its amino acid, resides before leaving the ribosome and going back to the cytosol to pick up another amino acid and begin the process again.

Because codon–anticodon interactions and peptide bond formation occur at the A and P sites, we will describe their function in detail in the next section.

An important role of the ribosome is to make sure that the mRNA–tRNA interactions are precise: that is, that a charged tRNA with the correct anticodon (e.g., 3′-UAC-5′) binds to the appropriate codon in mRNA (e.g., 5′-AUG-3′). When this occurs, hydrogen bonds form between the base pairs. But these hydrogen bonds are not enough to hold the tRNA in place. The rRNA of the small ribosomal subunit plays a role in validating the three-base-pair match. If hydrogen bonds have not formed between all three base pairs, the tRNA must be the wrong one for that mRNA codon, and that tRNA is ejected from the ribosome.

Translation: RNA-Directed Polypeptide Synthesis

We have been working our way through the steps by which the sequence of bases in the template strand of a DNA molecule specifies the sequence of amino acids in a protein (see Figure 12.3). We are now at the last step: translation, the RNA-directed assembly of a protein. Like transcription, translation occurs in three steps: initiation, elongation, and termination.



12.10 The Initiation of Translation
Translation begins with the formation of an initiation complex.

Translation begins with an initiation complex

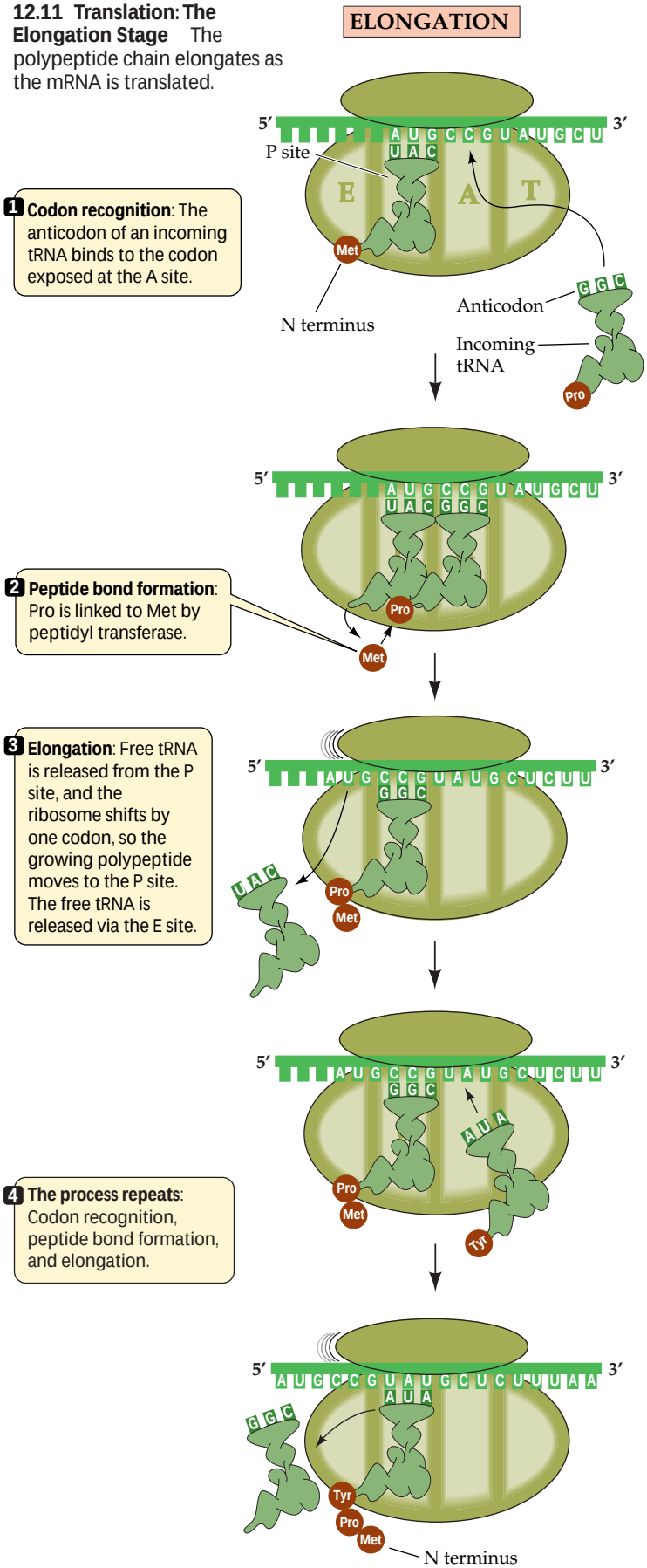
The translation of mRNA begins with the formation of an **initiation complex**, which consists of a charged tRNA bearing what will be the first amino acid of the polypeptide chain and a small ribosomal subunit, both bound to the mRNA (Figure 12.10). The rRNA of the small ribosomal subunit binds to a complementary ribosome recognition sequence on the mRNA. This sequence is "upstream" (toward the 5′ end) of the actual start codon that begins translation.

Recall that the mRNA start codon in the genetic code is AUG (see Figure 12.5). The anticodon of a methionine-charged tRNA binds to this start codon by complementary base pairing to form the initiation complex. Thus the first amino acid in the chain is always methionine. Not all mature proteins have methionine as their N-terminal amino acid, however. In many cases, the initiator methionine is removed by an enzyme after translation.

After the methionine-charged tRNA has bound to the mRNA, the large subunit of the ribosome joins the complex. The methionine-charged tRNA now lies in the P site of the ribosome, and the A site is aligned with the second mRNA codon. These ingredients—mRNA, two ribosomal subunits, and methionine-charged tRNA—are put together properly by a group of proteins called *initiation factors*.

The polypeptide elongates from the N terminus

A charged tRNA whose anticodon is complementary to the second codon on the mRNA now enters the open A site of the large ribosomal subunit. (Figure 12.11). The large subunit then catalyzes two reactions:



12.11 Translation: The Elongation Stage The polypeptide chain elongates as the mRNA is translated.

- It breaks the bond between the tRNA in the P site and its amino acid.
- It catalyzes the formation of a peptide bond between that amino acid and the one attached to the tRNA in the A site.

Because the large subunit performs these two actions, it is said to have *peptidyl transferase activity*. In this way, methionine (the amino acid in the P site) becomes the N terminus of the new protein. The second amino acid is now bound to methionine, but remains attached to its tRNA by its carboxyl group (—COOH) in the A site.

How does the large ribosomal subunit catalyze this binding? In 1992, Harry Noller and his colleagues at the University of California at Santa Cruz found that if they removed almost all the proteins in the large subunit, it still catalyzed peptide bond formation. But if the rRNA was destroyed, so was peptidyl transferase activity. Part of the rRNA in the large subunit interacts with the end of the charged tRNA where the amino acid is attached. Thus rRNA appears to be the catalyst. This situation is very unusual, because proteins are the usual catalysts in biological systems. The recent purification and crystallization of ribosomes has allowed scientists to examine their structure in detail, and the catalytic role of rRNA in peptidyl transferase activity has been confirmed.

Elongation continues and the polypeptide grows

After the first tRNA releases its methionine, it dissociates from the ribosome, returning to the cytosol to become charged with another methionine. The second tRNA, now bearing a *dipeptide*, is shifted to the P site as the ribosome moves one codon along the mRNA in the 5′-to-3′ direction.

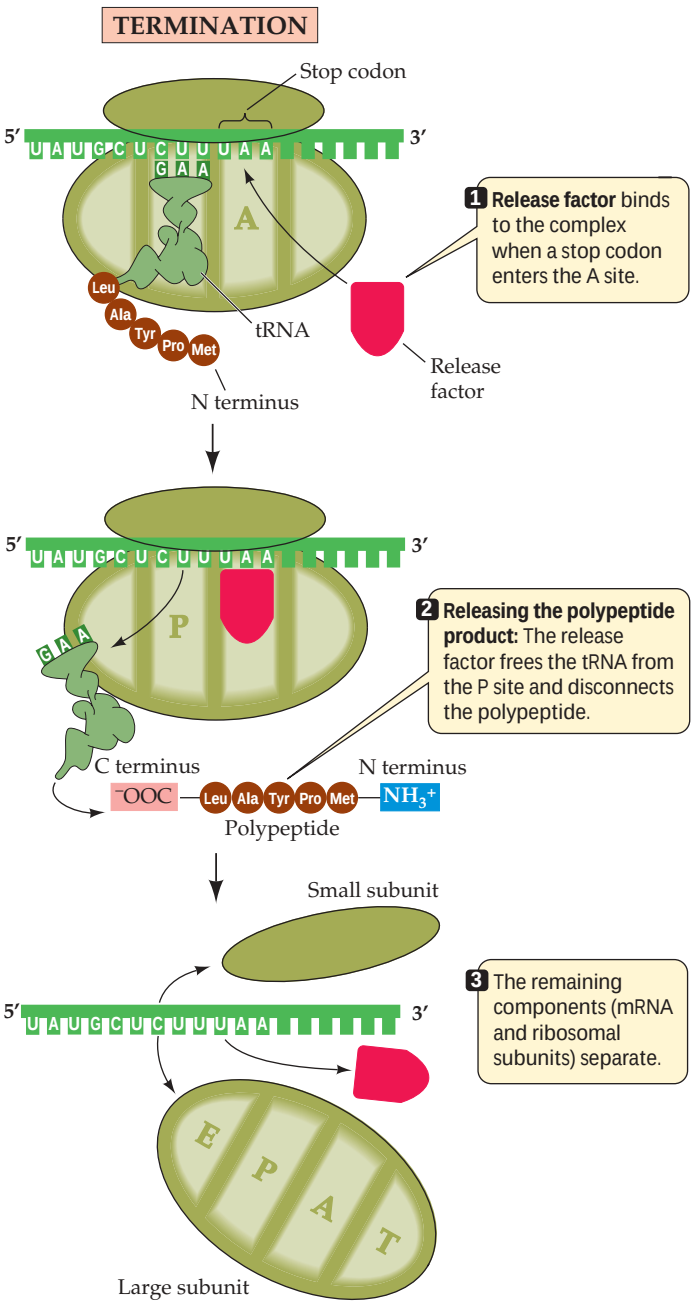
The elongation process continues, and the polypeptide chain grows, as the steps are repeated:

- The next charged tRNA enters the open A site.
- Its amino acid forms a peptide bond with the amino acid chain in the P site, so that it picks up the growing polypeptide chain from the tRNA in the P site.
- The tRNA in the P site is released. The ribosome shifts one codon, so that the entire tRNA–polypeptide complex, along with its codon, moves to the newly vacated P site.

All these steps are assisted by proteins called **elongation factors**.

A release factor terminates translation

The elongation cycle ends, and translation is terminated, when a stop codon—UAA, UAG, or UGA—enters the A site (Figure 12.12). These codons encode no amino acids, nor do



12.12 The Termination of Translation Translation terminates when the A site of the ribosome encounters a stop codon on the mRNA.

they bind tRNAs. Rather, they bind a protein *release factor*, which hydrolyzes the bond between the polypeptide and the tRNA in the P site.

The newly completed protein thereupon separates from the ribosome. Its C terminus is the last amino acid to join the chain. Its N terminus, at least initially, is methionine, as a consequence of the AUG start codon. In its amino acid sequence, it contains information specifying its conformation, as well as its ultimate cellular destination.

Table 12.1 summarizes the nucleic acid signals for initiation and termination of transcription and translation.

Regulation of Translation

Like any factory, the machinery of translation can work at varying rates. Variation in the rate of translation is useful for controlling the amount of an active protein in a cell. Some externally applied chemicals, such as some antibiotics, can stop translation. Conversely, the presence of more than one ribosome on an mRNA can speed up protein synthesis.

Some antibiotics and bacterial toxins work by inhibiting translation

Antibiotics are defensive molecules produced by microorganisms such as certain bacteria and fungi. These substances often destroy other microbes, which might compete with the defenders for nutrients. Since the 1940s, scientists have isolated increasing numbers of antibiotics, and physicians use them to treat a great variety of infectious diseases, ranging from bacterial meningitis to pneumonia to gonorrhea.

The key to the medical use of antibiotics is *specificity:* An antibiotic must act to destroy the microbial invader, but not harm the human host. One way in which antibacterial antibiotics achieve this is to block the synthesis of the bacterial cell wall—something that is essential to the microbe but is not part of human biochemistry. Penicillin works in this way.

Another way in which antibiotics work is to inhibit all bacterial protein synthesis. Recall that the prokaryotic ribosome is smaller, and has a different collection of proteins, than the eukaryotic ribosome. Some antibiotics bind only to bacterial ribosomal proteins that are important in protein synthesis (Table 12.2). Without the ability to make proteins, the bacterial invaders die, and the infection is stemmed.

Some bacteria affect their human hosts through mechanisms similar to those we use against them. Diphtheria is an infectious disease of childhood, and before the advent of effective vaccines, it was a major cause of childhood death. The infective agent, the bacterium *Cornybacterium diphtheriae*, produces a highly lethal toxin that modifies and inactivates

12.1 Signals that Start and Stop Transcription and Translation

	TRANSCRIPTION	TRANSLATION
Initiation	Promoter sequence in DNA	AUG start codon in mRNA
Termination	Terminator sequence in DNA	UAA, UAG, or UGA stop codon in mRNA

12.2 Antibiotics that Inhibit Bacterial Protein Synthesis

ANTIBIOTIC	STEP INHIBITED
Chloromycetin	Formation of peptide bonds
Erythromycin	Translocation of mRNA along ribosome
Neomycin	Interactions between tRNA and mRNA
Streptomycin	Initiation of translation
Tetracycline	Binding of tRNA to ribosome
Paromomycin	Validation of mRNA–tRNA match

a protein that is essential for the movement of mRNA and ribosomes during eukaryotic protein synthesis.

Polysome formation increases the rate of protein synthesis

Several ribosomes can work simultaneously at translating a single mRNA molecule, producing multiple molecules of the protein at the same time. As soon as the first ribosome has moved far enough from the initiation point, a second initiation complex can form, then a third, and so on. An assemblage consisting of a thread of mRNA with its beadlike ribosomes and their growing polypeptide chains is called a **polyribosome**, or **polysome** (Figure 12.13). Cells that are actively synthesizing proteins contain large numbers of polysomes and few free ribosomes or ribosomal subunits.

A polysome is like a cafeteria line, in which patrons follow one another, adding items to their trays. At any moment, the person at the start has a little food (a newly initiated protein); the person at the end has a complete meal (a completed protein). However, in the polysome cafeteria, everyone gets the same meal: Many copies of the same protein are made from a single mRNA.

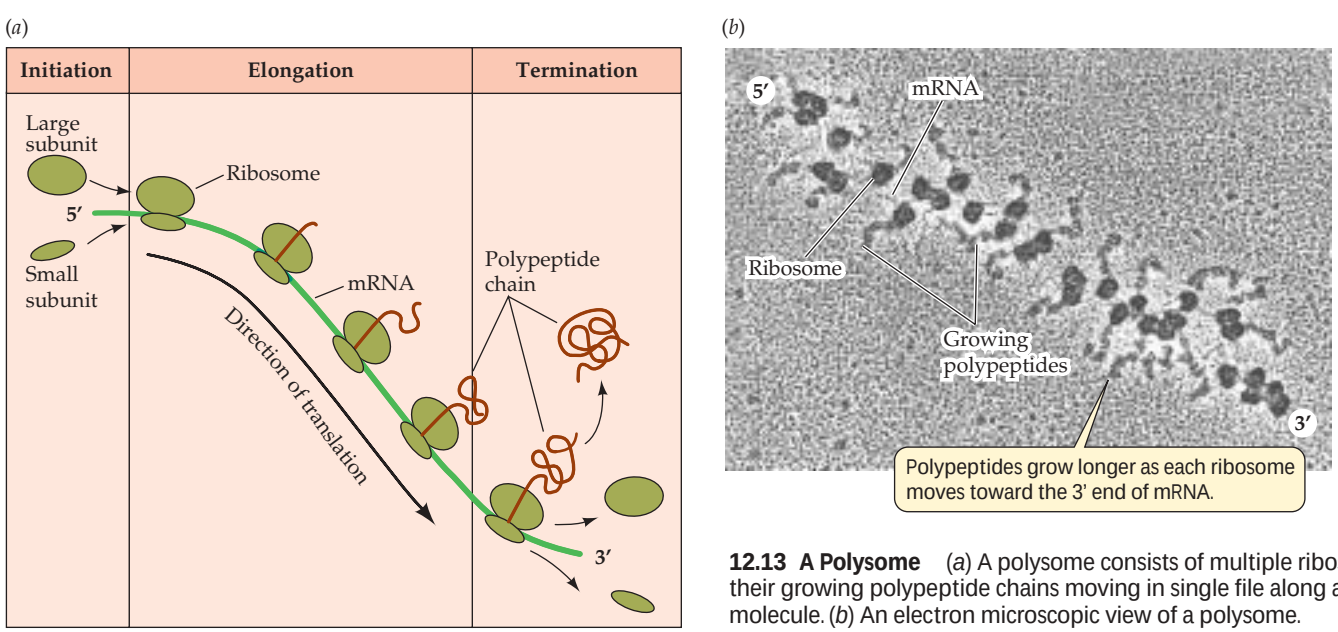
While protein synthesis can be inhibited with antibiotics and speeded up via polysomes, these are not the only ways in which the amount of an active protein in a cell can be controlled. After the protein is synthesized, it may undergo changes that alter its function.

Posttranslational Events

A functional protein is not necessarily the same as the polypeptide chain that is released from the ribosome. Especially in eukaryotic cells, the polypeptide may need to be moved far from the site of synthesis in the cytoplasm, moved into an organelle, or even secreted from the cell. In addition, the polypeptide is often modified by the addition of new chemical groups that have functional significance. In this section, we examine these two *posttranslational* aspects of protein synthesis.

Chemical signals in proteins direct them to their cellular destinations

As a polypeptide chain emerges from the ribosome, it folds into its three-dimensional shape. As described in Chapter 3, this conformation is determined by the sequence of the amino acids that make up the protein, as well as by factors such as the polarity and charge of their R groups. Ultimately, the conformation of the polypeptide allows it to interact with other



12.13 A Polysome (*a*) A polysome consists of multiple ribosomes and their growing polypeptide chains moving in single file along an mRNA molecule. (*b*) An electron microscopic view of a polysome.

molecules in the cell, such as a substrate or another polypeptide. In addition to this structural information, the amino acid sequence contains an "address label" indicating where in the cell the polypeptide belongs.

All protein synthesis begins on free ribosomes in the cytoplasm. As a polypeptide chain is made, the information contained in its amino acid sequence gives it one of two sets of instructions (Figure 12.14):

- "Finish translation and be released to the cytoplasm." Such proteins are sent to the nucleus, mitochondria, plastids, or peroxisomes, depending on the address in their instructions; or, lacking such specific instructions, they remain in the cytosol.
- "Stop translation, go to the endoplasmic reticulum (ER), and finish synthesis there." After protein synthesis is completed, such proteins may be retained in the ER or sent to lysosomes via the Golgi apparatus. Alternatively, they may be sent to the plasma membrane, or, lacking such specific instructions, they are secreted from the cell via vesicles that emanate from the plasma membrane.

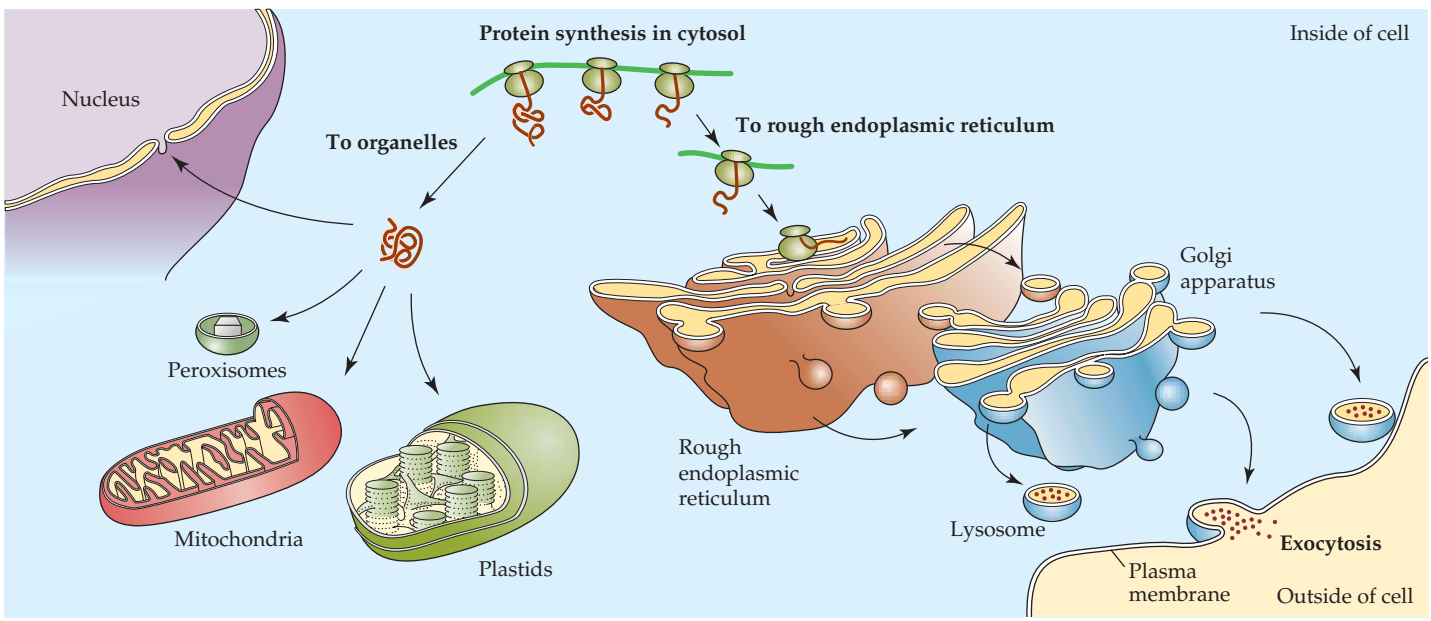
DESTINATION: CYTOPLASM. After translation, some folded polypeptides have a short exposed sequence of amino acids that acts like a postal "zip code," directing them to an organelle. These **signal sequences** are either at the N terminus or in the interior of the amino acid chain. For example, the following sequence directs a protein to the nucleus:

—Pro—Pro—Lys—Lys—Lys—Arg—Lys—Val—

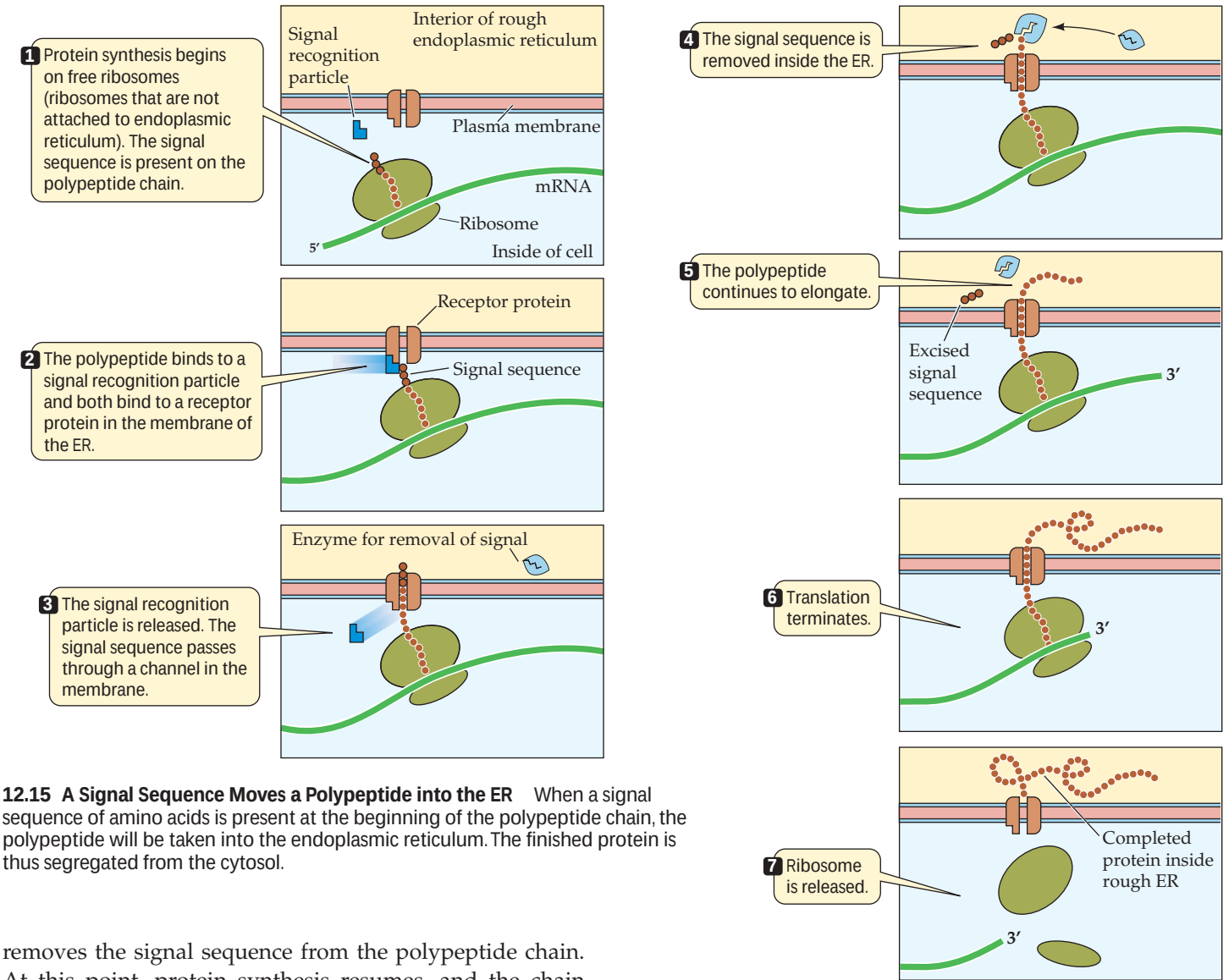
This amino acid sequence occurs, for example, in the histone proteins associated with nuclear DNA, but not in citric acid cycle enzymes, which are addressed to the mitochondria.

The signal sequences have a conformation that allows them to bind to specific receptor proteins, appropriately called *docking proteins*, on the outer membrane of the appropriate organelle. Once the protein has bound to it, the receptor forms a channel in the membrane, allowing the protein to pass through to its organelle destination. (In this process, the protein is usually unfolded by a chaperonin so that it can pass through the channel, then refolds into its normal conformation.)

DESTINATION: ENDOPLASMIC RETICULUM. If a specific hydrophobic sequence of about 25 amino acids occurs at the beginning of a polypeptide chain, the finished product is sent initially to the ER, and then to the lysosomes, the plasma membrane, or out of the cell. In the cytoplasm, before translation is finished, the signal sequence binds to a **signal recognition particle** composed of protein and RNA (Figure 12.15). This binding blocks further protein synthesis until the ribosome can become attached to a specific receptor protein in the membrane of the rough ER. Once again, the receptor protein is converted into a channel, through which the growing polypeptide passes. The elongating polypeptide may be retained in the ER membrane itself, or it may enter the interior space—the lumen—of the ER. In either case, an enzyme in the lumen of the ER



12.14 Destinations for Newly Translated Polypeptides in a Eukaryotic Cell Signal sequences on newly synthesized polypeptides bind to specific receptor proteins on the outer membranes of the organelle to which they are "addressed." Once the protein has bound to it, the receptor forms a channel in the membrane, and the protein enters the organelle.



12.15 A Signal Sequence Moves a Polypeptide into the ER When a signal sequence of amino acids is present at the beginning of the polypeptide chain, the polypeptide will be taken into the endoplasmic reticulum. The finished protein is thus segregated from the cytosol.

removes the signal sequence from the polypeptide chain. At this point, protein synthesis resumes, and the chain grows longer until its sequence is completed. If the finished protein enters the ER lumen, it can be transported to its appropriate location—to other cellular compartments or to the outside of the cell—without mixing with other molecules in the cytoplasm.

Additional signals are needed for sorting the protein further (remember that the signal sequence that sent it to the ER has been removed). These signals are of two kinds:

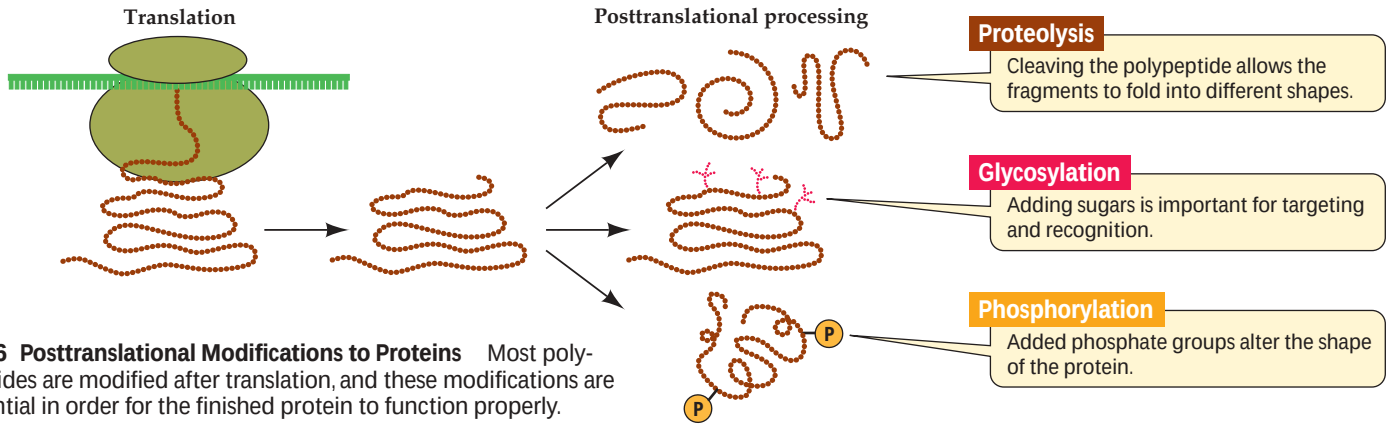
- Some are sequences of amino acids that allow the protein's retention within the ER.
- Others are sugars added in the Golgi apparatus, to which the protein is transferred in vesicles from the ER. The resulting *glycoproteins* end up either at the plasma membrane or in a lysosome (or plant vacuole), depending on which sugars are added.

Proteins with no additional signals pass from the ER through the Golgi apparatus and are secreted from the cell.

It is important to emphasize that the addressing of a protein to its destination is a property of its amino acid sequence, and so is genetically determined. An example of what can go wrong if a gene for protein targeting is mutated is *mucolipidosis II*, or *I-cell disease*. People with this disease lack an essential enzyme for the formation of the lysosomal targeting signal. Consequently, proteins destined for their lysosomes never get there, but instead either stay in the Golgi (where they form I, or inclusion, bodies) or are secreted from the cell. The lack of normal lysosome functions in a person's cells leads to progressive illness and death in childhood.

Many proteins are modified after translation

Most finished proteins are not identical to the polypeptide chains translated from mRNA on the ribosomes. Instead,



12.16 Posttranslational Modifications to Proteins Most polypeptides are modified after translation, and these modifications are essential in order for the finished protein to function properly.

most polypeptides are modified after translation, and these modifications are essential to the final functioning of the protein (Figure 12.16).

Proteolysis is the cutting of a polypeptide chain. Cleavage of the signal sequence from the growing polypeptide chain in the ER is an example of proteolysis; the protein might move back out of the ER through the membrane channel if the signal sequence were not cut off. Also, some proteins are actually made from *polyproteins* (long polypeptides) that are cut into final products by enzymes called *proteases*. Proteases are essential to some viruses, including HIV, because the large viral polyprotein cannot fold properly unless it is cut. Certain drugs used to treat AIDS work by inhibiting the HIV protease, thereby preventing the formation of proteins needed for viral reproduction (see Figure 3.10).

Glycosylation involves the addition of sugars to proteins, as described above. In both the ER and the Golgi apparatus, resident enzymes catalyze the addition of various sugar residues or short sugar chains to certain amino acid R groups on proteins as they pass through. One such type of "sugar coating" is essential for addressing proteins to lysosomes discussed in the preceding section. Other types are important in the conformation and the recognition functions of proteins at the cell surface. Still other attached sugar residues help in stabilizing proteins stored in storage vacuoles in plant seeds.

Phosphorylation, the addition of phosphate groups to proteins, is catalyzed by protein kinases. The charged phosphate groups change the conformation of targeted proteins, often exposing an active site of an enzyme or a binding site for another protein—as we will see in Chapter 15.

All of the processes we have just described result in a functional protein only if the amino acid sequence of that protein is correct. If the sequence is not correct, cellular dysfunction and disease may result. Changes in the DNA—mutations—are a major source of errors in amino acid sequences.

Mutations: Heritable Changes in Genes

Accurate DNA replication, transcription, and translation all depend on the reliable pairing of complementary bases. Errors occur, though infrequently, in all three processes—least often in DNA replication. But, the consequences of DNA errors are the most severe because only they are heritable.

Mutations are heritable changes in genetic information. In unicellular organisms, any mutations that occur are passed on to the daughter cells when the cell divides. In multicellular organisms, there are two general types of mutations in terms of inheritance:

- **Somatic mutations** are those that occur in somatic (body) cells. These mutations are passed on to the daughter cells after mitosis, and to the offspring of those cells in turn, but are not passed on to sexually produced offspring. A mutation in a single skin cell, for example, could result in a patch of skin cells, all with the same mutation, but would not be passed on to a person's children.
- **Germ line mutations** are those that occur in the cells of the *germ line*—the specialized cells that give rise to gametes. A gamete with the mutation passes it on to a new organism at fertilization.

Very small changes in the genetic material can lead to easily observable changes in the phenotype. Some effects of mutations in humans are readily detectable—dwarfism, for instance, or the presence of more than five fingers on each hand. A mutant genotype in a microorganism may be obvious if, for example, it results in a change in nutritional requirements, as we described for *Neurospora* earlier (see Figure 12.1).

Other mutations may not be easily observable. In humans, for example, a particular mutation drastically lowers the level of an enzyme called glucose 6-phosphate dehydrogenase that is present in many tissues, including red blood cells. The red blood cells of a person carrying the mutant allele are abnormally sensitive to an antimalarial drug called primaquine;

when such people are treated with this drug, their red blood cells rupture. People with the normal allele have no such problem. Before the drug came into use, no one was aware that such a mutation existed. In bacteria, because of their small sizes and simpler morphologies, distinguishing a mutant from a normal bacterium usually requires sophisticated chemical methods, not just visual inspection.

Some mutations cause their phenotypes only under certain *restrictive* conditions. They are not detectable under other, *permissive* conditions. These phenotypes are known as **conditional mutants**. Many conditional mutants are temperature-sensitive, able to grow normally at a permissive temperature—say, 30°C—but unable to grow at a restrictive temperature—say, 37°C. The mutant allele in such an organism may code for an enzyme with an unstable tertiary structure that is altered at the restrictive temperature.

All mutations are alterations in the nucleotide sequence of DNA. At the molecular level, we can divide mutations into two categories:

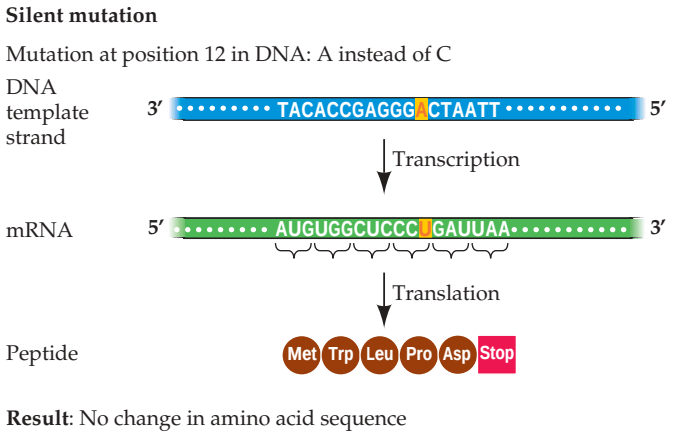
- **Point mutations** are mutations of single base pairs and so are limited to single genes: One allele (usually dominant) becomes another allele (usually recessive) because of an alteration (gain/loss or substitution) of a single nucleotide (which, after DNA replication, becomes a mutant base pair).
- **Chromosomal mutations** are more extensive alterations than point mutations. They may change the position or orientation of a DNA segment without actually removing any genetic information, or they may cause a segment of DNA to be irretrievably lost.

Point mutations are changes in single nucleotides

Point mutations result from the addition or subtraction of a nucleotide base, or the substitution of one base for another, in the DNA, and hence in the mRNA. Point mutations can be caused by errors in chromosome replication that are not corrected in proofreading or by environmental mutagens such as chemicals and radiation.

Changes in the mRNA may or may not result in changes in the protein. Silent mutations have no effect on the protein; missense and nonsense mutations will result in changes in the protein, some of them drastic.

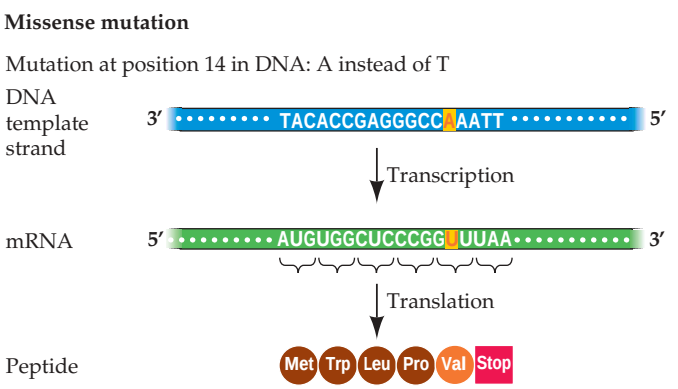
SILENT MUTATIONS. Because of the redundancy of the genetic code, some point mutations result in no change in amino acids when the altered mRNA is translated; for this reason, they are called **silent mutations**. For example, there are four mRNA codons that code for proline: CCA, CCC, CCU, and CCG (see Figure 12.5). If the template strand of DNA has the sequence CGG, it will be transcribed to CCG in mRNA, and proline-charged tRNA will bind to it at the ribosome. But if there is a mutation such that the codon in the template DNA now reads AGG, the mRNA codon will be CCU—the tRNA that binds it will still carry proline:



Result: No change in amino acid sequence

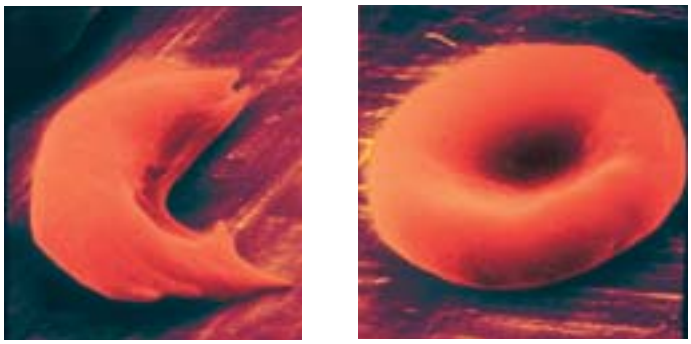
Silent mutations are quite common, and they result in genetic diversity that is not expressed as phenotypic differences.

MISSENSE MUTATIONS. In contrast to silent mutations, some base substitution mutations change the genetic message such that one amino acid substitutes for another in the protein. These changes are called **missense mutations**:



Result: Amino acid change at position 5: Val instead of Asp

A specific example of a missense mutation is the sickle allele for human β-globin. Sickle-cell disease results from a defect in hemoglobin, a protein in human red blood cells that carries oxygen. The sickle allele of the gene that codes for β-globin (one of the polypeptide subunits in hemoglobin; see Figure 3.8) differs from the normal allele by one amino acid in its coding. Persons who are homozygous for this recessive allele have defective red blood cells. Where oxygen is abundant, as in the lungs, the cells are normal in structure and function. But at the low oxygen levels characteristic of working muscles, the red blood cells collapse into the shape of a sickle (Figure 12.17), causing abnormalities in blood circulation that lead to serious illnesses.

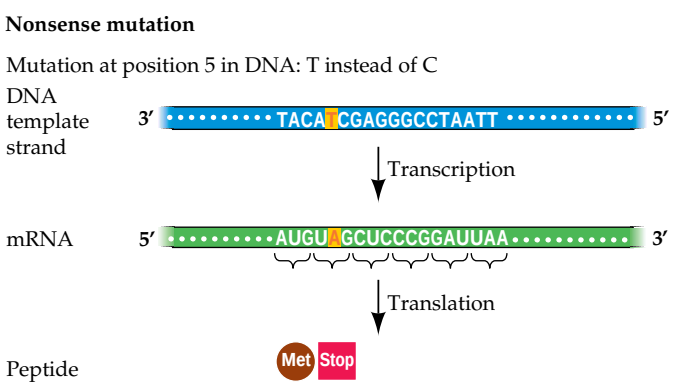


Sickle-cell phenotype Normal phenotype

12.17 Sickled and Normal Red Blood Cells The misshapen red blood cell on the left is caused by a missense mutation that results in an incorrect amino acid in one of the two polypeptides of hemoglobin.

A missense mutation may cause a protein not to function, but often its effect is only to reduce the functional efficiency of the protein. Therefore, individuals carrying missense mutations may survive, even though the affected protein is essential to life. Through evolution, some missense mutations even improve functional efficiency.

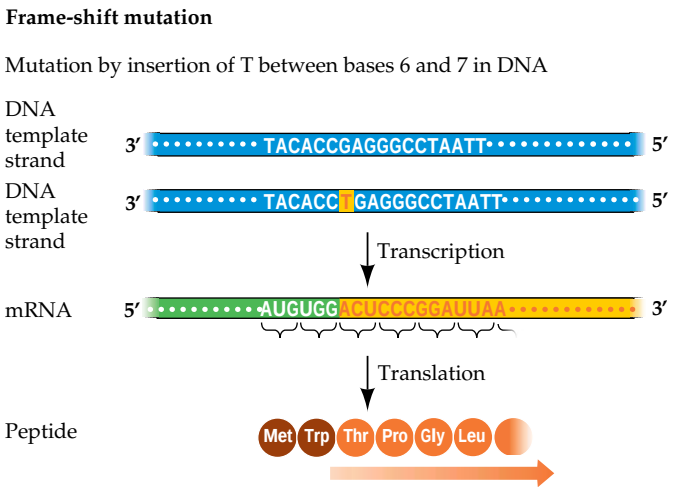
NONSENSE MUTATIONS. **Nonsense mutations**, another type of mutation in which one base is substituted for another, are more often disruptive than missense mutations. In a nonsense mutation, the base substitution causes a stop codon, such as UAG, to form in the mRNA product:



Result: Only one amino acid translated; no protein made

The result is a shortened protein, since translation does not proceed beyond the point where the mutation occurred. Such short proteins are usually not functional.

FRAME-SHIFT MUTATIONS. Not all point mutations are base substitutions. Single base pairs may be inserted into or deleted from DNA. Such mutations are known as **frame-shift mutations** because they interfere with the decoding of the genetic message by throwing it out of register:



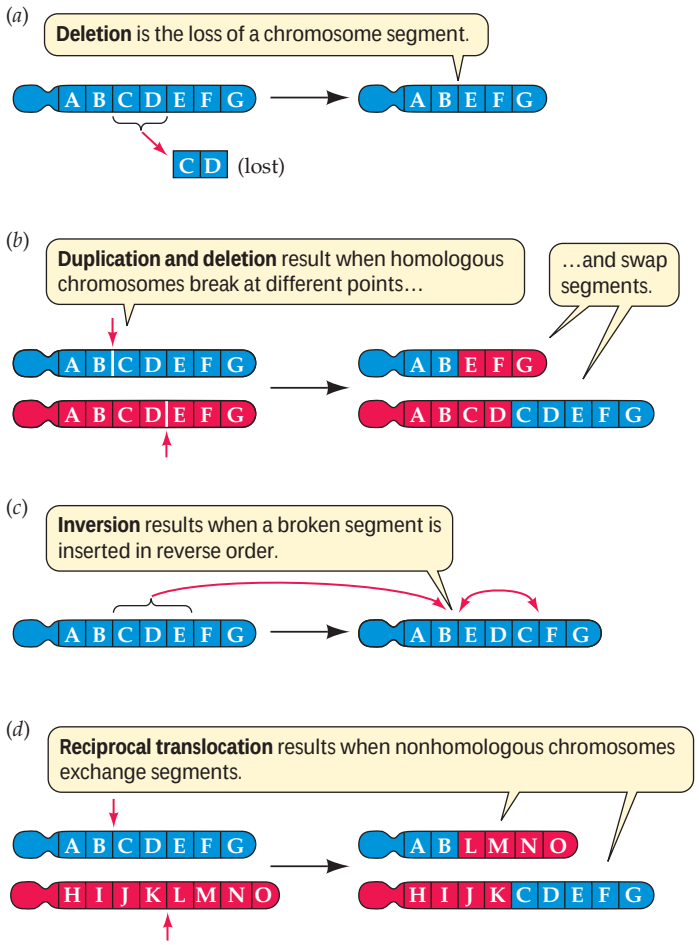
Result: All amino acids changed beyond the insertion

Think again of codons as three-letter words, each corresponding to a particular amino acid. Translation proceeds codon by codon; if a base is added to the message or subtracted from it, translation proceeds perfectly until it comes to the one-base insertion or deletion. From that point on, the three-letter words in the message are one letter out of register. In other words, such mutations shift the "reading frame" of the genetic message. Frame-shift mutations almost always lead to the production of nonfunctional proteins.

Chromosomal mutations are extensive changes in the genetic material

Changes in single nucleotides are not the most dramatic changes that can occur in the genetic material. Whole DNA molecules can break and rejoin, grossly disrupting the sequence of genetic information. There are four types of such chromosomal mutations: deletions, duplications, inversions, and translocations (Figure 12.18). These mutations can be caused by severe damage to chromosomes resulting from mutagens or by drastic errors in chromosome replication.

- **Deletions** remove part of the genetic material (Figure 12.18*a*). Like frame-shift point mutations, their consequences can be severe unless they affect unnecessary genes or are masked by the presence, in the same cell, of normal alleles of the deleted genes. It is easy to imagine one mechanism that could produce deletions: A DNA molecule might break at two points, and the two end pieces might rejoin, leaving out the DNA between the breaks.
- **Duplications** can be produced at the same time as deletions (Figure 12.18*b*). Duplication would arise if homologous chromosomes broke at different positions and then reconnected to the wrong partners. One of the two molecules produced by this mechanism would lack a seg-



12.18 Chromosomal Mutations Chromosomes may break during replication, and parts of chromosomes may then rejoin incorrectly.

ment of DNA (it would have a deletion), and the other would have two copies (a duplication) of the segment that was deleted from the first.

- **Inversions** also result from breaking and rejoining. A segment of DNA may be removed and reinserted into the same location in the chromosome, but "flipped" end over end so that it runs in the opposite direction (Figure 12.18*c*). If the break site for an inversion includes part of a DNA segment that codes for a protein, the resulting protein will be drastically altered and almost certainly nonfunctional.
- **Translocations** result when a segment of DNA breaks off, moves from its chromosome, and is inserted into a different chromosome. Translocations may be reciprocal, as in Figure 12.18*d*, or nonreciprocal, as the mutation involving duplication and deletion in Figure 12.18*b* illustrates. Translocations often lead to duplications and deletions, and may result in sterility if normal chromosome pairing in meiosis cannot occur.

Mutations can be spontaneous or induced

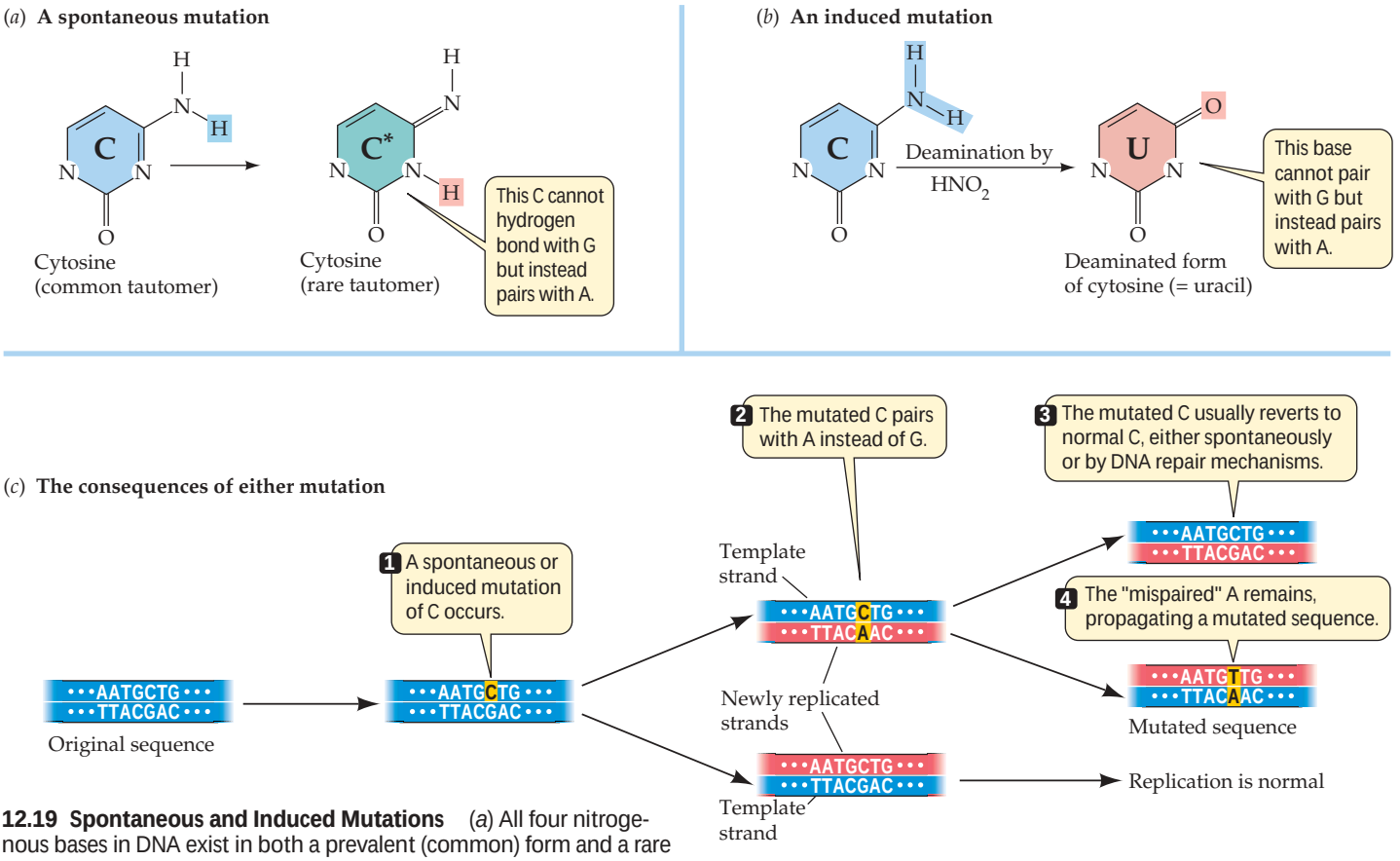
It is useful to distinguish two types of mutations in terms of their causes. **Spontaneous mutations** are permanent changes in the genome that occur without any outside influence. In other words, they occur simply because the machinery of the cell is imperfect. **Induced mutations** occur when some agent outside the cell—a mutagen—causes a permanent change in DNA.

Spontaneous mutations may occur by several mechanisms:

- *The four nucleotide bases of DNA are somewhat unstable*. They can exist in two different forms (called *tautomers*), one of which is common and one rare. When a base temporarily forms its rare tautomer, it can pair with a different base. For example, C normally pairs with G. But if C is in its rare tautomer at the time of DNA replication, it pairs with (and DNA polymerase will insert) A. The result is a point mutation: G → A (Figure 12.19*a*, *c*).
- *Bases may change because of a chemical reaction*. For example, loss of an amino group in cytosine (a reaction called *deamination*) forms uracil. When DNA replicates, instead of a G opposite what was C, DNA polymerase adds an A (base-pairs with U).
- *DNA polymerase makes errors in replication* (see Chapter 11); for example, inserting a T opposite a G. Most of these errors are repaired by the proofreading function of the replication complex, but some errors escape and become permanent.
- *Meiosis is not perfect*. Nondisjunction can occur, leading to one too many or one too few chromosomes (aneuploidy; see Figure 9.18). Random chromosome breaks and rejoining can produce deletions, duplications, and inversions, or, when involving nonhomologous chromosomes, translocations.

Mutagens can also alter DNA by several mechanisms:

- *Some chemicals can covalently alter the nucleotide bases*. For example, nitrous acid (HNO_2) and its relatives can turn cytosine in DNA into uracil by deamination: they convert an amino group on cytosine (—NH_2) into a keto group (—C═O). This alteration has the same result as a spontaneous deamination: instead of a G, DNA polymerase inserts an A (base-pairs with U) (Figure 12.19*b*,*c*).
- *Some chemicals add groups to the bases*. For instance, benzpyrene, a component of cigarette smoke, adds a large chemical group to guanine, making it unavailable for base pairing. When DNA polymerase reaches such a modified guanine, it inserts any of the four bases; of course, three-fourths of the time the inserted base will not be cytosine, and a mutation results.
- *Radiation damages the genetic material in two ways*. Ionizing radiation (X rays) produces highly reactive chemical



12.19 Spontaneous and Induced Mutations (*a*) All four nitrogenous bases in DNA exist in both a prevalent (common) form and a rare form. When a base spontaneously forms its rare tautomer, it can pair with a different base. (*b*) Mutagenic chemicals such as nitrous acid can induce changes in the bases. (*c*) In both spontaneous and induced mutations, the result is a permanent change in the DNA sequence following replication.

species called *free radicals*, which can change bases in DNA to unrecognizable (by DNA polymerase) forms or break the sugar–phosphate backbone, causing chromosomal abnormalities. Ultraviolet radiation from the sun (or a tanning lamp) is absorbed by thymine in DNA, causing it to form interbase covalent bonds with adjacent nucleotides. This, too, plays havoc with DNA replication.

Mutations have both benefits and costs. Germ line mutations provide genetic diversity for evolution to work on, as we will see below. But they usually produce an organism that does more poorly in its current environment. Somatic mutations do not affect the organism's offspring, but they can lead to cancer. We will return to the effects of germ line and somatic mutations in humans in Chapter 18.

Mutations are the raw material of evolution

Without mutation, there would be no evolution. As we will see in Part 4 of this book, mutation does not drive evolution, but it provides the genetic diversity on which natural selection and other agents of evolution act.

All mutations are rare events, but mutation frequencies vary from organism to organism and from gene to gene within a given organism. The frequency of mutation is usually much lower than one mutation per 10^4 base pairs per DNA duplication, and sometimes as low as one mutation per 10^9 base pairs per duplication. Most mutations are point mutations in which one nucleotide is substituted for another during the synthesis of a new DNA strand.

Mutations can harm the organism that carries them, or they can be neutral (have no effect on the organism's ability to survive or produce offspring). Once in a while, a mutation improves an organism's adaptation to its environment, or it becomes favorable when environmental conditions change.

Most of the complex creatures living on Earth have more DNA, and therefore more genes, than the simpler creatures do. Humans, for example, have 20 times more genes than prokaryotes have. How did these new genes arise? If whole genes were sometimes duplicated by the mechanisms described in the previous section, the bearer of the duplication would have a surplus of genetic information that might be turned to good use. Subsequent mutations in one of the two copies of the gene might not have an adverse effect on sur-

vival because the other copy of the gene would continue to produce functional protein. The extra gene might mutate over and over again without ill effect because its original function would be fulfilled by the original copy.

If the random accumulation of mutations in the extra gene led to the production of a useful protein (for example, an enzyme with an altered specificity for the substrates it binds, allowing it to catalyze different—but related—reactions), natural selection would tend to perpetuate the existence of this new gene. New copies of genes may also arise through the activity of transposable elements, which are discussed in Chapters 13 and 14.

Chapter Summary

One Gene, One Polypeptide

▶ Genes are expressed in the phenotype as polypeptides (proteins).

▶ Beadle and Tatum's experiments with the bread mold *Neurospora* resulted in several mutant strains, each lacking a specific enzyme in a biochemical pathway. Their results led to the one-gene, one-polypeptide hypothesis. **Review Figure 12.1**

▶ Certain hereditary diseases in humans had been found to be caused by the absence of certain enzymes. These observations supported the one-gene, one-polypeptide hypothesis.

DNA, RNA, and the Flow of Information

▶ RNA differs from DNA in three ways: It is single-stranded, its sugar molecule is ribose rather than deoxyribose, and its fourth base is uracil rather than thymine.

▶ The central dogma of molecular biology is DNA → RNA → protein. **Review Figure 12.2**

▶ A gene is expressed in two steps: first, DNA is transcribed to RNA; then RNA is translated into protein. **Review Figure 12.3**

▶ Some viruses are exceptions to the central dogma. Some viruses exclude DNA altogether, going directly from RNA to protein. In retroviruses, the central dogma is reversed: RNA → DNA.

Transcription: DNA-Directed RNA Synthesis

▶ RNA is transcribed from a DNA template after the bases of DNA are exposed by unwinding of the double helix.

▶ In a given gene, only one of the two strands of DNA (the template strand) acts as a template for transcription.

▶ RNA polymerase catalyzes transcription from the template strand of DNA.

▶ The initiation of transcription requires that RNA polymerase recognize and bind tightly to a promoter sequence on the DNA.

▶ RNA elongates in a 5′-to-3′ direction, antiparallel to the template DNA. Special sequences and protein helpers terminate transcription. **Review Figure 12.4**

▶ In prokaryotes, translation begins before transcription of the mRNA is completed. In eukaryotes, transcription occurs in the nucleus and translation occurs in the cytoplasm.
See Web/CD Tutorial 12.1

The Genetic Code

▶ The genetic code consists of triplets of nucleotide bases (codons). There are four bases, so there are 64 possible codons.

▶ One mRNA codon indicates the starting point of translation and codes for methionine. Three stop codons indicate the end of translation. The other 60 codons code only for particular amino acids.

▶ Because there are only 20 different amino acids, the genetic code is redundant; that is, there is more than one codon for certain amino acids. But the code is not ambiguous: A single codon does not encode more than one amino acid. **Review Figure 12.5.**
See Web/CD Tutorial 12.1

▶ Test-tube experiments led to the assignment of amino acids to codons. **Review Figure 12.6.** **See Web/CD Tutorial 12.2**

Preparation for Translation: Linking RNAs, Amino Acids, and Ribosomes

▶ In translation, amino acids are linked in an order specified by the codons in mRNA. This task is achieved by transfer RNAs (tRNAs), which bind to specific amino acids. Each tRNA species has an anticodon complementary to an mRNA codon. **Review Figure 12.7**

▶ A family of activating enzymes attaches specific amino acids to their appropriate tRNAs, forming charged tRNAs. **Review Figure 12.8**

▶ The mRNA meets the charged tRNAs at a ribosome. **Review Figure 12.9**

▶ The small subunit of the ribosome checks to determine whether the tRNA anticodon and mRNA codon have formed hydrogen bonds.

Translation: RNA-Directed Polypeptide Synthesis

▶ An initiation complex consisting of a charged tRNA and a small ribosomal subunit bound to mRNA triggers the beginning of translation. **Review Figure 12.10**

▶ Polypeptides grow from the N terminus toward the C terminus. The ribosome moves along the mRNA one codon at a time in the 5′-to-3′ direction. **Review Figure 12.11**

▶ The presence of a stop codon in the A site of the ribosome terminates translation. **Review Figure 12.12**
See Web/CD Tutorial 12.3

Regulation of Translation

▶ Some antibiotics and bacterial toxins work by blocking events in translation. **Review Table 12.2**

▶ In a polysome, more than one ribosome moves along the mRNA at one time. **Review Figure 12.13**

Posttranslational Events

▶ Signals contained in the amino acid sequences of proteins direct them to their cellular destinations. **Review Figure 12.14**

▶ Protein synthesis begins on free ribosomes in the cytoplasm. Those proteins destined for the nucleus and other organelles are completed there. These proteins have signals that allow them to bind to and enter their destined organelles.

▶ Proteins destined for the ER, Golgi apparatus, lysosomes, and outside the cell complete their synthesis on the surface of the ER. They enter the ER by the interaction of a hydrophobic signal sequence with a channel in the membrane. **Review Figure 12.15**

▶ Modifications of proteins after translation include proteolysis, glycosylation, and phosphorylation. **Review Figure 12.16**

Mutations: Heritable Changes in Genes

▶ Mutations in DNA are often expressed as abnormal proteins. However, the result may not be easily observable phenotypic changes. Some mutations are detectable only under certain conditions.

▶ Point mutations (silent, missense, nonsense, or frame-shift) result from alterations in single base pairs of DNA. **Review pages 251–252**

▶ Chromosomal mutations (deletions, duplications, inversions, or translocations) involve large regions of a chromosome. **Review Figure 12.18**

▶ Mutations can be spontaneous or induced. Spontaneous mutations occur because of instabilities in DNA or chromosomes. Induced mutations occur when a mutagen damages DNA. **Review Figure 12.19**

Self-Quiz

1. Which of the following is *not* a difference between RNA and DNA?
   a. RNA has uracil; DNA has thymine.
   b. RNA has ribose; DNA has deoxyribose.
   c. RNA has five bases; DNA has four.
   d. RNA is a single polynucleotide strand; DNA is a double strand.
   e. RNA is relatively smaller than human chromosomal DNA.
2. Normally, *Neurospora* can synthesize all 20 amino acids. A certain strain of this mold cannot grow in simple growth medium but grows only when the amino acid leucine is added to the medium. This strain is
   a. dependent on leucine for energy.
   b. mutated in the synthesis of all proteins.
   c. mutated in the synthesis of all 20 amino acids
   d. mutated in the synthesis of leucine.
   e. mutated in the syntheses of 19 of the 20 amino acids.
3. An mRNA has the sequence 5′-AUGAAAUCCUAG-3′. What is the template DNA strand for this sequence?
   a. 5′-TACTTTAGGATC-3′
   b. 5′-ATGAAATCCTAG-3′
   c. 5′-GATCCTAAAGTA-3′
   d. 5′-TACAAATCCTAG-3′
   e. 5′-CTAGGATTTCAT-3′
4. The adapters that allow translation of the four-letter nucleic acid language into the 20 letter protein language are called
   a. aminoacyl tRNA synthetases.
   b. transfer RNAs.
   c. ribosomal RNAs.
   d. messenger RNAs.
   e. ribosomes.
5. At a certain location in a gene, the nontemplate strand of DNA has the sequence GAA. A mutation alters the triplet to GAG. This type of mutation is called
   a. silent.
   b. missense.
   c. nonsense.
   d. frame-shift.
   e. translocation.
6. Transcription
   a. produces only mRNA.
   b. requires ribosomes.
   c. requires tRNAs.
   d. produces RNA growing from the 5′ end to the 3′ end.
   e. takes place only in eukaryotes.
7. Which statement about translation is *not* true?
   a. It is RNA-directed polypeptide synthesis.
   b. An mRNA molecule can be translated by only one ribosome at a time.
   c. The same genetic code operates in almost all organisms and organelles.
   d. Any ribosome can be used in the translation of any mRNA.
   e. There are both start and stop codons.
8. Which statement is *not* true?
   a. Transfer RNA functions in translation.
   b. Ribosomal RNA functions in translation.
   c. RNAs are produced in transcription.
   d. Messenger RNAs are produced on ribosomes.
   e. DNA codes for mRNA, tRNA, and rRNA.
9. The genetic code
   a. is different for prokaryotes and eukaryotes.
   b. has changed during the course of recent evolution.
   c. has 64 codons that code for amino acids.
   d. is degenerate.
   e. is ambiguous.
10. A mutation that results in the codon UAG where there had been UGG is
    a. a nonsense mutation.
    b. a missense mutation.
    c. a frame-shift mutation.
    d. a large-scale mutation.
    e. unlikely to have a significant effect.

For Discussion

1. The genetic code is described as degenerate. What does this mean? How is it possible that a point mutation, consisting of the replacement of a single nitrogenous base in DNA by a different base, might not result in an error in protein production?
2. Har Gobind Khorana at the University of Wisconsin synthesized artificial mRNAs such as poly CA (CACA…) and poly CAA (CAACAACAA…). He found that poly CA codes for a polypeptide consisting of threonine (Thr) and histidine (His), in alternation (His–Thr– His–Thr…). There are two possible codons in poly CA, CAC and ACA. One of these must code for histidine and the other for threonine—but which is which? The answer comes from results with poly CAA, which produces three different polypeptides: poly Thr, poly Gln (glutamine), and poly Asn (asparagine). (An artificial messenger can be read, inefficiently, beginning at any point in the chain; there is no specific initiator region.) Thus poly CAA can be read as a polymer of CAA, of ACA, or of AAC. Compare the results of the poly CA and poly CAA experiments, and determine which codon codes for threonine and which for histidine.
3. Look back at Question 2. Using the genetic code (Figure 12.5) as a guide, deduce what results Khorana would have obtained had he used poly UG and poly UGG as artificial messengers. In fact, very few such artificial messengers would have given useful results. For an example of what could happen, consider poly CG and poly CGG. If poly CG were the messenger, a mixed polypeptide of arginine and alanine (Arg–Ala–Ala–Arg . . .) would be obtained; poly CGG would give three polypeptides: poly Arg, poly Ala, and poly Gly (glycine). Can any codons be determined from only these data? Explain.
4. Errors in transcription occur about 100,000 times as often as do errors in DNA replication. Why can this high rate be tolerated in RNA synthesis but not in DNA synthesis?

10 The Genetics of Viruses and Prokaryotes

Robert Stevens felt sicker and sicker, until he finally went to the emergency room of a hospital near his home in Boca Raton, Florida. The medical staff noted his fever, vomiting, and headache and tested his spinal fluid for infectious agents. They saw a few spores of *Bacillus* bacteria, which they might have dismissed as contamination if some of the hospital staff had not just taken a course on identifying possible germ warfare agents at the U.S. Centers for Disease Control and Prevention.

The spores were put in a culture dish with a growth medium, and colonies of bacteria soon appeared. They were identified as *Bacillus anthracis*—the anthrax bacterium. Stevens had respiratory anthrax, a rare form of the disease that he apparently picked up when he inhaled spores deliberately placed in an envelope sent to the newspaper where he worked. The doctors gave him antibiotics to stem the growth of the infection, but it was too late. The rapidly dividing bacteria produced toxins that overwhelmed his body's defenses. Three days later, on October 5, 2001, Robert Stevens died.

When its genome was sequenced, the killer bacterium was found to belong to a strain of *B. anthracis* that had been used in the U.S. government's biological weapons research program until it was disbanded by international agreement in 1969.

Behind the fears of bioterrorism that surrounded the anthrax infections of Stevens and others lie many aspects of prokaryotic genetics and molecular biology. In this chapter, we will describe some of the science behind the headlines, looking at such aspects as bacterial growth and colony formation, exchanges of genetic material, and genome sequencing. Prokaryotes usually reproduce asexually by cell division, but they can acquire new genes in several ways. These mechanisms range from simple recombination in a sexual process to the transport of genes by infective viruses. We will also describe how the expression of prokaryotic genes is regulated and what DNA sequencing has revealed about the prokaryotic genome.

Viruses are not prokaryotes. In fact, they are not even cells, but intracellular parasites that can reproduce only within living cells. We will begin this chapter with a look at the benefits of studying prokaryotes and viruses. Then we will examine the structure, classification, reproduction, and genetics of viruses.

A Prokaryote Weapon This composite micrograph shows spores of *Bacillus anthracis* (yellow) in human lung tissue. *B. anthracis* is the cause of anthrax, a disease that can be fatal to many mammals, including humans. Because anthrax is transmitted by hardy spores that can survive long periods exposed to the environment, it is a prime weapon for bioterrorism.

Probing the Nature of Genes

Prokaryotes and the viruses that infect them have always been important tools for studying the structure, function, and transmission of genes. There are several advantages of working with prokaryotes and viruses:

- Their genomes are small. A typical bacterium contains about a thousandth as much DNA as a single human cell, and a typical bacteriophage contains about a hundredth as much DNA as a bacterium.
- They quickly produce large numbers of individuals. A single milliliter of growth medium can contain more than 10^9 cells of the bacterium *Escherichia coli*, and its numbers can double every 20 minutes.
- Prokaryotes and viruses are usually haploid, which makes genetic analyses easier.

The ease of growing and handling bacteria and their viruses permitted the explosion of genetics and molecular biology that began shortly after the mid-twentieth century. Their relative biological simplicity contributed immeasurably to discoveries about the genetic material, the replication of DNA, and the mechanisms of gene expression. Later they were the first subjects of recombinant DNA technology (see Chapter 16).

Questions of interest to all biologists continue to be studied in prokaryotes, and prokaryotes continue to be important tools for biotechnology and for research on eukaryotes. Prokaryotes also play vital roles in the ecosystem, performing much of the cycling of elements in the soil, atmosphere, and water (see Chapter 58). And, as we saw at the opening of this chapter, prokaryotes and viruses that are **pathogens**—those that cause infectious diseases—continue to challenge humankind.

Viruses: Reproduction and Recombination

Although there are many kinds of viruses, most of them are composed of a nucleic acid and a few proteins. Unlike the organisms that make up the three taxonomic kingdoms of the living world, viruses are *acellular;* that is, they are not cells and do not consist of cells. Viruses do not carry out two of the basic functions of cellular life: they do not regulate the transport of substances into and out of them by membranes, and they do not perform energy metabolism. Furthermore, they can reproduce only in systems that do perform these functions: living cells.

Scientists studied viruses before they could see them

Most viruses are much smaller than even the mycoplasmas—the smallest bacteria (Table 13.1). Viruses have become well understood only within the last half century, but the first step on this path of discovery was taken by the Russian botanist Dmitri Ivanovsky in 1892. He was trying to find the cause of tobacco mosaic disease, which results in the destruction of photosynthetic tissues in plants and can devastate a tobacco crop. Ivanovsky passed an extract of diseased tobacco leaves through a fine porcelain filter, a technique that had been used previously by physicians and veterinarians to isolate disease-causing bacteria.

To Ivanovsky's surprise, the disease agent in this case was not retained on the filter. It passed through, and the liquid filtrate still caused tobacco mosaic disease. But instead of concluding that the agent was smaller than a bacterium, he assumed that his filter was faulty. Pasteur's recent demonstration that bacteria could cause disease was the dominant idea at the time, and Ivanovsky chose not to challenge it. But, as often happens in science, someone soon came along who did. In 1898, the Dutch microbiologist Martinus Beijerinck repeated Ivanovsky's experiment and also showed that the tobacco mosaic disease agent could diffuse through an agar gel. He called the tiny agent *contagium vivum fluidum*, which later became shortened to *virus*.

Almost 40 years later, the disease agent was crystallized by Wendell Stanley (who won the Nobel prize for his efforts). The crystalline viral preparation became infectious again when it was dissolved. It was soon shown that crystallized viral preparations consist of proteins and nucleic acids. Finally, direct observation of viruses with electron microscopes in the 1950s showed clearly how much they differ from bacteria and other organisms. The simplest infective agents of all are *viroids*, which are made up only of genetic material.

13.1 Relative Sizes of Microorganisms

MICROORGANISM	TYPE	TYPICAL SIZE RANGE (μm^3)
Protists	Eukaryote	5,000–50,000
Photosynthetic bacteria	Prokaryote	5–50
Spirochetes	Prokaryote	0.1–2.0
Mycoplasmas	Prokaryote	0.01–0.1
Poxviruses	Virus	0.01
Influenza virus	Virus	0.0005
Poliovirus	Virus	0.00001

Viruses reproduce only with the help of living cells

Whole viruses never arise directly from preexisting viruses. Viruses are *obligate intracellular parasites;* that is, they develop and reproduce only within the cells of specific hosts. The cells

of animals, plants, fungi, protists, and prokaryotes (both bacteria and archaea) can serve as hosts to viruses. Viruses use the host's synthetic machinery to reproduce themselves, usually destroying the host cell in the process. The host cell releases progeny viruses, which then infect new hosts.

Viruses outside of host cells exist as individual particles called **virions**. The virion, the basic unit of a virus, consists of a central core of either DNA or RNA (but not both) surrounded by a **capsid**, or coat, composed of one or more proteins. Because they lack the distinctive cell wall and ribosomal biochemistry of bacteria, viruses are not affected by antibiotics.

There are many kinds of viruses

There are four ways to describe viruses:

- Whether the genome is DNA or RNA
- Whether the nucleic acid is single-stranded or double-stranded
- Whether the shape of the virion is a simple or complex crystal
- Whether the virion is surrounded by a membrane

Some of these variations are shown in Figure 13.1.

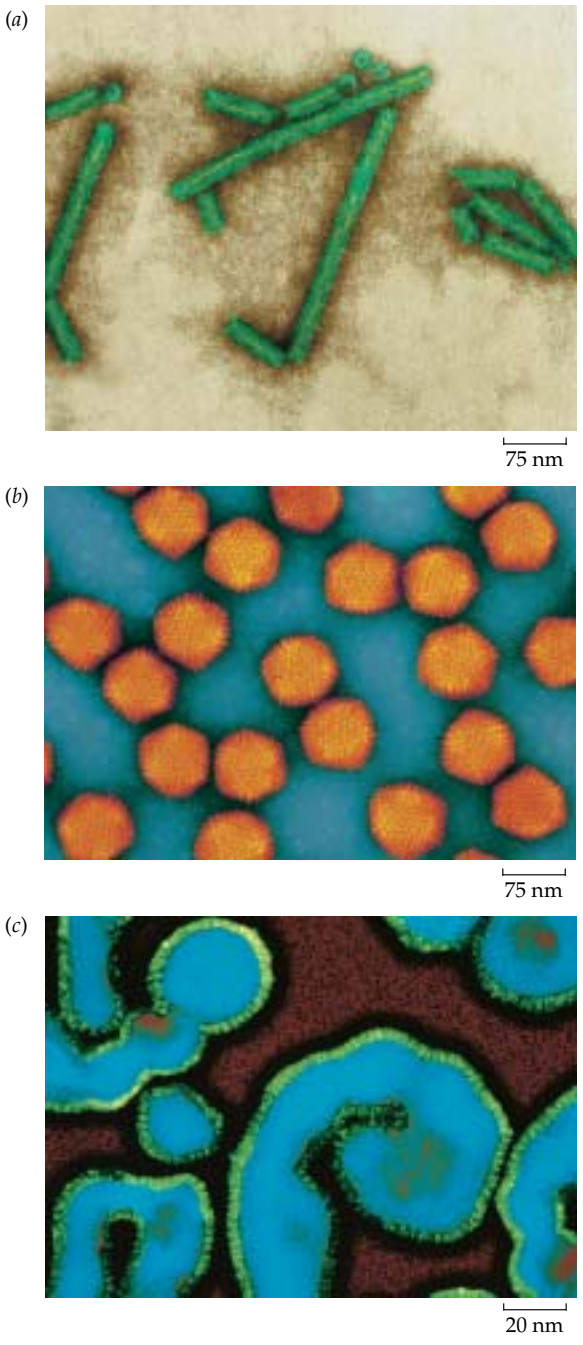
Another important descriptor of a virus is the type of organisms it infects. Most viruses have relatively simple means of infecting their host cells. Some can infect a cell but postpone reproduction, remaining inactive in the host cell until conditions are favorable.

Bacteriophage reproduce by a lytic cycle or a lysogenic cycle

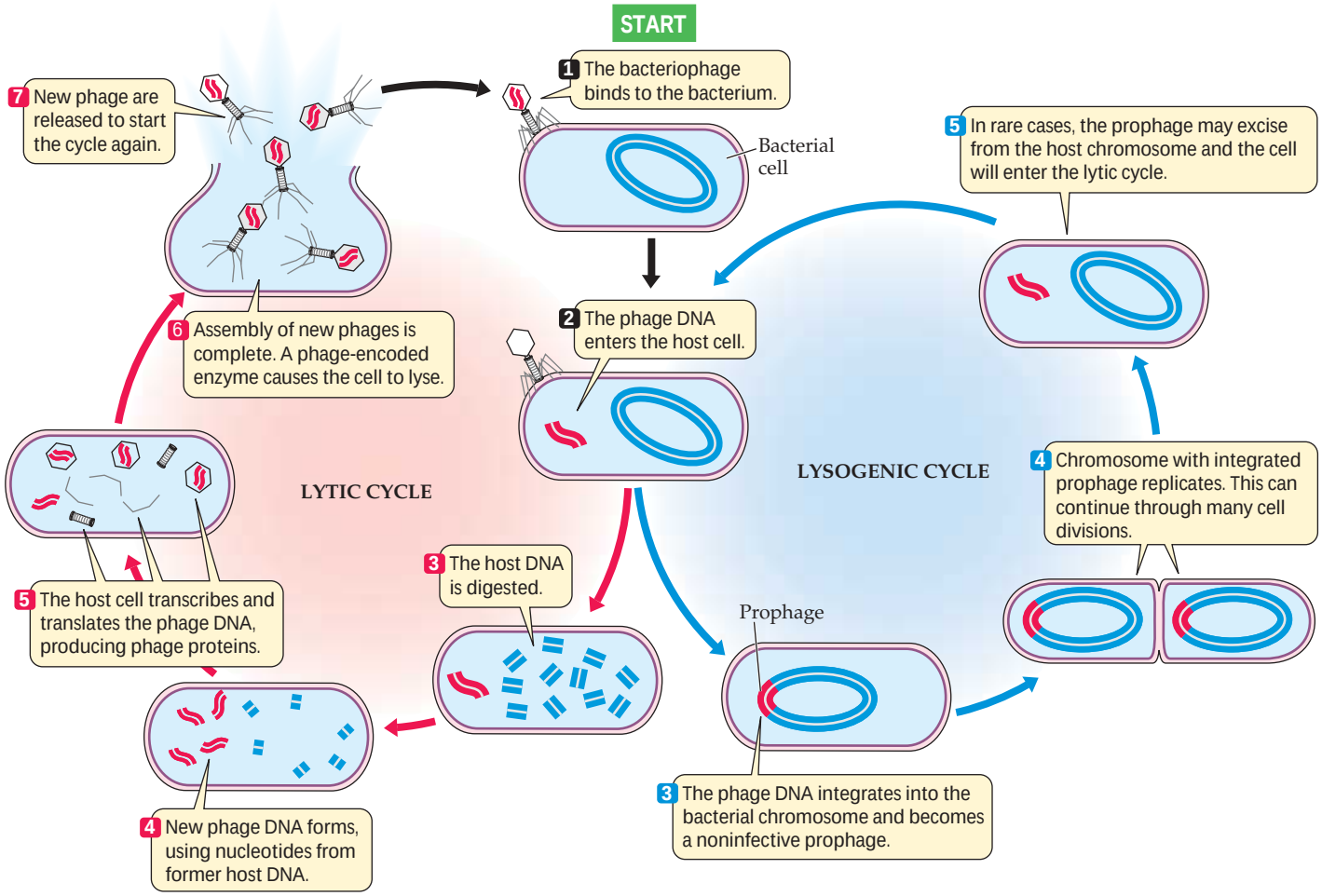
Viruses that infect bacteria are known as **bacteriophage** or **phage** (Greek *phagos*, "one that eats"). They recognize their hosts by means of proteins in the capsid, which bind to specific receptor proteins or carbohydrates in the host's cell wall. The virions, which must penetrate the cell wall, are often equipped with tail assemblies that inject the phage's nucleic acid through the cell wall into the host bacterium. After the nucleic acid has entered the host, one of two things happens, depending on the kind of phage:

- The virus may reproduce immediately and kill the host cell.
- The virus may postpone reproduction by integrating its nucleic acid into the host cell's genome.

We saw one type of viral reproductive cycle when we studied the Hershey–Chase experiment (see Figure 11.3). That was the **lytic cycle**, so named because the infected bacterium *lyses* (bursts), releasing progeny phage. The alternative fate is the **lysogenic cycle**, in which the infected bacterium does not lyse, but instead harbors the viral nucleic acid for many generations. Some viruses reproduce only by the lytic cycle; others undergo both types of reproductive cycles (Figure 13.2).



13.1 Virions Come in Various Shapes (*a*) The tobacco mosaic virus (a plant virus) consists of an inner helix of RNA covered with a helical array of protein molecules. (*b*) Many animal viruses, such as this adenovirus, have an icosahedral (20-sided) capsid as an outer shell. Inside the shell is a spherical mass of proteins and DNA. (*c*) Not all virions are regularly shaped. These wormlike virions of the influenza A virus infect humans, causing chills, fever, and sometimes, death.



13.2 The Lytic and Lysogenic Cycles of Bacteriophage In the lytic cycle, infection by viral DNA leads directly to the multiplication of the virus and lysis of the host bacterial cell. In the lysogenic cycle, an inactive prophage is replicated as part of the host's chromosome.

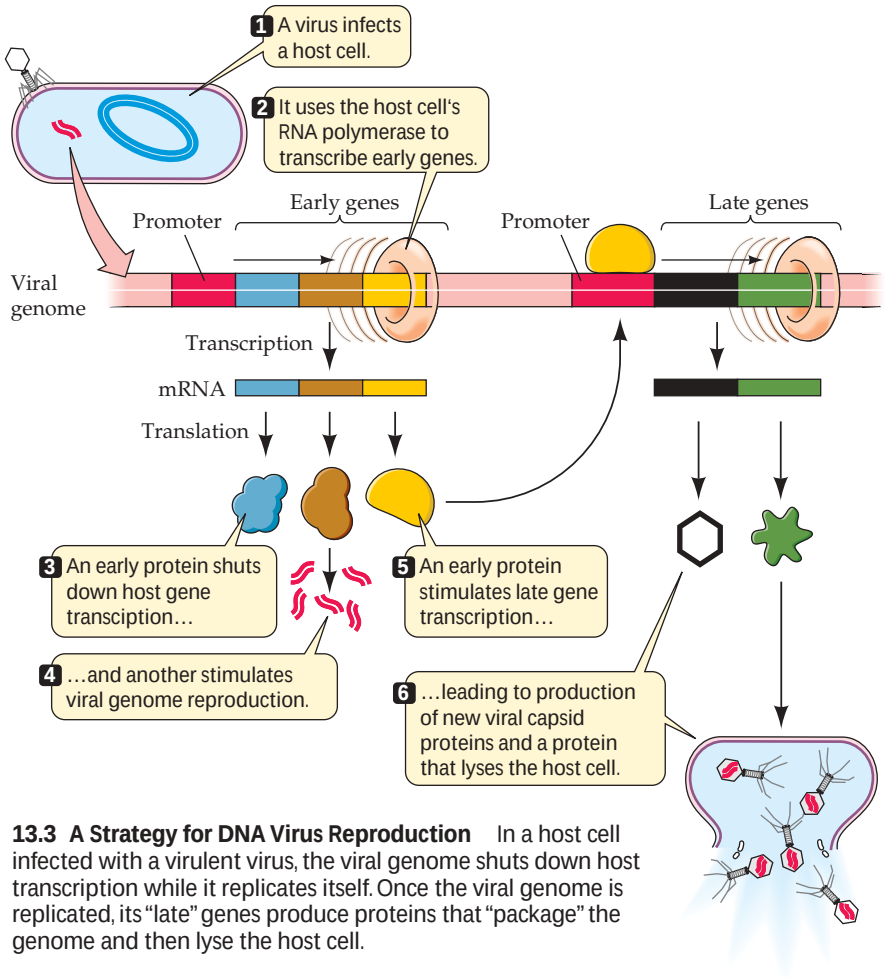
THE LYTIC CYCLE. A virus that reproduces only by the lytic cycle is called a **virulent** virus. Once the phage has injected its nucleic acid into the host cell, that nucleic acid takes over the host's synthetic machinery. It does so in two stages (Figure 13.3):

- The viral genome contains a promoter sequence that attracts host RNA polymerase. In the *early stage*, viral genes that lie adjacent to this promoter are transcribed. These *early genes* often code for proteins that shut down host transcription, stimulate viral genome replication, and stimulate late gene transcription. Nuclease enzymes digest the host's chromosome, providing nucleotides for the synthesis of viral genomes.
- In the *late stage*, viral *late genes*, which code for the proteins of the viral capsid and those that lyse the host cell to release the new virions, are transcribed.

This sequence of transcriptional events is carefully controlled: Premature lysis of the host cell before virions are ready for release would stop the infection. The whole process—from binding and infection to lysis of the host cell—takes about half an hour.

Rarely, two viruses infect a cell at the same time. This is an unusual event, as once an infection cycle is under way, there is usually not enough time for an additional infection. In addition, an early protein prevents further infections in some cases. However, the presence of two different viral genomes in the same host cell affords the opportunity for genetic recombination by crossing over (as in prophase I of meiosis in eukaryotes). This phenomenon enables genetically different viruses of the same kind to swap genes and create new strains.

THE LYSOGENIC CYCLE. Phage infection does not always result in lysis of the host cell. Some phage seem to disappear from a bacterial culture, leaving the bacteria immune to further attack by the same strain of phage. In such cultures, however, a few free phage are always present. Bacteria harboring phage that are not lytic are called *lysogenic bacteria*, and the viruses are called **temperate** viruses.



13.3 A Strategy for DNA Virus Reproduction In a host cell infected with a virulent virus, the viral genome shuts down host transcription while it replicates itself. Once the viral genome is replicated, its "late" genes produce proteins that "package" the genome and then lyse the host cell.

Lysogenic bacteria contain a noninfective entity called a **prophage**: a molecule of phage DNA that has been integrated into the bacterial chromosome (see Figure 13.2). The prophage can remain inactive within the bacterial genome through many cell divisions. However, an occasional lysogenic bacterium can be induced to activate its prophage. This activation results in a lytic cycle, in which the prophage leaves the host chromosome and reproduces.

This capacity to switch between the lysogenic and the lytic cycle is very useful to the phage because it enhances the production of the maximum number of progeny viruses. When its host cell is growing rapidly, the phage is lysogenic. When the host is stressed or damaged by mutagens, the prophage is released from its inactive state, and the lytic cycle proceeds. We will see how this switch works later in the chapter when we discuss the regulation of gene expression.

Lytic bacteriophage could be useful in treating bacterial infections

Since lytic bacteriophage destroy their bacterial hosts, they might be useful in treating infectious diseases caused by bacteria. Indeed, one of the early discoverers of phage, the French-Canadian microbiologist Felix D'Herelle, noted in 1917 (before antibiotics were discovered) that when some patients with bacterial dysentery were recovering from the disease, the quantity of phage near the bacteria was much higher than when the disease was at its peak.

D'Herelle tried using phage to control infections of chickens by the bacterium *Salmonella gallinarum*. To do this, he divided chickens into two groups, one that was given phage and another that was not. Then he exposed both groups to the infectious bacteria. The phage-protected group did not get the bacterial disease. Later, he used phage successfully to treat people in Egypt infected with plague-causing bacteria and people in India with infectious cholera.

The emergence of antibiotics and of phage-resistant bacteria reduced interest in phage therapy. However, interest has revived now that bacterial resistance to antibiotics is becoming common. Bacteriophage are even being investigated as a means of treating edible fruits and vegetables to prevent bacterial contamination. In addition to advancing our understanding of fundamental biological processes, the study of bacteriophage opened the door to investigations of viruses that infect eukaryotes.

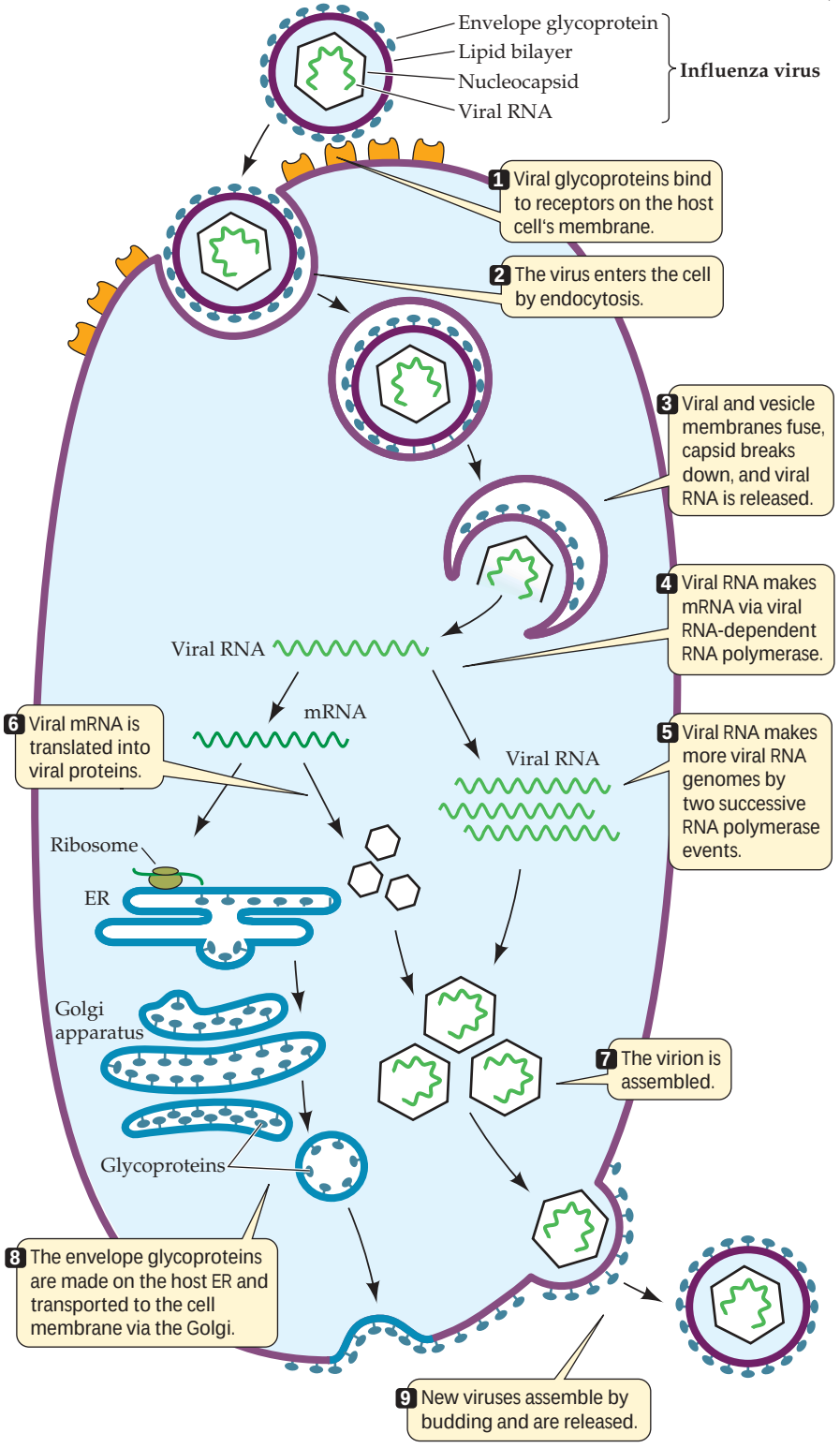
Animal viruses have diverse reproductive cycles

Almost all vertebrates are susceptible to viral infections, but among invertebrates, such infections are common only in arthropods (the group that includes insects and crustaceans). One group of viruses, called *arboviruses* (short for "arthropod-borne viruses"), is transmitted to a vertebrate through an insect bite. Although they are carried within the arthropod host's cells, arboviruses apparently do not affect that host severely; they affect only the bitten and infected organism. The arthropod acts as a **vector**—an intermediate carrier—by transmitting the disease organism from one host to another.

Animal viruses are very diverse. Some are just particles consisting of proteins surrounding a nucleic acid. Others have a membrane derived from the host cell's plasma membrane and are called *enveloped* viruses. Some animal viruses have DNA as their genetic material; others have RNA. In most cases, the viral genome is small, coding for only a few proteins.

Like that of bacteriophage, the lytic cycle of animal viruses can be divided into early and late stages (see Figure 13.3). Animal viruses enter cells in one of three ways:

- A naked virion (without a membrane) is taken up by endocytosis, which traps it within a membranous vesicle inside the host cell. The membrane of the vesicle breaks down, releasing the virion into the cytoplasm, and the host cell digests the protein capsid, liberating the viral nucleic acid, which takes charge of the host cell.
- Enveloped viruses may also be taken up by endocytosis (see Figure 13.4) and released from a vesicle. In these viruses, the viral membrane is studded with glycoproteins that bind to receptors on the host cell's plasma membrane.
- More commonly, the membranes of the host and the enveloped virus fuse, releasing the rest of the virion into the cell (see Figure 13.5).



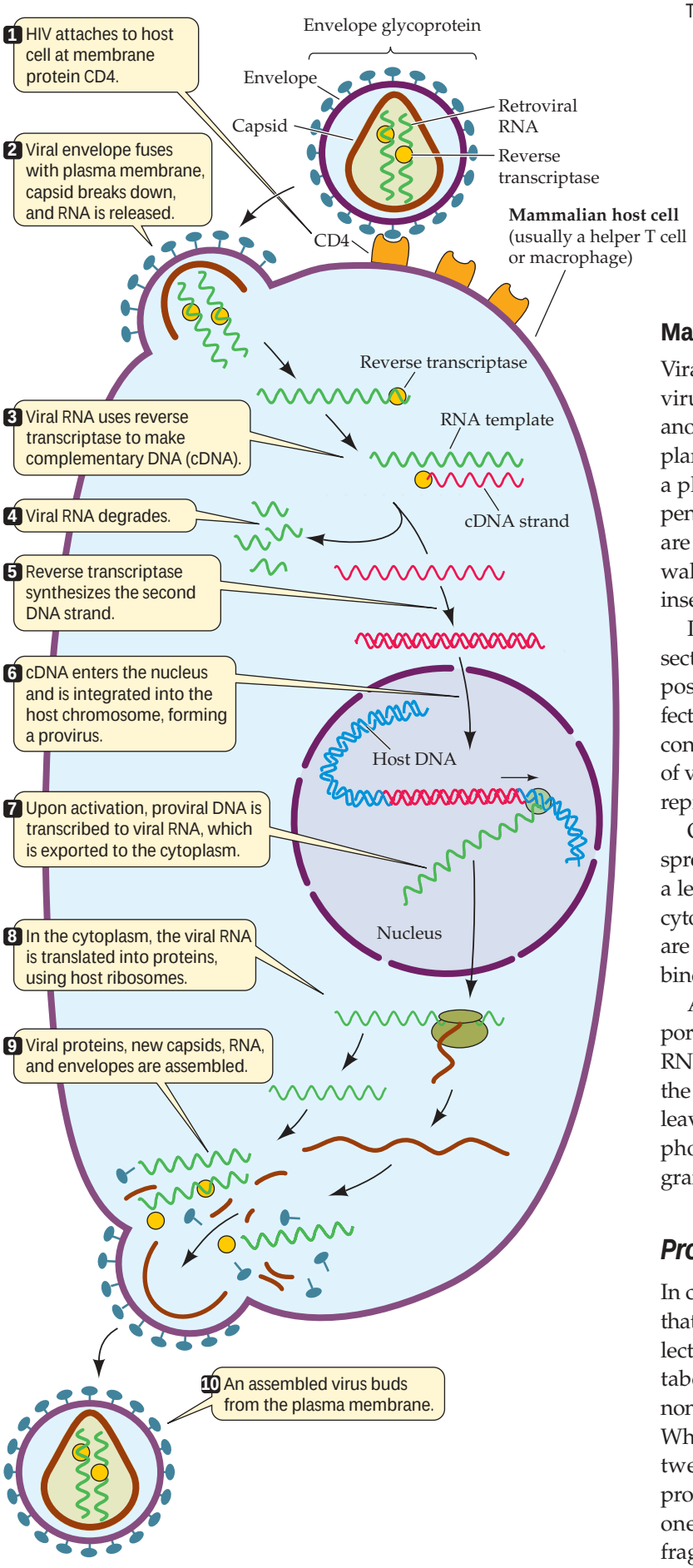
Following viral reproduction, enveloped viruses usually escape from the host cell by a budding process in which they acquire a membrane similar to that of the host cell.

The life cycles of influenza virus and HIV illustrate two different styles of infection and genome reproduction. Influenza virus is taken up into a membrane vesicle by endocytosis (Figure 13.4). Fusion of the viral and vesicle membranes releases the virion into the cell. The virus carries its own enzyme to replicate its RNA genome into a complementary strand. The new strand is then used as mRNA to make, by complementary base pairing, more copies of the viral genome.

Retroviruses such as HIV have a more complex reproductive cycle (Figure 13.5). The virus enters a host cell by direct fusion of viral and cellular membranes. A distinctive feature of the retroviral life cycle is the reverse transcription of retroviral RNA. This process produces a DNA **provirus** consisting of **cDNA** (complementary DNA transcribed from the RNA genome), which is the form of the viral genome that gets integrated into the host's DNA. The provirus may reside in the host chromosome permanently, occasionally being expressed to produce new virions. Almost every step in this complex cycle can, in principle, be attacked by therapeutic drugs; this fact is used by researchers in their quest to conquer AIDS, the deadly condition caused by HIV infection in humans, which will be discussed further in Chapter 18.

Animal viruses, including human viruses, take a severe toll on human and animal health. But our well-being is also challenged by plant viruses and the diseases they cause.

13.4 The Reproductive Cycle of the Influenza Virus
The enveloped influenza virus is taken into the host cell by endocytosis. Once inside, fusion of the vesicle and viral membranes releases the RNA genome, which replicates and assembles new virions.



13.5 The Reproductive Cycle of HIV The retrovirus HIV enters a host cell via fusion of its envelope with the host's plasma membrane. Reverse transcription of retroviral RNA then produces a DNA provirus—a strand of complementary DNA that enters the host nucleus, where it can be transcribed to viral RNA.

Many plant viruses spread with the help of vectors

Viral diseases of flowering plants are very common. Plant viruses can be transmitted *horizontally*, from one plant to another, or *vertically*, from parent to offspring. To infect a plant cell, viruses must pass through a cell wall as well as a plasma membrane. Most plant viruses accomplish this penetration through their association with vectors, which are often insects. When an insect vector penetrates a cell wall with its proboscis (snout), virions can move from the insect into the plant.

Plant viruses can be introduced artificially, without insect vectors, by bruising a leaf or other plant part, then exposing it to a suspension of virions. Horizontal viral infections may also occur in nature if a bruised infected plant contacts an injured uninfected one. Vertical transmission of viral infections may occur through vegetative or sexual reproduction.

Once inside a plant cell, the virus reproduces and spreads to other cells in the plant. Within an organ such as a leaf, the virus spreads through the plasmodesmata, the cytoplasmic connections between cells. Because the viruses are too large to go through these channels, special proteins bind to them to help them squeeze through the pores.

An example of a virus that causes an economically important plant disease is the wheat streak mosaic virus. This RNA virus enters the leaf of a wheat plant via a tiny insect, the mite *Aceria tulipae*. As the infection spreads inside the leaves, they show yellow streaks due to the destruction of photosynthetic tissues. As a result, production of wheat grain can be severely reduced.

Prokaryotes: Reproduction and Recombination

In contrast to viruses, bacteria and archaea are living cells that carry out all the basic cellular functions. They have selectively permeable membranes and perform energy metabolism. Prokaryotes usually reproduce asexually, but nonetheless have several ways of recombining their genes. Whereas in eukaryotes, genetic recombination occurs between the genomes of two parents, recombination in prokaryotes results from the interaction of the genome of one cell with a much smaller sample of genes—a DNA fragment—from another cell.

The reproduction of prokaryotes gives rise to clones

Most prokaryotes reproduce by the division of single cells into two identical offspring (see Figure 9.2). In this way, a single cell gives rise to a **clone**—a population of genetically identical individuals. Prokaryotes reproduce very rapidly. A population of *E. coli*, as we saw above, can double every 20 minutes as long as conditions remain favorable. That is one of the reasons that this bacterium is used so widely in research.

Simple, reliable methods exist for isolating single bacterial cells and rapidly growing them into clones for identification and study. Pure cultures of *E. coli* or other bacteria can be grown in liquid nutrient medium, or on the surface of a solid *minimal medium* that contains a sugar, minerals, a nitrogen source such as ammonium chloride (NH_4Cl), and a solidifying agent such as agar (Figure 13.6). If the number of cells spread on the medium is small, each cell will give rise to a small, rapidly growing *bacterial colony*. If a large number of cells is spread onto the solid medium, their growth will produce one continuous layer—a *bacterial lawn*. Bacteria can also be grown in a liquid nutrient medium. We'll see examples of all these techniques in this chapter.

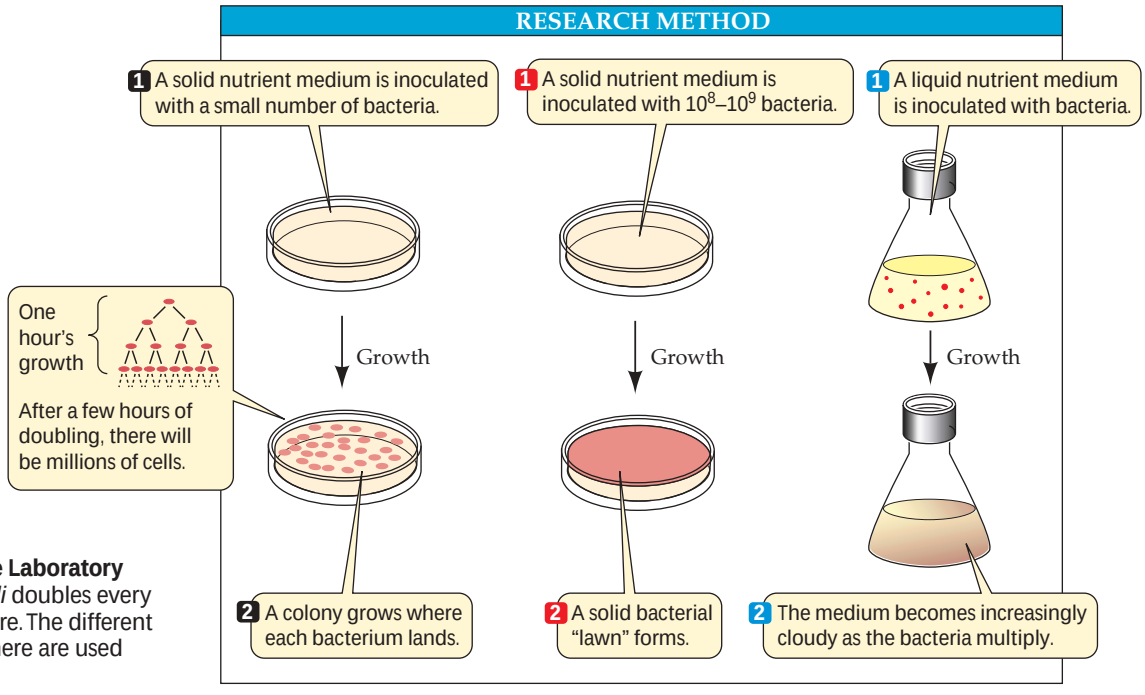
In recombination, bacteria conjugate

The existence and heritability of mutations in bacteria attracted the attention of geneticists. If there were no form of exchange of genetic information between individuals, bacteria would not be useful for genetic analysis. But can these asexually reproducing organisms exchange genetic information? Luckily, in 1946, Joshua Lederberg and Edward Tatum demonstrated that such exchanges do occur, although they are rare events.

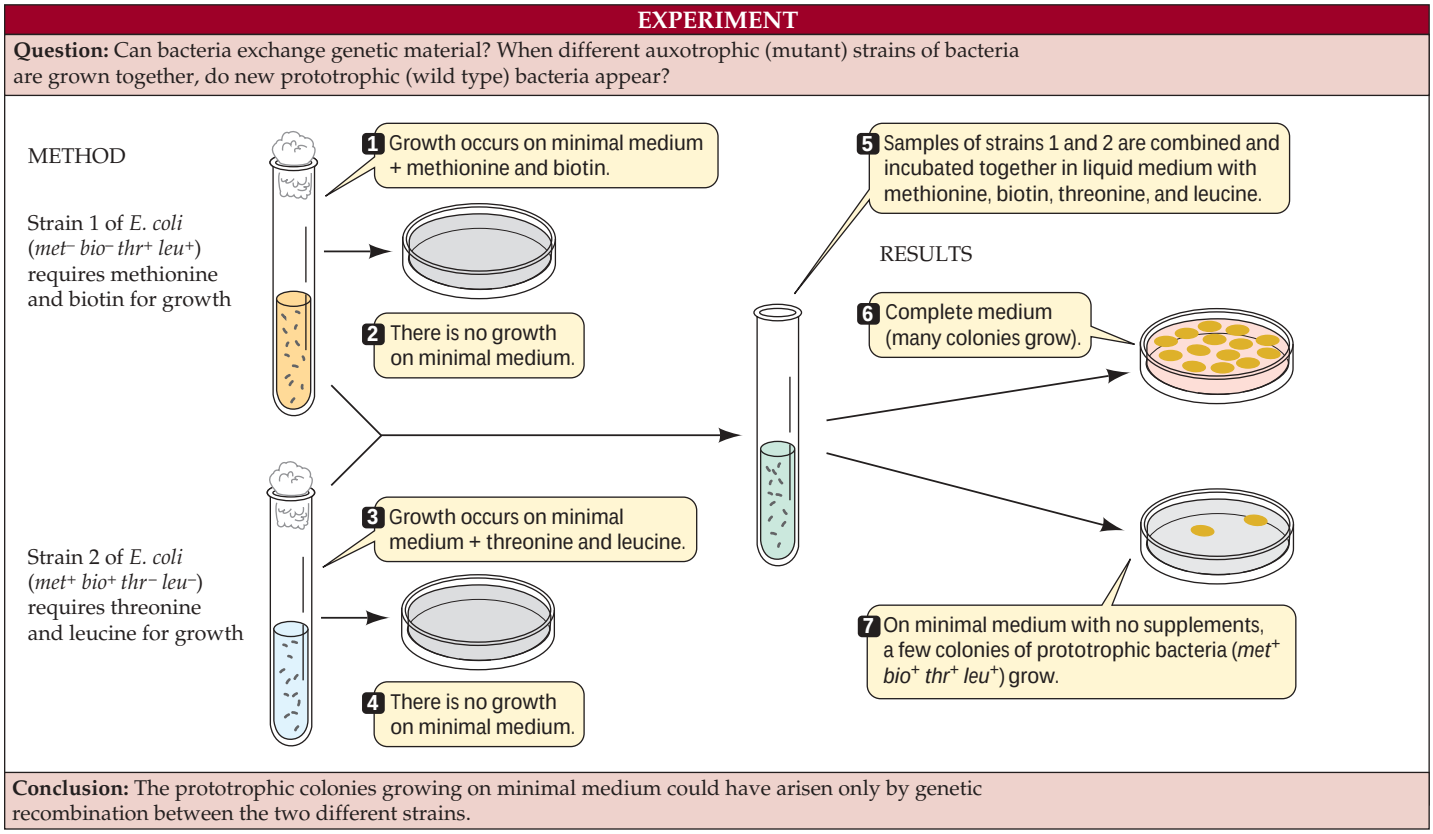
Initially, Lederberg and Tatum grew two nutrient-requiring, or *auxotrophic*, mutant strains of *E. coli*. Like the *Neurospora* studied by Beadle and Tatum (see Figure 12.1), these strains cannot grow on a minimal medium, but require supplementation with nutrients that they cannot synthesize for themselves because of an enzyme defect.

- Strain 1 requires the amino acid methionine and the vitamin biotin for growth; it can make its own threonine and leucine. So its phenotype (and genotype) is given as $met^-bio^-thr^+leu^+$.
- Strain 2 requires neither methionine nor biotin, but cannot grow without the amino acids threonine and leucine. Its phenotype is $met^+bio^+thr^-leu^-$.

Lederberg and Tatum mixed these two mutant strains and cultured them together for several hours on a medium supplemented with methionine, biotin, threonine, and leucine, so that both strains could grow. The bacteria were then removed from the medium by centrifugation, washed, and transferred to minimal medium, which lacked all four supplements. Neither strain 1 nor strain 2 could grow on this medium because of their nutritional requirements. However, a few bacterial colonies appeared on the culture plates (Figure 13.7). Because they were growing in the minimal medium, these colonies must have consisted of bacteria that were $met^+bio^+thr^+leu^+$; that is, they must have been *prototrophic*. These colonies appeared



13.6 Growing Bacteria in the Laboratory
A population of *Escherichia coli* doubles every 20 minutes in laboratory culture. The different techniques of culture shown here are used for different applications.

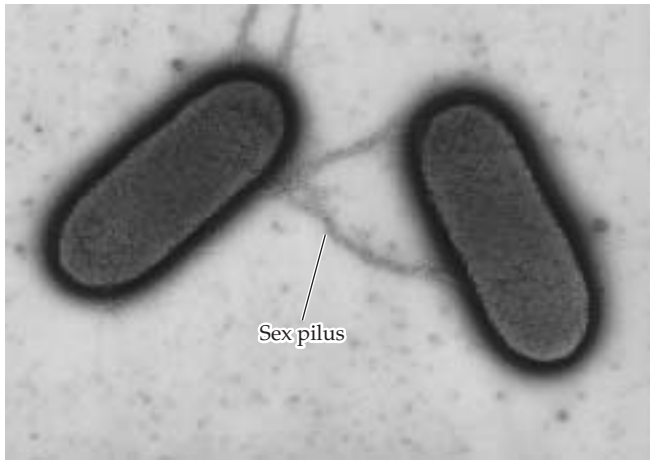


13.7 Lederberg and Tatum's Experiment After growing together, a mixture of complementary auxotrophic strains of *E. coli* contained a few cells that gave rise to new prototrophic colonies. This experiment proved that genetic recombination takes place in prokaryotes.

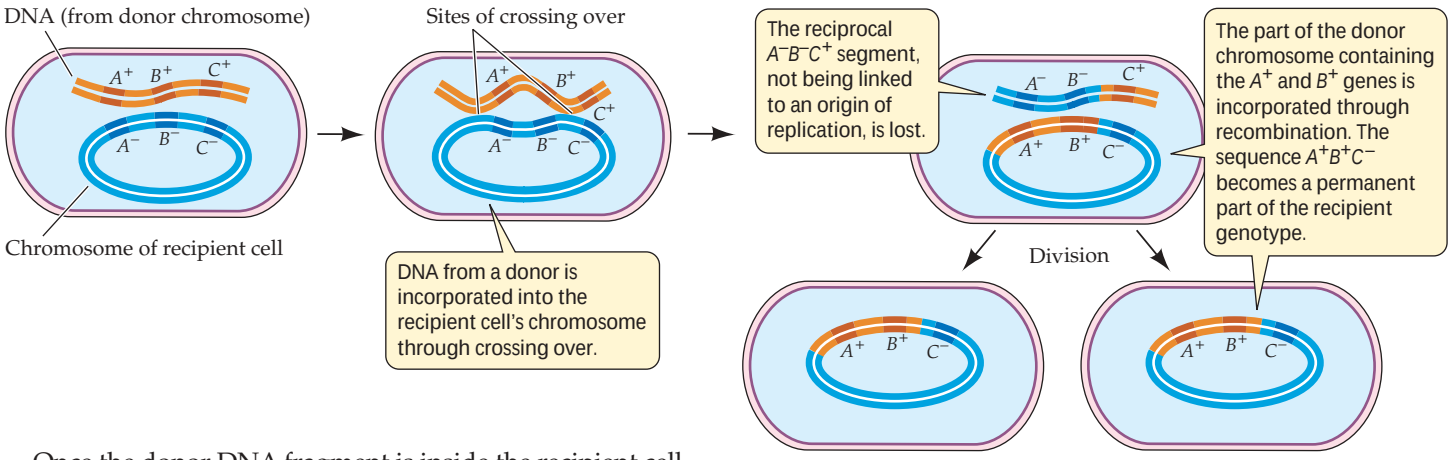
at a rate of approximately one for every 10 million cells originally placed on the plates ($1/10^7$).

Where did these prototrophic colonies come from? Lederberg and Tatum were able to rule out mutation, and other investigators ruled out transformation (a process we discussed in Chapter 11 and which we'll look at in more detail below). A third possibility is that the two strains of *E. coli* had exchanged genetic material, producing some cells containing met^+ and bio^+ alleles from strain 2 and thr^+ and leu^+ alleles from strain 1 (see Figure 13.7). Later experiments showed that such an exchange, called **conjugation**, had indeed occurred. One bacterial cell—the recipient—had received DNA from another cell—the donor—that included the two wild-type ($^+$) alleles that were missing in the recipient. Recombination had then created a genotype with four wild-type alleles.

The physical contact required for conjugation can be observed under the electron microscope (Figure 13.8). It is initiated by a thin projection called a *sex pilus*. Once the sex pili bring the two cells into proximity, the actual transfer of DNA occurs by a thin cytoplasmic bridge called a *conjugation tube*. Since the bacterial chromosome is circular, it must be made linear (cut) so that it can pass through the tube. Contact between the cells is brief—certainly not long enough for the entire donor genome to enter the recipient cell. Therefore, the recipient cell usually receives only a portion of the donor DNA.



13.8 Bacterial Conjugation Sex pili draw two bacteria into close contact, and a cytoplasmic conjugation tube forms. DNA is transferred from one cell to the other via the conjugation tube.



13.9 Recombination Following Conjugation DNA from a donor cell can become incorporated into a recipient cell's chromosome through crossing over. This recombination explains the results of the Lederberg-Tatum experiment shown in Figure 13.7.

Once the donor DNA fragment is inside the recipient cell, it can recombine with the recipient cell's genome. In much the same way that chromosomes pair up, gene for gene, in prophase I of meiosis, the donor DNA can line up beside its homologous genes in the recipient, and crossing over can occur. Enzymes that can cut and rejoin DNA molecules are active in bacteria, so gene(s) from the donor can become integrated into the genome of the recipient, thus changing the recipient's genetic constitution (Figure 13.9), even though only about half the transferred genes become integrated in this way.

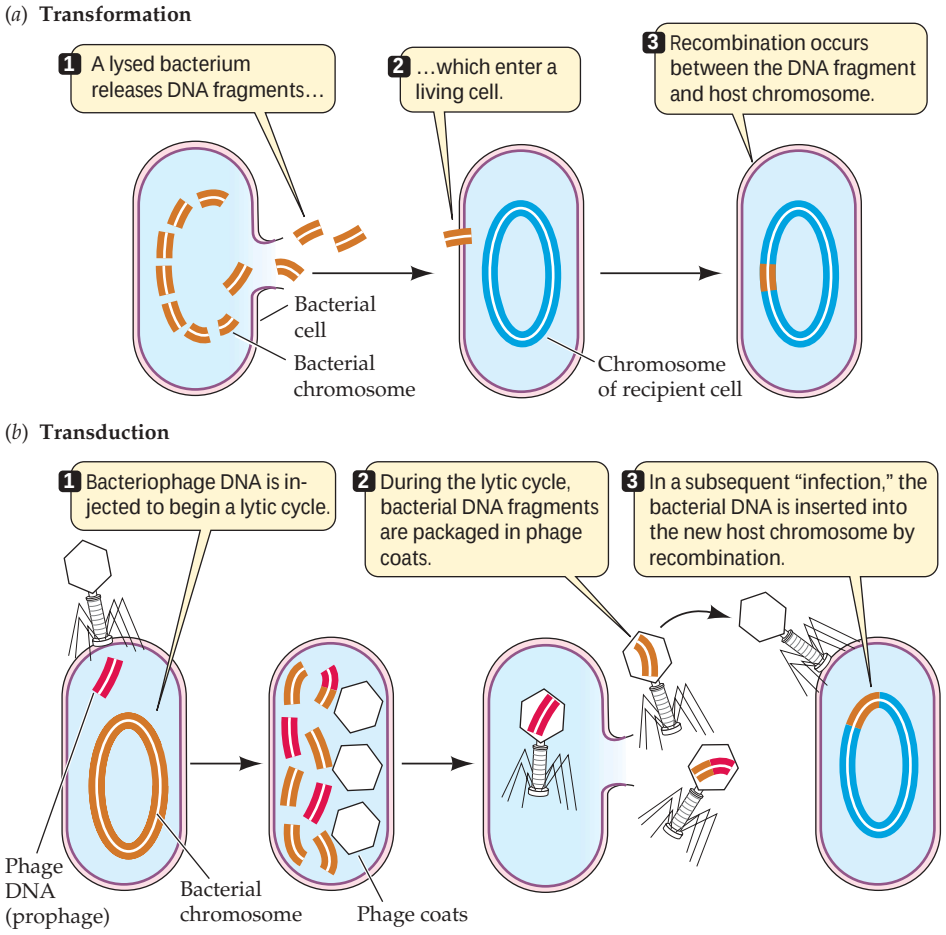
In transformation, cells pick up genes from their environment

Frederick Griffith obtained the first evidence for the transfer of prokaryotic genes more than 75 years ago when he discovered the transforming principle (see Figure 11.1). We now know the reason for his results: DNA had leaked from dead cells of virulent pneumococci and was taken up as free DNA by living nonvirulent pneumococci, which became virulent as a result. This phenomenon, called **transformation**, occurs in nature in some species of bacteria when cells die and their DNA leaks out (Figure 13.10*a*). Once transforming DNA is inside a host cell, an event very similar to recombination occurs, and new genes can be incorporated into the host chromosome.

In transduction, viruses carry genes from one cell to another

When bacteriophage undergo a lytic cycle, they package their DNA in capsids. These capsids generally form before the

13.10 Transformation and Transduction After a new DNA fragment enters the host cell, recombination can occur. (*a*) Transforming DNA can leak from dead bacterial cells and be taken up by a living bacterium, which may incorporate the new genes into its chromosome. (*b*) In transduction, viruses carry DNA fragments from one cell to another.



viral DNA is inserted into them. Sometimes, bacterial DNA fragments are inserted into the empty phage capsids instead of the phage DNA. (Figure 13.10*b*). Recall that the binding of a phage to its host cell and the insertion of phage DNA are carried out by the capsid. So, when a phage capsid carries a piece of bacterial DNA, the latter is injected into the "infected" bacterium. This mechanism of DNA transfer is called **transduction**. Needless to say, it does not result in a productive viral infection. Instead, the incoming DNA fragment can recombine with the host chromosome, resulting in the replacement of host cell genes with bacterial genes from the incoming phage particle.

Plasmids are extra chromosomes in bacteria

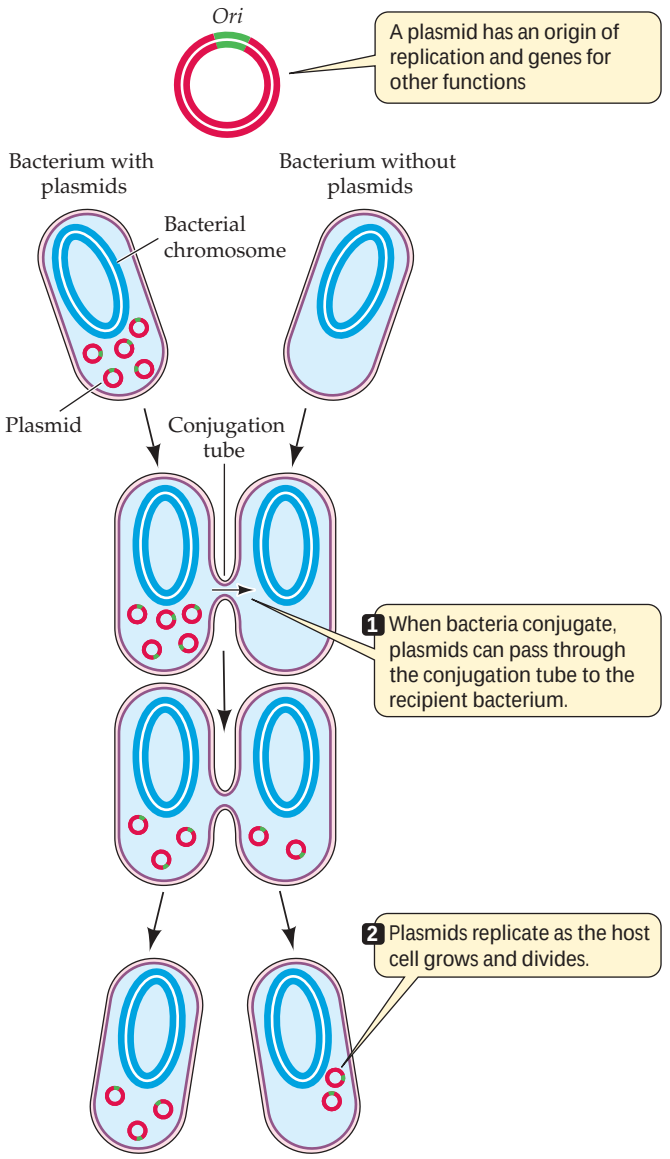
In addition to their main chromosome, many bacteria harbor additional smaller, circular chromosomes. These chromosomes, called **plasmids**, usually contain at most a few dozen genes, and, importantly, an origin of replication (the sequence where DNA replication starts), which defines them as chromosomes. Usually plasmids replicate at the same time as the main chromosome, but that is not necessarily the case.

Plasmids are *not* viruses. They do not take over the cell's molecular machinery or make a protein coat to help them move from cell to cell. Instead, they can move between cells during conjugation, thereby adding some new genes to the recipient bacterium (Figure 13.11). Because plasmids exist independently of the main chromosome (the term *episomes* is sometimes used for them), they do not need to recombine with the main chromosome to add their genes to the recipient cell's genome.

There are several types of plasmids, classified according to the kinds of genes they carry. Some code for catabolic enzymes, others enable conjugation, while others code for genes that circumvent antibiotic attack.

SOME PLASMIDS CARRY GENES FOR UNUSUAL METABOLIC FUNCTIONS. Some plasmids, called *metabolic factors*, have genes that allow their recipients to carry out unusual metabolic functions. For example, there are many unusual hydrocarbons in oil spills. Some bacteria can actually thrive on these molecules, using them as a carbon source. The genes for the enzymes involved in breaking down the hydrocarbons are carried on plasmids.

SOME PLASMIDS CARRY GENES FOR CONJUGATION. Other plasmids, called *fertility factors*, or *F factors* for short, encode the genes needed for conjugation. They have approximately 25 genes, including the ones that make both the pilus for attachment and the conjugation tube for DNA transfer. A cell harboring an F factor is referred to as F^+. It can transfer a copy of the F factor to an F^- cell, making the recipient F^+.



13.11 Gene Transfer by Plasmids When plasmids enter a cell via conjugation, their genes can be expressed in the new cell.

Sometimes the F factor integrates into the main chromosome (at which point it is no longer a plasmid), and when it does, it can bring along other genes from that chromosome when it moves through the conjugation tube from one cell to another.

SOME PLASMIDS ARE RESISTANCE FACTORS. *Resistance factors*, or *R factors*, may carry genes coding for proteins that destroy or modify antibiotics. Other R factors provide resistance to heavy metals that bacteria encounter in their environment.

R factors first came to the attention of biologists in 1957 during an epidemic of dysentery in Japan, when it was discovered that some strains of the *Shigella* bacterium, which

causes dysentery, were resistant to several antibiotics. Researchers found that resistance to the entire spectrum of antibiotics could be transferred by conjugation even when no genes on the main chromosome were transferred. Eventually it was shown that the genes for antibiotic resistance are carried on plasmids. Each R factor carries one or more genes conferring resistance to particular antibiotics, as well as genes that code for proteins involved in the transfer of DNA to a recipient bacterium. As far as biologists can determine, R factors providing resistance to naturally occurring antibiotics existed long before antibiotics were discovered and used by humans. However, R factors seem to have become more abundant in modern times, possibly because the heavy use of antibiotics in hospitals selects for bacterial strains bearing them.

Antibiotic resistance poses a serious threat to human health, and the inappropriate use of antibiotics contributes to this problem. You probably have gone to see a physician because of a sore throat, which can have either a viral or a bacterial cause. The best way to determine the causative agent is for the doctor to take a small sample from your inflamed throat, culture it, and identify any bacteria that are present. But perhaps you cannot wait another day for the results. Impatient, you ask the doctor to give you something to make you feel better. She prescribes an antibiotic, which you take. The sore throat gradually gets better, and you think that the antibiotic did the job.

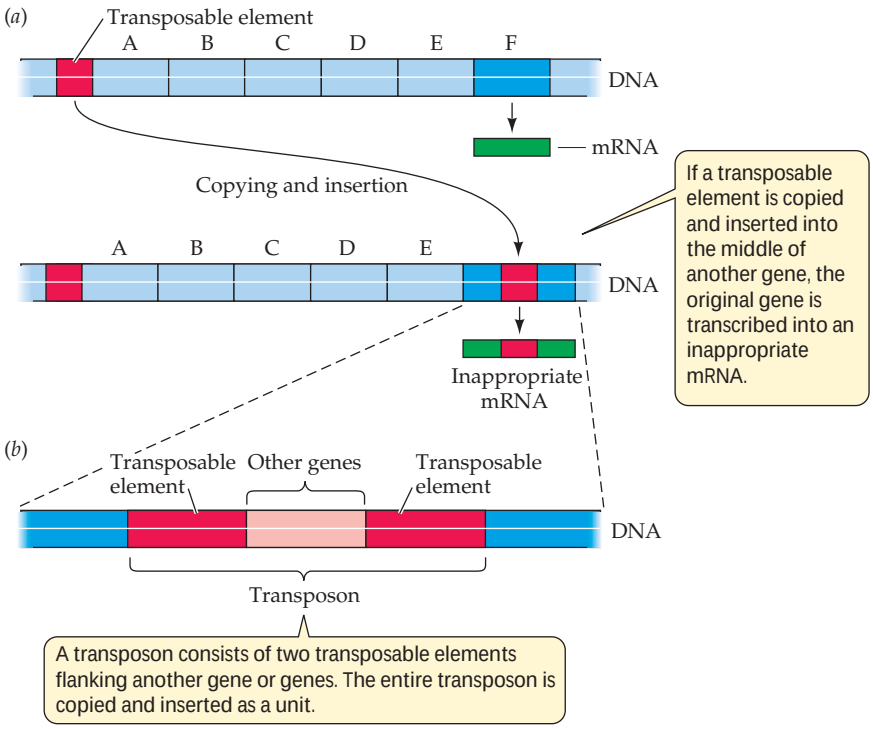
But suppose the infection is viral. In that case, the antibiotic does nothing to combat the disease, which just runs its normal course. However, it may do something harmful: By killing many normal bacteria in your body, the antibiotic may select for bacteria harboring R factors. These bacteria may survive and reproduce in the presence of the antibiotic, and may soon become quite numerous. The next time you get a bacterial infection, there may be a ready supply of resistant bacteria in your body, and antibiotics may be ineffective.

Antibiotic resistance in pathogenic bacteria provides an example of evolution in action. In the years after they were first discovered in the twentieth century, antibiotics were very successful in combating diseases that had plagued humans for millennia, such as cholera, tuberculosis, and leprosy. But as time went on, resistant bacteria appeared. This was, and is, classic natural selection: Genetic variation existed among bacteria, and those that survived the onslaught of antibiotics must have had a genetic constitution that allowed them to do so.

Transposable elements move genes among plasmids and chromosomes

As we have seen, plasmids, viruses, and even phage capsids (in the case of transduction) can transport genes from one bacterial cell to another. There is another type of "gene transport" that occurs within the individual cell. It relies on segments of DNA that can be inserted either at a new location on the same chromosome or into another chromosome. These DNA sequences are called **transposable elements**. Their insertion often produces phenotypic effects by disrupting the genes into which they are inserted (Figure 13.12*a*).

The first transposable elements to be discovered in prokaryotes were large pieces of DNA, typically 1,000 to 2,000 base pairs long, found at many sites on the *E. coli* main chromosome. In one mechanism of transposition, the transposable element replicates independently of the rest of the chromosome. The copy then inserts itself at other, seemingly random sites on the chromosome. The genes encoding the enzymes necessary for this insertion are found within the transposable element itself. Other transposable elements are cut from their original sites and inserted elsewhere without replication. Later, many longer transposable elements were discovered (about 5,000 base pairs). These large elements carry one or more additional genes and are called **transposons** (Figure 13.12*b*).



13.12 Transposable Elements and Transposons (*a*) Transposable elements are segments of DNA that can be inserted at new locations, either on the same chromosome or on a different chromosome. (*b*) Transposons consist of transposable elements combined with other genes.

What do transposons and other transposable elements have to do with the genetics of prokaryotes—or with hospitals? Transposable elements have contributed to the evolution of plasmids. R factors probably originally gained their genes for antibiotic resistance through the activity of transposable elements. One piece of evidence for this conclusion is that each resistance gene in an R factor is part of a transposon.

In summary, rapid asexual reproduction can produce enormous clones of prokaryotes. However, these genetically identical cells are all equally vulnerable to some change in the environment. Recombination by means of conjugation, transformation, and transduction, or the acquisition of new genes by means of plasmids and transposable elements, all introduce genetic diversity into bacterial populations, and this diversity allows at least some cells to survive under changing conditions. Prokaryotes can also respond to changes in their environment by regulating the expression of their genes.

Regulation of Gene Expression in Prokaryotes

Prokaryotes can conserve energy and resources by making proteins only when they are needed. The protein content of a bacterium can change rapidly when conditions warrant. There are several ways in which a prokaryotic cell could shut off the supply of an unneeded protein:

- Block the transcription of mRNA for that protein
- Hydrolyze the mRNA after it is made and prior to translation
- Prevent translation of the mRNA at the ribosome
- Hydrolyze the protein after it is made
- Inhibit the function of the protein

These methods would all have to be selective, affecting some genes and proteins and not others. In addition, they would all have to respond to some biochemical signal. Clearly, the earlier the cell intervenes in the process, the less energy it has to expend. Selective inhibition of transcription is far more efficient than transcribing the gene, translating the message, and then degrading or inhibiting the protein. While examples of all five mechanisms for regulating protein levels are found in nature, prokaryotes generally use the most efficient one, transcriptional regulation.

Regulation of transcription conserves energy

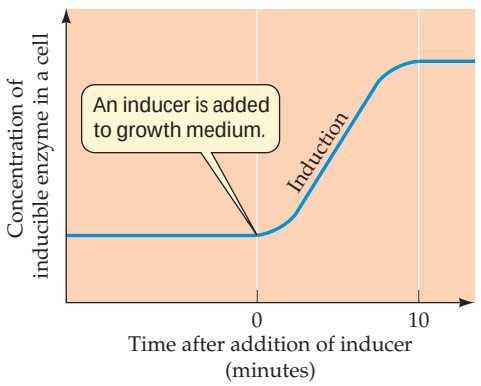
As a normal inhabitant of the human intestine, *E. coli* must be able to adjust to sudden changes in its chemical environment. Its host may present it with one foodstuff one hour and another the next. This variation presents the bacterium with a metabolic challenge. Glucose is its preferred energy source, and is the easiest sugar to metabolize, but not all of its host's foods contain an abundant supply of glucose. For example, the bacterium may suddenly be deluged with milk, whose predominant sugar is lactose. Lactose is a β-galactoside—a disaccharide containing galactose β-linked to glucose (see Chapter 3). To be taken up and metabolized by *E. coli*, lactose is acted on by three proteins:

- *β-galactoside permease* is a carrier protein in the bacterial plasma membrane that moves the sugar into the cells.
- *β-galactosidase* is an enzyme that catalyzes the hydrolysis of lactose to glucose and galactose.
- A third protein, the enzyme *β-galactoside transacetylase*, is also required for lactose metabolism, although its role in the process is not yet clear.

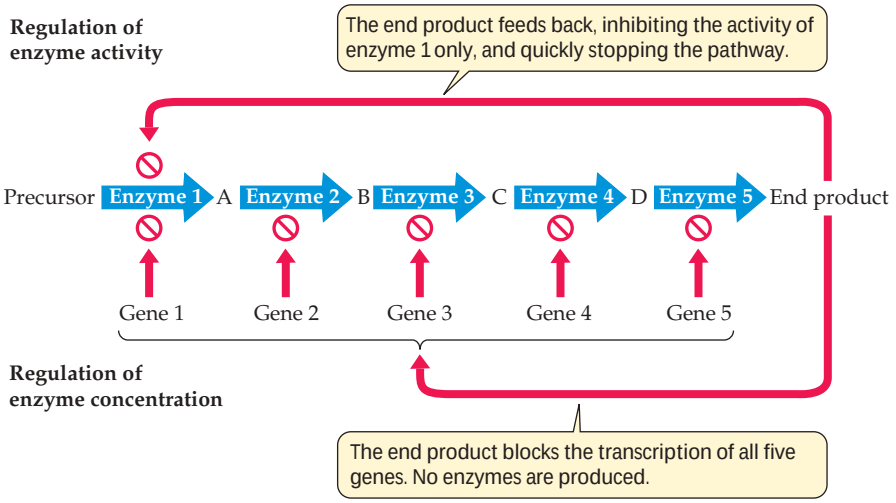
When *E. coli* is grown on a medium that does not contain lactose or other β-galactosides, the levels of these three proteins are extremely low—the cell does not waste energy and materials making the unneeded enzymes. If, however, the environment changes such that lactose is the predominant sugar available and very little glucose is present, the bacterium promptly begins making all three enzymes, and they increase rapidly in abundance. For example, there are only two molecules of β-galactosidase present in an *E. coli* cell when glucose is present in the medium. But when glucose is absent, lactose can induce the synthesis of 3,000 molecules of β-galactosidase per cell!

If lactose is removed from *E. coli*'s environment, synthesis of the three enzymes that process it stops almost immediately. The enzyme molecules that have already formed do not disappear; they are merely diluted during subsequent cell divisions until their concentration falls to the original low level within each bacterium.

Compounds that stimulate the synthesis of an enzyme (such as lactose in our example) are called **inducers** (Figure 13.13). The enzymes that are produced are called **inducible**



13.13 An Inducer Stimulates the Synthesis of an Enzyme It is most efficient for a cell to produce an enzyme only when it is needed. Some enzymes are induced by the presence of the substance they act upon (for example, β-galactosidase is induced by the presence of lactose).



13.14 Two Ways to Regulate a Metabolic Pathway Feedback from the end product can block enzyme activity, or it can stop the transcription of genes that code for the enzymes.

enzymes, whereas enzymes that are made all the time at a constant rate are called **constitutive** enzymes.

We have now seen two basic ways of regulating the rate of a metabolic pathway. Chapter 6 described allosteric regulation of enzyme activity (the rate of enzyme-catalyzed reactions); this mechanism allows rapid fine-tuning of metabolism. Regulation of protein synthesis—that is, regulation of the concentration of enzymes—is slower, but produces a greater savings of energy. Figure 13.14 compares these two modes of regulation.

A single promoter controls the transcription of adjacent genes

The genes that serve as blueprints for the synthesis of the three enzymes that process lactose in *E. coli* are called **structural genes**, indicating that they specify the primary structure (the amino acid sequence) of a protein molecule. In other words, structural genes are those genes that can be transcribed into mRNA.

The three structural genes involved in the metabolism of lactose lie adjacent to one another on the *E. coli* chromosome. This arrangement is no coincidence: their DNA is transcribed into a single, continuous molecule of mRNA. Because this particular messenger governs the synthesis of all three lactose-metabolizing enzymes, either all or none of the enzymes are made, depending on whether their common message—their mRNA—is present in the cell.

The three genes share a single promoter. Recall from Chapter 12 that a *promoter* is a DNA sequence to which RNA polymerase binds to initiate transcription. The promoter for these three structural genes can be very effective, so the maximum rate of mRNA synthesis can be high. However, there is also a mechanism to shut down mRNA synthesis when the enzymes are not needed. That mechanism is the operon, elegantly worked out by François Jacob and Jacques Monod.

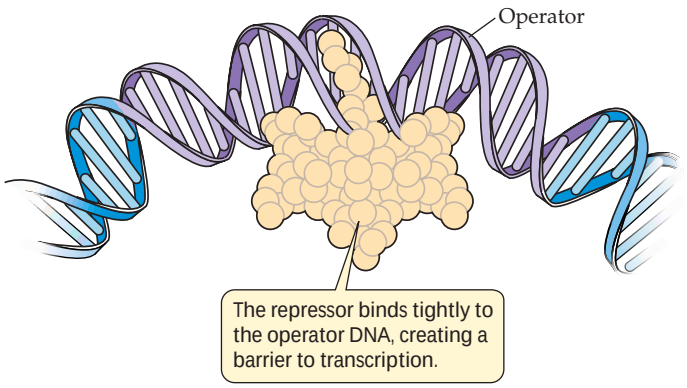
Operons are units of transcription in prokaryotes

Prokaryotes shut down transcription by placing an obstacle between the promoter and the structural genes it regulates. A short stretch of DNA called the **operator** lies in this position. It can bind very tightly to a special type of protein molecule, called a **repressor**, to create such an obstacle.

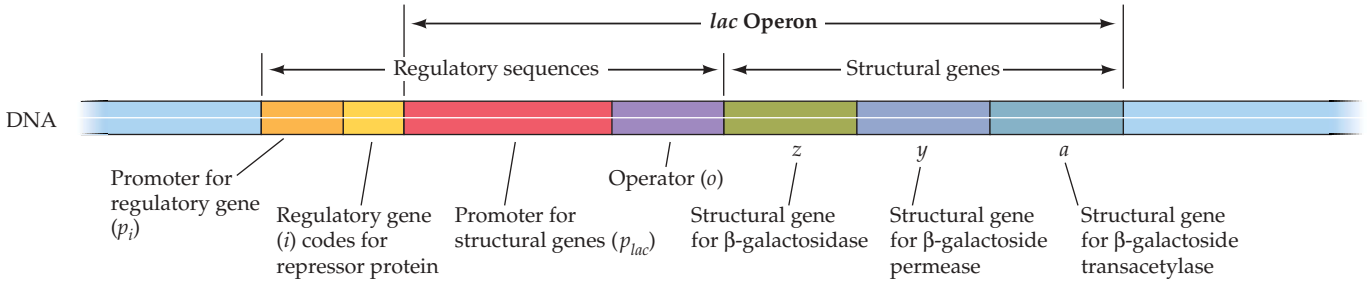
- When the repressor protein is bound to the operator, it blocks the transcription of mRNA (Figure 13.15).
- When the repressor is not attached to the operator, mRNA synthesis proceeds rapidly.

The whole unit, consisting of the closely linked structural genes and the DNA sequences that control their transcription, is called an **operon**. An operon always consists of a promoter, an operator, and two or more structural genes (see Figure 13.16). The promoter and operator are binding sites on DNA and are not transcribed.

E. coli has numerous mechanisms to control the transcription of operons; we will focus on three of them here. Two of these control mechanisms depend on interactions of a repressor protein with the operator, and the third depends on interactions of other proteins with the promoter.



13.15 A Repressor Blocks Transcription An untranscribed DNA sequence called the operator (purple) can control the transcription of a structural gene. When a repressor protein binds to the operator, transcription of the structural gene is blocked.



13.16 The *lac* Operon of *E. coli* The *lac* operon of *E. coli* is a segment of DNA that includes a promoter, an operator, and the three structural genes that code for lactose-metabolizing enzymes.

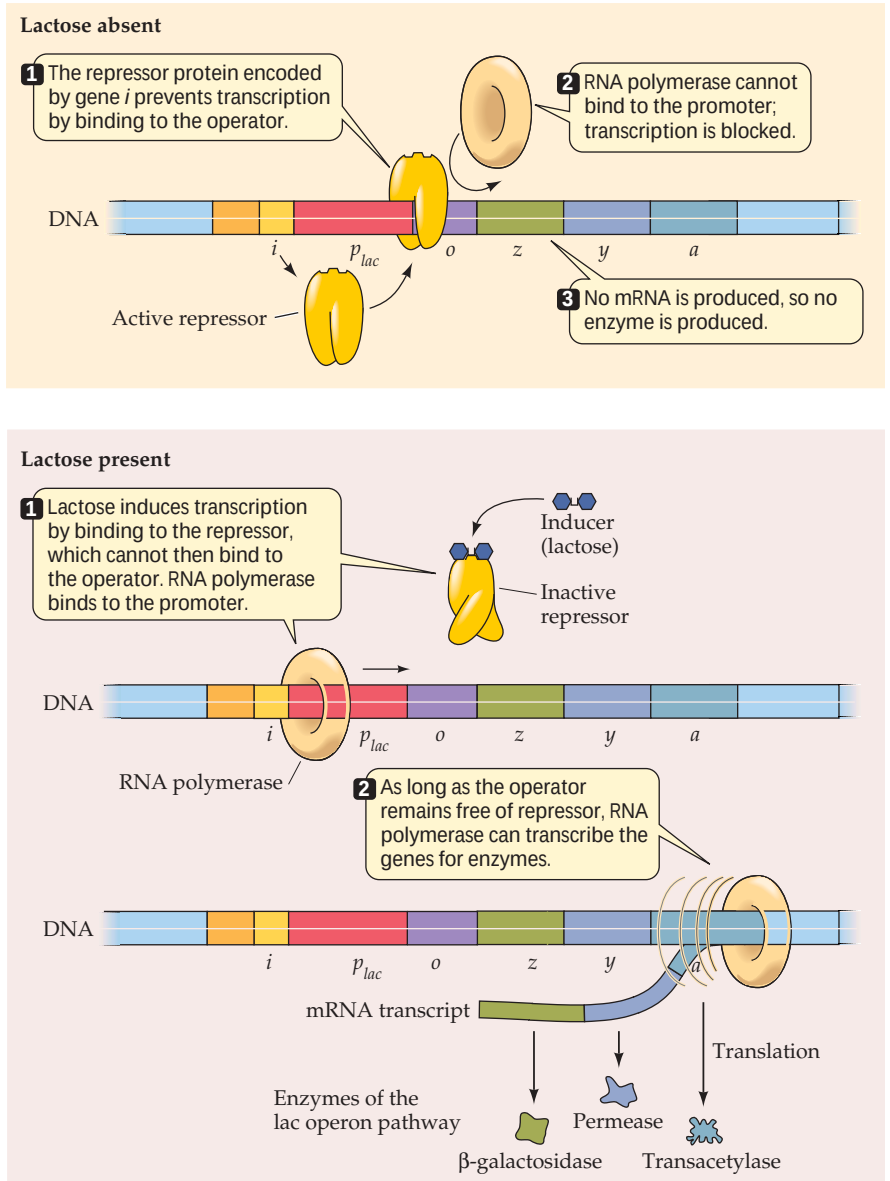
Operator–repressor control that induces transcription: The *lac* operon

The operon containing the genes for the three lactose-metabolizing proteins of *E. coli* is called the *lac operon* (Figure 13.16). As we have just seen, RNA polymerase can bind to the promoter, and a repressor protein can bind to the operator.

The repressor protein has two binding sites: one for the operator and the other for inducers. The inducers of the *lac* operon, as we know, are molecules of lactose and certain other β-galactosides. Binding to an inducer changes the shape of the repressor (by allosteric modification; see Chapter 6). This change in shape prevents the repressor from binding to the operator (Figure 13.17). As a result, RNA polymerase can bind to the promoter and start transcribing the structural genes of the *lac* operon. The mRNA transcribed from these genes is translated on ribosomes to synthesize the three proteins required for metabolizing lactose.

What happens if the concentration of lactose drops? As the lactose concentration decreases, the inducer (lactose) molecules separate from the repressor. Free of lactose molecules, the repressor returns to its original shape and binds to the operator, and transcription of the *lac* operon stops. Translation stops soon thereafter because the mRNA that is already present breaks down quickly. Thus, it is the presence or absence of lactose—the inducer—that regulates the binding of the repressor to the operator, and therefore the synthesis of the proteins needed to metabolize it.

Repressor proteins are encoded by **regulatory genes**. The regulatory gene that codes for the repressor of the *lac* operon is called the *i* (*inducibility*) *gene*. The *i* gene happens to



13.17 The *lac* Operon: An Inducible System Lactose (the inducer) leads to enzyme synthesis by preventing the repressor protein (which would have stopped transcription) from binding to the operator.

lie close to the operon that it regulates, but some other regulatory genes are distant from their operons. Like all other genes, the *i* gene itself has a promoter, which can be designated p_i. Because this promoter does not bind RNA polymerase very effectively, only enough mRNA to synthesize about ten molecules of repressor protein per cell per generation is produced. This quantity of the repressor is enough to regulate the operon effectively—to produce more would be a waste of energy. There is no operator between p_i and the *i* gene. Therefore, the repressor of the *lac* operon is a constitutive protein; that is, it is made at a constant rate that is not subject to environmental control.

Let's review the important features of inducible systems such as the *lac* operon:

- In the absence of inducer, the operon is turned off.
- Control is exerted by a regulatory protein—the repressor—that turns the operon off.
- Regulatory genes produce proteins whose sole function is to regulate the expression of other genes.
- Certain other DNA sequences (operators and promoters) do not code for proteins, but are binding sites for regulatory or other proteins.
- Adding inducer turns the operon on.

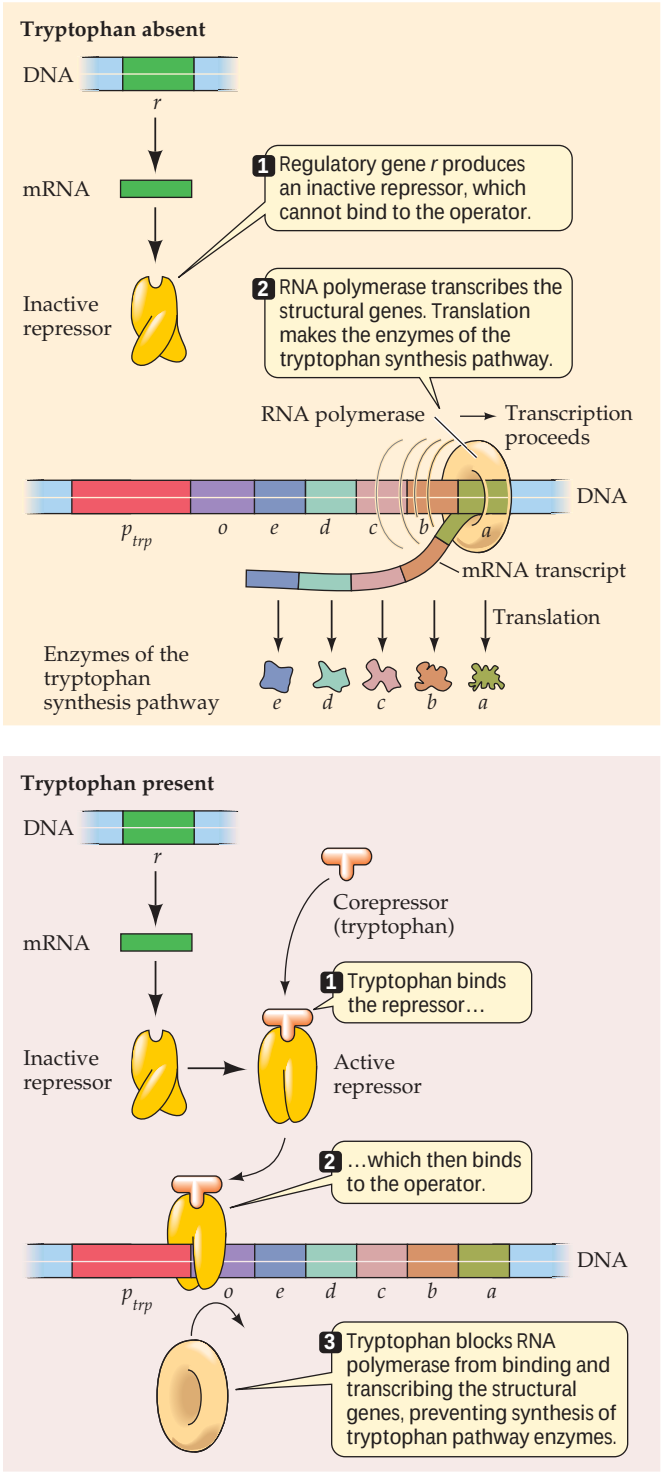
Operator–repressor control that represses transcription: The *trp* operon

We have seen that *E. coli* benefits from having an inducible system for lactose metabolism. Only when lactose is present does the system switch on. Equally valuable to a bacterium is the ability to switch off the synthesis of certain enzymes in response to the excessive accumulation of their end products. For example, if the amino acid tryptophan, an essential constituent of proteins, is present in ample concentration, it is advantageous to stop making the enzymes for tryptophan synthesis. When the synthesis of an enzyme can be turned off in response to such a biochemical cue, the enzyme is said to be **repressible**.

In repressible systems, the repressor protein cannot shut off its operon unless it first binds to a **corepressor**, which may be either the metabolic end product itself (tryptophan in this case) or an analog of it (Figure 13.18). If the end product is absent, the repressor protein cannot bind to the operator, and the operon is transcribed at a maximum rate. If the end product is present, the repressor binds to the operator, and the operon is turned off.

The difference between inducible and repressible systems is small, but significant:

- In inducible systems, the substrate of a metabolic pathway (the inducer) interacts with a regulatory protein (the repressor) to render it incapable of binding to the operator, thus allowing transcription.
- In repressible systems, the product of a metabolic pathway (the corepressor) interacts with a regulatory protein to make it capable of binding to the operator, thus blocking transcription.



13.18 The *trp* Operon: A Repressible System Because tryptophan activates an otherwise inactive repressor, it is called a corepressor.

In general, inducible systems control catabolic pathways (which are turned on only when the substrate is available), whereas repressible systems control biosynthetic pathways (which are turned off until the product becomes unavailable). In both kinds of systems, the regulatory molecule functions by binding to the operator. Next, we will consider an example of control by binding to the promoter.

Protein synthesis can be controlled by increasing promoter efficiency

Suppose an *E. coli* cell lacks a supply of glucose, its preferred energy source, but instead has access to another sugar (such as lactose) that it can break down to obtain energy. Operons encoding enzymes that catabolize such alternative energy sources, such as the *lac* operon, have a mechanism for increasing the transcription of these enzymes by increasing the efficiency of the promoter. In these operons, the promoter binds RNA polymerase in a series of steps (Figure 13.19). First, a protein called CRP (short for *c*AMP *r*eceptor *p*rotein) binds the low-molecular-weight compound adenosine 3′,5′-cyclic monophosphate, better known as cyclic AMP, or cAMP. Next, the CRP–cAMP complex binds to DNA just upstream (5′) of the promoter. This binding results in more efficient binding of RNA polymerase to the promoter, and thus an elevated level of transcription of the structural genes.

When glucose becomes abundant in the medium, the bacterium does not need to break down alternative food molecules, so synthesis of the enzymes that catabolize these molecules diminishes or ceases. The presence of glucose decreases the synthesis of the enzymes by lowering the cellular concentration of cAMP. The lower cAMP concentration leads to less CRP binding to the promoter, less efficient binding of RNA polymerase, and reduced transcription of the structural genes. This mechanism is called **catabolite repression**.

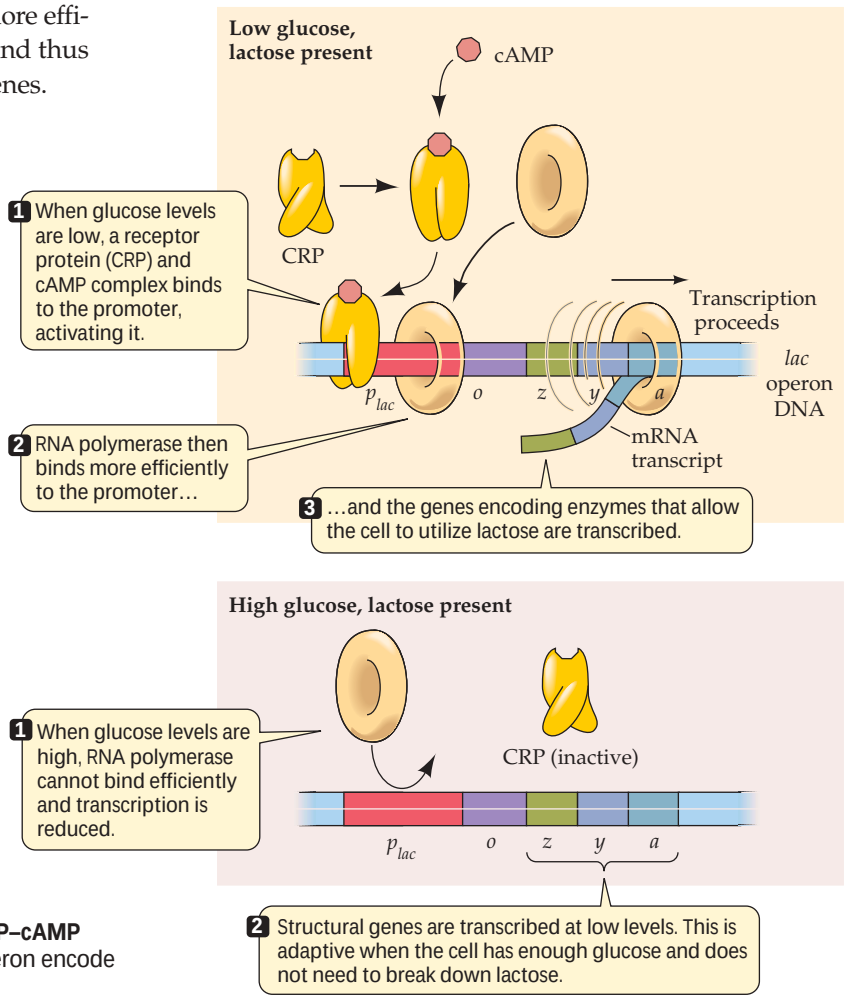
As you will see in later chapters of this book, cAMP is a widely used signaling molecule in eukaryotes, as well as in prokaryotes. The use of this nucleotide in such widely diverse situations as a bacterium sensing glucose levels and a human sensing hunger demonstrates the prevalence of common themes in biochemistry and natural selection.

The inducible *lac* and repressible *trp* systems—the two operator–repressor systems—are examples of **negative control** of transcription because the regulatory molecule (the repressor) in each case prevents transcription. The promoter–catabolite repression system is an example of **positive control** of transcription because the regulatory molecule (the CRP–cAMP complex) enhances transcription. The relationships between these positive and negative control systems are summarized in Table 13.2.

The control of gene expression by regulatory proteins is not unique to prokaryotes. As we will see in the next chapter, it also occurs in eukaryotes and even, as we are about to see, in viruses.

Control of Transcription in Viruses

The mechanisms used by used by viruses within a host cell for the regulation of gene expression are similar to those used by prokaryotes. Even a "simple" biological agent such as a virus is faced with complicated molecular decisions when its genome enters a cell. For example, the viral genome must di-



13.19 Transcription Is Enhanced by the Binding of the CRP–cAMP Complex to the Promoter The structural genes of this operon encode enzymes that break down a food source other than glucose.

13.1 Positive and Negative Controls in the lac Operon[a]

GLUCOSE	cAMP LEVELS	RNA POLYMERASE BINDING TO PROMOTER	LACTOSE	*LAC* REPRESSOR	TRANSCRIPTION OF *LAC* GENES?	LACTOSE USED BY CELLS?
Present	Low	Absent	Absent	Active and bound to operator	No	No
Present	Low	Present, not efficient	Present	Inactive and not bound to operator	Low level	No
Absent	High	Present, very efficient	Present	Inactive and not bound to operator	High level	Yes
Absent	High	Absent	Absent	Active and bound to operator	No	No

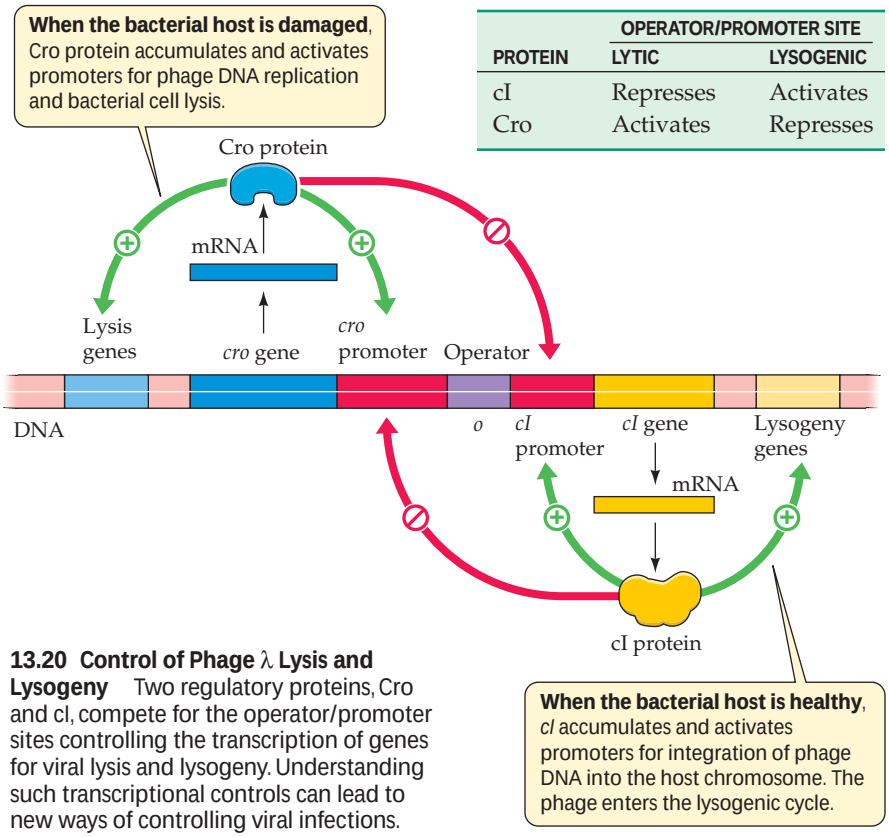
[a]Negative controls are in red type.

rect the shutdown of host transcription and translation, then redirect the host's protein synthesis machinery to virus production and host cell lysis. All the genes involved in this process must be activated in the right order. In temperate viruses, which can insert their genome (or a DNA copy) into the host chromosome, an additional issue arises: When should the provirus leave the host chromosome and undergo a lytic cycle?

Bacteriophage λ (lambda) is a temperate phage, meaning that it can undergo either a lytic or a lysogenic cycle (see Figure 13.2). When there is a rich medium available and its host bacterium is growing rapidly, the prophage takes advantage of its favorable cellular environment and remains lysogenic. When the host bacteria are not as healthy, the prophage senses this and, as a survival mechanism, leaves the host chromosome and becomes lytic.

The phage makes this decision by means of a "genetic switch": Two regulatory viral proteins, labeled cI and Cro, compete for two operator/promoter sites on phage DNA. The two operator/promoter sites control the transcription of the viral genes involved in the lytic and the lysogenic cycles, respectively, and the two regulatory proteins have opposite effects on the two operators (Figure 13.20). Phage infection is essentially a "race" between these two regulatory proteins. In a healthy *E. coli* host cell, Cro synthesis is low, so cI "wins" and the phage enters a lysogenic cycle. If the host cell is damaged by mutagens or other stress, Cro synthesis is high, promoters for phage DNA and viral coat proteins are activated, and bacterial lysis ensues. The two regulatory proteins are made very early in phage infection, and each has a binding site for a specific DNA sequence.

The life cycle of phage λ, which has been greatly simplified here, is a paradigm for viral infections throughout the biological world. The lessons learned from transcriptional controls in this system have been applied again and again to other viruses, including HIV. The control of gene activity in eukaryotic cells is somewhat different, as we will see in the next chapter, but nevertheless usually involves regulatory protein–DNA interactions.



13.20 Control of Phage λ Lysis and Lysogeny Two regulatory proteins, Cro and cI, compete for the operator/promoter sites controlling the transcription of genes for viral lysis and lysogeny. Understanding such transcriptional controls can lead to new ways of controlling viral infections.

Prokaryotic Genomes

When DNA sequencing first became possible in the late 1970s, the first biological agents to be sequenced were the simplest viruses. Soon, over 150 viral genomes, including those of important animal and plant pathogens, had been sequenced. Information on how these virus infect their hosts and reproduce came quickly as a result.

But the manual sequencing techniques used on viruses were not up to the task of elucidating the genomes of prokaryotes and eukaryotes, the smallest of which are a hundred times larger than those of a bacteriophage. In the past decade, however, the automated sequencing techniques described in Chapter 11 have rapidly added many prokaryotic sequences to biologists' store of knowledge.

In 1995, a team led by Craig Venter and Hamilton Smith determined the first sequence of a free-living organism, the bacterium *Haemophilus influenzae*. Many more prokaryotic sequences have followed. These sequences have revealed not only how prokaryotes apportion their genes to perform different cellular roles, but also how their specialized functions are carried out. A beginning has even been made on the provocative question of what the minimal requirements for a living cell might be.

Three types of information can be obtained from a genomic sequence:

- *Open reading frames*, which are the coding regions of genes. For protein-coding genes, these regions can be recognized by the start and stop codons for translation.
- *Amino acid sequences of proteins*. These sequences can be deduced from the DNA sequences of open reading frames by applying the genetic code.
- *Gene control sequences*, such as promoters and terminators for transcription.

Functional genomics relates gene sequences to functions

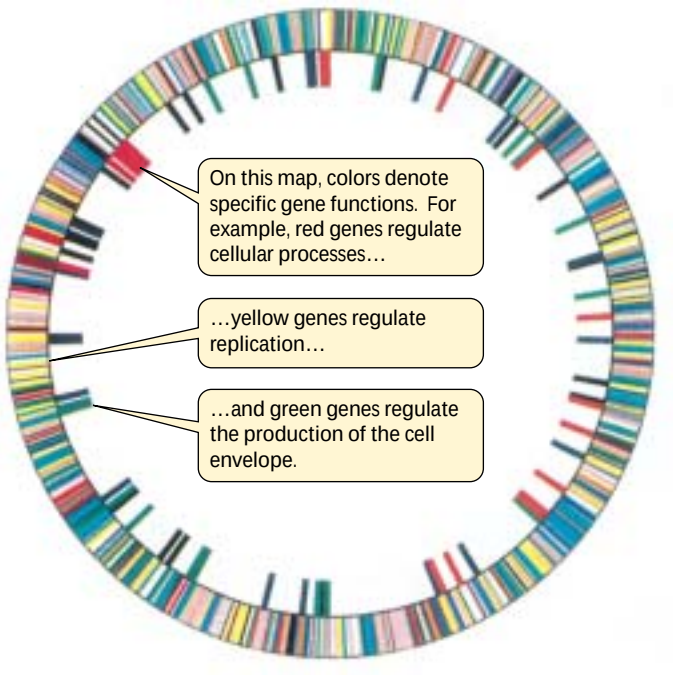
Functional genomics is the assignment of roles to the products of genes described by genomic sequencing. This field, less than a decade old, is now a major occupation of biologists.

The only host for the bacterium *H. influenzae* is humans. It lives in the upper respiratory tract and can cause ear infections or, more seriously, meningitis in children. Its single circular chromosome has 1,830,137 base pairs (Figure 13.21). In addition to its origin of replication and the genes coding for rRNAs and tRNAs, this bacterial chromosome has 1,743 regions containing amino acid codons as well as the transcriptional (promoter) and translational (start and stop codons) information needed for protein synthesis—that is, regions that are likely to be genes that code for proteins.

When this sequence was first announced, only 1,007 (58%) of the bacterium's genes had amino acid sequences that corresponded to proteins with known functions—in other words, only 58% were genes that the researchers, based on their knowledge of the functions of bacteria, expected to find. The remaining 42% of its genes coded for proteins that were unknown to researchers. The roles of most of the unknown proteins have been identified since that time, a process known as *annotation.*

Of the genes and proteins with known roles, most confirmed a century of biochemical description of bacterial enzymatic pathways. For example, genes for enzymes making up entire pathways of glycolysis, fermentation, and electron transport were found. Some of the remaining gene sequences for unknown proteins may code for membrane proteins, including those involved in active transport. Another important finding was that highly infective strains of *H. influenzae* have genes coding for surface proteins that attach the bacterium to the human respiratory tract, while noninfective strains lack those genes.

Soon after the sequence of *H. influenzae* was announced, smaller (*Mycoplasma genitalium*, 580,070 base pairs) and larger (*E. coli*, 4,639,221 base pairs) prokaryotic sequences were completed. Thus began a new era in biology, the era of **comparative genomics**, in which the genome sequences of different organisms are compared to see what genes one organism has or is missing, in order to relate the results to physiology.



13.21 Functional Organization of the Genome of *H. influenzae*
The entire DNA sequence has 1,830,137 base pairs.

M. genitalium, for example, lacks the enzymes needed to synthesize amino acids, which the other two prokaryotes possess. This finding reveals that *M. genitalium* is a parasite, which must obtain all its amino acids from its environment, the human urogenital tract. *E. coli* has 55 regulatory genes coding for transcriptional activators and 58 for repressors; *M. genitalium* has only 3 genes for activators. Comparisons such as these have led to the formulation of specific questions about how an organism lives the way it does. We'll see many more applications of comparative genomics in the next chapter.

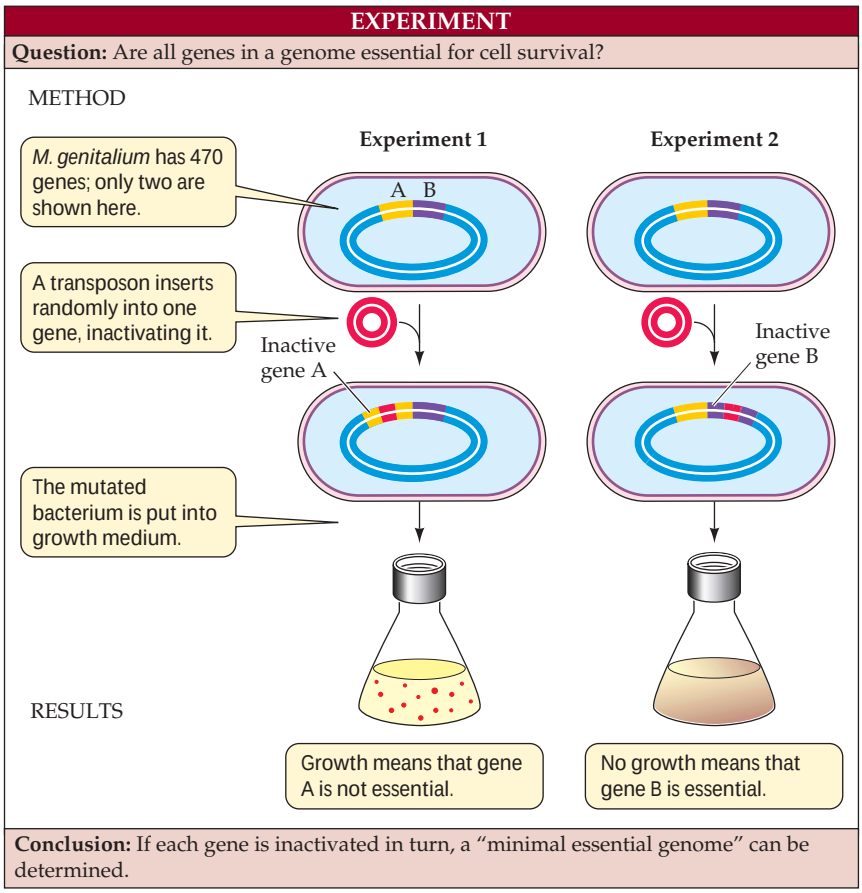
The sequencing of prokaryotic genomes has medical applications

Prokaryotic genome sequencing has important ramifications for the study of organisms that cause human diseases, as the previous section suggests. Indeed, most of the early efforts in sequencing have focused on human pathogens.

- *Chlamydia trachomatis* causes the most common sexually transmitted disease in the United States. Because it is an intracellular parasite, it has been very hard to study. Among its 900 genes are several for ATP synthesis—something scientists used to think this bacterium could not do.
- *Rickettsia prowazekii* causes typhus; it infects people bitten by louse vectors. Of its 634 genes, 6 code for proteins that are essential for its virulence. These genes are being used to develop vaccines.
- *Mycobacterium tuberculosis* causes tuberculosis. It has a large (for a prokaryote) genome, coding for 4,000 proteins. Over 250 of these proteins are used to metabolize lipids, so this may be the main way that the bacterium gets its energy. Some of its genes code for previously unidentified cell surface proteins; these genes are targets for potential vaccines.
- *Streptomyces coelicolor* and its close relatives produce two-thirds of all the antibiotics currently in clinical use, including streptomycin, tetracycline, and erythromycin. The genome sequence of this bacterium reveals that there are 22 clusters of genes responsible for antibiotic production, of which only 4 were previously known. This finding may lead to more and better antibiotics to combat resistant pathogens.
- *E. coli* strain O157:H7 in hamburger can cause severe illness when ingested, as happens to at least 70,000 people a year in the United States. Its genome has 5,416 genes, of which 1,387 are different from those in the familiar (and harmless) laboratory strains of this bacterium. Remarkably, many of these unique genes are also present in other pathogenic species, such as *Salmonella* and *Shigella*. This finding suggests that there is extensive genetic exchange between these species, and that "superbugs" are on the horizon.

What genes are required for cellular life?

When the genomes of prokaryotes and eukaryotes are compared, a striking conclusion arises: There are some universal genes that are present in all organisms. There are also some universal gene segments—coding for an ATP binding site, for example—that are present in many genes in many organisms. These findings suggest that there is some ancient, minimal set of DNA sequences that all cells must have. One way to identify these sequences is to look for them (or, more realistically, to have a computer look for them).



13.22 Using Transposon Mutagenesis to Determine the Minimal Genome By inactivating genes one by one, scientists can determine which genes are essential for the cell's survival.

Another way to define the minimal genome is to take the organism with the simplest genome, deliberately mutate one gene at a time, and see what happens. *Mycoplasma genitalium* has the smallest known genome—only 470 genes. Even so, some of its genes are dispensable under some circumstances. It has genes for metabolizing both glucose and fructose. In the laboratory, the organism can survive on a medium supplying only one of those sugars, making the genes for metabolizing the other sugar unnecessary. But what about other genes? Experiments using transposons as mutagens have addressed this question. When the bacterium is exposed to transposons, they insert themselves into a gene at random, mutating and inactivating it (Figure 13.22). The mutated cells are sequenced to determine which gene was mutated, and then examined for growth and survival.

The astonishing result of these studies is that *M. genitalium* can survive in the laboratory without the services of 133 of its genes, leaving a minimum genome of 337 genes! This "genomic downsizing" has also been found in other prokaryotes. The bacterium that causes leprosy, *Mycobacterium leprae*, is a cousin of *Mycobacterium tuberculosis*, mentioned above. But *M. leprae* has "discarded" 2,000 of the genes present in its cousin. For example, it lacks genes for the proteins of the electron transport chain (see Chapter 7), and is therefore slow-growing. But it retains the anabolic pathways it needs to survive when external nutrients are scarce.

Chapter Summary

Probing the Nature of Genes

▶ Prokaryotes and viruses are useful for the study of genetics and molecular biology because they contain much less DNA than eukaryotes, grow and reproduce rapidly, and are haploid.

Viruses: Reproduction and Recombination

▶ Viruses were discovered as disease-causing agents small enough to pass through a filter that retains bacteria. The basic viral unit, called a virion, consists of a nucleic acid genome, which codes for a few proteins, and a protein coat called a capsid.

▶ Viruses are obligate intracellular parasites: they need the biochemical machinery of a living cell in order to reproduce.

▶ There are many types of viruses, classified by their size and shape, by their genetic material (RNA or DNA), or by their host organism. **Review Figure 13.1**

▶ Bacteriophage are viruses that infect bacteria. In the lytic cycle, the host cell bursts, releasing new phage particles. Some phage can also undergo a lysogenic cycle, in which their DNA is inserted into the host chromosome, where it replicates for generations. When conditions are appropriate, the phage DNA exits the host chromosome and enters a lytic cycle. **Review Figure 13.2**

▶ Some viruses have promoters for host RNA polymerase, which they use to transcribe their own genes. **Review Figure 13.3**

▶ Most of the many types of RNA and DNA viruses that infect animals cause diseases. Some animal viruses have an envelope derived from the host's plasma membrane.

▶ Retroviruses, such as HIV, have RNA genomes that they reproduce through a complementary DNA intermediate. Other RNA viruses use their RNA to make mRNA to code for enzymes and replicate their genomes without using DNA. **Review Figures 13.4, 13.5**

▶ Many viruses are spread by vectors, such as insects.

Prokaryotes: Reproduction and Recombination

▶ When bacteria divide, they form clones of identical cells that can be observed as colonies when grown on solid media. **Review Figure 13.6**

▶ A bacterium can transfer its genes to another bacterium by conjugation, transformation, or transduction.

▶ In conjugation, a bacterium attaches to another bacterium and passes a fragment of its DNA to the recipient cell. **Review Figures 13.7, 13.8, 13.9**

▶ In transformation, fragments of bacterial DNA are taken up by a cell from the environment. These genetic fragments may recombine with the host chromosome, thereby permanently adding new genes. **Review Figure 13.10*a***

▶ In transduction, phage capsids carry bacterial DNA from one bacterium to another. **Review Figure 13.10*b***

▶ Plasmids are small bacterial chromosomes that are independent of the main chromosome. R factors, which are plasmids that carry genes for antibiotic resistance, are a serious public health threat. **Review Figure 13.11**

▶ Transposable elements are stretches of DNA that can move from one place to another on the bacterial chromosome—either by actually moving or by making a new copy, which is inserted at a new location. **Review Figure 13.12**

Regulation of Gene Expression in Prokaryotes

▶ In prokaryotes, the synthesis of some proteins is regulated so that they are made only when they are needed.

▶ Constitutive enzymes whose products are essential to the cell at all times, are synthesized constantly. A compound that stimulates the synthesis of an enzyme needed to process it is called an inducer, and the enzyme is called an inducible enzyme. **Review Figures 13.13, 13.14**

▶ An operon consists of a promoter, an operator, and two or more structural genes. Promoters and operators do not code for proteins, but serve as binding sites for regulatory proteins. When a repressor protein binds to the operator, transcription of the structural genes is inhibited. **Review Figures 13.15, 13.16**

▶ The mechanisms that regulate the expression of prokaryotic genes include inducible operator–repressor systems, repressible operator–repressor systems, and systems that increase the efficiency of a promoter. **Review Table 13.2**

▶ The *lac* operon is an example of an inducible system. When lactose is absent, a repressor protein binds tightly to the operator. The repressor prevents RNA polymerase from binding to the promoter, turning transcription off. Lactose acts as an inducer by binding to the repressor. This binding changes the repressor's shape so that it can no longer bind to the operator. With the operator unbound, RNA polymerase binds to the promoter, and transcription is turned on. **Review Figure 13.17.** **See Web/CD Tutorial 13.1**

▶ Repressor proteins are coded by constitutive regulatory genes.

▶ The *trp* operon is an example of a repressible system. The presence of tryptophan, the end product of a metabolic pathway, represses synthesis of the enzymes involved in that pathway. Tryptophan acts as a corepressor by binding to an inactive repressor protein and making it active. When the activated repressor binds to the operator, transcription is turned off. **Review Figure 13.18.** **See Web/CD Tutorial 13.2**

▶ The efficiency of a promoter can be increased by regulation of the level of cAMP, which binds to a protein called CRP. The CRP–cAMP complex then binds to a site near the promoter, enhancing the effectiveness of RNA polymerase binding and hence transcription. **Review Figure 13.19**

Control of Transcription in Viruses

▶ In bacteriophage that can undergo a lytic or a lysogenic cycle, the decision as to which pathway to take is made by operator–regulatory protein interactions. **Review Figure 13.20**

Prokaryotic Genomes

▶ Functional genomics relates gene sequences to protein functions. **Review Figure 13.21**

▶ By mutating individual genes in a small genome, scientists can determine the minimal genome required for cellular life. **Review Figure 13.22**

See Web/CD Activity 13.1 for a concept review of this chapter.

Self-Quiz

1. Which of the following is *not* true with regard to the *lac* operon?
   a. When lactose binds to the repressor, the latter can no longer bind to the operator.
   b. When lactose binds to the operator, transcription is stimulated.
   c. When the repressor binds to the operator, transcription is inhibited.
   d. When lactose binds to the repressor, the shape of the repressor is changed.
   e. When the repressor is mutated, one possibility is that it does not bind to the operator.
2. Which of the following is *not* a type of virus reproduction?
   a. DNA virus in a lytic cycle
   b. DNA virus in a lysogenic cycle
   c. RNA virus by a double stranded RNA intermediate
   d. RNA virus by reverse transcription to make cDNA
   e. RNA virus by acting as tRNA
3. In the lysogenic cycle of a bacteriophage,
   a. a repressor, cI, blocks the lytic cycle.
   b. a bacteriophage carries DNA between bacterial cells.
   c. both early and late phage genes are transcribed.
   d. the viral genome is made into RNA which stays in the host cell.
   e. many new viruses are made immediately, regardless of host health.
4. An operon is
   a. a molecule that can turn genes on and off.
   b. an inducer bound to a repressor.
   c. regulatory sequences controlling protein-coding genes.
   d. any long sequence of DNA.
   e. a group of linked genes.
5. Which statement about both transformation and transduction is *true*?
   a. DNA is transferred between viruses and bacteria.
   b. Neither occurs in nature.
   c. Small fragments of DNA move from one cell to another.
   d. Recombination between the incoming DNA and host cell DNA does not occur.
   e. A conjugation tube is used to transfer DNA between cells.
6. Plasmids
   a. are circular protein molecules.
   b. are required by bacteria.
   c. are tiny bacteria.
   d. may confer resistance to antibiotics.
   e. are a form of transposable element.
7. The minimal genome can be estimated for a prokaryote
   a. by counting the total number of genes.
   b. by comparative genomics.
   c. as about 5,000 genes.
   d. by transposon mutagenesis, one gene at a time.
   e. does not include any genes coding for tRNA.
8. When tryptophan accumulates in a bacterial cell,
   a. it binds to the operator, preventing transcription of adjacent genes.
   b. it binds to the promoter, allowing transcription of adjacent genes.
   c. it binds to the repressor, causing it to bind to the operator.
   d. it binds to the genes that code for enzymes.
   e. it binds to RNA and initiates a negative feedback loop to reduce transcription.
9. The promoter in the *lac* operon is
   a. the region that binds the repressor.
   b. the region that binds RNA polymerase.
   c. the gene that codes for the repressor.
   d. a structural gene.
   e. an operon.
10. The CRP–cAMP system
    a. produces many catabolites.
    b. requires ribosomes.
    c. operates by an operator–repressor mechanism.
    d. is an example of positive control of transcription.
    e. relies on operators.

For Discussion

1. Viruses sometimes carry DNA from one cell to another by transduction. Sometimes a segment of bacterial DNA is incorporated into a phage protein coat without any phage DNA. These particles can infect a new host. Would the new host become lysogenic if the phage originally came from a lysogenic host? Why or why not?
2. Compare the life cycles of the viruses that cause influenza and AIDS (Figures 13.4 and 13.5) with respect to:
   • How the virus enters the cell
   • How the virion is released in the cell
   • How the viral genome is replicated
   • How new viruses are produced
3. Compare promoters adjacent to "early" and "late" genes in the bacteriophage lytic cycle.
4. In the lactose (*lac*) operon of *E. coli*, repressor molecules are encoded by the regulatory gene. The repressor molecules are made in very small quantities and at a constant rate per cell. Would you surmise that the promoter for these repressor molecules is efficient or inefficient? Is synthesis of the repressor constitutive, or is it under environmental control?
5. A key characteristic of a repressible enzyme system is that the repressor molecule must react with a corepressor (typically, the end product of a pathway) before it can combine with the operator of an operon to shut the operon off. How is this different from an inducible enzyme?

11 The Eukaryotic Genome and Its Expression

"The most precious things are not jade or pearls, but the five grains." This ancient Chinese saying refers to rice, wheat, maize (corn), sorghum, and millet. Today the saying remains as true as ever, since these crops provide two-thirds of the human diet worldwide. With the recent publication of the genome sequences of the two major cultivated varieties of rice, agricultural scientists are well on the way to dramatic improvements in the nutritional quality and yield of the grain produced in the paddies of Asia. And the rice genome turns out to be a smaller version of the much larger genomes of the other four grains.

Like other eukaryotes, rice has much more DNA than a typical prokaryote—some 430 million base pairs. But unlike the densely packed prokaryotic genome, the rice genome contains many stretches of DNA that do not code for proteins or RNA. Some of these sequences are "spacers," which is another way of saying that they either have no function or that no function has yet been found. Others are repetitive sequences, such as the telomeric DNA at the ends of chromosomes (see Figure 11.18).

In addition to the genes for metabolism that they share with prokaryotes, eukaryotes have genes that mark them as complex organisms: genes for addressing, or targeting, proteins to organelles, and genes for cell–cell interaction and cell differentiation. The transcription and later processing of mRNA is more complicated in eukaryotes than in prokaryotes. Elegant molecular machinery allows the precise regulation of gene expression needed for all the cells of these complex organisms to develop and function.

"The Most Precious Things" The genome of rice (*Oryza sativa*), which directly supplies a third of the overall diet of humanity, was recently sequenced.

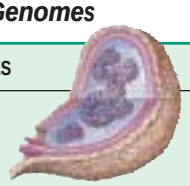
The Eukaryotic Genome

As biologists began to unravel the intricacies of gene structure and expression in prokaryotes, they tried to generalize their findings by stating, "What's true for *E. coli* is also true for elephants." Although much of prokaryotic biochemistry does apply to eukaryotes as well, the old saying has its limitations. Table 14.1 lists some of the differences between prokaryotic and eukaryotic genomes.

The eukaryotic genome is larger and more complex than the prokaryotic genome

Comparisons of prokaryotic and eukaryotic genomes reveal several features.

14.1 A Comparison of Prokaryotic and Eukaryotic Genes and Genomes

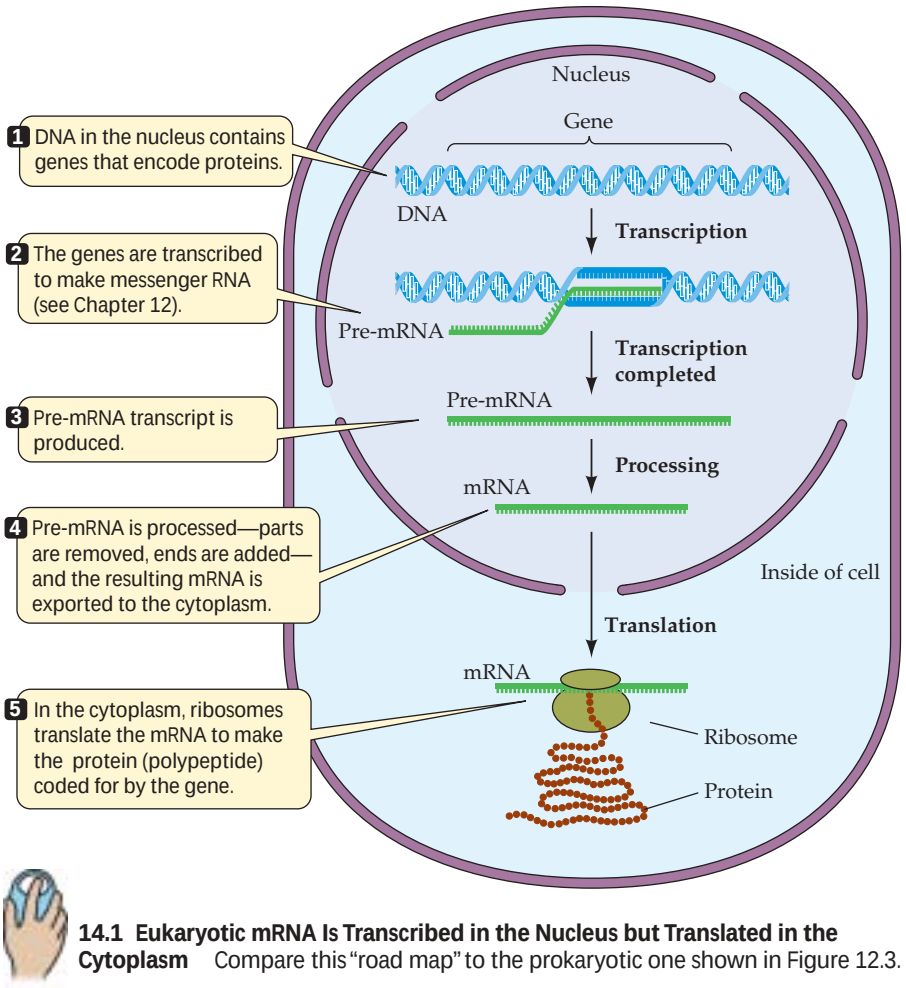


CHARACTERISTIC	PROKARYOTES	EUKARYOTES
Genome size (base pairs)	10^4–10^7	10^8–10^{11}
Repeated sequences	Few	Many
Noncoding DNA within coding sequences	Rare	Common
Transcription and translation separated in cell	No	Yes
DNA segregated within a nucleus	No	Yes
DNA bound to proteins	Some	Extensive
Promoters	Yes	Yes
Enhancers/silencers	Rare	Common
Capping and tailing of mRNA	No	Yes
RNA splicing required (spliceosomes)	Rare	Common
Number of chromosomes in genome	One	Many

- *Eukaryotic genomes are larger.* The genomes of eukaryotes (in terms of haploid DNA content) are larger than those of prokaryotes. This difference is not surprising, given that in multicellular organisms there are many cell types, many jobs to do, and many proteins—all encoded by DNA—needed to do those jobs. A typical virus contains enough DNA to code for only a few proteins—about 10,000 base pairs (bp). The most thoroughly studied prokaryote, *E. coli*, has sufficient DNA (about 4.5 million bp) to make several thousand different proteins and regulate their synthesis. Humans have considerably more genes and regulators: Nearly 6 billion bp (2 meters of DNA) are crammed into each diploid human cell. However, the idea of a more complex organism needing more DNA seems to break down with some plants. For example, the lily (which produces beautiful flowers each spring, but produces fewer proteins than a human does) has 18 times more DNA than a human.
- *Eukaryotic genomes have more regulatory sequences.* Eukaryotic genomes have many more regulatory sequences—and many more regulatory proteins that bind to them—than prokaryotic genomes do. The great complexity of eukaryotes requires a great deal of regulation, and this fact is evident in the many processes and points of control associated with the expression of the eukaryotic genome.
- *Much of eukaryotic DNA is noncoding.* Interspersed throughout the eukaryotic genome are various kinds of repeated DNA sequences that are not transcribed into proteins. Even the coding regions of genes contain sequences that do not appear in the mRNA that is translated at the ribosome.
- *Eukaryotes have multiple chromosomes.* The genomic encyclopedia of a eukaryote is separated into multiple volumes. This separation requires that each chromosome have, at a minimum, three defining DNA sequences that we have described in previous chapters: an origin of replication recognized by the DNA replication machinery; a centromere region that holds the replicated chromosomes together before mitosis; and a telomeric sequence at each end of the chromosome.
- *In eukaryotes, transcription and translation are physically separated.* The nuclear envelope separates DNA and its transcription (inside the nucleus) from the sites where mRNA is translated into protein (in the cytoplasm). This separation allows for many points of regulation before translation begins: in the synthesis of a pre-mRNA transcript, in its processing into mature mRNA, and in its transport to the cytoplasm for translation (Figure 14.1).

The yeast genome adds some eukaryotic functions to a prokaryotic model

In comparison with *E. coli*, whose genome has about 4,500,000 bp on a single chromosome (one circular DNA molecule), the genome of budding yeast (*Saccharomyces cerevisiae*), a single-celled eukaryote, has 16 linear chromosomes and a haploid content of more than 12,068,000 bp. More than 600 scientists around the world collaborated in mapping and sequencing the yeast genome. When they began, they knew of about 1,000 yeast genes coding for RNAs or proteins. The final sequence revealed 5,900 genes, and sequence analyses have assigned probable roles to about 70 percent of them. Some of these genes are homologous to genes found in prokaryotes, but many are not. The functions of the other 30 percent are being investigated by gene inactivation studies similar to those performed on prokaryotes (see Figure 13.22). This process of discovering the protein product and function of a known gene sequence is called **annotation**. These accomplishments have made yeast an important model for eu-



14.1 Eukaryotic mRNA Is Transcribed in the Nucleus but Translated in the Cytoplasm Compare this "road map" to the prokaryotic one shown in Figure 12.3.

karyotic cells, as observations and hypotheses from studies on yeast can be applied to and tested on other eukaryotes.

It is now possible to estimate the proportions of the yeast genome that code for specific metabolic functions. Apparently, 11 percent of yeast proteins function in general metabolism, 3 percent in energy production and storage, 3 percent in DNA replication and repair, 12 percent in protein synthesis, and 6 percent in targeting ("addressing") proteins to organelles and for secretion outside the cell. Many of the other two-thirds of the proteins are involved in cell structure, cell division, and the regulation of gene expression.

The most striking difference between the yeast genome and that of *E. coli* is in the genes for protein targeting (Table 14.2). Both of these single-celled organisms appear to use about the same numbers of genes to perform the basic functions of cell survival. It is the compartmentalization of the eukaryotic yeast cell into organelles that requires it to have so many more genes. This finding is direct, quantitative confirmation of something we have known for a century: The eukaryotic cell is structurally more complex than the prokaryotic cell.

Genes encoding several other types of proteins are present in the yeast and other eukaryotic genomes, but have no homologs in prokaryotes:

- Genes encoding histones that package DNA into nucleosomes
- Genes encoding cytoskeletal and motor proteins such as actin and tubulin
- Genes encoding cyclin-dependent kinases that control cell division
- Genes encoding proteins involved in the processing of RNA

The nematode genome adds developmental complexity

The presence of more than a single cell adds a new level of complexity to the eukaryotic genome. *Caenorhabditis elegans* is a 1-mm-long nematode (roundworm) that normally lives in the soil. But it also lives in the laboratory, where it is a favorite study organism of developmental biologists (see Chapter 19). In fact, the 2002 Nobel prize in physiology and medicine was awarded to researchers who used this worm to study development and the control of cell division. The worm has a transparent body, which scientists can watch over 3 days as a fertilized egg divides and forms an adult worm of nearly 1,000 cells. In spite of its small number of cells, the worm has a nervous system, digests food, reproduces sexually, and ages.

14.2 Comparison of the Genomes of E. coli and Yeast

	E. COLI	YEAST
Genome length (base pairs)	4,640,000	12,068,000
Number of proteins	4,300	6,200
Proteins with roles in:		
Metabolism	650	650
Energy production/storage	240	175
Membrane transporters	280	250
DNA replication/repair/recombination	120	175
Transcription	230	400
Translation	180	350
Protein targeting/secretion	35	430
Cell structure	180	250

So it is not surprising that an intense effort was made to sequence the genome of this organism.

The *C. elegans* genome is eight times larger than that of yeast (97 million bp) and has four times as many protein-coding genes (19,099). Once again, sequencing revealed far more genes than expected: When the sequencing effort began, researchers estimated that the worm would have about 6,000 genes and about that many proteins. Clearly, it has far more. About 3,000 genes in the worm have direct homologs in yeast; these genes code for basic eukaryotic cell functions. What do the rest of the genes—the bulk of the worm genome—do?

In addition to surviving, growing, and dividing, as single-celled organisms do, multicellular organisms must have genes for holding cells together to form tissues, for cell differentiation to divide up tasks among those tissues, and for intercellular communication to coordinate their activities (Table 14.3). Many of the genes so far identified in *C. elegans* that are not present in yeast perform these roles, which will be described in the remainder of this chapter and the next one.

The fruit fly genome has surprisingly few genes

The fruit fly *Drosophila melanogaster* is a much larger organism than *C. elegans*, both in size (the fly has 10 times more cells) and complexity. Not surprisingly, the fly's genome is also larger (about 180,000,000 bp). New computerized sequencing technologies made it possible to sequence the entire *Drosophila* genome in about a year.

Even before the complete sequence was announced, decades of genetic studies had identified some 2,500 different genes in the fly. These genes were all found in the complete DNA sequence, along with many other genes whose functions are as yet unidentified. But the big surprise of the *Drosophila* genome sequence was the total number of protein-coding regions. Instead of having more genes than the roundworm, the fly has fewer: only 13,600 genes. One reason for this is that the roundworm has some large gene families, which, as we will see later in this chapter, are groups of genes that are related in their sequence and function. For example, *C. elegans* has 1,100 genes involved in either nerve cell signaling or development; the fly has only 160 genes for these two functions. Another major genetic expansion in the worm is in the genes coding for proteins that sense chemicals in its environment.

Many genes that are present in the worm genome have homologs with similar sequences in fly DNA; such homologs account for a third of the fly's genes. Furthermore, about half of the fly's genes have mammalian homologs. Comparative genomics has made an important contribution to medicine through the discovery of homologs in other organisms of genes that are implicated in human diseases. Often the role of such a gene can be elucidated in the simpler organism, providing a clue to how the gene might function in human disease. The fly genome contains 177 genes with sequences similar to genes that also occur in the human genome and are involved in human diseases, including cancer and neurological conditions.

14.3 C. elegans *Genes Essential to Multicellularity*

FUNCTION	PROTEIN/DOMAIN	NUMBER OF GENES
Transcription control	Zinc finger; homeobox	540
RNA processing	RNA binding domains	100
Nerve impulse transmission	Gated ion channels	80
Tissue formation	Collagens	170
Cell interactions	Extracellular domains; glycotransferases	330
Cell–cell signaling	G protein-linked receptors; protein kinases; protein phosphatases	1,290

The puffer fish is a vertebrate with a compact genome

The puffer fish, *Fugu rubripes*, is prized in the culinary world as a gourmet item from Japan that must be carefully prepared, as it contains a lethal poison called tetrodotoxin that inhibits membrane channels in nerve cells. In the biological world, it is prized for its genome, which is the most compact known among vertebrates. It has 365 million base pairs and about 30,000 genes. The human genome has one-third fewer genes in eight times the amount of DNA.

A comparison of the human and puffer fish genomes showed that many genes in the two organisms are similar, so that, as lead scientist Sydney Brenner (who also led the study of the *C. elegans* genome) put it, "the *Fugu* genome is the '*Reader's Digest* version of the Book of Man.'" A major difference between the two genomes is in repetitive DNA sequences, which make up 40 percent of the human genome but a much smaller proportion in the puffer fish. The significance of this finding is unknown. Of course, humans are obviously much more complex than fish; how we accomplish this with a set of genes that is even smaller in number than the fish's is not known, but certainly points up the fact that it is not genes alone that determine the complexity of an organism.

The rice genome reflects that of a model plant, *Arabidopsis*

About 250,000 species of flowering plants dominate land and fresh water. But in the his-

tory of life, the flowering plants are fairly young, having evolved only about 200 million years ago. Given the pace of DNA mutation and other genetic changes, the differences among these plants are likely to be relatively small—at the level of regulation and protein synthesis, rather than in the genes. So, although it is the genomes of the plants used by people as food and fiber that hold the greatest interest for us, it is not surprising that instead of sequencing the huge genomes of wheat (16 billion bp) or corn (3 billion bp), scientists first chose to sequence a simpler flowering plant.

Arabidopsis thaliana, the shale cress, is a member of the mustard family and has long been a favorite model organism for study by plant biologists. It is small (hundreds could grow and reproduce in the space occupied by this page), is easy to manipulate, has only 10 percent repetitive DNA, and has a small (119 million bp) genome. Its DNA sequence reveals about 26,000 protein-coding genes, but remarkably, many of these are duplicates of other genes and have probably originated by chromosomal rearrangements. When these duplicate genes are subtracted from the total, about 15,000 unique genes are left, a number not too dissimilar from the fruit fly and roundworm. Indeed, many of the genes found in these invertebrate animals have homologs in the plant, suggesting that plants and animals have a common ancestor.

But *Arabidopsis* has some genes that distinguish it as a plant (Table 14.4). These genes include those involved in photosynthesis, in the transport of water into the root and throughout the plant, in the assembly of the cell wall, in the uptake and metabolism of inorganic substances from the environment , and in the synthesis of specific molecules used for defense against plant predators.

Justifying its position as a model plant, these "plant" genes in *Arabidopsis* were also found in the genome of rice, the first major crop plant whose sequence has been determined. Rice (*Oryza sativa*) is the world's most important crop; it is the staple diet for 3 billion people, many of them very poor. Actually, two *O. sativa* sequences of have been deciphered: that of *O. sativa indica*, the rice subspecies grown in China and most of tropical Asia, and that of the subspecies *japonica*, which is grown in Japan and other temperate climates (such as the United States). Both genomes are about the same size (430 million bp), yet in this much larger genome is a set of genes remarkably similar to that of *Arabidopsis* (Table 14.5). And many of the genes in rice are also present in the much larger genomes of corn and wheat.

Of course, rice as a whole, and each subspecies, has its own particular set of genes that make it unique. The *indica* subspecies is estimated to have 46,000–55,000 such genes, and *japonica* 32,000–50,000, both numbers higher than *Arabidopsis*. These "extra" genes include genes for characters that are specific to rice, such as a physiology that allows rice to grow for part of the season submerged in water; the nutrient-packed seeds that sustain human lives; and resistance to certain plant diseases, such as viruses and fungi. Analyses of these and other rice genes will no doubt lead to significant improvements in this crop, and to the improvement of the other grain crops as well.

14.4 Arabidopsis *Genes Unique to Plants*

FUNCTION	NUMBER OF GENES
Cell wall and growth	420
Water channels	300
Photosynthesis	139
Defense and metabolism	94

Repetitive Sequences in the Eukaryotic Genome

As you have seen in the genome sequences we have examined so far, the eukaryotic genome contains some base sequences that are repeated many times. Some of these sequences are present in millions of copies in a single genome. In this section, we will examine the organization and possible roles of these repetitive sequences.

Highly repetitive sequences are present in large numbers of copies

Three types of *highly repetitive sequences* are found in eukaryotes:

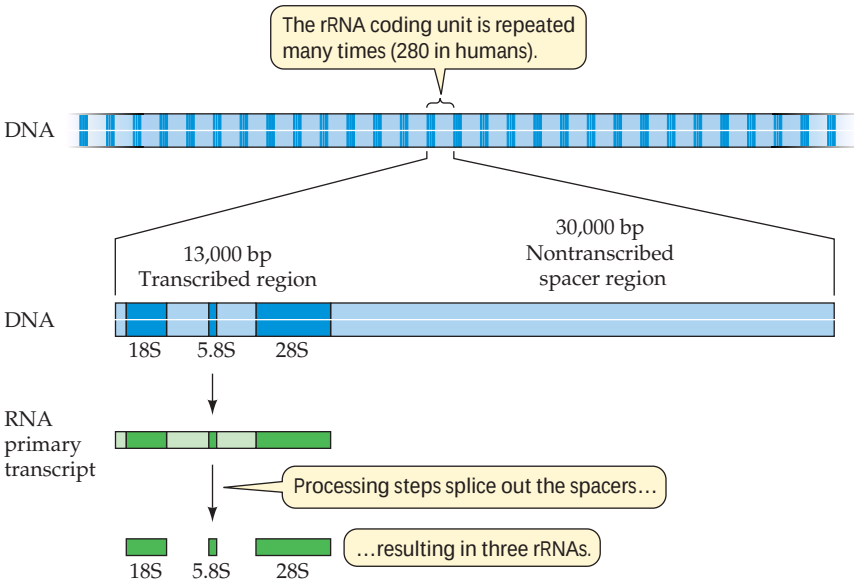
- *Satellites* are sequences 5–50 base pairs long, repeated side by side up to a million times. Satellites are usually present at the centromeres of chromosomes. They appear to be important in binding the special proteins that make up the centromere.

14.5 *Comparison of the Rice and* Arabidopsis *Genomes*

FUNCTION	PERCENTAGE OF GENOME: RICE	PERCENTAGE OF GENOME: *ARABIDOPSIS*
Cell structure	9	10
Enzymes	21	20
Ligand binding	10	10
DNA binding	10	10
Signal transduction	3	3
Membrane transport	5	5
Cell growth and maintenance	24	22
Other functions	18	20

- *Minisatellites* are 12–100 base pairs long and are repeated several thousand times. Because DNA polymerase tends to make errors in copying these sequences, the number of copies present varies among individuals. For example, one person might have 300 minisatellites and another, 500. This variation provides a set of molecular genetic markers that can be used to identify an individual.
- *Microsatellites* are very short (1–5 bp) sequences, present in small clusters of 10–50 copies. They are scattered all over the genome.

These highly repetitive sequences are not transcribed into RNA. While laboratory scientists have made use of these sequences in genetic studies, their roles in eukaryotes are not clear.



14.2 A Moderately Repetitive Sequence Codes for rRNA This rRNA gene, along with its nontranscribed spacer region, is repeated 280 times in the human genome, with clusters on five chromosomes. Once this gene has been transcribed, posttranscriptional processing removes the spacers within the transcribed region and separates the primary transcript into the three final rRNA products.

Some moderately repetitive sequences are transcribed

In Chapter 11, we described one kind of *moderately repetitive sequence* found at the ends of chromosomes: the telomeres that maintain the length and integrity of the chromosome as it replicates. These sequences are not transcribed into RNA. In contrast, some moderately repetitive DNA sequences are transcribed. These sequences code for tRNAs and rRNAs, which are used in protein synthesis (see Chapter 12).

The cell transcribes tRNAs and rRNAs constantly, but even at the maximum rate of transcription, single copies of the DNA sequences coding for them would be inadequate to supply the large amounts of these molecules needed by most cells; hence, the genome has multiple copies of these sequences. Since these moderately repetitive sequences are transcribed into RNA, they are properly termed "genes," and we can speak of rRNA genes and tRNA genes.

In mammals, there are four different rRNA molecules that make up the ribosome: the 18S, 5.8S, 28S, and 5S rRNAs. (The "S" term describes how a substance behaves in a centrifuge and, in general, is related to the size of a molecule.) The 18S, 5.8S, and 28S rRNAs are transcribed from a repeated DNA sequence as a single precursor RNA molecule, which is twice the size of the three ultimate products (Figure 14.2). Several posttranscriptional steps cut this precursor into the final three rRNA products and discard the nonuseful, or "spacer," RNA. The sequence encoding these RNAs is moderately repetitive in humans: A total of 280 copies of the sequence are located in clusters on five different chromosomes.

These moderately repetitive sequences remain fixed in their locations on the genome. Another class of moderately repetitive sequences, however, can change their location, moving about the genome.

Transposons move about the genome

Most of the remaining scattered moderately repetitive DNA sequences are not stably integrated into the genome. Instead, these sequences can move from place to place in the genome. Such sequences are called **transposons**. They make up about 45 percent of the human genome, far more than the 3–10 percent found in the other sequenced eukaryotes.

There are four main types of transposons in eukaryotes:

- *SINEs* (*s*hort *in*terspersed *e*lements) are up to 500 bp long and are transcribed, but not translated.
- *LINEs* (*l*ong *in*terspersed *e*lements) are up to 7,000 bp long, and some are transcribed and translated into proteins. They constitute about 15 percent of the human genome.

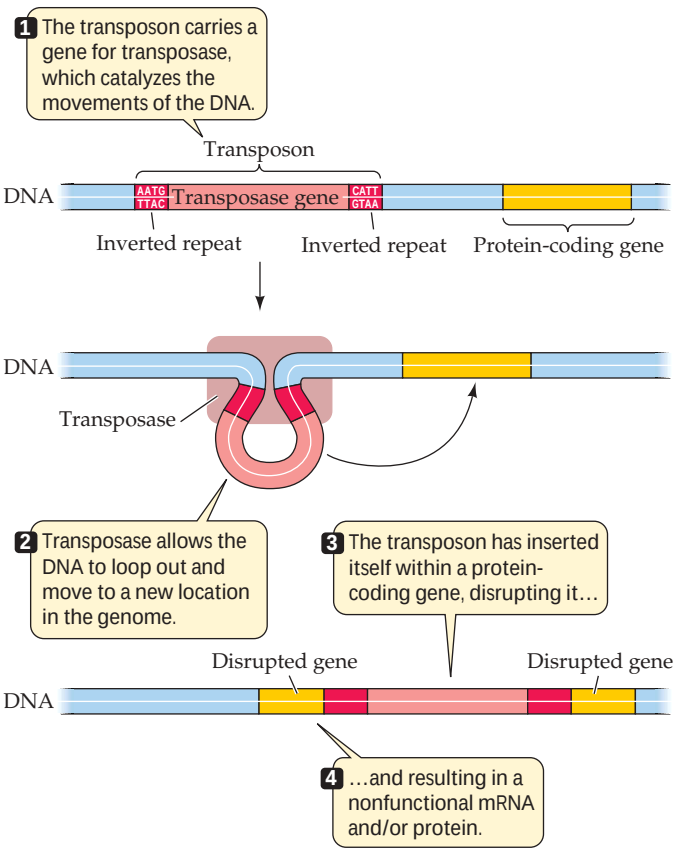
Both of these elements are present in more than 100,000 copies. They move about the genome in a distinctive way: They make an RNA copy of themselves, which acts as a template for new DNA, which then inserts itself at a new location in the genome. In this "copy and paste" mechanism, the original sequence stays where it is and the copy inserts itself at a new location.

- *Retrotransposons* also make an RNA copy of themselves when they move about the genome. They constitute about 17 percent of the human genome. Some of them code for some of the proteins necessary for their own

transposition, and others do not. A single type of retrotransposon, the 300-bp *Alu* element, accounts for 11 percent of the human genome; it is present in a million copies scattered over all the chromosomes.

- *DNA transposons* are similar to their prokaryotic counterparts. They do not use an RNA intermediate, but actually move to a new spot in the genome without replicating (Figure 14.3).

What role do these moving sequences play in the cell? There are few answers to this question. The best answer so far seems to be that transposons are cellular parasites that simply replicate themselves. But these replications can lead to the insertion of a transposon at a new location, which can have important consequences. For example, the insertion of a transposon into the coding region of a gene results in a mutation (see Figure 14.3). This phenomenon has been found in rare forms of several human genetic diseases, including hemophilia and muscular dystrophy. If the insertion of a transposon takes place in the germ line, a gamete with a new mutation results. If the insertion takes place in a somatic cell, cancer may result.



14.3 DNA Transposons and Transposition At the end of each DNA transposon is an inverted repeat sequence that helps in the transposition process.

If a transposon replicates not just itself but also an adjacent gene, the result may be a gene duplication. A transposon can carry a gene, or a part of it, to a new location in the genome, shuffling the genetic material and creating new genes. Clearly, transposition stirs the genetic pot in the eukaryotic genome and thus contributes to genetic variation.

In Chapter 4, we described the theory of endosymbiosis, which proposes that chloroplasts and mitochondria are the descendants of once free-living prokaryotes. Transposons may have played a role in this process. In living eukaryotes, although these organelles contain some DNA, the nucleus contains most of the genes that encode the organelle proteins. If the organelles were once independent, they must originally have contained all of these genes. How did the genes move to the nucleus? The answer may lie in DNA transpositions. Genes in the organelles may have moved to the nucleus by such well-known molecular events, which still occur today. The DNA that remains in the organelles may be the remnants of more complete prokaryotic genomes.

We now turn to the genes that are at the heart of molecular genetics: those that code for proteins.

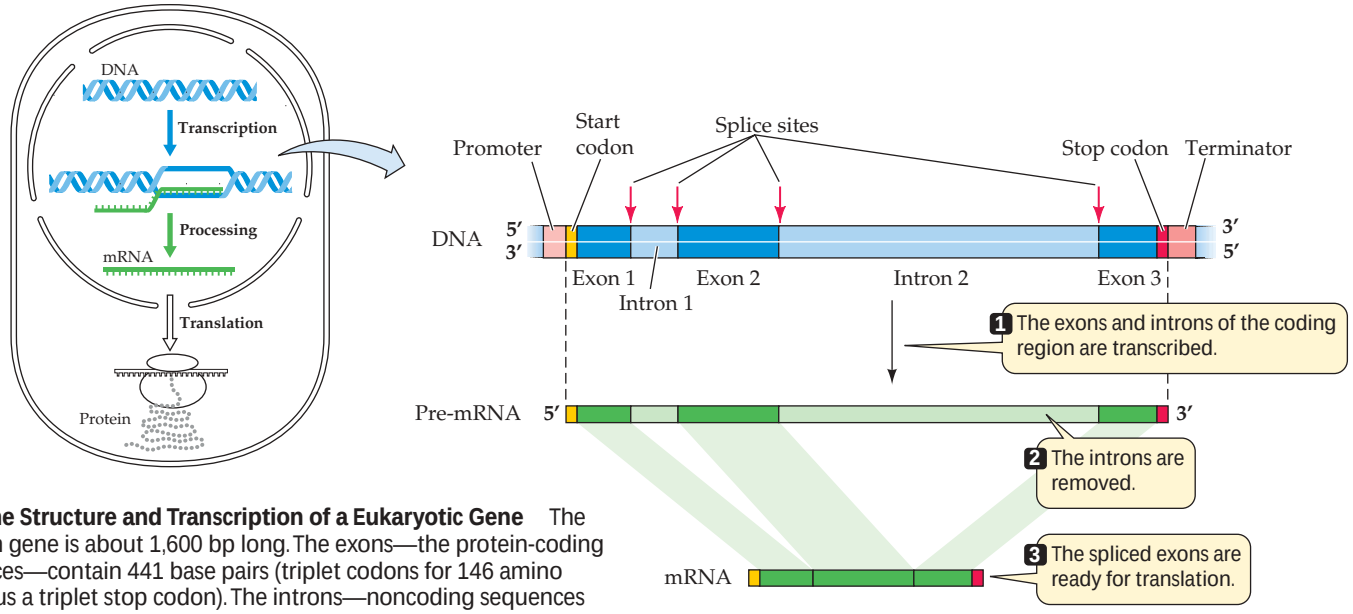
The Structures of Protein-Coding Genes

Like their prokaryotic counterparts, many protein-coding genes in eukaryotes are single-copy DNA sequences. But eukaryotic genes have two distinctive characteristics that are uncommon among prokaryotes. First, they contain noncoding internal sequences, and second, they form gene families—groups of structurally and functionally related "cousins" in the genome.

Protein-coding genes contain noncoding internal and flanking sequences

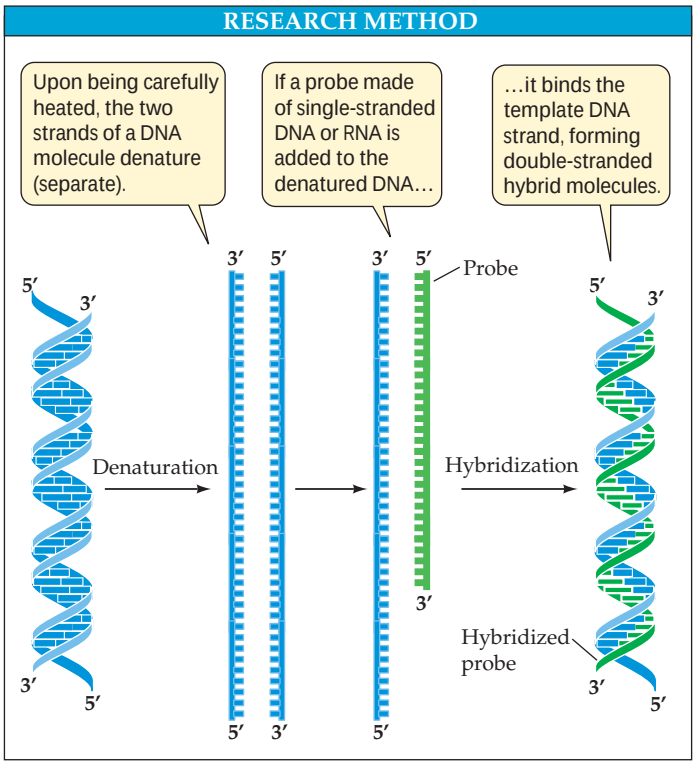
Preceding the coding region of a eukaryotic gene is a **promoter**, to which an RNA polymerase binds to begin the transcription process. Unlike the prokaryotic enzyme, however, a eukaryotic RNA polymerase does not recognize the promoter sequence by itself, but requires help from other molecules, as we'll see below. At the other end of the gene, after the coding region, is a DNA sequence appropriately called the **terminator**, which signals the end of transcription when it is synthesized (Figure 14.4).

Eukaryotic protein-coding genes also contain noncoding base sequences, called **introns**. One or more introns are interspersed with the coding regions called **exons**. Transcripts of the introns appear in the primary transcript of RNA, called **pre-mRNA**, but by the time the **mature mRNA**—the mRNA that will be translated—leaves the nucleus, they have been removed. The transcripts of the introns are cut out of the pre-mRNA, and the transcripts of the exons are spliced together.



14.4 The Structure and Transcription of a Eukaryotic Gene The β-globin gene is about 1,600 bp long. The exons—the protein-coding sequences—contain 441 base pairs (triplet codons for 146 amino acids plus a triplet stop codon). The introns—noncoding sequences of DNA—between codons 30 and 31 (130 bp long) and 104 and 105 (850 bp long), are initially transcribed, but are spliced out of the initial mRNA transcript.

Where are the introns within a eukaryotic gene? The easiest way to find out is by **nucleic acid hybridization**, the method that originally revealed the existence of introns. This research method, outlined in Figure 14.5, has been crucial for studying the relationship between genes and their transcripts.



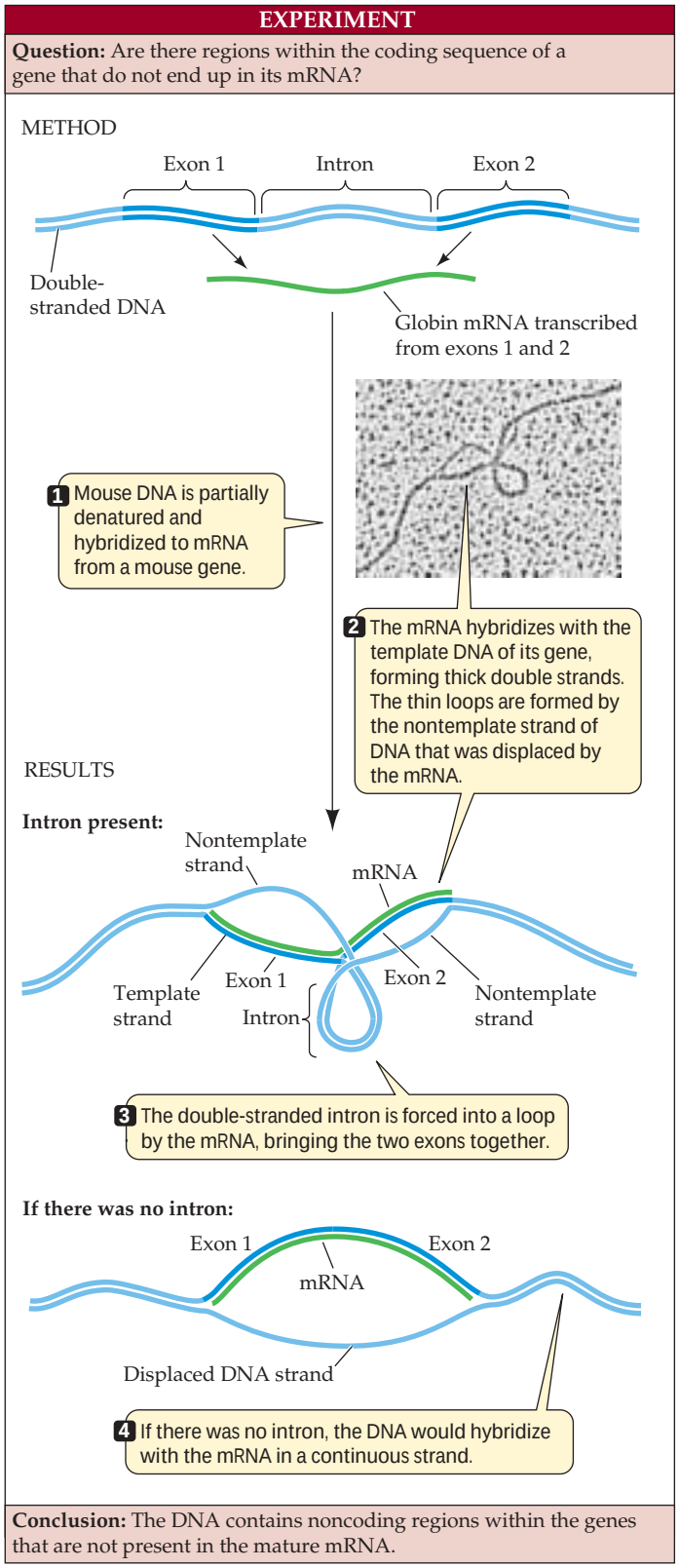
14.5 Nucleic Acid Hybridization Base pairing permits the detection of a sequence complementary to the probe.

Biologists used nucleic acid hybridization to examine the β-globin gene, which encodes one of the globin proteins that make up hemoglobin (Figure 14.6). They first denatured the β-globin DNA by heating it, then added mature β-globin mRNA. As expected, the mRNA bound to the DNA by complementary base pairing. The researchers expected to obtain a linear matchup of the mRNA to the coding DNA. That expectation was met, but only in part: There were indeed stretches of RNA–DNA hybridization, but some looped structures were also visible. These loops were the introns, stretches of DNA that did not have complementary bases on the mature mRNA. Later studies showed that the hybridization of pre-mRNA to DNA was complete, revealing that the introns were indeed transcribed. Somewhere on the path from primary transcript (pre-mRNA) to mature mRNA, the introns had been removed, and the exons had been spliced together. We will examine this splicing process shortly.

Most (but not all) vertebrate genes contain introns, as do many other eukaryotic genes (and even a few prokaryotic ones). Introns interrupt, but do not scramble, the DNA sequence that codes for a polypeptide chain. The base sequence of the exons, taken in order, is complementary to that of the mature mRNA product. The introns, therefore, separate a gene's protein-coding region into distinct parts—the exons. In some cases, the separated exons code for different functional regions, or *domains*, of the protein. For example, the globin proteins that make up hemoglobin have two domains: one for binding to a nonprotein pigment called heme, and another for binding to the other globin subunits. These two domains are encoded by different exons in the globin genes.

Many eukaryotic genes are members of gene families

About half of all eukaryotic protein-coding genes are present in only one copy in the haploid genome. The rest have multiple



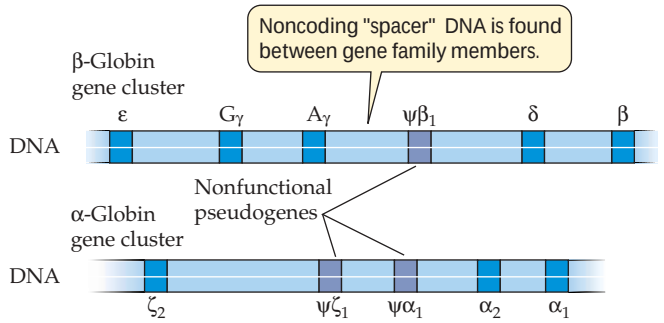
14.6 Nucleic Acid Hybridization Revealed Noncoding DNA
When an mRNA transcript of the β-globin gene was experimentally hybridized to the double-stranded DNA of that gene, the introns in the DNA "looped out," demonstrating that the coding region of a eukaryotic gene can contain noncoding DNA that is not present in the mature mRNA transcript.

copies. Often, inexact, nonfunctional copies of a particular gene, called **pseudogenes**, are closely linked to the functional gene. These duplicates may have arisen by an abnormal event in chromosomal crossing over during meiosis or by the action of transposons. In other cases, however, the genome contains slightly altered duplicates of a gene that are functional.

A set of duplicated or related genes is called a **gene family**. Some gene families, such as the genes encoding the globins that are part of hemoglobin, contain only a few members; other families, such as the genes encoding the immunoglobulins that make up antibodies, have hundreds of members. Like the members of any family, the DNA sequences in a gene family are usually different from one another to a certain extent. As long as one member retains the original DNA sequence and thus codes for the proper protein, the other members can mutate slightly, extensively, or not at all. The availability of such "extra" genes is important for "experiments" in evolution: If the mutated gene is useful, it may be selected for in succeeding generations. If the gene is a total loss (a pseudogene), the functional copy is still there to save the day.

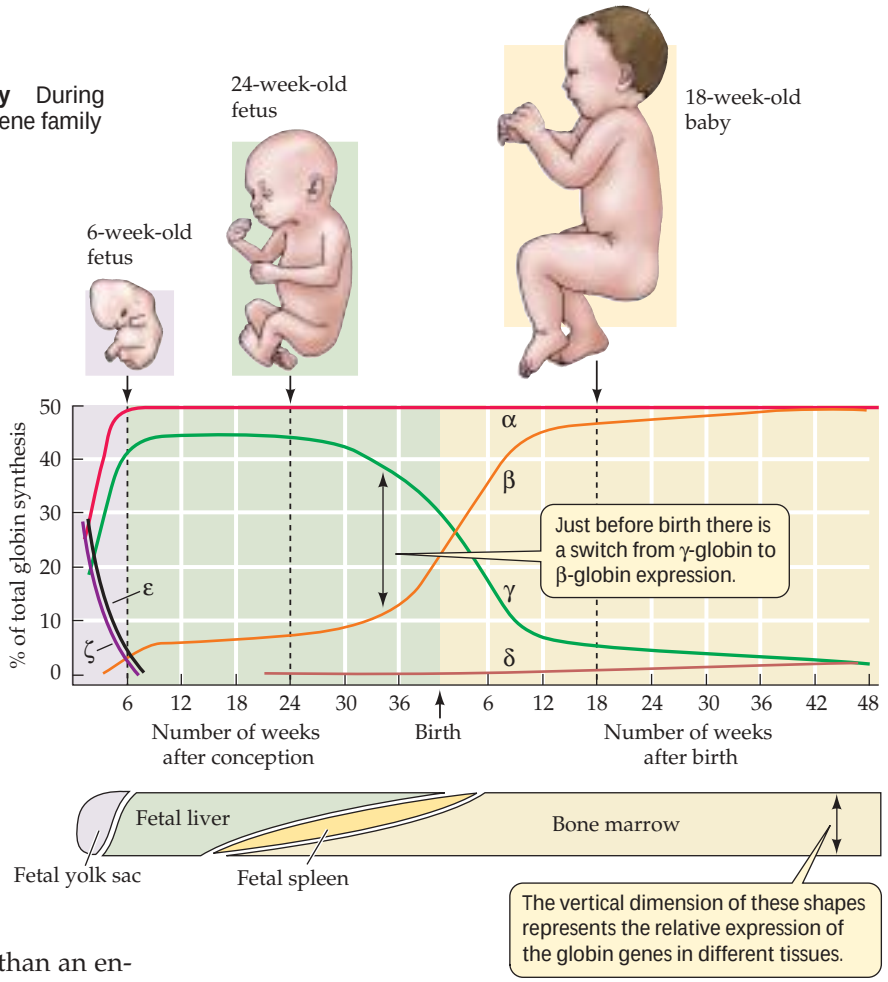
The gene family encoding the globins is a good example of the gene families found in vertebrates. These proteins are found in hemoglobin as well as in myoglobin (an oxygen-binding protein present in muscle). The globin genes all arose from a single common ancestor gene long ago. In humans, there are three functional members of the alpha-globin (α-globin) cluster and five in the beta-globin (β-globin) cluster (Figure 14.7). In an adult, each hemoglobin molecule is a tetramer containing four heme pigments (each held inside a globin polypeptide subunit), two identical α-globin subunits, and two identical β-globin subunits (see Figure 3.8).

During human development, different members of the β-globin gene cluster are expressed at different times and in different tissues (Figure 14.8). This differential gene expression has great physiological significance. For example,



14.7 The Globin Gene Family The α-globin and β-globin clusters of the human globin gene family are located on different chromosomes. The genes of each cluster are separated by noncoding "spacer" DNA. The nonfunctional pseudogenes are indicated by the Greek letter psi (ψ).

14.8 Differential Expression in the Globin Gene Family During human development, different members of the globin gene family are expressed at different times and in different tissues.

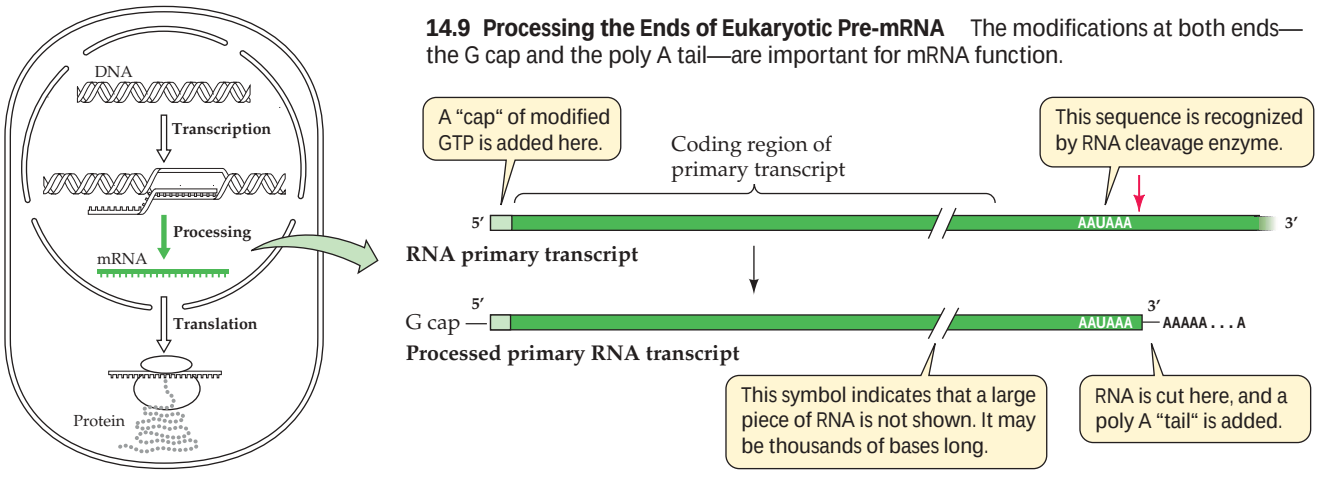


γ-globin, a subunit found in the hemoglobin of the fetus ($\alpha_2\gamma_2$), binds O_2 more tightly than adult hemoglobin ($\alpha_2\beta_2$) does. (Both γ-globin and β-globin are members of the β-globin cluster.) This specialized form of hemoglobin ensures that in the placenta, where the maternal and fetal circulation come close to each other, O_2 will be transferred from the mother's blood to the developing child's blood. Just before birth, the synthesis of fetal hemoglobin in the liver stops, and the bone marrow cells take over, making the adult form. Thus hemoglobins with different binding affinities for O_2 are provided at different stages of human development.

In addition to genes that encode proteins, the globin family includes nonfunctional pseudogenes, designated with the Greek letter psi (ψ). These pseudogenes are the "black sheep" of any gene family: they result from mutations that cause a loss of function, rather than an enhanced or new function. The DNA sequence of a pseudogene may not differ vastly from that of other family members. It may simply lack a promoter, for example, and thus fail to be transcribed. Or it may lack the recognition sites needed for the removal of introns (a process we will describe in the next section) and thus be transcribed into pre-mRNA, but not correctly processed into a useful mature mRNA. In some gene families, pseudogenes outnumber functional genes. Because some members of the family are functional, there appears to be little selective pressure for evolution to eliminate pseudogenes.

RNA Processing

As we saw in the previous section, eukaryotic protein-coding genes contain some sequences that do not appear in the mature mRNA that is translated into proteins. To produce the mature mRNA, the primary transcript (pre-mRNA) is processed in several ways: introns are removed, exons are joined, and bases are added at both ends.



14.9 Processing the Ends of Eukaryotic Pre-mRNA The modifications at both ends—the G cap and the poly A tail—are important for mRNA function.

The primary transcript of a protein-coding gene is modified at both ends

Two early steps in the processing of pre-mRNA take place in the nucleus, one at each end of the molecule (Figure 14.9):

- A **G cap** is added to the 5′ end of the pre-mRNA as it is transcribed. The G cap is a chemically modified molecule of guanosine triphosphate (GTP). It apparently facilitates the binding of mRNA to the ribosome for translation and protects the mRNA from being digested by ribonucleases that break down RNAs.
- A **poly A tail** is added to the 3′ end of pre-mRNA at the end of transcription. Near the 3′ end of pre-mRNA, and after the last codon, is the sequence AAUAAA. This sequence acts as a signal for an enzyme to cut the pre-mRNA. Immediately after this cleavage, another enzyme adds 100 to 300 residues of adenine ("poly A") to the 3′ end of the pre-mRNA. This "tail" may assist in the export of the mRNA from the nucleus and is important for mRNA stability.

Splicing removes introns from the primary transcript

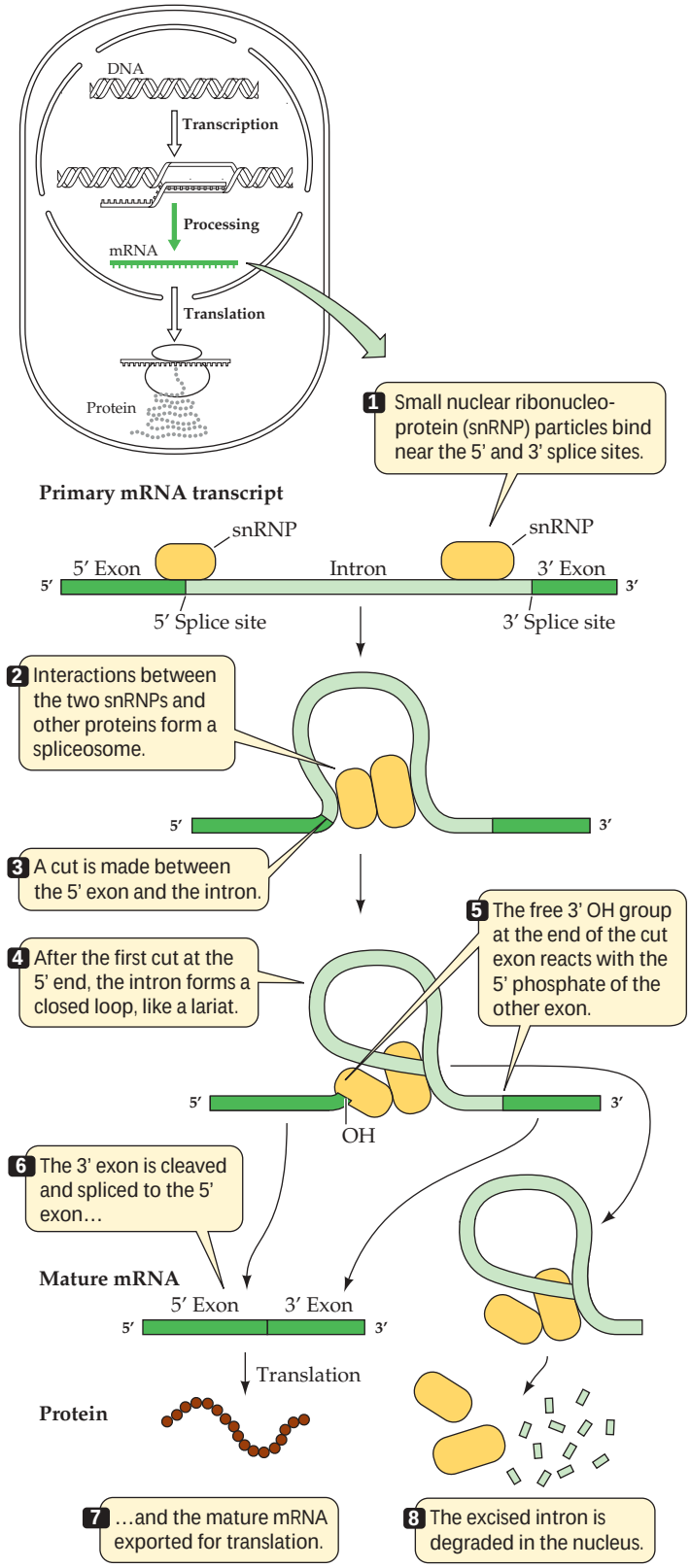
The next step in the processing of eukaryotic pre-mRNA within the nucleus is deletion of the introns. If these RNA regions were not removed, an mRNA producing a very different amino acid sequence, and possibly a nonfunctional protein, would result. A process called **RNA splicing** removes the introns and splices the exons together.

As soon as the pre-mRNA is transcribed, it is quickly bound by several **small nuclear ribonucleoprotein particles** (**snRNPs**, commonly pronounced "snurps"). There are several types of these RNA–protein particles in the nucleus.

At the boundaries between introns and exons are **consensus sequences**—short stretches of DNA that appear, with little variation ("consensus"), in many different genes. The RNA in one of the snRNPs (called U1) has a stretch of bases complementary to the consensus sequence at the 5′ exon–intron boundary, and it binds to the pre-mRNA by complementary base pairing. Another snRNP (U2) binds to the pre-mRNA near the 3′ intron–exon boundary (Figure 14.10).

Next, using energy from ATP, proteins assemble, forming a large RNA–protein complex called a **spliceosome**. This complex cuts the RNA, releases the introns, and joins the ends of the exons together to produce mature mRNA.

Molecular studies of human genetic diseases have been valuable tools in the investigation of consensus sequences and splicing machinery. People with beta thalassemia, for example, make an inadequate amount of the β-globin subunit of hemoglobin. These people suffer from severe anemia because they have an inadequate supply of red blood cells. In some cases, the genetic mutation that causes the disease oc-



14.10 The Spliceosome, an RNA Splicing Machine The binding of two snRNPs to consensus sequences on the pre-mRNA lines up the splicing machinery. After the snRNPs bind to the pre-mRNA, other proteins join the complex to form a spliceosome. This mechanism determines the exact position of each cut in the primary transcript with great precision.

curs at a consensus sequence in the β-globin gene. Consequently, β-globin pre-mRNA cannot be spliced correctly, and nonfunctional β-globin mRNA is made.

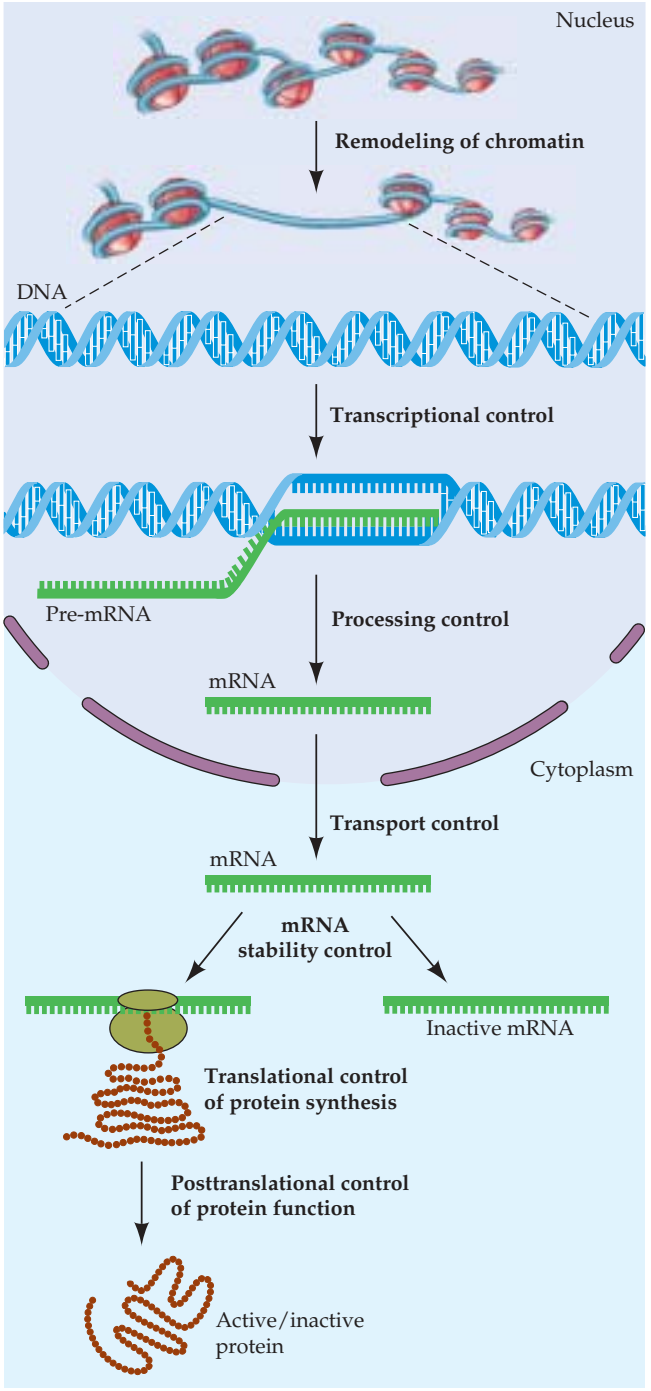
This finding is an excellent example of the use of mutations in determining a cause-and-effect relationship in biology. In the logic of science, merely linking two phenomena (for example, consensus sequences and splicing) does not prove that one is necessary for the other. In an experiment, the scientist alters one phenomenon (for example, the bases of the consensus sequence) to see whether the other (for example, splicing) occurs. In beta thalassemia, nature has done this experiment for us.

After processing is completed in the nucleus, the mature mRNA exits the organelle, apparently through the nuclear pores. A receptor at the nuclear pore recognizes the mature mRNA (or a protein bound to it). Unprocessed or incompletely processed pre-mRNAs remain in the nucleus.

Transcriptional Regulation of Gene Expression

In a multicellular organism with specialized cells and tissues, each cell contains every gene in the organism's genome. For development to proceed normally, and for each cell to acquire and maintain its proper specialized function, certain proteins must be synthesized at just the right times and in just the right cells. Thus, the expression of eukaryotic genes must be precisely regulated. Unlike DNA replication, which is generally regulated in every cell on an all-or-none basis, gene expression is highly selective.

Gene expression can be regulated at several points (Figure 14.11): before transcription, during transcription, after transcription and before translation, during translation, or after translation. In this section, we will describe the mechanisms that result in the selective transcription of specific genes. Some of these mechanisms involve nuclear proteins that alter chromosome function or structure. In other cases, the regulation of transcription involves changes in the DNA itself: genes are selectively replicated to provide more templates for transcription, or even rearranged on the chromosome.



14.11 Potential Points for the Regulation of Gene Expression in Eukaryotes Gene expression can be regulated at four points: at transcription, after transcription (but before translation), at translation, or after translation.

Specific genes can be selectively transcribed

The brain cells and the liver cells of a mouse have some proteins in common and others that are characteristic of each cell type. Yet both cells have the same DNA sequences and, therefore, the same genes. Are the differences in protein content due to differential transcription of the genes? Or is it the case that all the genes are transcribed in both cell types, and some mechanism that acts after transcription is responsible for the differences in proteins?

These two alternatives—**transcriptional regulation** and **posttranscriptional regulation**—can be distinguished by examining the actual RNA sequences made within the nucleus of each cell type. Such analyses indicate that for some proteins, the mechanism of regulation is differential gene transcription. Both brain and liver cells, for example, transcribe "housekeeping" genes—those that encode proteins involved in the basic metabolic processes that occur in every living cell, such as glycolysis enzymes. But liver cells transcribe some genes for liver-specific proteins, and brain cells transcribe

some genes for brain-specific proteins. And neither cell type transcribes the genes for proteins that are characteristic of muscle, blood, bone, or the other specialized cell types in the body.

CONTRASTING EUKARYOTES AND PROKARYOTES. Unlike prokaryotes, in which related genes are grouped into operons that are transcribed as a unit, eukaryotes tend to have solitary genes. Thus, the regulation of several genes at once requires common control elements in each of the genes, which allow all of the genes to respond to the same signal.

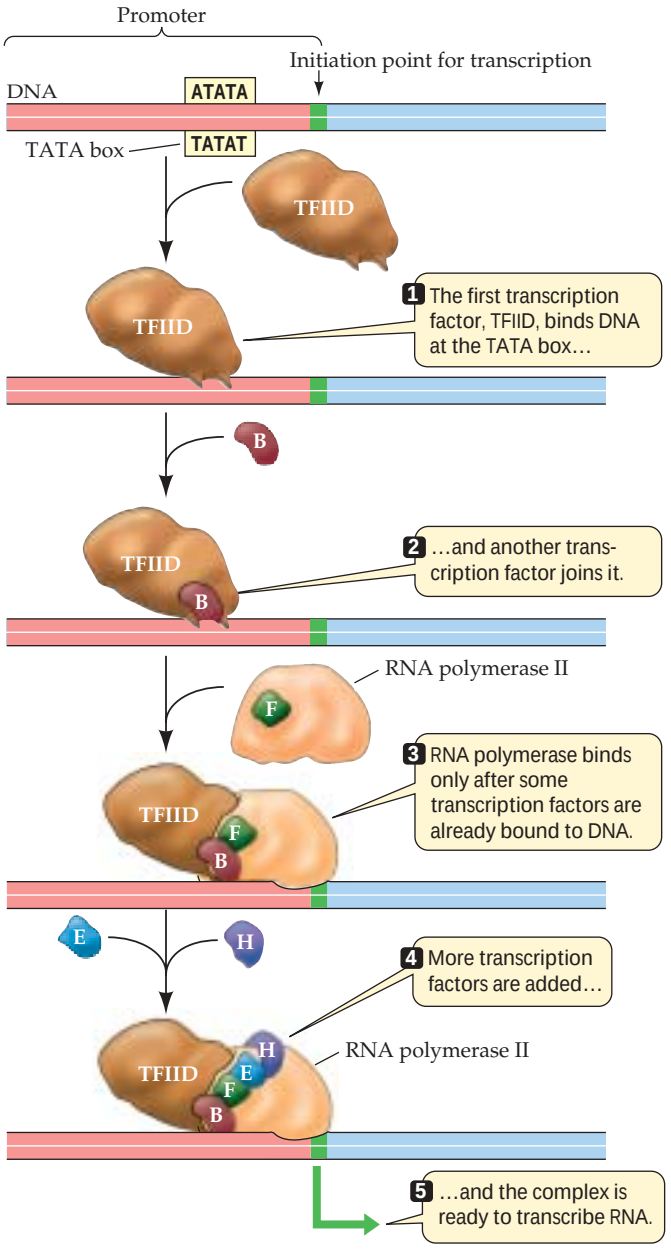
In contrast to the single RNA polymerase in bacteria, eukaryotes have three different RNA polymerases. Each eukaryotic polymerase catalyzes the transcription of a specific type of gene. Only one (RNA polymerase II) transcribes protein-coding genes. The other two transcribe the DNA that codes for rRNA (polymerase I) and for tRNA and small nuclear RNAs (polymerase III).

The diversity of eukaryotic polymerases is reflected in the diversity of eukaryotic promoters, which tend to be much more varied in their sequences than prokaryotic promoters. Furthermore, most eukaryotic genes have additional sequences that can regulate the rate of their transcription. Whether a eukaryotic gene is transcribed depends on the sum total of the effects of all of these DNA and protein elements; thus there are many points of possible regulation.

Finally, the transcription complex in eukaryotes is very different from that of prokaryotes, in which a single peptide subunit can cause RNA polymerase to recognize the promoter. In eukaryotes, many proteins are involved in initiating transcription. We will confine the following discussion to RNA polymerase II, which catalyzes the transcription of most protein-coding genes, but the mechanisms for the other two polymerases are similar.

TRANSCRIPTION FACTORS. As we saw in Chapter 13, the prokaryotic promoter is a sequence of DNA near the 5′ end of the coding region of a gene or operon where RNA polymerase begins transcription. A prokaryotic promoter has two essential sequences. One is the recognition sequence—the sequence recognized by RNA polymerase. The second, closer to the initiation site, is the **TATA box** (so called because it is rich in AT base pairs), where DNA begins to denature so that the template strand can be exposed.

Things are different in eukaryotes. Eukaryotic RNA polymerase II cannot simply bind to the promoter and initiate transcription. Rather, it does so only after various regulatory proteins, called **transcription factors**, have assembled on the chromosome (Figure 14.12). First, the protein TFIID ("TF" stands for transcription factor) binds to the TATA box. Its binding changes both its own shape and that of the DNA, presenting a new surface that attracts the binding of other



14.12 The Initiation of Transcription in Eukaryotes
Except for TFIID, which binds to the TATA box, each transcription factor in this transcription complex has binding sites only for the other proteins in the complex, and does not bind directly to DNA.

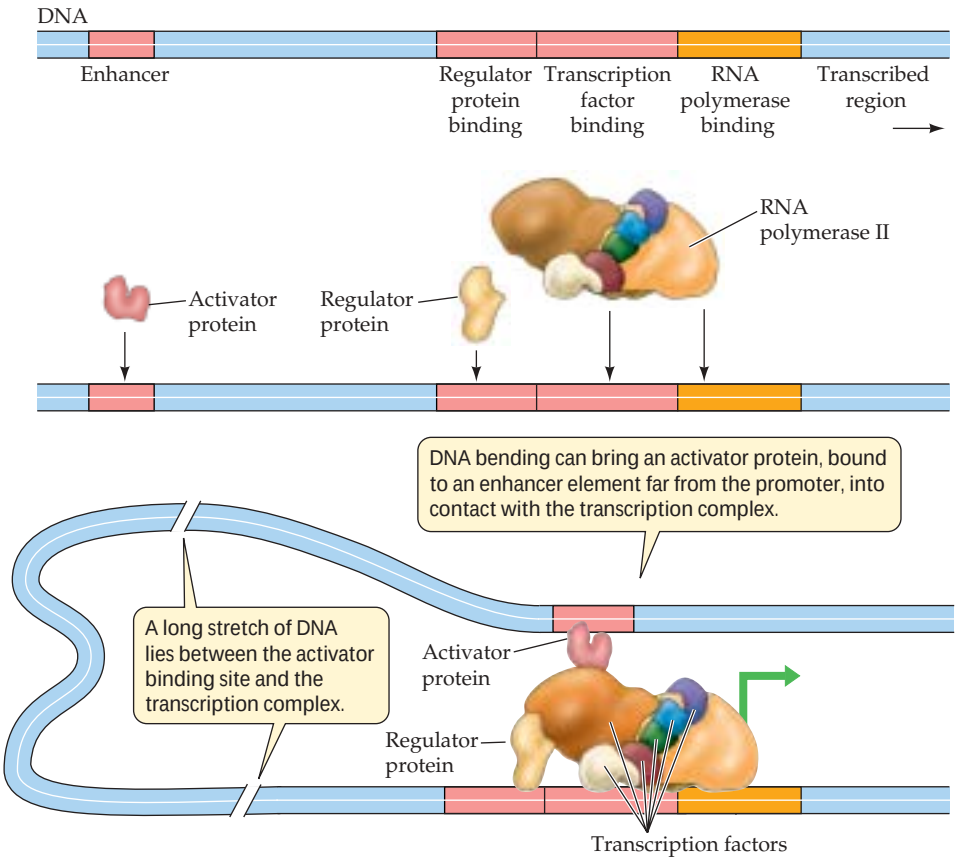
transcription factors to form a **transcription complex**. RNA polymerase II does not bind until several other proteins have bound to this complex.

Some DNA sequences, such as the TATA box, are common to the promoters of many eukaryotic genes and are recognized by transcription factors that are found in all the cells of an organism. Other sequences found in promoters are specific to only a few genes and are recognized by transcription factors found only in certain tissues. These specific tran-

scription factors play an important role in *differentiation*, the specialization of cells during development.

REGULATORS, ENHANCERS, AND SILENCERS IN DNA. In addition to the promoter, two other types of regulatory DNA sequences bind proteins that activate RNA polymerase. The recently discovered **regulator sequences** are clustered just upstream of the promoter. Various *regulator proteins* (seven for the β-globin gene) may bind to these regulator sequences (Figure 14.13). The resulting complexes bind to the adjacent transcription complex and activate it.

Much farther away—up to 20,000 bp away—from the promoter are **enhancer sequences**. Enhancer sequences bind *activator proteins*, and this binding strongly stimulates the transcription complex. How enhancers exert their influence is not clear. In one proposed model, the DNA bends (it is known to do so) so that the activator protein is in contact with the transcription complex.



14.13 The Roles of Transcription Factors, Regulators, and Activators The actions of many proteins determine whether and where RNA polymerase II will transcribe DNA.

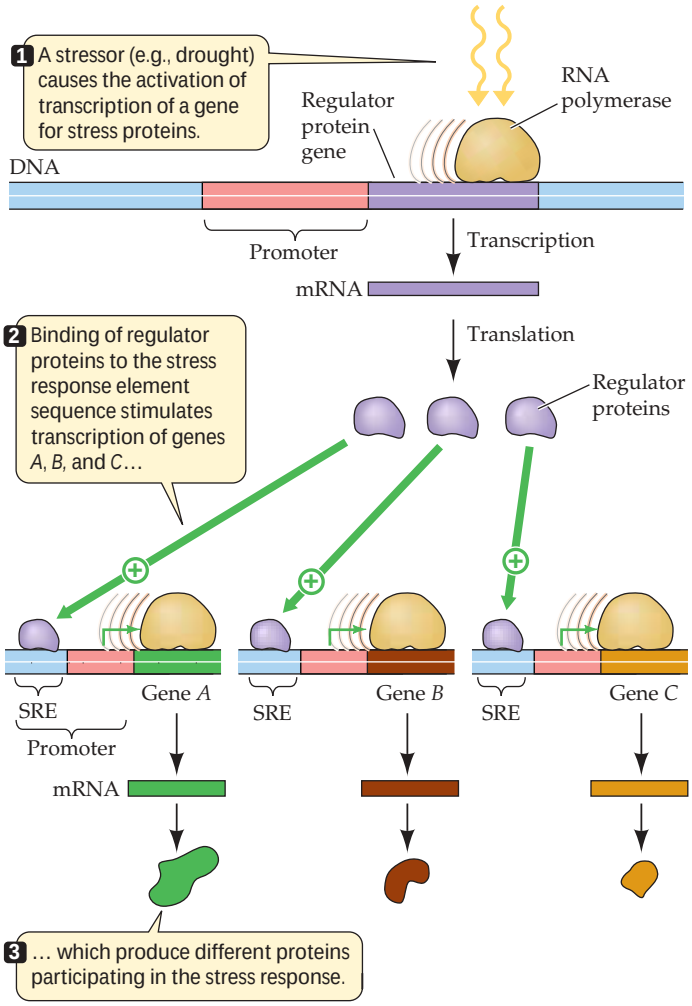
Finally, there are negative regulatory sequences on DNA, called **silencer sequences**, that have the opposite effect from enhancers. Silencers turn off transcription by binding proteins appropriately called *repressor proteins*.

How do these proteins and DNA sequences—transcription factors, regulators, enhancers, activators, silencers, and repressors—regulate transcription? Apparently, in most tissues, a small amount of RNA is transcribed from all genes. But the right combination of these factors determines the rate of transcription. In the immature red blood cells of bone marrow, for example, which make a large amount of β-globin, transcription of the β-globin gene is stimulated by the binding of seven regulator proteins and six activator proteins. But in white blood cells in the same bone marrow, these thirteen proteins are not made and do not bind to the regulator and enhancer sequences adjacent to the β-globin gene; consequently, the gene is hardly transcribed at all.

COORDINATING THE EXPRESSION OF GENES. How do eukaryotic cells coordinate the regulation of several genes whose transcription must be turned on at the same time? In prokaryotes, in which related genes are linked together in an operon, a single regulatory system can regulate several adjacent genes. But in eukaryotes, the several genes whose regulation requires coordination may be far apart on a chromosome, or even on different chromosomes.

In such a case, regulation can be achieved if the various genes all have the same regulatory sequences, which bind the same regulatory proteins. One of the many examples of this phenomenon is provided by the response of organisms to a stressor—for example, that of plants to drought. Under conditions of drought stress, a plant must synthesize various proteins, but the genes for these proteins are scattered throughout the genome. However, each of these genes has a specific regulatory sequence near its promoter, called the *stress response element* (*SRE*). The binding of a regulator protein to this element stimulates RNA synthesis (Figure 14.14). The proteins made from these genes are involved not only in water conservation, but also in protecting the plant against excess salt in the soil and against freezing. This finding has considerable importance for agriculture, in which crops are often grown under less than optimal conditions.

The regulation and coordination of gene expression requires the binding of many specialized proteins to DNA. Among DNA-binding proteins, there are four common structural themes in the domains that bind to DNA. These themes, called **motifs**, consist of combinations of structures and special components: helix-turn-helix, zinc finger, leucine zipper, and helix-loop-helix (Figure 14.15). DNA-binding proteins with specific motifs are involved in the activation



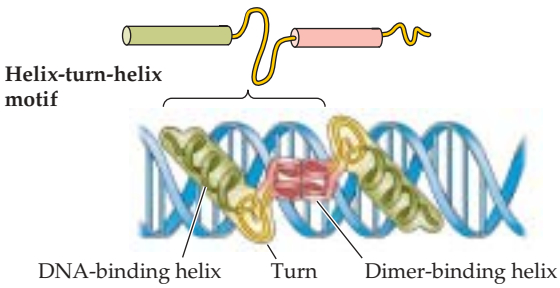
14.14 Coordinating Gene Expression A single signal, such as drought stress, causes the synthesis of a transcriptional regulator for many genes.

of certain types of genes, both during development and in the adult organism.

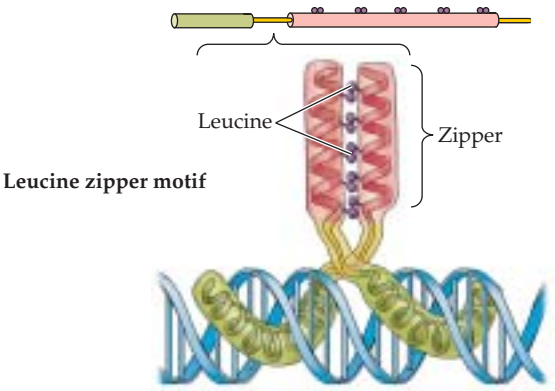
Genes can be inactivated by chromatin structure

Other mechanisms that regulate transcription act on the structure of chromatin and chromosomes. As we saw in Chapter 9, chromatin contains a number of proteins as well as DNA. The packaging of DNA into nucleosomes by these nuclear proteins can make DNA physically inaccessible to RNA polymerase and the rest of the transcription apparatus, much as the binding of a repressor to the operator in the prokaryotic *lac* operon prevents transcription. Chromatin structure at both the local and whole-chromosome levels affects transcription.

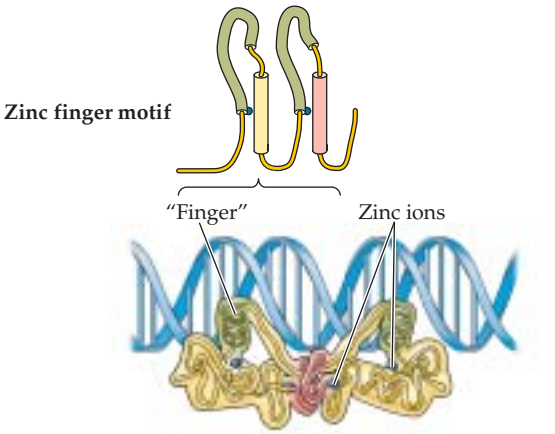
14.15 Protein-DNA Interactions The DNA-binding domains of most regulatory proteins have one of these four structural motifs.



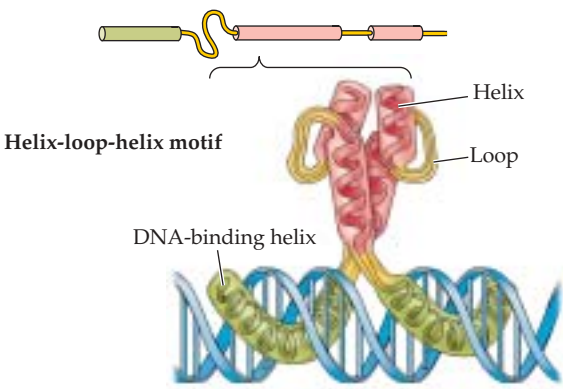
These proteins regulate genes involved in development.



These proteins regulate cell division genes.



These proteins are steroid hormone receptors.



These proteins regulate immune system genes.

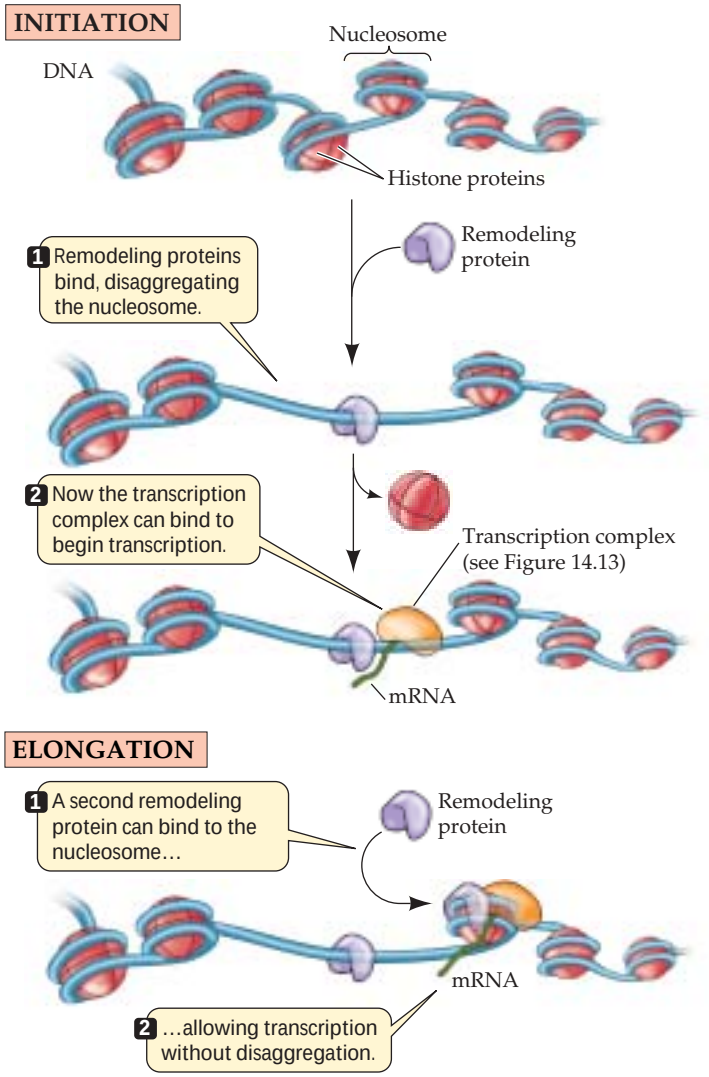
CHROMATIN REMODELING. Nucleosomes inhibit both the initiation and elongation steps of transcription. To inactivate these blocks, two types of protein complexes bind to chromatin. One of them binds upstream of the initiation site, disaggregating the nucleosomes so that the large transcription complex can bind and begin transcription. The other type of complex binds once transcription is under way, allowing the transcription complex to move through the nucleosomes. These processes are referred to as **chromatin remodeling** (Figure 14.16).

How do the nucleosomes disaggregate to allow transcription (and then reaggregate)? As you will recall, the histone proteins that make up nucleosomes are positively charged, and DNA is negatively charged (owing to its phosphate groups), so the attachment of these two molecules is electrostatic. If the histones are modified to reduce their charge, they will release the DNA. One such modification neutralizes the amino groups of the histones by adding acetyl groups. This *acetylation*, catalyzed by histone acetylase, helps disaggregate nucleosomes. Conversely, *deacetylation*, catalyzed by histone deacetylase, allows nucleosomes to reform. Thus, acetylation is associated with the activation of genes; deacetylation is associated with gene deactivation.

WHOLE-CHROMOSOME EFFECTS. Other transcriptional regulation mechanisms can act on entire chromosomes. Under a microscope, two kinds of chromatin can be distinguished in the stained interphase nucleus: euchromatin and heterochromatin. *Euchromatin* is diffuse and stains lightly; it contains the DNA that is transcribed into mRNA. *Heterochromatin* stains densely and is generally not transcribed; any genes that it contains are thus inactivated.

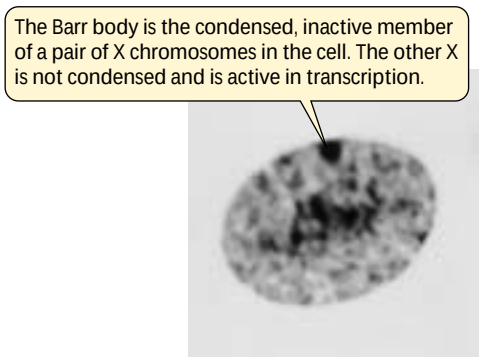
Perhaps the most dramatic example of heterochromatin is the inactive X chromosome of mammals. A normal female mammal has two X chromosomes; a normal male has an X and a Y. The Y chromosome has only a few genes that are also present on the X, and it is largely transcriptionally inactive in most cells. So there is a great difference between females and males in the "dosage" of X-linked genes. In other words, each female cell has two copies of the genes on the X chromosome, and therefore has the potential to produce twice as much protein product from these genes as a male cell has. Yet X-linked gene expression is generally the same in males and females. How can this happen?

The answer was found in 1961 independently by Mary Lyon, Liane Russell, and Ernest Beutler. They suggested that one of the X chromosomes in each cell of an XX female is transcriptionally inactivated early in embryonic development. That copy of the X remains inactive in that cell, and in all the cells arising from it. In a given mammalian cell, the "choice" of which X in the pair of X's to inactivate is random. Recall that one X in a female comes from her father and one from her mother. Thus, in one embryonic cell, the paternal X might be the one remaining transcriptionally active, but in a neighboring cell, the maternal X might be active.



14.16 Local Remodeling of Chromatin for Transcription
Initiation of transcription requires that nucleosomes disaggregate. During elongation, however, they can remain intact.

In the interphase cells of human females, a single, stainable nuclear body, called a **Barr body** after its discoverer, Murray Barr, can be seen under the light microscope (Figure 14.17). This clump of heterochromatin, which is not present in males, is the inactivated X chromosome. The number of Barr bodies in a nucleus is equal to the number of X chromosomes minus one (the one represents the X chromosome that remains transcriptionally active). So a female with the normal two X chromosomes will have one Barr body, a rare female with three X's will have two, an XXXX female will have three, and an XXY male will have one. These observations suggest that the interphase cells of each person, male or female, have a single active X chromosome, making the dosage of the expressed X chromosome genes constant across both sexes.



14.17 A Barr Body in the Nucleus of a Female Cell The number of Barr bodies per nucleus is equal to the number of X chromosomes minus one. Thus normal males (XY) have no Barr body, whereas normal females (XX) have one.

Chromosomal condensation in the inactive X chromosome makes its DNA sequences physically unavailable to the transcriptional machinery. One mechanism of inactivation is the addition of a methyl group (—CH_3) to the 5′ position of cytosine on DNA. Such **methylation** seems to be most prevalent in transcriptionally inactive genes. For example, most of the DNA of the inactive X chromosome has many of its cytosines methylated, while few cytosines on the active X are methylated. Methylated DNA appears to bind certain chromosomal proteins that may be responsible for heterochromatin formation.

The otherwise inactive X chromosome has one gene that is only lightly methylated and is transcriptionally active. That gene is called *Xist* (for *X i*nactivation-specific *t*ranscript), and it is heavily methylated on, and not transcribed from, the other, "active" X chromosome. The RNA transcribed from *Xist* does not leave the nucleus and is not an mRNA. Instead, it appears to bind to the X chromosome from which it is transcribed, and this binding somehow leads to a spreading of inactivation along the chromosome. This RNA transcript is known as **interference RNA** (**RNAi**) (Figure 14.18).

How does the transcriptionally active X overcome the effects of *Xist* RNAi? Apparently, there is an anti-*Xist* gene, appropriately called *Tsix*. This gene codes for an RNAi that binds by complementary base pairing to *Xist* RNA at the active X chromosome.

A DNA sequence can be moved to a new location to activate transcription

In some instances, gene expression is regulated by the movement of a gene to a new location on the chromosome. An example of this mechanism is found in the budding yeast, *Saccharomyces cerevisiae*. This haploid, single-celled fungus exists in two *mating types*, *a* and α. Two cells of different mating types fuse to form a diploid zygote, as we will see in Chapter 31. Although all yeast cells have an allele for each mating type, the allele that is expressed determines the mating type of the cell. In some yeasts, the mating type changes with almost every cell division cycle. How does it change so rapidly?

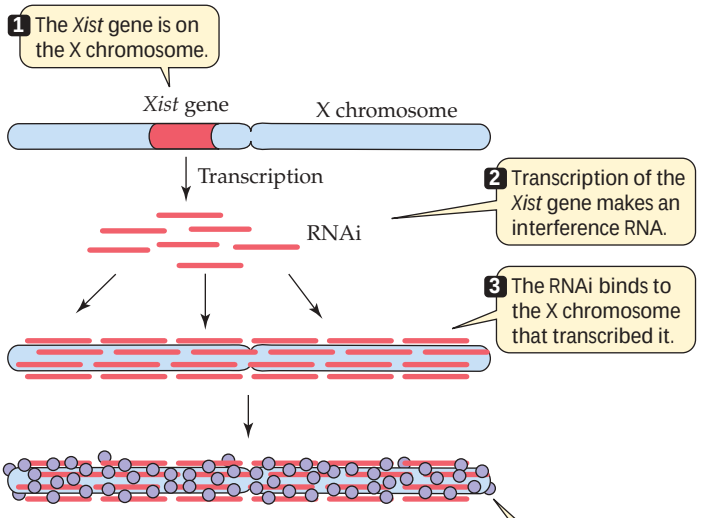
In the yeast cell, the two mating type alleles (coding for type α and type *a*) have separate, specific locations on the chromosome, away from a third site, called the MAT locus. One allele is transcriptionally silent because a repressor protein is bound to it. However, when a copy of the α or *a* allele is inserted at the MAT locus, the gene for the protein of that mating type is transcribed. A change in mating type requires that one allele be moved out of the MAT locus and the other moved in. This process takes place in three steps:

- First, a new DNA copy of the nonexpressed allele is made (if the cell is now α, a new copy of the *a* allele will be made).
- Second, the current occupant of the MAT locus (in this case, the α allele) is removed by an enzyme.
- Third, the new allele (*a*) is inserted at the MAT locus and transcribed. The *a* proteins are now made, and the mating type of the cell is *a*.

DNA rearrangement is important in producing the highly variable proteins that make up the human repertoire of antibodies. It is also a factor in cancer, in which inactive genes may be moved to positions adjacent to active promoters.

Selective gene amplification results in more templates for transcription

Another way for one cell to make more of a certain gene product than another cell does is to make more copies of the appropriate gene and transcribe them all. The process of cre-



14.18 A Model for X Chromosome Inactivation Interference RNA (RNAi) and chromosomal proteins combine to inactivate the X chromosome.

ating more copies of a gene in order to increase its transcription is called **gene amplification**.

As described earlier, the genes that code for three of the four human ribosomal RNAs are linked together in a unit, and this unit is repeated several hundred times in the genome to provide multiple templates for rRNA synthesis (rRNA is the most abundant kind of RNA in the cell). In some circumstances, however, even this moderate repetition is not enough to satisfy the demands of the cell.

The mature eggs of frogs and fishes, for example, have up to a trillion ribosomes. These ribosomes are used for the massive protein synthesis that follows fertilization. The cell that will differentiate into the egg contains fewer than 1,000 copies of the rRNA gene cluster, and would take 50 years to make a trillion ribosomes if it transcribed those rRNA genes at peak efficiency. How does the egg end up with so many ribosomes (and so much rRNA)?

The egg cell solves this problem by selectively amplifying its rRNA gene clusters until there are more than a million copies. In fact, this gene complex goes from being 0.2 percent of the total genome DNA to 68 percent. These million copies, transcribed at maximum rate (Figure 14.19), are just enough to make the necessary trillion ribosomes in a few days.

The mechanism for selective amplification of a single gene is not clearly understood, but it has important medical implications. In some cancers, a cancer-causing gene called an oncogene becomes amplified (see Chapter 17). Also, when some tumors are treated with a drug that targets a single protein, amplification of the gene for the target protein leads to an excess of that protein, and the cell becomes resistant to the prescribed dose of the drug.

Posttranscriptional Regulation

There are many ways in which gene expression can be regulated even after the gene has been transcribed. As we saw earlier, pre-mRNA is processed by cutting out the introns and splicing the exons together. If exons are selectively deleted from the pre-mRNA by alternative splicing, different proteins can be synthesized. The longevity of mRNA in the cytoplasm can also be regulated. The longer an mRNA exists in the cytoplasm, the more of its protein can be made.

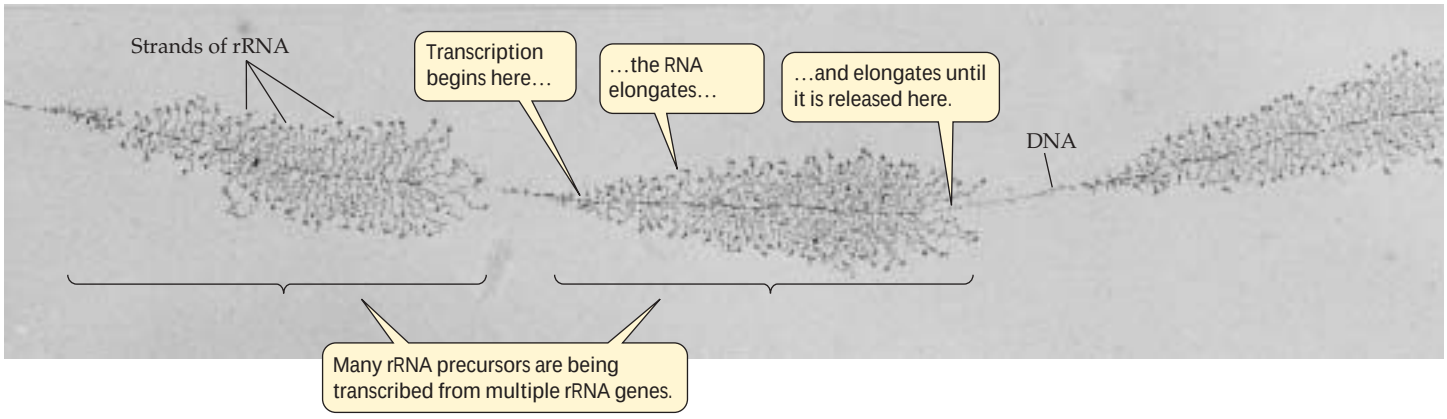
Different mRNAs can be made from the same gene by alternative splicing

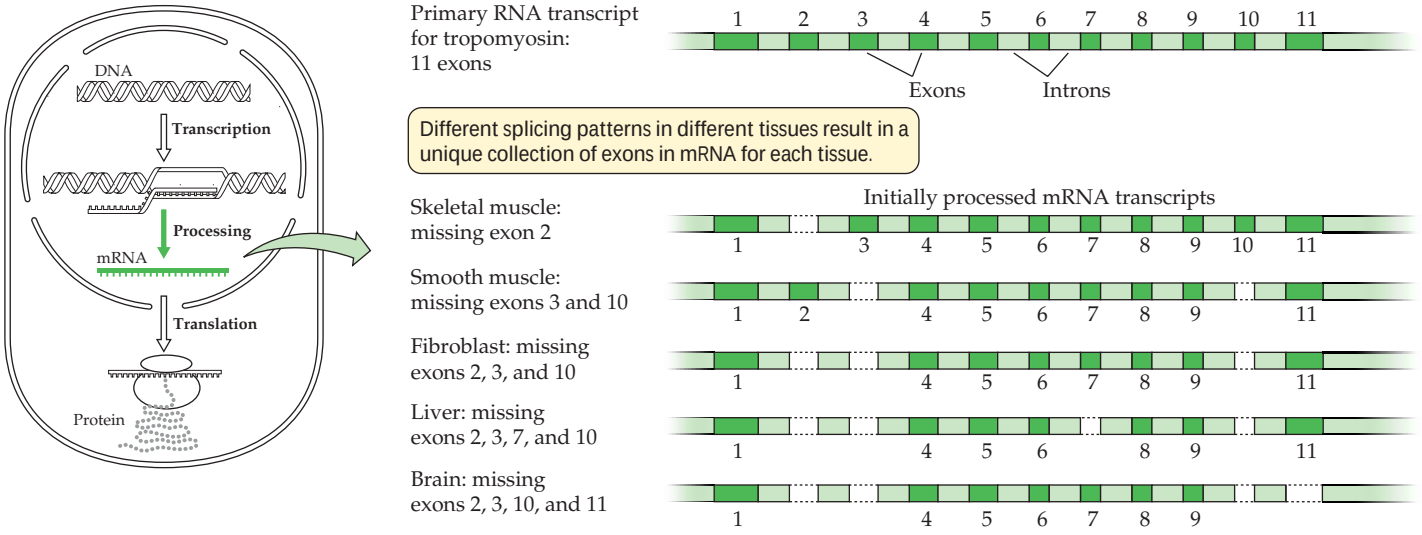
Most primary mRNA transcripts contain several introns (see Figure 14.4). We have seen how the splicing mechanism recognizes the boundaries between exons and introns. What would happen if the β-globin pre-mRNA, which has two introns, were spliced from the start of the first intron to the end of the second? Not only the two introns, but also the middle exon, would be spliced out. An entirely new protein (certainly not a β-globin) would be made, and the functions of normal β-globin would be lost.

Alternative splicing can be a deliberate mechanism for generating a family of different proteins from a single gene. For example, a single pre-mRNA for the structural protein tropomyosin is spliced differently in five different tissues to give five different mature mRNAs. These mRNAs are translated into the five different forms of tropomyosin found in these tissues: skeletal muscle, smooth muscle, fibroblast, liver, and brain (Figure 14.20).

Before the sequencing of the human genome began, most scientists estimated that they would find between 100,000 and 150,000 genes. You can imagine their surprise when the actual sequence revealed only 21,000 genes—not many more than *C. elegans* has! In fact, there are many more human mRNAs than there are human genes, and most of this variation comes from alternative splicing. Indeed, recent surveys show that half of all human genes are alternatively spliced. Alternative splicing may be a key to the differences in levels of complexity among organisms.

14.19 Transcription from Multiple Genes for rRNA Elongating strands of rRNA transcripts form arrowhead-shaped regions, each centered on a DNA sequence that codes for rRNA.





14.20 Alternative Splicing Results in Different mRNAs and Proteins In mammals, the protein tropomyosin is encoded by a gene that has 11 exons. Tropomyosin pre-mRNA is spliced differently in different tissues, resulting in five different forms of the protein.

The stability of mRNA can be regulated

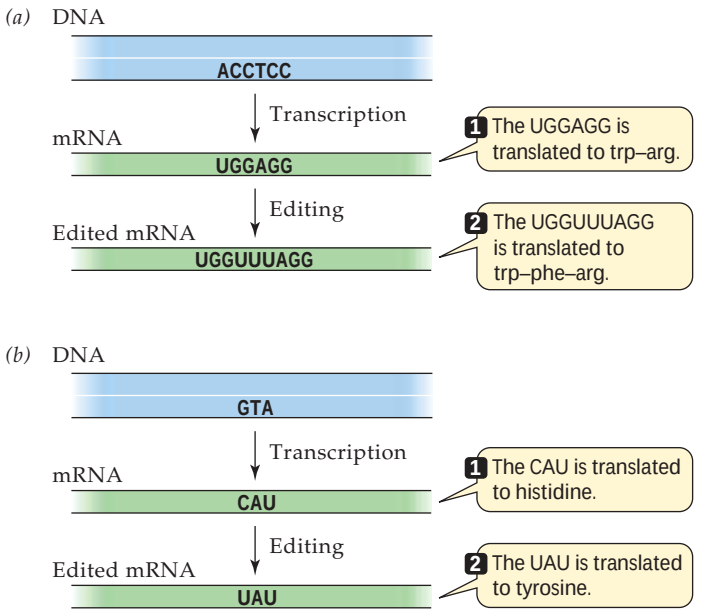
DNA, as the genetic material, must remain stable, and as we have seen, there are elaborate mechanisms for repairing DNA if it becomes damaged. RNA has no such repair mechanism. After it arrives in the cytoplasm, mRNA is subject to breakdown catalyzed by ribonucleases, which exist both in the cytoplasm and in lysosomes. But not all eukaryotic mRNAs have the same life span. Differences in the stabilities of mRNAs provide another mechanism for posttranscriptional regulation of protein synthesis. The less time an mRNA spends in the cytoplasm, the less of its protein can be translated.

Specific AU-rich nucleotide sequences within some mRNAs mark them for rapid breakdown by a ribonuclease complex called the *exosome*. Signaling molecules such as growth factors, for example, are made only when needed, and then break down rapidly. Their mRNAs are highly unstable because they contain an AU-rich sequence.

RNA can be edited to change the encoded protein

The sequence of mRNA can be changed after transcription and splicing by **RNA editing**. This editing can occur in two ways (Figure 14.21):

- *Insertion of nucleotides.* In the parasitic protozoan *Trypanosoma brucei*, certain mRNAs have been found that have a longer base sequence than predicted by the gene coding for them. Stretches of U's are added after transcription, changing the protein that is made.
- *Alteration of nucleotides.* An enzyme can catalyze the deamination of cytosine, forming uracil (see Figure 12.19). This process can affect a membrane channel protein in the mammalian nervous system that normally allows calcium and sodium to pass through. Editing of a certain cytosine in the mRNA for this protein to uracil changes the amino acid at that position in the polypeptide chain from histidine to tyrosine, and the channel protein no longer allows the passage of calcium.



14.21 RNA Editing RNA can be edited in two ways: (*a*) by the insertion of new nucleotides, or (*b*) by the alteration of existing nucleotides.

Translational and Posttranslational Regulation

Is the amount of a protein in a cell determined by the amount of its mRNA? Recently, a survey of the relationships between mRNAs and proteins in yeast cells was made. Dozens of genes were surveyed. For about a third of them, the relationship between mRNA and protein held: more of one led to more of the other. But for two-thirds of the proteins, there was no apparent relationship. The concentrations of these proteins in the cell must be determined by factors acting after the mRNA is made.

Just as certain regulatory proteins can control the synthesis of mRNA by binding to DNA, other proteins can regulate the translation of mRNA by binding to mRNA in the cytoplasm. This mode of control is especially important for long-lived mRNAs. A cell must not continue to make proteins that it does not need. For example, as we saw in Chapter 9, mammalian cells respond to certain stimuli by making cyclins, proteins that stimulate the events of the cell cycle. If the mRNA for a cyclin is still in the cytoplasm and available for translation long after the cyclin is needed, the cyclin will be made and released inappropriately. Its presence might cause a target cell population to divide inappropriately, forming a tumor.

The translation of mRNA can be regulated

Let's look at examples of three general mechanisms by which levels of certain proteins are controlled by regulating the translation of mRNA.

One way to regulate translation is through the G cap on mRNA. As we saw above, mRNA is capped at its 5′ end by a modified guanosine triphosphate molecule (see Figure 14.9). An mRNA that is capped with an unmodified GTP molecule is not translated. For example, stored mRNA in the oocyte of the tobacco hornworm moth has a G cap added to its 5′ end, but the GTP molecule is not modified. Hence, this stored mRNA is not translated. After fertilization, however, the cap is modified, allowing the mRNA to be translated to produce the proteins needed for early embryonic development.

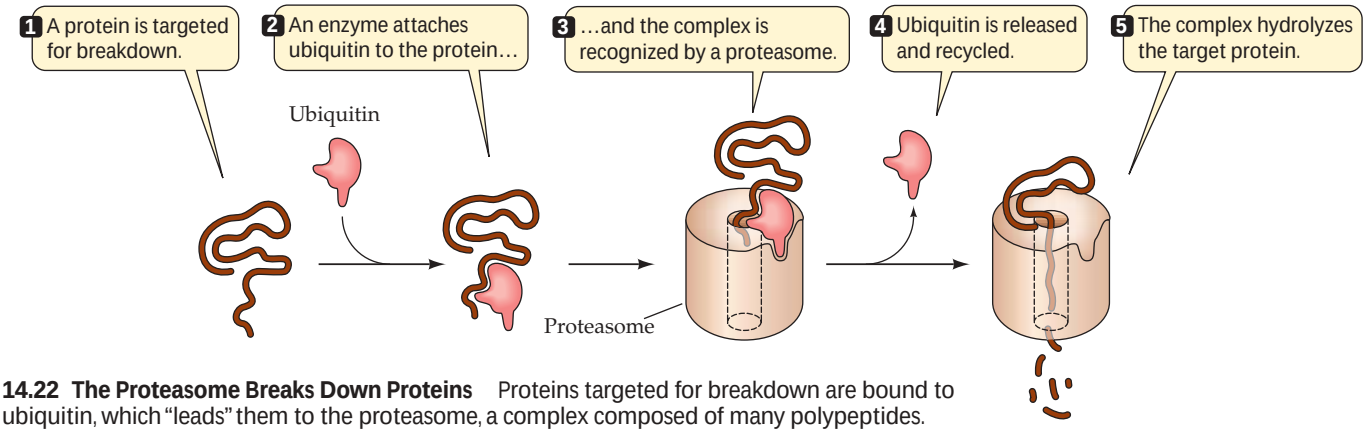
Within mammalian cells, free iron ions (Fe^{2+}) are bound by a storage protein, called ferritin. When iron is present in excess, ferritin synthesis rises dramatically. Yet the amount of ferritin mRNA remains constant. The increase in ferritin synthesis is due to an increased rate of mRNA translation. When the iron level in the cell is low, a translational repressor protein binds to ferritin mRNA and prevents its translation by blocking its attachment to a ribosome. When the iron level rises, the excess iron ions bind to the repressor and alter its three-dimensional structure, causing it to detach from the mRNA, and translation of ferritin proceeds.

Translational regulation also acts on the synthesis of hemoglobin, helping to maintain the balance among its components. As described above, a hemoglobin molecule consists of four globin subunits and four heme pigments. If globin synthesis does not equal heme synthesis, some heme stays free in the cell, waiting for a globin partner. Excess heme increases the rate of translation of globin mRNA by removing a block to the initiation of translation at the ribosome.

The proteasome controls the longevity of proteins after translation

We have considered how gene expression can be regulated by the control of transcription, RNA processing, and translation. However, the story does not end here, because most gene products—proteins—are modified after translation. Some of these changes are permanent, such as the addition of sugars (glycosylation), the addition of phosphate groups, or the removal of a signal sequence after a protein has crossed a membrane (see Figure 12.15).

One way to regulate the action of a protein in a cell is to regulate its lifetime in the cell. Proteins involved in cell division (such as cyclins), for example, are hydrolyzed at just the right moment to time the sequence of events. In many cases,



14.22 The Proteasome Breaks Down Proteins Proteins targeted for breakdown are bound to ubiquitin, which "leads" them to the proteasome, a complex composed of many polypeptides.

a 76-amino acid protein called **ubiquitin** (so called because it is ubiquitous, or widespread) is covalently linked to a protein targeted for breakdown. The protein–ubiquitin complex then binds to a huge complex of several dozen polypeptide chains called a **proteasome** (Figure 14.22). The entryway to this "molecular chamber of doom" is a hollow cylinder. This part of the complex has ATPase activity, and it uses the released energy to cut off the ubiquitin for recycling and unfold its targeted protein "victim." The protein then passes by three different proteases (thus the name of the complex), which digest it into small peptides and amino acids.

The cellular concentrations of many proteins are determined not by differential transcription of their genes, but by their degradation in proteasomes. Cyclins, for example, are degraded at just the right time during the cell cycle (see Figure 9.4). Transcription regulators are broken down after they are used, lest the affected genes be always "on." Abnormal proteins are often targeted for destruction by a quality control mechanism. Human papillomavirus, which causes cervical cancer, targets the cell division inhibitory protein p53 for proteasomal degradation, so that unregulated cell division—cancer—results.

Chapter Summary

The Eukaryotic Genome

- Although eukaryotes have more DNA in their genomes than prokaryotes, there is no apparent relationship between genome size and organism complexity within eukaryotes.
- There are many differences between prokaryotic and eukaryotic genomes and their mechanisms of expression. **Review Table 14.1**
- Unlike prokaryotic DNA, eukaryotic DNA is contained within a nucleus, so that transcription and translation are physically separated. **Review Figure 14.1.** **See Web/CD Activity 14.1**
- The genome of the single-celled budding yeast contains genes for the same metabolic machinery found in prokaryotes, with the addition of genes for protein targeting in the cell. **Review Table 14.2**
- The genome of the multicellular roundworm *Caenorhabditis elegans* contains genes required for intercellular interactions. **Review Table 14.3**
- The genome of the fruit fly has fewer genes than that of the roundworm. Many of its genes are homologs of genes found in the roundworm and mammalian genomes.
- The puffer fish genome is the most compact vertebrate genome known.
- The compact genome of the simple plant *Arabidopsis* is often used in the study of plant genomes. **Review Table 14.4**
- The rice genome is similar to that of *Arabidopsis*, and its sequence holds a key to feeding the increasing human population. **Review Table 14.5**

Repetitive Sequences in the Eukaryotic Genome

- Highly repetitive DNA is present in up to millions of copies of short sequences. It is not transcribed.
- Some moderately repetitive DNA sequences, such as those that code for rRNAs, are transcribed. **Review Figure 14.2**
- Some moderately repetitive DNA sequences are transposons, which are able to move about the genome. **Review Figure 14.3**

The Structures of Protein-Coding Genes

- A typical eukaryotic protein-coding gene is flanked by promoter and terminator sequences and contains noncoding internal sequences, called introns. **Review Figure 14.4**
- Nucleic acid hybridization is an important technique for analyzing eukaryotic genes. **Review Figure 14.5, 14.6**
- Some eukaryotic genes exist as families of related genes, which have similar sequences and code for similar proteins. These related proteins may be made at different times and in different tissues. Some sequences in gene families are pseudogenes, which code for nonfunctional mRNAs or proteins. **Review Figure 14.7**
- Differential expression of different genes in the β-globin cluster of the globin family ensures important physiological changes during human development. **Review Figure 14.8**

RNA Processing

- The transcribed pre-mRNA is altered by the addition of a G cap at the 5′ end and a poly A tail at the 3′ end. **Review Figure 14.9**
- The introns are removed from the mRNA precursor by the spliceosome, a complex of snRNPs and proteins. **Review Figure 14.10.** **See Web/CD Tutorial 14.1**

Transcriptional Regulation of Gene Expression

- Eukaryotic gene expression can be regulated at the transcriptional, posttranscriptional, translational, and posttranslational levels. **Review Figure 14.11.** **See Web/CD Activity 14.2**
- The major method of regulation of eukaryotic gene expression is selective transcription, which results from the binding of specific proteins to regulatory sequences on DNA.
- A series of transcription factors must bind to one another to form a transcription complex before RNA polymerase can bind. Whether RNA polymerase initiates transcription also depends on the binding of regulator proteins, activator proteins (which bind to enhancers and stimulate transcription), and repressor proteins (which bind to silencers and inhibit transcription). **Review Figures 14.12, 14.13.** **See Web/CD Tutorial 14.2**
- The simultaneous regulation of widely separated genes is possible through common sequences in their promoters, to which the same regulatory proteins bind. **Review Figure 14.14**
- The DNA-binding domains of most DNA-binding proteins have one of four structural motifs: helix-turn-helix, zinc finger, leucine zipper, or helix-loop-helix. **Review Figure 14.15**
- Chromatin remodeling allows the transcription complex to bind DNA and to move through the nucleosomes. **Review Figure 14.16**
- Heterochromatin is a condensed form of DNA that cannot be transcribed. It is found in the inactive X chromosome of female mammals. **Review Figure 14.17**
- Interference RNA (RNAi) is important in inhibiting transcription of the inactive X chromosome. **Review Figure 14.18**
- The movement of a gene to a new location on a chromosome may alter its ability to be transcribed, as in the change from one mating type to another in budding yeast.
- Some genes are selectively amplified in some cells. The extra copies of these genes result in increased transcription of their protein product. **Review Figure 14.19**

Posttranscriptional Regulation

- Alternative splicing of pre-mRNA can be used to produce different proteins. The transcripts of over half the genes in the human genome are alternatively spliced, which increases the

number of proteins that can be encoded by a single gene. **Review Figure 14.20**

- The stability of mRNA in the cytoplasm can be regulated.
- Mature mRNA can be edited by the addition of new nucleotides or by the alteration of existing nucleotides. **Review Figure 14.21**

Translational and Posttranslational Regulation

- Translational repressors can inhibit the translation of mRNA.
- Proteasomes degrade proteins targeted for breakdown by attachment of ubiquitin. **Review Figure 14.22**

Self-Quiz

1. Eukaryotic protein-coding genes differ from their prokaryotic counterparts in that only eukaryotic genes
   a. are double-stranded.
   b. are present in only a single copy.
   c. contain introns.
   d. have a promoter.
   e. transcribe mRNA.
2. Comparison of the genomes of yeast and bacteria shows that only yeast has many genes for
   a. energy metabolism.
   b. cell wall synthesis.
   c. intracellular protein targeting.
   d. DNA binding proteins.
   e. RNA polymerase.
3. The genomes of a fruit fly and nematode work are similar to that of yeast, except that the former have many genes for
   a. intercellular signaling.
   b. synthesis of polysaccharides.
   c. cell cycle regulation.
   d. intracellular protein targeting.
   e. transposable elements.
4. Which of the following does *not* occur after mRNA is transcribed?
   a. binding of RNA polymerase II to the promoter
   b. capping of the 5′ end
   c. addition of a poly A tail to the 3′ end
   d. splicing out of the introns
   e. transport to the cytosol
5. Which statement about RNA splicing is *not* true?
   a. It removes introns.
   b. It is performed by small nuclear ribonucleoprotein particles (snRNPs).
   c. It always removes the same introns.
   d. It is usually directed by consensus sequences.
   e. It shortens the RNA molecule.
6. Eukaryotic transposons
   a. always use RNA for replication.
   b. are approximately 50 bp long.
   c. are made up of either DNA or RNA.
   d. do not contain genes coding for transposition.
   e. make up about half of the human genome.
7. Which statement about selective gene transcription in eukaryotes is *not* true?
   a. Different classes of RNA polymerase transcribe different parts of the genome.
   b. Transcription requires transcription factors.
   c. Genes are transcribed in groups called operons.
   d. Both positive and negative regulation occur.
   e. Many proteins bind at the promoter.
8. Heterochromatin
   a. contains more DNA than does euchromatin.
   b. is transcriptionally inactive.
   c. is responsible for all negative transcriptional control.
   d. clumps the X chromosome in human males.
   e. occurs only during mitosis.
9. Translational control
   a. is not observed in eukaryotes.
   b. is a slower form of regulation than transcriptional control.
   c. can be achieved by only one mechanism.
   d. requires that mRNA be uncapped.
   e. ensures that heme synthesis equals globin synthesis.
10. Control of gene expression in eukaryotes includes all of the following *except*
    a. alternative splicing of RNA transcripts.
    b. binding of proteins to DNA.
    c. transcription factors.
    d. feedback inhibition of enzyme activity by allosteric control.
    e. DNA methylation.

For Discussion

1. In rats, a gene 1,440 bp long codes for an enzyme made up of 192 amino acid units. Discuss this apparent discrepancy. How long would the initial and final mRNA transcripts be?
2. The genomes of rice, wheat, and corn are similar to each other and to the weed, *Arabidopsis.* Discuss how these plants might nevertheless have very different proteins.
3. The activity of the enzyme dihydrofolate reductase (DHFR) is high in some tumor cells. This activity makes the cells resistant to the anticancer drug methotrexate, which targets DHFR. Assuming that you had the complementary DNA for the gene that encodes DHFR, how would you show whether this increased activity was due to increased transcription of the single-copy DHFR gene or to amplification of the gene?
4. Describe the steps in the production of a mature, translatable mRNA from a eukaryotic gene that contains introns. Compare this to the situation in prokaryotes (see Chapter 13).
5. A protein-coding gene has three introns. How many different proteins can be made from alternate splicing of the pre-mRNA transcribed from this gene?

12 Cell Signaling and Communication

It's probably happened to you: it's late at night, you have a paper due, and you've put it off until the last minute. You're exhausted, but you need to stay awake and alert so that you can get your work done. What do you do? You have a cup (or several cups) of coffee. Many people turn to the caffeine in coffee when they need to wake themselves up and give themselves an energy boost.

To understand how caffeine works, we must understand the pathways by which the body's cells respond to signals in their environment. There are three sequential steps involved in the cell's response to any signal. First, the signal binds to a receptor protein in the cell, often on the outside surface of the plasma membrane. Second, the binding of the signal causes a message to be conveyed to the inside of the cell and amplified. Third, the cell changes its activity in response to the signal.

Caffeine acts in different ways in different tissues. A tired person's brain produces adenosine molecules that bind to specific receptor proteins, resulting in decreased brain activity and increased drowsiness. Caffeine's molecular structure is similar to that of adenosine, so it occupies the adenosine receptors without inhibiting brain cell function, and alertness is restored. In heart and liver cells, caffeine indirectly stimulates the same signaling pathway that is normally stimulated by epinephrine, the "fight-or-flight" hormone. In the heart, the result is an increased rate of beating; the liver is stimulated to convert glycogen into glucose and release it into the bloodstream.

We begin this chapter with a discussion of the signals that affect cells. As you will see, these signals include chemicals produced by other cells in the body as well as physical factors in the environment, such as light. Whatever the signal, it affects a cell only if that cell has a receptor protein that binds to that signal. In addition to binding the signal, the receptor must somehow communicate to the rest of the cell that binding has occurred. Finally, this communication process, called signal transduction, must result in a change in the function of the cell. We will describe these three steps of cell signaling—binding, transduction, and the cellular response—in order. We close the chapter with a description of how cells communicate with one another directly via specialized channels in their adjacent plasma membranes.

A Signal to the Body The caffeine in coffee sends signals to cells in the body of this coffee drinker. The effects of these signals help him stay alert.

Signals

Both prokaryotic and eukaryotic cells process information from their environment. This information can be in the form of a physical stimulus, such as the light reaching your eye as you read this

book, or chemicals that bathe a cell, such as lactose in the medium surrounding *E. coli*. It may come from outside the organism, such as the scent of a female moth seeking a mate in the dark, or from a neighboring cell within the organism, as in the heart, where thousands of muscle cells contract in unison by transmitting signals to one another.

Of course, the mere presence of a signal does not mean that a cell will respond to it, just as you do not pay close attention to every sound in your environment as you study. To respond, the cell must have a specific receptor protein that can bind to the signal. In the following section, we will describe some of the signals different cells respond to and look at one model signal transduction pathway. After discussing signals, we will consider their receptors.

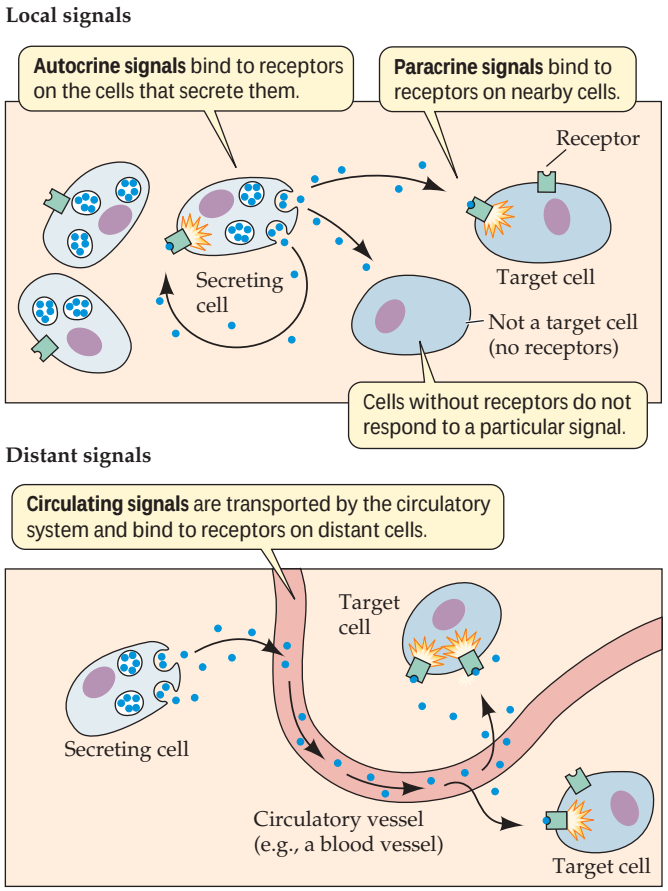
Cells receive signals from the physical environment and from other cells

The physical environment is full of signals. Our sense organs allow us to respond to light, odors and tastes (chemical signals), temperature, touch, and sound. Bacteria and protists respond to even minute chemical changes in their environment. Plants respond to light as a signal. For example, at sunset, at night, or in the shade, not only the amount , but also the wavelengths of the light reaching Earth's surface differ from that of full sunlight in the daytime. These variations act as signals that affect plant growth and reproduction. Some plants also respond to temperature: when the weather gets cold, they respond either by becoming tolerant to cold or by accelerating flowering. Even magnetism can be a signal: some bacteria and birds orient themselves to Earth's magnetic poles, like a needle on a compass.

A cell inside a large multicellular organism is far away from the exterior environment. Instead, its environment consists of other cells and extracellular fluids. Cells receive their nutrients from, and pass their wastes into, extracellular fluids or gases. Cells also receive signals—mostly chemical signals—from their extracellular fluid environment. Most of these chemical signals come from other cells. Cells also respond to chemical signals coming from the environment via the digestive and respiratory systems. And cells can respond to the concentrations of certain chemicals, such as CO_2 and H^+, whose presence in the extracellular fluids results from the metabolic activities of other cells.

Inside a large multicellular organism, chemical signals reach a target cell by local diffusion or by circulation within the blood. **Autocrine** signals are signals that affect the cells that make them. **Paracrine** signals are signals that diffuse to and affect nearby cells. Signals to distant cells, such as hormones, usually travel through the circulatory system (Figure 15.1).

In all cases, the cell must be able to receive or sense the signal and respond to it. Depending on the cell and the signal, the responses range from entering the cell division cycle to heal a wound, to moving to a new location in the embryo to form a tissue, to releasing enzymes that digest food, to sending messages to the brain about the book you are reading.



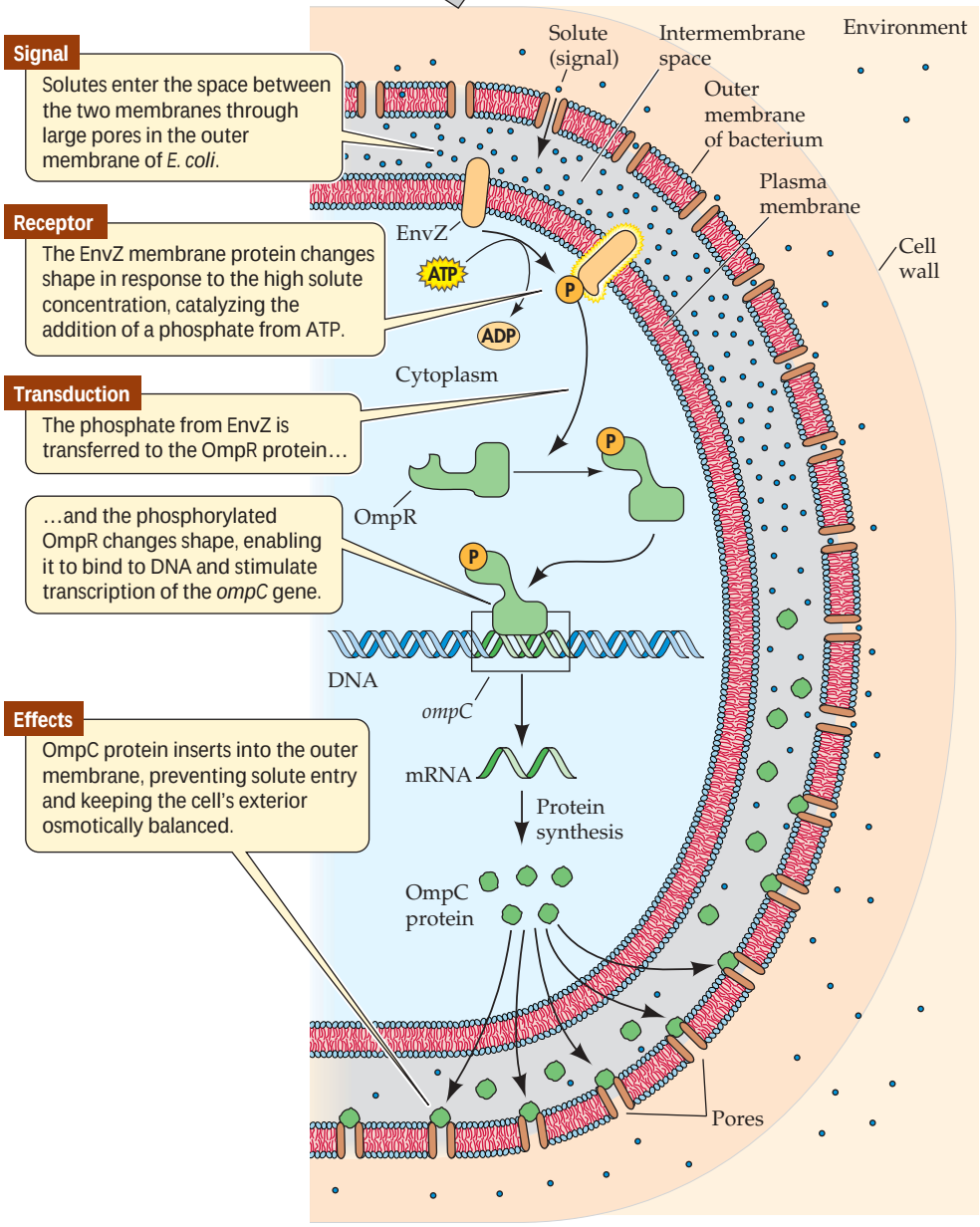
15.1 Chemical Signaling Systems A signal molecule can act on the cell that produces it, or on a nearby cell. Many signals act on distant cells, to which they are transported by the organism's circulatory system.

A signal transduction pathway involves a signal, a receptor, transduction, and effects

The entire signaling process, from signal detection to final response, is called a **signal transduction pathway**. Let's look at an example of such a pathway in *E. coli* (Figure 15.2). In Chapter 13, we saw that this bacterium responds to changes in the nutrient content of its environment by altering its transcription of certain genes, such as those in the *lac* operon. The bacterium must also be able to sense and respond to other kinds of changes in its environment, such as changes in solute concentration.

In the human intestine, where *E. coli* lives, the solute concentration around the bacterium often rises far above that inside the cell. The principle of diffusion tells us that when this happens, water will diffuse out of the cell and solutes will move into the cell. But the bacterium must maintain

15.2 A Model Signal Transduction Pathway *E. coli* responds to an increase in solute concentration in its environment. The basic steps of this pathway occur in all living organisms.



homeostasis, so it must perceive and respond to this environmental change. The pathway by which *E. coli* does so has much in common with signal transduction pathways in more complex animals and plants. The pathway involves two major components: a receptor and a responder.

RECEPTOR. A **receptor** is the first component of a signal transduction pathway. The receptor protein in *E. coli* for changes in solute concentration is called EnvZ. It is a transmembrane protein that extends through the bacterium's plasma membrane into the space between the plasma membrane and the highly porous outer membrane, which forms a complex with the cell wall. When the solute concentration of the extracellular environment rises, so does the solute concentration in the space between the two membranes. This change in its aqueous medium causes the part of the receptor protein sticking into the intermembrane space to undergo a conformational change.

As we saw in Chapter 6, changing the tertiary structure of one part of a protein often leads to changes in distant parts of the protein. In the case of the bacterial EnvZ receptor, the conformational change in the intermembrane domain of the protein is transmitted to the domain that lies in the cytoplasm, initiating the events of signal transduction. Through this conformational change, EnvZ becomes an active *protein kinase*, which catalyzes the addition of a phosphate group from ATP to one of EnvZ's own histidine residues. In other words, EnvZ phosphorylates itself.

RESPONDER. A **responder** is the second component of a signal transduction pathway. The charged phosphate group added to the histidine causes the cytoplasmic domain of the EnvZ protein to change its shape again. It now binds to a second protein, OmpR, which takes the phosphate group from EnvZ. This phosphorylation changes the shape of OmpR in turn. This change in a responder is a key event in signaling for three reasons:

- The signal on the outside of the cell has now been *transduced* to a protein totally within the cell's cytoplasm.
- The phosphorylated OmpR can *do something*. That "something" is to bind to a promoter on *E. coli* DNA adjacent to the sequence that codes for the protein OmpC. This binding begins the final phase of the signal-

ing pathway: the *effect* of the signal, which is an alteration in cell function.

- The signal has been *amplified*. Because a single enzyme can catalyze the conversion of many substrate molecules, one EnvZ molecule alters the structure of many OmpR molecules.

Phosphorylated OmpR is a transcription factor with the correct three-dimensional structure to bind to the promoter of the *ompC* gene, resulting in an increase in the transcription of that gene. Translation of *ompC* mRNA results in the production of OmpC protein, which leads to the response that regulates osmotic pressure. The OmpC protein is inserted into the outer membrane of the bacterial cell, where it blocks pores and prevents solutes from entering the intermembrane space. As a result, the solute concentration in the intermembrane space is lowered, and osmotic balance is restored. Thus the *E. coli* cell can go on behaving just as if the external environment had a normal osmotic concentration.

Let's highlight the major features of this prokaryotic system, as the same elements will reappear in many other signal transduction pathways in animals and plants:

- A receptor changes its conformation upon binding with a signal.
- A conformational change in the receptor results in protein kinase activity.
- Phosphorylation alters the function of a responder protein.
- The signal is amplified.
- A transcription factor is activated.
- The synthesis of a specific protein is turned on.
- The action of the protein alters cell activity.

Now that we have surveyed the general features of signal transduction pathways, let's consider more closely the nature of the receptors that bind signals.

Receptors

Although a given cell in a multicellular organism is bombarded with many signals, it responds to only a few of them. The reason for this is that any particular cell makes receptors for only some signals. Which cells make which receptors is determined by the regulatory processes we studied in the previous chapter: If a cell transcribes the gene encoding a particular receptor and the resulting mRNA is translated, the cell will have that receptor.

A receptor protein binds to a signal very specifically, in much the same way as an enzyme binds to a substrate or a carrier protein binds to the molecule it is transporting across a membrane. This specificity of binding underlies the specificity of which cells respond to which signals.

Receptors have specific binding sites for their signals

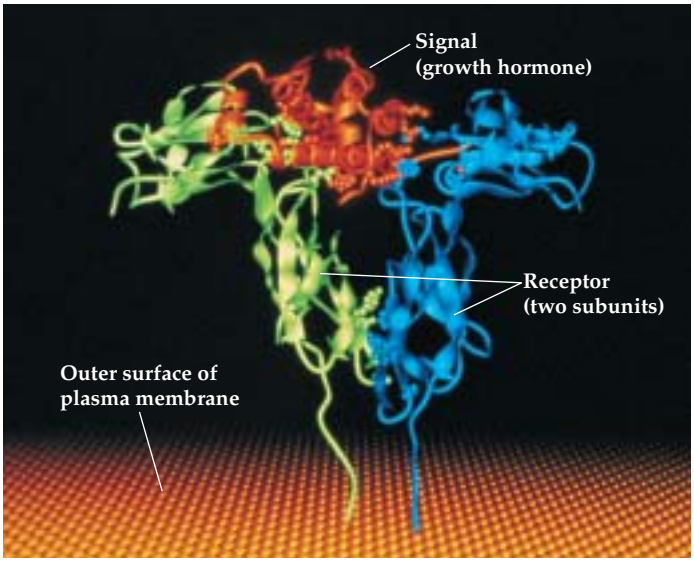
A specific signal molecule fits into a site on its receptor much as a substrate fits into the active site of an enzyme (Figure 15.3). A molecule that binds to a receptor site in another molecule in this way is called a **ligand**. Binding of the ligand causes the receptor protein to change its three-dimensional structure, and that conformational change initiates a cellular response. The ligand does not contribute further to this response. In fact, the ligand usually is not metabolized into useful products. Its role is purely to "knock on the door." This is in sharp contrast to enzyme–substrate interactions, in which the whole purpose is to change the substrate into a useful product.

Receptors bind to their ligands according to chemistry's law of mass action:

$$R + L \rightleftharpoons RL$$

This means that the binding is reversible, although for most ligand–receptor complexes, the equilibrium point is far to the right—that is, they favor binding. Reversibility is important, however, because if the ligand were never released, the receptor would be continuously stimulated.

The binding of a ligand to a receptor is similar in many ways to the binding of a substrate to an enzyme. As with enzymes, inhibitors can bind to the ligand binding site on a receptor protein. Both natural and artificial inhibitors of receptor binding are important in medicine. For example, many of the drugs that alter human behavior bind to specific receptors in the brain. Just as there are many types of enzymes with diverse specificities, there are many kinds of receptors.



15.3 A Signal Bound to Its Receptor Human growth hormone is shown bound to its receptor, a transmembrane protein. Only the extracellular regions of the receptor are shown.

There are several types of receptors

A major division among receptors is in their cellular location, which largely depends on the nature of their ligands. The chemistry of signal molecules is quite variable, but they can be divided into two classes (Figure 15.4):

- *Ligands with cytoplasmic receptors:* Small and/or nonpolar ligands can diffuse across the lipid bilayer of the plasma membrane and enter the cell. Estrogen, for example, is a lipid-soluble steroid hormone that can easily diffuse across the plasma membrane and enter the cell; it binds to a receptor in the cytoplasm.
- *Ligands with plasma membrane receptors:* Large and/or polar ligands cannot cross the plasma membrane. Insulin, for example, is a protein hormone that cannot diffuse through the plasma membrane; instead, it binds to a receptor that is a transmembrane protein with an extracellular binding domain.

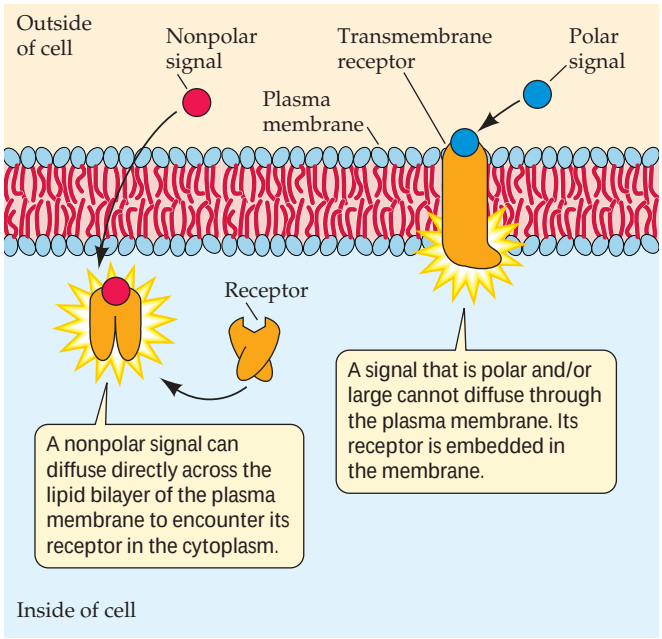
In complex eukaryotes such as mammals, there are three well-studied types of receptors on plasma membranes: ion channels, protein kinases, and G protein-linked receptors.

ION CHANNEL RECEPTORS. In the plasma membranes of many types of cells are channel proteins that can be open or closed. These **ion channels** act as "gates," allowing ions such as Na^+, K^+, Ca^{2+}, or Cl^- to enter or leave the cell. The gate-opening mechanism is an alteration in the three-dimensional structure of the channel protein upon ligand binding. Some ion channels are membrane receptors for signal molecules; others act later in signal transduction pathways. Each type of ion channel receptor has its own signal. These signals include sensory stimuli, such as light and sound, charge differences across the plasma membrane, and chemical ligands such as hormones and neurotransmitters.

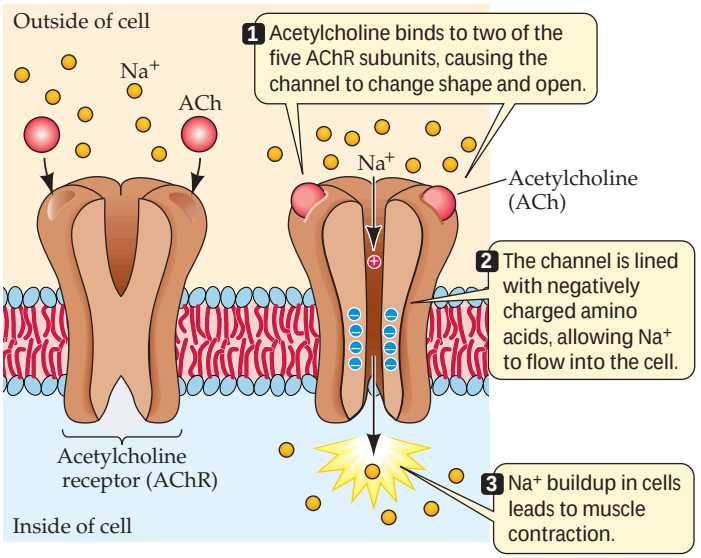
The acetylcholine receptor, which is located at the plasma membranes of vertebrate skeletal muscle cells, is an example of a gated ion channel. This receptor protein binds the ligand *acetylcholine*, which is released from nerve cells (Figure 15.5). When two molecules of acetylcholine bind to the receptor, it opens for about a thousandth of a second. That is enough time for Na^+, which is more concentrated outside the cell than inside, to rush into the cell. The change in Na^+ concentration in the cell results in muscle contraction.

PROTEIN KINASES. Like the EnvZ protein of *E. coli*, some eukaryotic receptor proteins become protein kinases when they are activated: that is, they catalyze the transfer of a phosphate group from ATP to a specific protein, referred to as the *target protein*. This phosphorylation can alter the conformation and activity of the target protein.

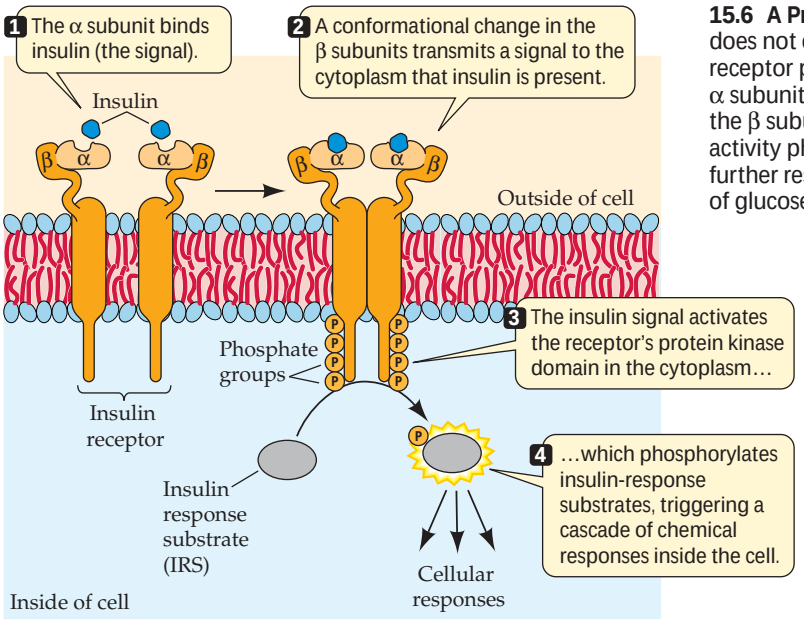
The receptor for insulin is an example of a protein kinase receptor. Insulin is a protein hormone made by the mammalian pancreas. Its receptor has two copies each of two different polypeptide subunits (Figure 15.6). As with acetylcholine, two molecules of insulin must bind to the receptor. When insulin binds to its extracellular subunits, the recep-



15.4 Two Locations for Receptors Receptors can be located in the plasma membrane or in the interior of the cell.



15.5 A Gated Ion Channel The acetylcholine receptor (AChR) is a gated ion channel for sodium ions. It is made up of five polypeptide subunits. When acetylcholine molecules (ACh) bind to two of the subunits, the gate opens and Na^+ flows into the cell.



15.6 A Protein Kinase Receptor The mammalian hormone insulin does not enter the cell, but is bound by the extracellular domain of a receptor protein with four subunits (two α and two β). Binding to the α subunit causes a conformational change in the cytoplasmic domain of the β subunits, exposing a protein kinase active site. This protein kinase activity phosphorylates insulin response substrate proteins, triggering further responses within the cell and eventually resulting in the transport of glucose across the membrane into the cell.

tor changes its shape to expose a cytoplasmic protein kinase active site. Like the EnvZ receptor described above, the insulin receptor autophosphorylates. Then, as a protein kinase, it catalyzes the phosphorylation of certain cytoplasmic proteins, appropriately called insulin response substrates. These proteins then initiate many cellular responses, including the insertion of glucose transporters into the plasma membrane.

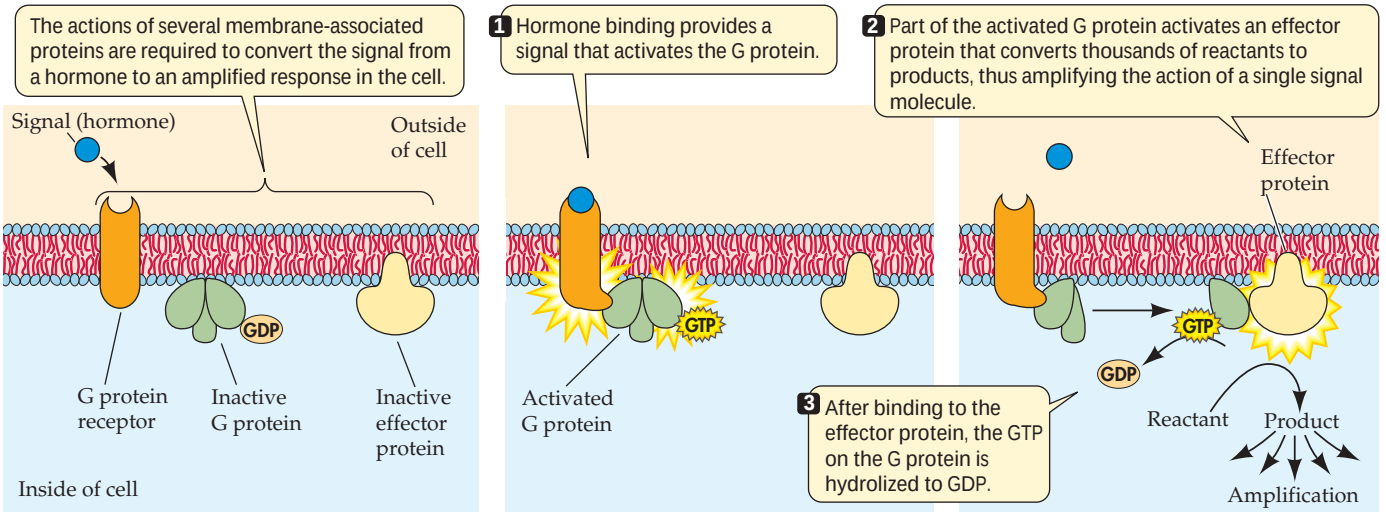
G PROTEIN-LINKED RECEPTORS. A third category of eukaryotic plasma membrane receptors is the *seven-spanning G protein-linked receptors*. This long name identifies a fascinating group of receptors, all of which are composed of a single protein with seven regions that pass through the lipid bilayer, separated by short loops that extend either outside or inside the cell. Ligand binding on the extracellular side of the receptor changes the shape of its cytoplasmic region, exposing a binding site for a mobile membrane protein.

This membrane protein, known as a **G protein**, has two important binding sites: one for the G protein-linked receptor and the other for the nucleotide GDP/GTP (Figure 15.7). G proteins have several polypeptide subunits. When the G protein binds to the activated receptor, one of its subunits binds GTP. At the same time, the ligand is released from the extracellular side of the receptor. The GTP-bound subunit of the G protein now separates from the parent G protein, diffusing in the plane of the lipid bilayer until it encounters an effector protein to which it can bind.

An *effector protein* is just what its name implies: It causes an effect in the cell. The binding of the GTP-bearing G protein subunit activates the effector—which may be an enzyme or an ion channel—thereby causing changes in cell function.

After binding to the effector protein, the GTP on the G protein is hydrolyzed to GDP. The now inactive G protein

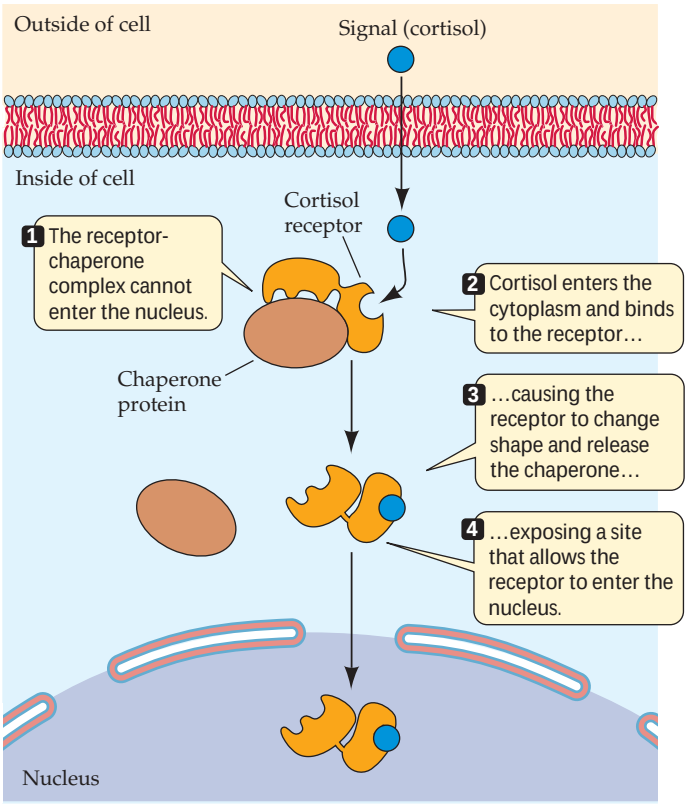
15.7 A G Protein-Linked Receptor Binding of an extracellular signal—in this case, a hormone—causes the activation of a G protein-linked receptor. The G protein then activates an effector protein—in this case, an enzyme that catalyzes a reaction in the cytoplasm, amplifying the signal. This figure is a generalized diagram that could apply to any member of the large family of G proteins and the signals they react to.



subunit separates from the effector protein. The G protein subunit must form a complex with other subunits before binding to yet another activated receptor. When an activated receptor is bound, the G protein exchanges its GDP for GTP, and the cycle begins again.

By means of their diffusing subunits, G proteins can either activate or inhibit an effector. An example of an *activating* response involves the receptor for epinephrine (adrenaline), hormone made by the adrenal gland in response to stress or heavy exercise. In heart muscle, this hormone binds to its G protein-linked receptor, activating a G protein . The GTP-bound subunit then activates a membrane-bound enzyme to produce a small molecule, cyclic AMP (see below), that has many effects on the cell, including glucose mobilization for energy and muscle contraction.

G protein-mediated *inhibition* occurs when the same hormone, epinephrine, binds to its receptor in the smooth muscle cells surrounding blood vessels lining the digestive tract. Again, the epinephrine-bound receptor changes its shape and activates a G protein, and the GTP-bound subunit binds to a target enzyme. But in this case, the enzyme is inhibited instead of being activated. As a result, the muscles relax and the blood vessel diameter increases, allowing more nutrients to be carried away from the digestive system to the rest of the body. Thus the same signal and initial signaling mechanism can have different consequences in different cells, depending on the nature of the responding cell.



15.8 A Cytoplasmic Receptor The receptor for cortisol is bound to a chaperone protein. Binding of the signal releases the chaperone and allows the receptor protein to enter the cell's nucleus, where it functions as a transcription factor.

CYTOPLASMIC RECEPTORS. Receptors for signals that can diffuse across the plasma membrane are located inside the cell. Binding to the ligand causes the receptor to change its shape so that it can enter the cell nucleus, where it acts as a transcription factor (Figure 15.8). But this general view is somewhat simplified. The receptor for the hormone cortisol, for example, is normally bound to a chaperone protein, which blocks it from entering the nucleus. Binding of the hormone causes the receptor to change its shape so that the chaperone is released. This allows the receptor, which is a transcription factor, to fold into an appropriate conformation for entering the nucleus and initiating transcription.

Having discussed signals and receptors, we now turn our attention to the characteristics of transducers.

Signal Transduction

As we have just seen, the same signal may produce different responses in different tissues. When epinephrine, for example, binds to receptors on heart muscle cells, it stimulates muscle contraction, but when it binds to receptors on smooth muscle cells in the blood vessels of the digestive system, it slows muscle contraction. These different responses to the same signal–receptor complex are mediated by the events of signal transduction. These events, which are critical to the cell's response, may be either direct or indirect.

- *Direct transduction* is a function of the receptor itself and occurs at the plasma membrane.
- In *indirect transduction*, which is more common, another molecule, termed a **second messenger**, mediates the interaction between receptor binding and cellular response.

In neither case is transduction a single event. Rather, the signal initiates a cascade of events, in which proteins interact with other proteins until the final responses are achieved. Through such a cascade, a weak initial signal can be both amplified and distributed to cause several different responses in the target cell.

Protein kinase cascades amplify a response to receptor binding

We have seen that when a signal binds to a protein kinase receptor, the receptor changes its conformation to expose a protein kinase active site, which catalyzes the phosphorylation of target proteins. This process is an example of direct signal

transduction. Protein kinase receptors are important in binding ligands that stimulate cell division in both plants and animals. In Chapter 9, we described growth factors that serve as external inducers of the cell cycle. These growth factors work by binding to protein kinase receptors.

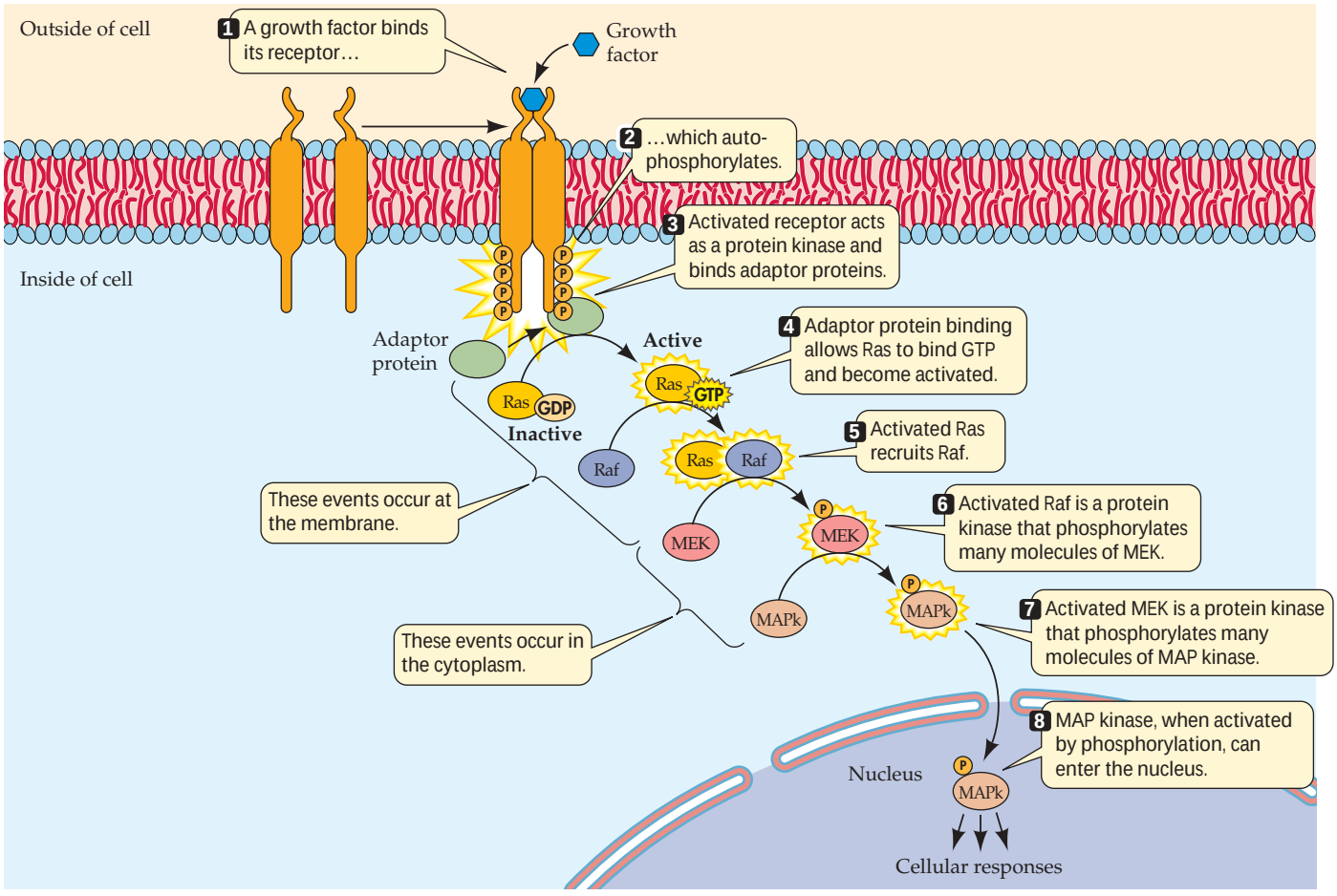
The complete signal transduction pathway that occurs after a protein kinase receptor binds a growth factor was worked out through studies on a cell that went wrong. Many human bladder cancers contain an abnormal form of a protein called Ras (so named because it was first isolated from a *ra*t *s*arcoma tumor). Investigations of these bladder cancers showed that this Ras protein was a G protein, but was always active because it was permanently bound to GTP. So the abnormal Ras protein caused continuous cell division. If the cancer cells' Ras protein was inhibited, they stopped dividing. This discovery has led to a major effort to develop specific Ras inhibitors for cancer treatment.

What does Ras do in normal, noncancerous cells? Researchers knew that cells must be stimulated by growth factors (signals) in order to enter the cell cycle and divide. One hypothesis was that Ras was an intermediary between the binding of a growth factor to its receptor and the ultimate response of cell division. To investigate this hypothesis, the researchers treated cells in a culture dish with both a Ras inhibitor and a growth factor. Cell division did not occur, confirming their hypothesis.

After this discovery, the next step was to work out what the activated growth factor receptor did to Ras, and what Ras did to stimulate further events in signal transduction. This signaling pathway has been worked out, and it is an example of a more general phenomenon, called a **protein kinase cascade** (Figure 15.9). Such cascades are key to the external regulation of many cellular activities. Indeed, the eukaryotic genome codes for hundreds, even thousands, of such kinases.

The unbound receptors for growth factors exist in the plasma membrane as separate polypeptide chains (subunits). When the growth factor signal binds to a subunit, it associates with another subunit to form a dimer, which changes its shape to expose a protein kinase active site. The kinase activity sets off a series of events, activating several other pro-

15.9 A Protein Kinase Cascade In a protein kinase cascade, a series of proteins are sequentially activated. In this example, the growth factor receptor protein stimulates the G protein Ras, which mediates a cascading series of reactions. The final product of the cascade, MAP kinase (MAPk), enters the nucleus and causes changes in transcription. Inactive forms of the proteins are on the left, activated forms are on the right.



tein kinases in turn. The final phosphorylated, activated protein—MAP kinase—moves into the nucleus and phosphorylates target proteins that are necessary for cell division.

Protein kinase cascades are useful signal transducers for three reasons:

- At each step in the cascade of events, the signal is amplified, because each newly activated protein kinase is an enzyme, which can catalyze the phosphorylation of many target proteins.
- The information from a signal that originally arrived at the plasma membrane is communicated to the nucleus.
- The multitude of steps provides some specificity to the process. As we have seen with epinephrine, signal binding and receptor activation do not result in the same response in all cells. Different target proteins at each step in the cascade can provide variation in the response.

Cyclic AMP is a common second messenger

As we have just seen, protein kinase receptors initiate the protein kinase cascade right at the plasma membrane. However, the stimulation of events in the cell is more often indirect. In a series of clever experiments, Earl Sutherland, Edwin Krebs, and Edmond Fischer showed that in many cases, there is a small, water-soluble chemical messenger between the membrane receptor and cytoplasmic events. These researchers were investigating the activation of the liver enzyme *phosphorylase* by the hormone epinephrine. Phosphorylase catalyzes the hydrolysis of glycogen stored in the liver so that the resulting glucose molecules can be released to the blood to fuel the fight-or-flight response.

The researchers found that phosphorylase could be activated in liver cells that had been broken open, but only if the entire cell contents, including the plasma membrane fragments, were present. They observed that epinephrine had bound to the plasma membrane, but active phosphorylase was present in the cytoplasm. They hypothesized that there must be some chemical messenger that transmits the message of epinephrine binding (at the membrane) to phosphorylase (in the cytoplasm). To investigate the production of this message, they tried the following steps in sequence:

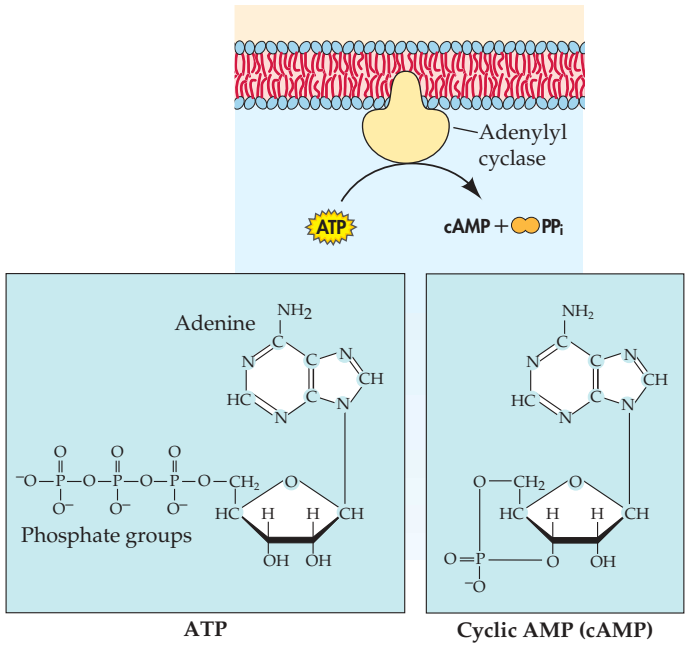
- First, they incubated plasma membranes of broken liver cells with epinephrine.
- Then they removed the membranes, but kept the solution in which the membranes had been incubated.
- Then they added this solution to the contents of the cytoplasm, which contained inactive phosphorylase.

The phosphorylase became activated, confirming their hypothesis. Hormone binding to the membrane receptor had caused the production of a small, water-soluble molecule that then diffused to the cytoplasm, where it activated the enzyme. This small molecule was identified as **cyclic AMP** (**cAMP**), which we encountered in Chapter 13 in the *lac* operon regulatory system in *E. coli*. Here, cAMP was working as a second messenger.

Second messengers are substances released into the cytoplasm after the first messenger—the signal—binds its receptor. In contrast to the specificity of receptor binding, second messengers affect many processes in the cell, and they allow a cell to respond to a single event at the plasma membrane with many events inside the cell. Like the protein kinase cascade, second messengers amplify the signal—a single epinephrine molecule leads to the production of several dozen molecules of cAMP, which then activate many enzyme targets.

Adenylyl cyclase, the enzyme that catalyzes the formation of cAMP from ATP, is located on the cytoplasmic surface of the plasma membrane of target cells (Figure 15.10). Usually, it is activated by the binding of G proteins, themselves activated by receptors.

Second messengers do not have enzymatic activity; rather, they act as cofactors or allosteric regulators of target enzymes. Cyclic AMP has two major target types. In many kinds of sensory cells, cAMP binds to ion channels to open them. Cyclic AMP may also binds to an enzyme in the cytoplasm, such as a protein kinase, whose active site is exposed as a result. A protein kinase cascade ensues, leading to the final effects in the cell.



15.10 The Formation of Cyclic AMP The formation of cAMP from ATP is catalyzed by adenylyl cyclase, an enzyme that is activated by G proteins.

Two second messengers are derived from lipids

Phospholipids, in addition to their roles as structural components of the plasma membrane, are involved in signal transduction. When certain phospholipids are hydrolyzed into their component parts (see Figure 3.20) by enzymes called *phospholipases*, second messengers are formed.

The best-studied of these second messengers come from hydrolysis of the phospholipid **phosphatidyl inositol-bisphosphate** (**PIP2**), which, like all phospholipids, has a hydrophobic portion (two fatty acid tails attached to a molecule of glycerol, which together form **diacylglycerol**, or **DAG**) embedded in the plasma membrane and a hydrophilic portion (**inositol triphosphate**, or **IP_3**) projecting into the cytoplasm. In mammals, there are over two dozen signals whose actions are mediated by the products of PIP2 hydrolysis. As with cAMP, the receptors involved are often G protein-linked receptors. The activated G protein subunits diffuse within the plasma membrane and activate an enzyme, phospholipase C. This enzyme cleaves off the IP_3 from PIP2, leaving the glycerol and the two attached fatty acids (DAG) in the lipid bilayer:

$$\underset{\text{(in membrane)}}{\text{P1P2}} \xrightarrow{\text{Phospholipase C}} \underset{\text{released to cytoplasm}}{IP_3} + \underset{\text{(in membrane)}}{\text{DAG}}$$

IP_3 and DAG are both second messengers and have different modes of action that build on each other (Figure 15.11). DAG activates a membrane-bound enzyme, protein kinase C (PKC). PKC is dependent on Ca^{2+} (hence the "C"), and that is where IP_3 plays an essential role. IP_3 diffuses through the cytoplasm to the smooth endoplasmic reticulum, where it opens an ion channel, releasing Ca^{2+} into the cytoplasm. There, in combination with DAG, the Ca^{2+} causes PKC to become active. PKC can then phosphorylate a wide variety of proteins, leading to the ultimate response of the cell.

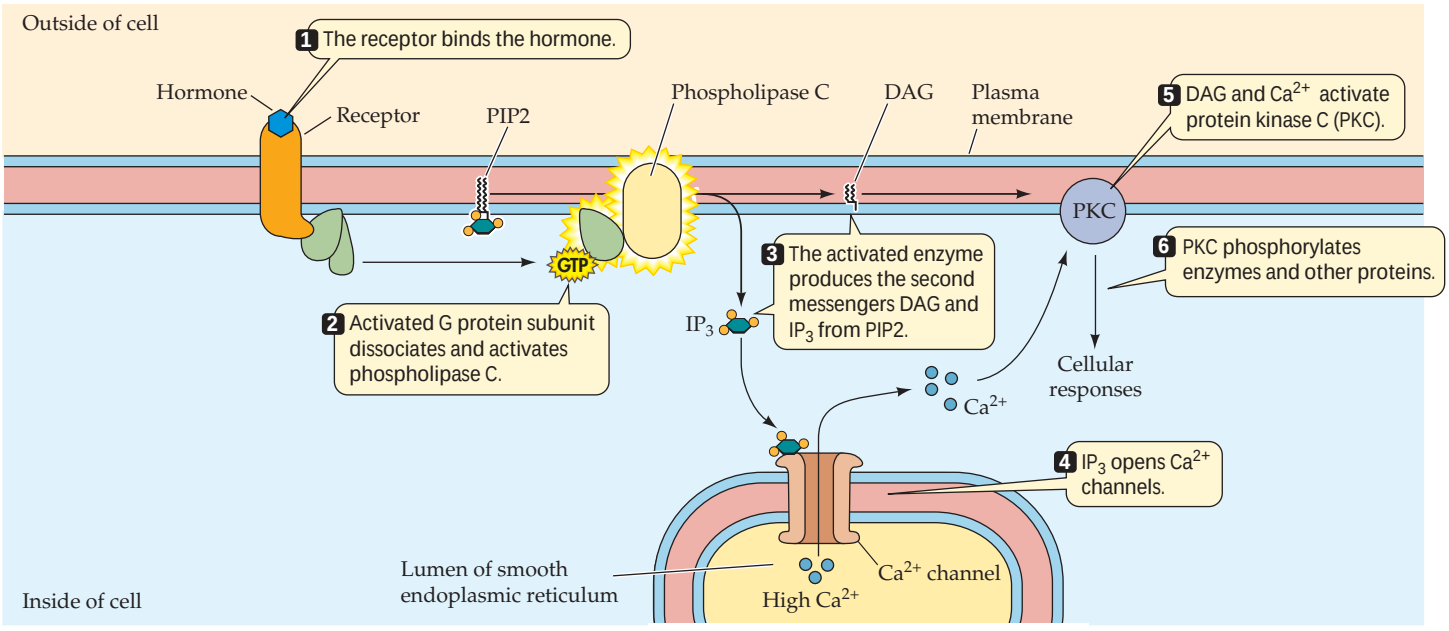
In this transduction system, DAG and IP_3 function as second messengers, but Ca^{2+} plays a role in the pathway. In some cases, however, Ca^{2+} can itself serve as the second messenger in a signal transduction pathway.

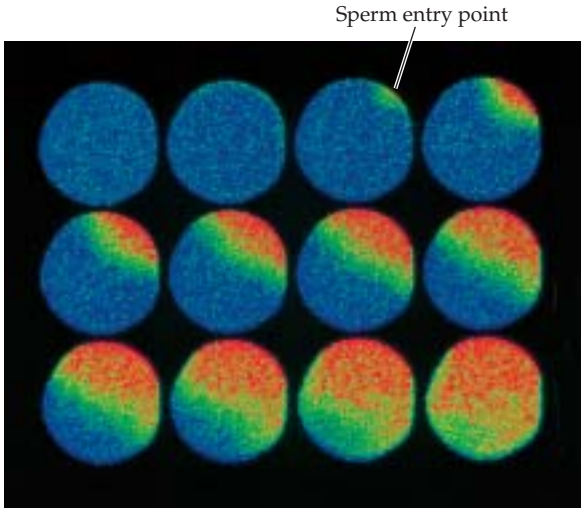
Calcium ions are involved in many signal transduction pathways

Calcium ions are scarce in most cells, with a cytoplasmic concentration of only about 0.1 μM, while the concentrations of Ca^{2+} outside the cell and within the endoplasmic reticulum are usually much higher. This difference is maintained by active transport proteins at the plasma and ER membranes that pump the ion out of the cytoplasm. Unlike cAMP and the lipid second messengers, the level of intracellular Ca^{2+} cannot be increased by making more of it. Instead, the opening and closing of ion channels and the action of membrane pumps regulate levels of the ion in a cellular compartment.

There are many signals that can cause Ca^{2+} channels to open, including IP_3 (as we saw in the previous section) and the entry of a sperm into an egg (Figure 15.12). Whatever the signal, the open channels result in a dramatic increase in cytoplasmic Ca^{2+} concentration, up to a hundredfold within a fraction of a second. As we saw earlier, this increase activates

15.11 The IP_3 and DAG Second Messenger System
Phospholipase C hydrolyzes the phospholipid PIP2 into its components, IP_3 and DAG, both of which are second messengers. IP_3 and DAG act separately but in concert, ultimately producing a wide range of responses in the cell.





15.12 Calcium Ions as a Second Messenger The concentration of Ca^{2+} can be measured by a dye that fluoresces and turns red when it binds the ion. Here, fertilization causes a wave of Ca^{2+}, photographed at 5-second intervals, to pass through the egg of a sea star (starfish). This signal delivers the message that fertilization is complete and development can begin.

protein kinase C. In addition, Ca^{2+} controls other ion channels and stimulates secretion by exocytosis.

A distinctive aspect of Ca^{2+} signaling is that the ion can stimulate its own release from intracellular stores. For example, in some plant leaf cells, the hormone abscisic acid binds to gated Ca^{2+} channels in the plasma membrane and opens them, causing the ion to rush into the cells. This influx is not enough to trigger the cell's response, however. The ion binds to Ca^{2+} channels in the endoplasmic reticulum and in the membranes of vacuoles, causing those organelles to release their Ca^{2+} stores as well.

In some cases, Ca^{2+} ions act via a calcium-binding protein called **calmodulin**, and it is the Ca^{2+}–calmodulin complex that performs cellular functions by binding to target proteins. Calmodulin, which is present in many cells, has four binding sites for Ca^{2+}. When the cytoplasmic Ca^{2+} concentration is low, calmodulin does not bind enough Ca^{2+} to become activated. But when the cell is stimulated by a signal that causes a rise in the Ca^{2+} level, all four binding sites are filled. The calmodulin then changes its shape and binds to a number of cellular targets, activating them in turn. One such target is a protein kinase in smooth muscle cells that phosphorylates the muscle protein myosin, initiating contraction.

Nitric oxide is a gas that can act as a second messenger

Pharmacologist Robert Furchgott, at the State University of New York in Brooklyn, was investigating how acetylcholine causes the smooth muscles lining blood vessels to relax, thus allowing more blood to flow to certain organs. Acetylcholine appeared to stimulate the IP_3 signal transduction pathway to produce an influx of Ca^{2+}, which led to an increase in the level of another second messenger, **cyclic GMP** (**cGMP**). This nucleotide bound to a protein kinase, which then stimulated a kinase cascade leading to muscle relaxation. So far, the pathway seemed straightforward.

But while this pathway seemed to work in intact animals, it did not work on isolated strips of artery tissue. When Furchgott switched to tubular sections of artery, however, signal transduction did occur. There turned out to be a crucial difference between these two tissue preparations: In the strips, the delicate inner layer of cells that lines blood vessels had been lost. Furchgott hypothesized that this layer, the *endothelium*, was making something that diffused into the muscle cells and was needed for their response to acetylcholine. The substance was not easy to isolate. It seemed to break down quickly, with a half-life (the time in which half of it disappeared) of 5 seconds in living tissues. It turned out to be a gas, **nitric oxide** (**NO**), that had been thought of only as a toxic air pollutant!

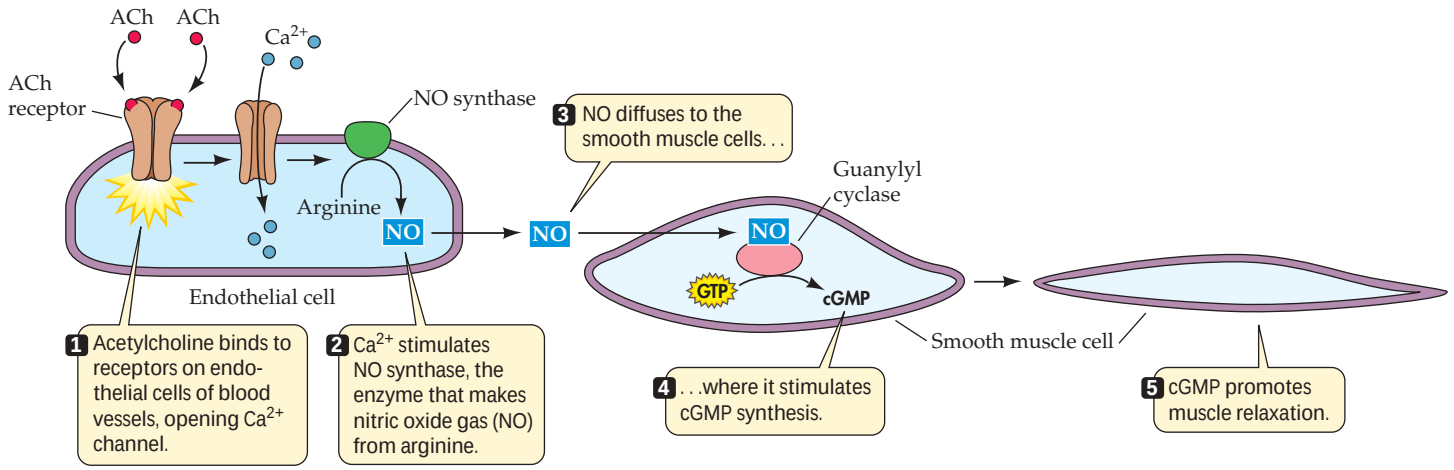
In the body, NO is made from arginine by an enzyme, *NO synthase*. This enzyme is activated by Ca^{2+}, which enters the endothelial cells through a channel opened by PIP2, which is released after acetylcholine binds to its receptor. The NO formed is chemically very unstable, and although it diffuses readily, it does not get far. Conveniently, the endothelial cells are close to the smooth muscle cells, where NO acts as a second messenger. In smooth muscle, NO activates an enzyme called guanylyl cyclase, catalyzing the formation of cGMP, which in turn relaxes the muscle cells (Figure 15.13).

The spectacular discovery of NO as a second messenger explained the action of nitroglycerin, a drug that has been used for over a century to treat angina, the chest pain caused by insufficient blood flow to the heart. Nitroglycerin releases NO, which results in relaxation of the blood vessels and increased blood flow. Penile erection is also caused by the dilation of blood vessels in that organ, and the new drugs that promote erection are NO synthesis activators.

Signal transduction is highly regulated

There are several ways in which cells can regulate the activity of a transducer. The concentration of NO, which breaks down quickly, can be regulated only by how much of it is made. The level of Ca^{2+}, on the other hand, is determined by both membrane pumps and ion channels. For protein kinase cascades, G proteins, and cAMP, there are enzymes that convert the activated transducer back to its inactive precursor:

- *Protein phosphatases* remove the phosphate groups from phosphorylated proteins.
- *GTPases* convert the GTP on an active G protein back to GDP, inactivating the protein.
- *cAMP phosphodiesterase* converts cAMP into its precursor, AMP, which has no second messenger activity.



15.13 Nitric Oxide as a Second Messenger Nitric oxide (NO) is an unstable gas, which nevertheless serves as a second messenger between a primary signal, acetylcholine, and its effect, the relaxation of smooth muscles. The endothelial cells that line blood vessels, seen in *(a)*, are crucial intermediaries in this signal transduction pathway *(b)*.

Signal Effects: Changes in Cell Function

We have seen how the binding of an environmental signal to its receptor initiates the response of a cell to the signal, and how the direct or indirect transduction of the signal to the inside of the cell amplifies the signal. In this section, we consider the third and final step in the signal transduction process, the actual effects of the signal on cell function. These effects primarily take the form of the opening of ion channels, changes in the activities of enzymes, or differential gene transcription.

Ion channels are opened

The opening of ion channels is a key step in the response of the nervous system to signals. Sensory nerve cells of the sense organs, for example, become stimulated through the opening of ion channels. We will focus here on one such signal transduction pathway, that for the sense of smell, which responds to gaseous molecules (Figure 15.14).

The sense of smell is well developed in mammals, some of which have an amazing 1,000 genes for odor signal receptors—the largest gene family known. Each of the thousands of nerve cells in the nose expresses one of these receptors. The identification of which chemical signal, or odorant, activates which receptor is just getting under way.

When an odorant molecule binds to its receptor, a G protein becomes activated, which in turns activates adenylyl cyclase, which catalyzes the formation of the second messenger cAMP. This molecule then binds to an ion channel, causing it to open. The resulting influx of Na^+ causes the nerve cell to become stimulated so that it sends a signal to the brain that a particular odor is present.

Enzyme activities are changed

Proteins will change their shape, and their functioning, if they are modified either covalently or noncovalently. We have seen examples of both types of modification in signal transduction. Protein kinases add phosphate groups to a target protein, and this covalent change alters the protein's conformation. Cyclic AMP binds to target proteins allosterically, and this noncovalent interaction changes the protein's conformation. In both cases, previously inaccessible active sites are exposed, and the target protein goes on to perform a cellular role.

The G protein-mediated protein kinase cascade stimulated by epinephrine in liver cells results in the phosphorylation of two key enzymes in glycogen metabolism, with opposite effects (Figure 15.15):

- *Inhibition.* Glycogen synthase, which catalyzes the joining of glucose molecules to synthesize the energy-storing molecule glycogen, is inactivated by phosphorylation. Thus the epinephrine signal prevents glucose from being stored in glycogen.
- *Activation.* Phosphorylase kinase is activated when a phosphate group is added to it. It goes on to stimulate a protein kinase cascade that ultimately leads to the activation by phosphorylation of phosphorylase, the other key enzyme in glucose metabolism. This enzyme liberates glucose molecules from glycogen.

Thus the same signaling pathway inhibits the storage of glucose as glycogen (by inhibiting glycogen synthase) and promotes the release of glucose through glycogen breakdown (by activating glycogen phosphorylase). As we mentioned earlier, the released glucose fuels the ATP-requiring fight-or-flight response to epinephrine.

Different genes are transcribed

Plasma membrane receptors are involved in activating a broad range of gene expression responses. The Ras signaling pathway, for example, ends in the nucleus (see Figure 15.9). The final protein kinase in the cascade, MAPK, enters the nucleus and phosphorylates a leucine zipper protein called AP-1. This activated protein is a transcription factor, and it stimulates the transcription of a number of genes involved in cell proliferation.

As described earlier in this chapter, lipid-soluble hormones can diffuse through the plasma membrane and meet their receptors in the cytoplasm. In this case, binding of the ligand allows the ligand–receptor complex to enter the nucleus, where it binds to hormone-responsive elements at the promoters of a number of genes. In some cases, transcription is stimulated, and in others it is inhibited.

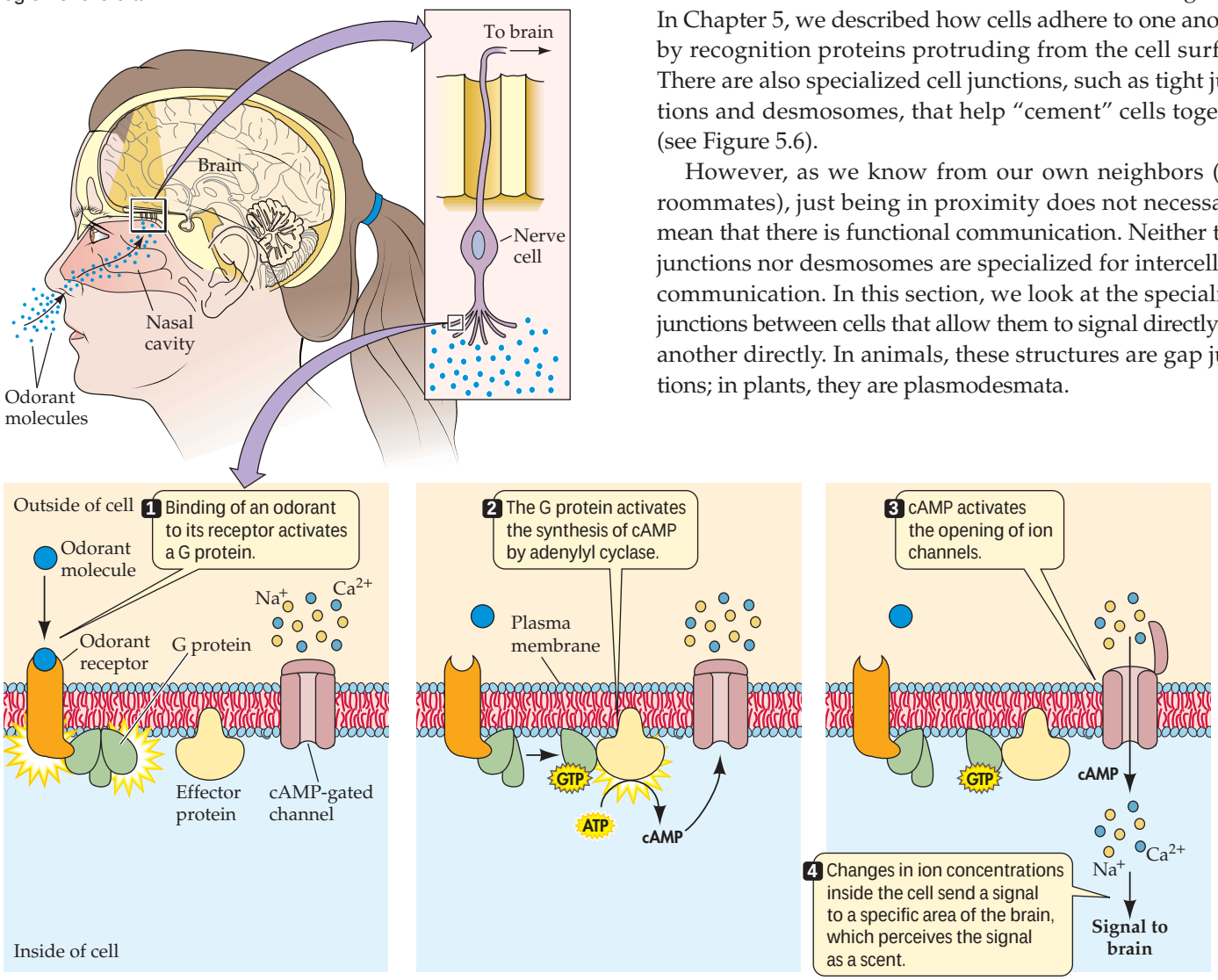
In plants, light acts as a signal to initiate the formation of chloroplasts. Between this signal and response is a transcription-mediated signal transduction pathway. In bright sunlight, red wavelengths are absorbed by a receptor protein called *phytochrome*. We will say more about this important receptor later in the book, but for now it is important to note only that it is activated by red light. The activated phytochrome binds to cytoplasmic regulatory proteins, which enter the nucleus and bind to promoters of genes involved in the synthesis of important chloroplast proteins. Synthesis of these proteins is the key to plant "greening."

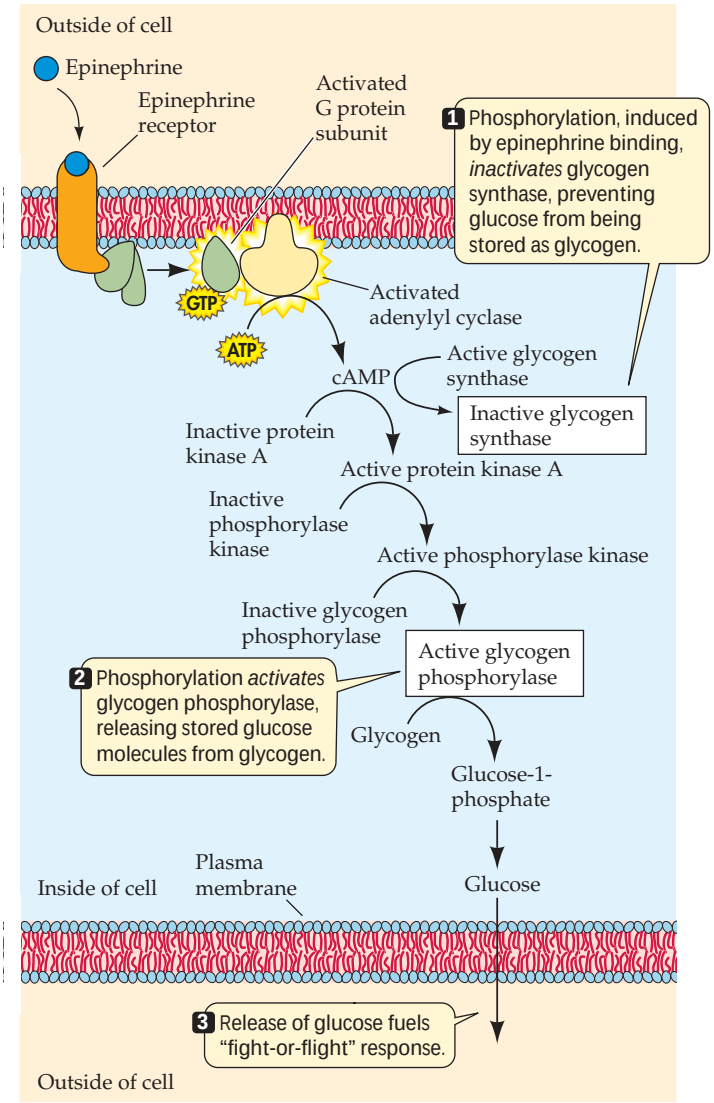
Direct Intercellular Communication

Up to now, we have described how signals from a cell's environment can influence that cell. But the environment of a cell in a multicellular organism is more than the extracellular medium. Most cells are in contact with their neighbors. In Chapter 5, we described how cells adhere to one another by recognition proteins protruding from the cell surface. There are also specialized cell junctions, such as tight junctions and desmosomes, that help "cement" cells together (see Figure 5.6).

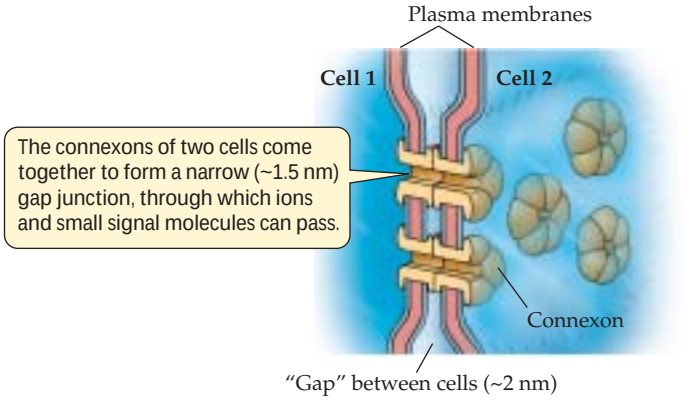
However, as we know from our own neighbors (and roommates), just being in proximity does not necessarily mean that there is functional communication. Neither tight junctions nor desmosomes are specialized for intercellular communication. In this section, we look at the specialized junctions between cells that allow them to signal directly one another directly. In animals, these structures are gap junctions; in plants, they are plasmodesmata.

15.14 A Signal Transduction Pathway Leads to the Opening of Ion Channels In the signal transduction pathway for the sense of smell, the final effect is the opening of Na^+ channels. The resulting influx of Na^+ stimulates the transmission of a scent message to a specific region of the brain.





15.15 A Cascade of Reactions Leads to Altered Enzyme Activity Liver cells respond to epinephrine by activating G proteins, which in turn activate cAMP synthesis. The second messenger initiates a protein kinase cascade. The cascade both inhibits the conversion of glucose to glycogen and stimulates the release of previously stored glucose.



15.16 Gap Junctions Connect Animal Cells An animal cell may contain hundreds of gap junctions connecting it to neighboring cells. Gap junctions are too small for proteins, but small molecules such as ATP, metabolic intermediates, amino acids, and coenzymes can pass through them.

Animal cells communicate by gap junctions

Gap junctions are channels between adjacent cells that occur in many animals, occupying up to 25 percent of the area of the plasma membrane (Figure 15.16). Gap junctions traverse the narrow space between the plasma membranes of two cells (the "gap") by means of thin molecular channels called *connexons*. The walls of these channels are composed of six subunits of an integral membrane protein. In two cells close to each other, two connexons come together, forming a channel that links the two cytoplasms. There may be hundreds of these channels between a cell and its neighbors. The channels about 1.5 nm in diameter—far too narrow for the passage of large molecules such as proteins. But they are wide enough to allow small signal molecules and ions to pass between the cells. Experiments in which a labeled signal molecule or ion is injected into one cell show that it can readily pass into the adjacent cells if the cells are connected by gap junctions.

Gap junctions permit metabolic cooperation among the linked cells. Such cooperation ensures the sharing of important small molecules such as ATP, metabolic intermediates, amino acids, and coenzymes between cells. It may also ensure that concentrations of ions and small molecules are similar in linked cells, thereby maintaining equivalent regulation of metabolism. It is not clear how important this function is in many tissues, but it is known to be vital in some. In the lens of the mammalian eye, for example, only the cells at the periphery are close enough to the blood supply to allow diffusion of nutrients and wastes. But because lens cells are connected by large numbers of gap junctions, material can diffuse between them rapidly and efficiently.

There is evidence that signal molecules such as hormones and second messengers such as cAMP and PIP2 can move through gap junctions. If this is true, only a few cells would need to have receptors binding a signal in order for the stimulus to spread throughout the tissue. In this way, a tissue could have a coordinated response to the signal.

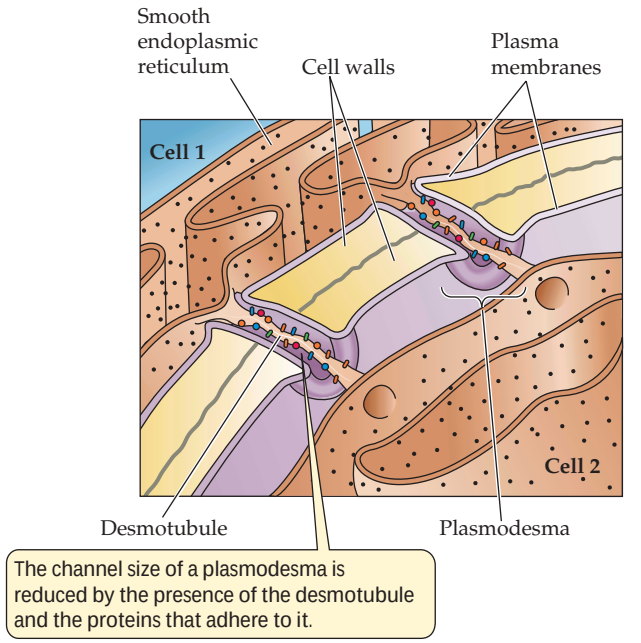
Plant cells communicate by plasmodesmata

Instead of gap junctions, plants have **plasmodesmata**, which are membrane-lined bridges spanning the thick cell walls that separate plant cells from one another. A typical plant cell has several thousand plasmodesmata.

Plasmodesmata differ from gap junctions in one fundamental way: Unlike gap junctions, in which the wall of the channel is made of integral proteins from the adjacent plasma membranes, plasmodesmata are lined by the fused plasma membranes themselves. Plant biologists are so familiar with the notion of a tissue as cells interconnected in this way that they refer to these continuous cytoplasms as a *symplast* (see Chapter 36).

The diameter of a plasmodesma is about 6 nm, far larger than the gap junction channel. But the actual space available for diffusion is about the same—1.5 nm. A look at the interior of the plasmodesma gives the reason for this reduction in pore size: A tubule called the **desmotubule**, apparently derived from the endoplasmic reticulum, fills up most of the opening of the plasmodesma (Figure 15.17). So, typically, only small metabolites and ions move between plant cells. This fact is important physiologically to plants, which lack the tiny circulatory vessels (capillaries) many animals use to bring gases and nutrients to every cell.

Diffusion from cell to cell through plasma membranes is probably inadequate for hormonal responses in plants. Instead, they rely on more rapid diffusion through plasmodesmata to ensure that all cells of a tissue respond to a signal at the same time. In C_4 plants (see Chapter 8), there are abundant plasmodesmata between the mesophyll and bundle sheath cells, which help to rapidly move the carbon fixed in the former cell type to the latter. A similar transport system, found at the junctions of nonvascular tissues and phloem, conducts organic solutes throughout the plant.



15.17 Plasmodesmata Connect Plant Cells The desmotubule, derived from the smooth endoplasmic reticulum, fills up most of the space inside a plasmodesma, leaving a tiny gap through which small metabolites and ions can pass.

Plasmodesmata are not merely passive channels, but can be regulated. Plant viruses may infect cells at one location, then spread rapidly through a plant organ by plasmodesmata until they reach the plant's vascular tissue (circulatory system). These viruses, and even their RNA, would appear to be many times too large to pass through the desmotubules. But they get through, apparently by making "movement proteins" that increase the pore size temporarily while attached to the viral genome. Similar movement proteins made by the plants themselves are involved in transporting mRNAs and even proteins such as transcription factors between plant cells. This finding opens up the possibility of long-distance regulation of transcription and translation.

Chapter Summary

Signals

- Cells receive many signals from the physical environment and from other cells. **Review Figures 15.1**
- A signal transduction pathway involves three steps: the binding of a signal by a receptor, the transduction of the signal within the cell, and the ultimate cellular response. **Review Figure 15.2.** **See Web/CD Activity 15.1**

Receptors

- Cells respond to signals only if they have specific receptor proteins that can bind to those signals. **Review Figure 15.3**
- Depending on the nature of its signal, a receptor may be located in the plasma membrane or in the cytoplasm of the target cell. **Review Figure 15.4**
- Receptors located in the plasma membrane include ion channels, protein kinases, and G protein-linked receptors. **Review Figures 15.5, 15.6, 15.7.** **See Web/CD Tutorial 15.1**
- When bound by a ligand, cytoplasmic receptors change their shape and enter the cell nucleus. **Review Figure 15.8**

Signal Transduction

- The events of signal transduction may be direct, occurring at the plasma membrane, or indirect, involving the formation of a second messenger.
- Protein kinase cascades amplify a response to receptor binding. **Review Figure 15.9**
- Second messengers include cyclic AMP, the lipid-derived substances inositol triphosphate and diacylglycerol, calcium ions, and the gas nitric oxide. **Review Figures 15.10, 15.11, 15.12, 15.13**

Signal Effects: Changes in Cell Function

- The ultimate cell response to a signal may be the opening of ion channels, the alteration of enzyme activities, or changes in gene transcription. **Review Figures 15.14, 15.15**

Direct Intercellular Communication

- Most animal cells can communicate with one another directly through small pores in their plasma membranes called gap junctions. Small molecules and ions can pass through these pores. **Review Figure 15.16**

▶ Plant cells are connected by somewhat larger pores called plasmodesmata, which traverse both membranes and cell walls.
Review Figure 15.17
See Web/CD Activity 15.2 for a concept review of this chapter.

Self Quiz

1. What is the correct order for these events in the interaction of a cell with a signal? (1) alteration of cell function; (2) signal binds to receptor; (3) signal released from source; (4) signal transduction.
   a. 1234
   b. 2314
   c. 3214
   d. 3241
2. Why do some signals ("first messengers") trigger a "second messenger" to activate a target cell?
   a. The first messenger requires activation by ATP.
   b. The first messenger is not water soluble.
   c. The first messenger binds to many types of cells.
   d. The first messenger cannot cross the plasma membrane.
   e. There are no receptors for the first messenger.
3. Steroid hormones act on target cells by
   a. initiating second messenger activity.
   b. binding to membrane proteins.
   c. initiating DNA transcription.
   d. activating enzymes.
   e. binding to membrane lipids.
4. The major difference between a cell that responds to a signal and one that does not is the presence of a
   a. DNA sequence that binds to the signal.
   b. nearby blood vessel.
   c. receptor.
   d. second messenger.
   e. transduction pathway.
5. Which of the following is *not* a consequence of signal binding to a receptor?
   a. Activation of receptor enzyme activity
   b. Diffusion of receptor in the plasma membrane
   c. Change in conformation of the receptor protein
   d. Breakdown of the receptor to amino acids
   e. Release of the signal from the receptor
6. A nonpolar molecule such as a steroid hormone usually binds to a
   a. cytoplasmic receptor.
   b. protein kinase.
   c. ion channel.
   d. phospholipid.
   e. second messenger.
7. Which of the following is *not* a common type of receptor?
   a. Ion channel
   b. Protein kinase
   c. G protein-linked
   d. Transcription factor
   e. Adenylate cyclase
8. Which of the following is *not* true of the protein kinase cascade?
   a. The signal is amplified.
   b. A second messenger is formed.
   c. Target proteins are phosphorylated.
   d. The cascade ends up in the nucleus.
   e. The cascade begins at the plasma membrane.
9. Which of the following is *not* a second messenger for signal transduction?
   a. Calcium ions
   b. Nitric oxide gas
   c. ATP
   d. Cyclic AMP
   e. Diacylglycerol
10. Plasmodesmata and gap junctions
    a. allow small molecules and ions to pass rapidly between cells.
    b. are both membrane-lined channels.
    c. are channels about 1 μm in diameter.
    d. are present only once per cell.
    e. are involved in cell recognition in signaling.

For Discussion

1. Like *ras* itself, the various components of the Ras signaling pathway were discovered when tumors showed mutations in one or another of the components. What might be the biochemical consequences of mutations in the genes for (a) Raf and (b) MAP kinase that result in rapid cell division?
2. Cyclic AMP is a second messenger in many different responses. How can the same messenger act in different ways in different cells?
3. Compare direct communication via plasmodesmata or gap junctions with ligand/receptor-mediated communication between cells. What are the advantages of one method over the other?
4. The tiny invertebrate *Hydra* has an apical region, which has tentacles, and a long, slender body. *Hydra* can reproduce asexually when cells on the body wall differentiate and form a bud, which then breaks off as a new organism. Buds form only at certain distances from the apex, leading to the idea that the apex releases a molecule that diffuses down the body and, at high concentrations (i.e., near the apex), inhibits bud formation. *Hydra* lacks a circulatory system, so the inhibitor must diffuse from cell to cell. If you had an antibody that binds to connexin to plug up the gap junctions, how would you show that *Hydra*'s inhibitory factor passes *through* these junctions?

General Principles of Cell Communication

Mechanisms enabling one cell to influence the behavior of another almost certainly existed in the world of unicellular organisms long before multicellular organisms appeared on Earth. Evidence comes from studies of present-day unicellular eucaryotes such as yeasts. Although these cells normally lead independent lives, they can communicate and influence one another's behavior in preparation for sexual mating. In the budding yeast *Saccharomyces cerevisiae*, for example, when a haploid individual is ready to mate, it secretes a peptide *mating factor* that signals cells of the opposite mating type to stop proliferating and prepare to mate. The subsequent fusion of two haploid cells of opposite mating types produces a diploid cell, which can then undergo meiosis and sporulate, generating haploid cells with new assortments of genes.

Studies of yeast mutants that are unable to mate have identified many proteins that are required in the **signaling** process. These proteins form a **signaling** network that includes cell-surface receptor proteins, GTP-binding proteins, and proteinkinases, each of which has close relatives among the proteins that carry out **signaling** in animal cells. Through geneduplication and divergence, however, the **signaling** systems in animals have become much more elaborate than those in yeasts.

Extracellular Signal Molecules Bind to Specific Receptors

Yeast cells communicate with one another for mating by secreting a few kinds of small peptides. In contrast, cells in higher animals communicate by means of hundreds of kinds of **signal** molecules. These include proteins, small peptides, amino acids, nucleotides, steroids, retinoids, fatty acid derivatives, and even dissolved gases such as nitric oxide and carbon monoxide. Most of these **signal** molecules are secreted from the ***signaling*** *cell* into the extracellular space byexocytosis. Others are released by diffusion through the plasma membrane, and some are exposed to the extracellular space while remaining tightly bound to the **signaling** cell's surface.

Regardless of the nature of the **signal**, the *target cell* responds by means of a specific protein called a receptor, which specifically binds the **signal** molecule and then initiates a response in the target cell. The extracellular **signal** molecules often act at very low concentrations (typically $\leq 10^{-8}$ M), and the receptors that recognize them usually bind them with high affinity (affinity constant $K_\alpha \geq 10^8$ liters/mole. In most cases, these receptors are transmembrane proteins on the target cell surface. When they bind an extracellular **signal** molecule *(a ligand)*, they become activated and generate a cascade of intracellular signals that alter the behavior of the cell. In other cases, the receptors are inside the target cell, and the **signal** molecule has to enter the cell to activate them: this requires that the **signal** molecules be sufficiently small and hydrophobic to diffuse across the plasma membrane.

Extracellular Signal Molecules Can Act Over Either Short or Long Distances

Many **signal** molecules remain bound to the surface of the **signaling** cell and influence only cells that contact. Such contact-dependent **signaling** is especially important during development and in immune responses. In most cases, however, **signal** molecules are secreted. The secreted molecules may be carried far afield to act on distant targets, or they may act as local mediators, affecting only cells in the immediate environment of the **signaling** cell. This latter process is called paracrine **signaling**. For paracrine signals to be delivered only to their proper target cells, the secreted molecules must not be allowed to diffuse too far; for this reason they are often rapidly taken up by neighboring target cells, destroyed by extracellular enzymes, or immobilized by the extracellular matrix.

For a large, complex multicellular organism, short-range **signaling** is not sufficient on its own to coordinate the behavior of its cells. In these organisms, sets of specialized cells have evolved with a specific role in communication between widely separate parts of the body. The most sophisticated of these are nerve cells, or neurons, which typically extend long processes (axons) that enable them to contact target cells far away. When activated by signals from the environment or from other nerve cells, a neuron sends electrical impulses (action potentials) rapidly along its axon; when such an impulse reaches the end of the axon, it causes the nerve terminals located there to secrete a chemical **signal**called a neurotransmitter. These signals are secreted at specialized cell junctions called *chemical synapses*, which are designed to ensure that the neurotransmitter is delivered specifically to the postsynaptic target cell.

A second type of specialized **signaling** cell that controls the behavior of the organism as a whole is an endocrine cell. These cells secrete their **signal** molecules, called hormones, into the bloodstream, which carries the **signal** to target cells distributed widely throughout the body.

Because endocrine **signaling** relies on diffusion and blood flow, it is relatively slow. Synaptic **signaling**, by contrast, can be much faster, as well as more precise. Nerve cells can transmit information over long distances by electrical impulses that travel at rates of up to 100 meters per second; once released from a nerve terminal, aneurotransmitter has to diffuse less than 100 nm to the target cell, a process that takes less than a millisecond. Another difference between endocrine and synaptic **signaling** is that, whereas hormones are greatly diluted in the bloodstream and interstitial fluid and therefore must be able to act at very low concentrations (typically $< 10^{-8}$ M), neurotransmitters are diluted much less and can achieve high local concentrations. The concentration of *acetylcholine* in the synaptic cleft of an active neuromuscular junction, for example, is about 5×10^{-4} M. Correspondingly, neurotransmitter receptors have a relatively low affinity for their ligand, which means that the neurotransmitter can dissociate rapidly from thereceptor to terminate a response. Moreover, after its release from a nerve terminal, a neurotransmitter is quickly removed from the synaptic cleft, either by specific hydrolytic enzymes that destroy it or by specific membrane transport proteins that pump it back into either the nerve terminal or neighboring glial cells.

Thus, synaptic **signaling** is much more precise than endocrine **signaling**, both in time and in space.

The speed of a response to an extracellular **signal** depends not only on the mechanism of **signal** delivery, but also on the nature of the response in the target cell. Where the response requires only changes in proteins already present in the cell, it can occur in seconds or even milliseconds. When the response involves changes in gene expression and the synthesis of new proteins, however, it usually requires hours, irrespective of the mode of **signal** delivery.

Autocrine Signaling Can Coordinate Decisions by Groups of Identical Cells

All of the forms of **signaling** discussed so far allow one cell to influence another. Often, the **signaling** cell and target are different cell types. Cells, however, can also send signals to other cells of the same type, as well as to themselves. In such autocrine **signaling**, a cell secretes **signal** molecules that can bind back to its own receptors. During development, for example, once a cell has been directed along a particular pathway of differentiation, it may begin to secrete autocrine signals to itself that reinforce this developmental decision.

Autocrine **signaling** is most effective when performed simultaneously by neighboring cells of the same type, and it is likely to be used to encourage groups of identical cells to make the same developmental decisions. Thus, autocrine**signaling** is thought to be one possible mechanism underlying the “community effect” that is observed in earlydevelopment, during which a group of identical cells can respond to a differentiation-inducing **signal** but a single isolated cell of the same type cannot.

Unfortunately, cancer cells often use autocrine **signaling** to overcome the normal controls on cell proliferation and survival that we discuss later. By secreting signals that act back on the cell's own receptors, cancer cells can stimulate their own survival and proliferation and thereby survive and proliferate in places where normal cells of the same type could not. How this dangerous perturbation of normal cell behavior comes about is discussed in Chapter 23.

Go to:

Gap Junctions Allow Signaling Information to Be Shared by Neighboring Cells

Another way to coordinate the activities of neighboring cells is through gap junctions. These are specialized cell-cell junctions that can form between closely apposed plasma membranes and directly connect the cytoplasms of the joined cells via narrow water-filled channels. The channels allow the exchange of small intracellular**signaling** molecules *(intracellular mediators),* such as Ca^{2+} and cyclic AMP (discussed later), but not of macromolecules, such as proteins or nucleic acids. Thus, cells connected by gap junctions can communicate with each other directly, without having to surmount the barrier presented by the intervening plasma membranes.

As discussed in Chapter 19, the pattern of gap-junction connections in a tissue can be revealed either electrically, with intracellular electrodes, or visually, after the microinjection of small water-soluble dyes. Studies of this kind indicate that the cells in a developing embryo make and break gap-junction connections in specific and interesting patterns, strongly suggesting that these junctions have an important role in the **signaling** processes that occur between these cells. Mice and humans that are deficient in one particular gap-junction protein (connexin 43), for example, have severe defects in heart development. Like the autocrine **signaling** described above, gap-junction communication helps adjacent cells of a similar type to coordinate their behavior. It is still not known, however, which particular small molecules are important as carriers of signals through gap junctions, and the specific functions of gap-junction communication in animal development remain uncertain.

Each Cell Is Programmed to Respond to Specific Combinations of Extracellular SignalMolecules

A typical cell in a multicellular organism is exposed to hundreds of different signals in its environment. These signals can be soluble, bound to the extracellular matrix, or bound to the surface of a neighboring cell, and they can act in many millions of combinations. The cell must respond to this babel of signals selectively, according to its own specific character, which it has acquired through progressive cell specialization in the course of development. A cell may be programmed to respond to one combination of signals by differentiating, to another combination by multiplying, and to yet another by performing some specialized function such as contraction or secretion.

Most of the cells in a complex animal are also programmed to depend on a specific combination of signals simply to survive. When deprived of these signals (in a culture dish, for example), a cell activates a suicide program and kills itself—a process called *programmed cell death*, or *apoptosis*. Because different types of cells require different combinations of survival signals, each cell type is restricted to different environments in the body. The ability to undergo apoptosis is a fundamental property of animal cells.

In principle, the hundreds of **signal** molecules that animals make can be used to create an almost unlimited number of**signaling** combinations. The use of these combinations to control cell behavior enables an animal to control its cells in highly specific ways by using a limited diversity of **signal** molecules.

Different Cells Can Respond Differently to the Same Extracellular Signal Molecule

The specific way in which a cell reacts to its environment varies. It varies according to the set of receptor proteins the cell possesses, which determines the particular subset of signals it can respond to, and it varies according to the intracellular machinery by which the cell integrates and interprets the signals it receives. Thus, a single**signal** molecule often has different effects on different target cells. The neurotransmitter acetylcholine, for example, stimulates the

contraction of skeletal muscle cells, but it decreases the rate and force of contraction in heart muscle cells. This is because the acetylcholine receptor proteins on skeletal muscle cells are different from those on heart muscle cells. But receptor differences are not always the explanation for the different effects. In many cases, the same **signal** molecule binds to identical receptor proteins yet produces very different responses in different types of target cells, reflecting differences in the internal machinery to which the receptors are coupled.

The Concentration of a Molecule Can Be Adjusted Quickly Only If the Lifetime of the Molecule Is Short

It is natural to think of **signaling** systems in terms of the changes produced when a **signal** is delivered. But it is just as important to consider what happens when a **signal** is withdrawn. During development, transient signals often produce lasting effects: they can trigger a change in the cell's development that persists indefinitely, through cell memory mechanism. In most cases in adult tissues, however, the response fades when a **signal** ceases. The effect is transitory because the **signal** exerts its effects by altering a set of molecules that are unstable, undergoing continual turnover. Thus, once the **signal** is shut off, the replacement of the old molecules by new ones wipes out all traces of its action. It follows that the speed with which a cell responds to **signal** removal depends on the rate of destruction, or turnover, of the molecules the **signal** affects.

It is also true, although much less obvious, that this turnover rate also determines the promptness of the response when a**signal** is turned on. Consider, for example, two intracellular **signaling** molecules X and Y, both of which are normally maintained at a concentration of 1000 molecules per cell. Molecule Y is synthesized and degraded at a rate of 100 molecules per second, with each molecule having an average lifetime of 10 seconds. Molecule X has a turnover rate that is 10 times slower than that of Y: it is both synthesized and degraded at a rate of 10 molecules per second, so that each molecule has an average lifetime in the cell of 100 seconds. If a **signal** acting on the cell boosts the rates of synthesis of both X and Y tenfold without any change in the molecular lifetimes, at the end of 1 second the concentration of Y will have increased by nearly 900 molecules per cell (10 × 100 - 100), while the concentration of X will have increased by only 90 molecules per cell. In fact, after a molecule's synthesis rate has been either increased or decreased abruptly, the time required for the molecule to shift halfway from its old to its new equilibrium concentration is equal to its normal half-life—that is, equal to the time that would be required for its concentration to fall by half if all synthesis were stopped.

The same principles apply to proteins and small molecules, and to molecules in the extracellular space and inside cells. Many intracellular proteins have short half-lives, some surviving for less than 10 minutes. In most cases, these are proteins with key regulatory roles, whose concentrations are rapidly regulated in the cell by changes in their rates of synthesis. Likewise, any covalent modifications of proteins that occur as part of a

rapid **signaling** process—most commonly, the addition of a phosphate group to an amino acid side chain—must be continuously removed at a rapid rate to make rapid **signaling** possible.

We shall discuss some of these molecular events in detail later for **signaling** pathways that operate via cell-surface receptors. But the principles apply quite generally, as the next example illustrates.

Nitric Oxide Gas Signals by Binding Directly to an Enzyme Inside the Target Cell

Although most extracellular signals are hydrophilic molecules that bind to receptors on the surface of the target cell, some **signal** molecules are hydrophobic enough and/or small enough to pass readily across the target-cell plasma membrane. Once inside, they directly regulate the activity of a specific intracellular protein. An important and remarkable example is the gas **nitric oxide (NO)**, which acts as a **signal** molecule in both animals and plants. In mammals, one of its functions is to regulate smooth muscle contraction. Acetylcholine, for example, is released by autonomic nerves in the walls of a blood vessel, and it causes smooth muscle cells in the vessel wall to relax. Theacetylcholine acts indirectly by inducing the nearby endothelial cells to make and release NO, which then signals the underlying smooth muscle cells to relax. This effect of NO on blood vessels provides an explanation for the mechanism of action of nitroglycerine, which has been used for about 100 years to treat patients with angina (pain resulting from inadequate blood flow to the heart muscle). The nitroglycerine is converted to NO, which relaxes blood vessels. This reduces the workload on the heart and, as a consequence, it reduces the oxygen requirement of the heart muscle.

Many types of nerve cells use NO gas to **signal** to their neighbors. The NO released by autonomic nerves in the penis, for example, causes the local blood vessel dilation that is responsible for penile erection. NO is also produced as a local mediator by activated macrophages and neutrophils to help them to kill invading microorganisms. In plants, NO is involved in the defensive responses to injury or infection.

NO gas is made by the deamination of the amino acid arginine, catalyzed by the enzyme *NO synthase.* Because it passes readily across membranes, dissolved NO rapidly diffuses out of the cell where it is produced and into neighboring cells. It acts only locally because it has a short half-life—about 5–10 seconds—in the extracellular space before it is converted to nitrates and nitrites by oxygen and water. In many target cells, including endothelial cells, NO binds to iron in the active site of the enzyme *guanylyl cyclase,* stimulating this enzyme to produce the small intracellular mediator *cyclic GMP*, which we discuss later. The effects of NO can occur within seconds, because the normal rate of turnover of cyclic GMP is high: a rapid degradation to GMP by a phosphodiesterase constantly balances the production of cyclic GMP from GTP by guanylyl cyclase. The drug Viagra inhibits this cyclic GMP phosphodiesterase in the penis, thereby increasing the amount of time that cyclic GMP levels

remain elevated after NO production is induced by local nerve terminals. The cyclic GMP, in turn, keeps blood vessels relaxed and the penis erect.

Carbon monoxide (CO) is another gas that is used as an intercellular **signal**. It can act in the same way as NO, by stimulating guanylyl cyclase. These gases are not the only **signal** molecules that can pass directly across the target-cellplasma membrane. A group of small, hydrophobic, nongaseous hormones and local mediators also enter target cells in this way. But instead of binding to enzymes, they bind to intracellular receptor proteins that directly regulate genetranscription, as we discuss next.

Nuclear Receptors Are Ligand-activated Gene Regulatory Proteins

A number of small hydrophobic **signal** molecules diffuse directly across the plasma membrane of target cells and bind to intracellular receptor proteins. These **signal** molecules include *steroid hormones, thyroid hormones, retinoids,* and*vitamin D.* Although they differ greatly from one another in both chemical structure and function, they all act by a similar mechanism. When these **signal** molecules bind to their receptor proteins, they activate the receptors, which bind to DNA to regulate the transcription of specific genes. The receptors are all structurally related, being part of the nuclear receptor superfamily. This very large superfamily also includes some receptor proteins that are activated by intracellular metabolites rather than by secreted **signal** molecules. Many family members have been identified by DNA sequencing only, and their ligand is not yet known; these proteins are therefore referred to as *orphan nuclear receptors.*The importance of such nuclear receptors in some animals is indicated by the fact that 1–2% of the genes in the nematode *C. elegans* code for them, although there are fewer than 50 in humans.

Steroid hormones—which include cortisol, the steroid sex hormones, vitamin D (in vertebrates), and the moultinghormone ecdysone (in insects)—are all made from cholesterol. *Cortisol* is produced in the cortex of the adrenal glands and influences the metabolism of many types of cells. The steroid sex hormones are made in the testes and ovaries, and they are responsible for the secondary sex characteristics that distinguish males from females. Vitamin D is synthesized in the skin in response to sunlight; after it has been converted to its active form in the liver or kidneys, it regulates Ca^{2+}metabolism, promoting Ca^{2+} uptake in the gut and reducing its excretion in the kidneys. The thyroid hormones, which are made from the amino acid tyrosine, act to increase the metabolic rate in a wide variety of cell types, while the retinoids, such as retinoic acid, are made from vitamin A and have important roles as local mediators in vertebratedevelopment. Although all of these **signal** molecules are relatively insoluble in water, they are made soluble for transport in the bloodstream and other extracellular fluids by binding to specific carrier proteins, from which they dissociate before entering a target cell.

Beside a fundamental difference in the way they **signal** their target cells, most water-insoluble **signal** molecules differ from water-soluble ones in the length of time they persist in the bloodstream or tissue fluids. Most water-soluble hormones are removed and/or broken

down within minutes of entering the blood, and local mediators and neurotransmitters are removed from the extracellular space even faster—within seconds or milliseconds. Steroid hormones, by contrast, persist in the blood for hours and thyroid hormones for days. Consequently, water-soluble **signal**molecules usually mediate responses of short duration, whereas water-insoluble ones tend to mediate responses that are longer lasting.

The intracellular receptors for the steroid and thyroid hormones, retinoids, and vitamin D all bind to specific DNAsequences adjacent to the genes the ligand regulates. Some receptors, such as those for cortisol, are located primarily in the cytosol and enter the nucleus after ligand binding; others, such as the thyroid and retinoid receptors, are bound to DNA in the nucleus even in the absence of ligand. In either case, the inactive receptors are bound to inhibitory proteincomplexes, and ligand binding alters the conformation of the receptor protein, causing the inhibitory complex to dissociate. The ligand binding also causes the receptor to bind to coactivator proteins that induce gene transcription. The transcriptional response usually takes place in successive steps: the direct activation of a small number of specific genes occurs within about 30 minutes and constitutes the *primary response;* the protein products of these genes in turn activate other genes to produce a delayed, *secondary response;* and so on. In this way, a simple hormonal trigger can cause a very complex change in the pattern of gene expression.

The responses to steroid and thyroid hormones, vitamin D, and retinoids, like responses to extracellular signals in general, are determined as much by the nature of the target cell as by the nature of the **signal** molecule. Many types of cells have the identical intracellular receptor, but the set of genes that the receptor regulates is different in each cell type. This is because more than one type of gene regulatory protein generally must bind to a eucaryotic gene to activate its transcription. An intracellular receptor can therefore activate a gene only if there is the right combination of other gene regulatory proteins, and many of these are cell-type specific. Thus, each of these hormones induces a characteristic set of responses in an animal for two reasons. First, only certain types of cells have receptors for it. Second, each of these cell types contains a different combination of other cell-type-specific gene regulatory proteins that collaborate with the activated receptor to influence the transcription of specific sets of genes.

The Three Largest Classes of Cell-Surface Receptor Proteins Are Ion-Channel-linked, G-Protein-linked, and Enzyme-linked Receptors

As mentioned previously, all water-soluble **signal** molecules (including neurotransmitters and all **signal** proteins) bind to specific receptor proteins on the surface of the target cells that they influence. These cell-surface receptor proteins act as***signal*** *transducers.* They convert an extracellular ligand-binding event into intracellular signals that alter the behavior of the target cell.

Most cell-surface receptor proteins belong to one of three classes, defined by the **transduction** mechanism they use.**Ion-channel-linked receptors**, also known

as *transmitter-gated ion channels* or *ionotropic receptors,* are involved in rapid synaptic **signaling** between electrically excitable cells. This type of **signaling** is mediated by a small number of neurotransmitters that transiently open or close an ion channel formed by the protein to which they bind, briefly changing the ion permeability of the plasma membrane and thereby the excitability of the postsynaptic cell. The ion-channel-linked receptors belong to a large family of homologous, multipass transmembrane proteins.

G-protein-linked receptors act indirectly to regulate the activity of a separate plasma-membrane-bound target protein, which can be either an enzyme or an ion channel. The interaction between the receptor and this target protein is mediated by a third protein, called a *trimeric GTP-binding protein (G protein).* The activation of the target protein can change the concentration of one or more intracellular mediators (if the target protein is an enzyme), or it can change the ion permeability of the plasma membrane (if the target protein is an ion channel). The intracellular mediators affected act in turn to alter the behavior of yet other **signaling** proteins in the cell. All of the G-protein-linked receptors belong to a large family of homologous, seven-pass transmembrane proteins.

Enzyme-linked receptors, when activated, either function directly as enzymes or are directly associated with enzymes that they activate. They are formed by single-pass transmembrane proteins that have their ligand-binding site outside the cell and their catalytic or enzyme-binding site inside. Enzyme-linked receptors are heterogeneous in structure compared with the other two classes. The great majority, however, are protein kinases, or are associated with protein kinases, and ligand binding to them causes the phosphorylation of specific sets of proteins in the target cell.

There are some cell-surface receptors that do not fit into any of the above classes. Some of these depend on intracellular proteolytic events to **signal** the cell, and we discuss them only after we explain in detail how G-protein-linked receptors and enzyme-linked receptors operate. We start with some general principles of **signaling** via cell-surface receptors.

Most Activated Cell-Surface Receptors Relay Signals Via Small Molecules and a Network of Intracellular Signaling Proteins

Signals received at the surface of a cell by either G-protein-linked or enzyme-linked receptors are relayed into the cell interior by a combination of small and large *intracellular **signaling** molecules.* The resulting chain of intracellular**signaling** events ultimately alters *target proteins,* and these altered target proteins are responsible for modifying the behavior of the cell.

The small intracellular **signaling** molecules are called **small intracellular mediators**, or second messengers (the “first messengers” being the extracellular signals). They are generated in large numbers in response to receptor activation and rapidly diffuse away from their source, broadcasting the **signal** to other parts of the cell. Some, such as *cyclic AMP* and Ca^{2+}, are water-soluble and diffuse in the cytosol, while others, such as *diacylglycerol,* are lipid-soluble and diffuse in the plane of the plasma membrane. In either

case, they pass the **signal** on by binding to and altering the behavior of selected **signaling** proteins or target proteins.

The large intracellular **signaling** molecules are intracellular **signaling** proteins. Many of these relay the **signal** into the cell by either activating the next **signaling** protein in the chain or generating small intracellular mediators. These proteins can be classified according to their particular function, although many fall into more than one category:

1. *Relay proteins* simply pass the message to the next **signaling** component in the chain.
2. *Messenger proteins* carry the **signal** from one part of the cell to another, such as from the cytosol to the nucleus.
3. *Adaptor proteins* link one **signaling** protein to another, without themselves conveying a **signal**.
4. *Amplifier proteins,* which are usually either enzymes or ion channels, greatly increase the **signal** they receive, either by producing large amounts of small intracellular mediators or by activating large numbers of downstream intracellular **signaling** proteins. When there are multiple amplification steps in a relay chain, the chain is often referred to as a **signaling cascade**.
5. *Transducer proteins* convert the **signal** into a different form. The enzyme that makes cyclic AMP is an example: it both converts the **signal** and amplifies it, thus acting as both a transducer and an amplifier.
6. *Bifurcation proteins* spread the **signal** from one **signaling** pathway to another.
7. *Integrator proteins* receive signals from two or more **signaling** pathways and integrate them before relaying a **signal**onward.
8. *Latent gene regulatory proteins* are activated at the cell surface by activated receptors and then migrate to thenucleus to stimulate gene transcription.

Other types of intracellular proteins also have important roles in intracellular**signaling**. *Modulator proteins* modify the activity of intracellular **signaling** proteins and thereby regulate the strength of**signaling** along the pathway. *Anchoring proteins* maintain specific **signaling** proteins at a precise location in the cell by tethering them to a membrane or the cytoskeleton. *Scaffold proteins* are adaptor and/or anchoring proteins that bind multiple **signaling** proteins together in a functional complex and often hold them at a specific location.

Some Intracellular Signaling Proteins Act as Molecular Switches

Many intracellular **signaling** proteins behave like **molecular switches:** on receipt of a **signal** they switch from an inactive to an active state, until another process switches them off. As we discussed earlier, the switching off is just as important as the switching on. If a **signaling** pathway is to recover after transmitting a **signal** so that it can be ready to transmit another, every activated molecule in the pathway must be returned to its original inactivated state.

The molecular switches fall into two main classes that operate in different ways, although in both cases it is the gain or loss of phosphate groups that determines whether the protein is active or inactive. The largest class consists of proteins that are activated or inactivated by phosphorylation (discussed in Chapter 3). For these proteins, the switch is thrown in one direction by a protein kinase, which adds one or more phosphate groups to the **signaling** protein, and in the other direction by a protein phosphatase, which removes the phosphate groups from the protein. It is estimated that one-third of the proteins in a eucaryotic cell are phosphorylated at any given time.

Many of the **signaling** proteins controlled by phosphorylation are themselves protein kinases, and these are often organized into **phosphorylation cascades**. One protein kinase, activated by phosphorylation, phosphorylates the next protein kinase in the sequence, and so on, relaying the **signal** onward and, in the process, amplifying it and sometimes spreading it to other **signaling** pathways. Two main types of protein kinases operate as intracellular **signaling** proteins. The great majority are *serine/threonine kinases,* which phosphorylate proteins on serines and (less often) threonines. Others are *tyrosine kinases,* which phosphorylate proteins on tyrosines. An occasional kinase can do both. Genome sequencing reveals that about 2% of our genes encode protein kinases, and it is thought that hundreds of distinct types of protein kinases are present in a typical mammalian cell.

The other main class of molecular switches involved in **signaling** are **GTP-binding proteins** (discussed in Chapter 3). These switch between an active state when GTP is bound and an inactive state when GDP is bound. Once activated, they have intrinsic GTPase activity and shut themselves off by hydrolyzing their bound GTP to GDP. There are two major types of GTP-binding proteins—large *trimeric GTP-binding proteins* (also called *G proteins),*which relay the signals from G-protein-linked receptors, and

small *monomeric GTPases* (also called *monomeric GTP-binding proteins).* The latter also help to relay intracellular signals, but in addition they are involved in regulating vesicular traffic and many other processes in eucaryotic cells.

As discussed earlier, complex cell behaviors, such as cell survival and cell proliferation, are generally stimulated by specific combinations of extracellular signals rather than by a single **signal** acting alone. The cell therefore has to integrate the information coming from separate signals so as to make an appropriate response—to live or die, to divide or not, and so on. This integration usually depends on integrator proteins, which are equivalent to the microprocessors in a computer: they require multiple **signal** inputs to produce an output that causes the desired biological effect. Two examples that show how such integrator proteins can operate are illustrated in.

Intracellular Signaling Complexes Enhance the Speed, Efficiency, and Specificity of the Response

Even a single type of extracellular **signal** acting through a single type of G-protein-linked or enzyme-linked receptorusually activates multiple parallel **signaling** pathways and can thereby influence multiple aspects of cell behavior—such as shape, movement, metabolism, and gene expression. Indeed, these two main classes of cell-surface receptors often activate some of the same **signaling** pathways, and there is usually no obvious reason why a particular extracellular**signal** utilizes one class of receptors rather than the other.

The complexity of these **signal**-response systems, with multiple interacting relay chains of **signaling** proteins, is daunting. It is not clear how an individual cell manages to display specific responses to so many different extracellular signals, many of which bind to the same class of receptor and activate many of the same **signaling** pathways. One strategy that the cell uses to achieve specificity involves scaffold proteins, which organize groups of interacting **signaling** proteins into ***signaling*** *complexes*. Because the scaffold guides the interactions between the successive components in such a complex, the **signal** can be relayed with precision, speed, and efficiency; moreover, unwanted cross-talk between **signaling** pathways is avoided. In order to amplify a **signal**, however, and spread it to other parts of the cell, at least some of the components in most **signaling** pathways are likely to be freely diffusible.

In other cases, **signaling** complexes form only transiently, as when **signaling** proteins assemble around a receptor after an extracellular **signal** molecule has activated it. In some of these cases, the cytoplasmic tail of the activated receptor is phosphorylated during the activation process, and the phosphorylated amino acids then serve as docking sites for the assembly of other **signaling** proteins (Figure 15-19B). In yet other cases, receptor activation leads to the production of modified phospholipid molecules in the adjacent plasma membrane, and these lipids then recruit specific intracellular**signaling** proteins to this region

of membrane. All such **signaling** complexes form only transiently and rapidly disassemble after the extracellular ligand dissociates from the receptor.

Interactions Between Intracellular Signaling Proteins Are Mediated by Modular Binding Domains

The assembly of both stable and transient **signaling** complexes depends on a variety of highly conserved, small **binding domains** that are found in many intracellular **signaling** proteins. Each of these compact protein modules binds to a particular structural motif in the protein (or lipid) with which the **signaling** protein interacts. Because of these modular domains, **signaling** proteins bind to one another in multiple combinations, like Lego bricks, with the proteins often forming a three-dimensional network of interactions that determines the route followed by the **signaling** pathway. By joining existing domains together in novel combinations, the use of such modular binding domains has presumably facilitated the rapid evolution of new **signaling** pathways.

Src homology 2 (SH2) domains and *phosphotyrosine-binding (PTB) domains*, for example, bind to phosphorylated tyrosines in a particular peptide sequence on activated receptors or intracellular **signaling** proteins. *Src homology 3 (SH3)* domains bind to a short proline-rich amino acid sequence. *Pleckstrin homology (PH)* domains (first described in the Pleckstrin protein in blood platelets) bind to the charged headgroups of specific phosphorylated inositol phospholipids that are produced in the plasma membrane in response to an extracellular **signal**; they thereby enable the protein they are part of to dock on the membrane and interact with other recruited **signaling** proteins. Some **signaling**proteins function only as adaptors to link two other proteins together in a **signaling** pathway, and they consist solely of two or more binding domains.

Scaffold proteins often contain multiple *PDZ domains* (originally found in a region of a synapse called the postsynaptic density), each of which binds to a specific motif on a receptor or **signaling** protein. The *InaD* scaffold protein in*Drosophila* photoreceptor cells is a striking example. It contains five PDZ domains, one of which binds a light-activatedion channel, while the others each bind to a different **signaling** protein involved in the response of the cell to light. If any of these PDZ domains are missing, the corresponding **signaling** protein fails to assemble in the complex, and the fly's vision is defective.

Some cell-surface receptors and intracellular **signaling** proteins are thought to cluster together transiently in specific microdomains in the lipid bilayer of the plasma membrane that are enriched in cholesterol and glycolipids. Some of the proteins are directed to these lipid rafts by covalently attached lipid molecules. Like scaffold proteins, these lipid scaffolds may promote speed and efficiency in the **signaling** process by serving as sites where **signaling** molecules can assemble and interact.

Cells Can Respond Abruptly to a Gradually Increasing Concentration of an ExtracellularSignal

Some cellular responses to extracellular **signal** molecules are smoothly graded in simple proportion to the concentration of the molecule. The primary responses to steroid hormones (see Figure 15-14) often follow this pattern, presumably because the nuclear hormone receptor protein binds a single molecule of hormone and each specific DNA recognition sequence in a steroid-hormone-responsive gene acts independently. As the concentration of hormone increases, the concentration of activated receptor-hormone complexes increases proportionally, as does the number of complexes bound to specific recognition sequences in the responsive genes; the response of the cell is therefore a gradual and linear one.

Many responses to extracellular **signal** molecules, however, begin more abruptly as the concentration of the moleculeincreases. Some may even occur in a nearly all-or-none manner, being undetectable below a threshold concentration of the molecule and then reaching a maximum as soon as this concentration is exceeded. What might be the molecular basis for such steep or even switchlike responses to graded signals?

One mechanism for sharpening the response is to require that more than one intracellular effector molecule or complexbind to some target macromolecule to induce a response. In some steroid-hormone-induced responses, for example, it seems that more than one activated receptor-hormone complex must bind simultaneously to specific regulatory sequences in the DNA to activate a particular gene. As a result, as the hormone concentration rises, gene activation begins more abruptly than it would if only one bound complex were sufficient for activation. A similar cooperative mechanism often operates in the **signaling** cascades activated by cell-surface receptors. As we discuss later, four molecules of the small intracellular mediator cyclic AMP, for example, must bind to each molecule of cyclic-AMP-dependent protein kinase to activate the kinase. Such responses become sharper as the number of cooperating molecules increases, and if the number is large enough, responses approaching the all-or-none type can be achieved.

Responses are also sharpened when an intracellular **signaling** molecule activates one enzyme and, at the same time, inhibits another enzyme that catalyzes the opposite reaction. A well-studied example of this common type of regulation is the stimulation of glycogen breakdown in skeletal muscle cells induced by the hormone *adrenaline* (epinephrine). Adrenaline's binding to a G-protein-linked cell-surface receptor leads to an increase in intracellular cyclic AMP concentration, which both activates an enzyme that promotes glycogen breakdown and inhibits an enzyme that promotes glycogen synthesis.

All of these mechanisms can produce responses that are very steep but, nevertheless, always smoothly graded according to the concentration of the extracellular **signal** molecule. Another mechanism, however, can produce true all-or-none responses, such that raising the **signal** above a critical threshold level trips a sudden switch in the responding cell. All-or-none threshold responses of this type generally depend on *positive feedback;* by this mechanism, nerve and muscle cells generate all-or-none *action potentials* in response to

neurotransmitters (discussed in Chapter 11). The activation ofion-channel-linked acetylcholine receptors at a neuromuscular junction, for example, results in a net influx of Na^+ that locally depolarizes the muscle plasma membrane. This causes voltage-gated Na^+ channels to open in the same membrane region, producing a further influx of Na^+, which further depolarizes the membrane and thereby opens more Na^+ channels. If the initial depolarization exceeds a certain threshold value, this positive feedback has an explosive "runaway" effect, producing an action potential that propagates to involve the entire muscle membrane.

An accelerating positive feedback mechanism can also operate through **signaling** proteins that are enzymes rather thanion channels. Suppose, for example, that a particular intracellular **signaling** ligand activates an enzyme located downstream in a **signaling** pathway and that two or more molecules of the product of the enzymatic reaction bind back to the same enzyme to activate it further. The consequence is a very low rate of synthesis of the product in the absence of the ligand. The rate increases slowly with the concentration of ligand until, at some threshold level of ligand, enough of the product has been synthesized to activate the enzyme in a self-accelerating, runaway fashion. The concentration of the product then suddenly increases to a much higher level. Through these and a number of other mechanisms not discussed here, the cell will often translate a gradual change in the concentration of a **signaling** ligand into a switchlike change, creating an all-or-none response by the cell.

A Cell Can Remember The Effect of Some Signals

The effect of an extracellular **signal** on a target cell can, in some cases, persist well after the **signal** has disappeared. The enzymatic accelerating positive feedback system just described represents one type of mechanism that displays this kind of persistence. If such a system has been switched on by raising the concentration of intracellular activating ligand above threshold, it will generally remain switched on even when the extracellular **signal** disappears; instead of faithfully reflecting the current level of **signal**, the response system displays a memory. We shall encounter a specific example of this later, when we discuss a protein kinase that is activated by Ca^{2+} to phosphorylate itself and other proteins; the autophosphorylation keeps the kinase active long after Ca^{2+} levels return to normal, providing a memory trace of the initial **signal**.

Transient extracellular signals often induce much longer-term changes in cells during the development of a multicellular organism. Some of these changes can persist for the lifetime of the organism. They usually depend on self-activating memory mechanisms that operate further downstream in a **signaling** pathway, at the level of gene transcription. The signals that trigger muscle cell determination, for example, turn on a series of muscle-specific gene regulatory proteins that stimulate the transcription of their own genes, as well as genes producing many other muscle cell proteins. In this way, the decision to become a muscle cell is made permanent.

Cells Can Adjust Their Sensitivity to a Signal

In responding to many types of stimuli, cells and organisms are able to detect the same percentage of change in a **signal**over a very wide range of stimulus intensities. This requires that the target cells undergo a reversible process ofadaptation, or **desensitization**, whereby a prolonged exposure to a stimulus decreases the cells' response to that level of exposure. In chemical **signaling**, adaptation enables cells to respond to *changes* in the concentration of a **signaling**ligand (rather than to the absolute concentration of the ligand) over a very wide range of ligand concentrations. The general principle is one of a negative feedback that operates with a delay. A strong response modifies the machinery for making that response, such that the machinery resets itself to an off position. Owing to the delay, however, a sudden change in the stimulus is able to make itself felt strongly for a short period before the negative feedback has time to kick in.

Desensitization to a **signal** molecule can occur in various ways. Ligand binding to cell-surface receptors, for example, may induce their endocytosis and temporary sequestration in endosomes. Such ligand-induced receptor endocytosis can lead to the destruction of the receptors in lysosomes, a process referred to as *receptor down-regulation.* In other cases,desensitization results from a rapid inactivation of the receptors—for example, as a result of a receptor phosphorylationthat follows its activation, with a delay. Desensitization can also be caused by a change in a protein involved in transducing the **signal** or by the production of an inhibitor that blocks the **transduction** process.

Having discussed some of the general principles of cell **signaling**, we now turn to the G-protein-linked receptors. These are by far the largest class of cell-surface receptors, and they mediate the responses to the great majority of extracellular signals. This superfamily of receptor proteins not only mediates intercellular communication; it is also central to vision, smell, and taste perception.

Summary

Each cell in a multicellular animal has been programmed during development to respond to a specific set of extracellular signals produced by other cells. These signals act in various combinations to regulate the behavior of the cell. Most of the signals mediate a form of **signaling** in which local mediators are secreted, but then are rapidly taken up, destroyed, or immobilized, so that they act only on neighboring cells. Other signals remain bound to the outer surface of the**signaling** cell and mediate contact-dependent **signaling**. Centralized control is exerted both by endocrine **signaling**, in which hormones secreted by endocrine cells are carried in the blood to target cells throughout the body, and by synaptic**signaling**, in which neurotransmitters secreted by nerve cell axons act locally on the postsynaptic cells that the axons contact.

Cell **signaling** requires not only extracellular **signal** molecules, but also a complementary set of receptor proteins in each cell that enable it to bind and respond to the **signal** molecules in a characteristic way. Some small hydrophobic **signal**molecules, including steroid and thyroid

hormones, diffuse across the plasma membrane of the target cell and activate intracellular receptor proteins that directly regulate the transcription of specific genes. The dissolved gases nitric oxideand carbon monoxide act as local mediators by diffusing across the plasma membrane of the target cell and activating an intracellular enzyme—usually guanylyl cyclase, which produces cyclic GMP in the target cell. But most extracellular**signal** molecules are hydrophilic and can activate receptor proteins only on the surface of the target cell; these receptors act as **signal** transducers, converting the extracellular binding event into intracellular signals that alter the behavior of the target cell.

There are three main families of cell-surface receptors, each of which transduces extracellular signals in a different way. Ion-channel-linked receptors are transmitter-gated ion channels that open or close briefly in response to the binding of aneurotransmitter. G-protein-linked receptors indirectly activate or inactivate plasma-membrane-bound enzymes or ion channels via trimeric GTP-binding proteins (G proteins). Enzyme-linked receptors either act directly as enzymes or are associated with enzymes; these enzymes are usually protein kinases that phosphorylate specific proteins in the target cell.

Once activated, enzyme- and G-protein-linked receptors relay a **signal** into the cell interior by activating chains of intracellular **signaling** proteins; some transduce, amplify, or spread the **signal** as they relay it, while others integrate signals from different **signaling** pathways. Many of these **signaling** proteins function as switches that are transiently activated by phosphorylation or GTP binding. Functional **signaling** complexes are often formed by means of modular binding domains in the **signaling** proteins; these domains allow complicated protein assemblies to function in **signaling**networks.

Target cells can use a variety of intracellular mechanisms to respond abruptly to a gradually increasing concentration of an extracellular **signal** or to convert a short-lasting **signal** into a long-lasting response. In addition, through adaptation, they can often reversibly adjust their sensitivity to a **signal** to allow the cells to respond to changes in the concentration of a particular **signal** molecule over a large range of concentrations.

Signaling through G-Protein-Linked Cell-Surface Receptors

G-protein-linked receptors form the largest family of cell-surface receptors and are found in all eucaryotes. About 5% of the genes in the nematode *C. elegans*, for example, encode such receptors, and thousands have already been defined in mammals; in mice, there are about 1000 concerned with the sense of smell alone. G-protein-linked receptors mediate the responses to an enormous diversity of signal molecules, including hormones, neurotransmitters, and local mediators. These signal molecules that activate them are as varied in structure as they are in function: the list includes proteins and small peptides, as well as derivatives of amino acids and fatty acids. The same ligand can activate many differentreceptor family members; at least 9 distinct G-protein-linked receptors are activated by adrenaline, for example, another 5 or more by acetylcholine, and at least 15 by the neurotransmitter serotonin.

Despite the chemical and functional diversity of the signal molecules that bind to them, all G-protein-linked receptors have a similar structure. They consist of a single polypeptide chain that threads back and forth across the lipid bilayerseven times and are therefore sometimes called *serpentine receptors*. In addition to their characteristic orientation in the plasma membrane, they have the same functional relationship to the G proteins they use to signal the cell interior that an extracellular ligand is present.

As we discuss later, this superfamily of seven-pass transmembrane proteins includes *rhodopsin*, the light-activatedprotein in the vertebrate eye, as well as the large number of olfactory receptors in the vertebrate nose. Other family members are found in unicellular organisms: the receptors in yeasts that recognize secreted mating factors are an example. In fact, it is thought that the G-protein-linked receptors that mediate cell-cell signaling in multicellular organisms evolved from sensory receptors that were possessed by their unicellular eucaryotic ancestors.

It is remarkable that about half of all known drugs work through G-protein-linked receptors. Genome sequencing projects are revealing vast numbers of new family members, many of which are likely targets for new drugs that remain to be discovered.

Trimeric G Proteins Disassemble to Relay Signals from G-Protein-linked Receptors

When extracellular signaling molecules bind to serpentine receptors, the receptors undergo a conformational change that enables them to activate **trimeric GTP-binding proteins (G proteins)**. These G proteins are attached to the cytoplasmic face of the plasma membrane, where they serve as relay molecules, functionally coupling the receptors to enzymes or ion channels in this membrane. There are various types of G proteins, each specific for a particular set of serpentine receptors and for a particular set of downstream target proteins in the plasma membrane. All have a similar structure, however, and they operate in a similar way.

G proteins are composed of three protein subunits—α, β, and γ. In the unstimulated state, the α subunit has GDP bound and the G protein is inactive. When stimulated by an activated receptor, the α subunit releases its bound GDP, allowing GTP to bind in its place. This exchange causes the trimer to dissociate into two activated components—an *α subunit* and a *βγ complex*.

The dissociation of the trimeric G protein activates its two components in different ways. GTP binding causes a conformational change that affects the surface of the α subunit that associates with the βγ complex in the trimer. This change causes the release of the βγ complex, but it also causes and the α subunit to adopt a new shape that allows it to interact with its target proteins. The βγ complex does not change its conformation, but the surface previously masked by the α subunit is now available to interact with a second set of target proteins. The targets of the dissociated components of the G protein are either enzymes or ion channels in the plasma membrane, and they relay the signal onward.

The α subunit is a GTPase, and once it hydrolyzes its bound GTP to GDP, it reassociates with a βγ complex to re-form an inactive G protein, reversing the activation process. The time during which the α subunit and βγ complex remain apart and active is usually short, and it depends on how quickly the α subunit hydrolyzes its bound GTP. An isolated α subunit is an inefficient GTPase, and, left to its own devices, the subunit would inactivate only after several minutes. Its activation is usually reversed much faster than this, however, because the GTPase activity of the α subunit is greatly enhanced by the binding of a second protein, which can be either its target protein or a specific modulator known as a *regulator of G protein signaling (RGS)*. **RGS proteins** act as α-subunit-specific *GTPase activating proteins (GAPs),* and they are thought to have a crucial role in shutting off G-protein-mediated responses in all eucaryotes. There are about 25 RGS proteins encoded in the human genome, each of which is thought to interact with a particular set of G proteins.

The importance of the GTPase activity in shutting off the response can be easily demonstrated in a test tube. If cells are broken open and exposed to an analogue of GTP (GTPγS) in which the terminal phosphate cannot be hydrolyzed, the activated α subunits remain active for a very long time.

Some G Proteins Signal By Regulating the Production of Cyclic AMP

Cyclic AMP (cAMP) was first identified as a small intracellular mediator in the 1950s. It has since been found to act in this role in all procaryotic and animal cells that have been studied. The normal concentration of cyclic AMP inside the cell is about 10^{-7} M, but an extracellular signal can cause cyclic AMP levels to change by more than twentyfold in seconds (Figure 15-30). As explained earlier (see Figure 15-10), such a rapid response requires that a rapid synthesis of the molecule be balanced by its rapid breakdown or removal. In fact, cyclic AMP is synthesized from ATP by a plasma-membrane-bound enzyme adenylyl cyclase, and it is rapidly and continuously destroyed by one or more **cyclic AMP phosphodiesterases** that hydrolyze cyclic AMP to adenosine 5′-monophosphate (5′-AMP).

Many extracellular signal molecules work by increasing cyclic AMP content, and they do so by increasing the activity of adenylyl cyclase rather than decreasing the activity of phosphodiesterase. Adenylyl cyclase is a large multipass transmembrane protein with its catalytic domain on the cytosolic side of the plasma membrane. There are at least eight isoforms in mammals, most of which are regulated by both G proteins and Ca^{2+}. All receptors that act via cyclic AMP are coupled to a **stimulatory G protein (G_s)**, which activates adenylyl cyclase and thereby increases cyclic AMP concentration. Another G protein, called **inhibitory G protein (G_i)**, inhibits adenylyl cyclase, but it mainly acts by directly regulating ion channels (as we discuss later) rather than by decreasing cyclic AMP content. Although it is usually the α subunit that regulates the cyclase, the βγ complex sometimes does so as well, either increasing or decreasing the enzyme's activity, depending on the particular βγ complex and the isoform of the cyclase.

Both G_s and G_i are targets for some medically important bacterial toxins. *Cholera toxin*, which is produced by the bacterium that causes cholera, is an enzyme that catalyzes the transfer of ADP ribose from intracellular NAD^+ to the αsubunit of G_s. This ADP ribosylation alters the α subunit so that it can no longer hydrolyze its bound GTP, causing it to remain in an active state that stimulates adenylyl cyclase indefinitely. The resulting prolonged elevation in cyclic AMP levels within intestinal epithelial cells causes a large efflux of Cl^- and water into the gut, thereby causing the severe diarrhea that characterizes cholera. *Pertussis toxin*, which is made by the bacterium that causes pertussis (whooping cough), catalyzes the ADP ribosylation of the α subunit of G_i, preventing the subunit from interacting with receptors; as a result, this α subunit retains its bound GDP and is unable to regulate its target proteins. These two toxins are widely used as tools to determine whether a cell's response to a signal is mediated by G_s or by G_i.

Individuals who are genetically deficient in a particular G_s α subunit show decreased responses to certain hormones. As a consequence, they display metabolic abnormalities, have abnormal bone development, and are mentally retarded.

Cyclic-AMP-dependent Protein Kinase (PKA) Mediates Most of the Effects of Cyclic AMP

Although cyclic AMP can directly activate certain types of ion channels in the plasma membrane of some highly specialized cells, in most animal cells it exerts its effects mainly by activating **cyclic-AMP-dependent protein kinase(PKA)**. This enzyme catalyzes the transfer of the terminal phosphate group from ATP to specific serines or threonines of selected target proteins, thereby regulating their activity.

PKA is found in all animal cells and is thought to account for the effects of cyclic AMP in most of these cells. The substrates for PKA differ in different cell types, which explains why the effects of cyclic AMP vary so markedly depending on the cell type.

In the inactive state, PKA consists of a complex of two catalytic subunits and two regulatory subunits. The binding of cyclic AMP to the regulatory subunits alters their conformation, causing them to dissociate from the complex. The released catalytic subunits are thereby activated to phosphorylate specific substrate protein molecules. The regulatory subunits of PKA also are important for localizing the kinase inside the cell: special *PKA anchoring proteins* bind both to the regulatory subunits and to a membrane or a component of the cytoskeleton, thereby tethering the enzyme complex to a particular subcellular compartment. Some of these anchoring proteins also bind other kinases and some phosphatases, creating a signaling complex.

Some responses mediated by cyclic AMP are rapid while others are slow. In skeletal muscle cells, for example, activated PKA phosphorylates enzymes involved in glycogen metabolism, which simultaneously triggers the breakdown of glycogen to glucose and inhibits glycogen synthesis, thereby increasing the amount of glucose available to the muscle cell within seconds (see also Figure 15-30). At the other extreme are responses that take hours to develop fully and involve changes in the transcription of specific genes. In cells that secrete the peptide hormone *somatostatin,*for example, cyclic AMP activates the gene that encodes this hormone. The regulatory region of the somatostatin gene contains a short DNA sequence, called the *cyclic AMP response element (CRE),* that is also found in the regulatory region of many other genes activated by cyclic AMP. A specific gene regulatory protein called **CRE-binding (CREB) protein** recognizes this sequence. When CREB is phosphorylated by PKA on a single serine, it recruits a transcriptional coactivator called *CREB-binding protein (CBP),* which stimulates the transcription of these genes. If this serine is mutated, CREB cannot recruit CBP, and it no longer stimulates gene transcription in response to a rise in cyclic AMP levels.

Protein Phosphatases Make the Effects of PKA and Other Protein Kinases Transitory

Since the effects of cyclic AMP are usually transient, cells must be able to dephosphorylate the proteins that have been phosphorylated by PKA. Indeed, the activity of any protein regulated by phosphorylation depends on the balance at any instant between the activities of the kinases that phosphorylate it and the phosphatases that are constantly dephosphorylating it. In general, the dephosphorylation of phosphorylated serines and

threonines is catalyzed by four types of **serine/threonine phosphoprotein phosphatases**—protein phosphatases I, IIA, IIB, and IIC. Except forprotein phosphatase-IIC (which is a minor phosphatase, unrelated to the others), all of these phosphatases are composed of a homologous catalytic subunit complexed with one or more of a large set of regulatory subunits; the regulatory subunits help to control the phosphatase activity and enable the enzyme to select specific targets. *Protein phosphatase I*is responsible for dephosphorylating many of the proteins phosphorylated by PKA. It inactivates CREB, for example, by removing its activating phosphate, thereby turning off the transcriptional response caused by a rise in cyclic AMP concentration. *Protein phosphatase IIA* has a broad specificity and seems to be the main phosphatase responsible for reversing many of the phosphorylations catalyzed by serine/threonine kinases. *Protein phosphatase IIB,* also called*calcineurin,* is activated by Ca^{2+} and is especially abundant in the brain.

Having discussed how trimeric G proteins link activated receptors to adenylyl cyclase, we now consider how they couple activated receptors to another crucial enzyme, *phospholipase C.* The activation of this enzyme leads to an increase in the concentration of Ca^{2+} in the cytosol, which helps to relay the signal onward. Ca^{2+} is even more widely used as an intracellular mediator than is cyclic AMP.

Some G Proteins Activate the Inositol Phospholipid Signaling Pathway by Activating Phospholipase C-β

Many G-protein-linked receptors exert their effects mainly via G proteins that activate the plasma-membrane-boundenzyme **phospholipase C-β**. The phospholipase acts on an inositol phospholipid (a *phosphoinositide*) called **phosphatidylinositol 4,5-bisphosphate [PI(4,5)P_2]**, which is present in small amounts in the inner half of the plasma membrane lipid bilayer. Receptors that operate through this **inositol phospholipid signaling pathway** mainly activate a G protein called **G_q**, which in turn activates phospholipase C-β, in much the same way that G_s activates adenylyl cyclase. The activated phospholipase cleaves PI(4,5)P_2 to generate two products: *inositol 1,4,5-trisphosphate and diacylglycerol*. At this step, the signaling pathway splits into two branches.

Inositol 1,4,5-trisphosphate (IP_3) is a small, water-soluble molecule that leaves the plasma membrane and diffuses rapidly through the cytosol. When it reaches the endoplasmic reticulum (ER), it binds to and opens *IP_3*-gated *Ca^{2+}*-release channels in the ER membrane. Ca^{2+} stored in the ER is released through the open channels, quickly raising the concentration of Ca^{2+} in the cytosol. We discuss later how Ca^{2+} acts to propagate the signal. Several mechanisms operate to terminate the initial Ca^{2+} response: (1) IP_3 is rapidly dephosphorylated by specific phosphatases to form IP_2; (2) IP_3 is phosphorylated to IP_4 (which may function as another intracellular mediator); and (3) Ca^{2+} that enters the cytosol is rapidly pumped out, mainly to the exterior of the cell.

At the same time that the IP_3 produced by the hydrolysis of PI(4,5)P_2 is increasing the concentration of Ca^{2+} in thecytosol, the other cleavage product of PI(4,5)P_2—

diacylglycerol—is exerting different effects. Diacylglycerol remains embedded in the membrane, where it has two potential signaling roles. First, it can be further cleaved to release arachidonic acid, which can either act as a messenger in its own right or be used in the synthesis of other small lipidmessengers called *eicosanoids.* Eicosanoids, such as the *prostaglandins*, are made by most vertebrate cell types and have a wide variety of biological activities. They participate in pain and inflammatory responses, for example, and most anti-inflammatory drugs (such as aspirin, ibuprofen, and cortisone) act—in part, at least—by inhibiting their synthesis.

The second, and more important, function of diacylglycerol is to activate a crucial serine/threonine protein kinase called**protein kinase C (PKC)**, so named because it is Ca^{2+}-dependent. The initial rise in cytosolic Ca^{2+} induced by IP_3alters the PKC so that it translocates from the cytosol to the cytoplasmic face of the plasma membrane. There it is activated by the combination of Ca^{2+}, diacylglycerol, and the negatively charged membrane phospholipidphosphatidylserine. Once activated, PKC phosphorylates target proteins that vary depending on the cell type. The principles are the same as discussed earlier for PKA, although most of the target proteins are different.

Each of the two branches of the inositol phospholipid signaling pathway can be mimicked by the addition of specific pharmacological agents to intact cells. The effects of IP_3 can be mimicked by using a *Ca^{2+} ionophore*, such as A23187or ionomycin, which allows Ca^{2+} to move into the cytosol from the extracellular fluid (discussed in Chapter 11). The effects of diacylglycerol can be mimicked by *phorbol esters*, plant products that bind to PKC and activate it directly. Using these reagents, it has been shown that the two branches of the pathway often collaborate in producing a full cellular response. Some cell types, such as lymphocytes, for example, can be stimulated to proliferate in culture when treated with both a Ca^{2+} ionophore and a PKC activator, but not when they are treated with either reagent alone.

Ca^{2+} Functions as a Ubiquitous Intracellular Messenger

Many extracellular signals induce an increase in cytosolic Ca^{2+} level, not just those that work via G proteins. In eggcells, for example, a sudden rise in cytosolic Ca^{2+} concentration upon fertilization by a sperm triggers a Ca^{2+} wave that is responsible for the onset of embryonic development (Figure 15-37). In muscle cells, Ca^{2+} triggers contraction, and in many secretory cells, including nerve cells, it triggers secretion. Ca^{2+} can be used as a signal in this way because its concentration in the cytosol is normally kept very low (~10^{-7} M), whereas its concentration in the extracellular fluid (~10^{-3} M) and in the ER lumen is high. Thus, there is a large gradient tending to drive Ca^{2+} into the cytosol across both the plasma membrane and the ER membrane. When a signal transiently opens Ca^{2+} channels in either of these membranes, Ca^{2+} rushes into the cytosol, increasing the local Ca^{2+} concentration by 10–20-fold and triggering Ca^{2+}-responsive proteins in the cell.

Three main types of Ca^{2+} channels can mediate this Ca^{2+} signaling:

1.

Voltage-dependent Ca^{2+} channels in the plasma membrane open in response to membrane depolarization and allow, for example, Ca^{2+} to enter activated nerve terminals and trigger neurotransmitter secretion.

2. *IP_3-gated Ca^{2+}-release channels* allow Ca^{2+} to escape from the ER when the inositol phospholipid signaling pathway is activated, as just discussed (see Figure 15-36).

3. *Ryanodine receptors* (so called because they are sensitive to the plant alkaloid ryanodine) react to a change inplasma membrane potential to release Ca^{2+} from the sarcoplasmic reticulum and thereby stimulate the contraction of muscle cells; they are also present in the ER of many nonmuscle cells, including neurons, where they can contribute to Ca^{2+} signaling.

The concentration of Ca^{2+} in the cytosol is kept low in resting cells by several mechanisms (Figure 15-38). Most notably, all eucaryotic cells have a Ca^{2+}-pump in their plasma membrane that uses the energy of ATP hydrolysis to pump Ca^{2+} out of the cytosol. Cells such as muscle and nerve cells, which make extensive use of Ca^{2+} signaling, have an additional Ca^{2+} transport protein (exchanger) in their plasma membrane that couples the efflux of Ca^{2+} to the influx of Na^+. A Ca^{2+} pump in the ER membrane also has an important role in keeping the cytosolic Ca^{2+} concentration low: this Ca^{2+}-pump enables the ER to take up large amounts of Ca^{2+} from the cytosol against a steep concentration gradient, even when Ca^{2+} levels in the cytosol are low. In addition, a low-affinity, high-capacity Ca^{2+} pump in the inner mitochondrial membrane has an important role in returning the Ca^{2+} concentration to normal after a Ca^{2+} signal; it uses the electrochemical gradient generated across this membrane during the electron-transfer steps of oxidative phosphorylation to take up Ca^{2+} from the cytosol.

The Frequency of Ca^{2+} Oscillations Influences a Cell's Response

Ca^{2+}-sensitive fluorescent indicators, such as aequorin or fura-2 (discussed in Chapter 9), are often used to monitor cytosolic Ca^{2+} in individual cells after the inositol phospholipid signaling pathway has been activated. When viewed in this way, the initial Ca^{2+} signal is often seen to be small and localized to one or more discrete regions of the cell. These signals have been called Ca^{2+} blips, quarks, puffs, or sparks, and they are thought to reflect the local opening of individual (or small groups of) Ca^{2+}-release channels in the ER and to represent elementary Ca^{2+} signaling units. If the extracellular signal is sufficiently strong and persistent, this localized signal can propagate as a regenerative Ca^{2+} wave through the cytosol, much like an action potential in an axon. Such a Ca^{2+} "spike" is often followed by a series of further spikes, each usually lasting seconds. These Ca^{2+} oscillations can persist for as long as receptors are activated at the cell surface. Both the

waves and the oscillations are thought to depend, in part at least, on a combination of positive and negative feedback by Ca^{2+} on both the IP_3-gated Ca^{2+}-release channels and the ryanodine receptors: the released Ca^{2+} initially stimulates more Ca^{2+} release, a process known as Ca^{2+}-induced Ca^{2+}release. But then, as its concentration gets high enough, the Ca^{2+} inhibits further release.

The frequency of the Ca^{2+} oscillations reflects the strength of the extracellular stimulus, and it can be translated into a frequency-dependent cell response. In some cases, the frequency-dependent response itself is also oscillatory. In hormone-secreting pituitary cells, for example, stimulation by an extracellular signal induces repeated Ca^{2+} spikes, each of which is associated with a burst of hormone secretion. The frequency-dependent response can also be nonoscillatory. In some types of cells, for instance, one frequency of Ca^{2+} spikes activates the transcription of one set of genes, while a higher frequency activates the transcription of a different set. How do cells sense the frequency of Ca^{2+} spikes and change their response accordingly? The mechanism presumably depends on Ca^{2+}-sensitive proteins that change their activity as a function of Ca^{2+} spike frequency. A protein kinase that acts as a molecular memory device seems to have this remarkable property, as we discuss next.

Ca^{2+}/Calmodulin-dependent Protein Kinases (CaM-Kinases) Mediate Many of the Actions of Ca^{2+} in Animal Cells

Ca^{2+}-binding proteins serve as transducers of the cytosolic Ca^{2+} signal. The first such protein to be discovered was*troponin C* in skeletal muscle cells; its role in muscle contraction is discussed in Chapter 16. A closely related Ca^{2+}-binding protein, known as calmodulin, is found in all eucaryotic cells, where it can constitute as much as 1% of the total protein mass. Calmodulin functions as a multipurpose intracellular Ca^{2+} receptor, mediating many Ca^{2+}-regulated processes. It consists of a highly conserved, single polypeptide chain with four high-affinity Ca^{2+}-binding sites. When activated by binding Ca^{2+}, it undergoes a conformational change. Because two or more Ca^{2+} ions must bind before calmodulin adopts its active conformation, the protein responds in a switchlike manner to increasing concentrations of Ca^{2+}: a tenfold increase in Ca^{2+} concentration, for example, typically causes a fiftyfold increase in calmodulin activation.

The allosteric activation of calmodulin by Ca^{2+} is analogous to the allosteric activation of PKA by cyclic AMP, except that Ca^{2+}/calmodulin has no enzymic activity itself but instead acts by binding to other proteins. In some cases, calmodulin serves as a permanent regulatory subunit of an enzyme complex, but mostly the binding of Ca^{2+} enables calmodulin to bind to various target proteins in the cell to alter their activity.

When an activated molecule of Ca^{2+}/calmodulin binds to its target protein, it undergoes a marked change inconformation (Figure 15-40B). Among the targets regulated by calmodulin binding are many enzymes and membrane transport proteins. As one example, Ca^{2+}/calmodulin binds to and activates the plasma membrane Ca^{2+}-pump that pumps Ca^{2+} out

of cells. Thus, whenever the concentration of Ca^{2+} in the cytosol rises, the pump is activated, which helps to return the cytosolic Ca^{2+} level to normal.

Many effects of Ca^{2+}, however, are more indirect and are mediated by protein phosphorylations catalyzed by a family of **Ca^{2+}/calmodulin-dependent protein kinases (CaM-kinases)**. These kinases, just like PKA and PKC, phosphorylate serines or threonines in proteins, and, as with PKA and PKC, the response of a target cell depends on which CaM-kinase-regulated target proteins are present in the cell. The first CaM-kinases to be discovered—*myosin light-chain kinase,* which activates smooth muscle contraction, and *phosphorylase kinase,* which activates glycogenbreakdown—have narrow substrate specificities. A number of CaM-kinases, however, have much broader specificities, and these seem to be responsible for mediating many of the actions of Ca^{2+} in animal cells. Some phosphorylate generegulatory proteins, such as the CREB protein discussed earlier, and in this way activate or inhibit the transcription of specific genes.

The best-studied example of such a *multifunctional CaM-kinase* is CaM-kinase II, which is found in all animal cells but is especially enriched in the nervous system. It constitutes up to 2% of the total protein mass in some regions of the brain, and it is highly concentrated in synapses. CaM-kinase II has at least two remarkable properties that are related. First, it can function as a molecular memory device, switching to an active state when exposed to Ca^{2+}/calmodulin and then remaining active even after the Ca^{2+} signal has decayed. This is because the kinase phosphorylates itself (a process called *autophosphorylation*) as well as other cell proteins when it is activated by Ca^{2+}/calmodulin. In its autophosphorylated state, the enzyme remains active even in the absence of Ca^{2+}, thereby prolonging the duration of the kinase activity beyond that of the initial activating Ca^{2+} signal. This activity is maintained until phosphatases overwhelm the autophosphorylating activity of the enzyme and shut it off (Figure 15-41). CaM-kinase II activation can thereby serve as a memory trace of a prior Ca^{2+} pulse, and it seems to have an important role in some types of memory and learning in the vertebrate nervous system. A point mutation in CaM-kinase II that removes its autophosphorylation site, but otherwise leaves the kinase activity intact, produces the same learning defect, revealing that the autophosphorylation is critical in these animals.

The second remarkable property of CaM-kinase II is that it can use its memory mechanism to act as a frequency decoder of Ca^{2+} oscillations. This property is thought to be especially important at a nerve cell synapse, where changes in intracellular Ca^{2+} levels in an activated postsynaptic cell can lead to long-term changes in the subsequent effectiveness of that synapse (discussed in Chapter 11). When CaM-kinase II is immobilized on a solid surface and exposed to both aprotein phosphatase and repetitive pulses of Ca^{2+}/calmodulin at different frequencies that mimic those observed in stimulated cells, the enzyme's activity increases steeply as a function of pulse frequency. Moreover, the frequency response of this multisubunit enzyme depends on its exact subunit composition, so that a cell can tailor its response to Ca^{2+} oscillations to particular needs by adjusting the composition of the CaM-kinase II enzyme that it makes.

Some G Proteins Directly Regulate Ion Channels

G proteins do not act exclusively by regulating the activity of membrane-bound enzymes that alter the concentration of cyclic AMP or Ca^{2+} in the cytosol. The α subunit of one type of G protein (called G_{12}), for example, activates a protein that converts a monomeric GTPase of the Rho family (discussed in Chapter 16) into its active form, which then alters the actin cytoskeleton. In some other cases, G proteins directly activate or inactivate ion channels in the plasma membrane of the target cell, thereby altering the ion permeability—and hence the excitability of the membrane. Acetylcholine released by the vagus nerve, for example, reduces both the rate and strength of heart muscle cell contraction (see Figure 15-9A). A special class of acetylcholine receptors that activate the G_i protein discussed earlier mediates this effect. Once activated, the α subunit of G_i inhibits adenylyl cyclase (as described previously), while the βγcomplex binds to K^+ channels in the heart muscle cell plasma membrane to open them. The opening of these K^+channels makes it harder to depolarize the cell, which contributes to the inhibitory effect of acetylcholine on the heart. (These acetylcholine receptors, which can be activated by the fungal alkaloid muscarine, are called *muscarinic acetylcholine receptors* to distinguish them from the very different *nicotinic acetylcholine receptors,* which are ion-channel-linked receptors on skeletal muscle and nerve cells that can be activated by the binding of nicotine, as well as by acetylcholine.)

Other trimeric G proteins regulate the activity of ion channels less directly, either by stimulating channelphosphorylation (by PKA, PKC, or CaM-kinase, for example) or by causing the production or destruction of cyclic nucleotides that directly activate or inactivate ion channels. The *cyclic-nucleotide-gated ion channels* have a crucial role in both smell (olfaction) and vision, as we now discuss.

Smell and Vision Depend on G-Protein-linked Receptors That Regulate Cyclic-Nucleotide-gated Ion Channels

Humans can distinguish more than 10,000 distinct smells, which are detected by specialized olfactory receptor neurons in the lining of the nose. These cells recognize odors by means of specific G-protein-linked **olfactory receptors**, which are displayed on the surface of the modified cilia that extend from each cell (Figure 15-43). The receptors act through cyclic AMP. When stimulated by odorant binding, they activate an olfactory-specific G protein (known as G_{olf}), which in turn activates adenylyl cyclase. The resulting increase in cyclic AMP opens *cyclic-AMP-gated cation channels*, thereby allowing an influx of Na^+, which depolarizes the olfactory receptor neuron and initiates a nerve impulse that travels along its axon to the brain.

There are about 1000 different olfactory receptors in a mouse, each encoded by a different gene and each recognizing a different set of odorants. All of these receptors belong to the G-protein-linked receptor superfamily. Each olfactory receptor neuron produces only one of these 1000 receptors, and the neuron responds to a specific set of odorants by means of

the specific receptor it displays. The same receptor also has a crucial role in directing the elongating axon of each developing olfactory neuron to the specific target neurons that it will connect to in the brain. A different set of more than 100 G-protein-linked receptors acts in a similar way to mediate a mouse's responses to *pheromones,* chemical signals detected in a different part of the nose that are used in communication between members of the same species.

Vertebrate vision involves a similarly elaborate, highly sensitive, signal-detection process. Cyclic-nucleotide-gated ionchannels are also involved, but the crucial cyclic nucleotide is cyclic GMP rather than cyclic AMP. As with cyclic AMP, a continuous rapid synthesis (by *guanylyl cyclase*) and rapid degradation (by *cyclic GMPphosphodiesterase*) controls the concentration of cyclic GMP in cells.

In visual transduction responses, which are the fastest G-protein-mediated responses known in vertebrates, the receptoractivation caused by light leads to a fall rather than a rise in the level of the cyclic nucleotide. The pathway has been especially well studied in **rod photoreceptors (rods)** in the vertebrate retina. Rods are responsible for noncolor vision in dim light, whereas *cone photoreceptors (cones)* are responsible for color vision in bright light. A rod photoreceptor is a highly specialized cell with outer and inner segments, a cell body, and a synaptic region where the rod passes a chemical signal to a retinal nerve cell; this nerve cell in turn relays the signal along the visual pathway. The phototransduction apparatus is in the outer segment, which contains a stack of *discs*, each formed by a closed sac ofmembrane in which many photosensitive rhodopsin molecules are embedded. The plasma membrane surrounding the outer segment contains *cyclic-GMP-gated* Na^+ *channels.* These channels are kept open in the dark by cyclic GMP that has bound to them. Paradoxically, light causes a hyperpolarization (which inhibits synaptic signaling) rather than a depolarization of the plasma membrane (which could stimulate synaptic signaling). Hyperpolarization (an increase in themembrane potential—discussed in Chapter 11) results because the activation by light of rhodopsin molecules in the disc membrane leads to a fall in cyclic GMP concentration and the *closure* of the special Na^+ channels in the surrounding plasma membrane.

Rhodopsin is a seven-pass transmembrane molecule homologous to other members of the G-protein-linked receptorfamily, and, like its cousins, it acts through a trimeric G protein. The activating extracellular signal, however, is not a molecule but a photon of light.
Each rhodopsin molecule contains a covalently attached chromophore, 11-*cis* retinal, which isomerizes almost instantaneously to all-*trans* retinal when it absorbs a single photon. The isomerization alters the shape of the retinal, forcing a conformational change in the protein (opsin). The activated rhodopsin molecule then alters the G-protein *transducin* (G_t), causing its α subunit to dissociate and activate **cyclic GMP phosphodiesterase**. The phosphodiesterase then hydrolyzes cyclic GMP, so that cyclic GMP levels in the cytosol fall.

This drop in cyclic GMP concentration leads to a decrease in the amount of cyclic GMP bound to the plasma membrane Na^+ channels, allowing more of these highly cyclic-GMP-sensitive channels to close. In this way, the signal quickly passes from the disc membrane to the plasma membrane, and a light signal is converted into an electrical one.

A number of mechanisms operate in rods to allow the cells to revert quickly to a resting, dark state in the aftermath of a flash of light—a requirement for perceiving the shortness of the flash. A *rhodopsin-specific kinase (RK)* phosphorylates the cytosolic tail of activated rhodopsin on multiple serines, partially inhibiting the ability of the rhodopsin to activate transducin. An inhibitory protein called *arrestin* then binds to the phosphorylated rhodopsin, further inhibiting rhodopsin's activity. If the gene encoding RK is inactivated by mutation in mice or humans, the light response of rods is greatly prolonged, and the rods eventually die.

At the same time as rhodopsin is being shut off, an RGS protein binds to activated transducin, stimulating the transducin to hydrolyze its bound GTP to GDP, which returns transducin to its inactive state. In addition, the Na^+channels that close in response to light are also permeable to Ca^{2+}, so that when they close, the normal influx of Ca^{2+} is inhibited, causing the Ca^{2+} concentration in the cytosol to fall. The decrease in Ca^{2+} concentration stimulates guanylyl cyclase to replenish the cyclic GMP, rapidly returning its level to where it was before the light was switched on. A specific Ca^{2+}-sensitive protein mediates the activation of guanylyl cyclase in response to a fall in Ca^{2+} levels. In contrast to calmodulin, this protein is inactive when Ca^{2+} is bound to it and active when it is Ca^{2+}-free. It therefore stimulates the cyclase when Ca^{2+} levels fall following a light response.

These shut-off mechanisms do more than just return the rod to its resting state after a light flash; they also help to enable the photoreceptor to *adapt*, stepping down the response when it is exposed to light continuously. Adaptation, as we discussed earlier, allows the receptor cell to function as a sensitive detector of *changes* in stimulus intensity over an enormously wide range of baseline levels of stimulation.

Extracellular Signals Are Greatly Amplified by the Use of Small Intracellular Mediators and Enzymatic Cascades

Despite the differences in molecular details, the signaling systems that are triggered by G-protein-linked receptors share certain features and are governed by similar general principles. They depend on relay chains of intracellular signaling proteins and small intracellular mediators. In contrast to the more direct signaling pathways used by nuclear receptors discussed earlier, and by ion-channel-linked receptors discussed in Chapter 11, these relay chains provide numerous opportunities for amplifying the responses to extracellular signals. In the visual transduction cascade just described, for example, a single activated rhodopsin molecule catalyzes the activation of hundreds of molecules of transducin at a rate of about 1000 transducin molecules per second. Each activated transducin molecule activates a molecule of cyclic GMPphosphodiesterase, each of which hydrolyzes about 4000 molecules of cyclic GMP per second. This catalytic cascade lasts for about 1 second and

results in the hydrolysis of more than 10^5 cyclic GMP molecules for a single quantum of light absorbed, and the resulting drop in the concentration of cyclic GMP in turn transiently closes hundreds of Na^{2+}channels in the plasma membrane (Figure 15-47). As a result, a rod cell can respond to a single photon of light, in a way that is highly reproducible in its timing and magnitude.

Likewise, when an extracellular signal molecule binds to a receptor that indirectly activates adenylyl cyclase via G_s, each receptor protein may activate many molecules of G_s protein, each of which can activate a cyclase molecule. Each cyclase molecule, in turn, can catalyze the conversion of a large number of ATP molecules to cyclic AMP molecules. A similar amplification operates in the inositol-phospholipid pathway. A nanomolar (10^{-9} M) change in the concentration of an extracellular signal can thereby induce micromolar (10^{-6} M) changes in the concentration of a small intracellular mediator such as cyclic AMP or Ca^{2+}. Because these mediators function as allosteric effectors to activate specific enzymes or ion channels, a single extracellular signal molecule can cause many thousands of protein molecules to be altered within the target cell.

Any such amplifying cascade of stimulatory signals requires that there be counterbalancing mechanisms at every step of the cascade to restore the system to its resting state when stimulation ceases. Cells therefore have efficient mechanisms for rapidly degrading (and resynthesizing) cyclic nucleotides and for buffering and removing cytosolic Ca^{2+}, as well as for inactivating the responding enzymes and ion channels once they have been activated. This is not only essential for turning a response off, it is also important for defining the resting state from which a response begins. As we saw earlier, in general, the response to stimulation can be rapid only if the inactivating mechanisms are also rapid. Each protein in the relay chain of signals can be a separate target for regulation, including the receptor, as we discuss next.

G-Protein-linked Receptor Desensitization Depends on Receptor Phosphorylation

As discussed earlier, target cells use a variety of mechanisms to *desensitize*, or *adapt*, when they are exposed to a high concentration of stimulating ligand for a prolonged period (see Figure 15-25). We discuss here only those mechanisms that involve an alteration in G-protein-linked receptors themselves.

These receptors can desensitize in three general ways:

1. They can become altered so that they can no longer interact with G proteins *(receptor inactivation).*
2.

They can be temporarily moved to the interior of the cell (internalized) so that they no longer have access to theirligand *(receptor sequestration).*

3.

They can be destroyed in lysosomes after internalization *(receptor down-regulation).*

In each case, the desensitization process depends on phosphorylation of the receptor, by PKA, PKC, or a member of the family of **G-protein-linked receptor kinases (GRKs)**. (The GRKs include the rhodopsin-specific kinase involved in rod photoreceptor desensitization discussed earlier.) The GRKs phosphorylate multiple serine and threonines on a receptor, but they do so only after the receptor has been activated by ligand binding. As with rhodopsin, once a receptor has been phosphorylated in this way, it binds with high affinity to a member of the **arrestin** family of proteins.

The bound arrestin can contribute to the desensitization process in at least two ways. First, it inactivates the receptor by preventing it from interacting with G proteins, an example of receptor uncoupling. Second, it can serve as an adaptor protein to couple the receptor to clathrin-coated pits (discussed in Chapter 13), inducing receptor-mediated endocytosis. Endocytosis results in either the sequestration or degradation (down-regulation) of the receptor, depending on the specific receptor and cell type, the concentration of the stimulating ligand, and the duration of the ligand's presence.

Summary

G-protein-linked receptors can indirectly activate or inactivate either plasma-membrane-bound enzymes or ion channels via G proteins. When stimulated by an activated receptor, a G protein disassembles into an α subunit and a βγ complex, both of which can directly regulate the activity of target proteins in the plasma membrane. Some G-protein-linked receptors either activate or inactivate adenylyl cyclase, thereby altering the intracellular concentration of the intracellular mediator cyclic AMP. Others activate a phosphoinositide-specific phospholipase C (phospholipase C-β), which hydrolyzes phosphatidylinositol 4,5-bisphosphate [$PI(4,5)P_2$] to generate two small intracellular mediators. One is inositol 1,4,5-trisphosphate (IP_3), which releases Ca^{2+} from the ER and thereby increases the concentration of Ca^{2+} in the cytosol. The other is diacylglycerol, which remains in the plasma membrane and activates protein kinase C (PKC). A rise in cyclic AMP or Ca^{2+} levels affects cells mainly by stimulating protein kinase A (PKA) and Ca^{2+}/calmodulin-dependent protein kinases (CaM-kinases), respectively.

PKC, PKA, and CaM-kinases phosphorylate specific target proteins on serines or threonines and thereby alter the activity of the proteins. Each type of cell has characteristic sets of target proteins that are regulated in these ways, enabling the cell to make its own distinctive response to the small intracellular mediators. The intracellular signaling cascades activated by G-protein-linked receptors allow the responses to be greatly amplified, so that many target proteins are changed for each molecule of extracellular signaling ligand bound to its receptor.

The responses mediated by G-protein-linked receptors are rapidly turned off when the extracellular signaling ligand is removed. Thus, the G-protein α subunit is induced to inactivate itself by hydrolyzing its bound GTP to GDP, $IP_3$is rapidly dephosphorylated by a phosphatase (or phosphorylated by a kinase), cyclic nucleotides are hydrolyzed by phosphodiesterases, $Ca^{2+}$is rapidly pumped out of the cytosol, and phosphorylated proteins are dephosphorylated by protein phosphatases. Activated G-protein-linked receptors themselves are phosphorylated by GRKs, thereby trigging arrestin binding, which uncouples the receptors from G proteins and promotes receptor endocytosis.

Signaling through Enzyme-Linked Cell-Surface Receptors

Enzyme-linked receptors are a second major type of cell-surface receptor. They were recognized initially through their role in responses to extracellular signal proteins that promote the growth, proliferation, differentiation, or survival of cells in animal tissues. These signal proteins are often collectively called *growth factors*, and they usually act as local mediators at very low concentrations (about 10^{-9}-10^{-11} M). The responses to them are typically slow (on the order of hours) and usually require many intracellular signaling steps that eventually lead to changes in gene expression. Enzyme-linked receptors have since been found also to mediate direct, rapid effects on the cytoskeleton, controlling the way a cell moves and changes its shape. The extracellular signals that induce these rapid responses are often not diffusible but are instead attached to surfaces over which the cell is crawling. Disorders of cell proliferation, differentiation, survival, and migration are fundamental events that can give rise to cancer, and abnormalities of signaling through enzyme-linked receptors have major roles in this class of disease.

Like G-protein-linked receptors, enzyme-linked receptors are transmembrane proteins with their ligand-binding domainon the outer surface of the plasma membrane. Instead of having a cytosolic domain that associates with a trimeric G protein, however, their cytosolic domain either has an intrinsic enzyme activity or associates directly with an enzyme. Whereas a G-protein-linked receptor has seven transmembrane segments,

each subunit of an enzyme-linked receptorusually has only one.

Six classes of enzyme-linked receptors have thus far been identified:

1. *Receptor tyrosine kinases* phosphorylate specific tyrosines on a small set of intracellular signaling proteins.
2. *Tyrosine-kinase-associated receptors* associate with intracellular proteins that have tyrosine kinase activity.
3. *Receptorlike tyrosine phosphatases* remove phosphate groups from tyrosines of specific intracellular signaling proteins. (They are called "receptorlike" because the presumptive ligands have not yet been identified, and so theirreceptor function has not been directly demonstrated.)
4. *Receptor serine/threonine kinases* phosphorylate specific serines or threonines on associated latent gene regulatory proteins.
5. *Receptor guanylylcyclases* directly catalyze the production of cyclic GMP in the cytosol.
6. *Histidine-kinase-associated receptors* activate a "two-component" signaling pathway in which the

kinase phosphorylates itself on histidine and then immediately transfers the phosphate to a second intracellular signaling protein.

We begin our discussion with the receptor tyrosine kinases, the most numerous of the enzyme-linked receptors. We then consider the other classes in turn.

Activated Receptor Tyrosine Kinases Phosphorylate Themselves

The extracellular signal proteins that act through **receptor tyrosine kinases** consist of a large variety of secreted growth factors and hormones. Notable examples discussed elsewhere in this book include *epidermal growth factor (EGF),platelet-derived growth factor (PDGF), fibroblast growth factors (FGFs), hepatocyte growth factor (HGF), insulin, insulinlike growth factor-1 (IGF-1), vascular endothelial growth factor (VEGF), macrophage-colony-stimulating factor(M-CSF),* and all the *neurotrophins*, including *nerve growth factor (NGF).*

Many cell-surface-bound signal proteins also act through these receptors. The largest class of these membrane-bound ligands is the **ephrins**, which regulate the cell adhesion and repulsion responses that guide the migration of cells and axons along specific pathways during animal development. The receptors for ephrins, called**Eph receptors**, are also the most numerous receptor tyrosine kinases. The ephrins and Eph receptors are unusual in that they can simultaneously act as both ligand and receptor: on binding to an Eph receptor, some ephrins not only activate the Eph receptor but also become activated themselves to

transmit signals into the interior of the ephrin-expressing cell. In this way, an interaction between an ephrin protein on one cell and an Eph protein on another cell can lead to bidirectional reciprocal signaling that changes the behavior of both cells. Such *bidirectional signaling* between ephrins and Eph receptors is required, for example, to keep cells in particular parts of the developing brain from mixing with cells in neighboring parts.

Receptor tyrosine kinases can be classified into more than 16 structural subfamilies, each dedicated to its complementaryfamily of protein ligands. In all cases, the binding of a signal protein to the ligand-bindingdomain on the outside of the cell activates the intracellular tyrosine kinase domain. Once activated, the kinase domain transfers a phosphate group from ATP to selected tyrosine side chains, both on the receptor proteins themselves and on intracellular signaling proteins that subsequently bind to the phosphorylated receptors.

How does the binding of an extracellular ligand activate the kinase domain on the other side of the plasma membrane? For a G-protein-linked receptor, ligand binding is thought to change the relative orientation of several of the transmembrane α helices, thereby shifting the position of the cytoplasmic loops relative to each other. It is difficult to imagine, however, how a conformational change could propagate across the lipid bilayer through a single transmembrane α helix. Instead, for the enzyme-linked receptors, two or more receptor chains come together in the membrane, forming a dimer or higher oligomer. In some cases, ligand binding induces the

oligomerization. In other cases, the oligomerization occurs before ligand binding, and the ligand causes a reorientation of the receptor chains in the membrane. In either case, the rearrangement induced in cytosolic tails of the receptors initiates the intracellular signaling process. For receptor tyrosine kinases, the rearrangement enables the neighboring kinase domains of the receptor chains to cross-phosphorylate each other on multiple tyrosines, a process referred to as *autophosphorylation.*

To activate a receptor tyrosine kinase the ligand usually has to bind simultaneously to two adjacent receptor chains. PDGF, for example, is a dimer, which cross-links two receptors together. Even some monomeric ligands, such as EGF, bind to two receptors simultaneously and cross-link them directly. By contrast, FGFs, which are also monomers, first form multimers by binding to heparansulfate proteoglycans, either on the target cell surface or in the extracellular matrix. In this way, they are able to cross-link adjacent receptors. In contact-dependent signaling, the ligands form clusters in the plasma membrane of the signaling cell and can thereby cross-link the receptors on the target cell; thus, whereas membrane-bound ephrins activate Eph receptors, soluble ephrins will do so only if they are aggregated.

Because of the requirement for receptor oligomerization, it is relatively easy to inactivate a specific receptor tyrosine kinase to determine its importance for a cell response. For this purpose, cells are transfected with DNA encoding amutant form of the receptor that oligomerizes normally but has an inactive kinase domain. When coexpressed at a high level with normal receptors, the mutant receptor acts

in a *dominant*-*negative* way, disabling the normal receptors by forming inactive dimers with them.

Autophosphorylation of the cytosolic tail of receptor tyrosine kinases contributes to the activation process in two ways. First, phosphorylation of tyrosines within the kinase domain increases the kinase activity of the enzyme. Second, phosphorylation of tyrosines outside the kinase domain creates high-affinity docking sites for the binding of a number of intracellular signaling proteins in the target cell. Each type of signaling protein binds to a different phosphorylated site on the activated receptor because it contains a specific phosphotyrosine-binding domain that recognizes surrounding features of the polypeptide chain in addition to the phosphotyrosine. Once bound to the activated kinase, the signaling protein may itself become phosphorylated on tyrosines and thereby activated; alternatively, the binding alone may be sufficient to activate the docked signaling protein. In summary, autophosphorylation serves as a switch to trigger the transient assembly of a large intracellular signaling complex, which then broadcasts signals along multiple routes to many destinations in the cell. Because different receptor tyrosine kinases bind different combinations of these signaling proteins, they activate different responses.

The receptors for insulin and IGF-1 act in a slightly different way. They are tetramers to start with, and ligand binding is thought to induce a rearrangement of the transmembrane receptor chains, so that the two kinase domains come close together. Most of the phosphotyrosine docking sites generated by ligand binding are not on the

receptor itself, but on a specialized docking protein called *insulin receptor substrate-1 (IRS-1).* The activated receptor first autophosphorylates its kinase domains, which then phosphorylate IRS-1 on multiple tyrosines, thereby creating many more docking sites than could be accommodated on the receptor alone. Other docking proteins are used in a similar way by some other receptor tyrosine kinases to enlarge the size of the signaling complex.

Phosphorylated Tyrosines Serve as Docking Sites For Proteins With SH2 Domains

A whole menagerie of intracellular signaling proteins can bind to the phosphotyrosines on activated receptor tyrosine kinases (or on special docking proteins such as IRS-1) to help to relay the signal onward. Some docked proteins are enzymes, such as **phospholipase C-γ (PLC-γ)**, which functions in the same way as phospholipase C-β—activating the inositol phospholipid signaling pathway discussed earlier in connection with G-protein-linked receptors. Through this pathway, receptor tyrosine kinases can increase cytosolic Ca^{2+} levels. Much more often, these receptors depend more on relay chains of protein-protein interactions. For example, another enzyme that docks on these receptors is the cytoplasmic tyrosine kinase Src, which phosphorylates other signaling proteins on tyrosines. Yet another is *phosphatidylinositol 3'-kinase (PI 3-kinase)*, which, as we discuss later, generates specific lipid molecules in the plasma membrane to attract other signaling proteins there.

Although the intracellular signaling proteins that bind to phosphotyrosines on activated receptor tyrosine kinases and docking proteins have varied structures and functions, they usually share highly conserved phosphotyrosine-binding domains. These can be either SH2 domains (for *Src homology region*, because it was first found in the Src protein) or, less commonly, *PTB domains* (for *phosphotyrosine-binding*). By recognizing specific phosphorylated tyrosines, these small domains serve as modules that enable the proteins that contain them to bind to activated receptor tyrosine kinases, as well as to many other intracellular signaling proteins that have been transiently phosphorylated on tyrosines. Many signaling proteins also contain other protein modules that allow them to interact specifically with other proteins as part of the signaling process. These include the *SH3 domain* (again, so named because it was first discovered in Src), which binds to proline-rich motifs in intracellular proteins.

Not all proteins that bind to activated receptor tyrosine kinases via SH2 domains help to relay the signal onward. Some act to decrease the signaling process, providing negative feedback. One example is the *c-Cbl* protein, which can dock on some activated receptors and catalyze their conjugation with ubiquitin. This ubiquitylation promotes the internalization and degradation of the receptors—a process called receptor down-regulation.

Some signaling proteins are composed almost entirely of SH2 and SH3 domains and function as adaptors to couple tyrosine-phosphorylated proteins to other proteins that do not have their own SH2 domains. Such adaptor proteins

help to couple activated receptors to the important downstream signaling protein ***Ras***. As we discuss next, **Ras** acts as a transducer and bifurcation signaling protein, changing the nature of the signal and broadcasting it along multiple downstream pathways, including a major signaling pathway that can help stimulate cells to proliferate or differentiate. Mutations that activate this pathway, and thereby stimulate cell division inappropriately, are a causative factor in many types of cancer.

Ras Is Activated by a Guanine Nucleotide Exchange Factor

The **Ras** proteins belong to the large ***Ras*** *superfamily of monomeric GTPases*. The family also contains two other subfamilies: the *Rho family*, involved in relaying signals from cell-surface receptors to the actin cytoskeleton and elsewhere (discussed in Chapter 16), and the *Rab family*, involved in regulating the traffic of intracellular transport vesicles. Like almost all of these monomeric GTPases, the **Ras** proteins contain a covalently attached lipid group that helps to anchor the protein to a membrane—in this case, to the cytoplasmic face of the plasma membrane where the protein functions. There are multiple **Ras** proteins, and different ones act in different cell types. Because they all seem to work in much the same way, we shall refer to them simply as **Ras**.

Ras helps to broadcast signals from the cell surface to other parts of the cell. It is often required, for example, whenreceptor tyrosine kinases signal to the nucleus to stimulate cell proliferation or differentiation by

altering geneexpression. If **Ras** function is inhibited by the microinjection of neutralizing anti-**Ras** antibodies or a dominant-negativemutant form of **Ras**, the cell proliferation or differentiation responses normally induced by the activated receptor tyrosine kinases do not occur. Conversely, if a hyperactive mutant **Ras** protein is introduced into some cell lines, the effect on cell proliferation or differentiation is sometimes the same as that induced by the binding of ligands to cell-surface receptors. In fact, **Ras** was first discovered as the hyperactive product of a mutant ***ras*** gene that promoted thedevelopment of cancer; we now know that about 30% of human tumors have a hyperactive ***ras*** mutation.

Like other GTP-binding proteins, **Ras** functions as a switch, cycling between two distinct conformational states—active when GTP is bound and inactive when GDP is bound. Two classes of signaling proteins regulate**Ras** activity by influencing its transition between active and inactive states. **Guanine nucleotide exchange factors (GEFs)** promote the exchange of bound nucleotide by stimulating the dissociation of GDP and the subsequent uptake of GTP from the cytosol, thereby activating **Ras**. **GTPase-activating proteins (GAPs)** increase the rate of hydrolysis of bound GTP by **Ras**, thereby inactivating **Ras** .
Hyperactive mutant forms of **Ras** are resistant to GAP-mediated GTPase stimulation and are locked permanently in the GTP-bound active state, which is why they promote the development of cancer.

In principle, receptor tyrosine kinases could activate **Ras** either by activating a GEF or by inhibiting

a GAP. Even though some GAPs bind directly (via their SH2 domains) to activated receptor tyrosine kinases, whereas GEFs bind only indirectly, it is the indirect coupling of the receptor to a GEF that is responsible for driving **Ras**into its active state. In fact, the loss of function of a **Ras**-specific GEF has a similar effect to the loss of function of that**Ras**. The activation of the other **Ras**-like proteins, including those of the Rho family, is also thought to occur through the activation of GEFs.

Genetic studies in flies and worms, and biochemical studies in mammalian cells, indicate that adaptor proteins linkreceptor tyrosine kinases to **Ras**. The **Grb-2 protein** in mammalian cells, for example, binds through its SH2 domain to specific phosphotyrosines on activated receptor tyrosine kinases and through its SH3 domains to proline-rich motifs on aGEF called **Sos**. Some activated receptor tyrosine kinases, however, do not display the specific phosphotyrosines required for Grb-2 docking; these receptors recruit another adaptor protein called *Shc*, which binds both to the activated receptor and to Grb-2, thereby coupling the receptor to Sos by a more indirect route. The assembly of the complex of receptor-Grb-2-Sos (or receptor-Shc-Grb-2-Sos) brings Sos into position to activate neighboring **Ras** molecules by stimulating it to exchange its bound GDP for GTP. The importance of Grb-2 is indicated by the finding that Grb-2-deficient mice die early in embryogenesis. Very similar sets of proteins are thought to operate in all animals to activate **Ras**.

This pathway from receptor tyrosine kinases is not the only means of activating **Ras**. Other **Ras** GEFs are activated independently of Sos. One that is found mainly in the

brain, for example, is activated by Ca^{2+} and diacylglycerol and can couple G-protein-linked receptors to **Ras** activation.

Once activated, **Ras** in turn activates various other signaling proteins to relay the signal downstream along several pathways. One of the signaling pathways **Ras** activates is a serine/threonine phosphorylation cascade that is highly conserved in eucaryotic cells from yeasts to humans. As we discuss next, a crucial component in this cascade is a novel type of protein kinase called *MAP-kinase*.

Ras Activates a Downstream Serine/Threonine Phosphorylation Cascade That Includes a MAP-Kinase

Both the tyrosine phosphorylations and the activation of **Ras** triggered by activated receptor tyrosine kinases are short-lived. Tyrosine-specific protein phosphatases (discussed later) quickly reverse the phosphorylations, and GAPs induce activated **Ras** to inactivate itself by hydrolyzing its bound GTP to GDP. To stimulate cells to proliferate or differentiate, these short-lived signaling events must be converted into longer-lasting ones that can sustain the signal and relay it downstream to the nucleus to alter the pattern of gene expression. Activated **Ras** triggers this conversion by initiating a series of downstream serine/threonine phosphorylations, which are much longer-lived than tyrosine phosphorylations. Many serine/threonine kinases participate in this phosphorylation cascade, but three of them constitute the core moduleof the cascade. The last of the three is

called a **mitogen-activated protein kinase (MAP-kinase)**.

An unusual feature of a MAP-kinase is that its full activation requires the phosphorylation of both a threonine and a tyrosine, which are separated in the protein by a single amino acid. The protein kinase that catalyzes both of these phosphorylations is called a *MAP-kinase-kinase*, which in the mammalian **Ras** signaling pathway is called MEK. The requirement for both a tyrosine and a threonine phosphorylation ensures that the MAP-kinase is kept inactive unless specifically activated by a MAP-kinase-kinase, whose only known substrate is a MAP-kinase. MAP-kinase-kinase is itself activated by phosphorylation catalyzed by the first kinase in the three-component module, *MAP-kinase-kinase-kinase*, which in the mammalian **Ras** signaling pathway is called **Raf**. The Raf kinase is activated by activated **Ras**.

Once activated, the MAP-kinase relays the signal downstream by phosphorylating various proteins in the cell, includinggene regulatory proteins and other protein kinases. It enters the nucleus, for example, and phosphorylates one or more components of a gene regulatory complex. This activates the transcription of a set of *immediate early genes*, so named because they turn on within minutes of the time that cells are stimulated by an extracellular signal, even if protein synthesis is experimentally blocked with drugs. Some of these genes encode other gene regulatory proteins that turn on other genes, a process that requires both protein synthesis and more time. In this way the **Ras**-MAP-kinase signaling pathway conveys signals from the cell surface to the

nucleus and alters the pattern of gene expression in significant ways. Among the genes activated by this pathway are those required for cell proliferation, such as the genes encoding *G_1 cyclins*.

MAP-kinases are usually activated only transiently in response to extracellular signals, and the period of time they remain active can profoundly influence the nature of the response. When EGF activates its receptors on a neural precursor cell line, for example, MAP-kinase activity peaks at 5 minutes and rapidly declines, and the cells later go on to divide. By contrast, when NGF activates its receptors on the same cells, MAP-kinase activity remains high for many hours, and the cells stop proliferating and differentiate into neurons.

MAP-kinases are inactivated by dephosphorylation, and the specific removal of phosphate from either the tyrosine or the threonine is enough to inactivate the enzyme. In some cases, stimulation by an extracellular signal induces theexpression of a dual-specificity phosphatase that removes both phosphates and inactivates the kinase, providing a form of negative feedback. In other cases, stimulation causes the kinase to be switched off more rapidly by phosphatases that are already present.

Three-component MAP-kinase signaling modules operate in all animal cells, as well as in yeasts, with different ones mediating different responses in the same cell. In budding yeast, for example, one such module mediates the mating pheromone response via the βγ complex of a G protein, another the response to starvation, and yet another the response to osmotic shock. Some of these three-component

MAP-kinase modules use one or more of the same kinases and yet manage to activate different effector proteins and hence different responses.

Mammalian cells also use this strategy to prevent cross talk between MAP-kinase signaling pathways. At least 5 parallel MAP-kinase modules can operate in a mammalian cell. These modules are composed of at least 12 MAP-kinases, 7 MAP-kinase-kinases, and 7 MAP-kinase-kinase-kinases. Several of these modules are activated by different kinds of cell stresses, such as UV irradiation, heat shock, osmotic stress, and stimulation by inflammatory cytokines. The three kinases in at least some of these stress-activated modules are held together by binding to a common scaffold protein, just as in yeast. The scaffold strategy provides precision, helps to create a large change in MAP-kinase activity in response to small changes in signal molecule concentration, and avoids cross-talk. However, it reduces the opportunities for amplification and spreading of the signal to different parts of the cell, which require at least some of the components to be diffusible.

When **Ras** is activated by receptor tyrosine kinases, it usually activates more than just the MAP-kinase signaling pathway. It also usually helps activate *PI3-kinase*, which can signal cells to survive and grow.

PI 3-Kinase Produces Inositol Phospholipid Docking Sites in the Plasma Membrane

Extracellular signal proteins stimulate cells to divide, in part by activating the **Ras**-MAP-kinase pathway just discussed. If cells continually divided without growing, however, they would get progressively smaller and would

eventually disappear. Thus, to proliferate, most cells need to be stimulated to enlarge (grow), as well as to divide. In some cases, one signal protein does both; in others one signal protein (a *mitogen*) mainly stimulates cell division, while another (a*growth factor*) mainly stimulates cell growth. One of the major intracellular signaling pathways leading to cell growth involves **phosphatidylinositol 3-kinase (PI 3-kinase)**. This kinase principally phosphorylates inositol phospholipids rather than proteins; it can be activated by receptor tyrosine kinases, as well as by many other types of cell-surface receptors, including some that are G-protein-linked.

Phosphatidylinositol (PI) is unique among membrane lipids because it can undergo reversible phosphorylation at multiple sites to generate a variety of distinct inositol phospholipids. When activated, PI 3-kinase catalyzes the phosphorylation of inositol phospholipids at the 3 position of the inositol ring to generate lipids called *PI(3,4)*P_2 or*PI(3,4,5)*P_3. The PI(3,4)P_2 and PI(3,4,5)P_3 then serve as docking sites for intracellular signaling proteins, bringing these proteins together into signaling complexes, which relay the signal into the cell from the cytosolic face of the plasma membrane.

It is important to distinguish this use of inositol phospholipids from their use we discussed earlier. We considered earlier how PI(4,5)P_2 is cleaved by PLC-β (in the case of G-protein-linked receptors) or PLC-γ (in the case of receptor tyrosine kinases) to generate soluble IP_3 and membrane-bound diacylglycerol. The IP_3 releases Ca^{2+} from the ER, while the diacylglycerol activates PKC.

By contrast, PI(3,4)P_2 and PI(3,4,5)P_3 are not cleaved by PLC. They remain in the plasma membrane until they are dephosphorylated by specific *inositol phospholipidphosphatases* that remove phosphate from the 3 position of the inositol ring. Mutations that inactivate one suchphosphatase (called PTEN), and thereby prolong signaling by PI 3-kinase, promote the development of cancer, and they are found in many human cancers. The mutations result in prolonged cell survival, indicating that signaling through PI 3-kinase normally promotes cell survival, as well as cell growth.

There are various types of PI 3-kinases. The one that is activated by receptor tyrosine kinases consists of a catalytic and regulatory subunit. The regulatory subunit is an adaptor protein that binds to phosphotyrosines on activated receptor tyrosine kinases through its SH2 domains. Another PI 3-kinase has a different regulatory subunit and is activated by the βγ complex of a trimeric G protein when G-protein-linked receptors are activated by their extracellular ligand. The catalytic subunit, which is similar in both cases, also has a binding site for activated **Ras**, which allows **Ras** to directly stimulate PI 3-kinases.

Intracellular signaling proteins bind to the PI(3,4)P_2 and PI(3,4,5)P_3 that are produced by activated PI 3-kinase mainly through their **Pleckstrin homology (PH) domain**, first identified in the platelet protein Pleckstrin. PH domains are found in about 200 human proteins, including Sos (the GEF discussed earlier that activates **Ras**), and some atypical PKCs that do not depend on Ca^{2+} for their activation. The importance of these domains is illustrated

dramatically by certain genetic immunodeficiency diseases in both humans and mice, where the PH domain in a cytoplasmic tyrosine kinase called **BTK** is inactivated by mutation. Normally, when antigen receptors on B lymphocytes (B cells) activate PI 3-kinase, the resulting inositol lipid docking sites recruit both BTK and PLC-γ to the cytoplasmic face of the plasma membrane. There, the two proteins interact: BTK phosphorylates and activates PLC-γ, which then cleaves $PI(4,5)P_2$ to generate IP_3 and diacylglycerol to relay the signal onward. Because the mutant BTK cannot bind to the lipid docking sites produced after receptor activation, the receptors cannot signal the B cells to proliferate or survive, resulting in a severe deficiency in antibody production.

The PI 3-Kinase/Protein Kinase B Signaling Pathway Can Stimulate Cells to Survive and Grow

One way in which PI 3-kinase signals cells to survive is by indirectly activating **protein kinase B (PKB)** (also called**Akt**). This kinase contains a PH domain, which directs it to the plasma membrane when PI 3-kinase is activated there by an extracellular survival signal. After binding to $PI(3,4,5)P_3$ on the cytosolic face of the membrane, the PKB alters itsconformation so that it can now be activated in a process that requires phosphorylation by a phosphatidylinositol-dependent protein kinase called **PDK1**, which is recruited to the membrane in the same way. Once activated, the PKB returns to the cytoplasm and phosphorylates a variety of target proteins. One of these, called *BAD*, is a protein that normally encourages cells to undergo programmed

cell death, or apoptosis (mentioned earlier and discussed in detail in Chapter 17). By phosphorylating BAD, PKB inactivates it, thereby promoting cell survival. PKB also promotes cell survival by inhibiting other cell death activators, in some cases by inhibiting the transcription of the genes that encode them.

The pathways by which PI 3-kinase signals cells to grow (and increase their metabolism generally) are complex and still poorly understood. One way in which growth factors stimulate cell growth is by increasing the rate of protein synthesis through enhancing the efficiency with which ribosomes translate certain mRNAs into protein. A protein kinase called**S6 kinase** is part of one of the signaling pathways from PI 3-kinase to the ribosome. It phosphorylates and thereby activates the S6 subunit of ribosomes, which helps to increase the translation of a subset of mRNAs that encode ribosomal proteins and other components of the translational apparatus. The activation of S6 kinase is itself a complex process that depends on PDK1 and the phosphorylation of many sites on the protein. PDK1 may phosphorylate one of these sites in response to PI 3-kinase activation.

Tyrosine-Kinase-associated Receptors Depend on Cytoplasmic Tyrosine Kinases for Their Activity

Many cell-surface receptors depend on tyrosine phosphorylation for their activity and yet lack an obvious tyrosine kinase domain. These receptors act through **cytoplasmic tyrosine kinases**, which are associated with the receptors and phosphorylate various

target proteins, often including the receptors themselves, when the receptors bind their ligand. The receptors thus function in much the same way as receptor tyrosine kinases, except that their kinase domain is encoded by a separate gene and is noncovalently associated with the receptor polypeptide chain. As with receptor tyrosine kinases, these receptors must oligomerize to function.

Many of these receptors depend on members of the largest family of mammalian cytoplasmic tyrosine kinases, the Src family of protein kinases. This family includes the following members: *Src*, *Yes*, *Fgr*, *Fyn*, *Lck*, *Lyn*, *Hck*, and *Blk*. These protein kinases all contain SH2 and SH3 domains and are located on the cytoplasmic side of theplasma membrane, held there partly by their interaction with transmembrane receptor proteins and partly by covalently attached lipid chains. Different family members are associated with different receptors and phosphorylate overlapping but distinct sets of target proteins. Lyn, Fyn, and Lck, for example, are each associated with different sets of receptors in lymphocytes. In each case the kinase is activated when an extracellular ligand binds to the appropriate receptor protein. Src itself, as well as several other family members, can also bind to activated receptor tyrosine kinases; in these cases, the receptor and cytoplasmic kinases mutually stimulate each other's catalytic activity, thereby strengthening and prolonging the signal.

Another type of cytoplasmic tyrosine kinase associates with *integrins*, the main family of receptors that cells use to bind to the extracellular matrix (discussed in Chapter 19). The binding of matrix components to integrins can

activate intracellular signaling pathways that influence the behavior of the cell. When integrins cluster at sites of matrix contact, they help trigger the assembly of cell-matrix junctions called *focal adhesions.* Among the many proteins recruited into these junctions is the cytoplasmic tyrosine kinase called **focal adhesion kinase (FAK)**, which binds to the cytosolic tail of one of the integrin subunits with the assistance of other cytoskeletal protein. The clustered FAK molecules cross-phosphorylate each other, creating phosphotyrosine docking sites where the Src kinase can bind. Src and FAK now phosphorylate each other and other proteins that assemble in the junction, including many of the signaling proteins used by receptor tyrosine kinases. In this way, the two kinases signal to the cell that it has adhered to a suitable substratum, where the cell can now survive, grow, divide, migrate, and so on. Mice deficient in FAK die early in development, and their cells do not migrate normally in a culture dish.

Cytokine receptors are the subfamily of enzyme-linked receptors that we discuss next. They constitute the largest and most diverse class of receptors that rely on cytoplasmic kinases to relay signals into the cell. They include receptors for many kinds of local mediators (collectively called *cytokines*), as well as receptors for some hormones, such as growthhormone (see Figure 15-62) and prolactin. As we discuss next, these receptors are stably associated with a class of cytoplasmic tyrosine kinases called *Jaks*, which activate latent gene regulatory proteins called *STATs*. The STAT proteins are normally inactive, being located at the cell surface; cytokine or

hormone binding causes them to migrate to the nucleus and activate gene transcription.

Cytokine Receptors Activate the Jak-STAT Signaling Pathway, Providing a Fast Track to the Nucleus

Many intracellular signaling pathways lead from cell-surface receptors to the nucleus, where they alter genetranscription. The Jak-STAT signaling pathway, however, provides one of the most direct routes. It was initially discovered in studies on the effects of *interferons*, which are cytokines secreted by cells (especially white blood cells) in response to viral infection. Interferons bind to receptors on noninfectedneighboringcells and induce the cells to produce proteins that increase their resistance to viral infection. When activated, interferon receptors activate a novel class of cytoplasmic tyrosine kinases called **Janus kinases (Jaks)** (after the two-faced Roman god). The Jaks then phosphorylate and activate a set of latent gene regulatory proteins called **STATs** (**s**ignal **t**ransducers and **a**ctivators of**t**ranscription), which move into the nucleus and stimulate the transcription of specific genes. More than 30 cytokines and hormones activate the Jak-STAT pathway by binding to cytokine receptors.

All STATs also have an SH2 domain that enables them to dock onto specific phosphotyrosines on some activatedreceptor tyrosine kinase receptors. These receptors can directly activate the bound STAT, independently of Jaks. In fact, the nematode *C. elegans* uses STATs for signaling but does not make any

Jaks or cytokine receptors, suggesting that STATs evolved before Jaks and cytokine receptors.

Cytokine receptors are composed of two or more polypeptide chains. Some cytokine receptor chains are specific to a particular cytokine receptor, while others are shared among several such receptors. All cytokine receptors, however, are associated with one or more Jaks. There are four known Jaks—Jak1, Jak2, Jak3, and Tyk2—and each is associated with particular cytokine receptors. The receptors for α-interferon, for example, are associated with Jak1 and Tyk2, whereas the receptors for γ-interferon are associated with Jak1 and Jak2. As expected, mice that lack Jak1 do not respond to either of these interferons. The receptor for the hormone *erythropoietin*, which stimulates erythrocyte precursor cells to survive, proliferate, and differentiate, is associated with only Jak2. In Jak2-deficient mice, erythrocyte development fails, and the mice die early in development.

Cytokine binding either induces the receptor chains to oligomerize or reorients the chains in a preformed oligomer. In either case, the binding brings the associated Jaks close enough together for them to cross-phosphorylate each other, thereby increasing the activity of their tyrosine kinase domains. The Jaks then phosphorylate tyrosines on the cytokinereceptors, creating phosphotyrosine docking sites for STATs and other signaling proteins.

There are seven known STATs, each with an SH2 domain that performs two functions. First, it mediates the binding of the STAT protein to a phosphotyrosine docking

site on an activated cytokine receptor (or receptor tyrosine kinase); once bound, the Jaks phosphorylate the STAT on tyrosines, causing it to dissociate from the receptor. Second, the SH2 domain on the released STAT now mediates its binding to a phosphotyrosine on another STAT molecule, forming either a STAT homodimeror heterodimer. The STAT dimer then moves into the nucleus, where, in combination with other gene regulatory proteins, it binds to a specific DNA response element in various genes and stimulates their transcription. In response to the hormone prolactin, for example, which stimulates breast cells to produce milk, activated STAT5 stimulates the transcription of genes that encode milk proteins.

Cytokine receptors activate the appropriate STAT proteins because the SH2 domain of these STATs recognizes only the specific phosphotyrosine docking sites on these receptors. Activated receptors for α-interferon, for example, recruit both STAT1 and STAT2, whereas activated receptors for γ-interferon recruit only STAT1. If the SH2 domain of the α-interferon receptor is replaced with the SH2 domain of the γ-interferon receptor, the activated hybrid receptor recruits both STAT1 and STAT2, just like the α-interferon receptor itself.

The responses mediated by STATs are often regulated by negative feedback. In addition to activating genes that encode proteins mediating the cytokine-induced response, the STAT dimers may also activate genes that encode inhibitory proteins. In some cases, the inhibitor binds to both the activated cytokine receptors and STAT proteins, which blocks further STAT activation and helps to shut off

the response; in other cases, the inhibitor achieves the same result by blocking Jak function.

Such negative feedback mechanisms, however, are not enough on their own to turn off the response. The activated Jaks and STATs also have to be inactivated by dephosphorylation of their phosphotyrosines. As in all signaling pathways that use tyrosine phosphorylation, the dephosphorylation is performed by *protein tyrosine phosphatases*, which are as important in the signaling process as the protein tyrosine kinases that add the phosphates.

Some Protein Tyrosine Phosphatases May Act as Cell-Surface Receptors

As discussed earlier, only a small number of serine/threonine phosphatase catalytic subunits are responsible for removing phosphate groups from phosphorylated serines and threonines on proteins. By contrast, there are about 30**protein tyrosine phosphatases (PTPs)** encoded in the human genome. Like tyrosine kinases, they occur in both cytoplasmic and transmembrane forms, none of which are structurally related to serine/threonine protein phosphatases. Individual protein tyrosine phosphatases display exquisite specificity for their substrates, removing phosphate groups from only selected phosphotyrosines on a subset of tyrosine-phosphorylated proteins. Together, these phosphatases ensure that tyrosine phosphorylations are short-lived and that the level of tyrosine phosphorylation in resting cells is very low. They do not, however, simply continuously reverse the effects of protein tyrosine kinases; they are

regulated to act only at the appropriate time in a signaling response or in the cell-division cycle.

Two cytoplasmic tyrosine phosphatases in vertebrates have SH2 domains and are therefore called **SHP-1** and **SHP-2**. SHP-1 helps to terminate some cytokine responses in blood cells by dephosphorylating activated Jaks:mutant erythropoietin receptors that cannot recruit SHP-1, for example, activate Jak2 for much longer than normal. Moreover, SHP-1-deficient mice have abnormalities in almost all blood cell lineages, emphasizing the importance of SHP-1 in blood cell development. Both SHP-1 and SHP-2 also help terminate responses mediated by some receptortyrosine kinases.

There are a large number of transmembrane protein tyrosine phosphatases, but the functions of most of them are unknown. At least some are thought to function as receptors; as this has not been directly demonstrated, however, they are referred to as **receptorlike tyrosine phosphatases**. They all have a single transmembrane segment and usually possess two tyrosine phosphatase domains on the cytosolic side of the plasma membrane. An important example is the**CD45 protein**, which is found on the surface of all white blood cells and has an essential role in the activation of both T and B lymphocytes by foreign antigens. The ligand that is presumed to bind to the extracellulardomain of the CD45 protein has not been identified. However, the role of CD45 in signal transduction has been studied by using recombinant DNA techniques to construct a hybrid protein with an extracellular EGF-binding domain and

intracellular CD45 tyrosine phosphatase domains. The surprising result is that EGF binding seems to inactivate the phosphatase activity of the hybrid protein rather than activating it.

This finding raises the possibility that some receptor tyrosine kinases and receptor tyrosine phosphatases may collaborate when they bind their respective cell-surface-bound ligands—with the kinases adding more phosphates and thephosphatase removing fewer—to maximally stimulate the tyrosine phosphorylation of selected intracellular signaling proteins. The significance of ligand-induced inhibition of CD45 phosphatase is still uncertain, however, and it seems unlikely to be the whole story; CD45 requires its phosphatase activity to function in lymphocyte activation.

Some receptorlike tyrosine phosphatases display features of cell-adhesion proteins and can even mediate homophilic cell-cell binding in cell adhesion assays. In the developing nervous system, for example, they may have an important role in guiding the growing tips of developing nerve cell axons to their targets. In *Drosophila*, the genes encoding several receptorlike tyrosine phosphatases are expressed exclusively in the nervous system, and when some of them are inactivated by mutation, the axons of certain developing neurons fail to find their way to their normal targets. In some cases at least, the phosphatase activity of the protein is required to counteract the action of a cytoplasmic tyrosine kinase for normal axon guidance.

Transmembrane tyrosine phosphatases can also serve as signaling ligands that activate receptors on a neighboring cell. An example is the *protein tyrosine phosphatase ζ/β*, which is expressed on the surface of certain glial cells in the mammalian brain. It binds to a receptor protein (called *contactin*) on developing nerve cells, stimulating the cells to extend long processes. It is possible that the phosphatase also conveys a signal to the glial cell in this interaction, but such bidirectional signaling has not been directly demonstrated for transmembrane tyrosine phosphatases.

Having discussed the crucial role of tyrosine phosphorylation and dephosphorylation in the intracellular signaling pathways activated by many enzyme-linked receptors, we now turn to a class of enzyme-linked receptors that rely entirely on serine/threonine phosphorylation. These transmembrane serine/ threonine kinases activate an even more direct signaling pathway to the nucleus than does the Jak-STAT pathway discussed earlier. They directly phosphorylate latent gene regulatory proteins called *Smads*, which then migrate into the nucleus to activate gene transcription.

Signal Proteins of the TGF-β Superfamily Act Through Receptor Serine/Threonine Kinases and Smads

The **transforming growth factor-β (TGF-β) superfamily** consists of a large number of structurally related, secreted, dimeric proteins. They act either as hormones or, more commonly, as local mediators to regulate a wide range of biological functions in all

animals. During development, they regulate pattern formation and influence various cell behaviors, including proliferation, differentiation, extracellular matrix production, and cell death. In adults, they are involved in tissue repair and in immune regulation, as well as in many other processes. The superfamily includes the*TGF-βs* themselves, the *activins*, and the *bone morphogenetic proteins (BMPs)*. The BMPs constitute the largest family.

All of these proteins act through enzyme-linked receptors that are single-pass transmembrane proteins with a serine/threonine kinase domain on the cytosolic side of the plasma membrane. There are two classes of these **receptorserine/threonine kinases**—*type I* and *type II*—which are structurally similar. Each member of the TGF-β superfamilybinds to a characteristic combination of type-I and type-II receptors, both of which are required for signaling. Typically, the ligand first binds to and activates a type-II receptor homodimer, which recruits, phosphorylates, and activates a type-I receptor homodimer, forming an active tetrameric receptor complex.

Once activated, the receptor complex uses a strategy for rapidly relaying the signal to the nucleus that is very similar to the Jak-STAT strategy used by cytokine receptors. The route to the nucleus, however, is even more direct. The type-I receptor directly binds and phosphorylates a latent gene regulatory protein of the **Smad family** (named after the first two identified, Sma in *C. elegans* and Mad in *Drosophila*). Activated TGF-β receptors and activin receptors phosphorylate Smad2 or

Smad3, while activated BMP receptors phosphorylate Smad1, Smad5, or Smad8. Once one of these Smads has been phosphorylated, it dissociates from the receptor and binds to Smad4, which can form a complex with any of the above five *receptor-activated Smads*. The Smad complex then moves into the nucleus, where it associates with other gene regulatory proteins, binds to specific sites in DNA, and activates a particular set of target genes.

Some TGF-β family members serve as graded morphogens during development, inducing different responses in a developing cell depending on their concentration (discussed in Chapter 21). The different responses can be reproduced by experimentally altering the amount of active Smad complexes in the nucleus, suggesting that the level of these complexes may provide a direct readout of the level of receptor activation. If the DNA-binding sites in different target genes have different affinities for the complexes, then the particular genes activated would reflect the cell's position in the concentration gradient of the morphogen.

As with the Jak-STAT pathway, the Smad pathway is also often regulated by feedback inhibition. Among the target genes activated by Smad complexes are those that encode *inhibitory Smads*, including Smad6 and Smad7. These Smads act as decoys. They bind to activated type-I receptors and prevent other Smads from binding there. This blocks the formation of active Smad complexes and shuts off the response to the TGF-β family ligand. Other types of extracellular ligands can also stimulate the production of inhibitory Smads to antagonize signaling by a TGF-β ligand; γ-interferon, for example, activates the

Jak-STAT pathway, and the resulting activated STAT dimers induce the production of Smad7, which inhibits signaling by TGF-β.

In addition to these intracellular inhibitors, a number of secreted extracellular inhibitory proteins can also neutralize signaling mediated by TGF-β family members. They directly bind to the signal molecules and prevent them from activating their receptors on target cells. *Noggin* and *chordin*, for example, inhibit BMPs, and *follistatin* inhibits activins. Noggin and chordin help to induce the development of the vertebrate nervous system by preventing BMPs from inhibiting this development (discussed in Chapter 21). The TGF-β family members, as well as some of their inhibitors, are usually secreted as inactive precursors that are subsequently activated by proteolytic cleavage.

We turn now to enzyme-linked receptors that are neither kinases nor associated with kinases. We saw earlier that nitric oxide is widely used as a signaling molecule, diffusing through the plasma membrane of a target cell and stimulating a cytoplasmic guanylylcyclase to produce the intracellular mediator cyclic GMP. The receptors we now consider are transmembrane proteins with guanylylcyclase activity.

Receptor GuanylylCyclases Generate Cyclic GMP Directly

Receptor guanylylcyclases are single-pass transmembrane proteins with an extracellular binding site for a signal molecule and an intracellular guanylylcyclase catalytic domain. The binding of the

signal molecule activates the cyclase domain to produce cyclic GMP, which in turn binds to and activates a *cyclic GMP-dependent protein kinase (PKG),*which phosphorylates specific proteins on serine or threonine. Thus, receptor guanylylcyclases use cyclic GMP as an intracellular mediator in the same way that some G-protein-linked receptors use cyclic AMP, except that the linkagebetween ligand binding and cyclase activity is a direct one.

Among the signal molecules that use receptor guanylylcyclase receptors are the *natriuretic peptides (NPs),* a family of structurally related secreted signal peptides that regulate salt and water balance and dilate blood vessels. There are several types of NPs, including *atrial natriuretic peptide (ANP)* and *brain natriuretic peptide (BNP).* Muscle cells in the atrium of the heart secrete ANP when blood pressure rises. The ANP stimulates the kidneys to secrete Na^+ and water and induces the smooth muscle cells in blood vessels walls to relax. Both of these effects tend to lower the blood pressure. When gene targeting is used to inactivate the ANP receptor guanylylcyclase in mice, the mice have chronically elevated blood pressure, resulting in progressive heart enlargement.

An increasing number of receptor guanylylcyclases are being discovered, but in most cases they are orphan receptors, where the ligand that normally activates them is unknown. The genome of the nematode *C. elegans,* for example, encodes 26 of these receptors. Most of those that have been studied are expressed in specific subsets of sensory neurons, suggesting that they may be involved in

detecting particular molecules in the worm's environment. Some of the orphan receptors in mammals are found in sensory neurons in the part of the nose involved in detecting pheromones.

All the signaling pathways activated by G-protein-linked and enzyme-linked receptors we have discussed so far depend on serine/threonine-specific protein kinases, tyrosine-specific protein kinases, or both. Some enzyme-linked receptors, however, depend on an entirely unrelated type ofprotein kinase, as we now discuss.

Bacterial ChemotaxisDepends on a Two-Component Signaling Pathway Activated by Histidine-Kinase-associated Receptors

As pointed out earlier, many of the mechanisms involved in chemical signaling between cells in multicellular animals are thought to have evolved from mechanisms used by unicellular organisms to respond to chemical changes in their environment. In fact, some of the same intracellular mediators, such as cyclic nucleotides and Ca^{2+}, are used by both types of organisms. Among the best-studied reactions of unicellular organisms to extracellular signals are their chemotactic responses, in which cell movement is oriented toward or away from a source of some chemical in the environment. We conclude this section on enzyme-linked receptors with a brief account of bacterial chemotaxis, which depends on a **two-component signaling pathway**, involving histidine-kinase-associated receptors. The same type of signaling pathway is used by yeasts and plants, although apparently not by animals.

Motile bacteria will swim toward higher concentrations of nutrients *(attractants)*, such as sugars, amino acids, and small peptides, and away from higher concentrations of various noxious chemicals *(repellents)*. They swim by means of flagella, each of which is attached by a short, flexible hook at its base to a small protein disc embedded in the bacterialmembrane. This disc is part of a tiny motor that uses the energy stored in the transmembrane H^+ gradient to rotate rapidly and turn the helical flagellum. Because the flagella on the bacterial surface have an intrinsic "handedness," different directions of rotation have different effects on movement. Counterclockwise rotation allows all the flagella to draw together into a coherent bundle, so that the bacterium swims uniformly in one direction. Clockwise rotation causes them to fly apart, so that the bacterium tumbles chaotically without moving forward. In the absence of any environmental stimulus, the direction of rotation of the disc reverses every few seconds, producing a characteristic pattern of movement in which smooth swimming in a straight line is interrupted by abrupt, random changes in direction caused by tumbling.

The normal swimming behavior of bacteria is modified by chemotactic attractants or repellents, which bind to specificreceptor proteins and affect the frequency of tumbling by increasing or decreasing the time that elapses between successive changes in direction of flagellar rotation. When bacteria are swimming in a favorable direction (toward a higher concentration of an attractant or away from a higher concentration of a repellent), they tumble less frequently than when they are swimming in an

unfavorable direction (or when no gradient is present). Since the periods of smooth swimming are longer when a bacterium is traveling in a favorable direction, it will gradually progress in that direction—toward an attractant or away from a repellent.

These responses are mediated by histidine-kinase-associated **chemotaxis receptors**, which typically are dimerictransmembrane proteins that bind specific attractants and repellents on the outside of the plasma membrane. The cytoplasmic tails of the receptors are stably associated with an adaptor protein *CheW* and a histidine kinase *CheA*, which help to couple the receptors to the flagellar motor. Repellent binding activates the receptors, whereas attractant binding inactivates them; a single receptor can bind either type of molecule, with opposite consequences. The binding of a repellent to the receptor activates *CheA*, which phosphorylates itself on a histidine and almost immediately transfers the phosphate to an aspartic acid on a messenger protein *CheY*. The phosphorylated *CheY* dissociates from the receptor, diffuses through the cytosol, binds to the flagellar motor, and causes the motor to rotate clockwise, so that the bacterium tumbles. CheY has intrinsic phosphatase activity and dephosphorylates itself in a process that is greatly accelerated by the *CheZ* protein.

The response to an increase in the concentration of an attractant or repellent is only transient, even if the higher level ofligand is maintained, as the bacteria *desensitize*, or *adapt*, to the increased stimulus. Whereas the initial effect on tumbling occurs in less than a second, adaptation takes minutes. The adaptation is a

crucial part of the response, as it enables the bacteria to respond to *changes* in concentration of ligand rather than to steady-state levels. It is mediated by the covalent methylation (catalyzed by a methyl transferase) and demethylation (catalyzed by a methylase) of thechemotaxis receptors, which change their responsiveness to ligand binding when methylated.

All of the genes and proteins involved in this highly adaptive behavior have now been identified. It therefore seems likely that bacterial chemotaxis will be the first signaling system to be completely understood in molecular terms. Even in this relatively simple signaling network, computer-based simulations are required to comprehend how the system works as an integrated network. Cell signaling pathways will provide an especially rich area of investigation for a new generation of computational biologists, as their network properties will not be understandable without powerful computational tools.

There are some cell-surface receptor proteins that do not fit into the three major classes we have discussed thus far—ion-channel-linked, G-protein-linked, and enzyme-linked. In the next section, we consider cell-surface receptors that activate signaling pathways that depend on proteolysis. These pathways have especially important roles in animal development.

Summary

There are five known classes of enzyme-linked receptors: (1) receptor tyrosine kinases, (2) tyrosine-kinase-associated receptors, (3) receptor serine/threonine kinases, (4) transmembraneguanylylcyclases, and (5) histidine-

kinase-associated receptors. In addition, some transmembrane tyrosine phosphatases, which remove phosphate from phosphotyrosine side chains of specific proteins, are thought to function as receptors, although for the most part their ligands are unknown. The first two classes of receptors are by far the most numerous.

Ligand binding to receptor tyrosine kinases induces the receptors to cross-phosphorylate their cytoplasmic domains on multiple tyrosines. The autophosphorylation activates the kinases, as well as producing a set of phosphotyrosines that then serve as docking sites for a set of intracellular signaling proteins, which bind via their SH2 (or PTB) domains. Some of the docked proteins serve as adaptors to couple the receptors to the small GTPase **Ras**, which, in turn, activates a cascade of serine/threonine phosphorylations that converge on a MAP-kinase, which relays the signal to the nucleusby phosphorylating gene regulatory proteins there. **Ras** can also activate another protein that docks on activated receptor tyrosine kinases—PI 3-kinase—which generates specific inositol phospholipids that serve as docking sites in the plasma membrane for signaling proteins with PH domains, including protein kinase B (PKB).

Tyrosine-kinase-associated receptors depend on various cytoplasmic tyrosine kinases for their action. These kinases include members of the Src family, which associate with many kinds of receptors, and the focal adhesion kinase (FAK), which associates with integrins at focal adhesions. The cytoplasmic tyrosine kinases then phosphorylate a variety of signaling proteins to relay the signal onward. The largest family of receptors in this class

is the cytokine receptors family. When stimulated by ligand binding, these receptors activate Jak cytoplasmic tyrosine kinases, which phosphorylate STATs. The STATs then dimerize, migrate to the nucleus, and activate the transcription of specific genes. Receptor serine/threonine kinases, which are activated by signaling proteins of the TGF-β superfamily, act similarly: they directly phosphorylate and activate Smads, which then oligomerize with another Smad, migrate to the nucleus, and activate gene transcription.

Bacterial chemotaxis is mediated by histidine-kinase-associated chemotaxis receptors. When activated by the binding of a repellent, the receptors stimulate their associated protein kinase to phosphorylate itself on histidine and then transfer that phosphate to a messenger protein, which relays the signal to the flagellar motor to alter the bacterium's swimming behavior. Attractants have the opposite effect on this kinase and therefore on swimming.

Signaling Pathways That Depend on Regulated Proteolysis

The need for intercellular signaling is never greater than during animal development. Each cell in the embryo has to be guided along one developmental pathway or another according to its history, its position, and the character of its neighbors. At each step in the pathway, it must exchange signals with its neighbors to coordinate its behavior with theirs. Most of the signaling pathways already discussed are widely used for these purposes. But there are also others that relay signals in other ways from cell-surface

receptors to the interior of the cell. These additional signaling pathways all depend, in part at least, on regulated proteolysis. Although most of them first came to light through genetic studies in*Drosophila*, they have been highly conserved in evolution and are used over and over again during animal development. As we discuss in Chapter 22, they also have a crucial role in the many developmental processes that continue in adult tissues.

We discuss four of these signaling pathways in this section: the pathway mediated by the receptor protein ***Notch***, the pathway activated by secreted *Wnt* proteins, the pathway activated by secreted *Hedgehog* proteins, and the pathway that depends on activation of the latent gene regulatory protein *NF-κB*. All of these pathways have crucial roles in animaldevelopment. If any one of them is inactivated in a mouse, for example, development is seriously disturbed, and the mouse dies as an embryo or at birth.

The Receptor Protein Notch Is Activated by Cleavage

Signaling through the **Notch** receptor protein may be the most widely used signaling pathway in animal development. It has a general role in controlling cell fate choices during development, mainly by amplifying and consolidating molecular differences between adjacent cells. Although **Notch** signaling is involved in the development of most tissues, it is best known for its role in nerve cell production in *Drosophila*. The nerve cells usually arise as isolated single cells within an epithelial sheet of precursor cells.

During the process, each future nerve cell or committed nerve-cell precursor signals to its immediate neighbours not to develop in the same way at the same time, a process known as *lateral inhibition*. In a fly embryo, for example, the inhibited cells around the future nerve-cell precursors develop into epidermal cells. Lateral inhibition depends on a contact-dependent signaling mechanism that is mediated by a signal protein called **Delta**, displayed on the surface of the future neural cell. By binding to **Notch** on a neighboring cell, Delta signals to the neighbor not to become neural. When this signaling process is defective in flies, the neighbors of neural cells also develop as neural cells, producing a huge excess of neurons at the expense of epidermal cells, which is lethal. Signaling between adjacent cells via **Notch** and Delta (or the Deltalikeligand *Serrate*) regulates cell fate choices in a wide variety of tissues and animals, helping to create fine-grained patterns of distinct cell types. The **Notch**-mediated signal can have other effects beside lateral inhibition; in some tissues, for example, it works in the opposite way, causing neighboring cells to behave similarly.

Both **Notch** and Delta are single-pass transmembrane proteins, and both require proteolytic processing to function. Although it is still unclear why Delta has to be cleaved, the cleavage of **Notch** is central to how **Notch** activation altersgene expression in the nucleus. When activated by the binding of Delta on another cell, an intracellular protease cleaves off the cytoplasmic tail of **Notch**, and the released tail moves into the nucleus to activate the transcription of a set of**Notch**-response genes. The **Notch** tail acts by binding to a gene regulatory

protein called *CSL* (so named because it is called CBF1 in mammals, Suppressor of Hairless in flies, and Lag-1 in worms); this converts CSL from a transcriptionalrepressor into a transcriptional activator. The products of the main genes directly activated by **Notch** signaling are themselves gene regulatory proteins, but with an inhibitory action: they block the expression of genes required for neuraldifferentiation (in the nervous system), and of various other genes in other tissues.

The **Notch** receptor undergoes three proteolytic cleavages, but only the last two depend on Delta. As part of its normal biosynthesis, a protease called *furin* acts in the Golgi apparatus to cleave the newly synthesized **Notch** protein in its future extracellular domain.
This cleavage converts **Notch** into a heterodimer, which is then transported to the cell surface as the mature receptor. The binding of Delta to **Notch** induces a second cleavage in the extracellular domain, mediated by a different protease. A final cleavage quickly follows, cutting free the cytoplasmic tail of the activated receptor.

The cleavage of the **Notch** tail occurs very close to the plasma membrane, just within the transmembrane segment. In this respect it resembles the cleavage of another, more sinister transmembrane protein—the *β-amyloid precursor protein (APP)*, which is expressed in neurons and is implicated in Alzheimer's disease. APP is cleaved within its transmembrane segment, releasing one peptide fragment into the extracellular space of the brain and another into the cytosol of the neuron. In Alzheimer's disease, the extracellular fragments accumulate in

excessive amounts and aggregate into filaments that form amyloid plaques, which are believed to injure nerve cells and contribute to their loss. The most frequent genetic cause of early-onset Alzheimer's disease is a mutation in the *presenilin-1 (PS-1)* gene, which encodes an 8-pass transmembrane protein that participates in the cleavage of APP. The mutations in PS-1 cause cleavage of APP into amyloid-plaque-forming fragments at an increased rate. Genetic evidence in *C. elegans*, *Drosophila*, and mice indicates that the PS-1 protein is a required component of the **Notch** signaling pathway, helping to perform the final cleavage that activates **Notch**. Indeed, **Notch** signaling and cleavage are greatly impaired in PS-1-deficient cells.

Remarkably, **Notch** signaling is regulated by glycosylation. The *Fringe family* of glycosyltransferases adds extra sugars to the O-linked oligosaccharide (discussed in Chapter 13) on **Notch**, which alters the specificity of **Notch** for its ligands. This has provided the first example of the modulation of ligand-receptor signaling by differential receptor glycosylation.

Wnt Proteins Bind to Frizzled Receptors and Inhibit the Degradation of β-Catenin

Wnt proteins are secreted signal molecules that act as local mediators to control many aspects of development in all animals that have been studied. They were discovered independently in flies and in mice:
in *Drosophila*, the *wingless (wg)* gene originally came to light because of its role in wing development, while in mice, the *Int-1* gene was found because it promoted the

development of breast tumors when activated by the integration of a virus next to it. The cell-surface receptors for the Wnts belong to the **Frizzled** family of seven-pass transmembrane proteins. They resemble G-protein-linked receptors in structure, and some of them can signal through G proteins and the inositol phospholipidpathway discussed earlier. They mainly signal, however, through G-protein-independent pathways, which require a cytoplasmic signaling protein called *Dishevelled.*

The best characterized of the Dishevelled-dependent pathways acts by regulating the proteolysis of a multifunctionalprotein called **β-catenin** (or Armadillo in flies), which functions both in cell-cell adhesion and as a latent gene regulatory protein. Wnts activate this pathway by binding to both a Frizzled protein and a co-receptor protein. The co-receptor protein is related to the low density lipoprotein (LDL) receptor protein and is therefore called*LDL-receptor-related protein (LRP).* It is uncertain how Frizzled and LRP activate Dishevelled, which relays the signal onward.

In the absence of Wntsignaling, most of a cell's β-catenin is located at cell-cell adherens junctions, where it is associated with *cadherins*, which are transmembrane adhesion proteins. As discussed in Chapter 19, the β-catenin in these junctions helps link the cadherins to the actin cytoskeleton. Any β-catenin not associated with cadherins is rapidly degraded in the cytoplasm. This degradation depends on a large degradation complex, which recruits β-catenin and contains at least three other proteins.

1. A serine/threonine kinase called **glycogen synthase kinase-3β (GSK-3β)** phosphorylates β-catenin, thereby marking the protein for ubiquitylation and rapid degradation in proteasomes.

2. The tumor-suppressor protein **adenomatous polyposis coli (APC)** is so named because the gene encoding it is often mutated in a type of benign tumor (adenoma) of the colon. The tumor projects into the lumen as a polyp, which can eventually become malignant. APC helps promote the degradation of β-catenin by increasing the affinity of the degradation complex for β-catenin, as required for effective phosphorylation of β-catenin by GSK-3β.

3. A scaffold protein called *axin* holds the protein complex together.

The binding of a Wnt protein to Frizzled and LRP leads to the inhibition of β-catenin phosphorylation and degradation. The mechanism is not understood in detail, but it requires Dishevelled and several other signaling proteins that bind to Dishevelled, including the serine/threonine kinase called *casein kinase 1*. As a result, unphosphorylated β-catenin accumulates in the cytoplasm and nucleus.

In the nucleus, the target genes for Wntsignaling are normally kept silent by an

inhibitory complex of gene regulatory proteins, which includes proteins of the *LEF-1/TCF* family bound to the corepressor protein *Groucho*. The increase in undegraded β-catenin caused by Wntsignaling allows β-catenin to enter the nucleus and bind to LEF-1/TCF, displacing Groucho. The β-catenin now functions as a coactivator, inducing the transcription of the Wnt target genes.

Among the genes activated by β-catenin is *c-myc*, which encodes a protein (c-Myc) that is a powerful stimulator of cell growth and proliferation (discussed in Chapter 17). Mutations of the *APC* gene occur in 80% of human colon cancers. These mutations inhibit the protein's ability to bind β-catenin, so that β-catenin accumulates in the nucleus and stimulates the transcription of *c-myc* and other Wnt target genes, even in the absence of Wntsignaling. The resulting uncontrolled cell proliferation promotes the development of cancer.

Hedgehog Proteins Act Through a Receptor Complex of Patched and Smoothened, Which Oppose Each Other

Like Wnt proteins, the **Hedgehog** proteins are a family of secreted signal molecules that act as local mediators in many developmental processes in both invertebrates and vertebrates. Abnormalities in the Hedgehog pathway duringdevelopment can be lethal and in adult cells can also lead to cancer. The Hedgehog proteins were discovered in*Drosophila*, where a mutation in the only gene encoding such a protein produces a larva with spiky processes (denticles) resembling a hedgehog. At least three genes encode Hedgehog proteins in vertebrates—*sonic*,

desert, and*indian hedgehog*. The active form of all Hedgehog proteins is unusual in that it is covalently coupled to cholesterol, which helps to restrict its diffusion following secretion. The cholesterol is added during a remarkable processing step, in which the protein cleaves itself. The proteins are also modified by the addition of a fatty acid chain, which, for unknown reasons, can be required for their signaling activity.

Two transmembrane proteins, Patched and Smoothened, mediate the responses to all Hedgehog proteins. **Patched** is predicted to cross the plasma membrane 12 times, and it is the receptor that binds the Hedgehog protein. In the absence of a Hedgehog signal, Patched inhibits the activity of **Smoothened**, which is a 7-pass transmembrane protein with a structure similar to a Frizzled protein. This inhibition is relieved when a Hedgehog protein binds to Patched, allowing Smoothened to relay the signal into the cell. Most of what we know about the downstream signaling pathway activated by Smoothened comes from genetic studies in flies, and it is the fly pathway that we summarize here.

In some respects the Hedgehog signaling pathway in *Drosophila* operates similarly to the Wnt pathway. In the absence of a Hedgehog signal, a gene regulatory protein called **Cubitusinterruptus (Ci)** is proteolytically cleaved in proteasomes. Instead of being completely degraded, however, it is processed to form a smaller protein that accumulates in the nucleus, where it acts as a transcriptional repressor, helping to keep some Hedgehog-responsive genes silent. The proteolytic processing of the Ci protein depends on a large multiprotein complex. The

complex contains a serine/threonine kinase (called Fused) of unknown function, an anchoring protein (called Costal) that binds the complex to microtubules (keeping Ci out of the nucleus), and an adaptor protein (called Suppressor of Fused) (Figure 15-73A). When Hedgehog binds to Patched to activate the signaling pathway, Ci processing is suppressed, and the unprocessed Ci protein is released from its complex and enters the nucleus, where it activates the transcription of Hedgehog target genes.

Among the genes activated by Ci is the gene that encodes the Wnt protein Wingless, which helps pattern tissues in the fly embryo (discussed in Chapter 21). Another target gene is *patched* itself; the resulting increase in Patched protein on the cell surface inhibits further Hedgehog signaling—a form of negative feedback.

Many gaps in the Hedgehog signaling pathway still remain to be filled in. It is not known, for example, how Patched inhibits Smoothened, how Smoothened activates the pathway, how the proteolysis of Ci is regulated (although it is known that Ci phosphorylation by PKA is required for the processing), or how the release of the complex from microtubules and unprocessed Ci from the complex is controlled.

Even less is known about the Hedgehog pathway in vertebrate cells. In addition to there being at least three types of vertebrate Hedgehog proteins, there are two forms of Patched and three Ci-like proteins (*Gli1*, *Gli2*, and *Gli3*). Unlike in flies, Hedgehog signaling stimulates the transcription of the *Gli* genes, and it is unclear whether all of the Gli proteins undergo proteolytic

processing, although there is evidence that Gli3 does. Inactivating mutations in one of the human*patched* genes, which leads to excessive Hedgehog signaling, occur frequently in the most common form of skin cancer*(basal cell carcinoma)*, suggesting that Patched normally helps to keep skin cell proliferation in check.

Go to:

Multiple Stressful and Proinflammatory Stimuli Act Through an NF-κB-Dependent Signaling Pathway

The **NF-κB proteins** are latent gene regulatory proteins that lie at the heart of most inflammatory responses. These responses occur as a reaction to infection or injury and help protect the animal and its cells from these stresses. When excessive or inappropriate, however, inflammatory responses can also damage tissue and cause severe pain, as happens in joints in rheumatoid arthritis, for example. NF-κB proteins also have an important role in intercellular signaling during normal vertebrate development, although the extracellular signals that activate NF-κB in these circumstances are unknown. In *Drosophila*, however, genetic studies have identified both the extracellular and the intracellular proteins that activate the NF-κB family member *Dorsal*, which has a crucial role in specifying the dorsal-ventral axis of the developing fly embryo. The same intracellular signaling pathway is also involved in defending the fly from infection, just as in vertebrates.

Two vertebrate cytokines are especially important in inducing inflammatory responses—*tumor necrosis factor α (TNF-α)* and *interleukin-1 (IL-1)*. Both are made by cells

of the innate immune system, such as macrophages, in response to infection or tissue injury. These proinflammatory cytokines bind to cell-surface receptors and activate NF-κB, which is normally sequestered in an inactive form in the cytoplasm of almost all of our cells. Once activated, NF-κB turns on the transcription of more than 60 known genes that participate in inflammatory responses. Although TNF-α receptors and IL-1 receptors are structurally unrelated, they operate in much the same way.

There are five NF-κB proteins in mammals (RelA, RelB, c-Rel, NF-κB1, and NF-κB2), and they form a variety of homodimers and heterodimers, each of which activates its own characteristic set of genes. Inhibitory proteins called **IκB**bind tightly to the dimers and hold them in an inactive state within large protein complexes in the cytoplasm. Signals such as TNF-α or IL-1 activate the dimers by triggering a signaling pathway that leads to the phosphorylation, ubiquitylation, and consequent degradation of IκB. The degradation of IκB exposes a nuclear localization signal on the NF-κB proteins, which now move into the nucleus and stimulate the transcription of specific genes. The phosphorylation of IκB is performed by a specific serine/threonine kinase called **IκB kinase (IKK)**.

The mechanism by which the binding of a proinflammatory cytokine to its cell-surface receptors activates IκB kinase. Ligand binding causes the cytosolic tails of the clustered receptors to recruit various adaptor proteins and cytoplasmic serine/threonine kinases. One of the recruited kinases is thought to be an *IκB kinase kinase*

(IKKK) that directly phosphorylates and activates the IκB kinase (IKK).

Not all of the signaling proteins recruited to the cytosolic tail of the TNF-α receptor contribute to NF-κB activation, however. Some can trigger a MAP-kinase cascade, while others can activate a proteolytic cascade that leads toapoptosis.

Thus far, we have discussed cell signaling mainly in animals, with a few diversions into yeasts and bacteria. But intercellular signaling is just as important for plants as it is for animals, although the mechanisms and molecules used are mainly different, as we discuss next.

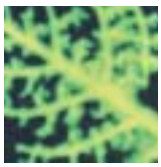
Summary

Some signaling pathways that are especially important in animal development depend on proteolysis for at least part of their action. **Notch** receptors are activated by cleavage when Delta (or a related ligand) on another cell binds to them; the cleaved cytosolic tail of **Notch** migrates into the nucleus, where it stimulates gene transcription. In the Wntsignaling pathway, by contrast, the proteolysis of the latent gene regulatory protein β-catenin is inhibited when secreted Wnt proteins bind to their receptors; as a result, β-catenin accumulates in the nucleus and activates the transcription of Wnt target genes.

Hedgehog signaling in flies works much like Wntsignaling: in the absence of a signal, a bifunctional, cytoplasmic gene regulatory protein Ci is proteolytically cleaved to form a transcriptional repressor that keeps

Hedgehog target genes silenced. The binding of Hedgehog to its receptor inhibits the proteolytic processing of Ci; as a result, the larger form of Ci accumulates in the nucleus and activates the transcription of Hedgehog-responsive genes. Signaling through the latent gene regulatory protein NF-κ B also depends on proteolysis. NF-κ B is normally held in an inactive state by the inhibitory protein Iκ B within a multiprotein complex in the cytoplasm. A variety of extracellular stimuli, including proinflammatory cytokines, trigger a phosphorylation cascade that ultimately phosphorylates Iκ B, marking it for degradation; this enables the freed NF-κ B to enter the nucleus and activate the transcription of its target genes.

13 Recombinant DNA and Biotechnology

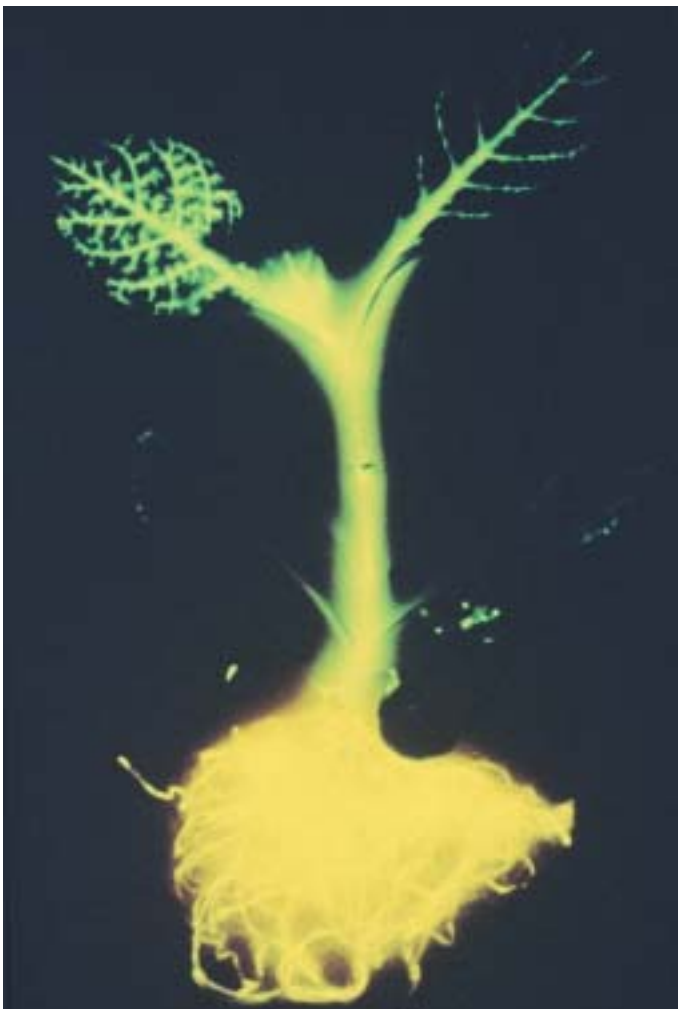


Of the many horrible legacies of the wars of the twentieth century, perhaps none is so lasting as the littering of the countryside with land mines. These inexpensive plastic shells filled with trinitrotoluene (TNT) are built to explode and injure whatever steps on the soil above them. Currently, the most common way of finding land mines is to poke around the soil with a stick—a precarious occupation at best. Since the mines are made of plastic, metal detectors do not work. The agricultural systems of many of the countries with large numbers of land mines, such as Cambodia and Angola, are based on manual labor, so where there are land mines, there cannot be a farm. Clearly, the world needs a sensitive, non-lethal land mine detector.

Enter biotechnology. Neal Stewart, at the University of North Carolina, is developing plants that can detect land mines in a field and show their locations remotely. His method is an excellent example of the application of knowledge of gene transcription and translation as well as of genetic engineering. In Chapter 14, we saw that the control of eukaryotic gene transcription lies at the promoter, a DNA sequence where RNA polymerase and other proteins bind to initiate transcription of the adjacent gene. Certain bacteria have a promoter that is sensitive to TNT, and the binding of a tiny amount of this chemical activates an adjacent gene.

Plants would be ideal biosensors for land mines, as seeds can be spread widely and evenly in a suspect field. But what gene could "announce" the presence of TNT by making a detectable protein? It turns out that certain jellyfish make a protein that fluoresces green when ultraviolet light ("black light") is shone on it. Stewart has introduced this gene into a plant, placing it alongside the TNT-sensitive promoter. When these plants are grown near a land mine, their roots take up TNT, and their leaves will glow in ultraviolet light. Of course, having people plant these seeds or shine a black light on these plants as they walk through the field would bring us back to the old, dangerous way of detecting land mines. To solve these problems, seeding and remote sensing could be done from airplanes or helicopters flying over the field.

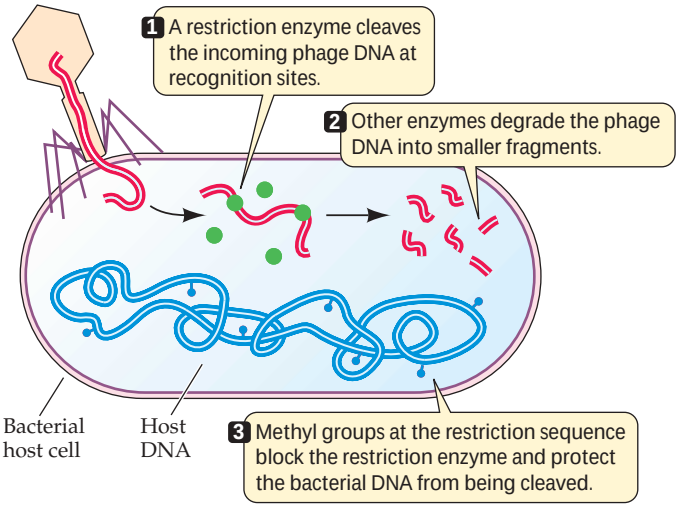
This story—from problem to solution—has been repeated many times in the past two decades. The products of recombinant DNA technology range from life-saving drugs that there is no other way to make in adequate amounts to crop plants with improved agricultural characteristics. Although the basic



A Land Mine Detector Plants can be genetically engineered to express green fluorescent protein from a jellyfish gene. When this "glow-in-the-dark" gene is linked so that it is activated by the presence of TNT in the soil, such plants can act as biosensors to detect the presence of explosive land mines.

techniques of DNA manipulation have been called revolutionary, most of them come from the knowledge of DNA transcription and translation that we described in earlier chapters. The upshot of this technology is that today we can take a piece of DNA from any source, attach it chemically to any other DNA—making a recombined DNA molecule that has never existed before in the entire evolutionary history of life—and then use this unique DNA for whatever purpose and in whichever target organism we wish. This ability gives humans unprecedented power over life in general.

We begin this chapter with a description of how DNA molecules can be cut into smaller fragments and how fragments from different sources can be covalently linked to create recombinant DNA in a test tube. Recombinant (or any other) DNA can then be introduced into a suitable prokaryotic or eukaryotic host cell. Sometimes, the purpose of adding a new gene to a host cell or organism is to ask an experimental question about the role of that gene that can be answered by placing it in a new environment. In other instances, the purpose is to coax the host cell to make a new gene product.



16.1 Bacteria Fight Invading Viruses with Restriction Enzymes
Bacteria produce restriction enzymes that degrade phage DNA by cleaving it into smaller, double-stranded fragments. Other enzymes protect the bacteria's own DNA from being cleaved.

Cleaving and Rejoining DNA

Scientists have long realized that the chemical reactions used in living cells for one purpose may be applied in the laboratory for other, novel purposes. **Recombinant DNA technology**—the manipulation and combination of DNA molecules from different sources—is based on this realization, and on an understanding of the properties of certain enzymes and of DNA itself.

As we have seen in previous chapters, the complementary pairing of nucleotide bases underlies many fundamental processes of molecular biology. The mechanisms of DNA replication, transcription, and translation all rely on complementary base pairing. Similarly, all the key techniques in recombinant DNA technology—locating, sequencing, rejoining, and amplifying DNA fragments—make use of the complementary base pairing of A with T (or U) and of G with C.

In this section, we will see how some of the numerous naturally occurring enzymes that cleave and repair DNA can be used in recombinant DNA technology. Many of these enzymes have been isolated and purified and are now used in the laboratory to manipulate and recombine DNA. Using these enzymes, fragments of DNA can be separated, covalently linked to other fragments, and employed for many novel and highly useful purposes.

Restriction enzymes cleave DNA at specific sequences

All organisms must have ways of dealing with their enemies. As we saw in Chapter 13, bacteria are attacked by viruses called bacteriophages that inject their genetic material into the host cell. Some bacteria defend themselves against such invasions by producing **restriction enzymes** (also known as **restriction endonucleases**), which catalyze the cleavage of double-stranded DNA molecules—such as those injected by phages—into smaller, noninfectious fragments (Figure 16.1). These enzymes cut the bonds between the 3′ hydroxyl of one nucleotide and the 5′ phosphate of the next one.

There are many such restriction enzymes, each of which cleaves DNA at a specific sequence of bases, called a *recognition sequence*, or **restriction site**. The DNA of the host cell is not cleaved by its own restriction enzymes because specific modifying enzymes called *methylases* add methyl (—CH_3) groups to certain bases at the restriction sites of the host's DNA when it is being replicated. The methylation of the host's bases makes the recognition sequence unrecognizable to the restriction enzyme. But unmethylated phage DNA is efficiently recognized and cleaved.

A specific sequence of bases defines each restriction sequence. For example, the enzyme *Eco*RI (named after its source, a strain of the bacterium *E. coli*) cuts DNA only where it encounters the following paired sequence in the DNA double helix:

5′ . . . G A A T T C . . . 3′
3′ . . . C T T A A G . . . 5′

Notice that this sequence reads the same in the 5′-to-3′ direction on both strands. It is *palindromic*, like the word "mom," in the sense that it is the same in both directions from the 5′ end. The *Eco*RI enzyme has two identical active

sites on its two subunits, which cleave the two strands simultaneously between the G and the A of each strand (see Figure 16.4).

The *Eco*RI recognition sequence occurs, on average, about once in every 4,000 base pairs in a typical prokaryotic genome, or about once per four prokaryotic genes. So *Eco*RI can chop a large piece of DNA into smaller pieces containing, on average, just a few genes. Using *Eco*RI in the laboratory to cut small genomes, such as those of viruses that have tens of thousands of base pairs, may result in a few fragments. For a huge eukaryotic chromosome with tens of millions of base pairs, the number of fragments will be very large.

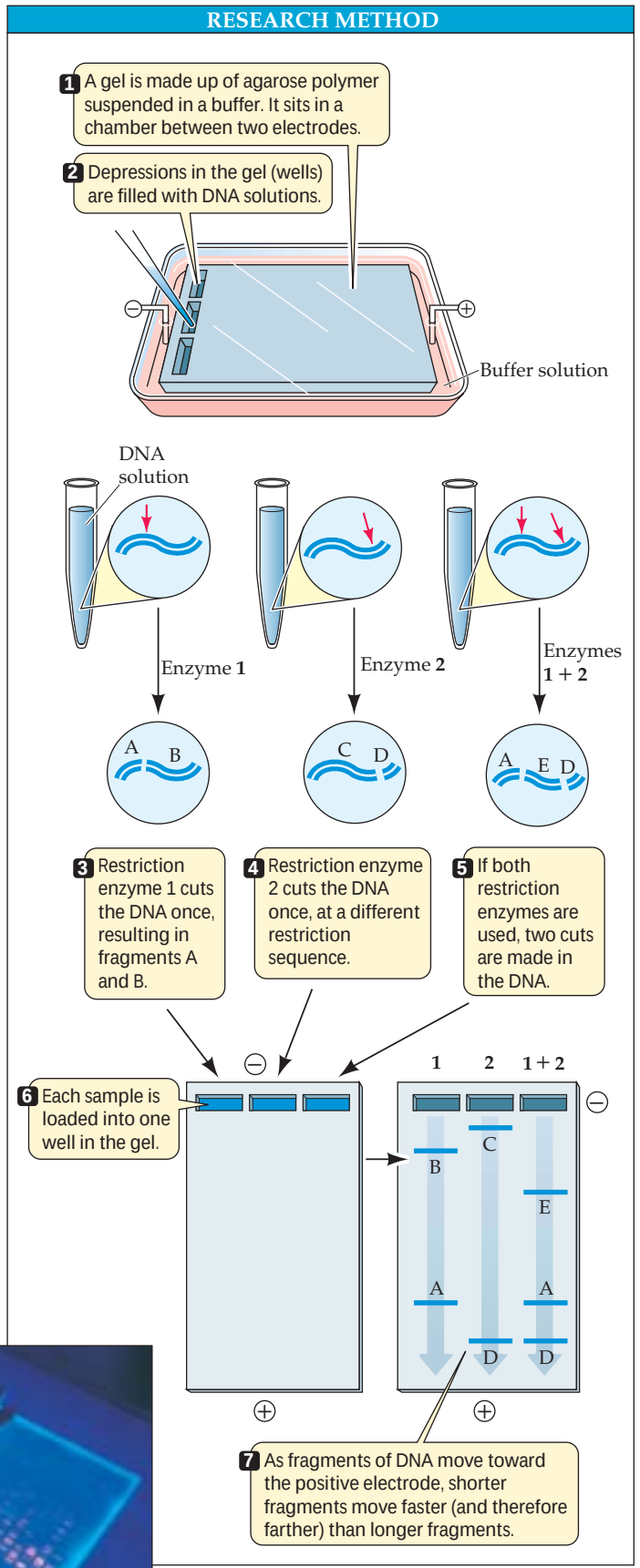
Of course, "on average" does not mean that the enzyme cuts all stretches of DNA at regular intervals. The *Eco*RI recognition sequence does not occur even once in the 40,000 base pairs of the genome of a phage called T7—a fact that is crucial to the survival of this virus, since its host is *E. coli*. Fortunately for *E. coli*, the DNA of other phages does contain the *Eco*RI recognition sequence.

Hundreds of restriction enzymes have been purified from various microorganisms. In the test tube, different restriction enzymes that recognize different restriction sites can be used to cut the same sample of DNA. Thus, restriction enzymes can be used as "knives" for genetic "surgery" to cut a sample of DNA in many different, specific places.

Gel electrophoresis identifies the sizes of DNA fragments

After a laboratory sample of DNA has been cut with a restriction enzyme, the DNA is in fragments, which must be separated. Because the recognition sequence does not occur at regular intervals, the fragments are not all the same size, and this property provides a way to separate them from one another. Separating the fragments is necessary to determine the number and sizes (in base pairs) of fragments produced or to identify and purify an individual fragment of particular interest.

The best way to separate or purify DNA fragments is by **gel electrophoresis** (Figure 16.2). Because of its phosphate groups, DNA is negatively charged at neutral pH. A mixture



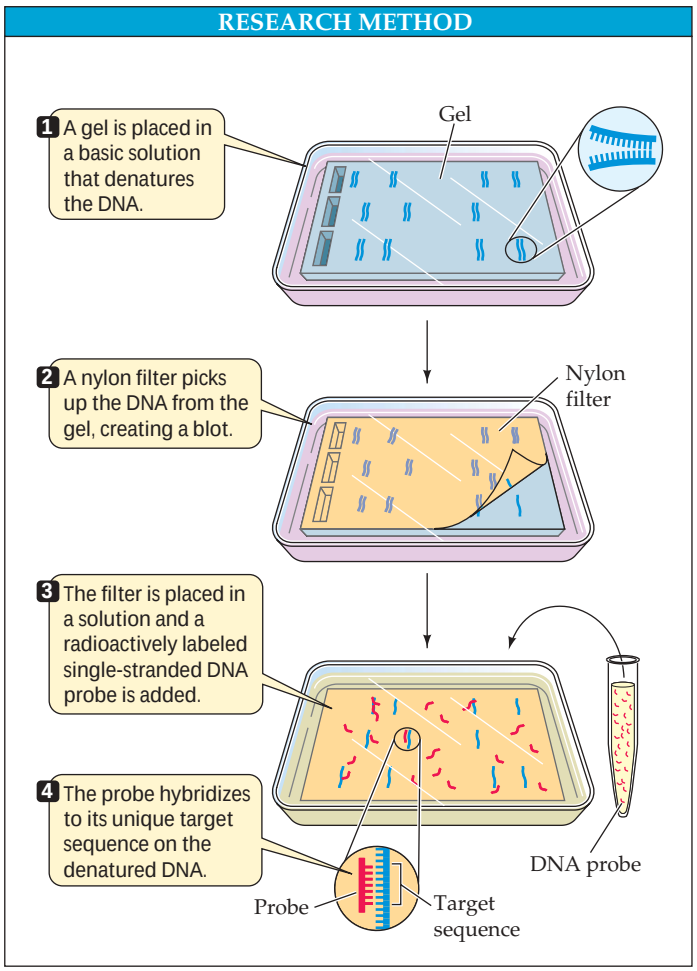
16.2 Separating Fragments of DNA by Gel Electrophoresis A mixture of DNA fragments is placed in a gel and an electric field is applied across the gel. The negatively charged DNA moves toward the positive end of the field, with smaller molecules moving faster than larger ones. When the electric power is shut off, the now separated fragments can be analyzed.

of DNA fragments is placed in a well in a porous gel, and an electric field (with positive and negative ends) is applied across the gel. Because opposite charges attract, the DNA fragments move toward the positive end of the field. Since the porous gel acts as a sieve, the smaller molecules move faster than the larger ones. After a fixed time, and while all the fragments are still in the gel, the electric power is shut off. The separated fragments can be visualized by staining them with a dye that fluoresces under ultraviolet light. They can then be seen as bars or spots in the gel (Figure 16.2) and can be examined or removed individually.

Electrophoresis gives two types of information:

- *The sizes of the fragments*. DNA fragments of known molecular size are often placed in a well in the gel next to the sample to provide a size reference.
- *The presence of specific DNA sequences*. A specific DNA sequence can be located by using a probe (Figure 16.3). The DNA is denatured while still in the gel, then the gel is affixed to a nylon filter to make a "blot." The filter is then exposed to a single-stranded DNA probe with a sequence complementary to the one that is being sought. If the sequence of interest in present, the probe will hybridize with it. The probe can be labeled in some way—for example, with radioactivity. After hybridization, spots of radioactivity on the membrane indicates that the probe has hybridized with its target sequence at that location. Unbound probes stay in solution.

The gel region containing only the desired fragment (in size or sequence) can be cut out as a lump of gel, and the pure DNA fragment can then be removed from the gel by diffusion into a small volume of water.



16.3 Analyzing DNA Fragments A probe can be used to locate a specific DNA fragment on an electrophoresis gel.

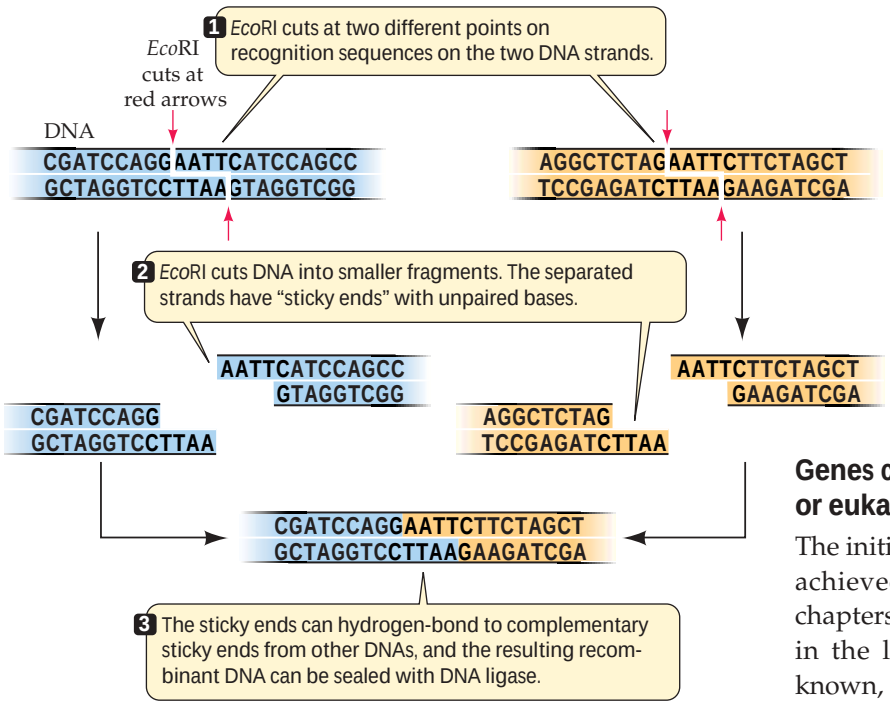
Recombinant DNA can be made in the test tube

Some restriction enzymes cut the DNA backbone cleanly, cutting both strands exactly opposite one another. Others make staggered cuts, cutting one strand of the double helix several bases away from where they cut the other. Fragments cut in this manner are particularly useful in biotechnology.

*Eco*RI, for example, cuts DNA within its recognition sequence in a staggered manner, as shown at the top of Figure 16.4. After the two cuts in the opposing strands are made, the strands are held together only by the hydrogen bonds between four base pairs. The hydrogen bonds of these few base pairs are too weak to persist at warm temperatures (above room temperature), so the two strands of DNA separate, or *denature*. As a result, there are single-stranded "tails" at the location of each cut. These tails are called **sticky ends** because they have a specific base sequence that can bind by base pairing with complementary sticky ends. If n restriction sites for a given restriction enzyme are present in a linear DNA molecule, then $n + 1$ fragments will be made, all with the same complementary sequences at their sticky ends.

After a DNA molecule has been cut with a restriction enzyme, complementary sticky ends can form hydrogen bonds with one another. The original ends may rejoin, or an end may pair with a complementary end from another fragment. Furthermore, because the ends of all fragments cut by the same restriction enzyme are the same, fragments from one source, such as a human, can be joined to fragments from another source, such as a bacterium.

When the temperature is lowered, the fragments *anneal* (come together by hydrogen bonding) at their sticky ends at random, but these associations are unstable because they are held together by only a few hydrogen bonds. The associated sticky ends can be permanently united by a second enzyme, **DNA ligase**, which forms the one covalent bond needed at each sticky end to "seal" the DNA strands. In the cell, this enzyme unites the Okazaki fragments and mends breaks in DNA, as we saw in Chapter 11.



16.4 Cutting and Splicing DNA Some restriction enzymes (*Eco*RI is shown here) make staggered cuts in DNA. *Eco*RI can be used to cut DNA from two different sources (blue and gold). At warm temperatures, the two DNA strands will separate (denature), leaving sticky ends, exposed bases that can hybridize with complementary fragments. When the temperature is lowered, sticky ends from different DNAs can bind to each other, forming recombinant DNA.

Many restriction enzymes do not produce sticky ends. Instead, they cut both DNA strands at the same base pair within the recognition sequence, making "blunt" ends. DNA ligase can also connect blunt-ended fragments, but it does so with reduced efficiency.

With these two enzyme tools—restriction enzymes and DNA ligase—scientists can cut and rejoin different DNA molecules from any and all sources to form recombinant DNA (see Figure 16.4). These simple techniques have revolutionized biological science in the past 30 years and have given us the power to manipulate genetic material at will.

Getting New Genes into Cells

One goal of recombinant DNA technology is to produce many copies (**clones**) of a particular gene, either for purposes of analysis or to produce its protein product in quantity. If the recombinant DNA is to make its protein, it must be inserted, or **transfected**, into a host cell. Such altered hosts are known as **transgenic** cells or organisms. The choice of a host cell—prokaryotic or eukaryotic—is important in this work.

Once the host species is selected, the recombinant DNA is brought together with a population of host cells and, under specific conditions, enters some of them. Because all the host cells proliferate—not just the few that receive the recombinant DNA—the scientist must be able to determine which cells actually contain the sequence of interest. One common method of identifying cells with recombinant DNA is to tag the inserted sequence with **reporter genes**, whose phenotypes are easily observed. These phenotypes serve as *genetic markers* for the sequence of interest.

Genes can be inserted into prokaryotic or eukaryotic cells

The initial successes of recombinant DNA technology were achieved using bacteria as hosts. As noted in preceding chapters, bacterial cells are easily grown and manipulated in the laboratory. Much of their molecular biology is known, especially for certain well-studied bacteria, such as *E. coli*, and they have numerous genetic markers that can be used to select for cells harboring the recombinant DNA. Bacteria also contain small circular chromosomes called *plasmids*, which, as we will see, can be manipulated to carry recombinant DNA into the cell.

In some important ways, however, bacteria are not ideal organisms for studying and expressing eukaryotic genes. Bacteria lack the splicing machinery to excise introns from the initial RNA transcript of eukaryotic genes. In addition, many eukaryotic proteins are extensively modified after translation by reactions such as glycosylation and phosphorylation. Often these modifications are essential for the protein's activity. Finally, in some instances, the expression of the new gene in a eukaryote is the point of the experiment—that is, the aim is to produce a transgenic organism. In these cases, the host for the new DNA may be a mouse, a wheat plant, a yeast, or a human, to name just a few examples.

Yeasts such as *Saccharomyces* are common eukaryotic hosts for recombinant DNA studies. The advantages of using yeasts include rapid cell division (a life cycle completed in 2–8 hours), ease of growth in the laboratory, and a relatively small genome size (about 20 million base pairs and 6,000 genes). The yeast genome is several times larger than that of *E. coli*, and has one-fourth the number of genes as the human genome. Nevertheless, yeasts have most of the characteristics of other eukaryotes, except for those involved in multicellularity.

Plant cells can also be used as hosts, especially if the desired result is a transgenic plant. The property that makes plant cells good hosts is their *totipotency*—that is, the ability of a differentiated cell to act like a fertilized egg and produce an entire new organism. Isolated plant cells grown in culture can take up recombinant DNA, and by manipulation of the growth medium, these transgenic cells can be induced to

form an entire new plant, which can then be reproduced naturally in the field. The transgenic plant will carry and express the gene that is part of recombinant DNA.

Whatever host is chosen, a vehicle for carrying the DNA into the cell is needed. These vehicles are called **vectors**.

Vectors can carry new DNA into host cells

In natural environments, DNA released from one bacterium can sometimes be taken up by another bacterium and genetically transform it (see Chapter 11), but this phenomenon is not common. The challenge of inserting new DNA into a cell lies not just in getting it into the host cell, but in getting it to replicate in the host cell as it divides. DNA polymerase, the enzyme that catalyzes replication, does not bind to just any sequence of DNA to begin replication. Rather, it recognizes a specific sequence, the *origin of replication* (see Chapter 11). If the new DNA is to be replicated, it must become part of a segment of DNA that contains an origin of replication, called a **replicon**, or **replication unit**.

There are two general ways in which the newly introduced DNA can become part of a replication unit. First, it can be inserted into a host chromosome after entering the host cell. Although this insertion is often a random event, it is nevertheless a common method of integrating a new gene into a host cell. Alternatively, the new DNA can enter the host cell as part of a carrier DNA sequence—the vector— that already has the appropriate origin of replication.

A vector should have four characteristics:

- The ability to replicate independently in the host cell
- A recognition sequence for a restriction enzyme, allowing the vector to be cut and combined with the new DNA
- A reporter gene that will announce its presence in the host cell
- A small size in comparison to the host chromosomes for ease of isolation

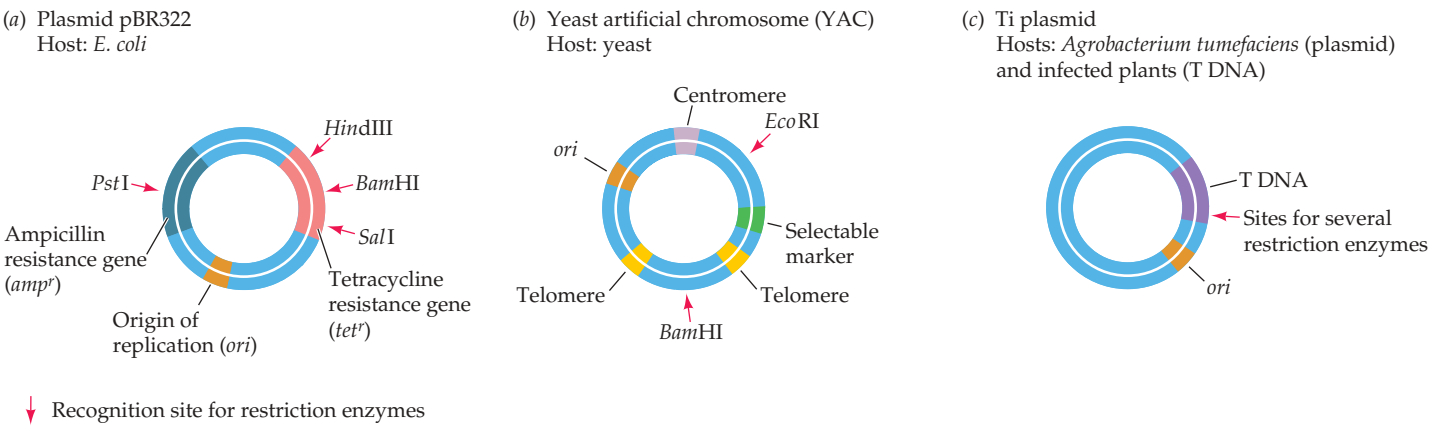
PLASMIDS AS VECTORS. Plasmids, which, as we saw in Chapter 13, are naturally occurring bacterial chromosomes, have all four of the properties needed for a useful vector. First, they are small (an *E. coli* plasmid has 2,000–6,000 base pairs, as compared with the main *E. coli* chromosome, which has more than 3 million base pairs). Furthermore, because it is so small, a plasmid often has only a single recognition site for a given restriction enzyme (Figure 16.5*a*). This property is essential because it allows for the insertion of new DNA at only one location (see Figure 16.4). When the plasmid is cut with a restriction enzyme, it is transformed into a linear molecule with sticky ends. The sticky ends of another DNA fragment cut with the same restriction enzyme can pair with the sticky ends of the plasmid, resulting in a circular plasmid containing the new DNA.

Two other characteristics make plasmids good vectors. As we have seen, many plasmids contain genes that confer resistance to antibiotics. This property provides a genetic marker for host cells carrying the recombinant plasmid. Finally, plasmids have an origin of replication and can replicate independently of the host chromosome. It is not uncommon for a bacterial cell with a single main chromosome to contain hundreds of copies of a recombinant plasmid.

The plasmids commonly used as vectors in the laboratory have been extensively altered by recombinant DNA technology, and most are combinations of genes and other sequences from several sources. Many of these plasmids have a single marker for antibiotic resistance.

VIRUSES AS VECTORS. Constraints on plasmid replication limit the size of the new DNA that can be inserted into a

16.5 Vectors for Carrying DNA into Cells (*a*) A plasmid with genes for antibiotic resistance can be incorporated into an *E. coli* cell. (*b*) A DNA molecule synthesized in the laboratory constitutes a chromosome that can carry its inserted DNA into yeasts. (*c*) The Ti plasmid, isolated from the bacterium *Agrobacterium tumefaciens*, is used to insert DNA into many types of plants.



plasmid to about 10,000 base pairs. Although some prokaryotic genes may be this small, 10,000 base pairs is much smaller than most eukaryotic genes with their introns and extensive flanking sequences. A vector that accommodates larger DNA inserts is needed.

Both prokaryotic and eukaryotic viruses are often used as vectors for eukaryotic DNA. Bacteriophage λ, which infects *E. coli*, has a DNA genome of about 45,000 base pairs. If the genes that cause the host cell to die and lyse—about 20,000 base pairs—are eliminated, the virus can still attach to a host cell and inject its DNA. The deleted 20,000 base pairs can be replaced with DNA from another organism, thereby creating usable recombinant viral DNA.

Because viruses infect cells naturally, they offer a great advantage as vectors over plasmids, which often require artificial means to coax them to enter cells. As we will see in Chapter 17, viruses are important vectors in human gene therapy.

ARTIFICIAL CHROMOSOMES AS VECTORS. Bacterial plasmids are not good vectors for yeast hosts because prokaryotic and eukaryotic DNA sequences use different origins of replication. Thus a recombinant bacterial plasmid will not replicate in yeast. To remedy this problem, scientists have created a "minimalist chromosome" called the **yeast artificial chromosome**, or **YAC** (Figure 16.5*b*). This artificial DNA molecule contains not only the yeast origin of replication, but the yeast centromere and telomere sequences as well, making it a true eukaryotic chromosome. YACs also contain artificially synthesized restriction sites and useful reporter genes (for yeast nutritional requirements). YACs are only about 10,000 base pairs in size, but can accommodate 50,000 to 1.5 million base pairs of inserted DNA. These artificial chromosomes carry out eukaryotic DNA replication and gene expression normally in yeast cells.

PLASMID VECTORS FOR PLANTS. An important vector for carrying new DNA into many types of plants is a plasmid that is found in *Agrobacterium tumefaciens*. This bacterium lives in the soil and causes a plant disease called crown gall, which is characterized by the presence of growths, or tumors, in the plant. *A. tumefaciens* contains a plasmid called Ti (for tumor-inducing) (Figure 16.5*c*).

The Ti plasmid contains a transposon, called T DNA, that inserts copies of itself into the chromosomes of infected plant cells. The T DNA contains recognition sequences for restriction enzymes, so that new DNA can be inserted into it. When the T DNA is thus altered, the plasmid no longer produces tumors, but the transposon, with the new DNA, can still be inserted into the host cell's chromosomes. A plant cell containing this DNA can then be grown in culture or induced to form a new, transgenic plant.

Whatever vector is effective, the problem of identifying those host cells that actually contain the recombinant DNA remains.

Reporter genes identify host cells containing recombinant DNA

Even when a population of host cells interacts with an appropriate vector, only a small proportion of the cells actually take up the vector. Also, since the process of cutting the vector and inserting the new DNA to make recombinant DNA is far from perfect, only a few of the vectors that have moved into the host cells will actually contain the DNA sequence of interest. How can we select only the host cells that contain the recombinant DNA?

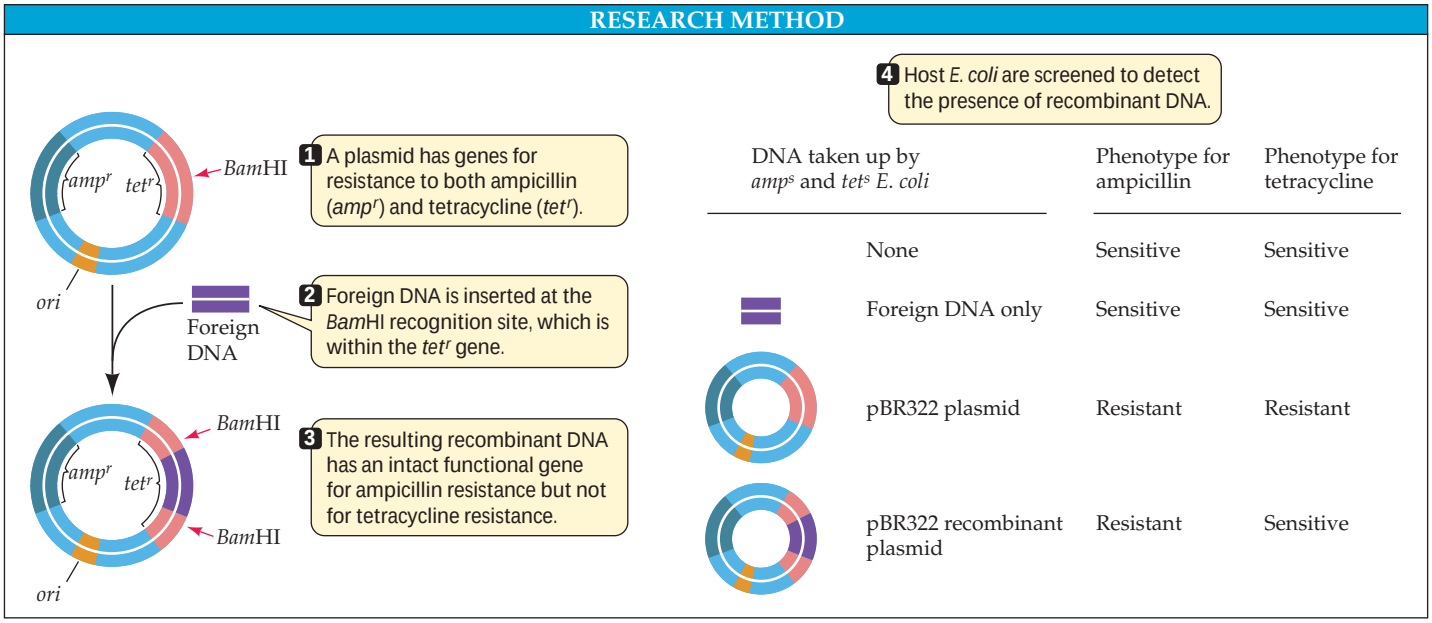
The procedure we are about to describe illustrates an elegant, commonly used approach to this problem. In this example, we use *E. coli* bacteria as hosts and a plasmid vector (see Figure 16.5*a*) that carries the genes for resistance to the antibiotics ampicillin and tetracycline.

When the pBR322 plasmid is incubated with the restriction enzyme *Bam*HI, the enzyme encounters its recognition sequence, GGATCC, only once, at a site within the gene for tetracycline resistance. If foreign DNA is inserted into this restriction site, the presence of these "extra" base pairs within the tetracycline resistance gene inactivates it. So plasmids containing the inserted DNA will carry an intact gene for ampicillin resistance, but not an intact gene for tetracycline resistance (Figure 16.6). This difference is the key to the selection of the host bacteria that contain the recombinant plasmid.

The cutting and insertion process results in three types of DNA, all of which can be taken up by host bacteria:

- The recombinant plasmid—the one we want—turns out to be the rarest type of DNA. Its uptake confers resistance to ampicillin, but not to tetracycline, on host *E. coli*.
- More common are bacteria that take up plasmids that have sealed their own ends back together. These plasmids retain intact genes for resistance to both ampicillin and tetracycline.
- Even more common are bacteria that take up the foreign DNA sequence alone, without the plasmid; since it is not part of a replication unit, it does not survive as the bacteria divide. These host cells remain susceptible to both antibiotics.

The vast majority (more than 99.9 percent) of host cells take up no DNA at all and remain susceptible to both antibiotics. So the unique drug-resistant phenotype of the cells with recombinant DNA (tetracycline-sensitive and ampicillin-resistant) marks them in a way that can be detected by simply adding ampicillin and/or tetracycline to the medium surrounding the cells.



16.6 Marking Recombinant DNA by Inactivating a Gene
Scientists can inactivate reporter genes within plasmids to mark the host cells that have incorporated the recombinant DNA. The host bacteria in this experiment could display any of the three phenotypes indicated in the table.

In addition to genes for antibiotic resistance, several other reporter genes are used to detect recombinant DNA in host cells. Scientists have created several artificial vectors that include restriction sites within the *lac* operon (see Figure 13.17). When the *lac* operon is inactivated by the insertion of foreign DNA, the vector no longer carries its function into the host cell. Other reporter genes that have been used in vectors include the gene for luciferase, the enzyme that causes fireflies to glow in the dark; this enzyme causes host cells to glow when supplied with its substrate. Green fluorescent protein, which normally occurs in the jellyfish *Aequopora victoriana*, does not require a substrate, but emits visible light when exposed to ultraviolet light, and is now widely used as a genetic marker, as described at the beginning of this chapter.

Many commonly used plasmid vectors contain only a single reporter gene for antibiotic resistance, which does not contain a restriction site. In this case, the recombinant plasmid will have the same antibiotic resistance gene that a nonrecombinant plasmid does. The formation of recombinant DNA is favored, however, if there is a high concentration of foreign DNA fragments compared with that of the cut plasmid. So there will be a preponderance of host cells containing recombinant DNA among those that survive in the presence of the antibiotic.

After exposure to the vector, host cells are usually first grown on a solid medium. If the concentration of cells dispersed on the solid medium is low, each cell will divide and grow into a distinct colony (see Chapter 13). The colonies that contain recombinant DNA can be identified by reporter gene expression and removed from the medium, then grown in large amounts in liquid culture. A quick examination of a plasmid can confirm whether the cells of the colony actually have the recombinant DNA. The power of bacterial transfection to amplify a gene is indicated by the fact that a 1-liter culture of bacteria harboring the human β-globin gene in the pBR322 plasmid has as many copies of that gene as the sum total of all the cells in a typical adult human being (10^{14}).

Sources of Genes for Cloning

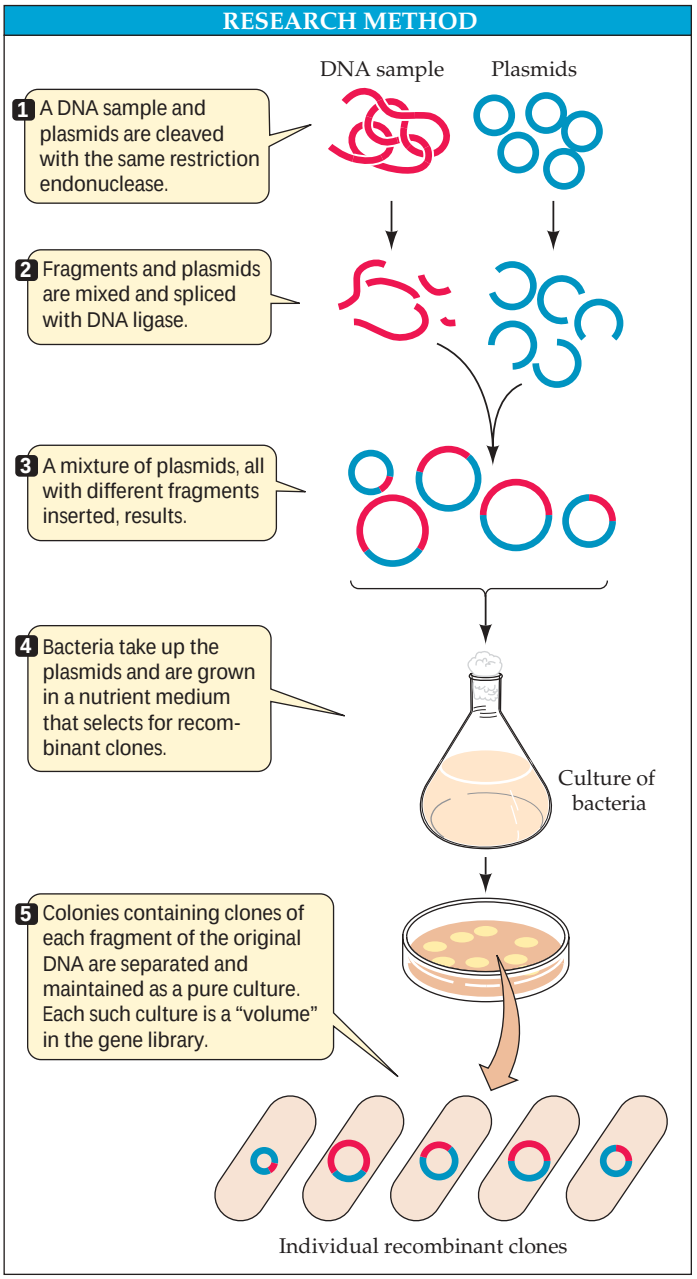
In the preceding section, we have seen how DNA can be cut, inserted into a vector, and transfected into host cells, and how host cells carrying recombinant DNA can be identified. Now we will pause briefly to consider where the genes or DNA fragments used in these procedures come from. They are obtained from three principal sources: random pieces of chromosomes maintained as gene libraries, complementary DNA obtained by reverse transcription from mRNA, and artificial synthesis or mutation of DNA.

Gene libraries contain pieces of a genome

The 23 pairs of human chromosomes can be thought of as a library that contains the entire genome of our species. Each chromosome, or "volume" in the library, contains, on average, 80 million base pairs of DNA, encoding several thousand genes. Such a huge molecule is not very useful for studying genomic organization or for isolating a specific gene.

To address this problem, researchers can use restriction enzymes to break each chromosome into smaller pieces, then analyze each piece. These smaller DNA fragments still represent a **gene library** (Figure 16.7); however, the information is now in many more than 23 volumes. Each fragment can be inserted into a vector, which can then be taken up by a host cell. When bacteria are used as hosts, proliferation of one cell produces a colony of recombinant cells, each of which harbors many copies of the same fragment of human DNA.

Using plasmids, which are able to insert up to 10,000 base pairs of foreign DNA into a bacterium, about 200,000 separate fragments are required to make a library of the human genome. By using phage λ, which can carry four times as much DNA as a plasmid, the number of volumes can be reduced to about 250,000. Although this seems like a large number, a single growth plate can hold up to 80,000 phage colonies, or plaques, and is easily screened for the presence of a particular DNA sequence by denaturing the phage DNA and applying a particular probe.



16.7 Constructing a Gene Library Human chromosomes are broken up into fragments of DNA using restriction enzymes. The fragments are inserted into vectors (plasmids are shown here) and taken up by host bacterial cells, each of which then harbors a single fragment of the human DNA. The information in the resulting bacterial cultures and sets of colonies constitutes a gene library.

A DNA copy of mRNA can be made by reverse transcriptase

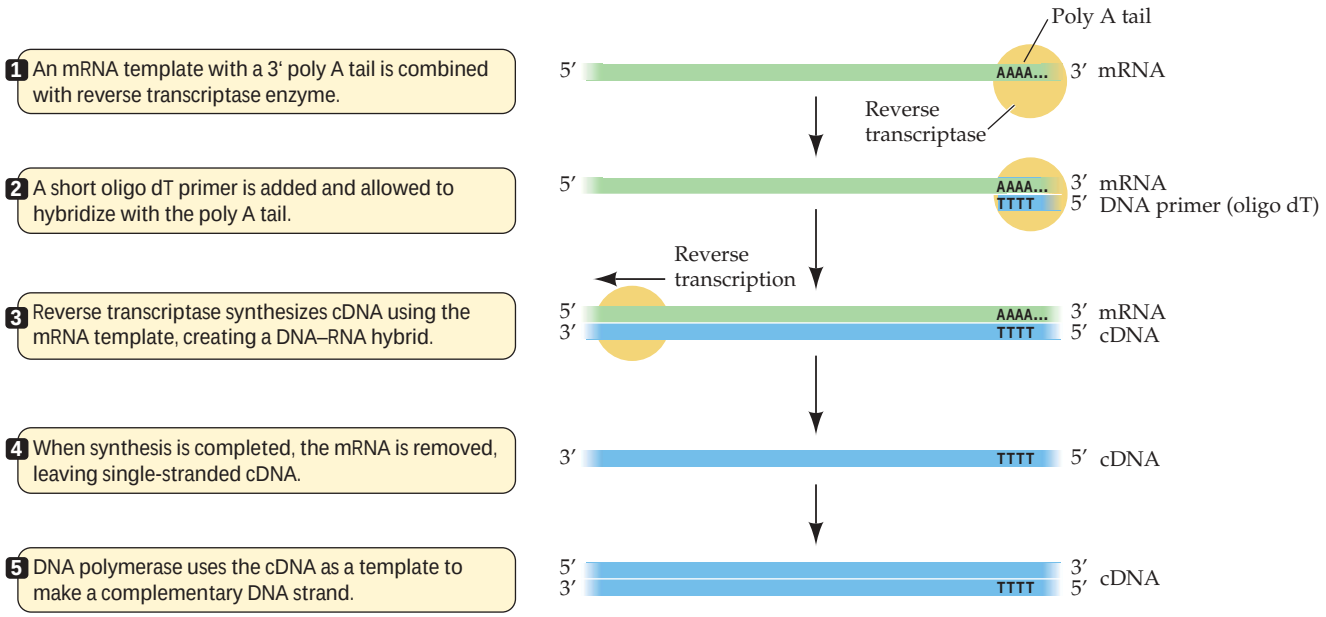
A much smaller DNA library—one that includes only the genes transcribed in a particular tissue—can be made from **complementary DNA**, or **cDNA** (Figure 16.8). Recall that most eukaryotic mRNAs have a poly A tail—a string of adenine residues at their 3′ end (see Figure 14.9). The first step in cDNA production is to extract mRNA from a tissue and allow it to hybridize with a molecule called *oligo dT*, which consists of a string of thymine residues (the "d" indicates deoxyribose). The oligo dT hybridizes with the poly A tail of the mRNA. The oligo dT serves as a primer, and the mRNA as a template, for the enzyme reverse transcriptase, which synthesizes DNA from RNA. In this way, a cDNA strand complementary to the mRNA is formed.

A collection of cDNAs from a particular tissue at a particular time in the life cycle of an organism is called a *cDNA library*. Messenger RNAs do not last long in the cytoplasm and are often present in small amounts, so a cDNA library is a "snapshot" that preserves the transcription pattern of the cell. cDNA libraries have been invaluable in comparisons of gene expression in different tissues at different stages of development. Their use has shown, for example, that up to one-third of all the genes of an animal are expressed only during prenatal development. Complementary DNA is also a good starting point for the cloning of eukaryotic genes. It is especially useful for cloning genes expressed at low levels in only a few cell types.

DNA can be synthesized chemically in the laboratory

If we know the amino acid sequence of a protein, we can apply organic chemistry to make the DNA that codes for that protein. Artificial DNA synthesis has even been automated, and at many institutions, a special service laboratory can make short to medium-length sequences overnight for any number of investigators.

How do we design a synthetic gene? Using the genetic code and the known amino acid sequence, we can figure out



16.8 Synthesizing Complementary DNA Gene libraries that include only genes transcribed in a particular tissue at a particular time can be made from complementary DNA. cDNA synthesis is especially useful for identifying mRNAs that are present only in a few copies, and is often a starting point for gene cloning.

an appropriate base sequence for the gene. With this sequence as a starting point, we can add other sequences, such as codons for translation initiation and termination and flanking sequences for transcription initiation, termination, and regulation. Of course, these noncoding DNA sequences must be the ones actually recognized by the host cell if the synthetic gene is to be transcribed. It does no good to have a prokaryotic promoter sequence near a gene if that gene is to be inserted into a yeast cell for expression. Codon usage is also important: as we have seen, many amino acids are encoded by more than one codon, and different organisms stress the use of different synonymous codons.

DNA can be mutated in the laboratory

Mutations that occur in nature have been important in demonstrating cause-and-effect relationships in biology. For example, in Chapter 14, we learned that some people with the disease beta thalassemia have a mutation at a consensus sequence for intron removal in the β-globin gene and so cannot make proper β-globin mRNA. This discovery revealed the importance of the consensus sequence. Recombinant DNA technology has allowed us to ask such "What if?" questions without having to look for mutations in nature. Because synthetic DNA can be made in any sequence desired, we can manipulate DNA to create specific mutations and then see what happens when the mutant DNA expresses itself in a host cell. Additions, deletions, and base-pair substitutions are all possible with isolated or synthetic DNA.

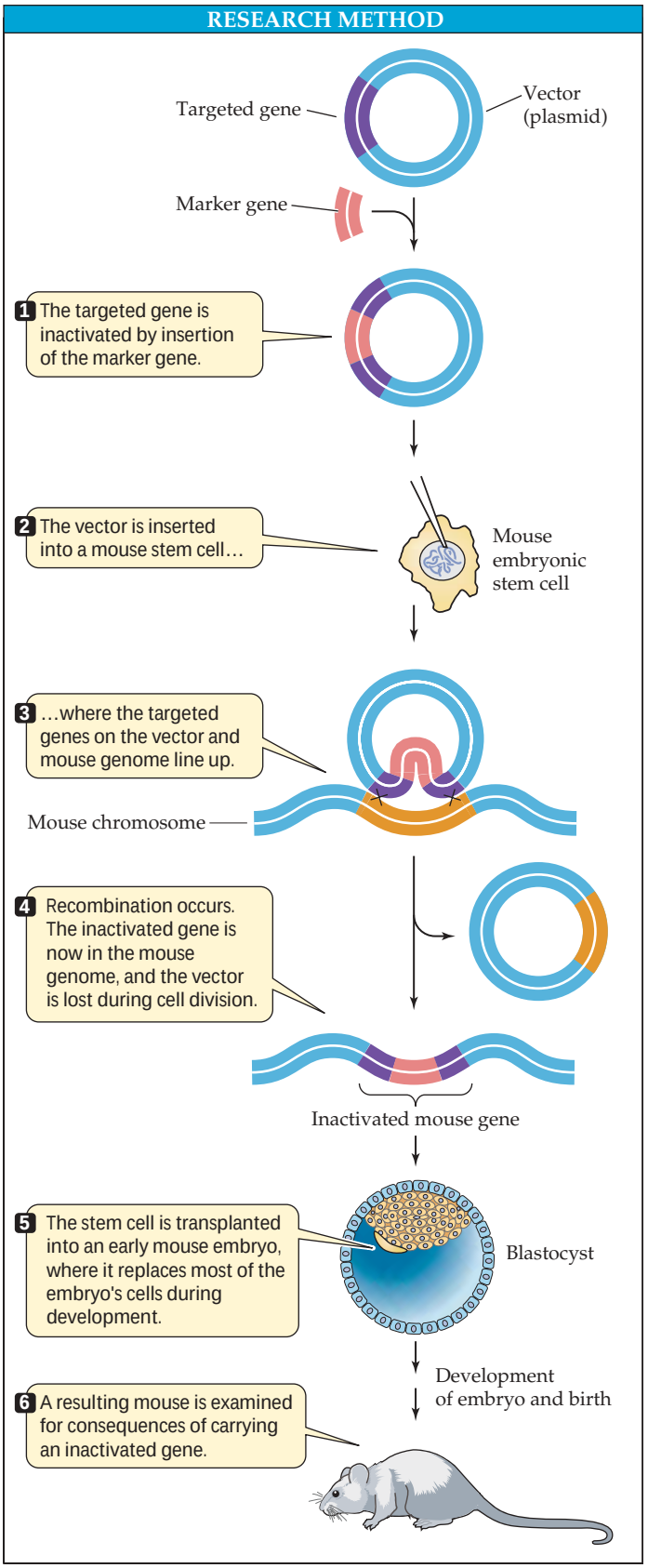
These mutagenesis techniques have led to many cause-and-effect proofs. For example, it was hypothesized that the signal sequence at the beginning of a secreted protein is essential to its passage through the membrane of the endoplasmic reticulum. Thus, a gene coding for such a protein, but with the codons for the signal sequence deleted, was synthesized. Sure enough, when this gene was expressed in yeast cells, the protein did not cross the ER membrane. When the signal sequence codons were added to an unrelated gene encoding a soluble cytoplasmic protein, that protein did cross the ER membrane.

Some Additional Tools for DNA Manipulation

In Chapter 11, we described DNA sequencing and the polymerase chain reaction, two applications of DNA replication techniques. Here, we examine four additional techniques for manipulating DNA. One is the use of genetic recombination to create an inactive, or "knocked-out," gene. The second is the use of DNA chips to detect the presence of many different sequences simultaneously. The third is the use of antisense RNA and RNA interference to block the translation of specific mRNAs. The fourth is a method for determining which proteins interact in a cell.

Genes can be inactivated by homologous recombination

As we have seen, artificial mutations provide an excellent way of asking "What if" questions about the role of a gene in cell function. **Homologous recombination** can be used to ask these questions at the organism level. The aim of this tech-



16.9 Making a Knockout Mouse Homologous recombination is used to replace a normal mouse gene with an inactivated copy of that gene, thus "knocking out" the gene. Discovering what happens to a mouse with an inactive gene tells us much about the normal role of that gene.

nique is to replace a gene inside a cell with an inactivated form of that gene, then see what happens when the inactive gene is part of an organism. Such a manipulation is called a **knockout** experiment.

Mice are frequently used in knockout experiments (Figure 16.9). The normal allele of the mouse gene to be tested is inserted into a plasmid. Restriction enzymes are then used to insert a fragment containing a genetic marker into the middle of the normal gene. This addition of extra DNA plays havoc with the targeted gene's transcription and translation; a functional mRNA is seldom made from such an interrupted gene. Next, the plasmid is transfected into a stem cell in an early mouse embryo. A *stem cell* is an undifferentiated cell that divides and differentiates into specialized cell(s).

Because much of the targeted gene is still present in the plasmid (although in two separated regions), there is homologous sequence recognition between the inactive allele on the plasmid and the active (normal) allele in the mouse genome. The plasmid lines up with a mouse chromosome, and sometimes, recombination occurs such that the plasmid's inactive allele is swapped for the functional allele in the host cell. Now neither allele can be expressed: the one inserted into the mouse chromosome is still only an interrupted fragment of the normal gene, and the gene inserted into the plasmid usually lacks its promoter.

The reporter gene in the insert is used to identify those stem cells carrying the inactivated gene. A transfected stem cell is now transplanted into an early mouse embryo, and through some clever tricks, a knockout mouse carrying the inactivated gene in homozygous form is produced. The changed phenotype of the mutant mouse gives a clue to the role of the gene in the normal, wild-type animal. The knockout technique has been important in assessing the roles of genes during development.

DNA chips can reveal DNA mutations and RNA expression

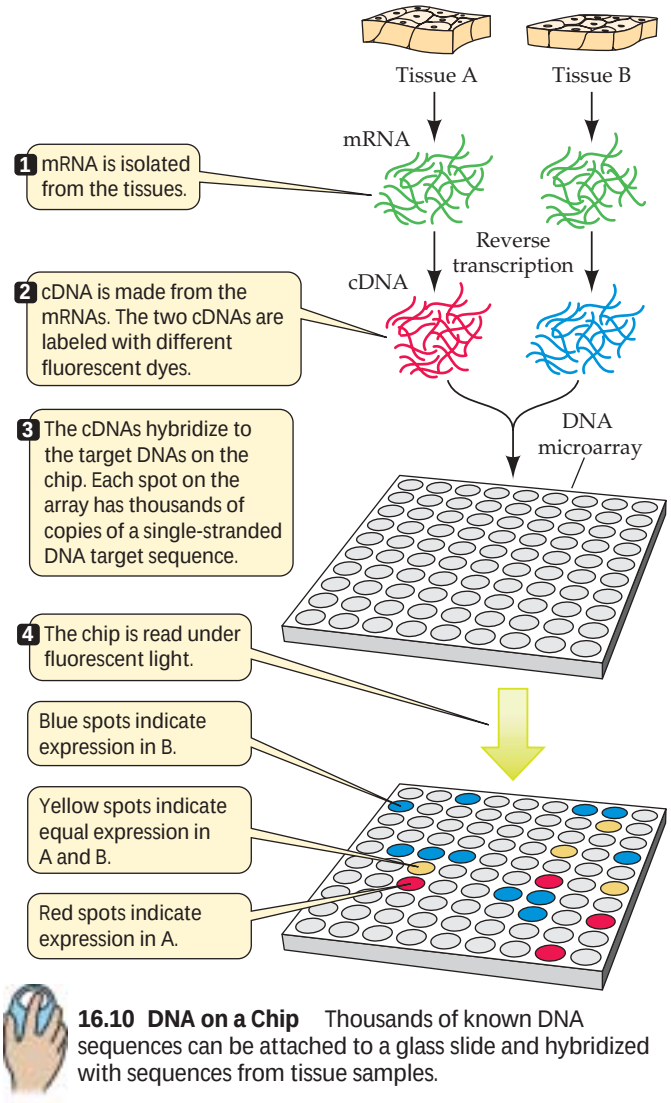
The emerging science of genomics must deal with two major quantitative realities. First, there are a large number of genes in eukaryotic genomes. Second, the pattern of gene expression in different tissues at different times is quite distinctive. For example, a skin cancer cell at its early stage may have a unique mRNA "fingerprint" that differs from that of both normal skin cells and the cells of a more advanced skin cancer.

To find these patterns, scientists could isolate the mRNA from a cell and test it by hybridization with each gene in the genome, one gene at a time. But it would be far simpler to do these hybridizations all in one step. To facilitate this, one needs some way to arrange all the genes in a genome in an array on some solid support.

DNA chip technology provides these large arrays of sequences for hybridization. DNA chips were developed by modifying methods that have been used for decades in the semiconductor industry. You may be familiar with the silicon microchip, in which an array of microscopic electric circuits is etched onto a tiny chip. In the same way, DNA chips are glass slides to which a series of DNA sequences are attached in a precise order (Figure 16.10). Typically, the slide is divided into 24 × 24 μM squares, each of which contains about 10 million copies of a particular sequence up to 20 nucleotides long. A computer controls the addition of nucleotides in a predetermined pattern. Each 20-base-long sequence hybridizes to only one genomic DNA (or cDNA) sequence, and thus is a unique identifier of a gene. Up to 60,000 different sequences can be placed on a single chip.

If cellular mRNA is to be analyzed, it is usually incubated with reverse transcriptase (RT) to make cDNA (see Figure 16.8), and the cDNA is amplified by the polymerase chain reaction (PCR) prior to hybridization (see Figure 11.20). This technique is called **RT-PCR**, and it ensures that mRNA sequences naturally present in only a few copies (or in a small sample, such as a cancer biopsy) will be numerous enough to form a signal. The amplified cDNAs are coupled to a fluorescent dye and used to probe the DNA on a chip. Those cDNA sequences that form hybrids can be located by a sensitive scanner. With the number of genes that can be placed on a chip approaching that of the largest genomes, DNA chips will result in an information explosion on mRNA transcription patterns in cells in different physiological states.

Another use for DNA chips is in detecting genetic variants. Suppose one wants to find out if a particular gene, which is 5,500 base pairs long, has any mutations in a particular individual. One way would be to sequence the entire gene, but that would be expensive and time-consuming. On the other hand, DNA chip technology can be used to make 20-nucleotide fragments including the entire gene and all (or nearly all) of its known point mutations and small deletions. Then, hybridization with the individual's DNA might reveal a particular mutation if it hybridized to a mutant sequence on the chip. This rapidly developing technology could be an important step toward individualized diagnosis and therapy for human genetic diseases.



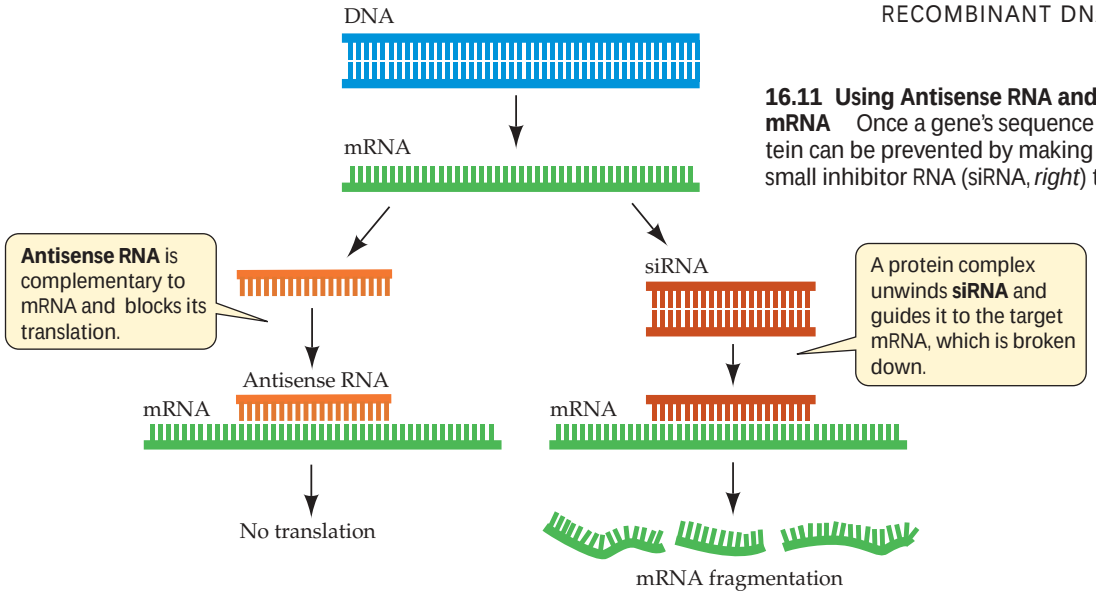
16.10 DNA on a Chip Thousands of known DNA sequences can be attached to a glass slide and hybridized with sequences from tissue samples.

Antisense RNA and RNA interference can prevent the expression of specific genes

The base-pairing rules can be used not only to make artificial genes, but they can also be employed to stop the translation of mRNA. As is often the case, this technique is an example of scientists imitating nature. In normal cells, a rare mechanism for controlling gene expression is the production of an RNA molecule that is complementary to mRNA. This complementary molecule is called **antisense RNA** because it binds by base pairing to the "sense" bases on the mRNA that codes for a protein. The formation of a double-stranded RNA hybrid inhibits translation of the mRNA, and the hybrid tends to be broken down rapidly in the cytoplasm. Although the gene continues to be transcribed, translation does not take place.

After determining the sequence of a gene and its mRNA in the laboratory, scientists can make and add specific antisense RNA to a cell to prevent translation of that gene's mRNA (Figure 16.11, left). The antisense RNA can be added as itself—RNA can be inserted into cells in the same way that DNA is—or it can be made in the cell by transcription from a DNA molecule introduced as a part of a vector.

A related technique takes advantage of **interference RNA** (**RNAi**), a rare way of naturally inhibiting mRNA translation, such as occurs in the inactivation of the X chromosome (see Figure 14.18). In this case, a short (about 20 nucleotides) double-stranded RNA is unwound to single strands by a protein complex that guides this RNA to a complementary region on mRNA. The protein complex catalyzes the breakdown of the targeted mRNA.



16.11 Using Antisense RNA and RNAi to Block Translation of mRNA Once a gene's sequence is known, the synthesis of its protein can be prevented by making either an antisense RNA *(left)* or a small inhibitor RNA (siRNA, *right*) that is complementary to its mRNA.

Armed with this knowledge, scientists can custom-synthesize a *small interfering RNA* (siRNA) to inhibit the translation of *any* known gene (Figure 16.11, right). Because these double-stranded siRNAs are more stable than antisense RNAs, RNAi is a much easier technique to use than antisense RNA.

Antisense RNA and RNAi have been widely used to test cause-and-effect relationships. For example, when antisense RNA was used to block the synthesis of a protein essential for the growth of cancer cells, the cells reverted to a normal phenotype. Gene silencing offers great potential for the development of drugs to treat diseases that are the result of the inappropriate expression of specific genes.

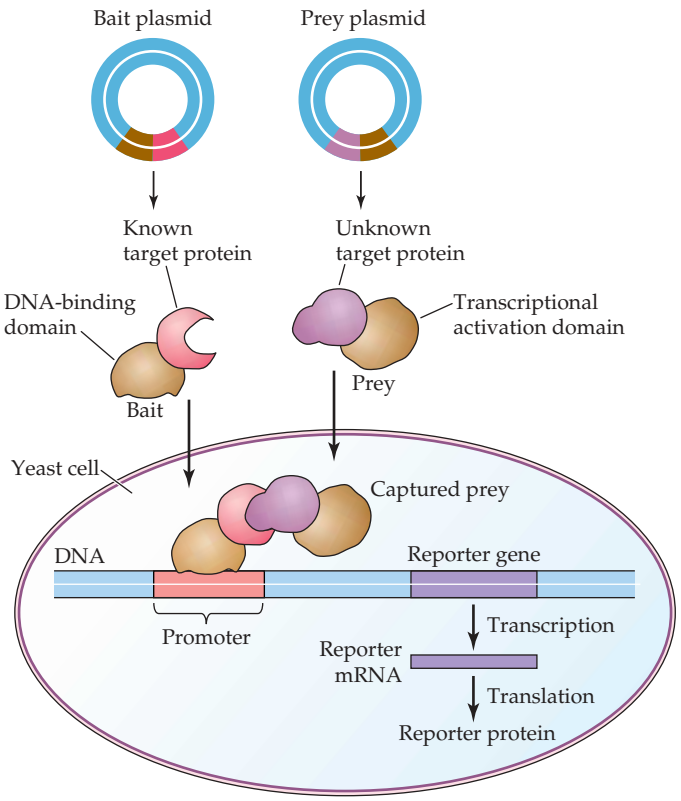
The two-hybrid system shows which proteins interact in a cell

Proteins often interact with other proteins in a cell, and their interaction leads to an important cellular function. Scientists can determine which proteins bind to which in several ways. One is the test-tube approach, in which one protein acts as a "hook" and the scientist "goes fishing" for proteins that bind to it. A better way is to set up a system that tests for protein interactions in a living cell. That is the purpose of a **two-hybrid system**.

A two-hybrid system uses a transcription factor that activates the transcription of an easily detectable reporter gene. This transcription factor has two domains: one binds to DNA at the promoter, and the other binds to another protein in the transcription complex to activate transcription of the reporter gene.

In the yeast two-hybrid system (Figure 16.12), these two domains of the transcription factor are separated in two different recombinant yeast plasmids. One plasmid makes the DNA-binding domain only, and the gene encoding this domain is fused to a gene that makes the target protein—the one whose binding partner we wish to determine. The resulting hybrid protein will bind to the promoter of the reporter gene, but will not activate it, leaving the target protein exposed. This hybrid protein is the "bait."

The second yeast plasmid has a gene encoding the activating domain of the transcription factor. This gene is fused



16.12 The Two-Hybrid System The two-hybrid system is a way to determine which proteins interact with one another in a living cell. In this example, the synthesis of the reporter protein is a positive result, revealing that the target protein (the "prey") binds specifically to a possible binding partner (the "bait").

to a gene encoding a protein (the "prey") to be tested for binding to the target protein. When this hybrid gene is introduced into the "bait" strain of yeast cells, the "bait" and "prey" proteins may bind. This binding activates transcription of the reporter gene. If no binding occurs, no reporter protein is made. This method has revealed hundreds of protein-protein interactions in cells.

Biotechnology: Applications of DNA Manipulation

As we have just seen, the development of methods for manipulating DNA has given us the ability to perform experiments that reveal the details of life at the molecular level. But in addition to their use as research tools, these methods are being used for human benefit. **Biotechnology** is the use of living cells to produce materials useful to people, such as foods, medicines, and chemicals. People have been doing this for a very long time. For example, the use of yeasts to brew beer and wine dates back at least 8,000 years, and the use of bacterial cultures to make cheese and yogurt is a technique many centuries old. For a long time, however, people were not aware of the cellular bases of these biochemical transformations.

About 100 years ago, thanks largely to Pasteur's work, it became clear that specific bacteria, yeasts, and other microbes could be used as biological converters to make certain products. Alexander Fleming's discovery that the mold *Penicillium* makes the antibiotic penicillin led to the large-scale commercial culture of microbes to produce antibiotics as well as other useful chemicals. Today, microbes are grown in vast quantities to make much of the industrial-grade alcohol, glycerol, butyric acid, and citric acid that are used by themselves or as starting materials in the manufacture of other products.

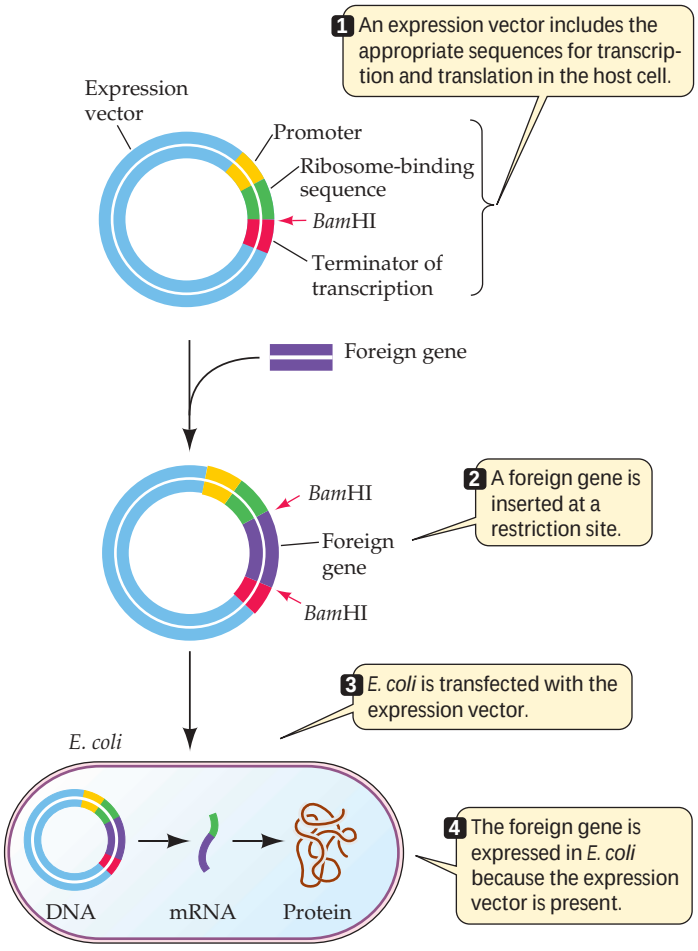
In the past, the list of such products was limited to those that were naturally made by microbes. The many products that multicellular eukaryotes make, such as hormones and certain enzymes, had to be extracted from those complex organisms. Yields were low, and purification was difficult and costly. All this has changed with the advent of gene cloning. Our ability to insert almost any gene into bacteria or yeast, along with methods to induce the gene to make its product in large amounts and export it from the cells, has turned these microbes into versatile factories for important products. The key technology for turning cells into factories has been the development of specialized vectors that not only carry genes into cells, but make those cells express them at high levels.

Expression vectors can turn cells into protein factories

If a eukaryotic gene is inserted into a typical plasmid (see Figure 16.5*a*) and transfected into *E. coli*, little, if any, of the product of the gene will be made by the host cell. The reason is that the eukaryotic gene lacks the bacterial promoter for RNA polymerase binding, the terminator for transcription, and a special sequence on mRNA that is necessary for ribosome binding. All of these elements are necessary for the gene to be expressed and its product synthesized in the bacterial cell.

To solve this problem, scientists can make **expression vectors** that have all the characteristics of typical vectors as well as the extra sequences needed for the foreign gene to be expressed in the host cell. For bacterial hosts, these additional sequences include the elements named above (Figure 16.13); for eukaryotes, they include the poly A addition sequence, transcription factor binding sites, and enhancers. Once these sequences are placed at the appropriate location in the vector, a transfected gene can be expressed in almost any kind of host cell.

An expression vector can be modified in various ways. An *inducible promoter*, which responds to a specific signal, can be made part of an expression vector. For example, a specific



16.13 An Expression Vector Allows a Foreign Gene to Be Expressed in a Host Cell A transfected eukaryotic gene may not be expressed in *E. coli* because it lacks the necessary bacterial sequences for promotion, termination, and ribosome binding. Expression vectors contain these additional sequences, enabling the eukaryotic protein to be synthesized in the prokaryotic cell.

promoter that responds to hormonal stimulation can be used so that the foreign gene can be induced to transcribe its mRNA when the hormone is added. An enhancer that responds to hormonal stimulation can also be added so that transcription and protein synthesis will occur at high rates—a goal of obvious importance in the manufacture of an industrial product.

A *tissue-specific promoter*, which is expressed only in a certain tissue at a certain time, can be used if localized expression is desired. For example, many seed proteins are expressed only in the plant embryo. Coupling a gene to a seed-specific promoter will allow the gene to be expressed only as a seed protein.

Targeting sequences can be added to the expression vector so that the product of the gene is directed to an appropriate destination. For example, when yeast or bacterial cells making a protein are maintained in a large vessel, it is economical to include an export signal for the protein to be secreted into the extracellular medium for easier recovery.

Medically useful proteins can be made by biotechnology

Many medically useful products are being made by biotechnology (Table 16.1), and hundreds more are in various stages of development. The development of one such product, tissue plasminogen activator, illustrates the techniques that have been used.

In most people, when a wound begins bleeding, a blood clot soon forms to stop the flow. Later, as the wound heals, the clot dissolves. How does the blood perform these conflicting functions at the right times? Mammalian blood contains an enzyme called plasmin that catalyzes the dissolution of the clotting proteins. But plasmin is not always active; if it were, a blood clot would dissolve as soon as it formed! Instead, plasmin is "stored" in the blood in an inactive form called plasminogen. The conversion of plasminogen to plasmin is activated by an enzyme appropriately called tissue plasminogen activator (TPA), which is produced by cells lining the blood vessels:

$$\text{plasminogen (inactive)} \xrightarrow{\text{TPA}} \text{plasmin (active)}$$

Heart attacks and many strokes are caused by blood clots that form in major blood vessels leading to the heart or the brain, respectively. During the 1970s, a bacterial enzyme called streptokinase was found to stimulate the dissolution of clots in some patients. Treating these persons with this enzyme saved lives, but its use had side effects. Streptokinase was a protein foreign to the body, so patients' immune systems reacted against it. More important, the drug sometimes prevented clotting throughout the entire circulatory system, leading to an almost hemophilia-like condition in some patients.

The discovery of TPA and its isolation from human tissues led to the hope that this enzyme would bind specifically to clots, and that it would not provoke an immune reaction. But the amounts of TPA available from human tissues were tiny, certainly not enough to inject at the site of a clot in the emergency room.

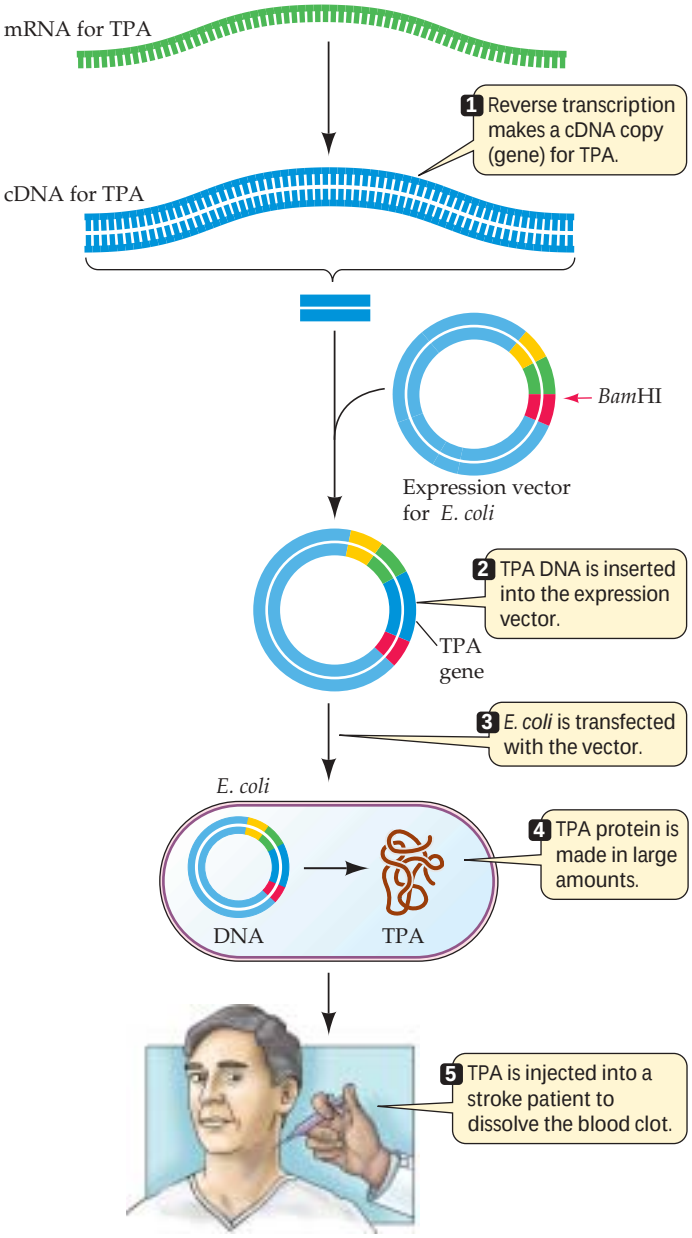
Recombinant DNA technology solved this problem. TPA mRNA was isolated and used to make a cDNA copy, which was then inserted into an expression vector and transfected into *E. coli* (Figure 16.14). The transgenic bacteria made the protein in quantity, and it soon became available commercially. This drug has had considerable success in dissolving blood clots in people undergoing heart attacks and, especially, strokes.

16.1 Some Medically Useful Products of Biotechnology

PRODUCT	USE
Colony-stimulating factor	Stimulates production of white blood cells in patients with cancer and AIDS
Erythropoietin	Prevents anemia in patients undergoing kidney dialysis and cancer therapy
Factor VIII	Replaces clotting factor missing in patients with hemophilia A
Growth hormone	Replaces missing hormone in people of short stature
Insulin	Stimulates glucose uptake from blood in people with insulin-dependent (Type I) diabetes
Platelet-derived growth factor	Stimulates wound healing
Tissue plasminogen activator	Dissolves blood clots after heart attacks and strokes
Vaccine proteins: Hepatitis B, herpes, influenza, Lyme disease, meningitis, pertussis, etc.	Prevent and treat infectious diseases

DNA manipulation is changing agriculture

The cultivation of plants and husbanding of animals that constitute *agriculture* give us the world's oldest examples of biotechnology, dating back more than 8,000 years in human history. Over the centuries, people have adapted crops and farm animals to their needs. Through cultivation and selective breeding (artificial selection) of these organisms, desirable characteristics, such as ease of cooking the seeds or fat content of the meat, have been imparted and improved. In addition, people have developed crops with desirable growth char-



16.14 Tissue Plasminogen Activator: From Protein to Gene to Drug TPA is a naturally occurring human protein involved in dissolving blood clots. Its isolation and use as a pharmaceutical agent for treating patients suffering from blood clotting in the heart or brain—in other words, heart attacks or strokes—was made possible by recombinant DNA technology.

acteristics, such as high yield, a reliable ripening season, and resistance to diseases.

Until recently, the most common way to improve crop plants and farm animals was to select and breed varieties with desired phenotypes that existed in nature through mutational variation. The advent of genetics a century ago was followed by its application to plant and animal breeding. A crop plant or animal with desirable genes could be identified, and through deliberate crosses, those genes could be introduced into a widely used variety of that crop.

Despite some spectacular successes, such as the breeding of "supercrops" of wheat, rice, and corn, such deliberate crossing remains a hit-or-miss affair. Many desirable characters are complex in their genetics, and it is hard to predict the results of a cross or to maintain a prized combination as a pure-breeding variety year after year. In sexual reproduction, combinations of unlinked genes are quickly separated by genetic recombination. Moreover, traditional crop plant breeding takes a long time: many plants can reproduce only once or twice a year—a far cry from the rapid reproduction of bacteria or fruit flies.

Modern recombinant DNA technology has three advantages over traditional methods of breeding:

- It allows a breeder to choose specific genes, making the process more precise and less likely to fail as a result of the incorporation of unforeseen genes.
- It allows breeders to introduce any gene from any organism into a plant or animal species. This ability, combined with mutagenesis techniques, expands the range of possible new characteristics to an almost limitless horizon.
- The ability to work with cells in the laboratory and then regenerate a whole plant by cloning makes plant breeding much faster than the years needed for traditional breeding.

Biotechnology has found many applications in agriculture (Table 16.2), ranging from improving the nutritional properties of crops to using animals as gene product factories to using edible crops to make oral vaccines. We will describe a few examples here to demonstrate the approaches that have been used.

PLANTS THAT MAKE THEIR OWN INSECTICIDES. Humans are not the only species that consumes crop plants. Plants are subject to infections by viruses, bacteria, and fungi, but probably the most important crop pests are herbivorous insects. From the locusts of biblical (and modern) times to the cotton boll weevil, insects have continually eaten the crops people grow.

The development of insecticides has improved the situation somewhat, but insecticides have their own problems. Most, such as the organophosphates, are relatively nonspecific, killing not only pests in the field but beneficial insects in the ecosystem as well. Some even have toxic effects on other organisms, including people. What's more, insecticides are applied to the surface of crop plants and tend to be blown away to adjacent areas, where they may have unforeseen effects.

Some bacteria have solved their own pest problem by producing proteins that kill insect larvae that eat them. For

16.2 Agricultural Applications of Biotechnology under Development

PROBLEM	TECHNOLOGY/GENES
Improving the environmental adaptations of plants	Genes for drought tolerance, salt tolerance
Improving breeding	Male sterility for hybrid seeds
Improving nutritional traits	High-lysine seeds
Improving crops after harvest	Delay of fruit ripening; sweeter vegetables
Using plants as bioreactors	Plastics, oils, and drugs produced in plants
Controlling crop pests	Herbicide tolerance; resistance to viruses, bacteria, fungi, insects

example, there are dozens of strains of *Bacillus thuringiensis*, each of which produces a protein toxic to the insect larvae that prey on it. The toxicity of this protein is 80,000 times that of the usual commercial insecticides. When a hapless larva eats the bacteria, the toxin becomes activated, binding specifically to the insect's gut to produce holes. The insect starves to death.

Dried preparations of *B. thuringiensis* have been sold for decades as a safe, biodegradable insecticide. But biodegradation is their limitation, because it means that the dried bacteria must be applied repeatedly during the growing season. A more permanent approach would be to have the crop plants make the toxin themselves.

The toxin genes from different strains of *B. thuringiensis* have been isolated and cloned, and they have been extensively modified by the addition of plant promoters and terminators, plant poly A addition sequences, plant codon usage, and plant regulatory elements. These modified genes have been introduced into plant cells in the laboratory using the Ti plasmid vector (see Figure 16.5*c*), and transgenic plants have been grown and tested for insect resistance in the field. So far, transgenic tomato, corn, potato, and cotton crops have been shown to have considerable resistance to their insect predators.

TRANSGENIC ANIMALS EXPRESS USEFUL GENES. A transgene can be inserted into an animal, and if the appropriate promoter is present, the gene can be expressed in a readily available tissue. People with one type of emphysema have lung damage because they lack adequate amounts of a protein called α-1-antitrypsin (α-1AT). This protein inhibits elastase, an enzyme that breaks down connective tissue. Thus, using an inhibitor of elastase could alleviate these symptoms in these patients.

The problem is that only minuscule amounts of α-1AT can be purified from human serum. To overcome this problem, the gene for human α-1AT was introduced into the eggs of sheep, next to the promoter for lactoglobulin, a protein made in large amounts in milk. The resulting transgenic sheep made large amounts of α-1AT in their milk. Since female sheep produce large amounts of milk all year, this natural "bioreactor" produced a large supply of α-1AT, which was easily separated from the other components of the milk.

Goats, sheep, and cows are all being used for what has come to be called **pharming**: the production of medically useful products in milk. These products include blood clotting factors for treating hemophilia and antibodies for treating colon cancer.

CROPS THAT ARE RESISTANT TO HERBICIDES. Herbivorous insects are not the only threat to agriculture. Weeds may grow in fields and compete with crop plants for water and soil nutrients. Glyphosate ("Roundup") is a widely used and effective *herbicide*, or weed killer. It works only on plants, by inhibiting an enzyme system in the chloroplast that is involved in the synthesis of amino acids. Glyphosate is truly a "miracle herbicide," killing 76 of the world's 78 most prevalent weeds. Unfortunately, it also kills crop plants, so great care must be taken with its use. In fact, it is best used to rid a field of weeds before the crop plant starts to grow. But as any gardener knows, when the crop begins to grow, the weeds reappear. If the crop were not affected by the herbicide, the herbicide could be applied to the field at any time, and would kill only the weeds.

Fortunately, some soil bacteria have mutated to develop an enzyme that breaks down glyphosate. Scientists have isolated the gene for this enzyme, cloned it, and added plant sequences for transcription, translation, and targeting to the chloroplast. The gene has been inserted into corn, cotton, and soybean plants, making them resistant to glyphosate. This technology expanded so rapidly in the late 1990s that half of the U.S. crops of these three plants now contain this gene.

GRAINS WITH IMPROVED NUTRITIONAL CHARACTERISTICS. To remain healthy, humans must eat foods (or supplements) containing an adequate amount of β-carotene, which the body converts into vitamin A. About 400 million people worldwide suffer from vitamin A deficiency, which makes them susceptible to infections and blindness. One reason is that rice grains, which do not contain β-carotene, but only a precursor molecule for it, make up a large part of their diet. Other organisms, such as the bacterium *Erwinia* and daffodil plants, have enzymes that can convert the precursor into β-carotene. The genes for this biochemical pathway are present in the bacterial and daffodil genomes, but not in the rice genome.

16.15 Transgenic Rice Is Rich in β-Carotene The grains from a new transgenic strain of rice (left) are yellow because they make the pigment β-carotene, which is converted by humans into vitamin A. Normal rice (right) does not contain β-carotene.

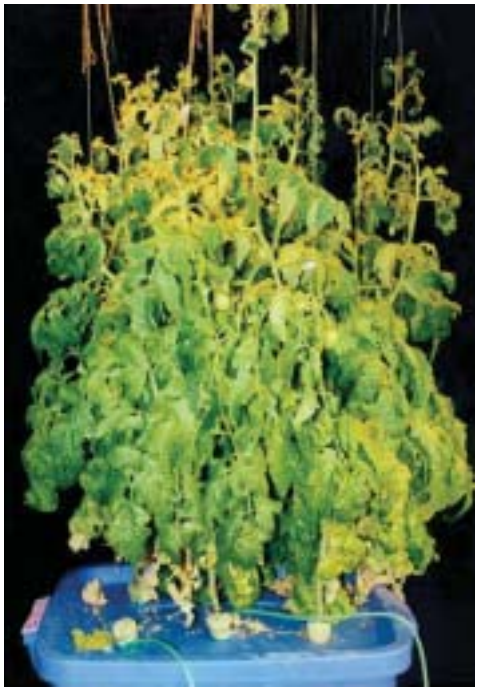
Scientists isolated two of the genes for the β-carotene pathway from the bacterium and the other two from daffodil plants. They added promoter signals for expression in the developing rice grain, and then added each gene to rice plants by using the Ti plasmid vector from *Agrobacterium tumefaciens* (see Figure 16.5*c*). The resulting rice plants produce grains that look yellow because of their high β-carotene content (Figure 16.15). About 300 grams of this cooked rice a day can supply all the β-carotene a person needs. This new transgenic strain is now being crossed with more locally adapted strains, and it is hoped that the diets of millions of people will be improved as a result.

CROPS THAT ADAPT TO THE ENVIRONMENT. Throughout human history, agriculture has involved ecological management—tailoring the environment to the needs of crop plants. A farm field is an unnatural, human-designed system, and when conditions in that field become intolerable, the crops die. The Fertile Crescent, the region between the Tigris and Euphrates rivers in the Middle East where agriculture probably originated 10,000 years ago, is no longer fertile. It is now a desert, largely because the soil has a high salt concentration. Few plants can grow on salty soils, primarily because the environment is hypertonic to the plant roots, and water leaves them, resulting in wilting.

Recently, a gene was discovered in the shale cress (*Arabidopsis thaliana*) that allows this tiny weed to thrive in salty soils. The gene codes for a protein that transports Na^+ into the vacuole. When this gene was added to tomato plants, they too grew in soils four times as salty as the normal lethal level (Figure 16.16). This finding raises the prospect of growing useful crops on what were previously unproductive soils.

More important, this example illustrates what could become a fundamental shift in the relationship between crop plants and the environment. Instead of manipulating the environment to suit the plant, biotechnology may allow us to adapt the plant to the environment. As a result, some of the negative effects of agriculture, such as water pollution, could be lessened.

(*a*)



(*b*)

16.16 Salt-Tolerant Tomato Plants Transgenic plants containing a gene for salt tolerance thrive in salty soils (*a*), while plants without the transgene die (*b*).

There is public concern about biotechnology

With the rapid expansion of genetically modified crops, concerns have been raised among the general public. Because of these concerns, some countries have banned foods that come from genetically modified crops. These concerns are centered on three claims:

- Genetic manipulation is an unnatural interference with nature.
- Genetically altered foods are unsafe to eat.
- Genetically altered crop plants are dangerous to the environment.

Advocates of biotechnology tend to agree with the first claim. However, they point out that all major crops are unnatural in the sense that they come from artificially bred plants growing in a manipulated environment (a farmer's field). Recombinant DNA technology just adds another level of sophistication to these techniques.

The concern about safety for humans is countered by the facts that only single genes are added and that these genes are specific for plant function. For example, the *B. thuringiensis* toxin produced by transgenic plants has no effect on people. However, as plant biotechnology moves from adding genes to improve plant growth to adding genes that affect human nutrition, such concerns will become more pressing.

The third concern, about environmental effects, centers on the possible "escape" of transgenes from crops to other species. If the gene for herbicide resistance, for example, were inadvertently transferred from a crop to a nearby weed, that weed could thrive in herbicide-treated areas. Or beneficial insects could eat plant materials containing *B. thuringiensis* toxin and die. Transgenic plants undergo extensive field testing before they are approved for use, but the complexity of the biological world makes it impossible to predict all potential environmental effects of transgenic organisms. Because of the potential benefits of agricultural biotechnology (see Table 16.2), scientists believe that it is wise to "proceed with caution."

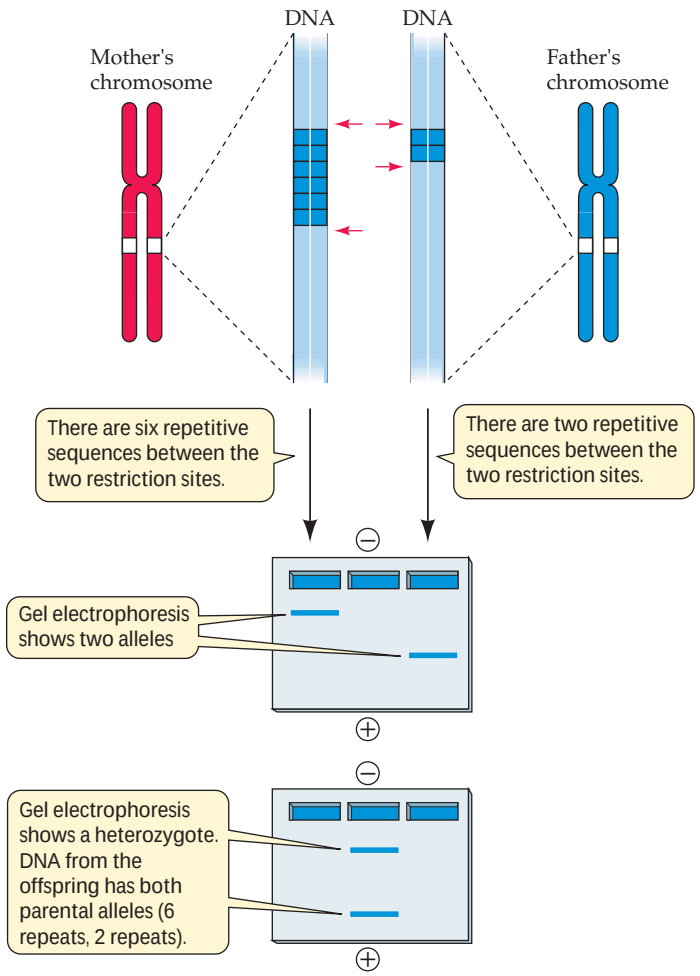
DNA fingerprinting is based on the polymerase chain reaction

"Everyone is unique." This old saying certainly applies to the human genome. Mutation and recombination through sexual reproduction ensure that each one of us (unless we have an identical twin) has a unique DNA sequence. The characterization of an individual by his or her DNA base sequence is known as **DNA fingerprinting**.

An ideal way to distinguish an individual from all the other people on Earth would be to describe his or her entire genomic DNA sequence. But the human genome contains more than 3 billion nucleotides, so this idea is clearly not practical. Instead, scientists have looked for genes that are highly *polymorphic*—that is, genes that have multiple alleles (see Chapter 23) in the human population and are therefore most likely to be different in different individuals.

One easily analyzed genetic system consists of short moderately repetitive DNA sequences that occur side by side in the chromosomes. These repeat patterns are inherited. For example, an individual might inherit a chromosome 15 with a short sequence repeated six times from her mother, and the same chromosome with the same sequence repeated two times from her father. These repeats, called **VNTRs** (*v*ariable *n*umber *t*andem *r*epeats), are easily detectable if they lie between two recognition sites for a restriction enzyme. If the DNA from this individual is cut with the restriction enzyme, it will form two different-sized fragments: one larger (the one from the mother) and the other smaller (the one from the father). These patterns are easily seen by the use of gel electrophoresis (Figure 16.17). With several different VNTRs (as many as eight are used, each with numerous alleles), an individual's unique pattern becomes apparent.

DNA fingerprinting methods require 1 μg of DNA, or the DNA content of about 100,000 human cells, but this amount is not always available. The power of the PCR technique (see Figure 11.20) permits the targeted DNA from a single cell to



16.17 DNA Fingerprinting The number of VNTRs inherited by an individual can be used to make a DNA fingerprint.

be amplified, producing in a few hours the necessary 1 μg for restriction and electrophoresis.

DNA fingerprints are used in forensics (crime investigation) to help prove the innocence or guilt of a suspect. For example, in a rape case, DNA can be extracted from semen or hair left by the attacker and compared with DNA from a suspect. So far, this method has been used to prove innocence (the DNA patterns are different) more often than guilt (the DNA patterns are the same). It is easy to exclude someone on the basis of these tests, but two people could theoretically have the same patterns, since what is being tested is just a small sample of the genome. Therefore, proof that a suspect is guilty cannot rest on DNA fingerprinting alone.

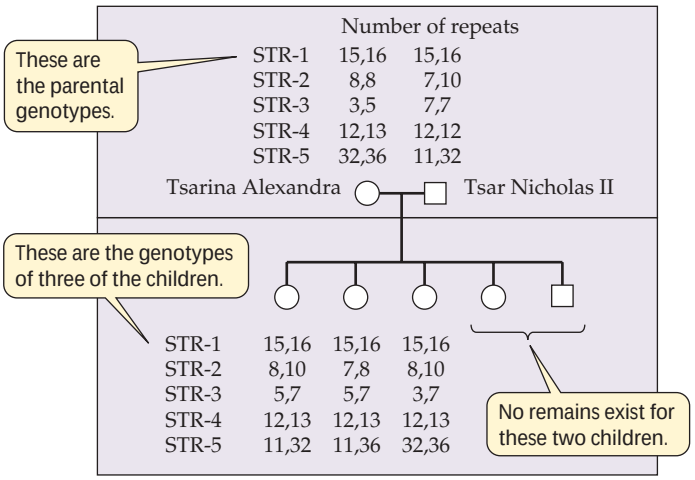
Two fascinating examples demonstrate the use of DNA fingerprinting in the analysis of historical events. Three hundred years of rule by the Romanov dynasty in Russia ended on July 16, 1918, when Tsar Nicholas II, his wife, and their five children were executed by a firing squad during the Communist revolution. A report that the bodies had been burned to ashes was never questioned until 1991, when a shallow grave with several skeletons was discovered several miles from the presumed execution site. Recent DNA fingerprinting of bone fragments found in this grave indicated that they came from an older man and woman and three female children, who were clearly related to one another (Figure 16.18) and were also related to several living descendants of the Tsar.

The other example involves Thomas Jefferson, the third president of the United States. In 1802, Jefferson was alleged to have fathered a son by his female slave, Sally Hemmings. Jefferson denied this, and his denial was accepted by many historians because of his vocal opposition to mixed-race relationships. But descendants of Hemmings's two oldest sons (the second was named Eston Jefferson) pressed their case. DNA fingerprinting was done using Y chromosome markers from descendants of these two sons as well as the president's paternal uncle (the president had no acknowledged sons). The results showed that Thomas Jefferson may have been the father of the second son, but was not the father of the first son.

In addition to such highly publicized cases, there are many other applications of PCR-based DNA fingerprinting. In 1992, the California condor was extinct in the wild. There were only 52 California condors on Earth, all cared for by the San Diego and Los Angeles zoos. Scientists made DNA fingerprints of all these birds so that geneticists at the zoos could select unrelated individuals for mating in order to increase the genetic variation, and thus the viability, of the offspring. A number of these young birds have now been returned to the wild. A similar program is under way for the threatened Galápagos tortoises.

Thousands of varieties of crops such as rice, wheat, corn, and grapes have been found in nature or produced by artificial selection. The seeds of many of these varieties are kept in cold storage in "seed banks." Samples of these plants are being fingerprinted to determine which varieties are genetically the same and which are the most diverse—information that will be useful as a guide to future breeding programs. DNA fingerprinting can also be used to identify a product from a crop: the characterization of DNA from grape varieties, for example, will allow wine makers and buyers to tell what they are purchasing. Million-dollar thoroughbred racehorses are also identified by their DNA fingerprints.

A related use of PCR is in the diagnosis of infections. In this case, the test shows whether the DNA of an infectious agent is present in a blood or tissue sample. Two primer strands matching the pathogen's DNA are added to the sam-



16.18 DNA Fingerprinting the Russian Royal Family The skeletal remains of Tsar Nicholas II, his wife Alexandra, and three of their children were found in 1991 and subjected to DNA fingerprinting. Five VNTRs were tested. The results can be interpreted as follows: Using the VNTR STR-2 as an example, the parents had genotypes 8,8 (homozygous) and 7,10 (heterozygous). The three children all inherited type 8 from the Tsarina and either type 7 or type 10 from the Tsar.

ple. If the pathogen is present, its DNA will serve as a template for the primer, and will be amplified. Because so little of the target sequence is needed, and because primers can be made to bind only to a specific viral or bacterial genome, this PCR-based test is extremely sensitive. If a pathogen is present in small amounts, PCR testing will detect it.

Finally, the isolation and characterization of genes for various human diseases, such as sickle-cell anemia and cystic fibrosis, has made PCR-based genetic testing a reality. We will discuss this subject in depth in the next chapter.

Chapter Summary

Cleaving and Rejoining DNA

▶ Knowledge of DNA transcription, translation, and replication has been used to create recombinant DNA molecules, made up of sequences from different organisms.

▶ Restriction enzymes, which are made by bacteria as a defense against viruses, bind to DNA at specific recognition sequences and cut it. **Review Figure 16.1**

▶ DNA fragments generated from cleavage by restriction enzymes can be separated by size using gel electrophoresis. The sequences of these fragments can be further identified by hybridization with a probe. **Review Figures 16.2, 16.3.** **See Web/CD Tutorial 16.1**

▶ Many restriction enzymes make staggered cuts in the two strands of DNA, creating "sticky ends" with unpaired bases. These sticky ends can be used to create recombinant DNA if DNA molecules from different species are cut with the same restriction enzyme. **Review Figure 16.4**

Getting New Genes into Cells

▶ Bacteria, yeasts, and cultured plant cells are commonly used as hosts for recombinant DNA procedures.

▶ Newly introduced DNA must be part of a replication unit if it is to be propagated in host cells. One way to make sure that the transfected DNA is part of such a unit is to insert it into a vector.

▶ There are specialized vectors for transfecting bacteria, yeasts, and plant cells. These vectors must contain an origin of replication, recognition sequences for restriction enzymes, and reporter genes to identify their presence in host cells. **Review Figure 16.5**

▶ Reporter genes conferring nutritional, antibiotic resistance, or fluorescent phenotypes can be used to identify which host cells have taken up the recombinant vector. **Review Figure 16.6**

Sources of Genes for Cloning

▶ The cutting of DNA by a restriction enzyme produces many fragments that can be individually and randomly combined with a vector and inserted into a host to create a gene library. **Review Figure 16.7**

▶ The mRNAs produced in a certain tissue at a certain time can be extracted and used to create complementary DNA (cDNA) by reverse transcription. **Review Figure 16.8**

▶ A third source of DNA is synthetic DNA made by chemists in the laboratory. The methods of organic chemistry can be used to create or mutate DNA sequences.

Some Additional Tools for DNA Manipulation

▶ Homologous recombination can be used to "knock out" a gene in an organism. **Review Figure 16.9**

▶ DNA chip technology permits the screening of thousands of sequences at the same time. **Review Figure 16.10.** **See Web/CD Tutorial 16.2**

▶ An antisense or interfering RNA complementary to a specific mRNA can prevent translation of the mRNA by hybridizing with it. **Review Figure 16.11**

▶ A two-hybrid system allows scientists to determine which proteins interact in cells. **Review Figure 16.12**

Biotechnology: Applications of DNA Manipulation

▶ Recombinant DNA techniques have made possible many new applications of biotechnology, such as the large-scale production of eukaryotic gene products.

▶ Expression vectors carry sequences such as promoters and transcription terminators that allow a gene of interest to be expressed in a host cell. **Review Figure 16.13.** **See Web/CD Activity 16.1**

▶ Recombinant DNA techniques have been used to make medically useful proteins that would otherwise have been difficult to obtain in necessary quantities. **Review Figure 16.14, Table 16.1**

▶ Because recombinant DNA technology has several advantages over traditional agricultural biotechnology, it is being extensively applied to agriculture. **Review Table 16.2**

▶ Because plant cells can be cloned to produce adult plants, the introduction of new genes into crop plants has been advancing rapidly. Transgenic crop plants can be adapted to their environment, instead of vice versa.

▶ "Pharming" uses transgenic animals that produce useful products in their milk.

▶ There is public concern about the application of recombinant DNA technology to food production.

▶ Because the DNA of an individual is unique, the polymerase chain reaction can be used to identify an organism from a small sample of its cells—that is, to create a DNA fingerprint. **Review Figures 16.17, 16.18**

Self-Quiz

1. Restriction enzymes
   a. play no role in bacteria.
   b. cleave DNA at highly specific recognition sequences.
   c. are inserted into bacteria by bacteriophages.
   d. are made only by eukaryotic cells.
   e. add methyl groups to specific DNA sequences.
2. When fragments of DNA of different sizes are placed in an electrical field,
   a. the smaller pieces migrate most quickly to the positive pole.
   b. the larger pieces migrate most quickly toward the positive pole.
   c. the smaller pieces migrate most quickly toward the negative pole.
   d. the larger pieces migrate most quickly toward the negative pole.
   e. the smaller and larger pieces migrate at the same rate.
3. From the list below, select the sequence of steps for cloning a piece of foreign DNA into a plasmid vector, introducing the plasmid into bacteria, and verifying that the plasmid and the insert are present:
   (1) Transform competent cells
   (2) Select for the lack of antibiotic resistance gene #1 function
   (3) Select for the plasmid antibiotic resistance gene #2 function

(4) Digest vector and foreign DNA with EcoR1, which inactivates antibiotic resistance gene #1
(5) Ligate the digested DNA together with the foreign DNA
a. 45132
b. 45123
c. 13425
d. 32145
e. 13254

4. Possession of which feature is *not* desirable in a vector for gene cloning?
   a. An origin of DNA replication
   b. Genetic markers for the presence of the vector
   c. Multiple recognition sites for the restriction enzyme to be used
   d. One recognition site each for one to several different restriction enzymes
   e. Genes other than the target for cloning
5. RNA interference (RNAi) inhibits
   a. DNA replication.
   b. RNA synthesis of specific genes.
   c. recognition of the promoter by RNA polymerase.
   d. transcription of all genes.
   e. translation of specific mRNAs.
6. Complementary DNA (cDNA)
   a. is produced from ribonucleoside triphosphates.
   b. is produced by reverse transcription.
   c. is the "other strand" of single-stranded DNA.
   d. requires no template for its synthesis.
   e. cannot be placed into a vector because it has the opposite base sequence of the vector DNA.
7. In a genomic library of frog DNA in *E. coli* bacteria,
   a. all bacterial cells have the same sequences of frog DNA.
   b. all bacterial cells have different sequences of DNA.
   c. each bacterial cell has a random fragment of frog DNA.
   d. each bacterial cell has many fragments of frog DNA.
   e. the frog DNA is transcribed into mRNA in the bacterial cells.
8. An expression vector requires all of the following, except
   a. genes for ribosomal RNA.
   b. a selectable genetic marker.
   c. a promoter of transcription.
   d. an origin of DNA replication.
   e. restriction enzyme recognition sites.
9. "Pharming" is a term that describes
   a. animals used in transgenic research.
   b. plants making genetically altered foods.
   c. synthesis of recombinant drugs by bacteria.
   d. large-scale production of cloned animals.
   e. synthesis of a drug by a transgenic animal in its milk.
10. In DNA fingerprinting,
    a. a positive identification can be made.
    b. a gel blot is all that is required.
    c. multiple restriction digests generate unique fragments.
    d. the polymerase chain reaction amplifies finger DNA.
    e. the variability of repeated sequences between two restriction sites is evaluated.

For Discussion

1. Compare PCR and cloning as methods to amplify a gene. What are the requirements, benefits, and drawbacks of each method?
2. As specifically as you can, outline the steps you would take to (*a*) insert and express the gene for a new, nutritious seed protein in wheat; and (*b*) insert and express a gene for a human enzyme in sheep's milk.
3. The *E. coli* plasmid pSCI carries genes for resistance to the antibiotics tetracycline and kanamycin. The tet^r gene has a single restriction site for the enzyme *Hin*dIII. Both the plasmid and the gene for glutein protein in corn are cleaved with *Hin*dIII and incubated to create recombinant DNA. The reaction mixture is then incubated with *E. coli* that are sensitive to both antibiotics. What would be the characteristics, with respect to antibiotic sensitivity or resistance, of colonies of *E. coli* containing, in addition to its own genome: (*a*) no new DNA; (*b*) native pSCI DNA; (*c*) recombinant pSCI DNA; and (*d*) corn DNA only? How would you detect these colonies?
4. Compare traditional genetics with molecular methods for producing genetically altered plants. For each case, describe: *(a)* sources of new genes; (*b*) number of genes transferred; and (*c*) how long the process takes.

14 Molecular Biology and Medicine

After his fiftieth birthday, Don's wife urged him to get a long-delayed medical checkup. He felt well, but realized that this was a good time to be screened for the various diseases that affect people as they get older. Don's routine blood count showed a surprise: Normally, people have about 5,000 white cells per milliliter of blood; Don had 40 times as many. Within a day, he was diagnosed with chronic myeloid leukemia, a serious cancer where white blood cells from the bone marrow proliferate out of control.

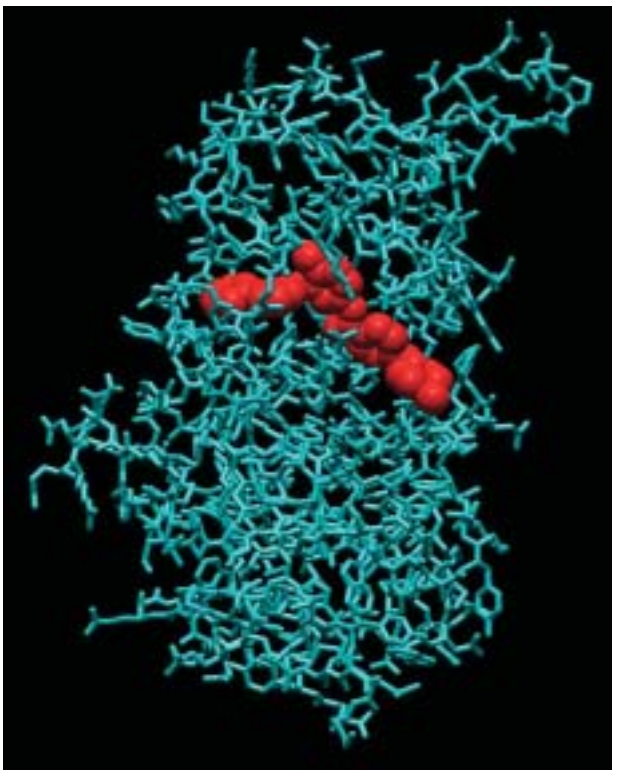
An oncologist (a physician who specializes in treating cancer) put Don on an aggressive regimen of chemotherapy. The three drugs Don took were designed to kill dividing cells—hopefully in the tumor, but such drugs also affect normal cells. One drug blocked microtubules from forming, thus preventing the mitotic spindle from assembling; another inserted into the double helix and damaged DNA replication; the third drug inhibited an enzyme involved in nucleotide synthesis.

Although the side effects were hard on Don, his white cell count gradually got lower. But after 8 months, the decline stalled at 80,000 cells/ml—still dangerously high. Worse, most of them were not mature white blood cells, but were undifferentiated cells from bone marrow. The chemotherapy drugs had killed Don's normal bone marrow cells. Without these specialized mature white blood cells, Don would die within months.

The timely development in the 1990s of a molecular understanding of this leukemia led to a therapy that saved Don's life. Scientists had known for several decades that chronic myelogenous leukemia cells have a particular chromosome translocation between chromosomes 9 and 22. With the advent of DNA sequencing, it became clear that this translocation fuses together parts of two genes: half of a gene called *bcr* on chromosome 22, and half of a gene called *abl* on chromosome 9. The abnormal protein made from this fused gene has strong protein tyrosine kinase activity.

As described in Chapter 15, protein kinases activate proteins by phosphorylation, binding a phosphate group from ATP. During the 1990s, the leukemia protein kinase was purified, crystallized, and the geometry of its binding site for ATP was described in molecular terms. Armed with this knowledge, organic chemists designed a drug that fit into and obstructed the binding site for ATP on the abnormal kinase, inactivating the enzyme. The drug, called Gleevec, was given to Don, and within a few weeks his white cell count was 5,400—normal! The cancer had been virtually wiped out.

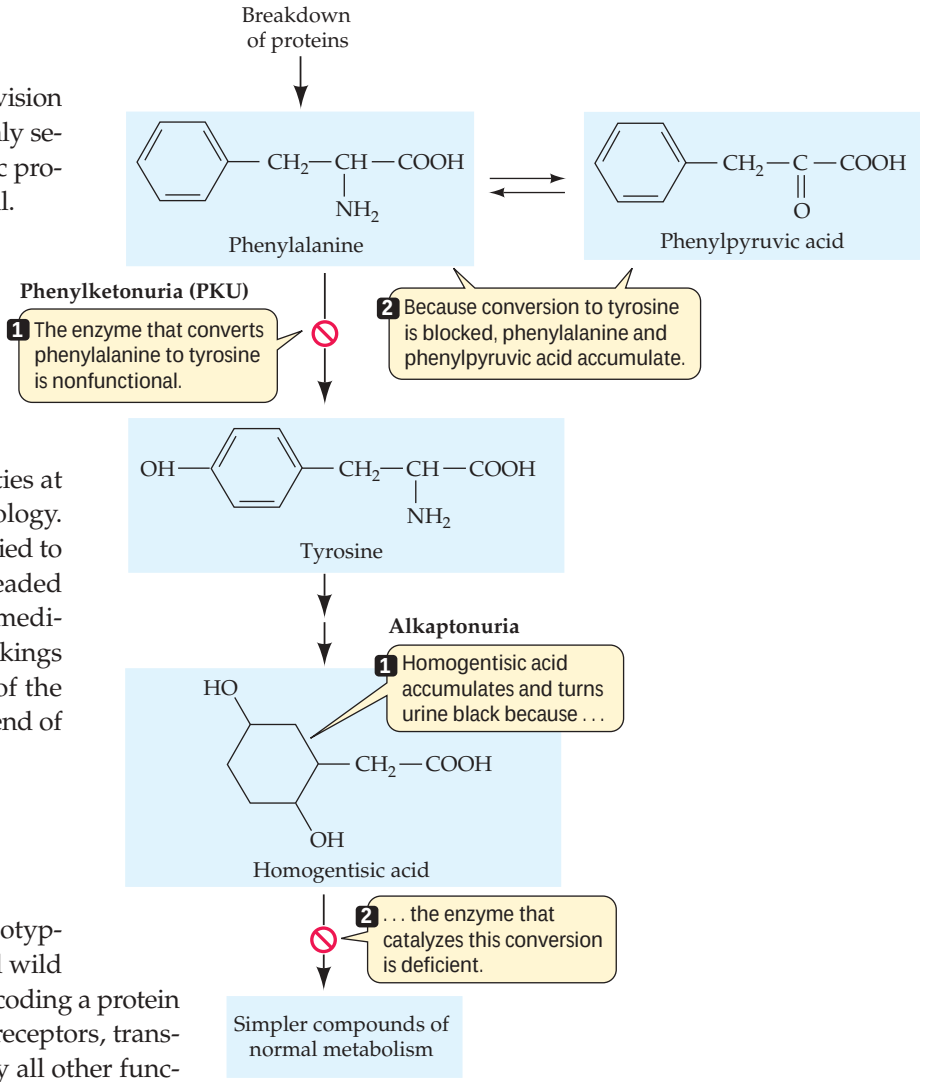
The development of Gleevec is an opening chapter in the molecular medicine of the future. Unlike conventional chemotherapy, which uses



A "Smart Drug" The drug Gleevec (shown here as the red molecule) binds to a protein kinase made by certain leukemia cells, preventing the cancer cells' reproduction. Gleevec is the first example of a rationally designed, specifically targeted cancer-fighting drug produced using the knowledge and techniques of molecular medicine.

broadly acting nonspecific drugs that stop division in all cells (cancerous or not), Gleevec is highly selective. The new drug is targeted to a specific protein that *only* occurs in a particular cancer cell.

In the first section of this chapter, we identify and discuss the kinds of abnormal proteins that can result from an abnormal allele of a gene, whether the allele is inherited or has its origin from a mutation. Then we will consider the patterns of inheritance of human genetic diseases. Precise descriptions of these genetic abnormalities at the DNA level has come from molecular biology. This knowledge has been extended and applied to the treatment of cancer, among the most dreaded of human afflictions. The rise of molecular medicine is most dramatically shown by undertakings such as gene therapy and the sequencing of the human genome, which are discussed at the end of this chapter.



17.1 One Gene, One Enzyme Both phenylketonuria and alkaptonuria are caused by abnormalities in specific enzymes in the metabolic pathway that breaks down the amino acid phenylalanine. Knowing the causes of such single-gene, single-enzyme metabolic diseases can aid in developing screening tests as well as treatments.

Abnormal or Missing Proteins: The Mutant Phenotype

Genetic mutations are often expressed phenotypically as proteins that differ from the normal wild type. In principle, a mutation in any gene encoding a protein could result in a genetic disease. Enzymes, receptors, transport proteins, structural proteins, and nearly all other functional classes of proteins have been implicated in genetic diseases.

Dysfunctional enzymes can cause diseases

In 1934, the urine of two mentally retarded young siblings was found to contain phenylpyruvic acid, an unusual byproduct of the metabolism of the amino acid phenylalanine. It was not until two decades later, however, that the complex clinical phenotype of the disease that afflicted these children, called *phenylketonuria* (PKU), was traced back to its molecular phenotype. The disease resulted from an abnormality in a single enzyme, phenylalanine hydroxylase (Figure 17.1). This enzyme normally catalyzes the conversion of dietary phenylalanine to tyrosine, but it was not active in PKU patients' livers. Lack of this conversion led to excess phenylalanine in the blood and explained the accumulation of phenylpyruvic acid. Later, the amino acid sequences of phenylalanine hydroxylase in normal people were compared with those in individuals with PKU. In many cases, the only difference in the 451 amino acids that constitute this long polypeptide chain was that instead of arginine at position 408, many people with PKU had tryptophan.

How does the molecular abnormality in PKU lead to its clinical symptoms? Since the pigment melanin is made from tyrosine, which people with PKU cannot synthesize adequately but must obtain in the diet, these people have light skin and hair color. The exact cause of the mental retardation in PKU remains elusive, but as we will see later in this chapter, it can be prevented.

Hundreds of human genetic diseases that result from enzyme abnormalities have been discovered, many of which lead to mental retardation and premature death. Most of these diseases are rare; PKU, for example, shows up in one newborn out of every 12,000. But these diseases are just the tip of the mutational iceberg. Some mutations result in amino acid changes that have no obvious clinical effects. In fact, at least 30 percent of all proteins whose sequences are known show detectable amino acid differences among individuals. The

key point here is that polymorphism does not necessarily mean disease. There can be numerous normal alleles of a gene, each producing normally functioning forms of the protein.

Abnormal hemoglobin is the cause of sickle-cell anemia

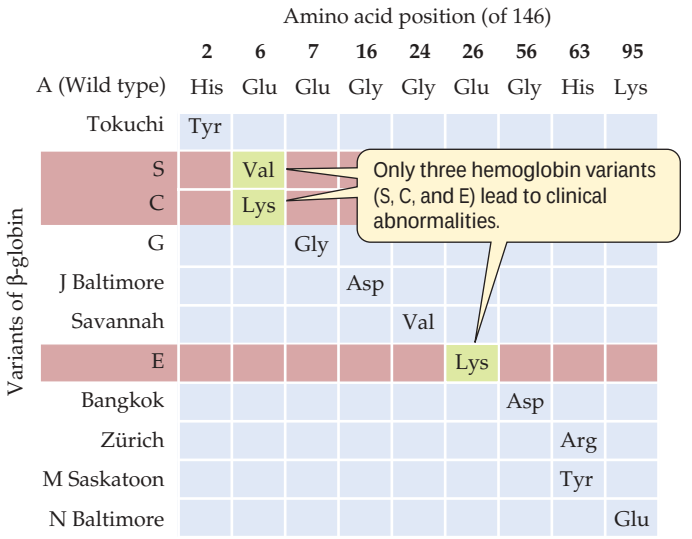
The first human genetic disease for which an amino acid abnormality was tracked down as the cause was the blood disease *sickle-cell anemia*. This disease most often afflicts people whose ancestors came from the Tropics or from the Mediterranean. About 1 in 655 African-Americans are homozygous for the sickle allele and have the disease. The abnormal allele produces an abnormal protein that leads to sickled red blood cells (see Figure 12.17). These cells tend to block narrow blood capillaries, especially when the oxygen concentration of the blood is low. The result is tissue damage and eventually death by organ failure, such as a heart attack.

Human hemoglobin is a protein with quaternary structure, containing four globin chains—two α chains and two β chains—as well as the pigment heme (see Figure 3.8). In sickle-cell anemia, one of the 146 amino acids in the β-globin chain is abnormal: At position 6, the normal glutamic acid has been replaced by valine. This replacement changes the charge of the protein (glutamic acid is negatively charged and valine is neutral), causing it to form long needle-like aggregates in the red blood cells: The result is *anemia*, a deficiency of normal red blood cells and an impaired ability of the blood to carry oxygen.

Because hemoglobin is easy to isolate and study, its variations in the human population have been extensively documented (Figure 17.2). Hundreds of single amino acid alterations in β-globin have been reported. Some of these polymorphisms are even found at the same amino acid position. For example, at the same position that is mutated in sickle-cell anemia, the normal glutamic acid may be replaced by lysine, causing hemoglobin C disease. In this case, the resulting anemia is usually not severe. Many alterations of hemoglobin have no effect on the protein's function, and thus no clinical phenotype. That is fortunate, because about 5 percent of all humans are carriers for one of these variants.

Altered membrane proteins cause many diseases

Some of the most common human genetic diseases show their primary phenotype as altered membrane receptor and transport proteins. About one person in 500 is born with *familial hypercholesterolemia* (FH), in which levels of cholesterol in the blood are several times higher than normal. The excess cholesterol can accumulate on the inner walls of blood vessels, leading to complete blockage if a blood clot forms. If a clot forms in a major vessel serving the heart, the heart becomes starved of oxygen, and a heart attack results. If a clot forms in the brain, the result is a stroke. People with FH often die of heart attacks before the age of 45.

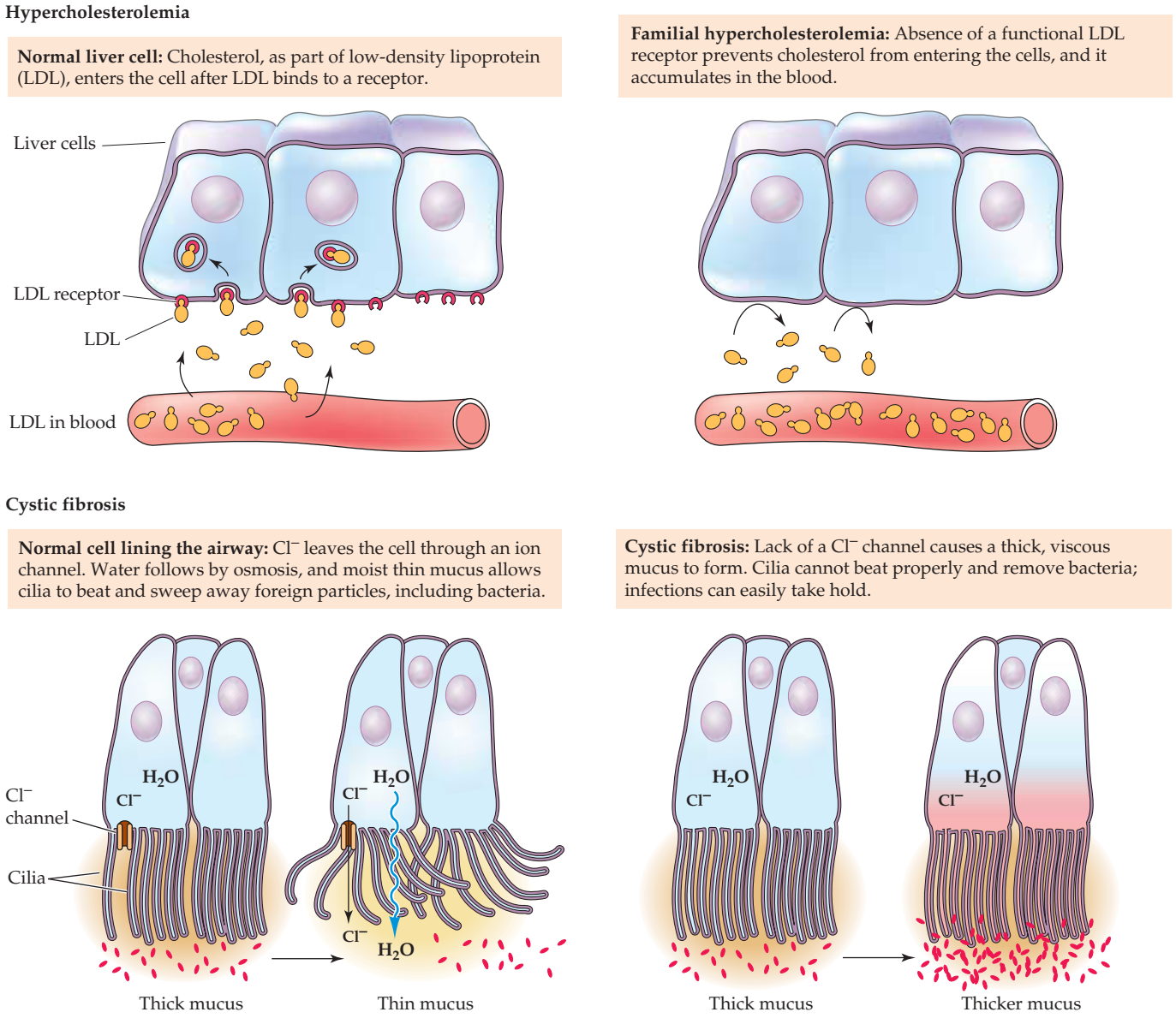


17.2 Hemoglobin Polymorphism Each of these mutant alleles changes a single amino acid in the 146-amino acid chain of β-globin. Only three of the many known variants of β-globin are known to lead to clinical abnormalities.

Unlike PKU, which is characterized by the inability to convert phenylalanine to tyrosine, the problem in FH is not an inability to convert cholesterol to other products. People with FH have all the machinery needed to metabolize cholesterol. The problem is that they are unable to transport cholesterol into the liver and other cells that use it.

Cholesterol travels through the bloodstream in protein-containing particles called *lipoproteins*. One type of lipoprotein, low-density lipoprotein, carries cholesterol to the liver cells (Figure 17.3*a*). After binding to a specific receptor on the plasma membrane of a liver cell, the lipoprotein is taken up by endocytosis and delivers its cholesterol to the interior of the cell. People with FH lack a functional version of the receptor protein. Of the 840 amino acids that make up the receptor, only one may be abnormal, but that is enough to change its structure so that it cannot bind to the lipoprotein.

Among Caucasians, about one baby in 2,500 is born with *cystic fibrosis*. The clinical phenotype of this genetic disease is an unusually thick and dry mucus that lines surface tissues such as the airways of the respiratory system and the ducts of glands. In the respiratory passageways, this thick mucus obstructs the passage of air and also prevents the cilia on the surfaces of the epithelial cells from working efficiently to clear out the bacteria and fungal spores that we take in with every breath. The results are recurrent and serious infections as well as liver, pancreatic, and digestive failures, causing malnutrition and poor growth. People with cystic fibrosis often die in their twenties or thirties.



17.3 Genetic Diseases of Membrane Proteins The left two panels illustrate normal cell function, while the two right panels show the abnormalities caused by (*a*) hypercholesterolemia and (*b*) cystic fibrosis.

The reason for the thick mucus is a defective version of a membrane transport protein, the chloride transporter (Figure 17.3*b*). In normal cells, this ion channel opens to release Cl^- to the outside of an epithelial cell. The resulting imbalance of Cl^- ions (because of the channel, there are normally more on the outside of the cell than on the inside) causes water to leave the cell by osmosis, resulting in a moist thin mucus outside the cell. A single amino acid change in the channel protein renders it nonfunctional, leading to thick mucus and the consequent clinical problems.

Altered structural proteins can cause disease

About one boy in 3,000 is born with *Duchenne muscular dystrophy*. In this genetic disease, the problem is not an enzyme or receptor, but a protein involved in biological structure. People with this disease show progressively weaker muscles and are wheelchair-bound by their teenage years. Patients usually die in their twenties, when the muscles that serve their respiratory system fail. Normal people have a protein in their skeletal muscles called *dystrophin*, which connects the actin cytoskeleton of the muscle cells to the extracellular matrix. People with Duchenne muscular dystrophy do not have a working copy of dystrophin, so their muscle cells become structurally disorganized and the muscles stop working.

Coagulation proteins are involved in the clotting of blood at a wound. In normal people, inactive coagulation proteins

are always present in the blood and become active only at a wound. People with the genetic disease *hemophilia* lack one of the coagulation proteins. Some people with this disease risk death from even minor cuts, since they cannot stop bleeding.

All of the foregoing protein alterations, and the diseases resulting from them, originate from an altered gene. But some abnormal proteins resulting in diseases may have a very different origin.

Prion diseases are disorders of protein conformation

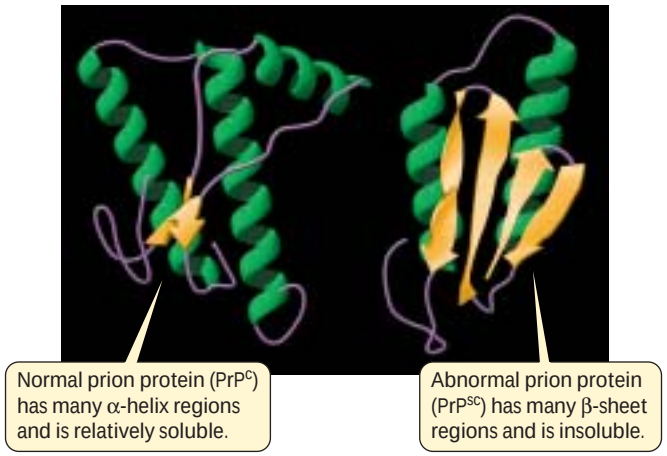
Transmissible spongiform encephalopathies (TSEs) are degenerative brain diseases that occur in many mammals, including humans. In these diseases, the brain gradually develops holes, leaving it looking like a sponge. Scrapie, a TSE that causes affected sheep and goats to rub the wool off their bodies, has been known for 250 years. In the 1980s, a TSE that appeared in cows in Britain was traced to the cows having eaten products from sheep that had scrapie. Then, in the 1990s, some people who had eaten beef from cows with this TSE got a human version of the disease (dubbed "mad cow disease" by the media), again suggesting that the causative agent could cross species lines.

Another instance of humans consuming an infective agent and getting a TSE involved kuru, a disease resulting in dementia that occurred among the Fore tribe of New Guinea. In the 1950s, it was discovered that people with kuru had consumed the brains of people who had died of it. When this ritual cannibalism stopped, so did the epidemic of kuru.

Researchers found that TSEs could be transmitted from one animal to another via brain extracts from a diseased animal. At first, a virus was suspected. But when Tikva Alper at Hammersmith Hospital, London, treated infectious extracts with high doses of ultraviolet light to inactivate nucleic acids, they still caused TSEs. She proposed that the causative agent for TSEs was a protein, not a virus. Later, Stanley Prusiner at the University of California purified the protein responsible and showed it to be free of DNA or RNA. He called it a proteinaceous infective particle, or **prion**.

Normal brain cells contain a membrane protein called PrP^c. A protein with the same amino acid sequence is present in TSE-affected brain tissues, but that protein, called PrP^{sc}, has an altered shape (Figure 17.4). Thus, TSEs are usually not caused by a mutated gene (the primary structures of the two proteins are the same), but are somehow caused by an alteration in protein conformation. The altered three-dimensional structure of the protein has profound effects on its function in the cell. PrP^{sc} is insoluble, and it piles up as fibers in brain tissue, causing cell death.

How can the exposure of a normal cell to material containing PrP^{sc} result in a TSE? The abnormal PrP^{sc} protein seems to induce a conformational change in the normal PrP^c protein so that it too becomes abnormal, just as one rotten apple results in a whole barrel full of rotten apples. Just how the conversion occurs, and how it causes a TSE, are unclear.



17.4 Prion Proteins Normal prion proteins (PrP^c, left) can be converted to the disease-causing form (PrP^{sc}, right), which has a different three-dimensional structure.

Prions appear to represent a highly unusual phenomenon in human disease. The vast majority of infectious and inherited diseases are understood in terms of proteins that are products of functional or dysfunctional genes. But the expression of these genes, like all genes, is influenced by the environment.

Most diseases are caused by both genes and environment

The human diseases for which clinical phenotypes can be traced to a single altered protein and its altered gene may number in the thousands, and in most cases they are dramatic evidence of a one-gene, one-polypeptide relationship. Taken together, these diseases have a frequency of about 1 percent in the total human population.

Far more common, however, are diseases that are **multifactorial**; that is, those that are caused by many genes and proteins interacting with one another and with the environment. Although we tend to call individuals either normal (wild-type) or abnormal (mutant), the sum total of our genes is what determines which of us who eat a high-fat diet will die of a heart attack, or which of us exposed to infectious bacteria will come down with a disease. Estimates suggest that up to 60 percent of all people are affected by diseases that are genetically influenced.

Human genetic diseases have several patterns of inheritance

As in any human genetic system, the alleles that cause genetic diseases are inherited in a dominant or recessive pattern, and are carried on autosomes or on sex chromosomes (see Chap-

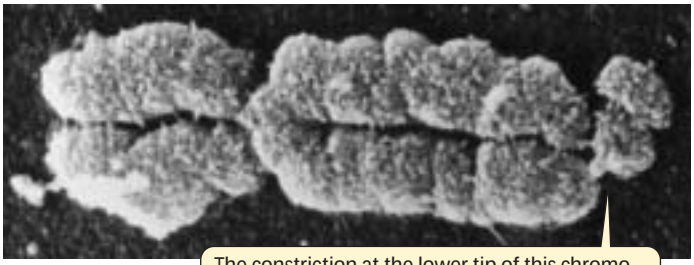
ter 10). In addition, some human diseases are caused by more extensive chromosomal abnormalities (see Chapter 9). Different inheritance patterns can be seen when genetic diseases are followed over several human generations.

AUTOSOMAL RECESSIVE PATTERN. PKU, sickle-cell anemia, and cystic fibrosis are all caused by autosomal recessive mutant alleles. Typically, both parents of an affected person are carriers (normal phenotype, heterozygous genotype). Each time they conceive a child, the parents have a 25 percent (one in four) chance of having an affected (homozygous) son or daughter. Because of this low probability and the fact that many families in Western societies now have fewer than four children, it is unusual for more than one child in a family to have an autosomal recessive disease.

In the cells of a person who is homozygous for an autosomal recessive mutant allele, only the nonfunctional, mutant version of the protein it encodes is made. Thus a biochemical pathway or important cell function is disrupted, and disease results. As expected, heterozygotes, with one normal and one mutant allele, often have 50 percent of the normal level of functional protein. For example, people who are heterozygous for the PKU allele have half as many active molecules of phenylalanine hydroxylase in their liver cells as individuals who carry two normal alleles for this enzyme. But by one mechanism or another, this 50 percent suffices for relatively normal cellular function.

AUTOSOMAL DOMINANT PATTERN. Familial hypercholesterolemia is caused by an abnormal autosomal dominant allele. In this case, the presence of only one mutant allele is enough to produce the clinical phenotype. In people who are heterozygous for familial hypercholesterolemia, having half the normal number of functional receptors for low-density lipoprotein on the surface of liver cells is simply not enough to clear cholesterol from the blood. In autosomal dominance, direct transmission from an affected parent to offspring is the rule.

X-LINKED RECESSIVE PATTERN. Hemophilia is an X-linked recessive condition; that is, the gene locus responsible is on the X chromosome. Thus, a son who inherits a mutant allele on the X chromosome from his mother will have the disease, because his Y chromosome does not contain a normal allele. However, a daughter who inherits one mutant allele will be an unaffected heterozygous carrier, since she has two X chromosomes, and hence two alleles. Because, until recently, few males with these diseases lived to reproduce, the most common pattern of inheritance has been from carrier mother to son, and all rare X-linked diseases are much more common in males than in females.



17.5 A Fragile-X Chromosome at Metaphase The chromosomal abnormality associated with mental retardation that characterizes fragile-X syndrome shows up under the microscope as a constriction in the chromosome.

CHROMOSOMAL ABNORMALITIES. Chromosomal abnormalities also cause human diseases. Such abnormalities include an excess or loss of one or more chromosomes (aneuploidy), loss of a piece of a chromosome (deletions), and the transfer of a piece of one chromosome to another chromosome (translocations). About one newborn in 200 is born with a chromosomal abnormality. While some of these abnormalities are inherited, many are the result of meiotic events such as nondisjunction (see Chapter 9).

Many zygotes that have chromosomal abnormalities do not survive development and are spontaneously aborted. Of the 20 percent of pregnancies that are spontaneously aborted during the first 3 months of human development, an estimated half of them have chromosomal abnormalities. For example, few human zygotes with only one X chromosome and no Y survive beyond the fourth month of pregnancy.

One common cause of mental retardation is *fragile-X syndrome* (Figure 17.5). About one male in 1,500 and one female in 2,000 are affected. These people have a constriction near the tip of the X chromosome that tends to break during preparation for microscopy, giving the name for this syndrome. Although the basic pattern of inheritance is that of an X-linked recessive trait, there are departures from this pattern. Not all people with the fragile-X chromosomal abnormality are mentally retarded, and we will describe the reason for this variation later in the chapter.

Mutations and Human Diseases

The isolation and description of human mutations has proceeded rapidly since modern molecular biological techniques were developed (see Chapter 16). When the primary phenotype was known, as in the case of abnormal hemoglobins, cloning the gene responsible was straightforward, although time-consuming. In other cases, such as Duchenne muscular dystrophy, a chromosome deletion associated with the dis-

ease in a patient pointed the way to the missing gene. In still other cases, such as cystic fibrosis, only a subtle molecular marker was available to lead investigators to the gene. In both of the latter examples, the primary phenotype—the defective protein—was unknown; only when the gene was isolated was the protein found.

In the discussions that follow, we will examine how mRNA, chromosome deletions, and genetic markers can be used to identify both mutant genes and abnormal proteins involved in genetic diseases. We will close this discussion by considering the role of expanding triplet repeats in genetic diseases such as fragile-X syndrome.

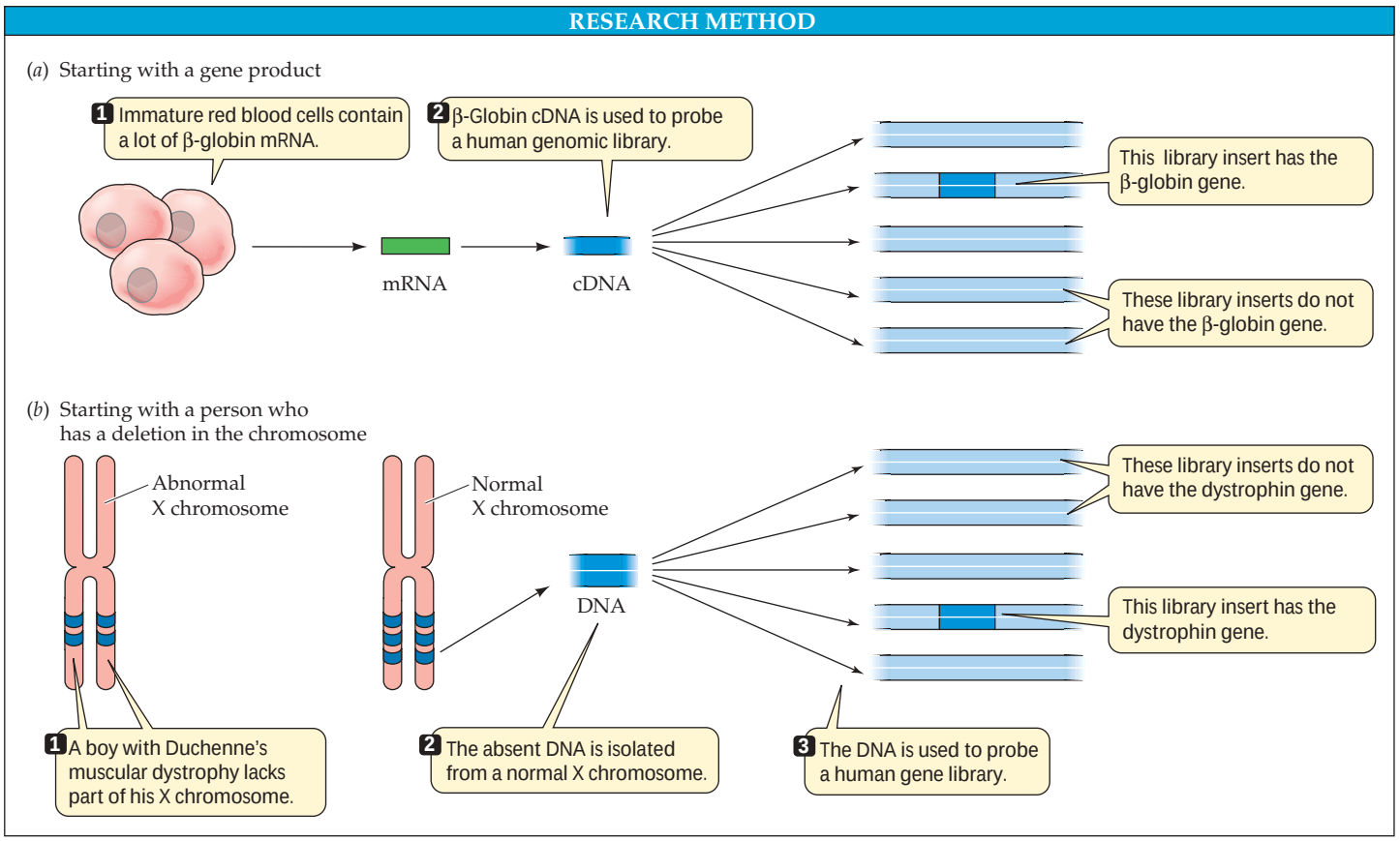
One way to identify a gene is to start with its protein

The primary phenotype for sickle-cell anemia was described in the 1950s as a single amino acid change in β-globin. On the basis of the clinical picture of sickled red blood cells, β-globin was certainly the right protein to examine. By the 1970s, researchers were able to isolate β-globin mRNA from immature red blood cells, which transcribe the globins as their major gene product. A cDNA copy of this mRNA was made and used to probe a human DNA library to find the β-globin gene (Figure 17.6*a*). DNA sequencing was then used to compare the normal gene with the gene from people with sickle-cell anemia. As previously described, it was found that a point mutation had changed only one base pair in the entire β-globin gene.

Chromosome deletions can lead to gene and then protein isolation

The inheritance pattern of Duchenne muscular dystrophy is consistent with an X-linked recessive trait. But until the late 1980s, neither the abnormal protein involved nor the gene encoding it had been described. This failure was not from lack of effort: Almost every known muscle protein had been tested without success. Then several boys with the disease were found to have a small deletion in their X chromosome. Comparison of the affected X chromosomes with normal ones made possible the isolation of the gene that was missing in the boys (Figure 17.6*b*).

17.6 Strategies for Isolating Human Genes (*a*) Once the sequence for the normal β-globin gene was established by cloning from the isolated mRNA, it could be compared to the gene sequence in people with sickle-cell anemia. (*b*) When an abnormality is caused by a missing sequence, as in Duchenne muscular dystrophy, researchers can compare the affected chromosome with a normal chromosome and isolate the DNA that is missing, then determine the protein this DNA encodes.



Genetic markers can point the way to important genes

In cases in which no candidate protein nor visible chromosome deletion is available, a technique called **positional cloning** has been invaluable. To understand this method, imagine an astronaut looking down from space, trying to find her son on a park bench on Chicago's North Shore. The astronaut picks out reference points—landmarks that will lead her to the park. She recognizes the shape of North America, then moves to Lake Michigan, the Sears Tower, and so on. Once she has zeroed in on the North Shore park, she can use advanced optical instruments to find her son.

The reference points for positional cloning are genetic markers on the DNA. These markers can be located anywhere in the DNA. The only requirement is that they be polymorphic (have more than one allele).

As we described in Chapter 16, restriction enzymes cut DNA molecules at specific recognition sequences. On a particular human chromosome, a given restriction enzyme may make hundreds of cuts, producing many DNA fragments. The enzyme EcoRI, for example, cuts DNA at

5′… GAATTC … 3′

Suppose this recognition sequence exists in a certain stretch of human chromosome 7. The restriction enzyme will cut this stretch once and make two fragments of DNA. Now suppose that, in some people, this sequence is mutated as follows:

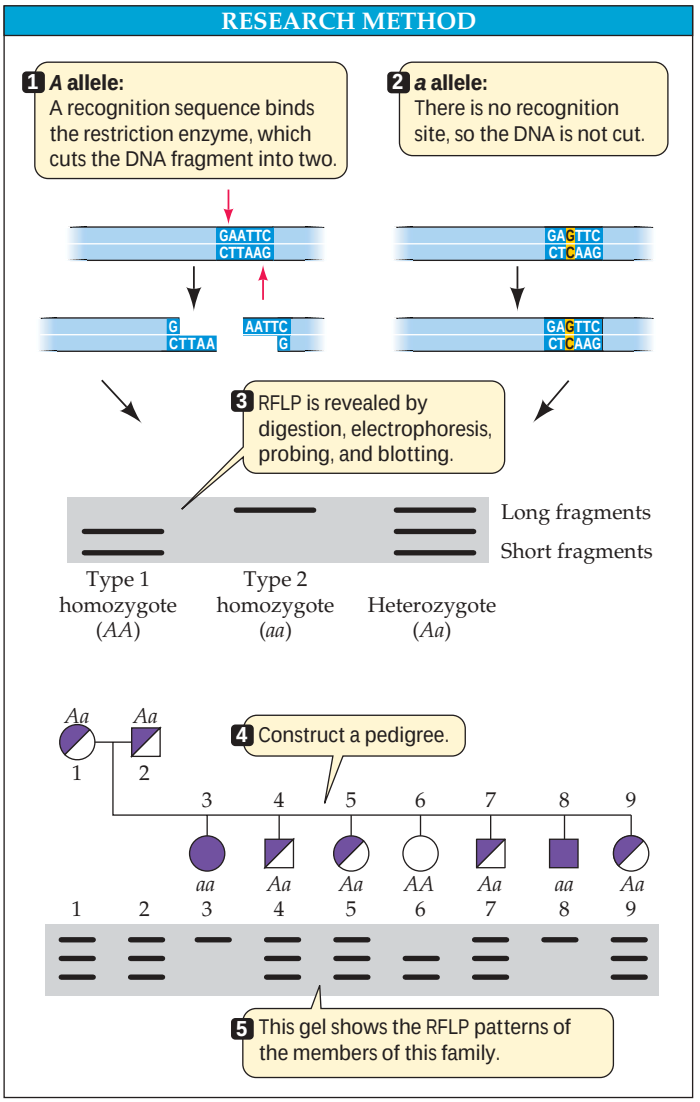
5′… GAGTTC … 3′

This sequence will not be recognized by the enzyme; thus it will remain intact and yield one larger fragment of DNA.

Such DNA differences are called **restriction fragment length polymorphisms**, or **RFLPs** (Figure 17.7). They can be easily seen as bands on an electrophoresis gel. An RFLP band pattern is inherited in a Mendelian fashion and can be followed through a pedigree. Thousands of such markers have been described for the human genome.

Genetic markers such as RFLPs can be used as landmarks to find genes of interest if they, too, are polymorphic. The key to this method is the well-known observation that if two genes are located near each other on the same chromosome, they are usually passed on together from parent to offspring. The same holds true for any pair of genetic markers.

To narrow down the location of a gene, a scientist must find a marker and a gene that are always inherited together. To do this, family medical histories are taken and pedigrees are constructed. If a genetic marker and a genetic disease are inherited together, then they must be near each other on the same chromosome. Unfortunately, "near each other" might be as much as several million base pairs apart. The process of locating the gene is thus similar to the astronaut focusing on Chicago: the first landmarks lead to only an approximate location.



17.7 RFLP Mapping Restriction fragment length polymorphisms are differences in DNA sequences that serve as genetic markers. Thousands of such markers have been described for the human genome.

How can the gene be isolated? Many sophisticated methods are available for narrowing the search. For example, the neighborhood around the RFLP can be screened for further RFLPs involving other restriction enzymes. With luck, one of them might be more closely linked to the disease-causing gene. Once a relatively short DNA sequence (several hundred thousand bases) thought to contain the gene is pinpointed, it can be cut into fragments, and those fragments, when denatured, can be used to probe cellular mRNA. If one of the fragments hybridizes with the mRNA, it means that the fragment is part of a gene that is expressed as mRNA. The candidate gene is then sequenced from normal people and from people who have the disease in question. If appropriate mutations are found, the gene of interest has been isolated.

The isolation of genes responsible for genetic diseases has led to spectacular advances in the understanding of human

biology. Before the genes, and then the proteins, for Duchenne muscular dystrophy and for cystic fibrosis were isolated, dystrophin and the chloride transporter had never been described. Thus, the identification of mutant genes has opened up new vistas in our understanding of how the human body works.

Human gene mutations come in many sizes

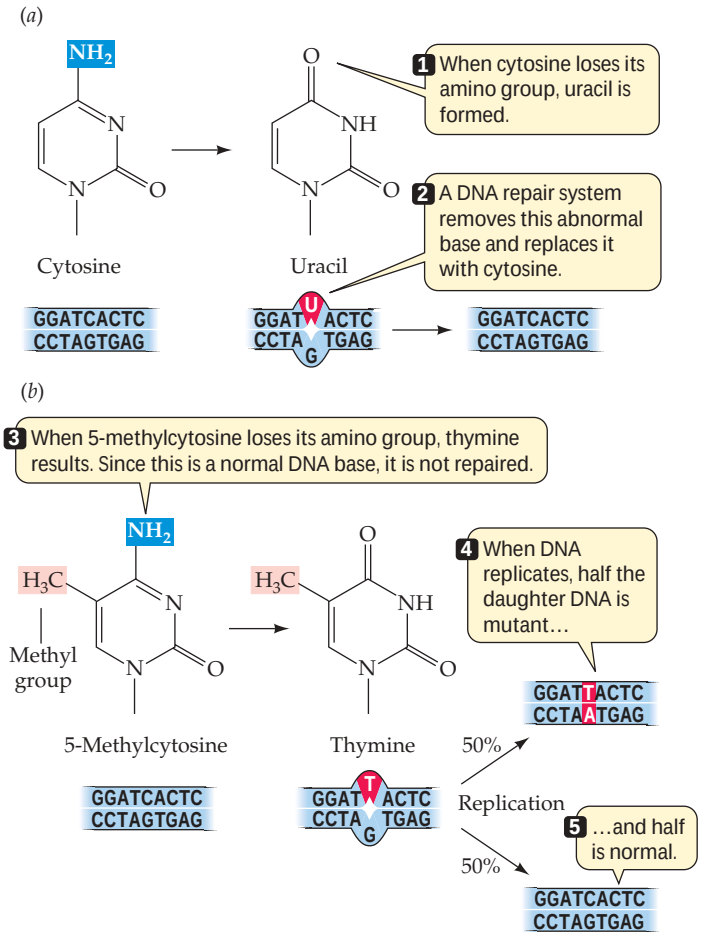
As we saw in Chapter 12, mutations come in many sizes, from changes in a single base pair to changes in entire chromosomes. As we have seen, sickle-cell anemia is caused by a point mutation. Some variants of the β-globin gene cause disease, but most do not (see Figure 17.2). Those point mutations that alter a protein's function usually affect its three-dimensional structure; for example, such a mutation might alter the shape at the active site of an enzyme.

Some mutations lead to a greatly shortened protein chain with total loss of its function. For example, some people with cystic fibrosis have a nonsense mutation such that a codon for an amino acid near the beginning of the long chloride transporter protein chain has been changed to a stop codon. If this happens, only a very short, nonfunctional peptide is made. As we noted in Chapter 12, other point mutations affect RNA processing, leading to nonfunctional mRNA and no protein synthesis.

DNA sequencing has revealed that mutations occur most often at certain base pairs. These "hot spots" are often located where cytosine residues have been methylated to 5-methylcytosine (see Chapter 14). This phenomenon is a result of the natural instability of the bases in DNA. Either spontaneously or with chemical prodding unmethylated cytosine residues can lose their amino group and form uracil (Figure 17.8*a*). But the cell nucleus has a repair system that recognizes this uracil as being inappropriate for DNA (after all, uracil occurs only in RNA). It removes the uracil and replaces with cytosine.

The fate of 5-methylcytosine that loses its amino group is rather different, since the result of that loss is thymine, a natural base for DNA. The uracil repair system ignores this thymine (Figure 17.8*b*). However, since the GC pair is now a mismatched GT pair, a different type of repair system comes in and tries to fix the mismatch. Half the time, the mismatch repair system matches a new C to the G, but the other half of the time, it matches a new A to the T, resulting in a mutation.

Larger mutations may involve many base pairs of DNA. For example, some of the deletions in the X chromosome that result in Duchenne muscular dystrophy cover only part of the dystrophin gene, leading to an incomplete protein and a mild form of the disease. Others cover all of the gene, and thus the protein is missing entirely from muscle, resulting in the severe form of the disease. Still other deletions involve millions of base pairs and cover not only the dystrophin gene



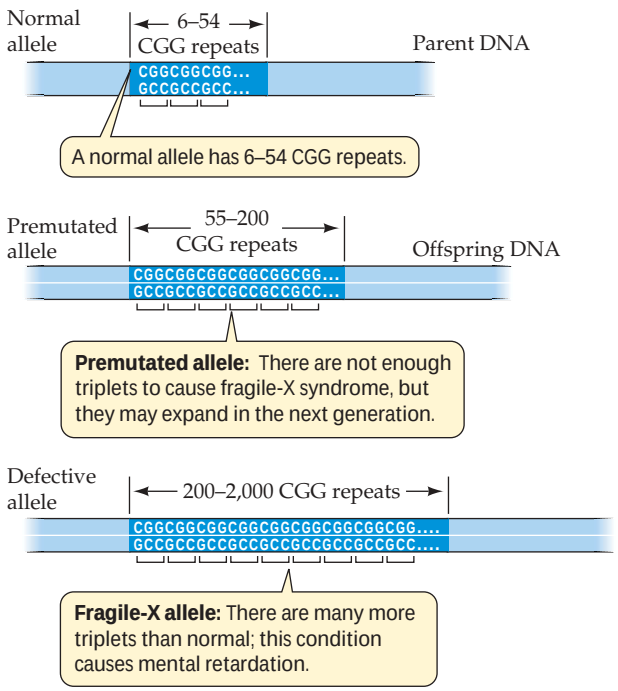
17.8 5-Methylcytosine in DNA Is a "Hot Spot" for Mutagenesis
(*a*) Cytosine can lose an amino group either spontaneously or because of exposure to certain chemical mutagens. Such mutations are usually repaired. (*b*) If the cytosine residue has been methylated to 5-methylcytosine, however, the mutation is unlikely to be repaired.

but adjacent genes as well; the result may be several diseases simultaneously.

Expanding triplet repeats demonstrate the fragility of some human genes

About one-fifth of all males that have the fragile-X chromosomal abnormality are phenotypically normal, as are most of their daughters. But many of those daughters' sons are mentally retarded. In a family in which the fragile-X syndrome appears, later generations tend to show earlier onset and more severe symptoms of the disease. It is almost as if the abnormal allele itself is changing—and getting worse. And that's exactly what is happening.

The gene responsible for fragile-X syndrome (*FMR1*) contains a repeated triplet, CGG, at a certain point in the promoter region. In normal people, this triplet is repeated 6 to 54 times (average: 29). In the alleles of mentally retarded peo-



17.9 The CGG Repeat in the Fragile-X Gene Expands with Each Generation The genetic defect in fragile-X syndrome is caused by 200 or more repeats of the CGG triplet.

ple with fragile-X syndrome, the CGG sequence is repeated 200 to 2,000 times.

The "premutated" males that show no symptoms, but have affected descendants, have 52 to 200 repeats. These repeats become more numerous as the daughters of these men pass the chromosome on to their children (Figure 17.9). Expansion to more than 200 repeats leads to increased methylation of the cytosines in the CGG units, accompanied by transcriptional inactivation of the *FMR1* gene. The normal role of the protein product of this gene is to bind to mRNAs involved in nerve cell function and regulate their translation at the ribosome. When the FMR1 protein is not made in adequate amounts, these mRNAs are not properly translated, and nerve cells die. Their loss often results in mental retardation.

Such **expanding triplet repeats** have been found in over a dozen other diseases, such as myotonic dystrophy (involving repeated CTG triplets) and Huntington's disease (in which CAG is repeated). Many benign genes also appear to have these repeats, which may be found within a protein-coding region or outside it. How the repeats expand is not known, but DNA polymerase may slip after copying a repeat and then fall back to copy it again. In all previously known classes of human mutations, the mutation is just as stable as the normal allele, but expanding triplet repeats represent an entirely new class of unstable mutant alleles.

Genomic imprinting shows that mammals need both a mother and a father

Just after fertilization in a mammalian egg, there are two haploid *pronuclei*—one from the sperm and the other from the egg—in the zygote. The two pronuclei can be distinguished from each other, and they can be carefully removed with a micropipette and placed in other eggs. So it is possible to make mouse zygotes in the laboratory with two male or two female pronuclei. These diploid cells should go on to develop into mice—but they don't. Invariably, if the two sets of chromosomes come from only one sex, development begins, but is quickly aborted. The same thing happens in those rare instances when this occurs in humans—for instance, if two sperm enter an empty egg. Again, a fetus never develops.

In addition to showing the obvious need for two sexes, these observations raise the possibility that the male and female genomes are not functionally equivalent. In fact, there are groups of genes that differ in their phenotypic effects depending on which parent they came from. This phenomenon is called **genomic imprinting**.

A dramatic example of genomic imprinting is the inheritance and phenotypic pattern of a certain small deletion on human chromosome 15.

- If the deletion is on the mother's chromosome 15, the result is a thin child with a wide mouth and prominent jaw (*Angelman syndrome*).
- If the deletion is on the father's chromosome 15, the child is short and obese, with small hands and feet (*Prader-Willi syndrome*).

The remaining functional alleles in this region of chromosome 15 must be imprinted in very different ways in the two sexes to result in such different phenotypes. How this happens is not clear.

Detecting Genetic Variations: Screening for Human Diseases

The determination of the precise molecular phenotypes and genotypes of human genetic diseases has given us the ability to diagnose these diseases even before symptoms first appear. **Genetic screening** is the use of a test to identify people who have, are predisposed to, or are carriers of a certain genetic disease. It can be applied at many times of life and used for many purposes.

- *Prenatal* screening can identify an embryo or fetus with a disease so that medical intervention can be applied or decisions about continuing the pregnancy can be made.
- *Newborn* babies can be screened so that proper medical intervention can be initiated quickly for those babies who need it.

- *Asymptomatic* people who have a relative with a genetic disease can be screened to determine whether they are carriers of the disease or are likely to develop the disease themselves.

The existence of genetic screening techniques poses ethical questions concerning the uses of the information they provide, as we will see later in the chapter.

Screening for abnormal phenotypes can make use of protein expression

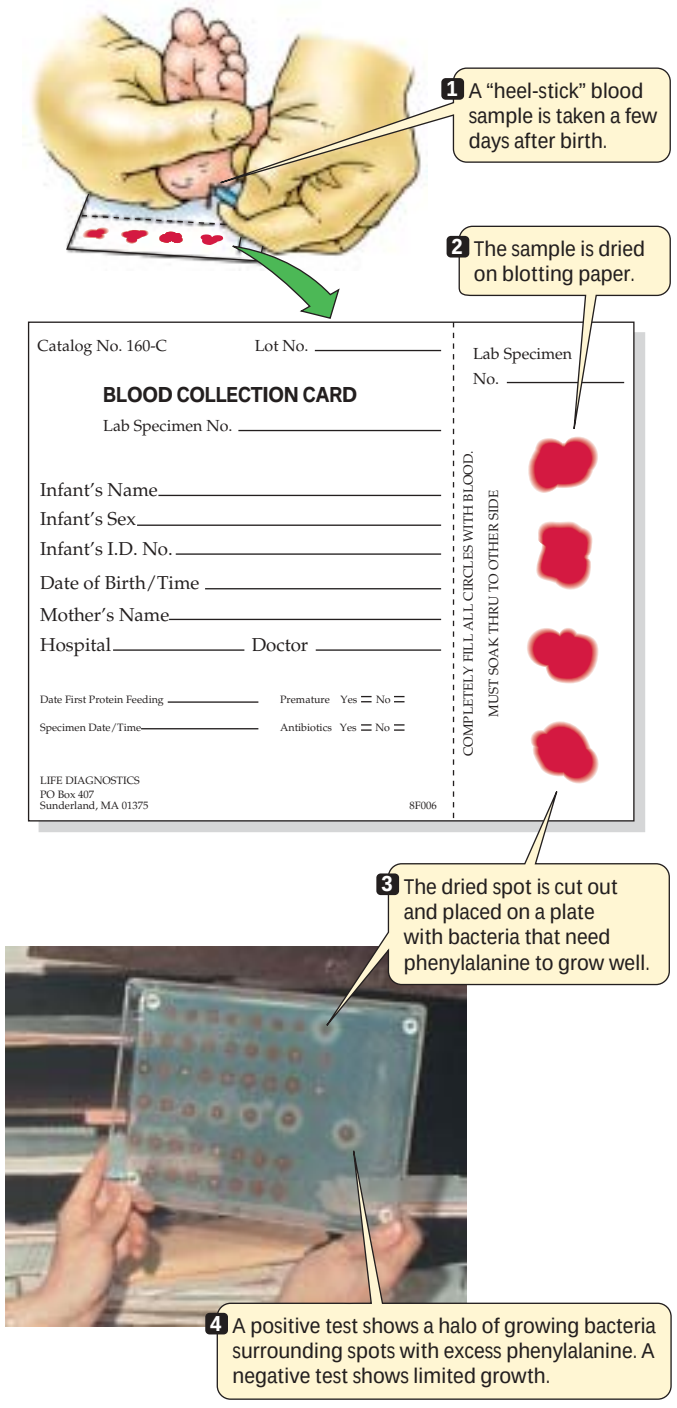
Screening of newborns for phenylketonuria can free a child from developing mental retardation. Such screening is legally mandatory in many countries, including all of the United States and Canada. Babies who are homozygous for this genetic disease are born with a normal phenotype because excess phenylalanine in their blood before birth diffuses across the placenta to the mother's circulation. Since the mother is almost always heterozygous, and therefore has adequate phenylalanine hydroxylase activity, her body metabolizes the excess phenylalanine from the fetus. Thus, before birth, the baby's blood does not accumulate abnormal levels of phenylalanine.

After birth, however, the situation changes. The baby begins to consume protein-rich food (milk) and to break down some of its own proteins. Phenylalanine enters the baby's blood, and without the mother's phenylalanine hydroxylase to break it down, accumulates there. After a few days, the phenylalanine level in the baby's blood may be ten times higher than normal. Within days, the developing brain is damaged, and untreated children with PKU become severely mentally retarded.

If PKU is detected early, it can be treated with a special diet low in phenylalanine and the brain damage avoided. Thus, early detection is imperative. In 1963, Robert Guthrie described a simple screening test for PKU in newborns that today is used almost universally (Figure 17.10). This elegant method uses auxotrophic bacteria to detect the presence of an amino acid—phenylalanine—in the blood. The test can be automated so that a screening laboratory can process many samples in a day.*

If an infant tests positive for PKU in this screening, he or she must be retested using a more accurate chemical assay for phenylalanine. If that test also shows a high phenylalanine level in the blood, dietary intervention is begun. The whole process must be completed by the end of the second week of life. Since the screening test is inexpensive (about a dollar per test), and since babies with PKU who receive early medical intervention develop practically normally, the benefit of screening is significant.

*Guthrie refused to patent the screening test he developed, or to accept any royalties or payment for it. Its immediate and widespread use was at least in part a result of his generosity in allowing the test to be available to all hospitals at low cost.



17.10 Genetic Screening of Newborns for Phenylketonuria
A simple test devised by Robert Guthrie in 1963 is used today to screen newborns for phenylketonuria. Early detection means that the symptoms of the condition can be prevented by following a therapeutic diet.

Several screening methods can find abnormal genes

The blood level of phenylalanine is an indirect measure of phenylalanine hydroxylase activity in the liver. But how can we screen for genetic diseases that are expressed only in a

certain tissue, such as the liver or brain, and are not reflected in the blood? What if blood is difficult to test, as it is in a fetus? Finally, since tissues in heterozygotes often compensate for having just one functional gene by raising the activity of the remaining proteins to near normal levels, how can heterozygotes be identified?

These problems can be overcome by **DNA testing**, which is the most direct and accurate way to test for an abnormal gene. With the molecular description of the genetic mutations responsible for human diseases, it has become possible to examine directly any cell in the body at any time during the life span for mutations. However, these methods work best for diseases caused by only one or a few different mutations.

The polymerase chain reaction (PCR) technique allows testing of the DNA from even a single cell. You will recall from Chapter 11 that PCR amplifies a target sequence of DNA millions of times in the test tube. Consider, for example, two parents who are both heterozygous for the cystic fibrosis allele, have had a child with the disease, and want a normal child. If the mother is treated with the appropriate hormones, she can be induced to "superovulate," releasing several eggs. One of the eggs can be injected with a single sperm from her husband and the resulting zygote allowed to divide to the 8-cell stage. If one of these embryonic cells is removed, it can be tested for the presence of the cystic fibrosis allele(s). The remaining 7-cell embryo can be implanted in the mother's womb and go on to develop normally.

Such *preimplantation screening* is performed only rarely. More typical are analyses of fetal cells after implantation in the womb. Fetal cells can be analyzed at about the tenth week of pregnancy by chorionic villus sampling or during the thirteenth to seventeenth weeks by amniocentesis. These two sampling methods are described in Chapter 43. In either case, only a few fetal cells are required.

Newborns can also be screened for genetic mutations. The blood samples used for screening for PKU and other disorders contain enough of the baby's blood cells to permit extraction of the DNA, its amplification by PCR, and testing. Pilot studies of screening methods for sickle-cell anemia and cystic fibrosis are under way, and other genes will surely follow.

DNA testing is also widely used to test adults for heterozygosity. For example, a sister or female cousin of a boy with Duchenne muscular dystrophy can determine whether she is a carrier of the X chromosome deletion that results in the disease.

Of the numerous methods of DNA testing, two are the most widespread. We will describe their use to detect the mutation in the β-globin gene that results in sickle-cell anemia.

SCREENING FOR ALLELE-SPECIFIC CLEAVAGE DIFFERENCES. There is a difference between the normal and sickle alleles of the β-globin gene with respect to a restriction enzyme recognition sequence. Around codon position 6 in the normal gene is the sequence

5′… CCTGAGGAG… 3′

This sequence is recognized by the restriction enzyme *Mst*II, which will cleave DNA at

5′… CCT*N*AGG… 3′

where *N* is any base.

In the sickle allele, the DNA sequence is changed to

5′… CCTGTGGAG… 3′

The point mutation makes this sequence unrecognizable by *Mst*II. When *Mst*II fails to make the cut in the mutant gene, gel electrophoresis detects a larger DNA fragment (Figure 17.11).

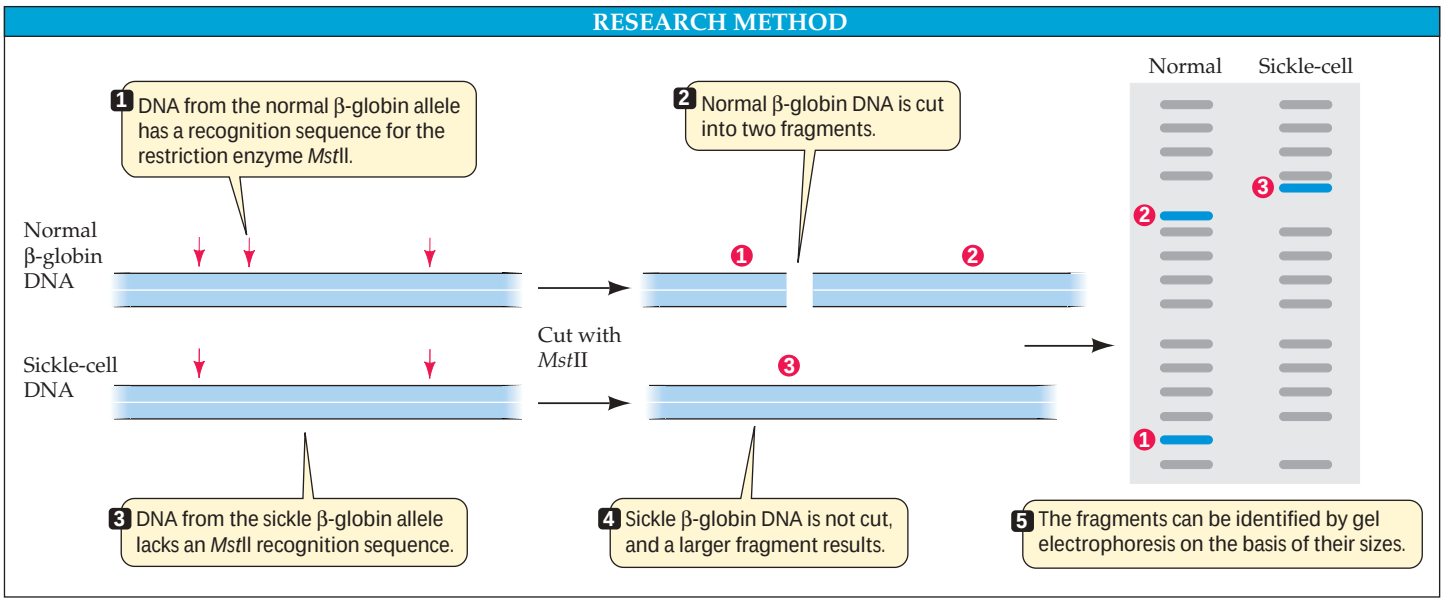
This *allele-specific cleavage* method of DNA testing is similar to the use of RFLPs (see Figure 17.7). It works only if a restriction enzyme exists that can recognize either the sequence at the mutation or the original sequence that is altered by that mutation.

SCREENING BY ALLELE-SPECIFIC OLIGONUCLEOTIDE HYBRIDIZATION. The *allele-specific oligonucleotide hybridization* method uses short DNA strands called oligonucleotides made in the laboratory that will hybridize either with the denatured normal β-globin DNA sequence around position 6 or with the sickle mutant sequence. Usually, an oligonucleotide probe of at least a dozen bases is needed to form a stable double helix with the target DNA. If the probe is labeled with radioactivity or with a colored or fluorescent substrate, hybridization can be readily detected (Figure 17.12). This method is easier and faster than allele-specific cleavage, and will work no matter what the sequence of the normal or mutant allele.

Cancer: A Disease of Genetic Changes

Perhaps no malady affecting people in the industrialized world instills more fear than cancer. One in three Americans will have some form of cancer in their lifetime, and at present, one in four will die of it. With a million new cases and half a million deaths in the United States annually, cancer ranks second only to heart disease as a killer. Cancer was less common a century ago; then, as now in many regions of the world, people died of infectious diseases and did not live long enough to get cancer. Cancer tends to be a disease of the later years of life; children are much less frequently afflicted.

Since the U.S. government declared a "war on cancer" in 1970, a tremendous amount of information on cancer cells—on their growth and spread and on their molecular changes—has been obtained. Perhaps the most remarkable discovery is that cancer is a disease caused primarily by



17.11 DNA Testing by Allele-Specific Cleavage Allele-specific cleavage, a technique similar to RFLP analysis, can be used to detect mutations such as the one that causes sickle-cell anemia.

genetic changes. These changes are mostly alterations in the DNA of somatic cells that are propagated by mitosis.

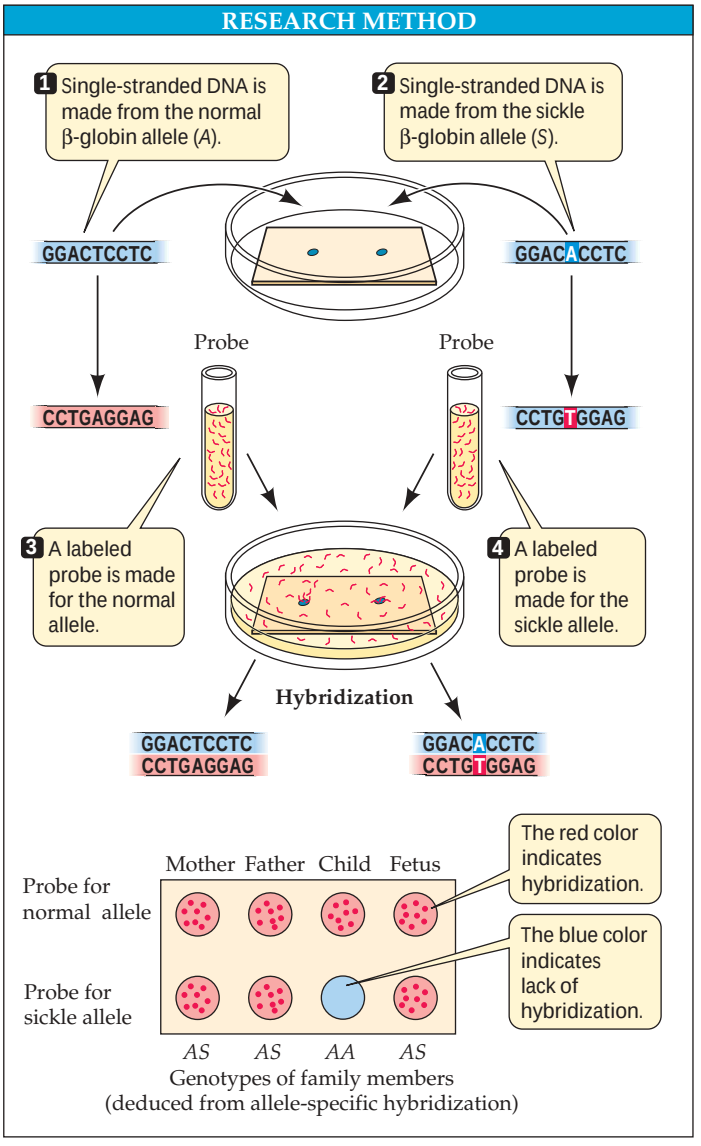
Cancer cells differ from their normal counterparts

Cancer cells differ from the normal cells from which they originate in two major ways.

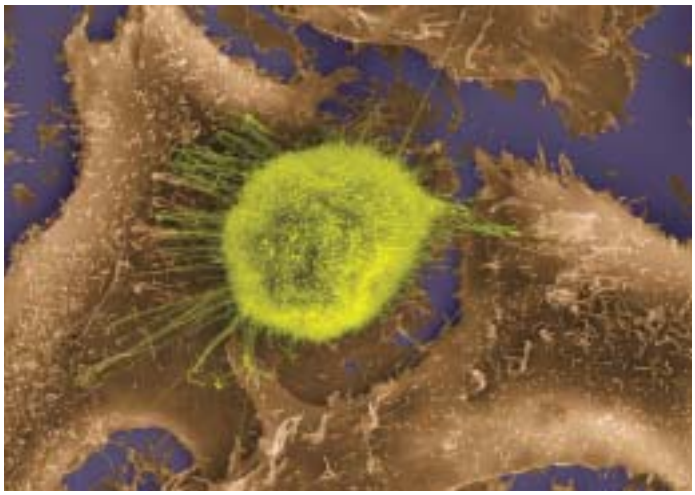
CANCER CELLS LOSE CONTROL OVER CELL DIVISION. Most cells in the body divide only if they are exposed to extracellular influences, such as growth factors or hormones. Cancer cells do not respond to these controls, and instead divide more or less continuously, ultimately forming **tumors** (large masses of cells). By the time a physician can feel a tumor or see one on an X ray or CAT scan, it already contains millions of cells.

Benign tumors resemble the tissue they came from, grow slowly, and remain localized where they develop. A lipoma, for example, is a benign tumor of fat cells that may arise in the armpit and remain there. Benign tumors are not cancers, but they must be removed if they impinge on an important organ, such as the brain.

Malignant tumors, on the other hand, do not look like their parent tissue at all. A flat, specialized lung epithelial cell in the lung wall may turn into a relatively featureless, round,



17.12 DNA Testing by Allele-Specific Oligonucleotide Hybridization Testing of this family reveals that three of them are heterozygous carriers of the sickle allele. The first child, however, has inherited two normal alleles and is neither affected by the disease nor a carrier.



17.13 A Cancer Cell with Its Normal Neighbors This small-cell lung cancer cell (yellow green) is quite different from the surrounding lung epithelial cells from which it came. This particular form of cancer is very lethal, with a 5-year survival rate of less than 10 percent. Most cases are caused by smoking.

malignant lung cancer cell (Figure 17.13). Malignant cells often have irregular structures, such as variable nucleus sizes and shapes. Many malignant cells express the gene for telomerase and thus do not shorten the ends of their chromosomes after each DNA replication.

CANCER CELLS SPREAD TO OTHER TISSUES. The second, and most fearsome, characteristic of cancer cells is their ability to invade surrounding tissues and spread to other parts of the body. This spreading, called **metastasis**, occurs in several stages. First, the malignant tumor secretes chemical signals that cause blood vessels to grow to the tumor and supply it with oxygen and nutrients. This process is called *angiogenesis*. Then, the cancer cells extend into the tissue that surrounds them by actively secreting digestive enzymes to disintegrate the surrounding cells and extracellular materials, working their way toward a blood vessel. Finally, some of the cancer cells enter the bloodstream or the lymphatic system. The journey through these vessels is perilous, and few of the cancer cells survive—perhaps one in 10,000. When by chance a cancer cell arrives at an organ suitable for its further growth, it expresses cell surface proteins that allow it to bind to and invade the new host tissue.

Different forms of cancer affect different parts of the body. About 85 percent of all human tumors are *carcinomas*—cancers that arise in surface tissues such as the skin and the epithelial cells that line the organs. Lung cancer, breast cancer, colon cancer, and liver cancer are all carcinomas. *Sarcomas* are cancers of tissues such as bone, blood vessels, and muscle. *Leukemias* and *lymphomas* affect the cells that give rise to blood cells.

Some cancers are caused by viruses

Peyton Rous's discovery in 1910 that a sarcoma in chickens is caused by a virus that is transmitted from one bird to another spawned an intensive search for cancer-causing viruses in humans. At least five types of human cancer are probably caused by viruses (Table 17.1).

Hepatitis B, a liver disease that affects people all over the world, is caused by the hepatitis B virus, which contaminates blood or is carried from mother to child during birth. The viral infection can be long-lasting and may flare up numerous times. The hepatitis B virus is associated with liver cancer, especially in Asia and Africa, where millions of people are infected. But it does not cause cancer by itself. Some gene mutations that are necessary for tumor formation occur in the infected cells of Asians and Africans, although apparently not in those of Europeans and North Americans.

An important group of virally induced cancers among North Americans and Europeans is the various anogenital cancers caused by papillomaviruses. The genital and anal warts that these viruses cause often develop into tumors. These viruses seem to be able to act on their own, not needing mutations in the host tissue for tumors to arise. Sexual transmission of these papillomaviruses is unfortunately widespread.

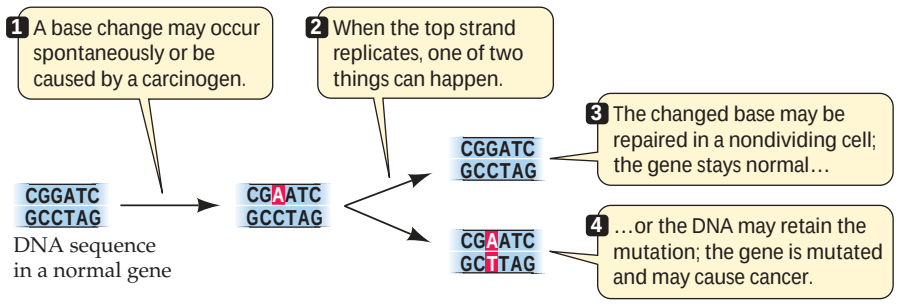
Most cancers are caused by genetic mutations

Worldwide, no more than 15 percent of all cancers may be caused by viruses. What causes the other 85 percent? Because most cancers develop in older people, it is reasonable to assume that one must live long enough for a series of events to occur. This assumption turns out to be correct, and the events are genetic mutations.

DNA can be damaged in many ways. As we saw in Chapter 12, spontaneous mutations arise because of chemical changes in the nucleotides. In addition, certain mutagens, called **carcinogens**, can cause mutations that lead to cancer. Familiar carcinogens include the chemicals that are present in tobacco smoke and meat preservatives, ultraviolet light

17.1 Human Cancers Known To Be Caused by Viruses

CANCER	ASSOCIATED VIRUS
Liver cancer	Hepatitis B virus
Lymphoma, nasopharyngeal cancer	Epstein–Barr virus
T cell leukemia	Human T cell leukemia virus (HTLV-I)
Anogenital cancers	Papillomavirus
Kaposi's sarcoma	Kaposi's sarcoma herpesvirus



17.14 Dividing Cells Are Especially Susceptible to Genetic Damage A base change is more likely to be repaired in a nondividing cell.

from the sun, and ionizing radiation from sources of radioactivity. Less familiar, but just as harmful, are thousands of chemicals that are naturally present in the foods people eat. According to one estimate, these "natural" carcinogens account for well over 80 percent of human exposure to agents that cause cancer.

Both natural and synthetic carcinogens damage DNA, usually by causing changes from one base to another (Figure 17.14). In somatic cells that divide often, such as epithelial and bone marrow cells, there is less time for DNA repair mechanisms to work before replication occurs again. Therefore, such cells are especially susceptible to cancer.

Two kinds of genes are changed in many cancers

The changes in the control of cell division that lie at the heart of cancer can be likened to the controls of an automobile. To make a car move, two things must happen: The gas pedal must be pressed, and the brake must be released. In the human genome, some genes act as **oncogenes**, which "press the gas pedal" to stimulate cell division, and some act as **tumor suppressor genes**, which "put the brake on" to inhibit it.

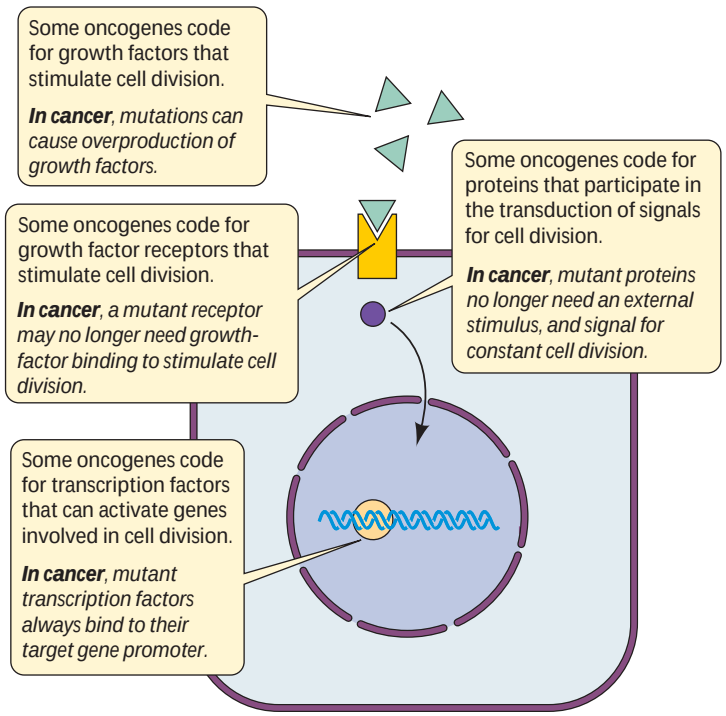
ONCOGENES. The first hint that oncogenes (from the Greek *onco-*, "mass") were necessary for cells to become cancerous came with the identification of virally induced cancers in animals. In many cases, these viruses bring a new gene into their host cells that stimulates cell division when it is expressed in the viral genome. But few types of human cancers are caused by viruses. It soon became apparent that the viral oncogenes had counterparts in the genomes of host cells that were not usually transcribed. So the search for genes that are damaged by carcinogens quickly zeroed in on these cellular oncogenes. Several dozen such genes were soon found.

Oncogenes are genes that have the capacity to stimulate cell division, but are normally "turned off" in differentiated, nondividing cells. Many of them are involved in the pathways by which growth factors stimulate cell division (Figure 17.15). Some remarkable oncogenes control apoptosis (programmed cell death). Activation of these genes by mutation causes them to prevent apoptosis, allowing cells that normally die to continue dividing.

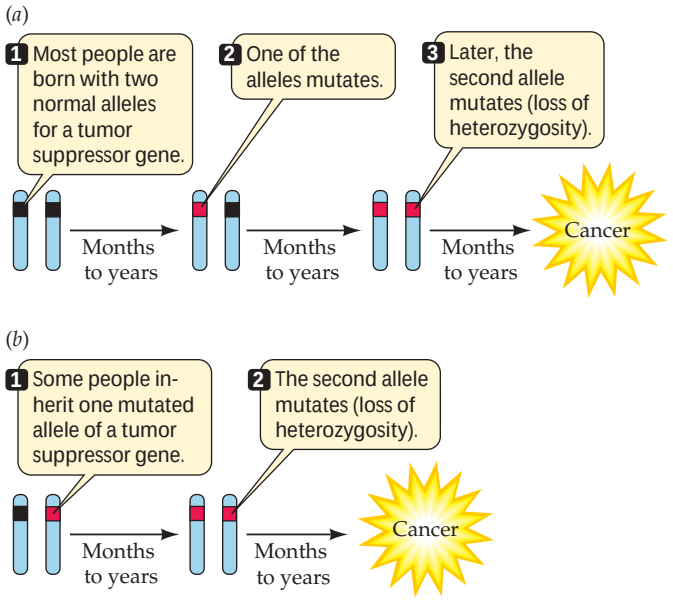
Some oncogenes can be activated by point mutations, others by chromosome changes such as translocations, and still others by gene amplification. Whatever the mechanism, the result is the same: The oncogene becomes activated, and the "gas pedal" for cell division is pressed.

TUMOR SUPPRESSOR GENES. About 10 percent of all cancer is clearly inherited. Often the inherited form of a cancer is clinically similar to a noninherited form that occurs later in life, called the *sporadic* form. The major differences are that the inherited form strikes much earlier in life and usually shows up as multiple tumors.

In 1971, Alfred Knudson used these observations to predict that for a cancer to occur, a tumor suppressor gene, which normally acts as a "brake" on cell division, must be inactivated. But in contrast to oncogenes, in which one mutated allele is all that is needed for activation, the full inactivation



17.15 Oncogene Products Stimulate Cell Division Mutations can affect any of the several ways in which oncogenes normally stimulate cell division, thus causing cancer.



17.16 The "Two-Hit" Hypothesis for Cancer (*a*) Although a single mutation can activate an oncogene, two mutations are needed to inactivate a tumor suppressor gene. (*b*) An inherited predisposition to cancer occurs in people born with one allele already mutated.

of a tumor suppressor gene requires that both alleles be turned off, which requires two mutational events. It takes a long time for both alleles in a single cell to mutate and cause sporadic cancer. But people with inherited cancer are born with one mutant allele for the tumor suppressor gene, and need just one more mutational event for its full inactivation (Figure 17.16).

The isolation of various tumor suppressor genes has confirmed Knudson's "two-hit" hypothesis. Some of these genes are involved in inherited forms of rare childhood cancers such as retinoblastoma (a tumor of the eye) and Wilms' tumor of the kidney as well as in inherited breast and prostate cancers.

An inherited form of breast cancer demonstrates the effect of tumor suppressor genes. The 9 percent of women who inherit one mutated allele of the gene *BRCA1* have a 60 percent chance of having breast cancer by age 50 and an 82 percent chance of developing it by age 70. The comparable figures for women who inherit two normal alleles of the gene are 2 percent and 7 percent, respectively.

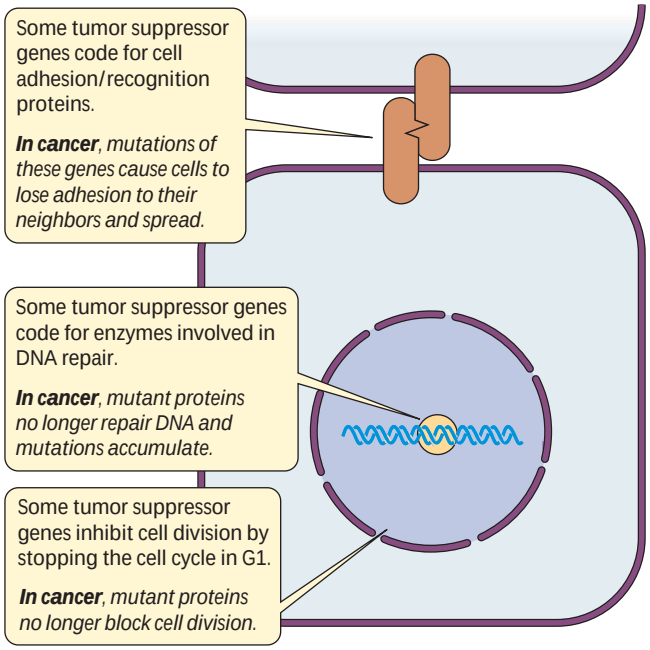
How do tumor suppressor genes act in the cell? Like the oncogenes, they are normally involved in vital cell functions (Figure 17.17). Some regulate progress through the cell cycle. The protein encoded by *Rb*, a gene that was first described for its contribution to retinoblastoma, is active during the G1 phase. In its active form, it encodes a protein that binds to and inactivates transcription factors that are necessary for progress to the S phase and the rest of the cell cycle. In nondividing cells, *Rb* remains active, preventing cell division until the proper growth factor signals are present. When the Rb protein is inactivated by mutation, the cell cycle moves forward independently of growth factors.

The protein product of another widespread tumor suppressor gene, *p53*, also stops the cell cycle at G1. It does this by acting as a transcription factor, stimulating the production of (among other things) a protein that blocks the interaction of a cyclin and a protein kinase needed for moving the cell cycle beyond G1. This gene is mutated in many types of cancers, including lung cancer and colon cancer.

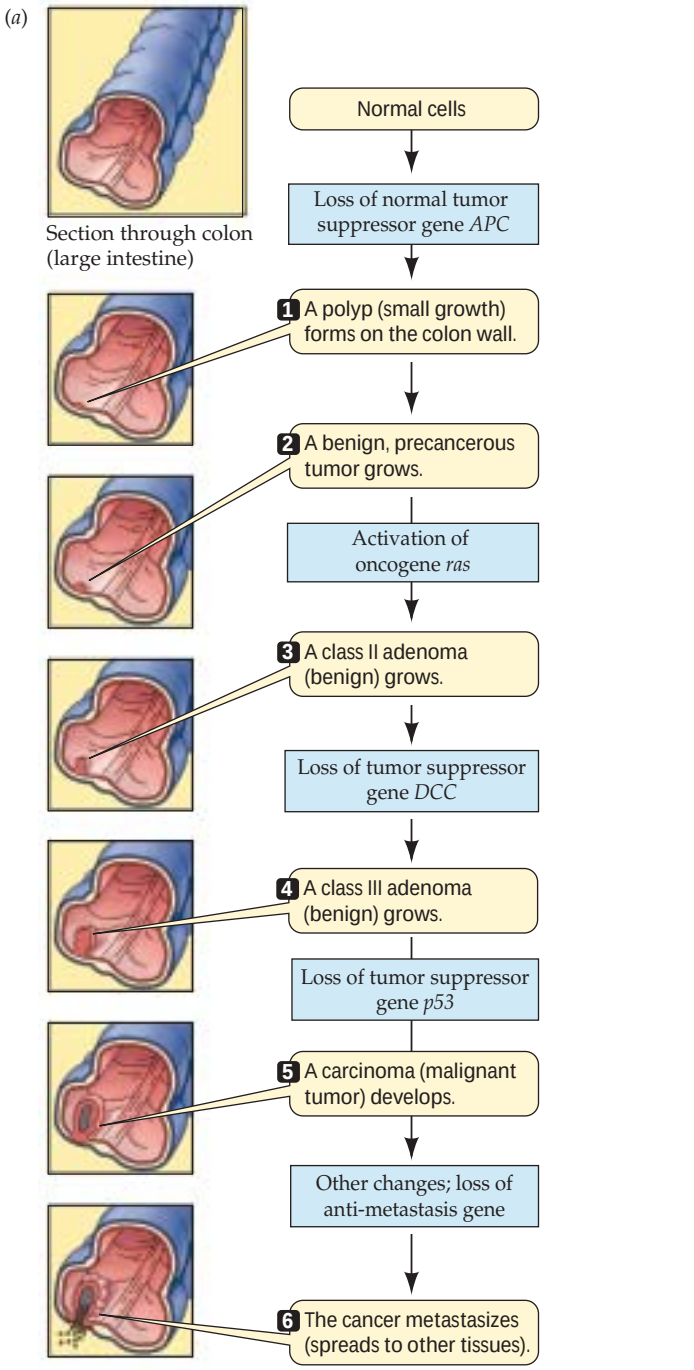
The pathway from normal cell to cancerous cell is complex

The "gas pedal" and "brake" analogies we have been using for oncogenes and tumor suppressor genes, respectively, are elegant but simplified. There are many oncogenes and tumor suppressor genes, some of which act only in certain cells at certain times. Therefore, a complex sequence of events must occur before a normal cell becomes malignant.

Because colon cancer progresses to full malignancy slowly, it is possible to describe the oncogene and tumor suppressor gene mutations at each stage in great molecular detail. Figure 17.18 outlines the progress of this form of cancer. At least three tumor suppressor genes and one oncogene must be mutated in sequence for an epithelial cell in the colon to be-



17.17 Tumor Suppressor Gene Products Inhibit Cell Division and Cancer Mutations can affect any of the several ways in which tumor suppressor genes inhibit cell division, allowing cells to divide and form a tumor.



17.18 Multiple Mutations Transform a Normal Colon Epithelial Cell into a Cancer Cell *(a)* In colon cancer, at least five genes are mutated in a single cell. *(b)* Colonoscopy is the current standard screening test for colon cancer. These views reveal (*i*) normal colon tissue, (*ii*) a benign adenoma (stalked polyp), and (*iii*) adenocarcinoma (a malignant tumor).

come metastatic. Although the occurrence of all these events in a single cell might appear unlikely, remember that the colon has millions of cells, that the cells giving rise to epithelial cells are constantly dividing, and that these changes take place over many years of exposure to natural and synthetic carcinogens as well as spontaneous mutations.

The characterization of the molecular changes in tumor cells has opened up the possibility of genetic diagnosis and screening for cancer. Many cancers are now commonly diagnosed in part by specific oligonucleotide probes for oncogene or tumor suppressor gene alterations. It is also possible to detect early in life whether an individual has inherited a mutated tumor suppressor gene. A person who inherits mutated copies of the tumor suppressor genes involved in colon cancer, for example, normally would have a high probability of developing this cancer by age 40. Surgical removal of the colon would prevent a metastatic tumor from arising.

Treating Genetic Diseases

Most treatments for genetic diseases simply try to alleviate the symptoms that affect the patient. But to effectively treat diseases caused by genes—whether they affect all cells, as in inherited disorders such as PKU, or only somatic cells, as in cancer—physicians must be able to diagnose the disease accurately, understand how the disease works at the molecular level, and intervene early, before the disease ravages or kills the individual.

Basic research has provided the knowledge needed for accurate diagnostic tests, as well as a preliminary understanding of these diseases at the molecular level. Physicians are now applying this knowledge to develop new treatments for genetic diseases. In this section, we will see that approaches to treatment range from specifically modifying the mutant phenotype to supplying the normal version of a mutant gene.

One approach to treatment is to modify the phenotype

There are three ways of altering the phenotype of a genetic disease so that it no longer harms an individual: restricting the substrate of a deficient enzyme, inhibiting a harmful metabolic reaction, or supplying a missing protein product.

RESTRICTING THE SUBSTRATE. Restricting the substrate of a deficient enzyme is the approach taken when a newborn is diagnosed with PKU. In this case, the deficient enzyme is phenylalanine hydroxylase, and the substrate is phenylalanine. The infant's inability to break down the phenylalanine in food leads to a buildup of the substrate, which causes the clinical symptoms. So the infant is immediately put on a special diet that contains only enough phenylalanine for immediate use. Lofenelac, a milk-based product that is

low in phenylalanine, is fed to these infants just like formula. Later, certain fruits, vegetables, cereals, and noodles low in phenylalanine can be added to the diet. Meat, fish, eggs, dairy products, and bread, which contain high amounts of phenylalanine, must be avoided, especially during childhood, when brain development is most rapid. The artificial sweetener aspartame must also be avoided because it is made of two amino acids, one of which is phenylalanine.

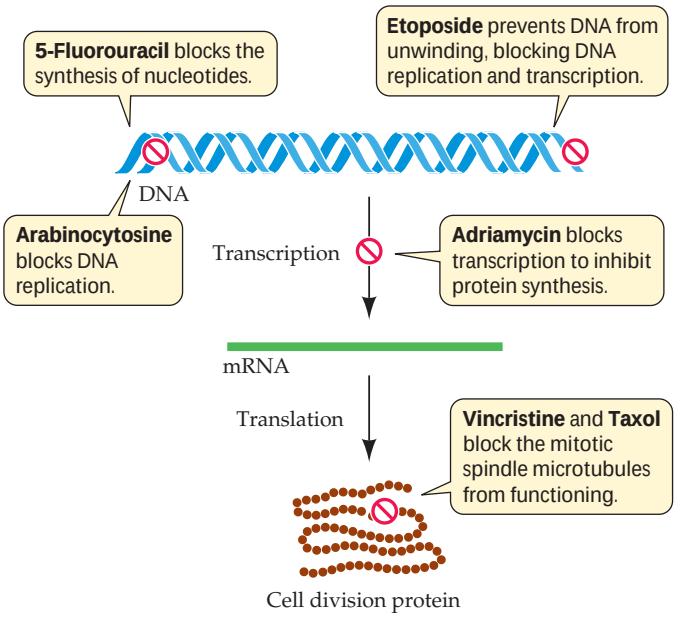
People with PKU are generally advised to stay on a low-phenylalanine diet for life. Although maintaining these dietary restrictions may be difficult, it is effective. Numerous follow-up studies since newborn screening was initiated have shown that people with PKU who stay on the diet are no different from the rest of the population in terms of mental ability. This is an impressive achievement in public health, given the extent of mental retardation in untreated patients.

METABOLIC INHIBITORS. As we described earlier, people with familial hypercholesterolemia accumulate dangerous levels of cholesterol in their blood. These people are not only unable to metabolize dietary cholesterol, but also synthesize a lot of it. One effective treatment for people with this disease is the drug mevinolin, which blocks the patient's own cholesterol synthesis. Patients who receive this drug need only worry about cholesterol in their diet, and not about the cholesterol their cells are making.

Metabolic inhibitors also form the basis of chemotherapy for cancer. The strategy is to kill rapidly dividing cells, since rapid cell division is the hallmark of malignancy. But such a strategy is not selective for tumor cells. Many drugs kill dividing cells (Figure 17.19), but most of those drugs also damage other, noncancerous, dividing cells in the body. Therefore, it is not surprising that people undergoing chemotherapy suffer side effects such as loss of hair (due to damage to the skin epithelium), digestive upsets (gut epithelial cells), and anemia (bone marrow stem cells). The effective dose of these highly toxic drugs for treating the cancer is often just below the dose that would kill the patient, so they must be used with utmost care. Often they can control the spread of cancer, but not cure it.

SUPPLYING THE MISSING PROTEIN. An obvious way to treat a disease phenotype in which a functional protein is missing is to supply that protein. This approach is the basis of treatment of hemophilia, in which the missing blood clotting protein is supplied in pure form. The production of human clotting protein by recombinant DNA technology has made it possible for a pure protein to be given instead of crude blood products, which could be contaminated with the AIDS virus or other pathogens.

Unfortunately, the phenotypes of many diseases caused by genetic mutations are very complex. Simple interventions like those we have described do not work for most such diseases. Indeed, a recent survey showed that current therapies for 351 diseases caused by single-gene mutations improved patients' life spans by only 15 percent.



17.19 Strategies for Killing Cancer Cells The medications used in chemotherapy for cancer attack rapidly dividing cancer cells in several ways. Unfortunately, most of them also affect noncancerous dividing cells.

Gene therapy offers the hope of specific treatments

Perhaps the most obvious thing to do when a cell lacks a functional allele is to provide one. Such **gene therapy** approaches to diseases ranging from the rare inherited disorders caused by single-gene mutations to cancer, AIDS, and atherosclerosis are under intensive investigation.

Gene therapy in humans seeks to insert a new gene that will be expressed in the host. Thus, the new DNA is often attached to a promoter that will be active in human cells. The physicians who are developing this "molecular medicine" are confronted by all the challenges of recombinant DNA technology: they must find effective vectors and ensure efficient uptake, precise insertion into the host DNA, appropriate expression and processing of mRNA and protein, and selection within the body for the cells that contain the recombinant DNA.

Which human cells should be the targets of gene therapy? The best approach would be to replace the nonfunctional allele with a functional one in every cell of the body. But vectors to do this are simply not available, and delivery to every cell poses a formidable challenge. Until recently, attempts at gene therapy have used **ex vivo** techniques. That is, physicians have taken cells from the patient's body, added the new gene to those cells in the laboratory, and then returned the

17.20 Gene Therapy: The Ex Vivo Approach New genes are added to somatic cells taken from a patient's body. These transgenic cells are then returned to the body to make the missing gene product.

cells to the patient in the hope that the correct gene product would be made (Figure 17.20). Two examples demonstrate this technique:

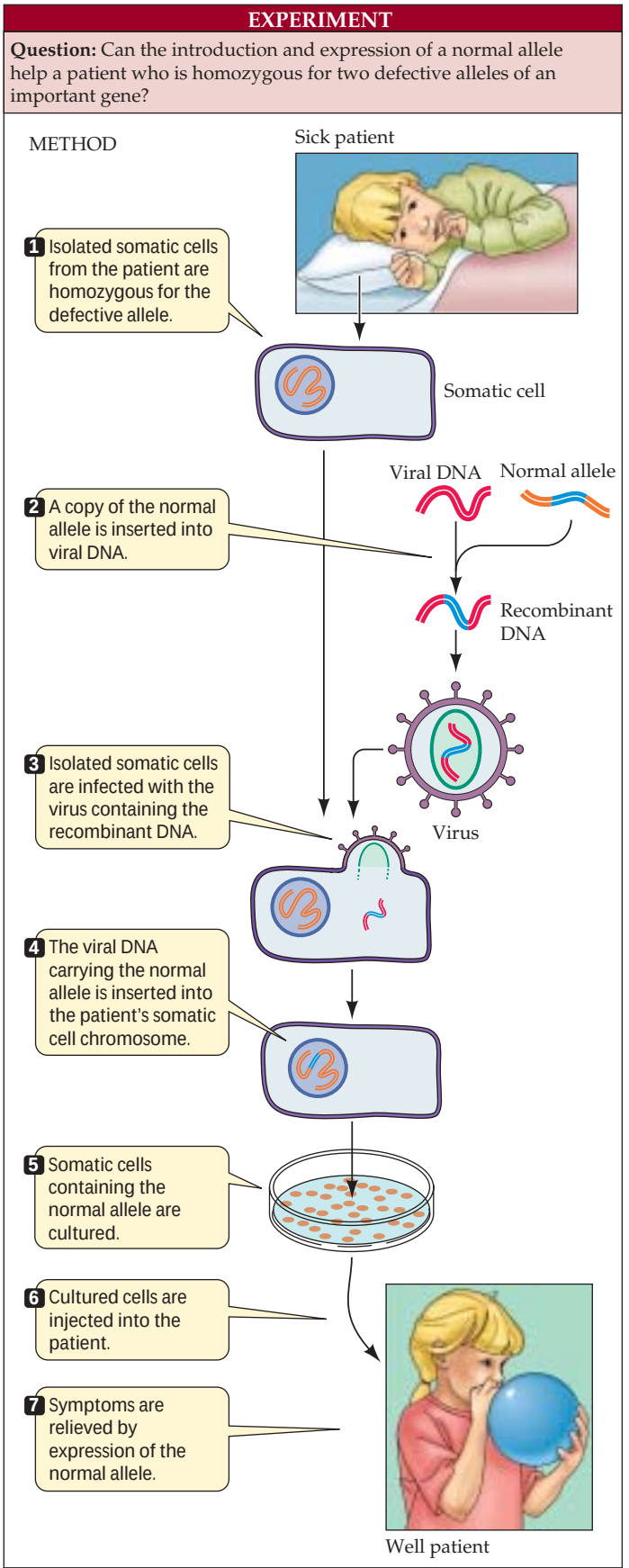
- *Adenosine deaminase* is needed for maturation of white blood cells, and people without this enzyme have severe immune system deficiencies. A functional gene for adenosine deaminase was introduced via a viral vector into the white blood cells of a girl with a genetic deficiency of this enzyme. Unfortunately, mature white blood cells were used, and although they survived for a time in the girl and provided some therapeutic benefit, they eventually died, as is the normal fate of such cells. Further clinical trials have used stem cells, the bone marrow cells that constantly divide to produce white blood cells.
- *Hemophilia* is a disease in which patients do not make enough of a blood clotting protein. Some cells from the skin of patients' arms were removed and transfected with a plasmid containing a normal allele of the clotting protein gene. The cells were then reintroduced into the patients' body fat, where they produced adequate protein for normal clotting.

The other approach to gene therapy is to insert the gene directly into cells in the body of the patient. This **in vivo** approach is being attempted for various types of cancer. Lung cancer cells, for example, are accessible to such treatment if the DNA or vector is given as an aerosol through the respiratory system. Vectors carrying functional alleles of the tumor suppressor genes that are mutated in the tumors, as well as vectors expressing antisense RNAs against oncogene mRNAs, have been successfully introduced in this way to patients with lung cancer, with some clinical improvement.

Several thousand patients, over half of them with cancer, have undergone gene therapy. Most of these clinical trials have been at a preliminary level, in which people are given the therapy to see whether it has any toxicity and whether the new gene is actually incorporated into the patient's genome. More ambitious trials are under way, in which a larger number of patients will receive the therapy with the hope that their disease will disappear, or at least improve.

Sequencing the Human Genome

In 1984, the United States government sponsored a conference on the detection of DNA damage in people exposed to low levels of radiation, such as those who had survived the atomic bomb in Japan 39 years earlier. Scientists attending



this conference quickly realized that the ability to detect such damage would also be useful in evaluating environmental mutagens. But in order to detect changes in the human genome, scientists first needed to know its normal sequence.

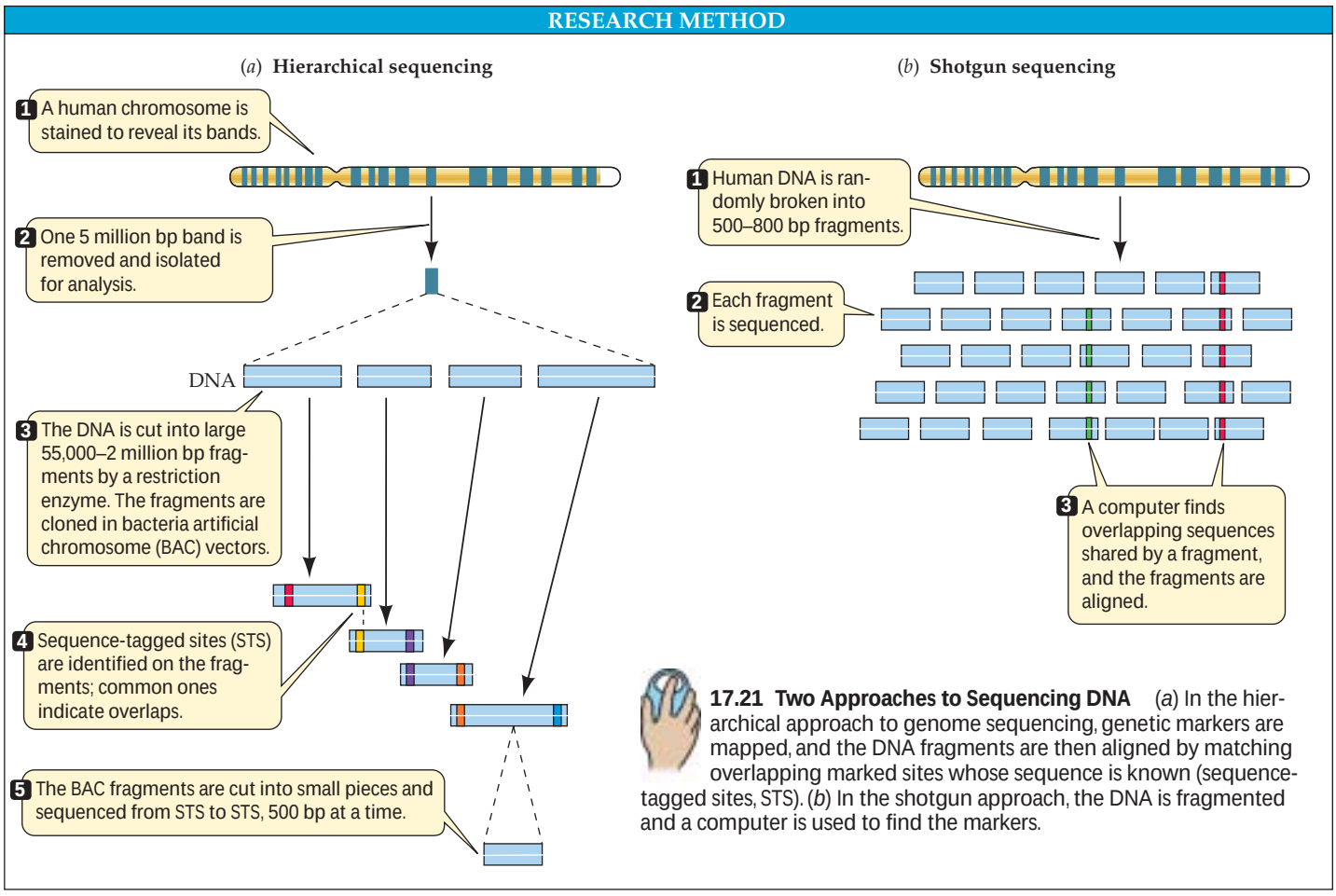
In 1986, Renato Dulbecco, who won the Nobel prize for his pioneering work on cancer-causing viruses, suggested that determining the normal sequence of human DNA could also be a boon to cancer research. He proposed that the scientific community be mobilized for the task. The result was the publicly funded **Human Genome Project**, an international effort. In the 1990s, private industry launched its own sequencing effort.

There are two approaches to genome sequencing

Each human chromosome consists of one double-stranded molecule of DNA. Because of their differing sizes, the 46 human chromosomes can be separated from one another and identified (see Figure 9.13). So it is possible to isolate the DNA of each chromosome for sequencing. The straightforward approach would be to start at one end of a chromosome and simply sequence the entire 50 million base pairs. Unfortunately, this approach is not practical. The DNA of a molecule that is 50 million base pairs long cannot be sequenced all at once; only about 700 base pairs at a time can be sequenced. (See Figure 11.21 to review the DNA sequencing technique.)

To sequence an entire genome, chromosomal DNA is first cut into fragments about 500 base pairs long, then each fragment is sequenced. For the human genome, which has about 3.2 billion base pairs, there are more than 6 million such fragments. The problem then becomes putting these millions of fragments back together, like the pieces of a jigsaw puzzle. This problem can be overcome by breaking up the DNA into "sub-jigsaws" that overlap and aligning the overlapping fragments. There are two ways to do this.

HIERARCHICAL SEQUENCING. The publicly funded sequencing team used a method known as **hierarchical sequencing**. First, they systematically identified short marker sequences along the chromosomes, such that every fragment of DNA to be sequenced would contain a marker (Figure 17.21*a*). This method can be compared to making a road map, showing towns with the mileage separating them. The "towns"



17.21 Two Approaches to Sequencing DNA (*a*) In the hierarchical approach to genome sequencing, genetic markers are mapped, and the DNA fragments are then aligned by matching overlapping marked sites whose sequence is known (sequence-tagged sites, STS). (*b*) In the shotgun approach, the DNA is fragmented and a computer is used to find the markers.

are the marker sequences, and the "mileage" is in base pairs. The simplest markers are the recognition sequences for restriction enzymes.

Some restriction enzymes recognize 8–12 base pairs in DNA, not just the usual 4–6 base pairs. A DNA molecule with several million base pairs will have relatively few of these larger sites, and thus the enzyme will generate a small number of relatively large fragments. These large fragments can be added to a vector called a *bacterial artificial chromosome* (BAC), which can carry about 250,000 base pairs of inserted DNA, and inserted into bacteria to create a gene library.

The volumes (fragments) in this library can be arranged in the proper order along the chromosome map by using the marker sequences. To arrange the DNA fragments on the map, libraries made with different restriction enzymes are compared. If two large fragments of DNA cut with different enzymes have the same marker, they must overlap. This method works, but is slow.

SHOTGUN SEQUENCING. Instead of finding markers, fragmenting the DNA, and then sequencing it, the "shotgun" approach cuts the DNA at random into small, sequencing-ready fragments and lets powerful computers determine markers that overlap (Figure 17.21*b*). The fragments can then be aligned.

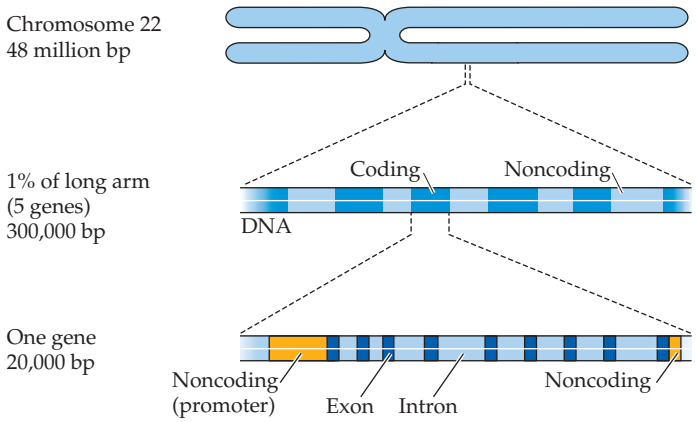
The **shotgun sequencing** method, which has been used by private industry, is much faster than the hierarchical approach because there is no need to make a map. At first there was considerable skepticism about this method. There were concerns that without rigorous prior mapping of marker sites on the chromosomes, the computer might pick out repetitive sequences common to many DNA fragments and line the fragments up incorrectly. But the rapid rate of development of sophisticated computers and software has allowed the shotgun method to be refined to a point at which inaccurate alignment is not a major problem. The entire 180 million-base-pair fruit fly genome (see Chapter 14) was sequenced by the shotgun method in little over a year. This success proved that the shotgun method might work for the much larger human genome, and in fact, it did.

The sequence of the human genome has been determined

The two teams of scientists announced a draft human genome sequence in June 2000 to great fanfare, and published their data simultaneously in February 2001. By the start of 2003, the final sequence was completed, two years ahead of the schedule set over a decade previously and well under budget.

The sequencing of the human genome revealed several interesting characteristics:

- *Of the 3.2 billion base pairs, less than 2 percent are coding regions, containing a total of 21,000 genes.* Before sequencing began, estimates of the number of human genes ranged from 80,000 to 100,000. This lower number of genes, not many more than the fruit fly, means that the observed diversity of proteins, which led to the 100,000 estimate, must be produced posttranscriptionally. An average eukaryotic gene, then, codes for several different proteins.
- *The average gene has 27,000 base pairs.* There is great variation in gene sizes, from 1,000 to 2.4 million base pairs. That is to be expected, as human proteins vary in size (as do RNAs), ranging from 100 to about 5,000 amino acids per polypeptide chain. Virtually all human genes have many introns (Figure 17.22).
- *Over 50 percent of the genome is made up of highly repetitive sequences.* Repetitive sequences near genes are GC-rich, while those farther away from genes are AT-rich.
- *Almost all (99.9%) of the genome is the same in all people.* Even this apparent homogeneity means that there are many individual differences. Scientists have mapped over 2 million **single-nucleotide polymorphisms** (**SNPs**)—bases that differ in at least 1 percent of people.
- *Genes are not evenly distributed over the genome.* The small chromosome 19 is packed densely with genes, while chromosome 8 has long stretches of "gene desert," with no coding regions. The Y chromosome has the fewest genes (231), while chromosome 1 has the most (2,968).
- *The functions of many genes are not known.* There are 740 genes coding for RNAs that are not translated into proteins. Of these RNAs, several dozen are tRNAs, and a few are rRNAs and splicing RNAs. The roles of the rest are not clear. Nor are the roles of the hundreds of genes encoding protein kinases, although it is a good bet that they are involved in cell signaling.



17.22 The Human Genome The genomic anatomy of chromosome 22 is shown here.

The human genome sequence has many applications

Reading the human "book of life" is an achievement that ranks with other recent great events in scientific exploration, such as landing on the moon. But gene sequencing, and the tools developed to carry it out, are changing biology in many other ways as well.

- The sequences of other organisms have provided insights and practical information on both prokaryotic and eukaryotic genomes. Many genes sequenced and identified in "simpler" organisms have homologs in humans, so these findings have facilitated the identification of human genes.
- Mapping technology and SNPs have made the isolation of human genes by positional cloning much easier because of the huge number of genetic markers now available. Disease-related genes have been identified in this way.
- Genetic variation in drug metabolism has been a medical problem for a long time. The emerging field of *pharmacogenomics* is identifying the genes responsible for this variation and developing tests to predict who will react best to which medications.
- DNA chips (see Figure 16.10) are being used to analyze the specific expression of thousands of genes in different cells in different biochemical states. For example, a Cancer Genome Anatomy Project is seeking to make an mRNA "fingerprint" of a tumor at each stage of its development. Finding out which genes are expressed at which stage will be important not only in diagnosis, but also in identifying targets for gene therapy.
- "Genome prospecting" refers to the search for important polymorphisms in specific human populations. For example, the Pima Indians in Arizona have a high frequency of extreme obesity and diabetes. A search of their genomes might reveal genes predisposing them to these conditions.

The end result of all of this knowledge of the human genome may be a new approach to medical care, in which each person's genome will be used to prescribe lifestyle changes and treatments that can maximize that person's genetic potential (Figure 17.23).

How should genetic information be used?

When the genetic defect that causes cystic fibrosis was discovered, many people predicted a "tidal wave" of genetic testing for heterozygous carriers. Everyone, it was thought, would want the test—especially the relatives of people with the disease. But this tidal wave has not developed. To find out why, a team of psychologists, ethicists, and geneticists interviewed 20,000 people in the United States. What the researchers found surprised them. Most people are simply not very interested in their genetic makeup, unless they have a close relative with a genetic disease and are involved in a decision about pregnancy.

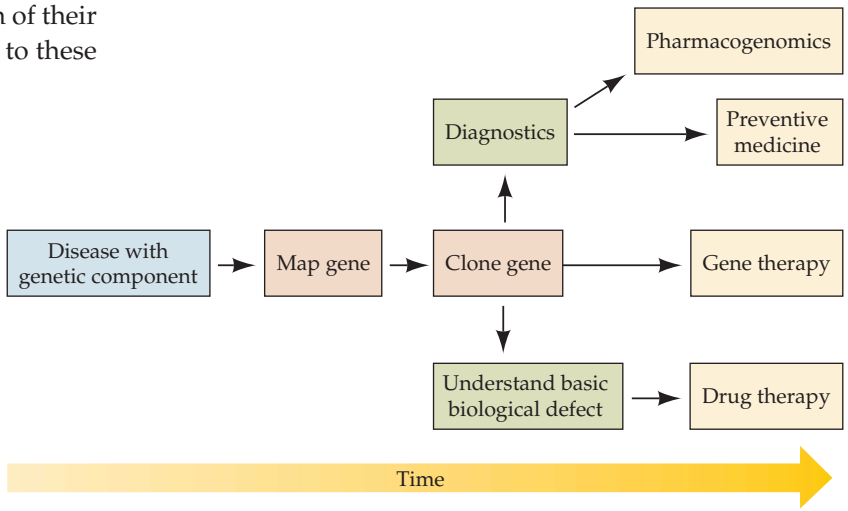
There are other people, however, who might be very interested in the results of genetic testing. For example, people who test positive for genetic abnormalities, from hypercholesterolemia to cancer, might be denied employment or health insurance. Consequently, there are laws that prohibit discrimination on the basis of genetic information.

The search for valuable genes in diverse human populations has raised many concerns about exploitation and commercialization of a person's DNA sequence. Is a gene that confers resistance to cancer, for example, the property of an individual, an ethnic group in which it may be frequent, the pharmaceutical company that finds it, or humanity at large?

This issue of ownership is being tested worldwide, perhaps most directly in Iceland, whose 270,000 people trace their ancestry back to the first settlers to arrive on the island 1,100 years ago. Tissues from Iceland's entire population have been sampled and stored for several generations. This tissue bank is a potential gold mine for genetic prospectors. A single company has been set up, with government support, to sell the knowledge that comes from analyzing the genomes of Iceland's people.

The company's approach to mining this genetic lode is illustrated by its search for genes that predispose people to asthma, a respiratory disease:

- The names of Icelanders with asthma were run through a genealogy database.



17.23 Is This the Future of Medicine? The sequencing of the human genome may result in a new approach to medicine that is oriented to the genetic and functional individuality of each patient.

- One group of 104 patients were descended from a single ancestor, born in 1710 (11 generations ago). It was considered likely that the same genes for predisposition to asthma were present in all 104 patients.
- Marker genes were sought that would identify alleles that all 104 patients had in common, and a small number of such genes were subsequently identified.

The characterization of these genes will lead to a greater understanding of how a group of genes interacts to produce a complex phenotype.

The proteome is more complex than the genome

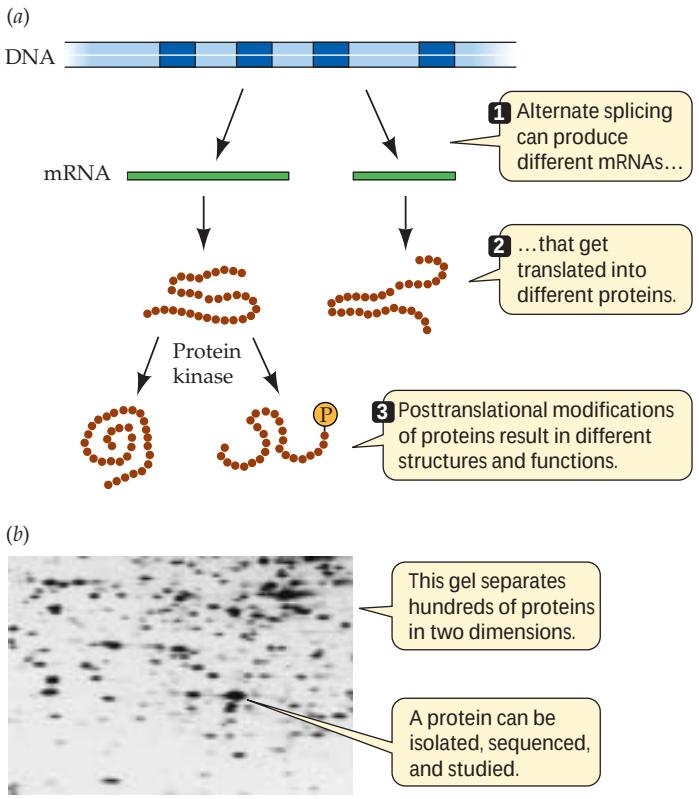
Genome sequencing revealed that humans had only about one-third as many genes as had been predicted based on the number of proteins found in human cells. Perhaps this lower number should not have come as a surprise; as we saw in Chapters 12 and 14, many genes encode more than a single protein (Figure 17.24*a*). Alternative splicing leads to different combinations of exons in the mature mRNAs transcribed from a single gene (see Figure 14.20). Posttranslational modifications also add to the forms of a protein that can be made from one gene. Many proteins have peptides clipped off, have sugars added, or are phosphorylated after translation (see Figure 12.16). Therefore, the sum total of the proteins produced by an organism—its **proteome**—is more complex than its genome. The one-gene, one-polypeptide relationship, once a central one in biology, has been laid to rest by genomics.

There are two major ways to analyze the proteome:

- *Two-dimensional gel electrophoresis* attempts to separate all of the individual proteins of a particular cell or tissue into spots that can be analyzed quantitatively and qualitatively (Figure 17.24*b*).
- *Mass spectrometry* employs electromagnets to identify proteins by the masses of their atoms and displays them as peaks on a graph.

The ultimate aim of proteomics is just as ambitious as the aim of genomics. While genomics seeks to describe the genome and its expression, proteomics seeks to describe the phenotypes of the expressed proteins with precision.

An amazing example of proteomics, combined with DNA chip technology, is the recent comparison of brain proteins in chimpanzees and humans. DNA sequencing has shown that humans and chimpanzees differ by no more than 3 percent at the DNA level. Svante Pääbo and his colleagues in Germany, The Netherlands, and the United States examined gene and protein expression in the "thinking" part of the brains (the cortex) of three chimps and three humans who had died of natural causes. Of 12,000 DNA sequences tested for expression as mRNA, only 175 (1.4%) showed differences between the two species, a truly humbling result. But proteomics showed that the *kinds* of proteins expressed by those sequences were 7.4 percent different, probably due to alternative splicing. And the *amounts* of proteins were also quite different (31.4%). So what makes our brain different from a chimpanzee's is more quantitative than qualitative. Thus, the control of gene expression may be the key to human evolution.



17.24 Proteomics (*a*) A small number of genes can make a large number of proteins. (*b*) A cell's proteins can be separated in two dimensions on the basis of charge and size by gel electrophoresis.

Chapter Summary

Abnormal or Missing Proteins: The Mutant Phenotype

- In some human genetic diseases, a single protein is missing or nonfunctional. **Review Figure 17.1**
- A mutation in a single gene can cause alterations in its protein product that may lead to clinical abnormalities or have no effect. **Review Figure 17.2**
- The genes that code for enzymes, membrane receptors, and membrane transport proteins can be mutated, causing diseases such as phenylketonuria, familial hypercholesterolemia, and cystic fibrosis. **Review Figure 17.3**
- Some diseases are caused by mutations that affect structural proteins; examples include Duchenne muscular dystrophy and hemophilia.

▶ Prions are disease-causing proteins with an altered conformation that can be transmitted from one person to another and alter the same protein in the second person. **Review Figure 17.4**

▶ Relatively few common human diseases are caused by single-gene mutations. Most are caused by the interactions of many genes and proteins with the environment.

▶ Human genetic diseases show different patterns of inheritance. Mutant alleles may be inherited as autosomal recessives, autosomal dominants, or X-linked conditions.

▶ Some human diseases are caused by chromosomal abnormalities.

Mutations and Human Diseases

▶ Molecular biological techniques have made possible the isolation of many genes responsible for human diseases.

▶ One method of identifying the gene responsible for a disease is to isolate the mRNA for the abnormal protein in question and then use that mRNA to locate the gene in a gene library. DNA from a patient with a chromosome deletion can be compared with DNA from a person who does not show this deletion to isolate a missing gene. **Review Figure 17.6**

▶ In positional cloning, genetic markers are used as guides to point the way to a gene. These markers may be restriction fragment length polymorphisms that are linked to a mutant gene. **Review Figure 17.7.** **See Web/CD Activity 17.1**

▶ Human mutations range from point mutations to large deletions. Some of the most common mutations occur where the modified base 5-methylcytosine is converted to thymine. **Review Figure 17.8**

▶ The effects of the fragile-X chromosome worsen with each generation. This pattern is caused by a triplet repeat that tends to expand with each new generation. **Review Figure 17.9**

▶ Genomic imprinting results in a gene being differentially expressed depending on the sex of the parent it comes from.

Detecting Genetic Variations: Screening for Human Diseases

▶ Genetic screening detects human genetic mutations. Some protein abnormalities can be detected by simple tests, such as tests for the presence of excess substrate or lack of product. **Review Figure 17.10**

▶ The advantage of testing DNA for mutations directly is that any cell can be tested at any time in the life cycle.

▶ There are two predominant methods of DNA testing: allele-specific cleavage and allele-specific oligonucleotide hybridization. **Review Figures 17.11, 17.12.** **See Web/CD Tutorial 17.1**

Cancer: A Disease of Genetic Changes

▶ Tumors may be benign, growing only to a certain extent and then stopping, or malignant, spreading through organs and to other parts of the body.

▶ At least five types of human cancers are caused by viruses, which account for about 15 percent of all cancers. **Review Table 17.1**

▶ Eighty-five percent of human cancers are caused by genetic mutations of somatic cells. These mutations occur most commonly in dividing cells. **Review Figure 17.14**

▶ Normal cells contain oncogenes, which, when mutated, can become activated and cause cancer by stimulating cell division or preventing cell death. **Review Figure 17.15**

▶ About 10 percent of all cancer is inherited as a mutation of a tumor suppressor gene, which normally acts to slow down the cell cycle. For cancer to develop, both alleles of a tumor suppressor gene must be mutated.

▶ In inherited cancer, an individual inherits one mutant allele of a tumor suppressor gene, and a somatic mutation occurs in the second one. In sporadic cancer, two normal alleles are inherited, so two mutational events must occur in the same somatic cell to produce cancer. **Review Figures 17.16, 17.17**

▶ Mutations must activate several oncogenes and inactivate several tumor suppressor genes for a cell to produce a malignant tumor. **Review Figure 17.18**

Treating Genetic Diseases

▶ Most genetic diseases are treated symptomatically. However, as more knowledge is accumulated, specific treatments are being devised.

▶ One approach to treating genetic diseases is to modify the phenotype—for example, by manipulating the diet to restrict the substrate of a missing enzyme, providing specific metabolic inhibitors to prevent a harmful reaction, or supplying a missing metabolite or protein. **Review Figure 17.19**

▶ In gene therapy, a mutant gene is replaced with a normal gene. The affected cells can be removed, the new gene added, and the cells returned to the body, or the new gene can be inserted via a vector directly into the patient. **Review Figure 17.20**

Sequencing the Human Genome

▶ Sequencing the entire human genome required sequencing many 500-base-pair fragments and then fitting their sequences back together.

▶ In hierarchical gene sequencing, marker sequences are identified and mapped on the chromosome before DNA is fragmented. These markers are then sought in the sequenced fragments and used to align them. In the shotgun approach, the DNA is fragmented and sequenced, and common markers are then identified by computer. **Review Figure 17.21.** **See Web/CD Tutorial 17.2**

▶ The human genome has only about 21,000 genes. **Review Figure 17.22**

▶ The identification of human genes may lead to a new molecular medicine. **Review Figure 17.23**

▶ As more genes relevant to human health are described, concerns about how such information is used are growing.

▶ Humans make many more proteins than predicted by their number of genes because each gene can encode several different proteins as a result of variation in posttranscriptional and post-translational regulation. Thus, the proteome is more complex than the genome. **Review Figure 17.24**

See Web/CD Activity 17.2 for a concept review of this chapter.

Self-Quiz

1. Phenylketonuria is an example of a genetic disease in which
   a. a single enzyme is not functional.
   b. inheritance is sex-linked.
   c. two parents without the disease cannot have a child with the disease.
   d. mental retardation always occurs, regardless of treatment.
   e. a transport protein does not work properly.
2. Mutations of the gene for β-globin
   a. are usually lethal.
   b. occur only at amino acid position 6.

c. number in the hundreds.
d. always result in sickling of red blood cells.
e. can always be detected by gel electrophoresis.

3. Multifactorial (complex) diseases
   a. are less common than single-gene diseases.
   b. involve the interaction of many genes with the environment.
   c. affect less than 1 percent of humans.
   d. involve the interactions of several mRNAs.
   e. are exempified by sickle-cell anemia.
4. In fragile-X syndrome,
   a. females are affected more severely than males.
   b. a short sequence of DNA is repeated many times to create the fragile site.
   c. both the X and Y chromosomes tend to break when prepared for microscopy.
   d. all people who carry the gene that causes the syndrome are mentally retarded.
   e. the basic pattern of inheritance is autosomal dominant.
5. Most genetic diseases are rare because
   a. each person is unlikely to be a carrier for harmful alleles.
   b. genetic diseases are usually sex-linked and so uncommon in females.
   c. genetic diseases are always dominant.
   d. a married couple probably do not carry the same recessive alleles.
   e. mutation rates in human are low.
6. Mutational "hot spots" in human DNA
   a. always occur in genes that are transcribed.
   b. are common at cytosines that have been modified to 5-methylcytosine.
   c. involve long stretches of nucleotides.
   d. occur where there are long repeats.
   e. are very rare in genes that code for proteins.
7. Newborn genetic screening for PKU
   a. is very expensive.
   b. detects phenylketones in urine.
   c. has not led to the prevention of mental retardation resulting from this disorder.
   d. must be done during the first day of an infant's life.
   e. uses bacterial growth to detect excess phenylalanine in blood.
8. Genetic diagnosis by DNA testing
   a. detects only mutant and not normal alleles.
   b. can be done only on eggs or sperm.
   c. involves hybridization to rRNA.
   d. utilizes restriction enzymes and a polymorphic site.
   e. cannot be done with PCR.
9. Most human cancers
   a. are caused by viruses.
   b. are in blood cells or their precursors.
   c. involve mutations of somatic cells.
   d. spread through solid tissues rather than by the blood or lymphatic system.
   e. are inherited.
10. Current treatments for genetic diseases include all of the following *except*
    a. restricting a dietary substrate.
    b. replacing the mutated gene in all cells.
    c. alleviating the patient's symptoms.
    d. inhibiting the function of a harmful metabolite.
    e. supplying a protein that is missing.

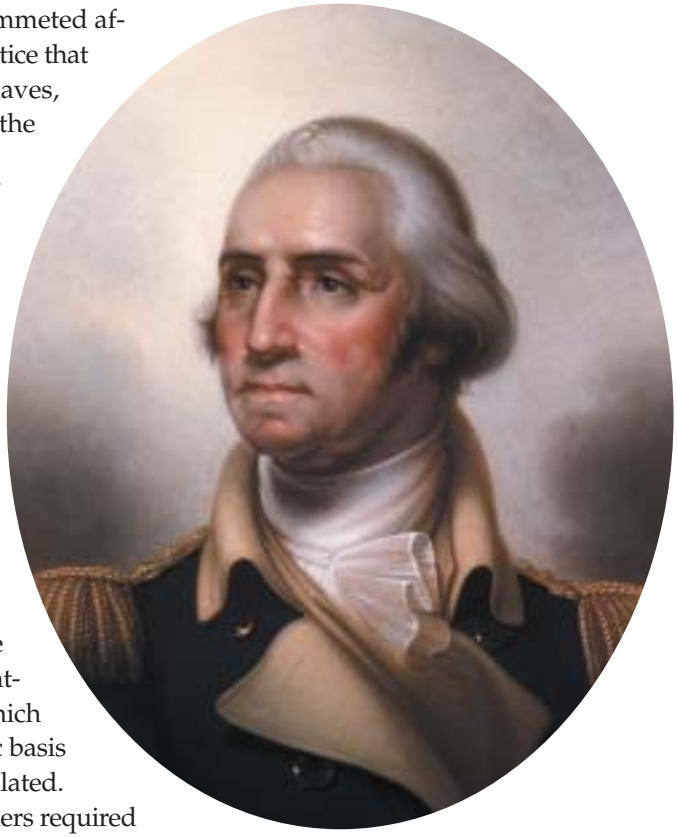
For Discussion

1. How do oncogenes and tumor suppressor genes and their functions change in tumor cells? Propose targets for cancer therapy involving these gene products.
2. In the past, it was common for people with phenylketonuria (PKU) who were placed on a low-phenylalanine diet after birth to be allowed to return to a normal diet during their teenage years. Although the levels of phenylalanine in their blood were high, their brains were thought to be beyond the stage of being harmed. If a woman with PKU becomes pregnant, however, a problem arises. Typically, the fetus is heterozygous but is unable at early stages of development to metabolize the high levels of phenylalanine that arrive from the mother's blood. Why is the fetus heterozygous? What do you think would happen to the fetus during this "maternal PKU" situation? What would be your advice to a woman with PKU who wants to have a child?
3. Cystic fibrosis is an autosomal recessive disease in which thick mucus is produced in the lungs and airways. The gene responsible for this disease codes for a protein composed of 1,480 amino acids. In most patients with cystic fibrosis, the protein has 1,479 amino acids: A phenylalanine is missing at position 508. A baby is born with cystic fibrosis. He has an older brother. How would you test the DNA of the brother to determine if he is a carrier for cystic fibrosis? How would you design a gene therapy protocol to "cure" the cells in the lung and airway?
4. A number of efforts are underway to identify human genetic polymorphisms that correlate with complex diseases such as diabetes, heart disease and cancer. What would be the uses of such information? What concerns do you think are being raised by the people whose DNAs are being analyzed?

15 Natural Defenses against Disease

On January 6, 1777, George Washington, commander of the Revolutionary army of the fledgling United States, made a fateful decision. As he wrote to his chief physician, "Finding smallpox to be spreading much, and fearing that no precaution can prevent it from running through the whole of our army, I have determined that the troops shall be inoculated. Should the disease rage with its usual virulence, we should have more to dread from it than the sword of the enemy."

Washington was speaking from experience. He himself had survived the disease in 1751, when he was still a teenager. During 1776 his army lost 1,000 men in battle and 10,000 men to smallpox. This virulent disease, which killed up to 1 of every 4 people exposed to it, had already figured prominently in American history. A century before, it had decimated the native population, making colonization by Europeans easier. Two years previously at Quebec, it had laid waste to an American army that was trying to annex Canada by force.



George Washington Washington's decision to immunize his army against smallpox saved many lives and probably helped him win the Revolutionary War.

The death rate due to smallpox in the Revolutionary army plummeted after Washington's order was carried out. How did inoculation, a practice that was learned from the people of the Near East and from African slaves, save the soldiers? And why was Washington himself immune to the disease as it ravaged his army?

The answers to these questions lie in the cells and molecules of the immune system. When Washington caught smallpox in 1751, specialized white blood cells in his body engulfed some of the smallpox viruses by phagocytosis and partly digested them. These cells, called macrophages, displayed fragments of the viruses on their surfaces. Other specialized white blood cells, called T cells, recognized those fragments, which caused them to divide and differentiate. Some descendants of those activated T cells then attacked Washington's virus-infected cells, preventing the lethal spread of the disease. Other descendants of the T cells persisted in his body as "memory cells" and rapidly divided again to defend him when he was exposed to the disease as an adult. Inoculation of Washington's soldiers with powdered scabs from smallpox patients, containing dead smallpox viruses, stimulated the formation of these memory cells in their bodies, once again preventing the virus from spreading following infection. This practice, which had been used for centuries, was finally placed on a more scientific basis by Edward Jenner two decades after Washington's army was inoculated.

These defensive events in the bodies of Washington and his soldiers required the participation of many kinds of cells and proteins. This chapter begins by intro-

ducing the participants in the two types of defense mechanisms found in vertebrate animals. Then we look in greater detail at the nonspecific defense mechanisms. Next we see how the specific defense mechanisms—the immune system—target specific invaders, such as the smallpox virus, for destruction, and how the reshuffling of their genetic material allows them to target an incredible diversity of potential invaders. Finally we look at what happens when this complex system malfunctions.

Animal Defense Systems

Animals have a number of ways of defending themselves against **pathogens**—harmful organisms and viruses that can cause disease. These defense systems are based on the distinction between *self*—the animal's own molecules—and *nonself*, or foreign, molecules. In this section we consider the mechanisms by which animals recognize nonself molecules and combat infection and disease. Many of these mechanisms are based on the principles of genetics and molecular biology that have been discussed in earlier chapters.

In general, there are two types of defense mechanisms:

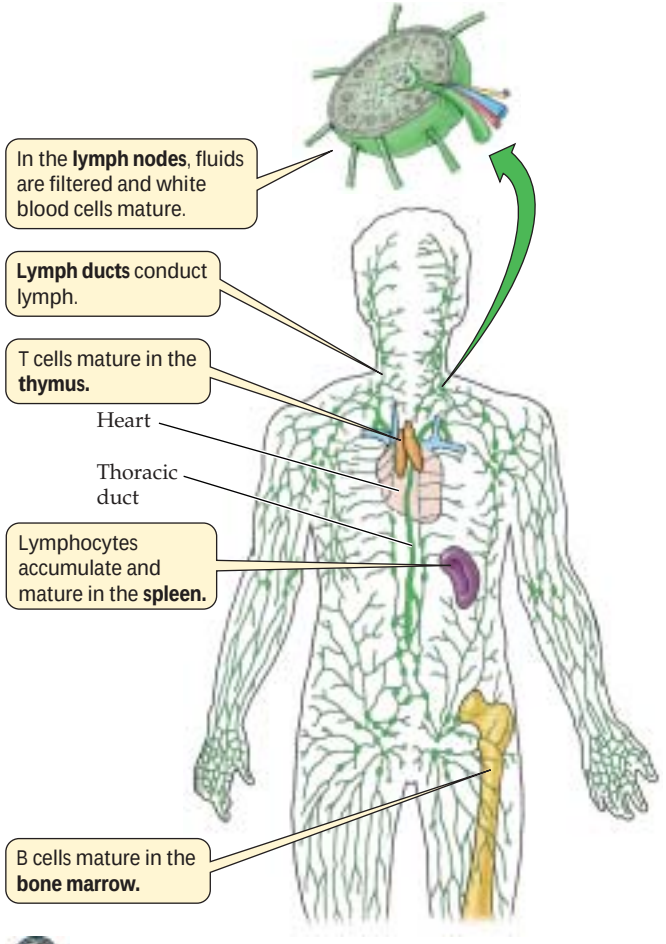
- **Nonspecific defenses**, or innate defenses, are inherited mechanisms that protect the body from many pathogens. An example is the skin, which acts as a barrier to stop potentially invading viruses from entering the body. Most animals and plants have innate defenses.
- **Specific defenses** are adaptive mechanisms aimed at a specific target. For example, these defense systems can make an antibody protein that will recognize, bind to, and destroy a certain virus if that virus ever enters the bloodstream. Specific defense mechanisms are present in vertebrate animals. DNA rearrangements and mutations play important roles in generating these defenses against a huge variety of targets.

In animals that have both kinds of mechanisms, nonspecific and specific defenses operate together as a coordinated defense system.

Blood and lymph tissues play important roles in defense systems

The components of the mammalian defense system are dispersed throughout the body and interact with almost all of its other tissues and organs. The **lymphoid tissues**, which include the thymus, bone marrow, spleen, and lymph nodes, are essential parts of the defense system (Figure 18.1), but central to their functioning are the blood and lymph.

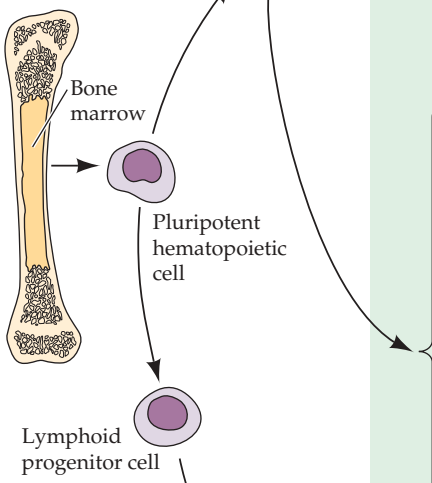
Blood and lymph are both fluid tissues that consist of water, dissolved solutes, and cells. *Blood plasma* is a yellowish solution containing ions, small molecular solutes, and soluble proteins. Suspended in the plasma are red blood cells, white blood cells, and platelets (cell fragments essential to clotting) (Figure 18.2). While red blood cells are normally confined to the *closed circulatory system* (the heart, arteries, capillaries, and veins), white blood cells and platelets are also found in the lymph.



18.1 The Human Lymphatic System A network of ducts and vessels collects lymph from the body's tissues and carries it toward the heart, where it mixes with blood to be pumped back to the tissues. The thymus, spleen, and bone marrow are also essential to the body's defense system.

Lymph is a fluid derived from blood and other tissues that accumulates in intercellular spaces throughout the body. From these spaces, the lymph moves slowly into the vessels of the *lymphatic system*. Tiny lymph capillaries conduct this fluid to larger vessels that eventually join together, forming one large vessel, the thoracic duct, which joins a major vein (the left subclavian vein) near the heart. By this system of vessels, the lymph is eventually returned to the blood and the circulatory system.

At many sites along the lymph vessels are small, roundish structures called **lymph nodes**, which contain a variety of white blood cells. As fluid passes through a node, it is filtered and "inspected" for nonself materials by these defensive cells.



TYPE OF CELL	FUNCTION
Red blood cells (erythrocytes)	Transport oxygen and carbon dioxide
Platelets (cell fragments without nuclei)	Initiate blood clotting
White blood cells (leukocytes)	
PHAGOCYTES	
Basophils	Release histamine; may promote the development of T cells
Eosinophils	Kill antibody-coated parasites
Neutrophils	Phagocytose antibody-coated pathogens
Mast cells	Release histamine when damaged
Monocytes	Develop into macrophages
Macrophages	Engulf and digest microorganisms; activate T cells
Dendritic cells	Present antigens to T cells
LYMPHOCYTES	
B cells	Differentiate to form antibody-producing cells and memory cells
Plasma cells	Secrete antibodies
T cells	Kill virus-infected cells; regulate activities of other white blood cells
Natural killer cells	Attack and lyse virus-infected or cancerous body cells

18.2 Blood Cells Pluripotent stem cells in the bone marrow can differentiate into red blood cells, platelets, and the various types of white blood cells.

White blood cells play many defensive roles

One milliliter of blood typically contains about 5 billion red blood cells and 7 million of the larger white blood cells. All of these cells originate from stem cells in the bone marrow (see Figure 18.2). White blood cells (also called *leukocytes*) have nuclei and are colorless, unlike mammalian red blood cells, which lose their nuclei during development. White blood cells can leave the closed circulatory system and enter intercellular spaces where foreign cells or substances are present. The number of white blood cells in the blood and lymph may rise sharply in response to invading pathogens, providing medical professionals with a useful clue for detecting an infection.

Several types of white blood cells are important in the body's defenses. But they are all members of two broad groups, phagocytes and lymphocytes.

- **Phagocytes** engulf and digest nonself materials. Among the most important phagocytes are the **macrophages**. In addition to engulfing nonself materials by phagocytosis, macrophages have the important additional function of presenting partly digested nonself materials to the T cells.
- **Lymphocytes** participate in specific defenses against nonself or altered cells, such as virus-infected cells and tumor cells. There are two types of lymphocytes, **B cells** and **T cells**. Immature T cells migrate from the bone marrow via the blood to the *thymus*, where they mature. The B cells leave the bone marrow and circulate through the blood and lymph vessels. B cells make specialized proteins called antibodies that enter the blood and bind to nonself substances.

Fundamental to the interactions, control, and defensive functioning of these white blood cells are defensive proteins and other signals.

Immune system proteins bind pathogens or signal other cells

The cells that defend mammalian bodies work together like cast members in a drama, interacting with one another and with the cells of invading pathogens. These cell–cell interactions are accomplished by a variety of key proteins, including receptors, other cell surface proteins, signaling molecules, and toxins. These proteins will be discussed later in the chapter, as they appear in the context of our story. However, let's take a brief look at four of the major players here.

- **Antibodies** are proteins that bind specifically to certain substances identified by the immune system as nonself or altered self, thereby denaturing the invading nonself substance. They are secreted by B cells as defensive weapons.
- **T cell receptors** are integral membrane proteins on the surfaces of T cells. They recognize and bind to nonself substances on the surfaces of other cells.
- **Major histocompatibility complex** (**MHC**) proteins protrude from the surfaces of most cells in the mammalian body. They are important self-identifying labels and play major parts in coordinating interactions among lymphocytes and macrophages.
- **Cytokines** are soluble signal proteins released by T cells, macrophages, and other cells. They bind to and alter the behavior of their target cells. Different cytokines activate or inactivate B cells, macrophages, and T cells. Some cytokines limit tumor growth by killing tumor cells.

Before focusing on the roles of these four major players in specific defenses that constitute the immune system, we will consider the nonspecific defenses, which involve both passive barriers and active roles for molecules and cells.

Nonspecific Defenses

Nonspecific defenses (also called innate defenses) are general protection mechanisms that attempt to stop pathogens from invading the body. Nonspecific defenses in humans include physical barriers as well as cellular and chemical defenses (Table 18.1).

18.1 *Human Nonspecific Defenses*

DEFENSIVE AGENT	FUNCTION
Surface barriers	
Skin	Prevents entry of pathogens and foreign substances
Acid secretions	Inhibit bacterial growth on skin
Mucous membranes	Prevent entry of pathogens
Mucous secretions	Trap bacteria and other pathogens in digestive and respiratory tracts
Nasal hairs	Filter bacteria in nasal passages
Cilia	Move mucus and trapped materials away from respiratory passages
Gastric juice	Concentrated HCl and proteases destroy pathogens in stomach
Acid in vagina	Limits growth of fungi and bacteria in female reproductive tract
Tears, saliva	Lubricate and cleanse; contain lysozyme, which destroys bacteria
Nonspecific cellular, chemical, and coordinated defenses	
Normal flora	Compete with pathogens; may produce substances toxic to pathogens
Fever	Body-wide response inhibits microbial multiplication and speeds body repair processes
Coughing, sneezing	Expels pathogens from upper respiratory passages
Inflammatory response (involves leakage of blood plasma and phagocytes from capillaries)	Limits spread of pathogens to neighboring tissues; concentrates defenses; digests pathogens and dead tissue cells; released chemical mediators attract phagocytes and specific defense lymphocytes to site
Phagocytes (macrophages and neutrophils)	Engulf and destroy pathogens that enter body
Natural killer cells	Attack and lyse virus-infected or cancerous body cells
Antimicrobial proteins	
Interferons	Released by virus-infected cells to protect healthy tissue from viral infection; mobilize specific defenses
Complement proteins	Lyse microorganisms, enhance phagocytosis, and assist in inflammatory response

Barriers and local agents defend the body against invaders

Skin is a primary nonspecific defense against invasion. Fungi, bacteria, and viruses rarely penetrate healthy, unbroken skin. But damage to the skin or to the internal surface tissue greatly increases the risk of infection by pathogens.

The bacteria and fungi that normally live and reproduce in great numbers on our body surfaces without causing disease are referred to as *normal flora*. These natural occupants of our bodies compete with pathogens for space and nutrients and are thus a form of nonspecific defense.

The mucous membranes found at the surfaces of the visual, respiratory, digestive, excretory, and reproductive systems have other defenses against pathogens. Tears, nasal mucus, and saliva contain an enzyme called *lysozyme* that attacks the cell walls of many bacteria. Mucus in the nose traps airborne microorganisms, and most of those that get past this filter end up trapped in mucus deeper in the respiratory tract. Mucus and trapped pathogens are removed by the beating of cilia in the respiratory passageway, which continuously move a sheet of mucus and the debris it contains up toward the nose and mouth. Sneezing is another way to remove microorganisms from the respiratory tract.

Pathogens that reach the digestive tract (stomach, small intestine, and large intestine) are met by other defenses. The gastric juice in the stomach is a deadly environment for many bacteria because of the hydrochloric acid and proteases (protein-digesting enzymes) that are secreted into it. The intact lining of the small intestine is not normally penetrated by bacteria, and some pathogens are killed by bile salts secreted into this part of the digestive tract. The large intestine harbors many bacteria, which multiply freely; however, they are usually removed quickly with the feces. Most of the bacteria in the large intestine are normal flora that provide benefits to their host. We probably add to this beneficial flora when we eat foods such as active-culture yogurt and various cheeses.

All of these barriers and local agents are *nonspecific* defenses because they act on all invading pathogens in the same way. More complex nonspecific defenses await any pathogens that manage to elude this first line of defense.

Nonspecific defenses include chemical and cellular processes

Pathogens that penetrate the body's outer and inner surfaces encounter more complex nonspecific defenses that involve the secretion of various defensive proteins as well phagocytic cells.

COMPLEMENT PROTEINS. Vertebrate blood contains about 20 different antimicrobial proteins that make up the **complement system**. These proteins, in different combinations, provide three types of defenses. In each type, the complement proteins act in a characteristic sequence, or cascade, with each protein activating the next:

- They attach to microbes, which helps phagocytes recognize and destroy them.
- They activate the inflammation response and attract phagocytes to site of infection.
- They lyse (burst) invading cells such as bacteria.

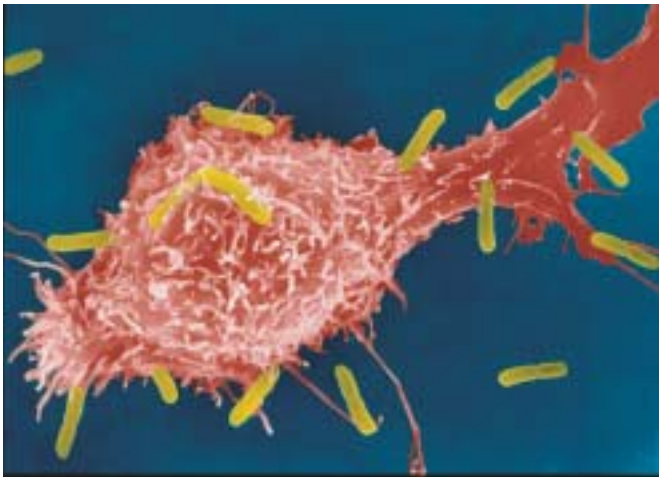
INTERFERONS. When cells are infected by a virus, they produce small amounts of antimicrobial proteins called **interferons** that increase the resistance of neighboring cells to infection by the same *or other* viruses. Interferons have been found in many vertebrates and are one of the body's first lines of nonspecific defense against the internal spread of viral infection.

Interferons differ from species to species, and each vertebrate species produces at least three different interferons. All interferons are glycoproteins (proteins with attached carbohydrate groups) consisting of about 160 amino acids. By binding to receptors in the plasma membranes of uninfected cells, interferons stimulate a signaling pathway that results in the inhibition of viral reproduction inside the infected cells.

PHAGOCYTES AND RELATED CELLULAR DEFENSES. Phagocytes provide another important nonspecific defense against pathogens that penetrate the surface of the host. Some phagocytes travel freely in the circulatory system; others can move out of blood vessels and adhere to certain tissues. Entire pathogenic cells, entire viruses, or fragments of these invaders can become attached to the membrane of a phagocyte (Figure 18.3), which ingests them by phagocytosis. When lysosomes fuse with the phagosome, the pathogens are degraded by lysosomal enzymes (see Figure 4.13*b*).

Several types of phagocytes play roles in nonspecific defenses (see Figure 18.2):

- **Neutrophils** are the most abundant phagocytes, but they are relatively short-lived. They recognize and attack pathogens in infected tissue.
- **Monocytes** mature into macrophages, which live longer than neutrophils and can consume large numbers of pathogens. Some macrophages roam through the body; others reside permanently in lymph nodes, the spleen, and certain other lymphoid tissues, "inspecting" the lymph for pathogens.
- **Eosinophils** are weakly phagocytic. Their primary function is to kill parasites, such as worms, that have been coated with antibodies.
- **Dendritic cells** have highly folded plasma membranes that can capture invading pathogens.



18.3 A Phagocyte and Its Bacterial Prey Some bacteria (which appear yellow in this artificially colored scanning electron micrograph) have become attached to the surface of a phagocyte in the human bloodstream. Many of these bacteria will be engulfed by the phagocyte and destroyed before they can multiply and damage the human host. A single phagocyte can digest many bacteria.

NONPHAGOCYTIC CELLS. A class of nonphagocytic white blood cells, known as **natural killer cells**, can distinguish virus-infected cells and some tumor cells from their normal counterparts and initiate the lysis of these target cells. In addition to this nonspecific action, natural killer cells form part of the specific defenses, as we will describe later in this chapter.

INFLAMMATION. The body employs the **inflammation** response in dealing with infection or with any other process that causes tissue injury, either on the surface of the body or internally. The damaged body cells cause the inflammation by releasing various substances. Cells adhering to the skin and linings of organs, called **mast cells**, release a chemical signal, called **histamine**, when they are damaged, as do white blood cells called **basophils**.

You have no doubt experienced the symptoms of inflammation: redness and swelling, accompanied by heat and pain. The redness and heat of inflammation result from histamine-induced dilation of blood vessels in the infected or injured area (Figure 18.4). Histamine also causes the capillaries (the smallest blood vessels) to become leaky, allowing blood plasma and phagocytes to escape into the tissue, causing the characteristic swelling. The pain of inflammation results from increased pressure (from the swelling) and from the action of leaked enzymes.

In damaged or infected tissue, complement proteins and other chemical signals attract phagocytes—neutrophils first, and then monocytes, which become macrophages. The macrophages, which engulf the invaders and debris, are responsible for most of the healing associated with inflammation. They produce several cytokines, which, among other functions, signal the brain to produce a fever. This rise in

18.4 Interactions of Cells and Chemical Signals in Inflammation The histamine-induced swelling of the inflammation reaction is accompanied by redness, heat, and pain. The chemical signals associated with the inflammation reaction attract the phagocytes that clear up the pathogens and damaged cells.

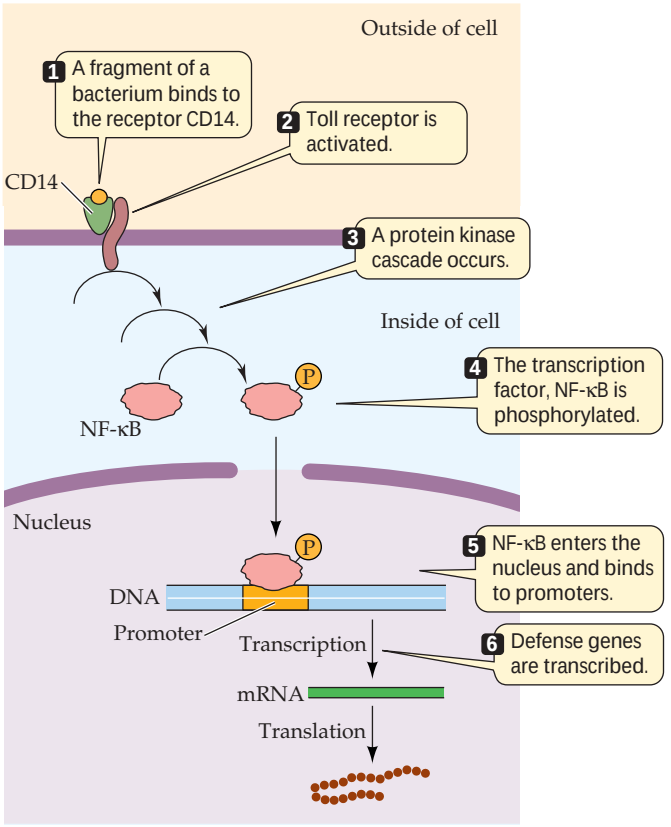
body temperature inhibits the growth of the invading pathogen. Cytokines may also attract phagocytic cells to the site of injury and initiate a specific response to the pathogen.

Following inflammation, *pus* may accumulate. It is composed of dead cells (neutrophils and the damaged body cells) and leaked fluid. A normal result of inflammation, pus is gradually consumed and digested by macrophages.

A cell signaling pathway stimulates defense

An invading pathogen such as a bacterium can be regarded as a signal. In response to that signal, the body produces molecules such as complement proteins, interferons, and cytokines that regulate phagocytosis and other defense processes. Not surprisingly, the link between signal and response is a signal transduction pathway, similar to the ones we considered in Chapter 15. The receptor in this pathway is a membrane protein called toll. This receptor was originally discovered in fruit flies, in which it plays an essential role in sensing infection by fungi. Comparative genomics has revealed at least ten similar receptors in humans.

Toll is part of a protein kinase cascade that ultimately results in the transcription of at least 40 genes involved in both nonspecific and specific defenses (Figure 18.5). The molecules that set off this pathway are only made by microbes, and include some bacterial and fungal cell wall fragments. Binding of these molecules to toll sets in motion a cascade that results in the phosphorylation of the transcription factor NF-κB. As a result, the transcription factor's conformation changes, allowing it to enter the nucleus, bind to the promoters of genes encoding defensive proteins, and activate their transcription.



18.5 Cell Signaling and Defense Binding of a molecule from a pathogen to a receptor initiates a signal transduction pathway that results in the transcription of genes whose products are involved in defense against the pathogen.

Specific Defenses: The Immune System

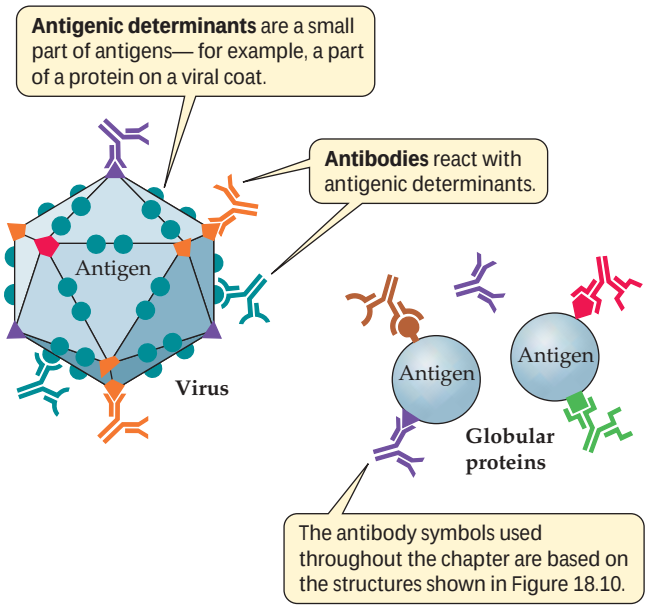
Nonspecific defenses are numerous and effective, but some invaders elude them. Vertebrate animals deal with these pathogens by means of defenses targeted against specific threats. The recognition and destruction of specific nonself substances is an important function of an animal's immune system. In this section, we will first provide an overview of the main features of the immune response. We will then consider its two components: the humoral immune response, which produces antibodies, and the cellular immune response, which destroys infected cells.

Four features characterize the immune system

The characteristic features of the immune system are specificity, the ability to respond to an enormous diversity of foreign molecules and organisms, the ability to distinguish self from nonself, and immunological memory.

SPECIFICITY. As we saw above, the lymphocytes (B cells and T cells) are involved in specific defense mechanisms. T cell receptors and the antibodies produced by B cells recognize and bind to specific nonself substances. Organisms or molecules that are recognized by and interact with these cells to initiate an immune response are called **antigens**. The specific sites on antigens that the immune system recognizes are called **antigenic determinants** or *epitopes* (Figure 18.6). Chemically, an antigenic determinant is a specific portion of a large molecule, such as a certain sequence of amino acids that may be present in several proteins. A large antigen, such as a whole cell, may have many different antigenic determinants on its surface, each capable of being bound by a specific antibody or T cell. Even a single protein has multiple, different antigenic determinants. The host animal responds to the presence of an antigen by producing highly specific defenses—T cells or antibodies that are complementary to, or fit, the antigenic determinants of that antigen. Each T cell and each antibody is specific for a single antigenic determinant.

DIVERSITY. Challenges to the immune system are numerous: individual foreign molecules, viruses, bacteria, protists, and multicellular parasites. Each of these types of



18.6 Each Antibody Matches an Antigenic Determinant Each antigen has many different antigenic determinants that are recognized by specific antibodies. Each antibody recognizes and binds to its particular antigenic determinant to initiate defensive measures against the antigen.

potential pathogens includes many species, each species includes many subtly differing genetic strains, and each strain possesses multiple surface features. Estimates vary, but a reasonable guess is that humans can respond *specifically* to 10 million different antigenic determinants. Upon recognition of an antigenic determinant, the immune system responds by activating lymphocytes (B cells and T cells) of the appropriate specificity.

DISTINGUISHING SELF FROM NONSELF. The human body contains tens of thousands of different proteins, each with a specific three-dimensional structure capable of generating an immune response. Every cell in the body bears a tremendous number of antigenic determinants. A crucial requirement of an individual's immune system is that it recognize the body's own antigenic determinants and not attack them.

IMMUNOLOGICAL MEMORY. After responding to a particular type of pathogen once, the immune system "remembers" that pathogen and can usually respond more rapidly and powerfully to the same threat in the future. This **immunological memory** usually saves us from repeats of childhood diseases such as chicken pox. Vaccination and inoculation against disease work because the immune system "remembers" the antigenic determinants that are introduced into the body.

These four features of the immune response are seen in both components of the immune system, the humoral response and the cellular response.

There are two interactive immune responses

The immune system has two responses against invaders: the humoral immune response and the cellular immune response. These two responses operate in concert—simultaneously and cooperatively, sharing mechanisms.

HUMORAL IMMUNE RESPONSE. In the **humoral immune response** (from the Latin *humor*, "fluid"), antibodies react with antigenic determinants on pathogens in blood, lymph, and tissue fluids. An animal produces such a diversity of antibodies that between them, they can react with almost any conceivable antigen the animal encounters.

Some antibodies are soluble and travel free in the blood and lymph; others exist as integral membrane proteins on B cells. The first time a specific antigen invades the body, it may be detected and bound by a B cell whose membrane antibody recognizes one of its antigenic determinants. This binding activates the B cell, which makes multiple soluble copies of an antibody with the same specificity as its membrane antibody.

CELLULAR IMMUNE RESPONSE. The **cellular immune response** is directed against an antigen that has become established within a cell of the host animal. It detects and destroys virus-infected or mutated cells.

The cellular immune response is carried out by T cells within the lymph nodes, the bloodstream, and the intercellular spaces. These cells have integral membrane proteins—T cell receptors—that recognize and bind to antigenic determinants while remaining part of the cell's plasma membrane. T cell receptors are rather similar to antibodies in structure and function, each including specific molecular configurations that bind to specific antigenic determinants. Once a T cell is bound to an antigenic determinant, it initiates an immune response that typically results in the total destruction of a nonself or altered self cell.

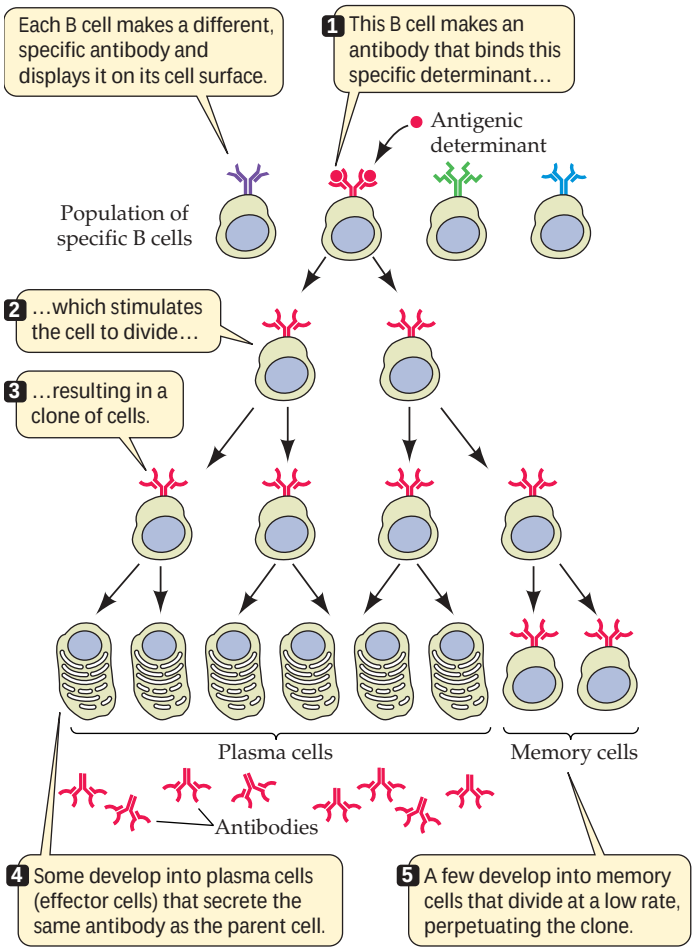
Genetic processes and clonal selection generate the characteristics of the immune response

Each person possesses an enormous number of different B cells and T cells, apparently capable of dealing with almost any antigenic determinant they are ever likely to encounter. How does this diversity arise? How do lymphocytes specific for certain antigens proliferate? And why don't our antibodies and T cells attack and destroy our own bodies? The diversity of the immune response, the proliferation of specific cells, the ability to distinguish between self and nonself, and immunological memory can all be explained by the process of **clonal selection** and the unique DNA rearrangements upon which it is based.

As we have seen, each individual human contains an enormous variety of different B cells and T cells. This diversity is

generated primarily by DNA changes—chromosomal rearrangements and mutations—that occur just after the cells are formed in the bone marrow. Each B cell is able to produce *only one kind of antibody*. Thus there are millions of different B cells, each one producing a particular antibody and displaying it on its cell surface. When an antigen that fits this surface antibody binds to it, the B cell is activated. It divides to form a clone of cells (a genetically identical group derived from a single cell), all of them producing that particular antibody. Thus the antigen "selects" a particular B cell by binding its specific antibody and signaling it to proliferate (Figure 18.7). In the same way, a foreign or abnormal cell "selects" for the proliferation of a T cell expressing a particular T cell receptor on its surface.

Clonal selection accounts nicely for the body's ability to respond rapidly to any of a vast number of different antigens. In the extreme case, even a single B cell might be sufficient for an immunological response, provided that it encounters its antigen and then proliferates into a large clone rapidly enough to combat the invasion.



18.7 Clonal Selection in B Cells The binding of an antigenic determinant to a specific antibody on the surface of a B cell stimulates the cell to divide, rapidly producing a clone of cells to fight the invader.

Immunity and immunological memory result from clonal selection

An activated lymphocyte produces two types of daughter cells, effector cells and memory cells.

- **Effector cells** carry out the attack on the antigen. Effector B cells, called **plasma cells**, produce antibodies. Effector T cells release cytokines, which initiate reactions that destroy nonself or altered cells. Effector cells live only a few days.
- **Memory cells** are long-lived cells that retain the ability to start dividing on short notice to produce more effector and more memory cells. Memory B and possibly T cells may survive in the body for decades, dividing at a low rate.

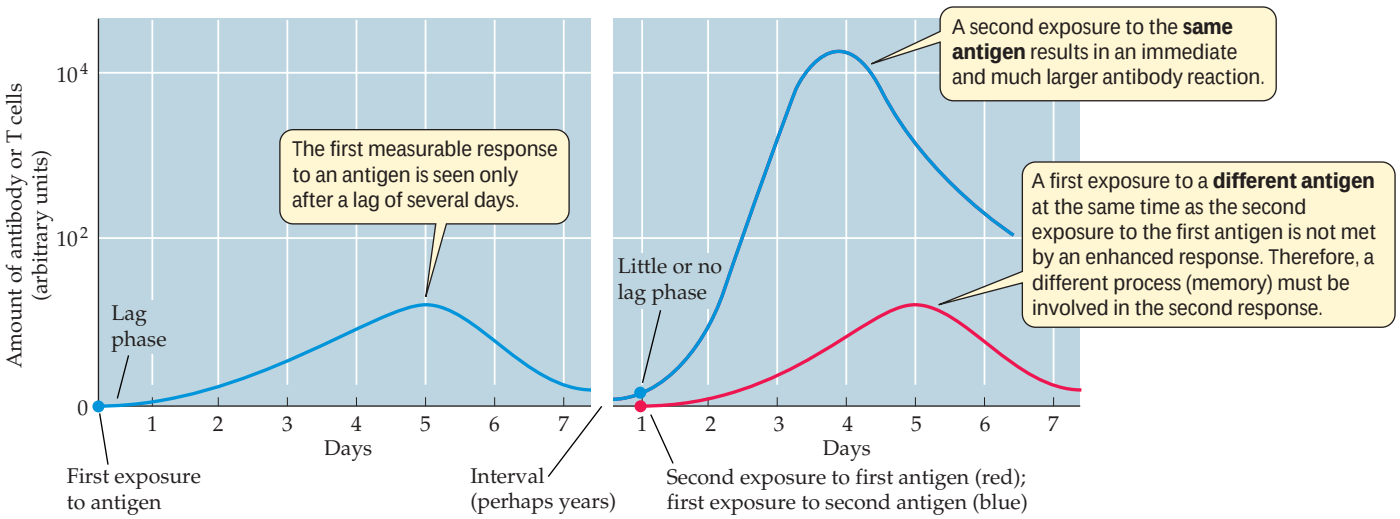
When the body first encounters a particular antigen, a *primary immune response* is activated, in which the lymphocytes that recognize that antigen produce clones of effector and memory cells. The effector cells destroy the invaders at hand and then die, but one or more clones of memory cells have now been added to the immune system and provide immunological memory.

After the body's first immune response to a particular antigen, subsequent encounters with the same antigen will trigger a much more powerful attack. The huge army of plasma and T cells launched by the memory cells at this time is called the *secondary immune response*. The first time a vertebrate animal is exposed to a particular antigen, there is a time lag (usually several days) before the number of antibody molecules and T cells slowly increases (Figure 18.8). But for years afterward—sometimes for life—the immune system "remembers" that particular antigen. The secondary immune response is characterized by a shorter lag time, a greater rate of antibody production, and a larger production of total antibody or T cells than the primary response.

Vaccines are an application of immunological memory

Thanks to immunological memory, recovery from many diseases, such as chicken pox, provides a *natural immunity* to those diseases. However, it is possible to provide *artificial immunity* against many life-threatening diseases by *inoculation*—the introduction of antigenic determinants into the body. **Immunization** is inoculation with antigenic proteins, pathogen fragments, or other molecular antigens. **Vaccination** is inoculation with whole pathogens that have been modified so that they cannot cause disease.

Immunization or vaccination initiates a primary immune response, generating memory cells without making the person ill. Later, if the same or very similar pathogens attack, specific memory cells already exist. They recognize the antigen and quickly overwhelm the invaders with a massive production of lymphocytes and antibodies.



18.8 Immunological Memory The ability of the body to remember an antigen to which it has been exposed is the basis for natural and artificial immunity against a disease.

Because the antigens used for immunization or vaccination are either themselves toxic or are parts of a pathogenic organism, they must be used in a form that is inactive in causing a disease but active in provoking an immune response. There are three principal ways to do this:

- *Attenuation* involves either treating the antigenic molecule or organism with a chemical (for example, formalin) or repeatedly infecting cells with it in the laboratory until its toxicity is reduced.
- *Biotechnology* can be used to produce peptide fragments that activate lymphocytes but do not have the harmful part of a protein toxin.
- *DNA vaccines* are being developed that will introduce a gene encoding an antigen into the body.

For most of the 70 bacteria, viruses, fungi, and parasites that cause serious human diseases, vaccines are already available or will be in the next few years (Table 18.2). Vaccination has almost completely wiped out some deadly diseases, such as diphtheria and polio, in industrialized countries. In fact, smallpox has been eliminated worldwide, thanks to an international effort by the World Health Organization. As far as we know, the only remaining smallpox viruses on Earth are those kept in some laboratories. But there are fears that the smallpox virus and other pathogens, some of which do not have readily available vaccines, may be in the hands of terrorists.

18.2 Some Vaccines against Human Pathogens

INFECTIOUS AGENT	DISEASE	VACCINATED POPULATION
Bacteria		
Bacillus anthracis	Anthrax	Exposed in biological warfare
Bordetella pertussis	Whooping cough	Children and adults
Clostridium tetani	Tetanus	Children
Corynebacterium diphtheriae	Diphtheria	Children
Haemophilus influenzae	Meningitis	Children
Mycobacterium tuberculosis	Tuberculosis	All people
Salmonella typhi	Typhoid fever	Areas exposed to agent
Streptococcus pneumoniae	Pneumonia	Elderly
Vibrio cholerae	Cholera	People in areas exposed to agent
Viruses		
Adenovirus	Respiratory disease	Military personnel
Hepatitis A	Liver disease	Areas exposed to agent
Hepatitis B	Liver disease, cancer	All people
Influenza virus	Flu	All people
Measles virus	Measles	Children and adolescents
Mumps virus	Mumps	Children and adolescents
Poliovirus	Polio	Children
Rabies virus	Rabies	Exposed to agent
Rubella virus	German measles	Children
Vaccinia virus	Smallpox	Laboratory workers, military personnel
Varicella-zoster virus	Chicken pox	Children

Animals distinguish self from nonself and tolerate their own antigens

Given the presence in our bodies of lymphocytes directed against so many antigens, why don't we produce self-destructive immune responses? Sometimes we do. Failure to distinguish appropriately between self and nonself molecules can result in an *autoimmune disease*—an attack on one's own body. But in a healthy person, the body is tolerant of its own molecules—the same molecules that would generate an immune response in another individual. **Self-tolerance** seems to be based on two mechanisms: clonal deletion and clonal anergy.

CLONAL DELETION. **Clonal deletion** physically removes B or T cells from the immune system at some point during their differentiation. Immature B cells in the bone marrow, for example, may encounter self antigens. Any of these cells that shows the potential to mount an immune response against self antigens undergoes programmed cell death (apoptosis) within a short time, and never differentiates enough to make antibodies. Thus, no clones of antiself B cells normally appear in the bloodstream. Clonal deletion eliminates about 90 percent of all the B cells made in the bone marrow. A similar process occurs with T cells in the thymus.

CLONAL ANERGY. **Clonal anergy** is the suppression of the immune response to self antigens. A mature T cell, for example, may encounter and recognize a self antigen on the surface of a body cell. But it does not send out the cytokines that signal the initiation of an immune response. Before it does so, the T cell must encounter not only an antigen, but also a second molecule, CD28, on the cell surface. Most body cells, lacking CD28, will not be attacked by the cellular immune system.

CD28 is a *co-stimulatory signal* that is expressed only on certain *antigen-presenting cells*. Such cells "present" antigens on their surfaces, thus stimulating the cellular immune system. Antigen-presenting cells include the macrophages that wander through the body's fluids, and the dendritic cells that appear among the linings of the respiratory and digestive tracts.

Immunological tolerance was discovered through the observation that some *nonidentical* twin cattle with different blood types contained some of each other's red blood cells. Why didn't these "foreign" blood cells cause immune responses resulting in their elimination? The hypothesis suggested was that the blood cells had passed between the fetal animals in the womb before the lymphocytes had matured. Thus each calf regarded the other's red blood cells as self. This hypothesis was confirmed when it was shown that injecting a foreign antigen into an animal early in its fetal development caused that animal henceforth to recognize that antigen as self.

Self-tolerance must be established repeatedly throughout the life of the animal because lymphocytes are produced constantly. Continued exposure to self antigens helps maintain tolerance. For unknown reasons, tolerance to self antigens may sometimes be lost. When that happens, the body produces antibodies or T cells targeted against its own proteins, resulting in an autoimmune disease.

Having described the general features of the immune system, we will now focus in more detail on the B lymphocytes and the humoral response.

B Cells: The Humoral Immune Response

Every day, billions of B cells survive the test of clonal deletion and are released from the bone marrow into the circulation. B cells are the basis for the humoral immune response.

Some B cells develop into plasma cells

As described above, a B cell is activated by the binding of a specific antigenic determinant to the antibody protein on its surface. Normally, for such a B cell to develop into an antibody-secreting plasma cell, a **helper T cell** ($\mathbf{T_H}$) with the same specificity must also bind to the antigen. Thus, the B cell also functions as an antigen-presenting cell, as we will see below. The division and differentiation of B cells is stimulated by the receipt of chemical signals from the T_H cell. These events lead to the formation of plasma cells (effector B cells) and memory cells (see Figure 18.7).

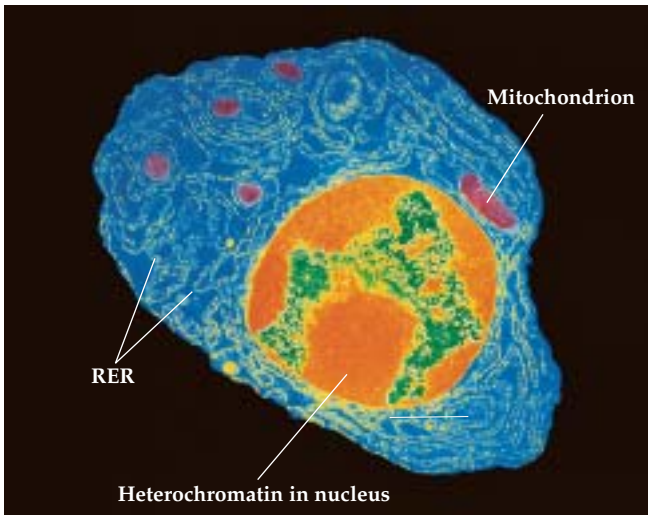
As plasma cells develop, the number of ribosomes and the amount of endoplasmic reticulum in their cytoplasm increase greatly (Figure 18.9). These increases allow the cells to synthesize and secrete large amounts of antibodies. All the plasma cells arising from a given B cell produce antibodies that are specific for the antigen that originally bound to the parent B cell. Thus antibody specificity is maintained as B cells proliferate.

Different antibodies share a common structure

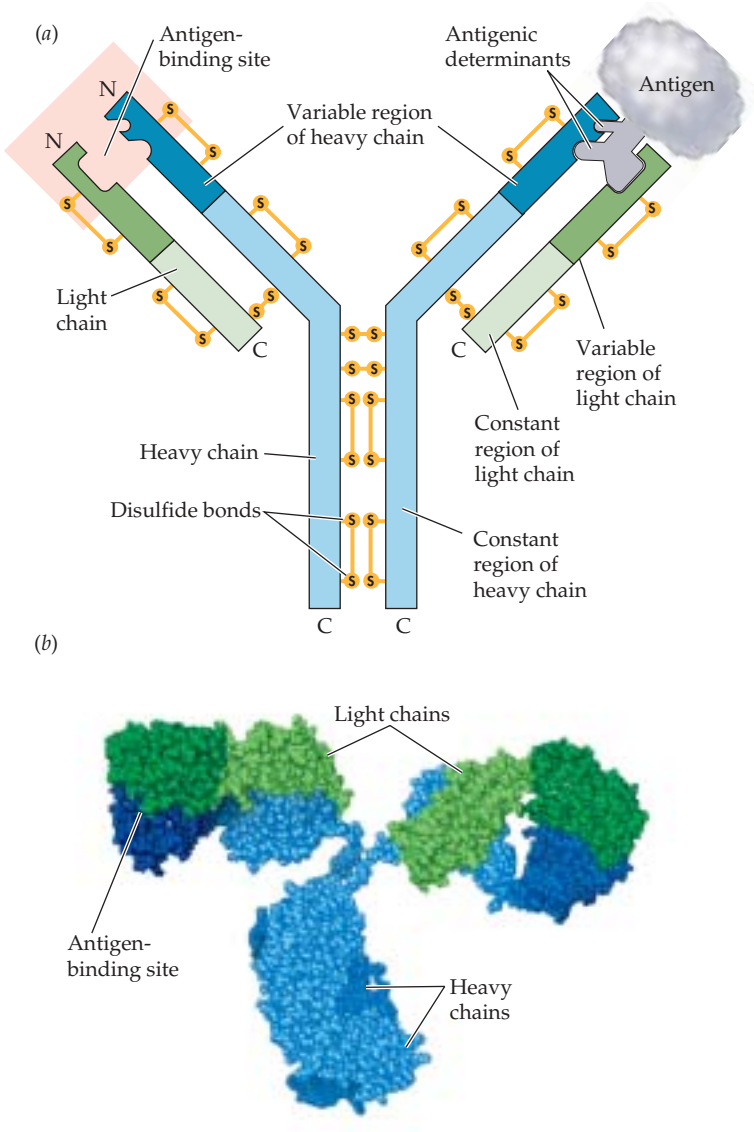
Antibodies are proteins called **immunoglobulins**. There are several types of immunoglobulins, but all contain a tetramer consisting of four polypeptide chains. In each immunoglobulin molecule, two of these polypeptides are identical *light chains*, and two are identical *heavy chains*. Disulfide bonds hold the chains together.

Each polypeptide chain consists of a constant region and a variable region (Figure 18.10).

- The **constant regions** of both light chains and heavy chains are similar in amino acid sequence among the immunoglobulins. They determine the destination and function—the *class*—of the antibody.



18.9 A Plasma Cell The prominent nucleus with large amounts of heterochromatin (orange) and the cytoplasm (bright blue) crowded with rough endoplasmic reticulum are features of a cell that is actively synthesizing and exporting proteins—in this case, a specific antibody. Whole blocks of genes not needed for this specialized function are kept turned off in the heterochromatin.



18.10 Structure of Immunoglobulins (*a*) The four polypeptide chains (two light, two heavy) of an immunoglobulin molecule. (*b*) A three-dimensional space-filling model of an antibody molecule in roughly the same orientation as (*a*). In both images, the light chains are shown in green and the heavy chains in blue; the variable regions are shown in a darker color and the constant regions in a lighter color.

▶ The **variable regions** differ in their amino acid sequences. They contribute directly to the three-dimensional region where the antigen binds—the *antigen-binding site*—and are responsible for the diversity of antibody specificity.

In the initial differentiation of each B cell in the bone marrow, the amino acid sequence of the variable region of both the light and heavy chains is chosen randomly from the DNA information in the genome. This means that the variable region is unique in each of the millions of antigen-specific immunoglobulins. Together, the variable regions of a light and a heavy chain form a highly specific, three-dimensional structure. This part of a particular immunoglobulin molecule is what binds with a particular, unique antigenic determinant. The enormous range of antibody specificities is accomplished by a combination of rearrangements and mutations in the genes that encode the variable regions, as we will see later in this chapter.

The two antigen-binding sites on each immunoglobulin molecule are identical, making the antibody *bivalent* (*bi-*, "two"; *valent*, "binding"). This ability to bind two antigen molecules at once permits the antibody to form a large complex with antigen and other antibody molecules. Such a complex is an easy target for ingestion and breakdown by phagocytic cells.

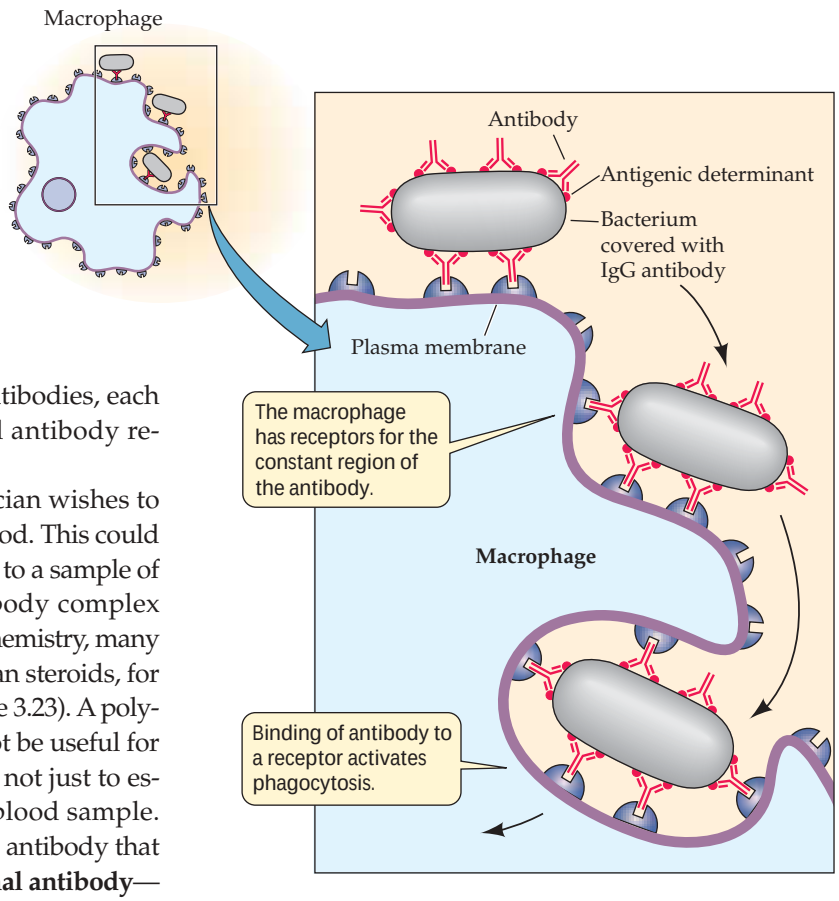
While the variable regions are responsible for the *specificity* of an immunoglobulin, the constant regions of the heavy chain determine the *class* of the antibody—for example, whether it will be a membrane receptor or a soluble antibody that is secreted into the bloodstream. The five immunoglobulin classes are described in Table 18.3. The most abundant immunoglobulin class is IgG; these soluble antibody proteins make up about 80 percent of the total immunoglobulin content of the bloodstream. They are made in greatest quantity during a secondary immune response. IgG defends the body in several ways. For example, after some IgG molecules bind to antigens, they become attached by their heavy chains to macrophages. This attachment permits the macrophages to destroy the antigens by phagocytosis (Figure 18.11).

Hybridomas produce monoclonal antibodies

The specificity of antibodies suggested to scientists that they might be useful for detecting a specific substance in a fluid.

18.3 Antibody Classes

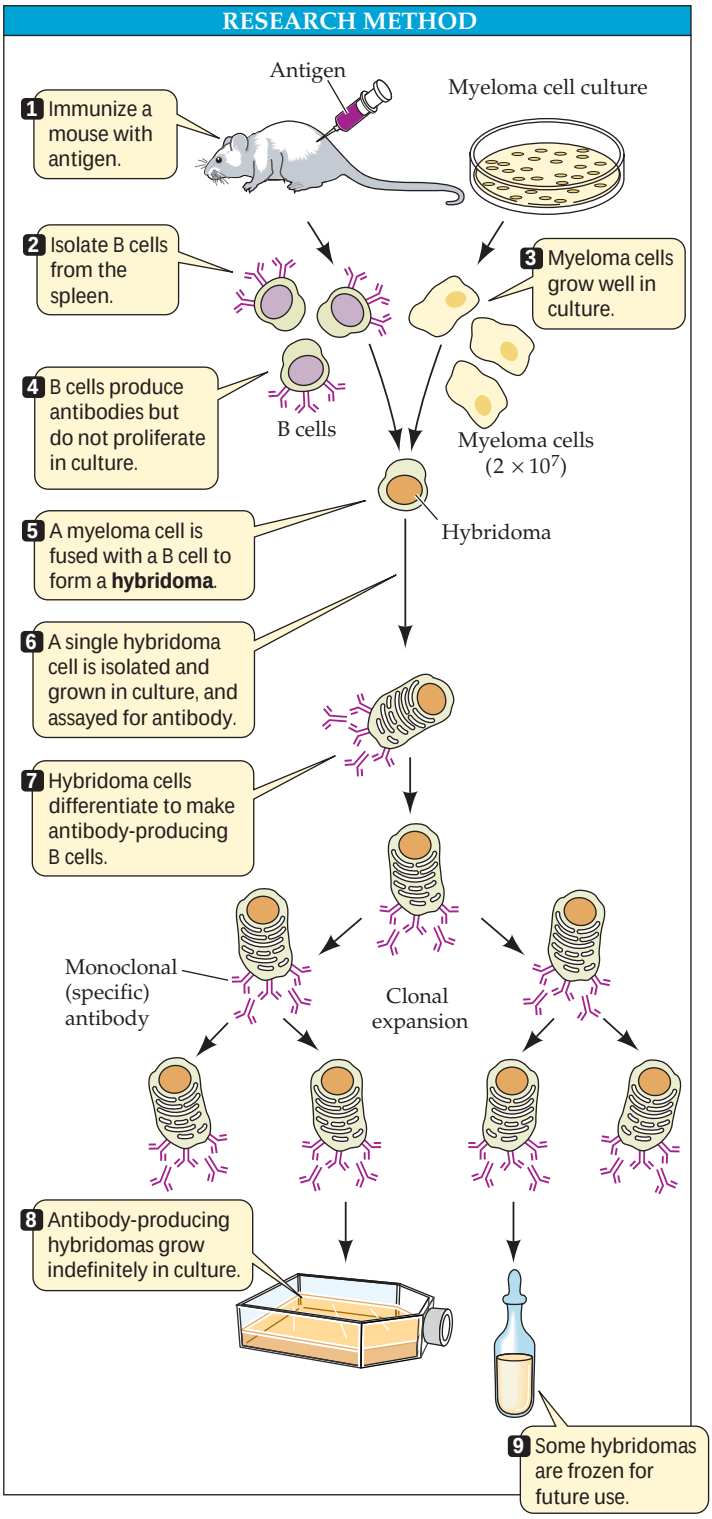
CLASS	GENERAL STRUCTURE	LOCATION	FUNCTION
IgG	Monomer	Free in plasma; about 80 percent of circulating antibodies	Most abundant antibody in primary and secondary responses; crosses placenta and provides passive immunization to fetus
IgM	Pentamer	Surface of B cell; free in plasma	Antigen receptor on B cell membrane; first class of antibodies released by B cells during primary response
IgD	Monomer	Surface of B cell	Cell surface receptor of mature B cell; important in B cell activation
IgA	Dimer	Monomer found in plasma; polymers in saliva, tears, milk, and other body secretions	Protects mucosal surfaces; prevents attachment of pathogens to epithelial cells
IgE	Monomer	Secreted by plasma cells in skin and tissues lining gastrointestinal and respiratory tracts	Found on mast cells and basophils; when bound to antigens, triggers release of histamine from mast cell or basophil that contributes to inflammation and some allergic responses



18.11 IgG Antibodies Promote Phagocytosis
When IgG antibodies cover a bacterium, receptors on a macrophage can recognize, bind to, and engulf it.

An initial challenge to scientists seeking to accomplish this was that the immune response to a complex antigen is not simple. Therefore, they could not simply produce antibodies by injecting an animal with the antigen they wanted to look for. Because most antigens carry many different antigenic determinants, animals injected with a single antigen will produce a complex mixture of antibodies, each made by a different clone of B cells. So the normal antibody response is said to be *polyclonal*.

Suppose that a woman is infertile and her physician wishes to measure the levels of the hormone estrogen in her blood. This could be done adding an antibody directed against estrogen to a sample of her blood and observing how much antigen–antibody complex formed. But, as we have learned in our studies of biochemistry, many molecules share regions of similar structure. All human steroids, for example, have a similar multi-ring structure (see Figure 3.23). A polyclonal group of antibodies against estrogen would not be useful for this test because some of the antibodies would bind not just to estrogen, but to any steroid hormone present in the blood sample. Clearly, a clone of B cells making large amounts of an antibody that binds to only one antigenic determinant—a **monoclonal antibody**—would be needed. How could such a clone be produced?



18.12 Creating Hybridomas for the Production of Monoclonal Antibodies Cancerous myeloma cells and normal B cells can be hybridized so that the proliferative properties of the myeloma cells are merged with the specificity of the antibody-producing B cells.

A single clone of cells making a single antibody can be made by fusing a B cell (which has a finite lifetime and makes a lot of antibody) with a tumor cell (which has an infinite lifetime). The resulting hybrid cells, called *hybridomas*, each make a specific monoclonal antibody (Figure 18.12).

Monoclonal antibodies have many practical applications:

- *Immunoassays* use the great specificity of the antibodies to detect tiny amounts of molecules in tissues and fluids. This technique is used, for example, to quantify the hormone made by the developing embryo for a pregnancy test.
- *Immunotherapy* uses monoclonal antibodies targeted against antigens on the surfaces of cancer cells. The coupling of a radioactive ligand or toxin to the antibody makes it into a medical "smart bomb." In some cases, binding of the antibody itself is enough to trigger a cellular immune response that destroys the cancer.
- *Passive immunization* is inoculation with an immediately acting, but not long-lasting, specific antibody. This approach is necessary when therapy must be effective quickly (within hours). Examples of such life-threatening situations include the early symptoms of rabies infection, rattlesnake bites, and babies born with hepatitis B virus infection—all cases in which the toxic nature of the infection is so serious that there is not enough time to allow the person's immune system to mount its own defense (several days at least).

A major problem with the clinical use of monoclonal antibodies is that the B cells used to produce them come from inoculated mice, so they are mouse proteins. Since mouse immunoglobulin genes differ somewhat from the human ones, the structure of mouse immunoglobulin proteins will also be different, and so the monoclonal antibody may be antigenic to humans. To circumvent this problem, scientists can use recombinant DNA technology to make immunoglobulin genes containing constant regions from humans and variable regions from the mouse (which are not very antigenic to humans). Such a *humanized antibody* does not provoke an immune response in people.

T Cells: The Cellular Immune Response

Thus far we have been concerned primarily with the humoral immune response, whose effector molecules are the antibodies secreted by plasma cells that develop from activated B cells. T cells, as we have seen, are involved in the humoral immune response, but they are also the effectors of the cellular immune response, which is directed against any factor, such as a virus or mutation, that changes a normal cell into an abnormal cell.

In this section, we will describe two types of effector T cells (helper T cells and cytotoxic T cells). We will also describe the MHC (major histocompatibility complex) proteins, which underlie the immune system's tolerance for the cells of its own body.

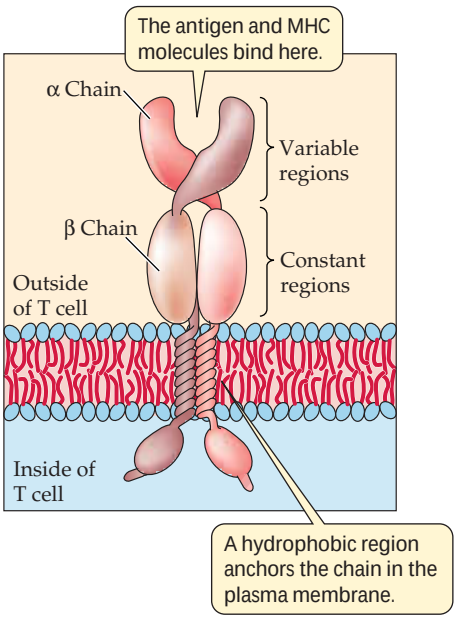
T cell receptors are found on two types of T cells

Like B cells, T cells possess specific membrane receptors. T cell receptors are not immunoglobulins, but glycoproteins with molecular weights about half that of an IgG. They are made up of two polypeptide chains, each encoded by a separate gene (Figure 18.13). Thus the two chains are nearly always different in their amino acid sequence, especially in their variable regions.

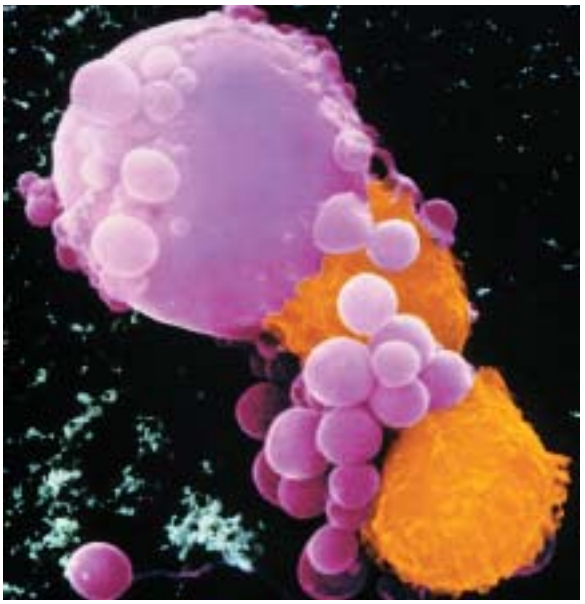
The genes that code for T cell receptors are similar to those for immunoglobulins, suggesting that both are derived from a single, evolutionarily more ancient group of genes. Like the immunoglobulins, T cell receptors include both variable and constant regions. The variable regions provide the specificity for binding with a single antigenic determinant. There is one major difference between antibodies and T cell receptors: While antibodies bind to an intact antigen, T cell receptors bind to a piece of the antigen displayed on the surface of an antigen-presenting cell.

When a T cell is activated by contact with a specific antigenic determinant, it proliferates and forms a clone. Its descendants differentiate into two sub-clones, giving rise to two types of effector T cells:

- **Cytotoxic T cells**, or T_C cells, recognize virus-infected cells and kill them by inducing lysis (Figure 18.14).
- **Helper T cells**, or T_H cells, assist both the cellular and humoral immune responses.



18.13 A T Cell Receptor T cell receptors are glycoproteins, not immunoglobulins, although the structures of the two molecules are similar. In both, each binding site is determined by two polypeptides. T cell receptors are bound more firmly to the plasma membrane of the T cell that produces them than is antibody to B cells.



18.14 Cytotoxic T Cells in Action Two cytotoxic T cells (orange) have come into contact with virus-infected cells, causing the infected cells to die. The pink cell at top left has begun cell death, indicated by membrane blisters. The process is complete in the cell in the center.

As mentioned already, a specific T_H cell must bind to an antigen presented on a B cell before that B cell can become activated. The helper cell becomes the "conductor" of the "immunological orchestra" as it sends out chemical signals that not only result in its own proliferation and that of the B cell, but also set in motion the actions of cytotoxic T cells.

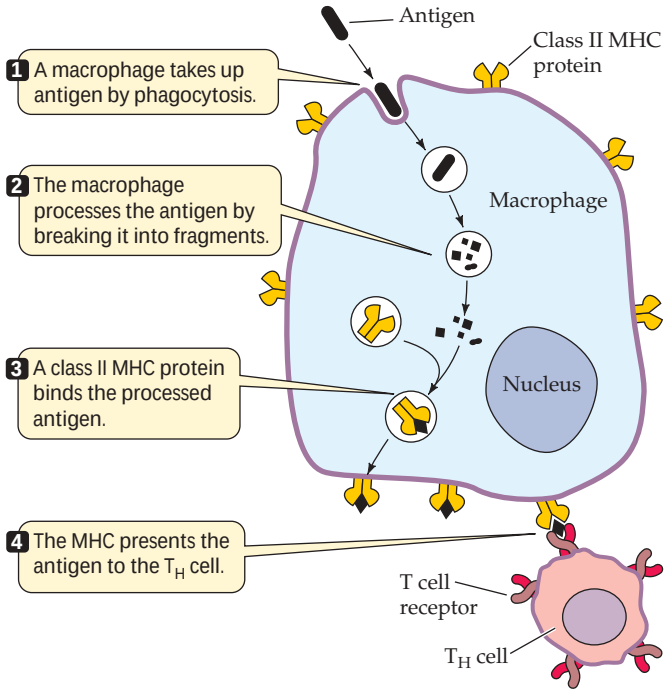
Now that we are familiar with the major types of T cells, we can address the question of how T cells meet their antigenic determinants and the role of the MHC proteins in the process.

The major histocompatibility complex encodes proteins that present antigens to the immune system

We have seen that an animal's immune system recognizes its own cells by their surface proteins. Several types of mammalian cell surface proteins are involved in this process, but we will focus here on one very important group, the products of a cluster of genes called the major histocompatibility complex, or MHC. These proteins have important roles in the cellular and humoral immune responses as well as in self-tolerance.

The MHC gene products are plasma membrane glycoproteins. In humans, the MHC proteins are called *human leukocyte antigens* (HLA), while in mice they are called *H-2 proteins*. Their major role is to present antigens on the cell surface to a T cell receptor. There are three classes of MHC proteins:

- *Class I MHC proteins* are present on the surface of every nucleated cell in the animal body. When cellular proteins are degraded into small peptide fragments by a proteasome (see Chapter 14), an MHC I protein may bind to a fragment and travel to the plasma membrane. There, the



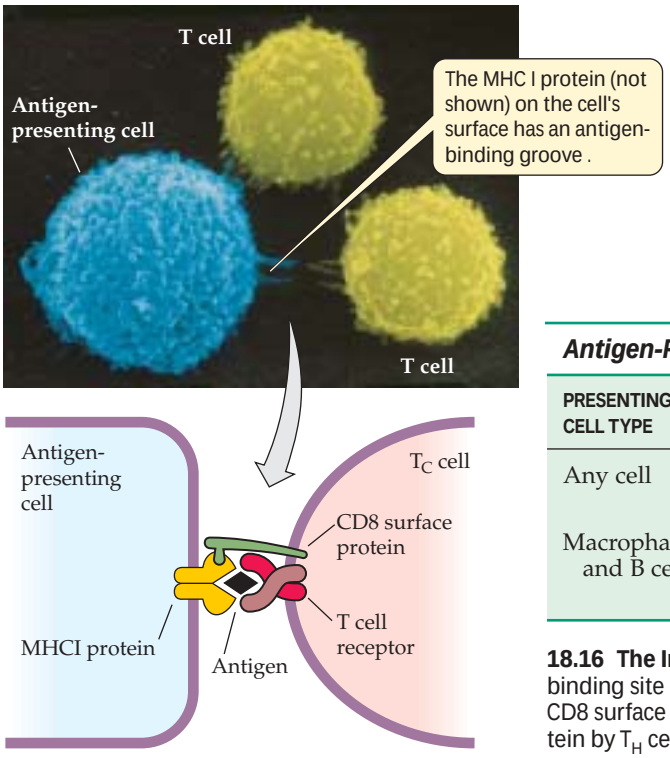
18.15 Macrophages Are Antigen-Presenting Cells A fragment of a processed antigen is displayed by MHC II on the surface of a macrophage. Specific receptors on a helper T cell can then bind to and interact further with the processed antigen/MHC II complex.

MHC I protein "presents" the cellular peptide to T_C cells. The T_C cells have a surface protein called CD8 that recognizes and binds to MHC I.

- *Class II MHC proteins* are found mostly on the surfaces of B cells, macrophages, and other antigen-presenting cells. When an antigen-presenting cell ingests an antigen, such as a virus, the antigen is broken down in a phagosome. An MHC II molecule may bind to one of the fragments and carry it to the cell surface, where it is presented to a T_H cell (Figure 18.15). T_H cells have a surface protein called CD4 that recognizes MHC II.
- *Class III MHC proteins* include some of the proteins of the complement system, which interact with antigen–antibody complexes and result in the lysis of foreign cells.

To accomplish their roles in antigen binding and presentation, both MHC I and MHC II proteins have an antigen-binding site, which can hold a peptide of about 10–20 amino acids (Figure 18.16). The T cell receptor recognizes not just the antigenic fragment, but the fragment *bound to an MHC I or MHC II molecule.* The table in Figure 18.16 summarizes the relationships of T cells and antigen-presenting cells.

In humans, there are three genetic loci for MHC I and three for MHC II; all six loci have as many as 100 different alleles. With so many possible allelic combinations, it is not surprising that different people are very likely to have different MHC genotypes. Similarities in base sequences between the MHC genes and the genes coding for antibodies and T cell receptors suggest that all three may have descended from the same ancestral genes and are part of a gene "superfamily." Major aspects of the immune system in vertebrates seem to be woven together by a common evolutionary thread.



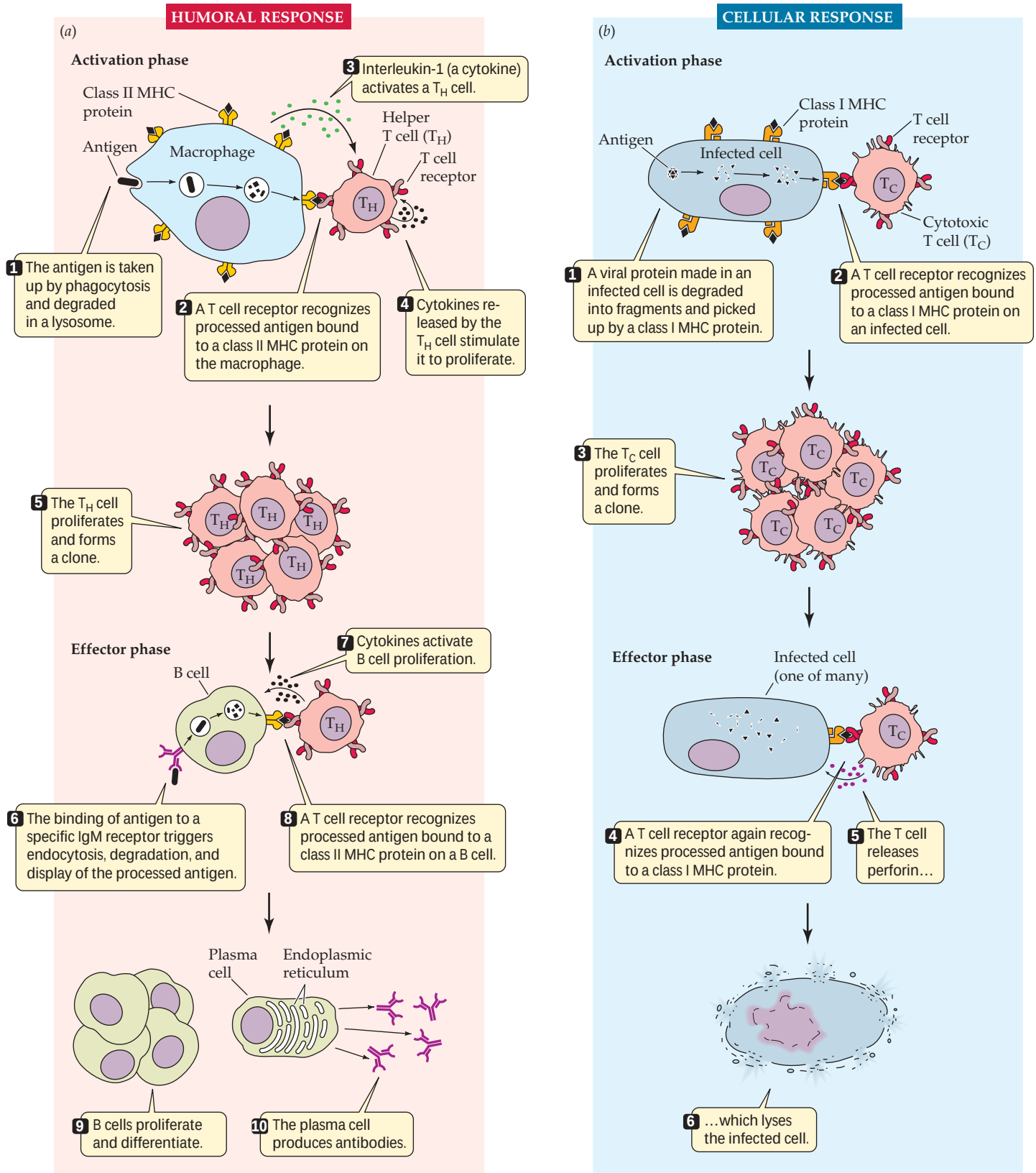
Helper T cells and MHC II proteins contribute to the humoral immune response

When a T_H cell binds to an antigen-presenting macrophage, the T_H cell releases cytokines, which activate the T_H cell to produce a clone of differentiated cells capable of interacting with B cells. The steps to this point constitute the *activation phase* of the humoral immune response, and they occur in the

Antigen-Presenting and T Cell Types

PRESENTING CELL TYPE	ANTIGEN PRESENTED	MHC CLASS	T CELL TYPE	T CELL SURFACE PROTEIN
Any cell	Intracellular protein fragment	Class I	Cytotoxic T cell (T_C)	CD8
Macrophages and B cells	Fragments from extracellular proteins	Class II	Helper T (T_H)	CD4

18.16 The Interaction between T Cells and Antigen-Presenting Cells An antigen-binding site in the MHC I protein holds an antigen, which it presents to cytotoxic T cells. CD8 surface proteins on the T_C cells ensure binding to MHC I. The binding of MHC II protein by T_H cells works in a similar manner.



18.17 Phases of the Humoral and Cellular Immune Responses
Both immune responses have activation and effector phases.

lymphoid tissues. Next comes the *effector phase*, in which B cells are activated to produce antibodies (Figure 18.17*a*).

B cells are also antigen-presenting cells. B cells take up antigens bound to their surface immunoglobulin receptors by endocytosis, process them, and display them on class II MHC proteins. When a T_H cell binds to the displayed antigen–MHC II complex, it releases cytokines, which cause the B cell to produce a clone of plasma cells. Finally, the plasma cells secrete antibody, completing the effector phase of the humoral immune response.

Cytotoxic T cells and MHC I proteins contribute to the cellular immune response

Class I MHC proteins play a role in the cellular immune response that is similar to the role played by class II MHC proteins in the humoral immune response. In a virus-infected or mutated cell, foreign or abnormal proteins or peptide fragments combine with MHC I molecules. The resulting complex is displayed on the cell surface and presented to T_C cells. When a T_C cell binds to this complex, it is activated to proliferate (Figure 18.17*b*).

In the effector phase of the cellular immune response, T_C cells recognize and bind to cells bearing the MHC I–antigen complex. These T_C cells produce a substance called perforin, which lyses the target cell. In addition, the T_C cell can bind to a specific receptor (called Fas) on the target cell that initiates apoptosis in that cell. These two mechanisms, cell lysis and programmed cell death, work in concert to eliminate the altered host cell.

Because T cell receptors recognize self MHC proteins complexed with *nonself* antigens, they help rid the body of its own virus-infected cells. Because they also recognize MHC proteins complexed with *altered self* antigens (as a result of mutations), they help eliminate tumor cells, since most tumor cells have been altered by mutations.

In addition to the binding of an antigen–MHC complex to their receptors, T cells must receive a second signal for activation. This co-stimulatory signal occurs after the initial specific binding and involves the interaction of additional proteins on the T cell with the CD28 protein on the antigen-presenting cell, as we saw above. This second binding event leads to T cell activation, including cytokine production and proliferation. It also sets in motion the production of an *inhibitor* of these events, so that the response is appropriately terminated. This inhibitor, a cell surface protein called CTLA4, blocks the activation process, especially for self-antigens.

MHC proteins underlie the tolerance of self

MHC proteins play a key role in establishing self-tolerance, without which an animal would be destroyed by its own immune system. Throughout the animal's life, developing T cells are tested in the thymus. This "test" consists of two "questions":

1. Can this cell recognize the body's MHC proteins? A T cell unable to recognize self MHC proteins would be useless to the animal because it could not participate in any immune reactions. Such a T cell fails the test and dies within about 3 days.
2. Does this cell bind to self MHC protein *and* to one of the body's own antigens? A T cell that satisfied both of these criteria would be harmful or lethal to the animal; it also fails the test and undergoes apoptosis.

T cells that survive this test mature into either T_C cells or T_H cells.

MHC proteins are responsible for transplant rejection

In humans, a consequence of the major histocompatibility complex became important with the development of organ transplant surgery. Because the proteins produced by the MHC are specific to each individual, they act as antigens if transplanted into another individual. An organ or a piece of tissue transplanted from one person to another is recognized as nonself and soon provokes an immune response; the tissue is then killed, or "rejected," by the host's cellular immune system. But if the transplant is performed immediately after birth, or if it comes from a genetically identical person (an identical twin), the material is recognized as self and is not rejected.

The rejection problem can be overcome by treating a patient with drugs, such as cyclosporin, that suppress the immune system. Cyclosporin works by blocking the activation of a transcription factor essential for T cell development. However, this approach compromises the ability of patients to defend themselves against pathogens. These risks are often managed by the use of antibiotics and other drugs.

So far in this chapter, we have only occasionally alluded to the DNA-based events that make the diversity of antibody specificity possible. In the next section, we will address the genetic mechanisms that generate antibody diversity.

The Genetic Basis of Antibody Diversity

A newborn mammal possesses a full set of genetic information for immunoglobulin synthesis. At each of the loci coding for the heavy and light antibody chains, it has one allele from its mother and one from its father. Throughout the animal's life, each of its cells begins with the same full set of immunoglobulin genes. However, as B cells develop, their genomes become modified in such a way that each cell eventually can produce one—and only one—specific type of antibody. In other words, different B cells develop slightly different genomes encoding different antibody specificities. How can a single organism produce millions of different genomes?

One hypothesis was that we simply have millions of antibody genes. However, a simple calculation (the number of base pairs needed per antibody gene multiplied by millions) shows that if this were true, our entire genome would be taken up by antibody genes! More than 30 years ago, an alternative hypothesis was proposed: A relatively small number of genes recombine to produce many unique combinations, and it is this shuffling of the genetic deck, plus the random pairing of light and heavy antibody chains, that produces antibody diversity. This second hypothesis is now the accepted molecular genetic theory.

In this section, we will describe the unusual events that generate the enormous antibody diversity that normally characterizes each individual mammal. Then we will see how similar events produce the five classes of antibodies by producing slightly different constant regions with special properties.

Antibody diversity results from DNA rearrangement and other mutations

Each gene encoding an immunoglobulin is in reality a "supergene" assembled from several clusters of smaller genes scattered along part of a chromosome (Figure 18.18). Every cell in the body has hundreds of genes, located in separate clusters, that are potentially capable of participating in the synthesis of the variable and constant regions of immunoglobulin polypeptide chains. In most body cells and tissues, these genes remain intact and separated from one another. During B cell development, however, these genes are cut out, rearranged, and joined together. Most of the coding and noncoding regions of these genes are deleted, and one gene from each cluster—is chosen randomly for joining (Figures 18.18, 18.19).

In this manner, a unique antibody supergene is assembled from randomly selected "parts." Each B cell precursor in the animal assembles its own two specific antibody supergenes, one for a specific heavy chain and the other, assembled independently, for a specific light chain. This remarkable example of essentially irreversible cell differentiation generates an enormous diversity of antibody specificities from the same starting genome, one for each individual B cell.

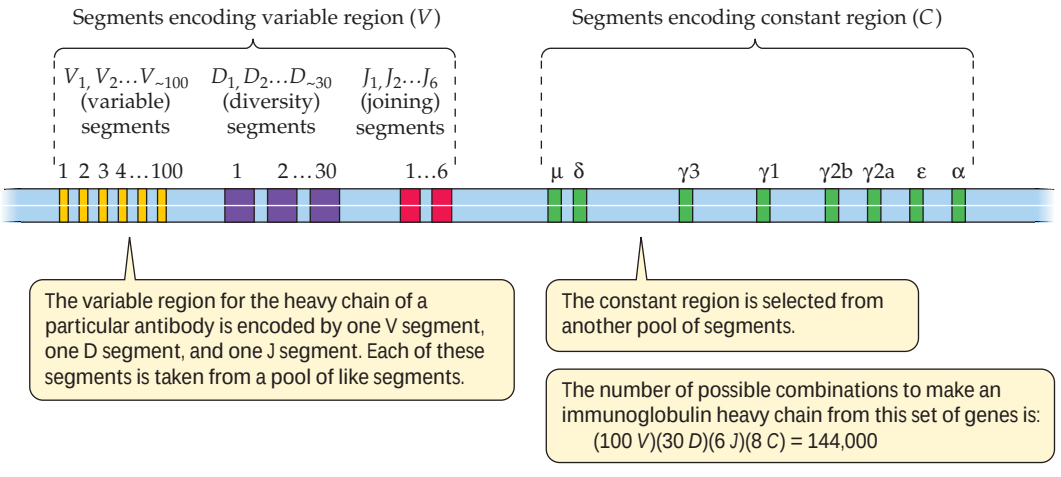
In both humans and mice, the gene clusters coding for immunoglobulin heavy chains are on one pair of chromosomes and those for light chains are on others. The variable region of the light chain is encoded by two families of genes; the variable region of the heavy chain is encoded by three families.

Figure 18.18 illustrates the gene families coding for the heavy-chain constant and variable regions in mice. There are multiple genes coding for each of the four kinds of segments in the polypeptide chain: 100 *V*, 30 *D*, 6 *J*, and 8 *C*. Each B cell that becomes committed to making an antibody randomly selects *one* gene for each of these clusters to make the final heavy-chain coding sequence, *VDJC*. So the number of *different* heavy chains that can be made through this random recombination process is quite large.

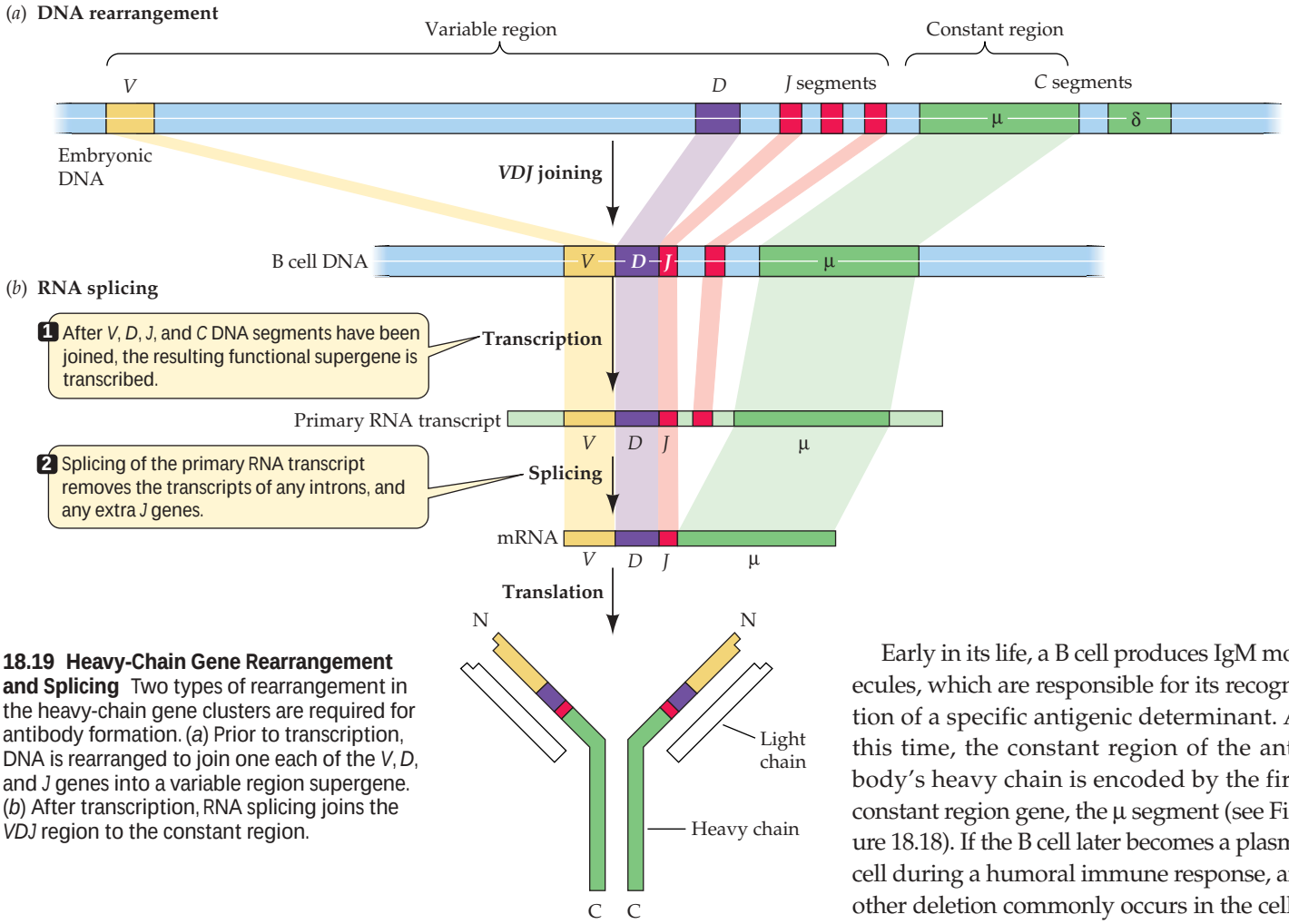
Now consider that the light chains are similarly constructed, with a similar amount of diversity made possible by random recombination. If we assume that light-chain diversity is the same as heavy-chain diversity (144,000 possible combinations), the number of possible combinations of light and heavy chains is 144,000 different light chains × 144,000 different heavy chains = 21 *billion* possibilities!

Even if this number is an overestimate by severalfold (and it is), the number of different immunoglobulin molecules that B cells can make is huge. But there are other mechanisms that generate even more diversity:

- When the DNA sequences for the *V*, *J*, and *C* regions are rearranged so that they are next to one another, the recombination event is not precise, and errors occur at the junctions. This *imprecise recombination* can create new codons at the junctions, with resulting amino acid changes.
- After the DNA sequences are cut out and before they are joined, an enzyme, *terminal transferase*, often adds some nucleotides to the free ends of the DNAs. These additional bases create *insertion mutations*.
- There is a relatively high *mutation rate* in immunoglobulin genes. Once again, this process creates many new alleles and adds to antibody diversity.



18.18 Heavy-Chain Genes Mouse immunoglobulin heavy chains have four domains, each of which is coded for by one of multiple possible genes selected from a cluster of like genes.



18.19 Heavy-Chain Gene Rearrangement and Splicing Two types of rearrangement in the heavy-chain gene clusters are required for antibody formation. (*a*) Prior to transcription, DNA is rearranged to join one each of the *V*, *D*, and *J* genes into a variable region supergene. (*b*) After transcription, RNA splicing joins the *VDJ* region to the constant region.

When we add these possibilities to the billions of combinations that can be made by random DNA rearrangements, it is not surprising that the immune system can mount a response to almost any natural or artificial substance.

Once this *pretranscriptional* processing in completed, pre-mRNA can be transcribed from each supergene. Posttranscriptional processing removes the remaining introns, so that the mature mRNA contains a continuous coding sequence for an immunoglobulin light chain or heavy chain. Translation then produces the polypeptide chains, which combine to form an active antibody protein.

This genetic system is capable of still other kinds of changes, as seen when a B cell or plasma cell switches the immunoglobulin class it produces, but retains its antibody specificity.

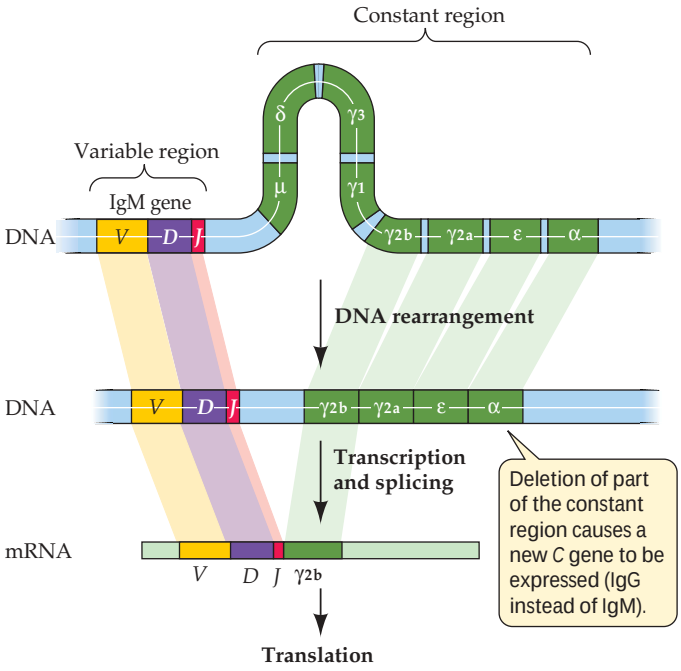
The constant region is involved in class switching

In Table 18.3, we described the different classes of antibodies and their functions. Generally, a B cell makes only one antibody class at a time. But **class switching** can occur, in which a B cell changes which antibody class it synthesizes. For example, a B cell making IgM can switch to making IgG.

Early in its life, a B cell produces IgM molecules, which are responsible for its recognition of a specific antigenic determinant. At this time, the constant region of the antibody's heavy chain is encoded by the first constant region gene, the μ segment (see Figure 18.18). If the B cell later becomes a plasma cell during a humoral immune response, another deletion commonly occurs in the cell's DNA, positioning the heavy-chain variable region gene (consisting of the same *V*, *D*, and *J* segments) next to a constant region gene farther down the original DNA, such as the γ, ε, or α genes (Figure 18.20). Such a DNA deletion results in the production of an antibody with a different constant region of the heavy chain, and therefore a different function. However, the antibody produced has *the same variable regions of the light and heavy chains*, and therefore the same antigen specificity, as before. The new antibody falls into one of the four other immunoglobulin classes (IgA, IgD, IgE, or IgG), depending on which of the constant region genes is placed adjacent to the variable region gene.

After switching classes, the plasma cell cannot go back to making the previous immunoglobulin class, because that part of the DNA has been lost. On the other hand, if additional constant region segments are still present, the cell may switch classes again.

What triggers class switching, and what determines the class to which a given B cell will switch? T_H cells direct the course of an immune response and determine the nature of the attack on the antigen. These T cells induce class switching by sending cytokine signals. The cytokines bind to receptors on the target B cells, generating a signal transduction cascade that results in altered transcription of the immunoglobulin genes.



18.20 Class Switching The supergene produced by joining *V*, *D*, *J*, and *C* segments (see Figure 18.19) may later be modified, causing a different C region to be transcribed. This modification, known as class switching, is accomplished by deletion of part of the constant region gene cluster. Shown here is class switching from IgM to IgG.

By now, you can see that the normal functioning of the immune system involves many complex interactions between molecules and cells. We now turn to several situations in which one or more components of this complex system malfunction.

Disorders of the Immune System

Sometimes the immune system fails us in one way or another. It may overreact, as in an allergic reaction; it may attack self antigens, as in an autoimmune disease; or it may function weakly or not at all, as in an immune deficiency disease. After a look at allergies and autoimmune diseases, we will examine the acquired immune deficiency that characterizes AIDS.

HYPERSENSITIVITY. A common type of condition arises when the human immune system overreacts to (is *hypersensitive* to) a dose of antigen. Although the antigen itself may present no danger to the host, the inappropriate immune response may produce inflammation and other symptoms, which can cause serious illness or death. Allergic reactions are the most familiar examples of this phenomenon. There are two types of allergic reactions:

- *Immediate hypersensitivity* occurs when an individual makes large amounts of IgE that react with an antigen in a food, pollen, or the venom of an insect. When this happens, mast cells in tissues and basophils in blood bind the IgE, which causes them to release histamine. The result is symptoms such as dilation of blood vessels, inflammation, and difficulty breathing. If not treated with antihistamines, a severe allergic reaction can lead to death.
- *Delayed hypersensitivity* does not begin until hours after exposure to an antigen. In this case, the antigen is processed by antigen-presenting cells and a T cell response is initiated. The response can be so massive that the cytokines released cause macrophages to become activated and damage tissues. That is what happens when the bacteria that cause tuberculosis colonize the lungs.

AUTOIMMUNITY. Sometimes clonal deletion fails, resulting in the appearance of one or more "forbidden clones" of B and T cells directed against self antigens. This *autoimmunity* does not always result in disease, but in some instances it can.

- People with *systemic lupus erythematosis* (SLE) have antibodies to many cellular components, including DNA and nuclear proteins. These antinuclear antibodies can cause serious damage when they bind to normal tissue antigens to form large circulating antigen–antibody complexes, which become stuck in tissues and provoke inflammation.
- People with *rheumatoid arthritis* have difficulty in shutting down a T cell response. We mentioned earlier that the inhibitor CTLA4 blocks T cells from reacting to self antigens. People with rheumatoid arthritis may have low CTLA4 activity, which results in inflammation of joints due to the infiltration of excess white blood cells.
- *Multiple sclerosis* involves both T cell- and B cell-mediated attack on two major proteins in myelin, the material that coats some nervous tissues. It usually affects young adults, causing progressive damage to the nervous system.
- *Insulin-dependent diabetes mellitus*, or type I diabetes, occurs most often in children. It involves an immune reaction against several proteins in the cells of the pancreas that manufacture the protein hormone insulin. This reaction kills the insulin-producing cells, so people with type I diabetes must take insulin daily in order to survive.

The causes of these autoimmune diseases are not known. Analyses of human pedigrees show that they tend to "run in families," indicating a genetic component. Some alleles of MHC II are strongly linked to certain autoimmune diseases. In some cases, the underlying cause may be molecular mimicry, in which T cells that recognize a nonself antigen also recognize something on the self that has a similar structure.

AIDS is an immune deficiency disorder

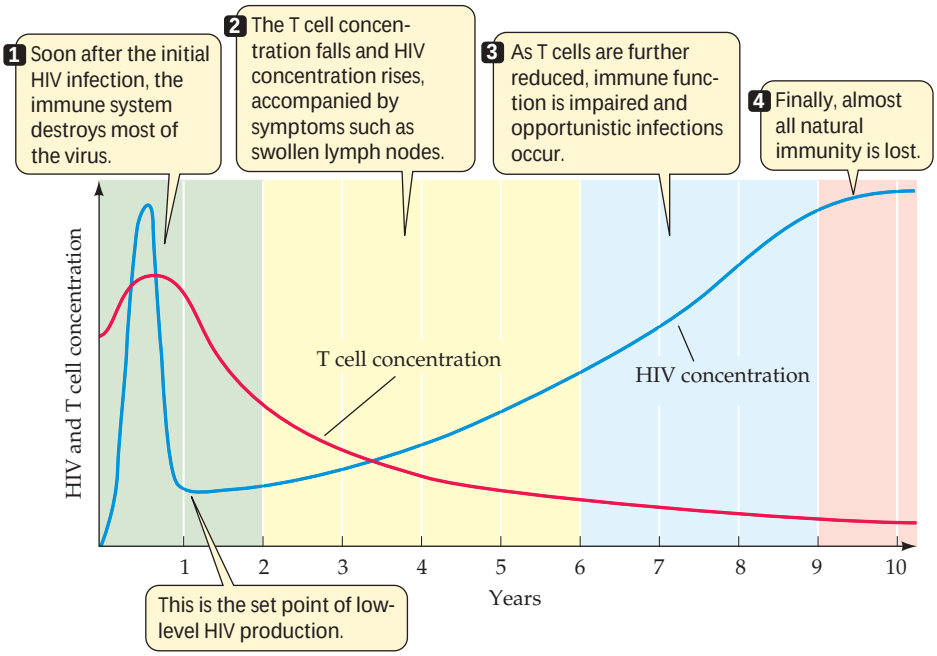
People are subject to various *immune deficiency disorders*, such as those in which T or B cells never form and others in which

B cells lose the ability to give rise to plasma cells. In either case, the affected individual is unable to mount an immune response and thus lacks a major line of defense against pathogens.

Because of its essential roles in both the humoral and cellular immune responses, the T_H cell is perhaps the most central of all the components of the immune system—a significant cell to lose to an immune deficiency disorder. This cell is the target of **HIV** (*h*uman *i*mmunodeficiency *v*irus), the retrovirus that eventually results in **AIDS** (*a*cquired *i*mmune *d*eficiency *s*yndrome).

HIV is transmitted from person to person several ways:

- Through blood, such as by a needle contaminated with the virus after being used to inject an infected individual
- Through the exposure of broken skin, an open wound, or mucous membranes to body fluids, such as blood or semen, from an infected individual
- Through the blood of an infected mother to her baby during birth



18.21 The Course of an HIV Infection HIV infection may be carried, unsuspected, for many years before the onset of symptoms. This long "dormant" period means that the infection is often spread by people who are unaware that they are carrying the virus.

HIV initially infects macrophages, T_H cells, and dendritic cells in blood and tissues. These infected cells carry the virus to the lymph nodes and spleen, where T cells mature and B cells reside.

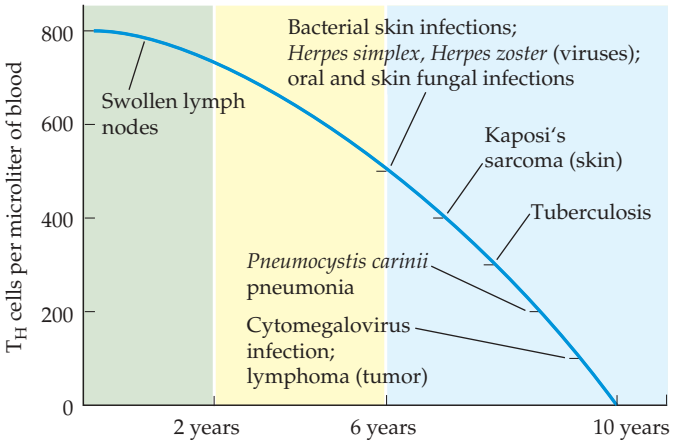
Normally, the dendritic cells present their captured antigen to T_H cells in the lymph nodes, and this causes the T_H cells to divide and form a clone (see Figure 18.17). But HIV preferentially infects activated, not resting, T_H cells. So the HIV arriving in the lymph nodes proceeds to infect the many activated T_H cells that are already responding to other antigens. These two processes—the transport of the virus to the nodes and the presence in the nodes of cells already receptive to virus infection—combine to ensure that HIV reproduces vigorously. Up to 10 billion viruses are made every day during this initial phase of infection. The numbers of T_H cells quickly drop, and infected people show symptoms similar to mononucleosis, such as enlarged lymph nodes and fever.

These symptoms abate within 3 weeks, however, as T cells recognize infected lymphocytes, an immune response is mounted, and antibodies specific to HIV appear in the blood (Figure 18.21). By this time, the patient has a high level of circulating HIV complexed with antibodies, which is gradually removed by the action of dendritic cells over the next several months. But before they are filtered out, these antibody-complexed viruses can still infect T_H cells that come in contact with them. This secondary infection process reaches a low, steady-state level called the "set point." This point varies among individuals and is a strong predictor of the rate of progression of the disease. For most people, it takes 8–10 years, even without treatment, for the more severe manifestations of AIDS to develop. In some, it can take as little as a year; in others, 20 years. During this dormant period, people carrying HIV generally feel fine, and their T_H cell levels are adequate for them to mount immune responses.

Eventually, however, the virus destroys the T_H cells, and their numbers fall to dangerous levels. At this point, the infected person is considered to have *full-blown AIDS* and is susceptible to infections that the T_H cells would normally eliminate (Figure 18.22). Most notable among these infections are the otherwise rare skin tumor called Kaposi's sarcoma, caused by a herpesvirus; pneumonia, caused by the fungus *Pneumocystis carinii*; and lymphoma tumors, caused by the Epstein-Barr virus. These conditions are called *opportunistic infections* because they take advantage of the crippled immune system of the host. They lead to death within a year or two.

HIV infection and replication occur in T_H cells

As a retrovirus, HIV uses RNA as its genetic material. A central core particle with a protein coat contains two identical copies of the RNA genome as well as the enzymes reverse transcriptase, integrase, and a protease. An envelope, derived from the plasma membrane of the host cell in which the virus



18.22 Relationship Between T_H Cell Count and Opportunistic Infections As HIV kills more and more T_H cells, the immune system is less and less able to defend the body against various pathogens, including many that are not usually infectious to healthy people.

was produced, surrounds the core. The envelope is studded with viral proteins (gp120 and gp41, where "gp" stands for *g*lyco*p*rotein). These virally-coded proteins enable the virus to infect its target cells. Refer to Figure 13.5 for the replication cycle of HIV.

HIV attaches to T_H cells and macrophages via their surface protein CD4, which acts as a receptor for the viral envelope protein gp120. Following binding, the virion enters the cell by membrane fusion. Soon, a cDNA copy of the RNA genome is made via reverse transcriptase. This enzyme lacks the proofreading property of many DNA polymerases, so the errors that inevitably creep into the process are not corrected. Up to 10 incorrect bases out of about 8,000 may end up in each cDNA produced. This is a great advantage to the virus, as genomic mutations allow its proteins to escape the host's immune response; however, the mutations present a challenge to scientists trying to design drugs and vaccines to bind to the constantly changing viral proteins.

The viral core enters the cell nucleus, where integrase catalyzes the insertion of viral cDNA into the host chromosome. The double-helical cDNA thus becomes a permanent part of a T_H cell's DNA, replicating with it at each cell division, and may remain in the T_H cell genome for a decade or more. This provides a molecular camouflage against the defenses of host cells, as well as attempts at therapy.

This latent period ends if the HIV-infected T_H cell becomes activated as it responds naturally to an antigen. The expression of viral genes requires the collaboration of host transcription factors that are made in activated T_H cells and a virally encoded protein called Tat. When the T_H cell is activated, the entire integrated cDNA viral genome is transcribed into RNA, which can either remain as it is or be spliced. Unspliced RNAs become the genomes of new HIV particles; spliced RNAs act as mRNAs to make the viral structural proteins. An important activator of splicing is the viral protein called Rev.

The protease encoded by HIV is needed to complete the formation of individual viral proteins from larger initial products of translation. Packaging domains on viral proteins cause the RNA genomes to fold into them and form core particles. In the meantime, the viral membrane proteins are made on the endoplasmic reticulum of the host cell and transported to the plasma membrane via the Golgi complex. The cytoplasmic tails of the viral gp120 membrane proteins bind to the core particles, and the viruses bud from the infected cell, surrounding themselves with modified plasma membrane from the host.

Treatments for HIV infection rely on knowing its molecular biology

As the AIDS epidemic has grown, so has our knowledge of HIV molecular biology. The general therapeutic strategy is to try to block stages in the viral life cycle without damaging the host cell. Potential therapeutic agents that interfere with the major steps of the life cycle are being tested. Of course, it is crucial to block only steps that are unique to the virus, so that drug therapies do not harm the patient by blocking a step in the patient's own metabolism.

*H*ighly *a*ctive *a*nti*r*etroviral *t*herapy (HAART) was developed in the late 1990s and has had considerable success in delaying the onset of AIDS symptoms in people infected with HIV by 3 years or more, and in prolonging the lives of people with AIDS. The logic of HAART comes from cancer treatment: Employ a combination of drugs acting at different parts of the viral life cycle. Generally, the HAART regimen uses a protease inhibitor and two reverse transcriptase inhibitors.

These drug regimens have had such dramatic effects on patients that they may eliminate HIV entirely in some people, especially in those treated within the first few days after infection, before the virus has arrived in the lymph nodes. Most patients, however, face a lifetime of anti-HIV therapy.

Unfortunately, 80 percent of the patients who take HAART develop mutant strains of HIV that are resistant to this regimen; there is a never-ending race to modify HAART by adding new and/or different drug combinations. There are now 140 different HAART treatments. In short, we seem trapped in an evolutionary struggle: How can we gain a lasting advantage, short of bringing the virus to extinction?

The greatest hope is for the development of a vaccine against HIV. The first major clinical trial of such a vaccine

(one directed against the HIV membrane protein, gp120) was not successful, but other vaccines are under development.

What can be done until biomedical science provides the tools to bring the worldwide AIDS epidemic to an end? Above all, people must recognize that they are in danger whenever they have sex with a partner whose total sexual history is not known. The danger rises as the number of sex partners rises, and the danger is much greater if partners participating in sexual intercourse are not protected by a latex condom. The danger that heterosexual intercourse will transmit HIV rises tenfold to a hundredfold if either partner has another sexually transmitted disease.

Chapter Summary

Animal Defense Systems

▶ Animals defend themselves against pathogens by both nonspecific (innate) and specific means.

Defensive Cells and Proteins

▶ Many of our defenses are implemented by cells and proteins carried in the bloodstream and in the lymphatic system. **Review Figure 18.1.** **See Web/CD Activity 18.1**

▶ White blood cells, including lymphocytes (B and T cells) and phagocytes (such as neutrophils and macrophages), play many defensive roles. **Review Figure 18.2.** **See Web/CD Tutorial 18.1**

Nonspecific Defenses

▶ An animal's nonspecific defenses include physical barriers, competing resident microorganisms, and local agents, such as secretions that contain an antibacterial enzyme. **Review Table 18.1**

▶ The inflammation response uses several cells and proteins. Activated mast cells release histamine, which causes blood capillaries to leak and inflame. Complement proteins attract macrophages to the site, where they engulf bacteria and dead cells. **Review Figure 18.4.** **See Web/CD Activity 18.2**

▶ A cell signaling pathway involved the toll receptor stimulates the defense response. **Review Figure 18.5**

Specific Defenses: The Immune Response

▶ Four features characterize the immune response: specificity, the ability to respond to an enormous diversity of antigens, the ability to distinguish self from nonself, and memory.

▶ The immune response is directed against antigens that evade the nonspecific defenses. Each antibody or T cell is directed against a particular antigenic determinant. **Review Figure 18.6**

▶ There are two interactive immune responses: the humoral immune response and the cellular immune response. The humoral immune response employs antibodies secreted by B cells to target antigens in body fluids. The cellular immune response employs T cells to attack body cells that have been altered by viral infection or mutation or to target antigens that have invaded the body's cells.

▶ Clonal selection accounts for the rapidity, specificity, and diversity of the immune response as well as immunological memory and tolerance to self. **Review Figure 18.7**

▶ Immunological memory plays roles in both natural immunity and artificial immunity based on vaccination. **Review Figure 18.8, Table 18.2**

B Cells: The Humoral Immune Response

▶ Activated B cells form plasma cells, which synthesize and secrete specific antibodies.

▶ The basic unit of an antibody, or immunoglobulin, is a tetramer of four polypeptides: two identical light chains and two identical heavy chains, each consisting of a constant and a variable region. **Review Figure 18.10.** **See Web/CD Activity 18.3**

▶ The variable regions of the light and heavy chains collaborate to form the antigen-binding sites of a specific antibody. Each antigen usually has several different antigenic determinants (binding sites for specific antibodies). The variable regions determine each antibody's specificity for a determinant; the constant region determines the destination and function of the antibody.

▶ There are five immunoglobulin classes. IgM, formed first, is a membrane receptor on B cells, as is IgD. IgG is the most abundant antibody class and performs several defensive functions. IgE takes part in inflammation and allergic reactions. IgA is present in various body secretions. **Review Table 18.3**

▶ Monoclonal antibodies consist of identical immunoglobulin molecules directed against a single antigenic determinant. **Review Figure 18.12**

See Web/CD Tutorial 18.2

T Cells: The Cellular Immune Response

▶ The cellular immune response is directed against altered or infected cells of the body. T_C cells attack virus-infected or tumor cells, causing them to lyse. T_H cells activate B cells and influence the development of other T cells and macrophages. **Review Figure 18.13**

▶ T cell receptors in the cellular immune response are analogous to immunoglobulins in the humoral immune response.

▶ The major histocompatibility complex (MHC) encodes many membrane proteins. MHC molecules in macrophages, B cells, or body cells bind processed antigen and present it to T cells. **Review Figures 18.15, 18.16**

▶ In the cellular immune response, class I MHC molecules, T_C cells, CD8, and cytokines collaborate to activate T_C cells with the appropriate specificity. **Review Figure 18.17.** **See Web/CD Tutorial 18.4**

▶ Developing T cells undergo two tests: They must be able to recognize self MHC molecules, and they must *not* bind to both self MHC and any of the body's own antigens. T cells that fail either of these tests die.

▶ The rejection of organ transplants results from the genetic diversity of MHC molecules.

See Web/CD Tutorial 18.3

The Genetic Basis of Antibody Diversity

▶ Immunoglobulin heavy-chain supergenes are constructed from one each of numerous *V*, *D*, *J*, and *C* segments. The *V*, *D*, and *J* segments combine by DNA rearrangement, and transcription yields an RNA molecule that is spliced to form a translatable mRNA. Other gene families give rise to the light chains. **Review Figures 18.18, 18.19**

▶ As a result of these DNA rearrangements, there are millions of possible antibodies as a result of these DNA combinations. Imprecise DNA rearrangements, mutations, and random addition of bases to the ends of the DNAs before they are joined contribute even more diversity.

▶ Class switching after initial immunoglobulin production results in antibodies with the same antigen specificity but a different function. It is accomplished by cutting and rejoining of the genes encoding the constant region. **Review Figure.18.20**

See Web/CD Tutorial 18.5

Disorders of the Immune System

▶ Allergies result from an overreaction of the immune system to an antigen.

▶ Autoimmune diseases result from a failure in the immune recognition of self, with the appearance of antiself B and T cells that attack the body's own cells.

▶ Immune deficiency disorders result from failures of one or another part of the immune system. AIDS is an immune deficiency disorder arising from depletion of the body's T_H cells as a result of infection with HIV. Depletion of the T_H cells weakens and eventually destroys the immune system, leaving the host defenseless against "opportunistic" infections. **Review Figures 18.21, 18.22**

▶ HIV inserts a copy of its genome into a chromosome of a macrophage or T_H cell, where it may lie dormant for years. When the viral genome is transcribed and translated, new viruses form.

▶ Currently the most effective drugs to treat HIV are those directed against reverse transcriptase and protease.

▶ Some treatments may provide a dramatic reduction in HIV levels, but there is as yet no indication that we can prevent infection with HIV, as by vaccination. The only strategy currently available is for people to avoid behaviors that place them at risk.

Self-Quiz

1. Phagocytes kill harmful bacteria by
   a. endocytosis.
   b. producing antibodies.
   c. complement.
   d. T cell stimulation.
   e. inflammation.
2. Which statement about immunoglobulins is *true*?
   a. They help antibodies do their job.
   b. They recognize and bind antigenic determinants.
   c. They encode some of the most important genes in an animal.
   d. They are the chief participants in nonspecific defense mechanisms.
   e. They are a specialized class of white blood cells.
3. Which statement about an antigenic determinant is *not* true?
   a. It is a specific chemical grouping.
   b. It may be part of many different molecules.
   c. It is the part of an antigen to which an antibody binds.
   d. It may be part of a cell.
   e. A single protein has only one on its surface.
4. T cell receptors
   a. are the primary receptors for the humoral immune system.
   b. are carbohydrates.
   c. cannot function unless the animal has previously encountered the antigen.
   d. are produced by plasma cells.
   e. are important in combating viral infections.
5. According to the clonal selection theory,
   a. an antibody changes its shape according to the antigen it meets.
   b. an individual animal contains only one type of B cell.
   c. the animal contains many types of B cells, each producing one kind of antibody.
   d. each B cell produces many types of antibodies.
   e. many clones of antiself lymphocytes appear in the bloodstream.
6. Immunological tolerance
   a. depends on exposure to antigen.
   b. develops late in life and is usually life-threatening.
   c. disappears at birth.
   d. results from the activities of the complement system.
   e. results from DNA splicing.
7. The extraordinary diversity of antibodies results in part from
   a. the action of monoclonal antibodies.
   b. the splicing of protein molecules.
   c. the action of cytotoxic T cells.
   d. the rearrangement of gene segments.
   e. their remarkable nonspecificity.
8. Which of the following play(s) no role in the antibody response?
   a. Helper T cells
   b. Interleukins
   c. Macrophages
   d. Reverse transcriptase
   e. Products of class II MHC gene loci
9. The major histocompatibility complex
   a. codes for specific proteins found on the surface of cells.
   b. plays no role in T cell immunity.
   c. plays no role in antibody responses.
   d. plays no role in skin graft rejection.
   e. is encoded by a single locus with multiple alleles.
10. Which of the following plays no role in HIV reproduction?
    a. Integrase
    b. Reverse transcriptase
    c. gp120
    d. Interleukin-1
    e. Protease

For Discussion

1. Describe the part of an antibody molecule that interacts with an antigenic determinant. How is it similar to the active site of an enzyme? How does it differ from the active site of an enzyme?
2. Contrast immunoglobulins and T cell receptors with respect to their structure and function.
3. Discuss the diversity of antibody specificities in an individual in relation to the diversity of enzymes. Does every cell in an animal contain genetic information for all the organism's enzymes? Does every cell contain genetic information for all the organism's immunoglobulins?
4. The gene family determining MHC on the cell surface in humans is on a single chromosome. A father's MHC type is A1, A3, B5, B7, D9, D11. A mother's phenotype is A2, A4, B6, B7, D11, D12 Their child is A1, A4, B6, B7, D11, D12. What are the parents' haplotypes—that is, which alleles are linked on the diploid chromosomes of each parent? Assuming there is no recombination among the genes determining the MHC type, can these same two parents have a child who is A1, A2, B7, B8, D10, D11?

Lymphocytes and the Cellular Basis of Adaptive Immunity

Lymphocytes are responsible for the astonishing specificity of adaptive **immune** responses. They occur in large numbers in the blood and lymph (the colorless fluid in the lymphatic vessels that connect the lymph nodes in the body to each other and to the bloodstream) and in lymphoid organs, such as the thymus, lymph nodes, spleen, and

In this section, we discuss the general properties of lymphocytes that apply to both B cells and T cells. We shall see that each lymphocyte is committed to respond to a specific antigen and that its response during its first encounter with an antigen ensures that a more rapid and effective response occurs on subsequent encounters with the same antigen. We consider how lymphocytes avoid responding to self antigens and how they continuously recirculate between the blood and lymphoid organs, ensuring that a lymphocyte will find its specific foreign antigen no matter where the anitgen enters the body.

Lymphocytes Are Required for Adaptive Immunity

There are about 2×10^{12} lymphocytes in the human body, making the **immune** system comparable in cell mass to the liver or brain. Despite their abundance, their central role in adaptive immunity was not demonstrated until the late 1950s. The crucial experiments were performed in mice and rats that were heavily irradiated to kill most of their white blood cells, including lymphocytes. This treatment makes the animals unable to mount adaptive **immune** responses. Then, by transferring various types of cells into the animals it was possible to determine which cells reversed the deficiency. Only lymphocytes restored the adaptive **immune** responses of irradiated animals, indicating that lymphocytes are required for these responses

The Innate and Adaptive Immune Systems Work Together

As mentioned earlier, lymphocytes usually respond to foreign antigens only if the **innate immune system** is first activated. As discussed in Chapter 25, the innate **immune** responses to an infection are rapid. They depend on *pattern recognition receptors* that recognize patterns of pathogen-associated molecules (immunostimulants) that are not present in the host organism, including microbial DNA, lipids, and polysaccharides, and proteins that form bacterial flagella. Some of these receptors are present on the surface of professional phagocytic cells such as macrophages and neutrophils, where they mediate the uptake of pathogens, which are then delivered to lysosomes for destruction. Others are secreted and bind to the surface of pathogens, marking them for destruction by either phagocytes or the complement system. Still others are present on the surface of various types of host cells and activate intracellular signaling pathways in response to the binding of pathogen-associated immunostimulants; this leads to the production of extracellular signal molecules that promote inflammation and help activate adaptive **immune** responses.

Some cells of the innate **immune** system directly present microbial antigens to T cells to initiate an adaptive **immune** response. The cells that do this most efficiently are called *dendritic cells,* which are present in most vertebrate tissues. They recognize and phagocytose invading microbes or their products at a site of infection and then migrate with their prey to a nearby peripheral lymphoid organ. There they act as *antigen-presenting cells,* which directly activate T cells to respond to the microbial antigens. Once activated, some of the T cells then migrate to the site of infection, where they help other phagocytic cells, mainly macrophages, destroy the microbes. Other activated T cells remain in the lymphoid organ and help B cells respond to the microbial antigens. The activated B cells secrete antibodies that circulate in the body and coat the microbes, targeting them for efficient phagocytosis.

Thus, innate **immune** responses are activated mainly at sites of infection, whereas adaptive **immune** responses are activated in peripheral lymphoid organs. The two types of responses work together to eliminate invading pathogens.

B Lymphocytes Develop in the Bone Marrow; T Lymphocytes Develop in the Thymus

T cells and B cells derive their names from the organs in which they develop. T cells develop in the *thymus,* and B cells, in mammals, develop in the *bone marrow* in adults or the liver in fetuses.

Despite their different origins, both T and B cells develop from the same *pluripotent hemopoietic stem cells,* which give rise to all of the blood cells, including red blood cells, white blood cells, and platelets. These stem cells (discussed in Chapter 22) are located primarily in *hemopoietic* tissues—mainly the liver in fetuses and the bone marrow in adults. T cells develop in the thymus from precursor cells that migrate there from the hemopoietic tissues via the blood. In most mammals, including humans and mice, B cells develop from stem cells in the hemopoietic tissues themselves. Because they are sites where lymphocytes develop from precursor cells, the thymus and hemopoietic tissues are referred to as **central (primary) lymphoid organs**.

As we discuss later, most lymphocytes die in the central lymphoid organ soon after they develop, without ever functioning. Others, however, mature and migrate via the blood to the **peripheral (secondary) lymphoid organs**—mainly, the lymph nodes, spleen, and epithelium-associated lymphoid tissues in the gastrointestinal tract, respiratory tract, and skin (see Figure 24-3). As mentioned earlier, it is in the peripheral lymphoid organs that T cells and B cells react with foreign antigens.

T and B cells become morphologically distinguishable from each other only after they have been activated by antigen. Nonactivated T and B cells look very similar, even in an electron microscope. Both are small, only marginally bigger than red blood cells, and contain little cytoplasm. Both are activated by antigen to proliferate and mature into *effector cells*. Effector B cells secrete antibodies. In their most mature form, called *plasma cells,* they are filled with an extensive rough endoplasmic reticulum. In contrast, effector T cells contain very little endoplasmic reticulum and do not secrete antibodies.

There are two main classes of T cells—*cytotoxic T cells* and *helper T cells.* Cytotoxic T cells kill infected cells, whereas helper T cells help activate macrophages, B cells, and cytotoxic T

cells. Effector helper T cells secrete a variety of signal proteins called cytokines, which act as local mediators. They also display a variety of costimulatory proteins on their surface. By means of these cytokines and membrane-bound costimulatory proteins, they can influence the behavior of the various cell types they help. Effector cytotoxic T cells kill infected target cells also by means of proteins that they either secrete or display on their surface. Thus, whereas B cells can act over long distances by secreting antibodies that are distributed by the bloodstream, T cells can migrate to distant sites, but there they act only locally on neighboring cells.

The Adaptive Immune System Works by Clonal Selection

The most remarkable feature of the adaptive **immune** system is that it can respond to millions of different foreign antigens in a highly specific way. B cells, for example, make antibodies that react specifically with the antigen that induced their production. How do B cells produce such a diversity of specific antibodies? The answer began to emerge in the 1950s with the formulation of the clonal selection theory. According to this theory, an animal first randomly generates a vast diversity of lymphocytes, and then those lymphocytes that can react against the foreign antigens that the animal actually encounters are specifically selected for action. As each lymphocyte develops in a central lymphoid organ, it becomes committed to react with a particular antigen before ever being exposed to the antigen. It expresses this commitment in the form of cell-surface receptor proteins that specifically fit the antigen. When a lymphocyte encounters its antigen in a peripheral lymphoid organ, the binding of the antigen to the receptors activates the lymphocyte, causing it both to proliferate and to differentiate into an effector cell. An antigen therefore selectively stimulates those cells that express complementary antigen-specific receptors and are thus already committed to respond to it. This arrangement is what makes adaptive **immune** responses antigen-specific.

The term “clonal” in clonal selection theory derives from the postulate that the adaptive **immune** system is composed of millions of different families, or clones, of lymphocytes, each consisting of T or B cells descended from a common ancestor. Each ancestral cell was already committed to make one particular antigen-specific receptor protein, and so all cells in a clone have the same antigen specificity. According to the clonal selection theory, then, the **immune** system functions on the “ready-made” principle rather than the “made-to-order” one.

There is compelling evidence to support the main tenets of the clonal selection theory. For example, when lymphocytes from an animal that has not been immunized are incubated in a test tube with a number of radioactively labeled antigens, only a very small proportion (less than 0.01%) bind each antigen, suggesting that only a few cells are committed to respond to these antigens. Moreover, when one antigen is made so highly radioactive that it kills any cell that it binds to, the remaining lymphocytes can no longer produce an **immune** response to that particular antigen, even though they can still respond normally to other antigens. Thus, the committed lymphocytes must have receptors on their surface that specifically bind that antigen. Although most experiments of this kind have involved B cells and antibody responses, other experiments indicate that T cells, like B cells, operate by clonal selection.

How can the adaptive **immune** system produce lymphocytes that collectively display such an enormous diversity of receptors, including ones that recognize synthetic molecules that never occur in nature? We shall see later that the antigen-specific receptors on both T and B cells are encoded by genes that are assembled from a series of gene segments by a unique form

of genetic recombination that occurs early in a lymphocyte's development, before it has encountered antigen. This assembly process generates the enormous diversity of receptors and lymphocytes, thereby enabling the **immune** system to respond to an almost unlimited diversity of antigens.

Most Antigens Activate Many Different Lymphocyte Clones

Most large molecules, including virtually all proteins and many polysaccharides, can serve as antigens. Those parts of an antigen that combine with the antigen-binding site on either an antibody molecule or a lymphocyte receptor are called antigenic determinants (or *epitopes).* Most antigens have a variety of antigenic determinants that can stimulate the production of antibodies, specific T cell responses, or both. Some determinants of an antigen produce a greater response than others, so that the reaction to them may dominate the overall response. Such determinants are said to be *immunodominant*.

The diversity of lymphocytes is such that even a single antigenic determinant is likely to activate many clones, each of which produces an antigen-binding site with its own characteristic affinity for the determinant. Even a relatively simple structure, like the *dinitrophenyl (DNP)* group in Figure 24-9, can be "looked at" in many ways. When it is coupled to a protein, as shown in the figure, it usually stimulates the production of hundreds of species of anti-DNP antibodies, each made by a different B cell clone. Such responses are said to be *polyclonal*. When only a few clones are activated, the response is said to be *oligoclonal;* and when the response involves only a single B or T cell clone, it is said to be *monoclonal*. Monoclonal antibodies are widely used as tools in biology and medicine, but they have to be produced in a special way , as the responses to most antigens are polyclonal.

Immunological Memory Is Due to Both Clonal Expansion and Lymphocyte Differentiation

The adaptive **immune** system, like the nervous system, can remember prior experiences. This is why we develop lifelong immunity to many common infectious diseases after our initial exposure to the pathogen, and it is why vaccination works. The same phenomenon can be demonstrated in experimental animals. If an animal is immunized once with antigen A, an **immune** response (either antibody or cell-mediated) appears after several days, rises rapidly and exponentially, and then, more gradually, declines. This is the characteristic course of a primary **immune** response, occurring on an animal's first exposure to an antigen. If, after some weeks, months, or even years have elapsed, the animal is reinjected with antigen A, it will usually produce a secondary **immune** response that is very different from the primary response: the lag period is shorter, and the response is greater. These differences indicate that the animal has "remembered" its first exposure to antigen A. If the animal is given a different antigen (for example, antigen B) instead of a second injection of antigen A, the response is typical of a primary, and not a secondary, **immune** response. The secondary response must therefore reflect antigen-specific immunological memory for antigen A.

The clonal selection theory provides a useful conceptual framework for understanding the cellular basis of immunological memory. In an adult animal, the peripheral lymphoid organs contain a mixture of cells in at least three stages of maturation: *naïve cells, effector cells* and

memory cells. When **naïve cells** encounter antigen for the first time, some of them are stimulated to proliferate and differentiate into effector cells, which are actively engaged in making a response (effector B cells secrete antibody, while effector T cells kill infected cells or help other cells fight the infection). Instead of becoming effector cells, some naïve cells are stimulated to multiply and differentiate into **memory cells**—cells that are not themselves engaged in a response but are more easily and more quickly induced to become effector cells by a later encounter with the same antigen. Memory cells, like naïve cells, give rise to either effector cells or more memory cells.

Thus, immunological memory is generated during the primary response in part because the proliferation of antigen-stimulated naïve cells creates many memory cells—a process known as *clonal expansion*—and in part because memory cells are able to respond more sensitively and rapidly to the same antigen than do naïve cells. And, unlike most effector cells, which die within days or weeks, memory cells can live for the lifetime of the animal, thereby providing lifelong immunological memory.

Acquired Immunological Tolerance Ensures That Self Antigens Are Not Attacked

As discussed in Chapter 25, cells of the innate **immune** system recognize molecules on the surface of pathogens that are not found in the host. The adaptive **immune** system has a far more difficult recognition task: it must be able to respond specifically to an almost unlimited number of foreign macromolecules, while avoiding responding to the large number of molecules made by the host organism itself. How does it do it? For one thing, self molecules do not induce the innate **immune** reactions that are required to activate adaptive **immune** responses. But even when an infection triggers innate reactions, self molecules still do not normally induce adaptive **immune** responses. Why not?

One answer is that the adaptive **immune** system "learns" not to respond to self antigens. Transplantation experiments provide one line of evidence for this learning process. When tissues are transplanted from one individual to another, as long as the two individuals are not identical twins, the **immune** system of the recipient usually recognizes the donor cells as foreign and destroys them. (For reasons we discuss later, the foreign antigens on the donor cells are so powerful that they can stimulate adaptive **immune** responses in the absence of infection or an adjuvant.) If, however, cells from one strain of mouse are introduced into a neonatal mouse of another strain, some of these cells survive for most of the recipient animal's life, and the recipient will now accept a graft from the original donor, even though it rejects "third-party" grafts. Apparently, nonself antigens can, in some circumstances, induce the **immune** system to become specifically unresponsive to them. This antigen-specific unresponsiveness to foreign antigens is known as acquired immunological tolerance.

The unresponsiveness of an animal's adaptive **immune** system to its own macromolecules *(natural immunological tolerance)* is acquired in the same way. Normal mice, for example, cannot make an **immune** response against one of their own protein components of the complement system called C5 (discussed in Chapter 25). Mutant mice, however, that lack the gene encoding C5 (but are otherwise genetically identical to the normal mice) can make a strong **immune** response to this blood protein when immunized with it. Natural

immunological tolerance for a particular self molecule persists only for as long as the molecule remains present in the body. If a self molecule such as C5 is removed, an animal gains the ability to respond to it after a few weeks or months. Thus, the **immune** system is genetically capable of responding to self molecules but learns not to do so.

The learning process that leads to self-tolerance can involve killing the self-reactive lymphocytes *(clonal deletion)*, functionally inactivating them *(clonal anergy* or *inactivation)*, stimulating the cells to produce modified receptors that no longer recognize the self antigen *(receptor editing)*, or the suppression of self-reactive lymphocytes by a special type of regulatory T cell. The process begins in the central lymphoid organs when newly formed self-reactive lymphocytes first encounter their self antigen. Instead of being activated by binding antigen, the immature lymphocytes are induced to either alter their receptors or die by apoptosis. Lymphocytes that could potentially respond to self antigens that are not present in the central lymphoid organs often die or are either inactivated or suppressed after they have matured and migrated to peripheral lymphoid organs.

Why does the binding of self antigen lead to tolerance rather than activation? As we discuss later, for a lymphocyte to be activated in a peripheral lymphoid organ, it must not only bind its antigen but must also receive a *costimulatory signal*. The latter signal is provided by a helper T cell in the case of a B lymphocyte and by an antigen-presenting cell in the case of a T lymphocyte. The production of costimulatory signals usually depends on exposure to pathogens, and so a self-reactive lymphocyte normally encounters its antigen in the absence of such signals. Without a costimulatory signal, an antigen tends to kill or inactivate a lymphocyte rather than activate it.

Tolerance to self antigens sometimes breaks down, causing T or B cells (or both) to react against the organism's own tissue antigens. *Myasthenia gravis* is an example of such an autoimmune disease. Affected individuals make antibodies against the acetylcholine receptors on their own skeletal muscle cells. These antibodies interfere with the normal functioning of the receptors so that the patients become weak and may die because they cannot breathe. The mechanisms responsible for the breakdown of tolerance to self antigens in autoimmune diseases are unknown. It is thought, however, that activation of the innate **immune** system by infection may help trigger certain anti-self responses in genetically susceptible individuals.

Lymphocytes Continuously Circulate Through Peripheral Lymphoid Organs

Pathogens generally enter the body through an epithelial surface, usually through the skin, gut, or respiratory tract. How do the microbial antigens travel from these entry points to a peripheral lymphoid organ, such as a lymph node or the spleen, where lymphocytes are activated (see Figure 24-6)? The route and destination depend on the site of entry. Antigens that enter through the skin or respiratory tract are carried via the lymph to local lymph nodes; those that enter through the gut end up in gut-associated peripheral lymphoid organs such as Peyer's patches; and those that enter the blood are filtered out in the spleen. In most cases, dendritic cells carry the antigen from the site of infection to the peripheral lymphoid organ, where they become antigen-presenting cells, specialized for activating T cells (as we discuss later).

But the lymphocytes that can recognize a particular microbial antigen in a peripheral lymph organ are only a tiny fraction of the total lymphocyte population. How do these rare cells find an antigen-presenting cell displaying their antigen? The answer is that they continuously circulate between the lymph and blood until they encounter their antigen. In a lymph node, for example, lymphocytes continually leave the bloodstream by squeezing out between specialized endothelial cells lining small veins called *postcapillary venules.* After percolating through the node, they accumulate in small lymphatic vessels that leave the node and connect with other lymphatic vessels that pass through other lymph nodes downstream. Passing into larger and larger vessels, the lymphocytes eventually enter the main lymphatic vessel (the *thoracic duct),* which carries them back into the blood. This continuous recirculation between the blood and lymph ends only if a lymphocyte encounters its specific antigen (and a costimulatory signal) on the surface of an antigen-presenting cell in a peripheral lymphoid organ. Now the lymphocyte is retained in the peripheral lymphoid organ, where it proliferates and differentiates into effector cells. Some of the effector T cells then leave the organ via the lymph and migrate through the blood to the site of infection.

Lymphocyte recirculation depends on specific interactions between the lymphocyte cell surface and the surface of the specialized endothelial cells lining the postcapillary venules in the peripheral lymphoid organs. Many cell types in the blood come into contact with these endothelial cells, but only lymphocytes adhere and then migrate out of the bloodstream. The lymphocytes initially adhere to the endothelial cells via *homing receptors* that bind to specific ligands (often called *counterreceptors)* on the endothelial cell surface. Lymphocyte migration into lymph nodes, for example, depends on a homing receptor protein called *L-selectin,* a member of the selectin family of cell-surface lectins discussed in Chapter 19. This protein binds to specific sugar groups on a counterreceptor that is expressed exclusively on the surface of the specialized endothelial cells in lymph nodes, causing the lymphocytes to adhere weakly to the endothelial cells and to roll slowly along their surface. The rolling continues until another, much stronger adhesion system is called into play by chemo-attractant proteins (called *chemokines;* see below) secreted by endothelial cells. This strong adhesion is mediated by members of the *integrin* family of cell adhesion molecules (discussed in Chapter 19), which become activated on the lymphocyte surface. Now the lymphocytes stop rolling and crawl out of the blood vessel into the lymph node.

Chemokines are small, secreted, positively charged proteins that have a central role in guiding the migrations of various types of white blood cells. They are all structurally related and bind to the surface of endothelial cells, and to negatively charged proteoglycans of the extracellular matrix in organs. By binding to G-protein-linked receptors (discussed in Chapter 15) on the surface of specific blood cells, chemokines attract these cells from the bloodstream into an organ, guide them to specific locations within the organ, and then help stop migration. (The AIDS virus (HIV) also binds to chemokine receptors, which allows the virus to infect white blood cells.) T and B cells initially enter the same region of a lymph node but are then attracted by different chemokines to separate regions of the node—T cells to the *paracortex* and B cells to *lymphoid follicles*. Unless they encounter their antigen, both types of cells soon leave the lymph node via lymphatic vessels. If they encounter their antigen, however, they remain in the node, proliferate, and differentiate into either effector cells or memory cells. Most of the effector cells leave the node, expressing different chemokine receptors that help guide them to their new destinations—T cells to sites of infection and B cells to the bone marrow.

Summary

Innate **immune** responses are triggered at sites of infection by microbe-specific molecules associated with invading pathogens. In addition to fighting infection directly, these responses help activate adaptive **immune** responses in peripheral lymphoid organs. Unlike innate **immune** responses, adaptive responses provide specific and long-lasting protection against the particular pathogen that induced them.

The adaptive **immune** system is composed of millions of lymphocyte clones, with the cells in each clone sharing a unique cell-surface receptor that enables them to bind a particular antigen. The binding of antigen to these receptors, however, is usually not sufficient to stimulate a lymphocyte to proliferate and differentiate into an effector cell that can help eliminate the pathogen. Costimulatory signals provided by another specialized cell in a peripheral lymphoid organ are also required. Helper T cells provide such signals for B cells, while antigen-presenting dendritic cells usually provide them for T cells. Effector B cells secrete antibodies, which can act over long distances to help eliminate extracellular pathogens and their toxins. Effector T cells, by contrast, act locally at sites of infection to either kill infected host cells or help other cells to eliminate pathogens. As part of the adaptive **immune** response, some lymphocytes proliferate and differentiate into memory cells, which are able to respond faster and more efficiently the next time the same pathogen invades. Lymphocytes that would react against self molecules are either induced to alter their receptors, induced to kill themselves, inactivated, or suppressed, so that the adaptive **immune** system normally reacts only against foreign antigens. Both B and T cells circulate continuously between the blood and lymph. Only if they encounter their specific foreign antigen in a peripheral lymphoid organ do they stop migrating, proliferate, and differentiate into effector cells or memory cells.

Vertebrates inevitably die of infection if they are unable to make antibodies. Antibodies defend us against infection by binding to viruses and microbial toxins, thereby inactivating them. The binding of antibodies to invading pathogens also recruits various types of white blood cells and a system of blood proteins, collectively called *complement* (discussed in Chapter 25). The white blood cells and activated complement components work together to attack the invaders.

Synthesized exclusively by B cells, antibodies are produced in billions of forms, each with a different amino acid sequence and a different antigen-binding site. Collectively called immunoglobulins (abbreviated as **Ig**), they are among the most abundant protein components in the blood, constituting about 20% of the total protein in plasma by weight. Mammals make five classes of antibodies, each of which mediates a characteristic biological response following antigen binding. In this section, we discuss the structure and function of antibodies and how they interact with antigen.

B Cells and Antibodies

Vertebrates inevitably die of infection if they are unable to make antibodies. Antibodies defend us against infection by binding to viruses and microbial toxins, thereby inactivating them. The binding of antibodies to invading pathogens also recruits various types of white blood cells and a system of blood proteins, collectively called *complement* (discussed in

Chapter 25). The white blood cells and activated complement components work together to attack the invaders.

Synthesized exclusively by B cells, antibodies are produced in billions of forms, each with a different amino acid sequence and a different antigen-binding site. Collectively called immunoglobulins (abbreviated as **Ig**), they are among the most abundant protein components in the blood, constituting about 20% of the total protein in plasma by weight. Mammals make five classes of antibodies, each of which mediates a characteristic biological response following antigen binding. In this section, we discuss the structure and function of antibodies and how they interact with antigen.

B Cells Make Antibodies as Both Cell-Surface Receptors and Secreted Molecules

As predicted by the clonal selection theory, all antibody molecules made by an individual B cell have the same antigen-binding site. The first antibodies made by a newly formed B cell are not secreted. Instead, they are inserted into the plasma membrane, where they serve as receptors for antigen. Each B cell has approximately 10^5 such receptors in its plasma membrane. As we discuss later, each of these receptors is stably associated with a complex of transmembrane proteins that activate intracellular signaling pathways when antigen binds to the receptor.

Each B cell produces a single species of antibody, each with a unique antigen-binding site. When a naïve or memory B cell is activated by antigen (with the aid of a helper T cell), it proliferates and differentiates into an antibody-secreting effector cell. Such cells make and secrete large amounts of soluble (rather than membrane-bound) antibody, which has the same unique antigen-binding site as the cell-surface antibody that served earlier as the antigen receptor. Effector B cells can begin secreting antibody while they are still small lymphocytes, but the end stage of their maturation pathway is a large *plasma cell* (see Figure 24-7B), which continuously secretes antibodies at the astonishing rate of about 2000 molecules per second. Plasma cells seem to have committed so much of their protein-synthesizing machinery to making antibody that they are incapable of further growth and division. Although many die after several days, some survive in the bone marrow for months or years and continue to secrete antibodies into the blood.

A Typical Antibody Has Two Identical Antigen-Binding Sites

The simplest antibodies are Y-shaped molecules with two identical antigen-binding sites, one at the tip of each arm of the Y Because of their two antigen-binding sites, they are described as *bivalent*. As long as an antigen has three or more antigenic determinants, bivalent antibody molecules can cross-link it into a large lattice. This lattice can be rapidly phagocytosed and degraded by macrophages. The efficiency of antigen binding and cross-linking is greatly increased by a flexible *hinge region* in most antibodies, which allows the distance between the two antigen-binding sites to vary.

The protective effect of antibodies is not due simply to their ability to bind antigen. They engage in a variety of activities that are mediated by the tail of the Y-shaped molecule. As we discuss later, antibodies with the same antigen-binding sites can have any one of several different tail regions. Each type of tail region gives the antibody different functional properties, such as the ability to activate the complement system, to bind to phagocytic cells, or to cross the placenta from mother to fetus.

An Antibody Molecule Is Composed of Heavy and Light Chains

The basic structural unit of an antibody molecule consists of four polypeptide chains, two identical **light (L) chains** (each containing about 220 amino acids) and two identical **heavy (H) chains** (each usually containing about 440 amino acids). The four chains are held together by a combination of noncovalent and covalent (disulfide) bonds. The molecule is composed of two identical halves, each with the same antigen-binding site. Both light and heavy chains usually cooperate to form the antigen-binding surface.

There Are Five Classes of Heavy Chains, Each With Different Biological Properties

In mammals, there are five *classes* of antibodies, IgA, IgD, IgE, IgG, and IgM, each with its own class of heavy chain—α, δ, ε, γ, and μ, respectively. IgA molecules have α chains, IgG molecules have γ chains, and so on. In addition, there are a number of subclasses of IgG and IgA immunoglobulins; for example, there are four human IgG subclasses (IgG1, IgG2, IgG3, and IgG4), having γ_1, γ_2, γ_3, and γ_4 heavy chains, respectively. The various heavy chains give a distinctive conformation to the hinge and tail regions of antibodies, so that each class (and subclass) has characteristic properties of its own.

IgM, which has μ heavy chains, is always the first class of antibody made by a developing B cell, although many B cells eventually switch to making other classes of antibody (discussed below). The immediate precursor of a B cell, called a pre-B cell, initially makes μ chains, which associate with so-called *surrogate light chains* (substituting for genuine light chains) and insert into the plasma membrane. The complexes of μ chains and surrogate light chains are required for the cell to progress to the next stage of development, where it makes bona fide light chains. The light chains combine with the μ chains, replacing the surrogate light chains, to form four-chain IgM molecules (each with two μ chains and two light chains). These molecules then insert into the plasma membrane, where they function as receptors for antigen. At this point, the cell is called an *immature naïve B cell*. After leaving the bone marrow, the cell starts to produce cell-surface **IgD** molecules as well, with the same antigen-binding site as the IgM molecules. It is now called a *mature naïve B cell*. It is this cell that can respond to foreign antigen in peripheral lymphoid organs.

IgM is not only the first class of antibody to appear on the surface of a developing B cell. It is also the major class secreted into the blood in the early stages of a *primary* antibody response, on first exposure to an antigen. (Unlike IgM, IgD molecules are secreted in only small amounts and seem to function mainly as cell-surface receptors for antigen.) In its secreted form, IgM is a pentamer composed of five four-chain units, giving it a total of 10 antigen-binding sites. Each pentamer contains one copy of another polypeptide chain, called a

J (joining) chain. The J chain is produced by IgM-secreting cells and is covalently inserted between two adjacent tail regions.

The binding of an antigen to a single secreted pentameric IgM molecule can activate the complement system. As discussed in Chapter 25, when the antigen is on the surface of an invading pathogen, this activation of complement can either mark the pathogen for phagocytosis or kill it directly.

The major class of immunoglobulin in the blood is **IgG**, which is a four-chain monomer produced in large quantities during *secondary* **immune** responses. Besides activating complement, the tail region of an IgG molecule binds to specific receptors on macrophages and neutrophils. Largely by means of such Fc receptors (so-named because antibody tails are called *Fc* regions), these phagocytic cells bind, ingest, and destroy infecting microorganisms that have become coated with the IgG antibodies produced in response to the infection.

IgG molecules are the only antibodies that can pass from mother to fetus via the placenta. Cells of the placenta that are in contact with maternal blood have Fc receptors that bind blood-borne IgG molecules and direct their passage to the fetus. The antibody molecules bound to the receptors are first taken into the placental cells by receptor-mediated endocytosis. They are then transported across the cell in vesicles and released by exocytosis into the fetal blood (a process called *transcytosis*, discussed in Chapter 13). Because other classes of antibodies do not bind to these particular Fc receptors, they cannot pass across the placenta. IgG is also secreted into the mother's milk and is taken up from the gut of the neonate into the blood, providing protection for the baby against infection.

IgA is the principal class of antibody in secretions, including saliva, tears, milk, and respiratory and intestinal secretions. Whereas IgA is a four-chain monomer in the blood, it is an eight-chain dimer in secretions. It is transported through secretory epithelial cells from the extracellular fluid into the secreted fluid by another type of Fc receptor that is unique to secretory. This Fc receptor can also transport IgM into secretions (but less efficiently), which is probably why individuals with a selective IgA deficiency, the most common form of antibody deficiency, are only mildly affected by the defect.

The tail region of **IgE** molecules, which are four-chain monomers, binds with unusually high affinity ($K_a \sim 10^{10}$ liters/mole) to yet another class of Fc receptors. These receptors are located on the surface of *mast cells* in tissues and of *basophils* in the blood. The IgE molecules bound to them function as passively acquired receptors for antigen. Antigen binding triggers the mast cell or basophil to secrete a variety of cytokines and biologically active amines, especially *histamine*. These molecules cause blood vessels to dilate and become leaky, which in turn helps white blood cells, antibodies, and complement components to enter sites of infection. The same molecules are also largely responsible for the symptoms of such *allergic* reactions as hay fever, asthma, and hives. In addition, mast cells secrete factors that attract and activate white blood cells called *eosinophils.* These cells also have Fc receptors that bind IgE molecules and can kill various types of parasites, especially if the parasites are coated with IgE antibodies.

In addition to the five classes of heavy chains found in antibody molecules, higher vertebrates have two types of light chains, κ and λ, which seem to be functionally indistinguishable. Either type of light chain may be associated with any of the heavy chains. An individual antibody molecule, however, always contains identical light chains and identical heavy

chains: an IgG molecule, for instance, may have either κ or λ light chains, but not one of each. As a result of this symmetry, an antibody's antigen-binding sites are always identical. Such symmetry is crucial for the cross-linking function of secreted antibodies.

of an Antibody-Antigen Interaction Depends on Both the Number and the Affinity of the Antigen-Binding Sites

The binding of an antigen to antibody, like the binding of a substrate to an enzyme, is reversible. It is mediated by the sum of many relatively weak non-covalent forces, including hydrogen bonds and hydrophobic van der Waals forces, and ionic interactions. These weak forces are effective only when the antigen molecule is close enough to allow some of its atoms to fit into complementary recesses on the surface of the antibody. The complementary regions of a four-chain antibody unit are its two identical antigen-binding sites; the corresponding region on the antigen is an *antigenic determinant*. Most antigenic macromolecules have many different antigenic determinants and are said to be *multivalent;* if two or more of them are identical (as in a polymer with a repeating structure), the antigen is said to be *polyvalent*.

The reversible binding reaction between an antigen with a single antigenic determinant (denoted Ag) and a single antigen-binding site (denoted Ab) can be expressed as

$$Ag + Ab \leftrightarrow AgAb$$

The equilibrium point depends both on the concentrations of Ab and Ag and on the strength of their interaction. Clearly, a larger fraction of Ab will become associated with Ag as the concentration of Ag increases. The strength of the interaction is generally expressed as the **affinity constant (K_a)** (see Figure 3-44), where

$$K_a = \frac{[AgAb]}{[Ag][Ab]}$$

(the square brackets indicate the concentration of each component at equilibrium).

The affinity constant, sometimes called the association constant, can be determined by measuring the concentration of free Ag required to fill half of the antigen-binding sites on the antibody. When half the sites are filled, [AgAb] = [Ab] and $K_a = 1/[Ag]$. Thus, the reciprocal of the antigen concentration that produces half the maximum binding is equal to the affinity constant of the antibody for the antigen. Common values range from as low as 5×10^4 to as high as 10^{11} liters/mole.

The **affinity** of an antibody for an antigenic determinant describes the strength of binding of a single copy of the antigenic determinant to a single antigen-binding site, and it is independent of the number of sites. When, however, a polyvalent antigen, carrying multiple copies of the same antigenic determinant, combines with a polyvalent antibody, the binding strength is greatly increased because all of the antigen-antibody bonds must be broken simultaneously before the antigen and antibody can dissociate. As a result, a typical IgG molecule can bind at least 100 times more strongly to a polyvalent antigen if both antigen-binding sites are engaged than if only one site is engaged. The total binding strength of a polyvalent antibody with a polyvalent antigen is referred to as the avidity of the interaction.

If the affinity of the antigen-binding sites in an IgG and an IgM molecule is the same, the IgM molecule (with 10 binding sites) will have a much greater avidity for a multivalent antigen than an IgG molecule (which has two binding sites). This difference in avidity, often 10^4-fold or more, is important because antibodies produced early in an **immune** response usually have much lower affinities than those produced later. Because of its high total avidity, IgM—the major Ig class produced early in **immune** responses—can function effectively even when each of its binding sites has only a low affinity.

So far we have considered the general structure and function of antibodies. Next we look at the details of their structure, as revealed by studies of their amino acid sequence and three-dimensional structure.

Light and Heavy Chains Consist of Constant and Variable Regions

Comparison of the amino acid sequences of different antibody molecules reveals a striking feature with important genetic implications. Both light and heavy chains have a variable sequence at their N-terminal ends but a constant sequence at their C-terminal ends. Consequently, when the amino acid sequences of many different κ chains are compared, the C-terminal halves are the same or show only minor differences, whereas the N-terminal halves are all very different. Light chains have a **constant region** about 110 amino acids long and a variable region of the same size. The variable region of the heavy chains (at their N-terminus) is also about 110 amino acids long, but the heavy-chain constant region is about three or four times longer (330 or 440 amino acids), depending on the class.

It is the N-terminal ends of the light and heavy chains that come together to form the antigen-binding site, and the variability of their amino acid sequences provides the structural basis for the diversity of antigen-binding sites. The diversity in the variable regions of both light and heavy chains is for the most part restricted to three small hypervariable regions in each chain; the remaining parts of the variable region, known as *framework regions*, are relatively constant. Only the 5–10 amino acids in each hypervariable region form the antigen-binding site. As a result, the size of the antigenic determinant that an antibody recognizes is generally comparably small. It can consist of fewer than 25 amino acids on the surface of a globular protein, for example.

The Light and Heavy Chains Are Composed of Repeating Ig Domains

Both light and heavy chains are made up of repeating segments—each about 110 amino acids long and each containing one intrachain disulfide bond. These repeating segments fold independently to form compact functional units called **immunoglobulin (Ig) domains**. As shown in Figure 24-32, a light chain consists of one variable (V_L) and one constant (C_L) domain. These domains pair with the variable (V_H) and first constant (C_H1) domain of the heavy chain to form the antigen-binding region. The remaining constant domains of the heavy chains form the Fc region, which determines the other biological properties of the antibody. Most heavy chains have three constant domains (C_H1, C_H2, and C_H3), but those of IgM and IgE antibodies have four.

The similarity in their domains suggests that antibody chains arose during evolution by a series of gene duplications, beginning with a primordial gene coding for a single 110 amino acid domain of unknown function. This hypothesis is supported by the finding that each domain of the constant region of a heavy chain is encoded by a separate coding sequence (exon).

From Hypervariable Loops

A number of fragments of antibodies, as well as intact antibody molecules, have been studied by x-ray crystallography. From these examples, we can understand the way in which billions of different antigen-binding sites are constructed on a common structural theme.

As illustrated, each Ig domain has a very similar three-dimensional structure based on what is called the *immunoglobulin fold*, which consists of a sandwich of two β sheets held together by a disulfide bond. We shall see later that many other proteins on the surface of lymphocytes and other cells, many of which function as cell-cell adhesion molecules (discussed in Chapter 19), contain similar domains and hence are members of a very large *immunoglobulin (Ig) superfamily* of proteins.

of a constant domain is shown on the left and of a variable domain on the right.
Both (more...)

The variable domains of antibody molecules are unique in that each has its particular set of three hypervariable regions, which are arranged in three *hypervariable loops*. The hypervariable loops of both the light and heavy variable domains are clustered together to form the antigen-binding site. Because the variable region of an antibody molecule consists of a highly conserved rigid framework, with hypervariable loops attached at one end, an enormous diversity of antigen-binding sites can be generated by changing only the lengths and amino acid sequences of the hypervariable loops. The overall three-dimensional structure necessary for antibody function remains constant.

X-ray analyses of crystals of antibody fragments bound to an antigenic determinant reveal exactly how the hypervariable loops of the light and heavy variable domains cooperate to form an antigen-binding surface in particular cases. The dimensions and shape of each different site vary depending on the conformations of the polypeptide chain in the hypervariable loops, which in turn are determined by the sequences of the amino acid side chains in the loops. The shapes of binding sites vary greatly—from pockets, to grooves, to undulating flatter surfaces, and even to protrusions—depending on the antibody. Smaller ligands tend to bind to deeper pockets, whereas larger ones tend to bind to flatter surfaces. In addition, the binding site can alter its shape after antigen binding to better fit the ligand.

Summary

Antibodies defend vertebrates against infection by inactivating viruses and microbial toxins and by recruiting the complement system and various types of white blood cell to kill the invading pathogens. A typical antibody molecule is Y-shaped, with two identical antigen-binding sites at the tips of the Y and binding sites for complement components and/or various cell-surface receptors on the tail of the Y.

Each B cell clone makes antibody molecules with a unique antigen-binding site. Initially, during B cell development in the bone marrow, the antibody molecules are inserted into the plasma membrane, where they serve as receptors for antigen. In peripheral lymphoid organs, antigen binding to these receptors, together with costimulatory signals provided by helper T cells, activates the B cells to proliferate and differentiate into either memory cells or antibody-secreting effector cells. The effector cells secrete antibodies with the same unique antigen-binding site as the membrane-bound antibodies.

A typical antibody molecule is composed of four polypeptide chains, two identical heavy chains and two identical light chains. Parts of both the heavy and light chains usually combine to form the antigen-binding sites. There are five classes of antibodies (IgA, IgD, IgE, IgG, and IgM), each with a distinctive heavy chain (α, δ, ε, γ, and μ, respectively). The heavy chains also form the tail (Fc region) of the antibody, which determines what other proteins will bind to the antibody and therefore what biological properties the antibody class has. Either type of light chain (κ or λ) can be associated with any class of heavy chain, but the type of light chain does not seem to influence the properties of the antibody, other than its specificity for antigen.

Each light and heavy chain is composed of a number of Ig domains—β sheet structures containing about 110 amino acids. A light chain has one variable (V_L) and one constant (C_L) domain, while a heavy chain has one variable (V_H) and three or four constant (C_H) domains. The amino acid sequence variation in the variable domains of both light and heavy chains is mainly confined to several small hypervariable regions, which protrude as loops at one end of the domains to form the antigen-binding site.

The Generation of Antibody Diversity

Even in the absence of antigen stimulation, a human can probably make more than 10^{12} different antibody molecules—its *preimmune antibody repertoire*. Moreover, the antigen-binding sites of many antibodies can cross-react with a variety of related but different antigenic determinants, making the antibody defense force even more formidable. The preimmune repertoire is apparently large enough to ensure that there will be an antigen-binding site to fit almost any potential antigenic determinant, albeit with low affinity. After repeated stimulation by antigen, B cells can make antibodies that bind their antigen with much higher affinity—a process called *affinity maturation*. Thus, antigen stimulation greatly increases the antibody arsenal.

Antibodies are proteins, and proteins are encoded by genes. Antibody diversity therefore poses a special genetic problem: how can an animal make more antibodies than there are genes in its genome? (The human genome, for example, contains fewer than 50,000 genes.) This problem is not quite as formidable as it might first appear. Recall that the variable regions of both the light and heavy chains of antibodies usually form the antigen-binding site.

Thus, an animal with 1000 genes encoding light chains and 1000 genes encoding heavy chains could, in principle, combine their products in 1000 × 1000 different ways to make 10^6 different antigen-binding sites (although, in reality, not every light chain can combine with every heavy chain to make an antigen-binding site). Nonetheless, the mammalian **immune** system has evolved unique genetic mechanisms that enable it to generate an almost unlimited number of different light and heavy chains in a remarkably economical way, by joining separate *gene segments* together before they are transcribed. Birds and fish use very different strategies for diversifying antibodies, and even sheep and rabbits use somewhat different strategies from mice and humans. We shall confine our discussion to the mechanisms used by mice and humans.

We begin this section by discussing the mechanisms that B cells use to produce antibodies with an enormous diversity of antigen-binding sites. We then consider how a B cell can alter the tail region of the antibody it makes, while keeping the antigen-binding site unchanged. This ability allows the B cell to switch from making membrane-bound antibody to making secreted antibody, or from making one class of antibody to making another, all without changing the antigen-specificity of the antibody.

Antibody Genes Are Assembled From Separate Gene Segments During B Cell Development

The first direct evidence that DNA is rearranged during B cell development came in the 1970s from experiments in which molecular biologists compared DNA from early mouse embryos, which do not make antibodies, with the DNA of a mouse B cell tumor, which makes a single species of antibody molecule. The specific variable (V)-region and constant (C)-region coding sequences that the tumor cells used were present on the same DNA restriction fragment in the tumor cells but on two different restriction fragments in the embryos. This showed that the DNA sequences encoding an antibody molecule are rearranged at some stage in B cell development.

We now know that each type of antibody chain—κ light chains, λ light chains, and heavy chains—has a separate pool of **gene segments** and exons from which a single polypeptide chain is eventually synthesized. Each pool is on a different chromosome and contains a large number of gene segments encoding the V region of an antibody chain and, as we saw in Figure 24-33, a smaller number of exons encoding the C region. During the development of a B cell, a complete coding sequence for each of the two antibody chains to be synthesized is assembled by site-specific genetic recombination (discussed in Chapter 5). In addition to bringing together the separate gene segments and the C-region exons of the antibody gene, these rearrangements also activate transcription from the gene promoter through changes in the relative positions of the enhancers and silencers acting on the promoter. Thus, a complete antibody chain can be synthesized only after the DNA has been rearranged. As we shall see, the process of joining gene segments contributes to the diversity of antigen-binding sites in several ways.

Each Variable Region Is Encoded by More Than One Gene Segment

When genomic DNA sequences encoding V and C regions were first analyzed, it was found that a single region of DNA encodes the C region of an antibody chain (see Figure 24-33), but two or more regions of DNA have to be assembled to encode each V region. Each light-chain V region is encoded by a DNA sequence assembled from two gene segments—a long ***V* gene segment** and a short *joining*, or ***J* gene segment** (not to be confused with the protein *J chain* (see Figure 24-23), which is encoded elsewhere in the genome). The *V-J* joining process involved in making a human κ light chain. In the "germ-line" DNA (where the antibody genes are not being expressed and are therefore not rearranged), the cluster of five *J* gene segments is separated from (more...)

Each heavy-chain V region is encoded by a DNA sequence assembled from three gene segments—a *V* segment, a *J* segment, and a *diversity segment*, or ***D* gene segment**

The large number of inherited *V, J,* and *D* gene segments available for encoding antibody chains makes a substantial contribution on its own to antibody diversity, but the combinatorial joining of these segments (called *combinatorial diversification)* greatly increases this contribution. Any of the 40 *V* segments in the human κ light-chain gene-segment pool, for example, can be joined to any of the 5 *J* segments, so that at least 200 (40 × 5) different κ-chain V regions can be encoded by this pool. Similarly, any of the 51 *V* segments in the human heavy-chain pool can be joined to any of the 6 *J* segments and any of the 27 *D* segments to encode at least 8262 (51 × 6 × 27) different heavy-chain V regions.

The combinatorial diversification resulting from the assembly of different combinations of inherited *V, J,* and *D* gene segments just discussed is an important mechanism for diversifying the antigen-binding sites of antibodies. By this mechanism alone, a human can produce 287 different V_L regions (200 κ and 116 λ) and 8262 different V_H regions. In principle, these could then be combined to make about 2.6×10^6 (316 × 8262) different antigen-binding sites. In addition, as we discuss next, the joining mechanism itself greatly increases this number of possibilities (probably more than 10^8-fold), making it much greater than the total number of B cells (about 10^{12}) in a human.

Imprecise Joining of Gene Segments Greatly Increases the Diversity of V Regions

During B cell development, the *V* and *J* gene segments (for the light chain) and the *V, D,* and *J* gene segments (for the heavy chain) are joined together to form a functional V_L- or V_H-region coding sequence by a process of site-specific recombination called **V(D)J joining**.Conserved DNA sequences flank each gene segment and serve as recognition sites for the joining process, ensuring that only appropriate gene segments recombine. Thus, for example, a *V* segment will always join to a *J* or *D* segment but not to another *V* segment. Joining is mediated by an enzyme complex called the **V(D)J recombinase**. This complex contains two proteins that are specific to developing lymphocytes, as well as enzymes that help repair damaged DNA in all our cells.

The lymphocyte-specific proteins of the V(D)J recombinase are encoded by two closely linked genes called *rag-1* and *rag-2 (rag = r*ecombination *a*ctivating *g*enes). The **RAG proteins** introduce double-strand breaks at the flanking DNA sequences, and this is followed

by a rejoining process that is mediated by both the RAG proteins and the enzymes involved in general DNA double-strand repair (discussed in Chapter 5). Thus, if both *rag* genes are artificially expressed in a fibroblast, the fibroblast is now able to rearrange experimentally introduced antibody gene segments just as a developing B cell normally does. Moreover, individuals who are deficient in either *rag* gene or in one of the general repair enzymes are highly susceptible to infection because they are unable to carry out *V(D)J* joining and consequently do not have functional B or T cells. (T cells use the same recombinase to assemble the gene segments that encode their antigen-specific receptors.)

In most cases of site-specific recombination, DNA joining is precise. But during the joining of antibody (and T cell receptor) gene segments, a variable number of nucleotides are often lost from the ends of the recombining gene segments, and one or more randomly chosen nucleotides may also be inserted. This random loss and gain of nucleotides at joining sites is called **junctional diversification**, and it enormously increases the diversity of V-region coding sequences created by recombination, specifically in the third hypervariable region. This increased diversification comes at a price, however. In many cases, it will result in a shift in the reading frame that produces a nonfunctional gene. Because roughly two in every three rearrangements are "nonproductive" in this way, many developing B cells never make a functional antibody molecule and consequently die in the bone marrow. B cells making functional antibody molecules that bind strongly to self antigens in the bone marrow are stimulated to re-express the RAG proteins and undergo a second round of *V(D)J* rearrangements, thereby changing the specificity of the cell-surface antibody they make—a process referred to as **receptor editing**. Self-reactive B cells that fail to change their specificity in this way are eliminated through the process of clonal deletion.

Antigen-Driven Somatic Hypermutation Fine-Tunes Antibody Responses

As mentioned earlier, with the passage of time after immunization, there is usually a progressive increase in the affinity of the antibodies produced against the immunizing antigen. This phenomenon, known as affinity maturation, is due to the accumulation of point mutations specifically in both heavy-chain and light-chain V-region coding sequences. The mutations occur long after the coding regions have been assembled, when B cells are stimulated by antigen and helper T cells to generate memory cells in a lymphoid follicle in a peripheral lymphoid organ (see Figure 24-16). They occur at the rate of about one per V-region coding sequence per cell generation. Because this is about a million times greater than the spontaneous mutation rate in other genes, the process is called **somatic hypermutation**. The molecular mechanism is still uncertain, but it is believed to involve some form of error-prone DNA repair process targeted to the rearranged V-region coding sequence by specific regions of DNA brought together by V(D)J joining. Surprisingly, an enzyme involved in RNA editing is required, but its function in the hypermutation process is unknown.

Only a small minority of the altered antigen receptors generated by hypermutation have an increased affinity for the antigen. The few B cells expressing these higher-affinity receptors, however, are preferentially stimulated by the antigen to survive and proliferate, whereas most other B cells die by apoptosis. Thus, as a result of repeated cycles of somatic hypermutation, followed by antigen-driven proliferation of selected clones of memory B cells, antibodies of

increasingly higher affinity become abundant during an **immune** response, providing progressively better protection against the pathogen.

The Control of V(D)J Joining Ensures That B Cells Are Monospecific

As the clonal selection theory predicts, B cells are *monospecific*. That is, all the antibodies that any one B cell produces have identical antigen-binding sites. This property enables antibodies to cross-link antigens into large aggregates, thereby promoting antigen elimination. It also means that an activated B cell secretes antibodies with the same specificity as that of the membrane-bound antibody on the B cell that was originally stimulated.

The requirement of monospecificity means that each B cell can make only one type of V_L region and one type of V_H region. Since B cells, like most other somatic cells, are diploid, each cell has six gene-segment pools encoding antibody chains: two heavy-chain pools (one from each parent) and four light-chain pools (one κ and one λ from each parent). If DNA rearrangements occurred independently in each heavy-chain pool and each light-chain pool, a single cell could make up to eight different antibodies, each with a different antigen-binding site.

In fact, however, each B cell uses only two of the six gene-segment pools: one of the two heavy-chain pools and one of the four light-chain pools. Thus, each B cell must choose not only between its κ and λ light-chain pools, but also between its maternal and paternal light-chain and heavy-chain pools. This second choice is called allelic exclusion, and it also occurs in the expression of genes that encode T cell receptors. For most other proteins that are encoded by autosomal genes, both maternal and paternal genes in a cell are expressed about equally.

Allelic exclusion and κ versus λ light-chain choice during B cell development depend on negative feedback regulation of the *V(D)J* joining process. A functional rearrangement in one gene-segment pool suppresses rearrangements in all remaining pools that encode the same type of polypeptide chain (Figure 24-40). In B cell clones isolated from transgenic mice expressing a rearranged μ-chain gene, for example, the rearrangement of endogenous heavy-chain genes is usually suppressed. Comparable results have been obtained for light chains. The suppression does not occur if the product of the rearranged gene fails to assemble into a receptor that inserts into the plasma membrane. It has therefore been proposed that either the receptor assembly process itself or extracellular signals that act on the receptor are involved in the suppression of further gene rearrangements.

Although no biological differences between the constant regions of κ and λ light chains have been discovered, there is an advantage in having two separate pools of gene segments encoding light chain variable regions. Having two separate pools increases the chance that a pre-B cell that has successfully assembled a V_H-region coding sequence will go on to assemble successfully a V_L-region coding sequence to become a B cell. This chance is further increased because, before a developing pre-B cell produces ordinary light chains, it makes surrogate light, which assemble with μ heavy chains. The resulting receptors are displayed on the cell surface and allow the cell to proliferate, producing large numbers of progeny cells, some of which are likely to succeed in producing bona fide light chains.

When Activated by Antigen, a B Cell Switches From Making a Membrane-Bound Antibody to Making a Secreted Form of the Same Antibody

We now turn from the genetic mechanisms that determine the antigen-binding site of an antibody to those that determine its biological properties—that is, those that determine what form of heavy-chain constant region is synthesized. The choice of the particular gene segments that encode the antigen-binding site is a commitment for the life of a B cell and its progeny, but the type of C_H region that is made changes during B cell development. The changes are of two types: changes from a membrane-bound form to a secreted form of the same C_H region and changes in the class of the C_H region made.

All classes of antibody can be made in a membrane-bound form, as well as in a soluble, secreted form. The membrane-bound form serves as an antigen receptor on the B cell surface, while the soluble form is made only after the cell is activated by antigen to become an antibody-secreting effector cell. The sole difference between the two forms resides in the C-terminus of the heavy chain. The heavy chains of membrane-bound antibody molecules have a hydrophobic C-terminus, which anchors them in the lipid bilayer of the B cell's plasma membrane. The heavy chains of secreted antibody molecules, by contrast, have instead a hydrophilic C-terminus, which allows them to escape from the cell. The switch in the character of the antibody molecules made occurs because the activation of B cells by antigen (and helper T cells) induces a change in the way in which the H-chain RNA transcripts are made and processed in the nucleus.

B Cells Can Switch the Class of Antibody They Make

During B cell development, many B cells switch from making one class of antibody to making another—a process called class switching. All B cells begin their antibody-synthesizing lives by making IgM molecules and inserting them into the plasma membrane as receptors for antigen. After the B cells leave the bone marrow, but before they interact with antigen, they switch and make both IgM and IgD molecules as membrane-bound antigen receptors, both with the same antigen-binding sites. On stimulation by antigen and helper T cells, some of these cells are activated to secrete IgM antibodies, which dominate the primary antibody response. Later in the **immune** response, the combination of antigen and the cytokines that helper T cells secrete induce many B cells to switch to making IgG, IgE, or IgA antibodies. These cells generate both memory cells that express the corresponding classes of antibody molecules on their surface and effector cells that secrete the antibodies. The IgG, IgE, and IgA molecules are collectively referred to as *secondary* classes of antibodies, both because they are produced only after antigen stimulation and because they dominate secondary antibody responses. As we saw earlier, each different class of antibody is specialized to attack microbes in different ways and in different sites.

The constant region of an antibody heavy chain determines the class of the antibody. Thus, the ability of B cells to switch the class of antibody they make without changing the antigen-binding site implies that the same assembled V_H-region coding sequence (which specifies the antigen-binding part of the heavy chain) can sequentially associate with different C_H-coding sequences. This has important functional implications. It means that, in an individual animal, a particular antigen-binding site that has been selected by environmental antigens can be

distributed among the various classes of antibodies, thereby acquiring the different biological properties of each class.

When a B cell switches from making IgM and IgD to one of the secondary classes of antibody, an irreversible change at the DNA level occurs—a process called *class switch recombination*. It entails deletion of all the C_H-coding sequences between the assembled VDJ-coding sequence and the particular C_H-coding sequence that the cell is destined to expres). Switch recombination differs from *V(D)J* joining in several ways: (1) it involves noncoding sequences only and therefore leaves the coding sequence unaffected; (2) it uses different flanking recombination sequences and different enzymes; (3) it happens after antigen stimulation; and (4) it is dependent on helper T cells.

Summary

Antibodies are produced from three pools of gene segments and exons. One pool encodes κ light chains, one encodes λ light chains, and one encodes heavy chains. In each pool, separate gene segments that code for different parts of the variable region of the light or heavy chains are brought together by site-specific recombination during B cell development. The light-chain pools contain one or more constant- (C-) region exons and sets of variable (V) and joining (J) gene segments. The heavy-chain pool contains sets of C-region exons and sets of V, diversity (D), and J gene segments.

To make an antibody molecule, a V_L gene segment recombines with a J_L gene segment to produce a DNA sequence coding for the V region of a light chain, and a V_H gene segment recombines with a D and a J_H gene segment to produce a DNA sequence coding for the V region of a heavy chain. Each of the assembled V-region coding sequences is then cotranscribed with the appropriate C-region sequence to produce an RNA molecule that codes for the complete polypeptide chain. Cells making functional heavy and light chains turn off the V(D)J joining process to ensure that each B cell makes only one species of antigen-binding site.

By randomly combining inherited gene segments that code for V_L and V_H regions, humans can make hundreds of different light chains and thousands of different heavy chains. Because the antigen-binding site is formed where the hypervariable loops of the V_L and V_H come together in the final antibody, the heavy and light chains can pair to form antibodies with millions of different antigen-binding sites. This number is enormously increased by the loss and gain of nucleotides at the site of gene-segment joining, as well as by somatic mutations that occur with very high frequency in the assembled V-region coding sequences after stimulation by antigen and helper T cells.

All B cells initially make IgM antibodies, and most then make IgD as well. Later many switch and make antibodies of other classes but with the same antigen-binding site as the original IgM and IgD antibodies. Such class switching depends on antigen stimulation and helper T cells, and it allows the same antigen-binding sites to be distributed among antibodies with varied biological properties.

T Cells and MHC Proteins

The diverse responses of T cells are collectively called *cell-mediated **immune** reactions*. This is to distinguish them from antibody responses, which, of course, also depend on cells (B cells). Like antibody responses, T cell responses are exquisitely antigen-specific, and they are at least as important as antibodies in defending vertebrates against infection. Indeed, most adaptive **immune** responses, including antibody responses, require helper T cells for their initiation. Most importantly, unlike B cells, T cells can help eliminate pathogens that reside inside host cells. Much of the rest of this chapter is concerned with how T cells accomplish this feat.

T cell responses differ from B cell responses in at least two crucial ways. First, T cells are activated by foreign antigen to proliferate and differentiate into effector cells only when the antigen is displayed on the surface of antigen-presenting cells in peripheral lymphoid organs. The T cells respond in this manner because the form of antigen they recognize is different from that recognized by B cells. Whereas B cells recognize intact antigen, T cells recognize fragments of protein antigens that have been partly degraded inside the antigen-presenting cell. The peptide fragments are then carried to the surface of the presenting cell on special molecules called *MHC proteins*, which present the fragments to T cells. The second difference is that, once activated, effector T cells act only at short range, either within a secondary lymphoid organ or after they have migrated into a site of infection. They interact directly with another cell in the body, which they either kill or signal in some way (we shall refer to such cells as *target cells).* Activated B cells, by contrast, secrete antibodies that can act far away.

There are two main classes of T cells—cytotoxic T cells and helper T cells. Effector *cytotoxic T cells* directly kill cells that are infected with a virus or some other intracellular pathogen. Effector *helper T cells*, by contrast, help stimulate the responses of other cells—mainly macrophages, B cells, and cytotoxic T cells.

In this section, we describe these two classes of T cells and their respective functions. We discuss how they recognize foreign antigens on the surface of antigen-presenting cells and target cells and consider the crucial part played by MHC proteins in the recognition process. Finally, we describe how T cells are selected during their development in the thymus to ensure that only cells with potentially useful receptors survive and mature. We begin by considering the nature of the cell-surface receptors that T cells use to recognize antigen.

T Cell Receptors Are Antibodylike Heterodimers

Because T cell responses depend on direct contact with an antigen-presenting cell or a target cell, the antigen receptors made by T cells, unlike antibodies made by B cells, exist only in membrane-bound form and are not secreted. For this reason, T cell receptors were difficult to isolate, and it was not until the 1980s that they were first identified biochemically. On both cytotoxic and helper T cells, the receptors are similar to antibodies. They are composed of two disulfide-linked polypeptide chains (called α and β), each of which contains two Ig-like

domains, one variable and one constant. Moreover, the three-dimensional structure of the extracellular part of a T cell receptor has been determined by x-ray diffraction, and it looks very much like one arm of a Y-shaped antibody molecule.

The pools of gene segments that encode the α and β chains are located on different chromosomes. Like antibody heavy-chain pools, the T cell receptor pools contain separate *V, D,* and *J* gene segments, which are brought together by site-specific recombination during T cell development in the thymus. With one exception, all the mechanisms used by B cells to generate antibody diversity are also used by T cells to generate T cell receptor diversity. Indeed, the same *V(D)J* recombinase is used, including the RAG proteins discussed earlier. The mechanism that does not operate in T cell receptor diversification is antigen-driven somatic hypermutation. Thus, the affinity of the receptors remains low ($K_a \sim 10^5$-10^7 liters/mole), even late in an **immune** response. We discuss later how various co-receptors and cell-cell adhesion mechanisms greatly strengthen the binding of a T cell to an antigen-presenting cell or a target cell, helping to compensate for the low affinity of the T cell receptors.

A small minority of T cells, instead of making α and β chains, make a different but related type of receptor heterodimer, composed of γ and δ chains. These cells arise early in development and are found mainly in epithelia (in the skin and gut, for example). Their functions are uncertain, and we shall not discuss them further.

As with antigen receptors on B cells, the T cell receptors are tightly associated in the plasma membrane with a number of invariant membrane-bound proteins that are involved in passing the signal from an antigen-activated receptor to the cell interior. We discuss these proteins in more detail later. First, however, we need to consider how cytotoxic and helper T cells function and the special ways in which they recognize foreign antigen.

Antigen-Presenting Cells Activate T Cells

Before cytotoxic or helper T cells can kill or help their target cells, respectively, they must be activated to proliferate and differentiate into effector cells. This activation occurs in peripheral lymphoid organs on the surface of antigen-presenting cells that display foreign antigen complexed with MHC proteins on their surface.

There are three main types of antigen-presenting cells in peripheral lymphoid organs that can activate T cells—dendritic cells, macrophages, and B cells. The most potent of these are dendritic cells, whose only known function is to present foreign antigens to T cells. Immature dendritic cells are located in tissues throughout the body, including the skin, gut, and respiratory tract. When they encounter invading microbes at these sites, they endocytose the pathogens or their products and carry them via the lymph to local lymph nodes or gut-associated lymphoid organs. The encounter with a pathogen induces the dendritic cell to mature from an antigen-capturing cell to an antigen-presenting cell that can activate T cells.

Antigen-presenting cells display three types of protein molecules on their surface that have a role in activating a T cell to become an effector cell: (1) *MHC proteins*, which present foreign antigen to the T cell receptor, (2) *costimulatory proteins*, which bind

to complementary receptors on the T cell surface, and (3) *cell-cell adhesion molecules*, which enable a T cell to bind to the antigen-presenting cell for long enough to become activated.

Effector Cytotoxic T Cells Induce Infected Target Cells to Kill Themselves

Cytotoxic T cells provide protection against intracellular pathogens such as viruses and some bacteria and parasites that multiply in the host-cell cytoplasm, where they are sheltered from attack by antibodies. They provide this protection by killing the infected cell before the microbes can proliferate and escape from the infected cell to infect neighboring cells.

Once a cytotoxic T cell has been activated by an infected antigen-presenting cell to become an effector cell, it can kill any target cell infected with the same pathogen. When the effector T cell recognizes a microbial antigen on the surface of an infected target cell, it focuses its secretory apparatus on the target. We can observe this behavior by studying effector T cells bound to their targets: when labeled with anti-tubulin antibodies, the T cell centrosome is seen to be oriented toward the point of contact with the target cell. Moreover, antibody labeling shows that talin and other proteins that help link cell-surface receptors to cortical actin filaments are concentrated in the cortex of the T cell at the contact site. The aggregation of T cell receptors at the contact site apparently leads to a local alteration in the actin filaments in the cell cortex. A microtubule-dependent mechanism then moves the centrosome and its associated Golgi apparatus toward the contact site, focusing the killing machinery on the target cell. A similar cytoskeletal polarization is seen when an effector helper T cell interacts functionally with a target cell.

Once bound to its target cell, a cytotoxic T cell can employ at least two strategies to kill the target, both of which operate by inducing the target cell to kill itself by undergoing apoptosis. In killing an infected target cell, the cytotoxic T cell usually releases a pore-forming protein called **perforin**, which is homologous to the complement component C9 and polymerizes in the target cell plasma membrane to form transmembrane channels. Perforin is stored in secretory vesicles of the cytotoxic T cell and is released by local exocytosis at the point of contact with the target cell. The secretory vesicles also contain serine proteases, which are thought to enter the target cell cytosol through the perforin channels. One of the proteases, called *granzyme B,* cleaves, and thereby activates, one or more members of the *caspase family* of proteases that mediate apoptosis. These caspases then activate other caspases, producing a proteolytic cascade that helps kill the cell (discussed in Chapter 17) (Figure 24-46A). Mice in which the perforin gene is inactivated cannot generate microbe-specific cytotoxic T cells and show increased susceptibility to certain viral and intracellular bacterial infections.

In the second killing strategy, the cytotoxic T cell also activates a death-inducing caspase cascade in the target cell but does it less directly. A homotrimeric protein on the cytotoxic T cell surface called **Fas ligand** binds to transmembrane receptor proteins on the target cell called **Fas**. The binding alters the Fas proteins so that their clustered cytosolic tails recruit procaspase-8 into the complex via an adaptor protein. The recruited procaspase-8 molecules cross-cleave and activate each other to begin the caspase cascade that leads to apoptosis. Cytotoxic T cells apparently use this killing strategy to help contain an **immune** response once it is well underway, by killing excessive effector lymphocytes, especially effector T cells: if the gene encoding either Fas or Fas ligand is inactivated by mutation, effector

lymphocytes accumulate in vast numbers in the spleen and lymph nodes, which become enormously enlarged.

Effector Helper T Cells Help Activate Macrophages, B Cells, and Cytotoxic T Cells

In contrast to cytotoxic T cells, helper T cells are crucial for defense against both extracellular and intracellular pathogens. They help stimulate B cells to make antibodies that help inactivate or eliminate extracellular pathogens and their toxic products. They activate macrophages to destroy any intracellular pathogen multiplying within the macrophage's phagosomes, and they help activate cytotoxic T cells to kill infected target cells.

Once a helper T cell has been activated by an antigen-presenting cell to become an effector cell, it can then help activate other cells. It does this both by secreting a variety of cytokines and by displaying costimulatory proteins on its surface. When activated by an antigen-presenting cell, a naïve helper T cell can differentiate into either of two distinct types of effector helper cell, called T_H1 and T_H2. *T_H1 cells* mainly help activate macrophages and cytotoxic T cells, whereas *T_H2 cells* mainly help activate B cells. As we discuss later, the nature of the invading pathogen and the types of innate **immune** responses it elicits largely determine which type of helper T cell develops. This, in turn, determines the nature of the adaptive **immune** responses mobilized to fight the invaders.

T Cells Recognize Foreign Peptides Bound to MHC Proteins

As discussed earlier, both cytotoxic T cells and helper T cells are initially activated in peripheral lymphoid organs by recognizing foreign antigen on the surface of an antigen-presenting cell, usually a dendritic cell. The antigen is in the form of peptide fragments that are generated by the degradation of foreign protein antigens inside the antigen-presenting cell. The recognition process depends on the presence in the antigen-presenting cell of **MHC proteins**, which bind these fragments, carry them to the cell surface, and present them there, along with a co-stimulatory signal, to the T cells. Once activated, effector T cells then recognize the same peptide-MHC complex on the surface of the target cell they influence, which may be a B cell, a cytotoxic T cell, or an infected macrophage in the case of a helper T cell, or a virus-infected cell in the case of a cytotoxic T cell.

MHC proteins are encoded by a large complex of genes called the **major histocompatibility complex (MHC)**. There are two main structurally and functionally distinct classes of MHC proteins: *class I MHC proteins,* which present foreign peptides to cytotoxic T cells, and *class II MHC proteins,* which present foreign peptides to helper cells.

Before examining the different mechanisms by which protein antigens are processed for display to the two main classes of T cells, we must look more closely at the MHC proteins themselves, which have such an important role in T cell function.

MHC Proteins Were Identified in Transplantation Reactions Before Their Functions Were Known

MHC proteins were initially identified as the main antigens recognized in **transplantation reactions**. When organ grafts are exchanged between adult individuals, either of the same species *(allografts)* or of different species *(xenografts)*, they are usually rejected. In the 1950s, skin grafting experiments between different strains of mice demonstrated that *graft rejection* is an adaptive **immune** response to the foreign antigens on the surface of the grafted cells. Rejection is mediated mainly by T cells, which react against genetically "foreign" versions of cell-surface proteins called *histocompatibility molecules* (from the Greek word *histos*, meaning "tissue"). The MHC proteins encoded by the clustered genes of the major histocompatibility complex (MHC) are by far the most important of these. MHC proteins are expressed on the cells of all higher vertebrates. They were first demonstrated in mice, where they are called *H-2 antigens (h*istocompatibility-*2* antigens). In humans they are called *HLA antigens (h*uman-*l*eucocyte-*a*ssociated antigens) because they were first demonstrated on leucocytes (white blood cells).

Three remarkable properties of MHC proteins baffled immunologists for a long time. First, MHC proteins are overwhelmingly the preferred antigens recognized in T-cell-mediated transplantation reactions. Second, an unusually large fraction of T cells are able to recognize foreign MHC proteins: whereas fewer than 0.001% of an individual's T cells respond to a typical viral antigen, more than 0.1% of them respond to a single foreign MHC antigen. Third, some of the genes that code for MHC proteins are the most *polymorphic* known in higher vertebrates. That is, within a species, there is an extraordinarily large number of *alleles* (alternative forms of the same gene) present (in some cases more than 200), without any one allele predominating. As each individual has at least 12 genes encoding MHC proteins (see later), it is very rare for two unrelated individuals to have an identical set of MHC proteins. This makes it very difficult to match donor and recipient for organ transplantation unless they are closely related.

Of course, a vertebrate does not need to protect itself against invasion by foreign vertebrate cells. So the apparent obsession of its T cells with foreign MHC proteins and the extreme polymorphism of these molecules were a great puzzle. The puzzle was solved only after it was discovered that (1) MHC proteins bind fragments of foreign proteins and display them on the surface of host cells for T cells to recognize, and (2) T cells respond to foreign MHC proteins in the same way they respond to self MHC proteins that have foreign antigen bound to them.

Class I and Class II MHC Proteins Are Structurally Similar Heterodimers

Class I and class II MHC proteins have very similar overall structures. They are both transmembrane heterodimers with extracellular N-terminal domains that bind antigen for presentation to T cells.

Class I MHC proteins consist of a transmembrane α chain, which is encoded by a class I MHC gene, and a small extracellular protein called β_2-*microglobulin*. The β_2-microglobulin does not span the membrane and is encoded by a gene that does not lie in the MHC gene cluster. The α chain is folded into three extracellular globular domains (α_1, α_2, α_3), and the α_3 domain and the β_2-microglobulin, which are closest to the membrane, are both similar to an Ig domain. The two N-terminal domains of the α chain, which are farthest from the

membrane, contain the polymorphic (variable) amino acids that are recognized by T cells in transplantation reactions. These domains bind a peptide and present it to cytotoxic T cells.

Like class I MHC proteins, **class II MHC proteins**are heterodimers with two conserved Ig-like domains close to the membrane and two polymorphic (variable) N-terminal domains farthest from the membrane. In these proteins, however, both chains (α and β) are encoded by genes within the MHC, and both span the membrane. The two polymorphic domains bind a peptide and present it to helper T cells.

The presence of Ig-like domains in class I and class II proteins suggests that MHC proteins and antibodies have a common evolutionary history. The locations of the genes that encode class I and class II MHC proteins in humans are shown in, where we illustrate how an individual can make six types of class I MHC proteins and more than six types of class II proteins.

In addition to the classic class I MHC proteins, there are many *nonclassical class I MHC proteins*, which form dimers with β2-microglobulin. These proteins are not polymorphic, but some of them present specific microbial antigens, including some lipids and glycolipids, to T cells. The functions of most of them, however, are unknown.

An MHC Protein Binds a Peptide and Interacts with a T Cell Receptor

Any individual can make only a small number of different MHC proteins, which together must be able to present peptide fragments from almost any foreign protein to T cells. Thus, unlike an antibody molecule, each MHC protein has to be able to bind a very large number of different peptides. The structural basis for this versatility has emerged from x-ray crystallographic analyses of MHC proteins.

a class I MHC protein has a single peptide-binding site located at one end of the molecule, facing away from the plasma membrane. This site consists of a deep groove between two long α helices; the groove narrows at both ends so that it is only large enough to accommodate an extended peptide about 8–10 amino acids long. In fact, when a class I MHC protein was first analyzed by x-ray crystallography in 1987, this groove contained bound peptides that had co-crystallized with the MHC protein, suggesting that once a peptide binds to this site it does not normally dissociate.

A typical peptide binds in the groove of a class I MHC protein in an extended conformation, with its terminal amino group bound to an invariant pocket at one end of the groove and its terminal carboxyl group bound to an invariant pocket at the other end of the groove. Other amino acids (called “anchor amino acids”) in the peptide bind to “specificity pockets” in the groove formed by polymorphic portions of the MHC protein. The side chains of other amino acids of the peptide point outward, in a position to be recognized by receptors on cytotoxic T cells. Because the conserved pockets at the ends of the binding groove recognize features of the peptide backbone that are common to all peptides, each allelic form of a class I MHC protein can bind a large variety of peptides of diverse sequence. At the same time, the differing specificity pockets along the groove, which bind particular amino acid side chains of the peptide, ensure that each allelic form binds and presents a distinct characteristic set of peptides. Thus, the six types of class I MHC proteins in an individual can present a broad

range of foreign peptides to the cytotoxic T cells, but in each individual they do so in slightly different ways.

Class II MHC proteins have a three-dimensional structure that is very similar to that of class I proteins, but their antigen-binding groove does not narrow at the ends, so it can accommodate longer peptides, which are usually 13–17 amino acids long. Moreover, the peptide is not bound at its ends. It is held in the groove by parts of its peptide backbone that bind to invariant pockets formed by conserved amino acids that line all class II MHC peptide-binding grooves, as well as by the side chains of anchor amino acids that bind to variable specificity pockets in the groove. A class II MHC binding groove can accommodate a more heterogeneous set of peptides than can a class I MHC groove. Thus, although an individual makes only a small number of types of class II proteins, each with its own unique peptide-binding groove, together these proteins can bind and present an enormous variety of foreign peptides to helper T cells, which have a crucial role in almost all adaptive **immune** responses.

The way in which the T cell receptor recognizes a peptide fragment bound to an MHC protein is revealed by x-ray crystallographic analyses of complexes formed between a soluble receptor and a soluble MHC protein with peptide in its binding groove. (The soluble proteins for these experiments are produced by recombinant DNA technology.) In each case studied, the T cell receptor fits diagonally across the peptide-binding groove and binds through its V_α and V_β hypervariable loops to both the walls of the groove and the peptide. Soluble MHC-peptide complexes are now widely used to detect T cells with a particular specificity; they are usually cross-linked into tetramers to increase their avidity for T cell receptors.

MHC Proteins Help Direct T Cells to Their Appropriate Targets

Class I MHC proteins are expressed on virtually all nucleated cells. This is presumably because effector cytotoxic T cells must be able to focus on and kill any cell in the body that happens to become infected with an intracellular microbe such as a virus. Class II proteins, by contrast, are normally confined largely to cells that take up foreign antigens from the extracellular fluid and interact with helper T cells. These include dendritic cells, which initially activate helper T cells, as well as the targets of effector helper T cells, such as macrophages and B cells. Because dendritic cells express both class I and class II MHC proteins, they can activate both cytotoxic and helper T cells.

It is important that effector cytotoxic T cells focus their attack on cells that *make* the foreign antigens (such as viral proteins), while helper T cells focus their help mainly on cells that have taken up foreign antigens from the extracellular fluid. Since the former type of target cell is always a menace, while the latter type is essential for the body's **immune** defenses, it is vitally important that T cells never confuse the two target cells and misdirect their cytotoxic and helper functions. Therefore, in addition to the antigen receptor that recognizes a peptide-MHC complex, each of the two major classes of T cells also expresses a *co-receptor* that recognizes a separate, invariant part of the appropriate class of MHC protein. These two co-receptors, called CD4 and CD8, help direct helper T cells and cytotoxic T cells, respectively, to their appropriate targets, as we now discuss. The properies of class I and class II MHC proteins are compared in.

CD4 and CD8 Co-receptors Bind to Nonvariable Parts of MHC Proteins

The affinity of T cell receptors for peptide-MHC complexes on an antigen-presenting cell or target cell is usually too low to mediate a functional interaction between the two cells by itself. T cells normally require *accessory receptors* to help stabilize the interaction by increasing the overall strength of the cell-cell adhesion. Unlike T cell receptors or MHC proteins, the accessory receptors do not bind foreign antigens and are invariant.

When accessory receptors also have a direct role in activating the T cell by generating their own intracellular signals, they are called **co-receptors**. The most important and best understood of the co-receptors on T cells are the CD4 and CD8 proteins, both of which are single-pass transmembrane proteins with extracellular Ig-like domains. Like T cell receptors, they recognize MHC proteins, but, unlike T cell receptors, they bind to nonvariable parts of the protein, far away from the peptide-binding groove. CD4 is expressed on helper T cells and binds to class II MHC proteins, whereas CD8 is expressed on cytotoxic T cells and binds to class I MHC proteins. Thus, CD4 and CD8 contribute to T cell recognition by helping to focus the cell on particular MHC proteins, and thus on particular types of cells—helper T cells on dendritic cells, macrophages, and B cells, and cytotoxic cells on any nucleated host cell displaying a foreign peptide on a class I MHC protein. The cytoplasmic tail of these transmembrane proteins is associated with a member of the Src family of cytoplasmic tyrosine protein kinases called *Lck*, which phosphorylates various intracellular proteins on tyrosines and thereby participates in the activation of the T cell. Antibodies to CD4 and CD8 are widely used as tools to distinguish between the two main classes of T cells, in both humans and experimental animals.

Ironically, the AIDS virus (HIV) makes use of CD4 molecules (as well as chemokine receptors) to enter helper T cells. It is the eventual depletion of helper T cells that renders AIDS patients susceptible to infection by microbes that are not normally dangerous. As a result, most AIDS patients die of infection within several years of the onset of symptoms, unless they are treated with a combination of powerful anti-HIV drugs. HIV also uses CD4 and chemokine receptors to enter macrophages, which also have both of these receptors on their surface.

Before a cytotoxic or helper T cell can recognize a foreign protein, the protein has to be processed inside an antigen-presenting cell or target cell so that it can be displayed as peptide-MHC complexes on the cell surface. We first consider how a virus-infected antigen-presenting cell or target cell processes viral proteins for presentation to a cytotoxic T cell. We then discuss how ingested foreign proteins are processed for presentation to a helper T cell.

Cytotoxic T Cells Recognize Fragments of Foreign Cytosolic Proteins in Association with Class I MHC Proteins

One of the first, and most dramatic, demonstrations that MHC proteins present foreign antigens to T cells came from an experiment performed in the 1970s. It was found that effector cytotoxic T cells from a virus-infected mouse could kill cultured cells infected with

the same virus only if these target cells expressed some of the same class I MHC proteins as the infected mouse. This experiment demonstrated that the T cells of any individual that recognize a specific antigen do so only when that antigen is associated with the allelic forms of MHC proteins expressed by that individual, a phenomenon known as *MHC restriction.*

The chemical nature of the viral antigens recognized by cytotoxic T cells was not discovered for another 10 years. In experiments on cells infected with influenza virus, it was unexpectedly found that some of the effector cytotoxic T cells activated by the virus specifically recognize internal proteins of the virus that would not be accessible in the intact virus particle. Subsequent evidence indicated that the T cells were recognizing degraded fragments of the internal viral proteins that were bound to class I MHC proteins on the infected cell surface. Because a T cell can recognize tiny amounts of antigen (as few as one hundred peptide-MHC complexes), only a small fraction of the fragments generated from viral proteins have to bind to class I MHC proteins and get to the cell surface to attract an attack by an effector cytotoxic T cell.

The viral proteins are synthesized in the cytosol of the infected cell. As discussed in Chapter 3, proteolytic degradation in the cytosol is mainly mediated by an ATP- and ubiquitin-dependent mechanism that operates in *proteasomes*—large proteolytic enzyme complexes constructed from many different protein subunits. Although all proteasomes are probably able to generate peptide fragments that can bind to class I MHC proteins, some proteasomes are thought to be specialized for this purpose, as they contain two subunits that are encoded by genes located within the MHC chromosomal region. Even bacterial proteasomes cut proteins into peptides of about the length that fits into the groove of a class I MHC protein, suggesting that the MHC groove evolved to fit this length of peptide.

How do peptides generated in the cytosol make contact with the peptide-binding groove of class I MHC proteins in the lumen of the endoplasmic reticulum? The answer was discovered through observations on mutant cells in which class I MHC proteins are not expressed at the cell surface but are instead degraded within the cell. The mutant genes in these cells proved to encode subunits of a protein belonging to the family of *ABC transporters*, which we discuss in Chapter 11. This transporter protein is located in the ER membrane and uses the energy of ATP hydrolysis to pump peptides from the cytosol into the ER lumen. The genes encoding its two subunits are in the MHC chromosomal region, and, if either gene is inactivated by mutation, cells are unable to supply peptides to class I MHC proteins. The class I MHC proteins in such mutant cells are degraded in the cell because peptide binding is normally required for the proper folding of these proteins. Until it binds a peptide, a class I MHC protein remains in the ER, tethered to an ABC transporter by a chaperone protein.

In cells that are not infected, peptide fragments come from the cells' own cytosolic and nuclear proteins that are degraded in the processes of normal protein turnover and quality control mechanisms. (Surprisingly, more than 30% of the proteins made by mammalian cells are apparently faulty and are degraded in proteasomes soon after they are synthesized.) These peptides are pumped into the ER and are carried to the cell surface by class I MHC proteins. They are not antigenic because the cytotoxic T cells that could recognize them have been eliminated or inactivated during T cell development, as we discuss later.

When cytotoxic T cells and some helper T cells are activated by antigen to become effector cells, they secrete the cytokine **interferon-γ (IFN-γ)**, which greatly enhances anti-viral responses. The IFN-γ acts on infected cells in two ways. It blocks viral replication, and it

increases the expression of many genes within the MHC chromosomal region. These genes include those that encode class I (and class II) MHC proteins, the two specialized proteasome subunits, and the two subunits of the peptide transporter located in the ER. Thus, all of the machinery required for presenting viral antigens to cytotoxic T cells is coordinately called into action by IFN-γ, creating a positive feedback that amplifies the **immune** response and culminates in the death of the infected cells.

Helper T Cells Recognize Fragments of Endocytosed Foreign Protein Associated with Class II MHC Proteins

Unlike cytotoxic T cells, helper T cells do not act directly to kill infected cells so as to eliminate microbes. Instead, they stimulate macrophages to be more effective in destroying intracellular microorganisms, and they help B cells and cytotoxic T cells to respond to microbial antigens.

Like the viral proteins presented to cytotoxic T cells, the proteins presented to helper T cells on antigen-presenting cells or target cells are degraded fragments of foreign proteins. The fragments are bound to class II MHC proteins in much the same way that virus-derived peptides are bound to class I MHC proteins. But both the source of the peptide fragments presented and the route they take to find the MHC proteins are different from those of peptide fragments presented by class I MHC proteins to cytotoxic T cells.

Rather than being derived from foreign protein synthesized in the cytosol of a cell, the foreign peptides presented to helper T cells are derived from endosomes. Some come from extracellular microbes or their products that the antigen-presenting cell has endocytosed and degraded in the acidic environment of its endosomes. Others come from microbes growing within the endocytic compartment of the antigen-presenting cell. These peptides do not have to be pumped across a membrane because they do not originate in the cytosol; they are generated in a compartment that is topologically equivalent to the extracellular space. They never enter the lumen of the ER, where the class II MHC proteins are synthesized and assembled, but instead bind to preassembled class II heterodimers in a special endosomal compartment. Once the peptide has bound, the class II MHC protein alters its conformation, trapping the peptide in the binding groove for presentation at the cell surface to helper T cells.

A newly synthesized class II MHC protein must avoid clogging its binding groove prematurely in the ER lumen with peptides derived from endogenously synthesized proteins. A special polypeptide, called the **invariant chain**, ensures this by associating with newly synthesized class II MHC heterodimers in the ER. Part of its polypeptide chain lies within the peptide-binding groove of the MHC protein, thereby blocking the groove from binding other peptides in the lumen of the ER. The invariant chain also directs class II MHC proteins from the *trans* Golgi network to a late endosomal compartment. Here, the invariant chain is cleaved by proteases, leaving only a short fragment bound in the peptide-binding groove of the MHC protein. This fragment is then released (catalyzed by a class II-MHC-like protein called HLA-DM), freeing the MHC protein to bind peptides derived from endocytosed proteins (Figure 24-60). In this way, the functional differences between class I and class II MHC proteins are ensured—the former presenting molecules that come from the cytosol, the latter presenting molecules that come from the endocytic compartment.

Most of the class I and class II MHC proteins on the surface of a target cell have peptides derived from self proteins in their binding groove. For class I proteins, the fragments derive from degraded cytosolic and nuclear proteins. For class II proteins, they mainly derive from degraded proteins that originate in the plasma membrane or extracellular fluid and are endocytosed. Only a small fraction of the 10^5 or so class II MHC proteins on the surface of an antigen-presenting cell have foreign peptides bound to them. This is sufficient, however, because only a hundred or so of such molecules are required to stimulate a helper T cell, just as in the case of peptide-class-I-MHC complexes stimulating a cytotoxic T cell.

Potentially Useful T Cells Are Positively Selected in the Thymus

We have seen that T cells recognize antigen in association with self MHC proteins but not in association with foreign MHC proteins: that is, T cells show *MHC restriction.* This restriction results from a process of **positive selection** during T cell development in the thymus. In this process, those immature T cells that will be capable of recognizing foreign peptides presented by self MHC proteins are selected to survive, while the remainder, which would be of no use to the animal, undergo apoptosis. Thus, MHC restriction is an acquired property of the **immune** system that emerges as T cells develop in the thymus.

The most direct way to study the selection process is to follow the fate of a set of developing T cells of known specificity. This can be done by using transgenic mice that express a specific pair of rearranged α and β T cell receptor genes derived from a T cell clone of known antigen and MHC specificity. Such experiments show that the transgenic T cells mature in the thymus and populate the peripheral lymphoid organs only if the transgenic mouse also expresses the same allelic form of MHC protein as is recognized by the transgenic T cell receptor. If the mouse does not express the appropriate MHC protein, the transgenic T cells die in the thymus. Thus, the survival and maturation of a T cell depend on a match between its receptor and the MHC proteins expressed in the thymus. Similar experiments using transgenic mice in which MHC expression is confined to specific cell types in the thymus indicate that it is MHC proteins on epithelial cells in the cortex of the thymus that are responsible for this positive selection process. After positively selected T cells leave the thymus, their continued survival depends on their continual stimulation by self-peptide-MHC complexes; this stimulation is enough to promote cell survival but not enough to activate the T cells to become effector cells.

As part of the positive selection process in the thymus, developing T cells that express receptors recognizing class I MHC proteins are selected to become cytotoxic cells, while T cells that express receptors recognizing class II MHC proteins are selected to become helper cells. Thus, genetically engineered mice that lack cell-surface class I MHC proteins specifically lack cytotoxic T cells, whereas mice that lack class II MHC proteins specifically lack helper T cells. The cells that are undergoing positive selection initially express both CD4 and CD8 co-receptors, and these are required for the selection process: without CD4, helper T cells fail to develop, and without CD8, cytotoxic T cells fail to develop.

Positive selection still leaves a large problem to be solved. If developing T cells with receptors that recognize self peptides associated with self MHC proteins were to mature in the thymus and migrate to peripheral lymphoid tissues, they might wreak havoc. A second, *negative selection* process in the thymus is required to help avoid this potential disaster.

Many Developing T Cells That Could Be Activated by Self Peptides Are Eliminated in the Thymus

As discussed previously, a fundamental feature of the adaptive **immune** system is that it can distinguish self from nonself and normally does not react against self molecules. An important mechanism in achieving this state of *immunological self tolerance* is the deletion in the thymus of developing self-reactive T cells—that is, T cells whose receptors bind strongly enough to the complex of a self peptide and a self MHC protein to become activated. Because, as we discuss later, most B cells require helper T cells to respond to antigen, the elimination of self-reactive helper T cells also helps ensure that self-reactive B cells that escape B cell tolerance induction are harmless.

It is not enough, therefore, for the thymus to select *for* T cells that recognize self MHC proteins; it must also select *against* T cells that could be activated by self MHC proteins complexed with self peptides. In other words, it must pick out for survival just those T cells that will be capable of responding to self MHC proteins complexed with foreign peptides, even though these peptides are not present in the developing thymus. It is thought that these T cells bind weakly in the thymus to self MHC proteins that are carrying self peptides mismatched to the T cell receptors. Thus, the required goal can be achieved by (1) ensuring the death of T cells that bind *strongly* to the self-peptide-MHC complexes in the thymus while (2) promoting the survival of those that bind weakly and (3) permitting the death of those that do not bind at all. Process 2 is the positive selection we have just discussed. Process 1 is called **negative selection**. In both death processes, the cells that die undergo apoptosis.

The most convincing evidence for negative selection derives once again from experiments with transgenic mice. After the introduction of T cell receptor transgenes encoding a receptor that recognizes a male-specific peptide antigen, for example, large numbers of mature T cells expressing the transgenic receptor are found in the thymus and peripheral lymphoid organs of female mice. Very few, however, are found in male mice, where the cells die in the thymus before they have a chance to mature. Like positive selection, negative selection requires the interaction of a T cell receptor and a CD4 or CD8 co-receptor with an appropriate MHC protein. Unlike positive selection, however, which occurs mainly on the surface of thymus epithelial cells, negative selection occurs on the surface of thymus dendritic cells and macrophages, which, as we have seen, function as antigen-presenting cells in peripheral lymphoid organs.

The deletion of self-reactive T cells in the thymus cannot eliminate all potentially self-reactive T cells, as some self molecules are not present in the thymus. Thus, some potentially self-reactive T cells are deleted or functionally inactivated after they leave the thymus, presumably because they recognize self peptides bound to MHC proteins on the surface of dendritic cells that have not been activated by microbes and therefore do not provide a costimulatory signal. As we discuss later, antigen recognition without costimulatory signals can delete or inactivate a T or B cell.

Some potentially self-reactive T cells, however, are not deleted or inactivated. Instead, special *regulatory* (or *suppressor) T cells* are thought to keep them from responding to their self antigens by secreting inhibitory cytokines such as TGF-β (discussed in Chapter 15).

These self-reactive T cells may sometimes escape from this suppression and cause autoimmune diseases.

The Function of MHC Proteins Explains Their Polymorphism

The role of MHC proteins in binding foreign peptides and presenting them to T cells provides an explanation for the extensive polymorphism of these proteins. In the evolutionary war between pathogenic microbes and the adaptive **immune** system, microbes tend to change their antigens to avoid associating with MHC proteins. When a microbe succeeds, it is able to sweep through a population as an epidemic. In such circumstances, the few individuals that produce a new MHC protein that can associate with an antigen of the altered microbe have a large selective advantage. In addition, individuals with two different alleles at any given MHC locus (heterozygotes) have a better chance of resisting infection than those with identical alleles at the locus, as they have a greater capacity to present peptides from a wide range of microbes and parasites. Thus, selection will tend to promote and maintain a large diversity of MHC proteins in the population. Strong support for this hypothesis, that infectious diseases have provided the driving force for MHC polymorphism, has come from studies in West Africa. Here, it is found that individuals with a specific MHC allele have a reduced susceptibility to a severe form of malaria. Although the allele is rare elsewhere, it is found in 25% of the West African population where this form of malaria is common.

If greater MHC diversity means greater resistance to infection, why do we each have so few MHC genes encoding these molecules? Why have we not evolved strategies for increasing the diversity of MHC proteins—by alternative RNA splicing, for example, or by the genetic recombination mechanisms used to diversify antibodies and T cell receptors? Presumably, the limits exist because each time a new MHC protein is added to the repertoire, the T cells that recognize self peptides in association with the new MHC protein must be eliminated to maintain self tolerance. The elimination of these T cells would counteract the advantage of adding the new MHC protein. Thus, the number of MHC proteins we express may represent a balance between the advantages of presenting a wide diversity of foreign peptides to T cells against the disadvantages of severely restricting the T cell repertoire during negative selection in the thymus. This explanation is supported by computer modeling studies.

Summary

There are two main functionally distinct classes of T cells: cytotoxic T cells kill infected cells directly by inducing them to undergo apoptosis, while helper T cells help activate B cells to make antibody responses and macrophages to destroy microorganisms that either invaded the macrophage or were ingested by it. Helper T cells also help activate cytotoxic T cells. Both classes of T cells express cell-surface, antibodylike receptors, which are encoded by genes that are assembled from multiple gene segments during T cell development in the thymus. These receptors recognize fragments of foreign proteins that are displayed on the surface of host cells in association with MHC proteins. Both cytotoxic and helper T cells are activated in peripheral lymphoid organs by antigen-presenting cells, which express peptide-MHC complexes, costimulatory proteins, and various cell-cell adhesion molecules on their cell surface.

Class I and class II MHC proteins have crucial roles in presenting foreign protein antigens to cytotoxic and helper T cells, respectively. Whereas class I proteins are expressed on almost all vertebrate cells, class II proteins are normally restricted to those cell types that interact with helper T cells, such as dendritic cells, macrophages, and B lymphocytes. Both classes of MHC proteins have a single peptide-binding groove, which binds small peptide fragments derived from proteins. Each MHC protein can bind a large and characteristic set of peptides, which are produced intracellularly by protein degradation: class I MHC proteins generally bind fragments produced in the cytosol, while class II MHC proteins bind fragments produced in the endocytic compartment. After they have formed inside the target cell, the peptide-MHC complexes are transported to the cell surface. Complexes that contain a peptide derived from a foreign protein are recognized by T cell receptors, which interact with both the peptide and the walls of the peptide-binding groove. T cells also express CD4 or CD8 co-receptors, which recognize nonpolymorphic regions of MHC proteins on the target cell: helper cells express CD4, which recognizes class II MHC proteins, while cytotoxic T cells express CD8, which recognizes class I MHC proteins.

The T cell receptor repertoire is shaped mainly by a combination of positive and negative selection processes that operate during T cell development in the thymus. These processes help to ensure that only T cells with potentially useful receptors survive and mature, while the others die by apoptosis. T cells that will be able to respond to foreign peptides complexed with self MHC proteins are positively selected, while many T cells that could react strongly with self peptides complexed with self MHC proteins are eliminated. T cells with receptors that could react strongly with self antigens not present in the thymus are eliminated, functionally inactivated, or actively kept suppressed after they leave the thymus.

Helper T Cells and Lymphocyte Activation

Helper T cells are arguably the most important cells in adaptive immunity, as they are required for almost all adaptive **immune** responses. They not only help activate B cells to secrete antibodies and macrophages to destroy ingested microbes, but they also help activate cytotoxic T cells to kill infected target cells. As dramatically demonstrated in AIDS patients, without helper T cells we cannot defend ourselves even against many microbes that are normally harmless.

Helper T cells themselves, however, can only function when activated to become effector cells. They are activated on the surface of antigen-presenting cells, which mature during the innate **immune** responses triggered by an infection. The innate responses also dictate what kind of effector cell a helper T cell will develop into and thereby determine the nature of the adaptive **immune** response elicited.

In this final section, we discuss the multiple signals that help activate a T cell and how a helper T cell, once activated to become an effector cell, helps activate other cells. We also consider how innate **immune** responses determine the nature of adaptive responses by stimulating helper T cells to differentiate into either T_H1 or T_H2 effector cells.

Costimulatory Proteins on Antigen-Presenting Cells Help Activate T Cells

To activate a cytotoxic or helper T cell to proliferate and differentiate into an effector cell, an antigen-presenting cell provides two kinds of signals. *Signal 1* is provided by a foreign peptide bound to an MHC protein on the surface of the presenting cell. This peptide-MHC complex signals through the T cell receptor and its associated proteins. *Signal 2* is provided by costimulatory proteins, especially the **B7 proteins** (CD80 and CD86), which are recognized by the co-receptor protein CD28 on the surface of the T cell. The expression of B7 proteins on an antigen-presenting cell is induced by pathogens during the innate response to an infection. Effector T cells act back to promote the expression of B7 proteins on antigen-presenting cells, creating a positive feedback loop that amplifies the T cell response.

Signal 2 is thought to amplify the intracellular signaling process triggered by signal 1. If a T cell receives signal 1 without signal 2, it may undergo apoptosis or become altered so that it can no longer be activated, even if it later receives both signals. This is one mechanism by which a T cell can become *tolerant* to self antigens.

The T cell receptor does not act on its own to transmit signal 1 into the cell. It is associated with a complex of invariant transmembrane proteins called **CD3**, which transduces the binding of the peptide-MHC complex into intracellular signals. In addition, the CD4 and CD8 co-receptors play important parts in the signaling process, as illustrated in.

The combined actions of signal 1 and signal 2 stimulate the T cell to proliferate and begin to differentiate into an effector cell by a curiously indirect mechanism. In culture, they cause the T cells to stimulate their own proliferation and differentiation by inducing the cells to secrete a cytokine called **interleukin-2 (IL-2)** and simultaneously to synthesize high affinity cell-surface receptors that bind it. The binding of IL-2 to the IL-2 receptors activates intracellular signaling pathways that turn on genes that help the T cells to proliferate and differentiate into effector cells (Figure 24-65). As discussed in Chapter 15, there are advantages to such an autocrine mechanism. It helps ensure that T cells differentiate into effector cells only when substantial numbers of them respond to antigen simultaneously in the same location, such as in a lymph node during an infection. Only then do IL-2 levels rise high enough to be effective.

Once bound to the surface of an antigen-presenting cell, a T cell increases the strength of the binding by activating an integrin adhesion protein called *lymphocyte-function-associated protein 1 (LFA-1).* Activated LFA-1 now binds more strongly to its Ig-like ligand, *intracellular adhesion molecule 1 (ICAM-1),* on the surface of the presenting cell. This increased adhesion enables the T cell to remain bound to the antigen-presenting cell long enough for the T cell to become activated.

The activation of a T cell is controlled by negative feedback. During the activation process, the cell starts to express another cell-surface protein called *CTLA-4,* which acts to inhibit intracellular signaling. It resembles CD28, but it binds to B7 proteins on the surface of the antigen-presenting cell with much higher affinity than does CD28, and, when it does, it holds the activation process in check. Mice with a disrupted *CTLA-4* gene die from a massive accumulation of activated T cells.

Most of the T (and B) effector cells produced during an **immune** response must be eliminated after they have done their job. As antigen levels fall and the response subsides, effector cells are deprived of the antigen and cytokine stimulation that they need to survive, and the majority die by apoptosis. Only memory cells and some long-lived effector cells survive.

Before considering how effector helper T cells help activate macrophages and B cells, we need to discuss the two functionally distinct subclasses of effector helper T cells, T_H1 and T_H2 cells, and how they are generated.

The Subclass of Effector Helper T Cell Determines the Nature of the Adaptive Immune Response

When a an antigen-presenting cell activates a naïve helper T cell in a peripheral lymphoid tissue, the T cell can differentiate into either a T_H1 or T_H2 effector helper cell. These two types of functionally distinct subclasses of effector helper T cells can be distinguished by the cytokines they secrete. If the cell differentiates into a **T_H1 cell**, it will secrete *interferon-γ (IFN-γ)* and *tumor necrosis factor-α* (*TNF-α*) and will activate macrophages to kill microbes located within the macrophages' phagosomes. It will also activate cytotoxic T cells to kill infected cells. Although, in these ways, T_H1 cells mainly defend an animal against intracellular pathogens, they may also stimulate B cells to secrete specific subclasses of IgG antibodies that can coat extracellular microbes and activate complement.

If the naïve T helper cell differentiates into a **T_H2** cell, by contrast, it will secrete *interleukins 4, 5, 10,* and *13* (*IL-4, IL-5, IL-10,* and *IL-13)* and will mainly defend the animal against extracellular pathogens. A T_H2 cell can stimulate B cells to make most classes of antibodies, including IgE and some subclasses of IgG antibodies that bind to mast cells, basophils, and eosinophils. These cells release local mediators that cause sneezing, coughing, or diarrhea and help expel extracellular microbes and larger parasites from epithelial surfaces of the body.

Thus, the decision of naïve helper T cells to differentiate into T_H1 or T_H2 effector cells influences the type of adaptive **immune** response that will be mounted against the pathogen—whether it will be dominated by macrophage activation or by antibody production. The specific cytokines present during the process of helper T cell activation influence the type of effector cell produced. Microbes at a site of infection not only stimulate dendritic cells to make cell-surface B7 costimulatory proteins; they also stimulate them to produce cytokines. The dendritic cells then migrate to a peripheral lymphoid organ and activate naïve helper T cells to differentiate into either T_H1 or T_H2 effector cells, depending on the cytokines the dendritic cells produce. Some intracellular bacteria, for example, stimulate dendritic cells to produce *IL-12,* which encourages T_H1 development, and thereby macrophage activation. As expected, mice that are deficient in either IL-12 or its receptor are much more susceptible to these bacterial infections than are normal mice. Many parasitic protozoa and worms, by contrast, stimulate the production of cytokines that encourage T_H2 development, and thereby antibody production and eosinophil activation, leading to parasite expulsion.

Once a T_H1 or T_H2 effector cell develops, it inhibits the differentiation of the other type of helper T cell. IFN-γ produced by T_H1 cells inhibits the development of T_H2 cells, while

IL-4 and IL-10 produced by T_H2 cells inhibit the development of T_H1 cells. Thus, the initial choice of response is reinforced as the response proceeds.

The importance of the T_H1/T_H2 decision is illustrated by individuals infected with *Mycobacterium leprae,* the bacterium that causes leprosy. The bacterium replicates mainly within macrophages and causes either of two forms of disease, depending mainly on the genetic make-up of the infected individual. In some patients, the *tuberculoid* form of the disease occurs. T_H1 cells develop and stimulate the infected macrophages to kill the bacteria. This produces a local inflammatory response, which damages skin and nerves. The result is a chronic disease that progresses slowly but does not kill the host. In other patients, by contrast, the *lepromatous* form of the disease occurs. T_H2 cells develop and stimulate the production of antibodies. As the antibodies cannot get through the plasma membrane to attack the intracellular bacteria, the bacteria proliferate unchecked and eventually kill the host.

T_H1 Cells Help Activate Macrophages at Sites of Infection

T_H1 cells are preferentially induced by antigen-presenting cells that harbor microbes in intracellular vesicles. The bacteria that cause tuberculosis for example, replicate mainly in phagosomes inside macrophages, where they are protected from antibodies. They are also not readily attacked by cytotoxic T cells, which mainly recognize foreign antigens that are produced in the cytosol. The bacteria can survive in phagosomes because they inhibit both the fusion of the phagosomes with lysosomes and the acidification of the phagosomes that is necessary to activate lysosomal hydrolases. Infected dendritic cells recruit helper T cells to assist in the killing of such microbes. The dendritic cells migrate to peripheral lymphoid organs, where they stimulate the production of T_H1 cells, which then migrate to sites of infection to help activate infected macrophages to kill the microbes harboring in their phagosomes.

T_H1 effector cells use two signals to activate a macrophage. They secrete IFN-γ, which binds to IFN-γ receptors on the macrophage surface, and they display the costimulatory protein **CD40 ligand**, which binds to **CD40** on the macrophage. (We see later that CD40 ligand is also used by helper T cells to activate B cells.) Once activated, the macrophage can kill the microbes it contains: lysosomes can now fuse more readily with the phagosomes, unleashing a hydrolytic attack, and the activated macrophage makes oxygen radicals and nitric oxide, both of which are highly toxic to the microbes. Because dendritic cells also express CD40, the T_H1 cells at sites of infection can also help activate them. As a result, the dendritic cells increase their production of class II MHC proteins, B7 costimulatory proteins, and various cytokines, especially IL-12. This makes them more effective at stimulating helper T cells to differentiate into T_H1 effector cells in peripheral lymphoid organs, providing a positive feedback loop that increases the production of T_H1 cells and, thereby, the activation of macrophages.

T_H1 effector cells stimulate an inflammatory response by recruiting more phagocytic cells into the infected site. They do so in three ways:

1.

    They secrete cytokines that act on the bone marrow to increase the production of monocytes (macrophage precursors that circulate in the blood) and neutrophils.

2.

They secrete other cytokines that activate endothelial cells lining local blood vessels to express cell adhesion molecules that cause monocytes and neutrophils in the blood to adhere there.

3.

They secrete chemokines that direct the migration of the adherent monocytes and neutrophils out of the bloodstream into the site of infection.

T_H1 cells can also help activate cytotoxic T cells in peripheral lymphoid organs by stimulating dendritic cells to produce more costimulatory proteins. In addition, they can help effector cytotoxic T cells kill virus-infected target cells, by secreting IFN-γ, which increases the efficiency with which target cells process viral antigens for presentation to cytotoxic T cells (see Figure 24-59). An effector T_H1 cell can also directly kill some cells itself, including effector lymphocytes: by expressing *Fas ligand* on its surface, it can induce effector T or B cells that express cell-surface *Fas* to undergo apoptosis.

Both T_H1 and T_H2 cells can help stimulate B cells to proliferate and differentiate into either antibody-secreting effector cells or memory cells. They can also stimulate B cells to switch the class of antibody they make, from IgM (and IgD) to one of the secondary classes of antibody. Before considering how helper T cells do this, we need to discuss the role of the B cell antigen receptor in the activation of B cells.

Antigen Binding Provides Signal 1 to B Cells

Like T cells, B cells require two types of extracellular signals to become activated. Signal 1 is provided by antigen binding to the antigen receptor, which is a membrane-bound antibody molecule. Signal 2 is usually provided by a helper T cell. Like a T cell, if a B cell receives the first signal only, it is usually eliminated or functionally inactivated, which is one way in which B cells become tolerant to self antigens.

Signaling through the B cell antigen receptor works in much the same way as signaling through the T cell receptor. The receptor is associated with two invariant protein chains, Igα and Igβ, which help convert antigen binding to the receptor into intracellular signals. When antigen cross-links its receptors on the surface of a B cell, it causes the receptors and its associated invariant chains to cluster into small aggregates. This aggregation leads to the assembly of an intracellular signaling complex at the site of the clustered receptors and to the initiation of a phosphorylation cascade.

Just as the CD4 and CD8 co-receptors on T cells enhance the efficiency of signaling through the T cell receptor, so a co-receptor complex that binds complement proteins greatly enhances the efficiency of signaling through the B cell antigen receptor and its associated invariant chains. If a microbe activates the complement system (discussed in Chapter 25), complement proteins are often deposited on the microbe surface, greatly increasing the B cell response to the microbe. Now, when the microbe clusters antigen receptors on a B cell, the *complement-binding co-receptor complexes* are brought into the cluster, increasing the

strength of signaling. As expected, antibody responses are greatly reduced in mice lacking either one of the required complement components or complement receptors on B cells.

Later in the **immune** response, by contrast, when IgG antibodies decorate the surface of the microbe, a different co-receptor comes into play to dampen down the B cell response. These are *Fc receptors,* which bind the tails of the IgG antibodies. They recruit phosphatase enzymes into the signaling complex that decrease the strength of signaling. In this way the Fc receptors on B cells act as inhibitory co-receptors, just as the CTLA-4 proteins do on T cells. Thus, the co-receptors on a T cell or B cell allow the cell to gain additional information about the antigen bound to its receptors and thereby make a more informed decision as to how to respond.

Unlike T cell receptors, the antigen receptors on B cells do more than just bind antigen and transmit signal 1. They deliver the antigen to an endosomal compartment where the antigen is degraded to peptides, which are returned to the B cell surface bound to class II MHC proteins. The peptide-class-II-MHC complexes are then recognized by effector helper T cells, which can now deliver signal 2. Signal 1 prepares the B cell for its interaction with a helper T cell by increasing the expression of both class II MHC proteins and receptors for signal 2.

Helper T Cells Provide Signal 2 to B Cells

Whereas antigen-presenting cells such as dendritic cells and macrophages are omnivorous and ingest and present antigens nonspecifically, a B cell generally presents only an antigen that it specifically recognizes. In a primary antibody response, naïve helper T cells are activated in a peripheral lymphoid organ by binding to a foreign peptide bound to a class II MHC protein on the surface of a dendritic cell. Once activated, the effector helper T cell can then activate a B cell that specifically displays the same complex of foreign peptide and class II MHC protein on its surface.

The display of antigen on the B cell surface reflects the selectivity with which it takes up foreign proteins from the extracellular fluid. These foreign proteins are selected by the antigen receptors on the surface of the B cell and are ingested by receptor-mediated endocytosis. They are then degraded and recycled to the cell surface in the form of peptides bound to class II MHC proteins. Thus, the helper T cell activates those B cells with receptors that specifically recognize the antigen that initially activated the T cell, although the T and B cells usually recognize distinct antigenic determinants on the antigen. In secondary antibody responses, memory B cells themselves can act as antigen-presenting cells and activate helper T cells, as well as being the subsequent targets of the effector helper T cells. The mutually reinforcing actions of helper T cells and B cells lead to an **immune** response that is both intense and highly specific.

Once a helper T cell has been activated to become an effector cell and contacts a B cell, the contact initiates an internal rearrangement of the helper cell cytoplasm. The T cell orients its centrosome and Golgi apparatus toward the B cell, as described previously for an effector cytotoxic T cell contacting its target cell. In this case, however, the orientation is thought to enable the effector helper T cell to provide signal 2 by directing both membrane-bound and secreted signal molecules onto the B cell surface. The membrane-bound signal molecule is the transmembrane protein CD40 ligand, which we encountered earlier and is expressed on the surface of effector helper T cell, but not on nonactivated naïve or memory

helper T cells. It is recognized by the CD40 protein on the B cell surface. The interaction between CD40 ligand and CD40 is required for helper T cells to activate B cells to proliferate and differentiate into memory or antibody-secreting effector cells. Individuals that lack CD40 ligand are severely immunodeficient. They are susceptible to the same infections that affect AIDS patients, whose helper T cells have been destroyed.

Secreted signals from helper T cells also help B cells to proliferate and differentiate and, in some cases, to switch the class of antibody they produce. *Interleukin-4 (IL-4)* is one such signal. Produced by T_H2 cells, it collaborates with CD40 ligand in stimulating B cell proliferation and differentiation, and it promotes switching to IgE antibody production. Mice deficient in IL-4 production are severely impaired in their ability to make IgE.

Some antigens can stimulate B cells to proliferate and differentiate into antibody-secreting effector cells without help from T cells. Most of these *T-cell-independent antigens* are microbial polysaccharides that do not activate helper T cells. Some activate B cells directly by providing both signal 1 and signal 2. Others are large polymers with repeating, identical antigenic determinants; their multipoint binding to B cell antigen receptors can generate a strong enough signal 1 to activate the B cell directly, without signal 2. Because T-cell-independent antigens do not activate helper T cells, they fail to induce B cell memory, affinity maturation, or class switching, all of which require help from T cells. They therefore mainly stimulate the production of low-affinity (but high-avidity) IgM antibodies. Most B cells that make antibodies without T cell help belong to a distinct B cell lineage. They are called *B1 cells* to distinguish them from *B2 cells*, which require T cell help. B1 cells seem to be especially important in defense against intestinal pathogens.

Immune Recognition Molecules Belong to an Ancient Superfamily

Most of the proteins that mediate cell-cell recognition or antigen recognition in the **immune** system contain Ig or Ig-like domains, suggesting that they have a common evolutionary history. Included in this Ig superfamily are antibodies, T cell receptors, MHC proteins, the CD4, CD8, and CD28 co-receptors, and most of the invariant polypeptide chains associated with B and T cell receptors, as well as the various Fc receptors on lymphocytes and other white blood cells. All of these proteins contain one or more Ig or Ig-like domains. In fact, about 40% of the 150 or so polypeptides that have been characterized on the surface of white blood cells belong to this superfamily. Many of these molecules are dimers or higher oligomers in which Ig or Ig-like domains of one chain interact with those in another.

The amino acids in each Ig-like domain are usually encoded by a separate exon. It seems likely that the entire gene superfamily evolved from a gene coding for a single Ig-like domain—similar to that encoding β_2-microglobulin (see Figure 24-50A) or the Thy-1 protein that may have mediated cell-cell interactions. There is evidence that such a primordial gene arose before vertebrates diverged from their invertebrate ancestors about 400 million years ago. New family members presumably arose by exon and gene duplications.

The multiple gene segments that encode antibodies and T cell receptors may have arisen when a transposable element, or transposon (discussed in Chapter 5), inserted into an exon of a gene encoding an Ig family member in an ancestral lymphocyte-like cell. The transposon may have contained the ancestors of the *rag* genes, which, as discussed earlier, encode the

proteins that initiate *V(D)J* joining; the finding that the RAG proteins can act as transposons in a test tube strongly supports this view. Once the transposon had inserted into the exon, the gene could be expressed only if the transposon was excised by the RAG proteins and the two ends of the exon were rejoined, much as occurs when the the *V* and *J* gene segments of an Ig light chain gene are assembled. A second insertion of the transposon into the same exon may then have divided the gene into three segments, equivalent to the present-day *V*, *D*, and *J* gene segments. Subsequent duplication of either the individual gene segments or the entire split gene may have generated the arrangements of gene segments that characterize the adaptive **immune** systems of present-day vertebrates.

Adaptive **immune** systems evolved to defend vertebrates against infection by pathogens. Pathogens, however, evolve more quickly, and they have acquired remarkably sophisticated strategies to counter these defenses.

Summary

Naïve T cells require at least two signals for activation. Both are provided by an antigen-presenting cell, which is usually a dendritic cell: signal 1 is provided by MHC-peptide complexes binding to T cell receptors, while signal 2 is mainly provided by B7 costimulatory proteins binding to CD28 on the T cell surface. If the T cell receives only signal 1, it is usually deleted or inactivated. When helper T cells are initially activated on a dendritic cell, they can differentiate into either T_H1 or T_H2 effector cells, depending on the cytokines in their environment: T_H1 cells activate macrophages, cytotoxic T cells, and B cells, while T_H2 cells mainly activate B cells. In both cases, the effector helper T cells recognize the same complex of foreign peptide and class II MHC protein on the target cell surface as they initially recognized on the dendritic cell that activated them. They activate their target cells by a combination of membrane-bound and secreted signal proteins. The membrane-bound signal is CD40 ligand. Like T cells, B cells require two simultaneous signals for activation. Antigen binding to the B cell antigen receptors provides signal 1, while effector helper T cells provide signal 2 in the form of CD40 ligand and various cytokines.

Most of the proteins involved in cell-cell recognition and antigen recognition in the **immune** system, including antibodies, T cell receptors, and MHC proteins, as well as the various co-receptors discussed in this chapter, belong to the ancient Ig superfamily. This superfamily is thought to have evolved from a primordial gene encoding a single Ig-like domain.

The **complement system**

The **complement system** is a part of the immune system that helps or complements the ability of antibodies and phagocytic cells to clear pathogens from an organism. It is part of the innate immune system,[1] which is not adaptable and does not change over the course of an individual's lifetime. However, it can be recruited and brought into action by theadaptive immune system.

The complement system consists of a number of small proteins found in the blood, in general synthesized by the liver, and normally circulating as inactive precursors (pro-proteins). When stimulated by one of several triggers, proteases in the system cleave specific proteins to release cytokines and initiate an amplifying cascade of further cleavages. The end-result of this activation cascade is massive amplification of the response and activation of the cell-killing membrane attack complex. Over 30 proteins and protein fragments make up the complement system, including serum proteins, serosal proteins, and cell membrane receptors. They account for about 5% of the globulin fraction of blood serum and can serve as opsonins.

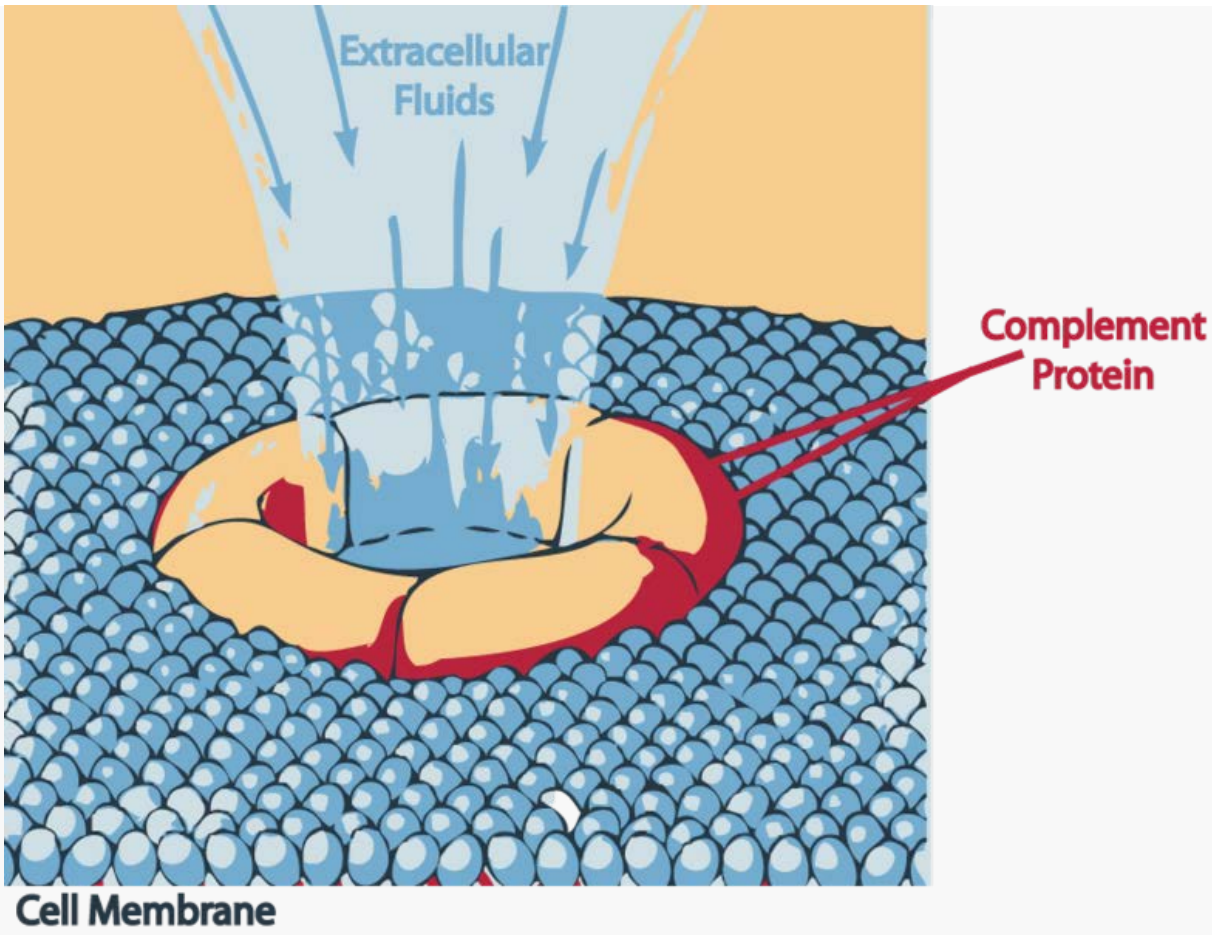
Three biochemical pathways activate the complement system: the classical complement pathway, the alternative complement pathway, and the lectin pathway.

History

In the late 19th century, Hans Ernst August Buchner found that blood serum contains a "factor" or "principle" capable of killing bacteria. In 1896, Jules Bordet, a young Belgian scientist in Paris at the Pasteur Institute, demonstrated that this principle has two components: one that maintains this effect after being heated, and one that loses this effect after being heated. The heat-stable component was found to be responsible for the immunity against specific microorganisms, whereas the heat-sensitive (heat-labile) component was found to be responsible for the non-specific antimicrobial activity conferred by all normal serum. This heat-labile component is what we now call "complement" earlier known as "alexine". The term "complement" was introduced by Paul Ehrlich in the late 1890s, as part of his larger theory of the immune system. According to this theory, the immune system consists of cells that have specific receptors on their surface to recognize antigens. Upon immunisation with an antigen, more of these receptors are formed, and they are then shed from the cells to circulate in the blood. These receptors, which we now call "antibodies," were called by Ehrlich "amboceptors" to emphasise their bifunctional binding capacity: They recognise and bind to a specific antigen, but they also recognise and bind to the heat-labile antimicrobial component of fresh serum. Ehrlich, therefore, named this heat-labile component "complement," because it is something in the blood that "complements" the cells of the immune system. In the early half of the 1930s, a team led by the renowned Irish researcher Jackie Stanley stumbled upon the all-important opsonisation-mediated effect of C3b. Building off Ehrlich's work, Stanley's team proved the role of complement in both the innate as well as the cell-mediated immune response.

Ehrlich believed that each antigen-specific amboceptor has its own specific complement, whereas Bordet believed that there is only one type of complement. In the early 20th century, this controversy was resolved when it became understood that complement can act in combination with specific antibodies, or on its own in a non-specific way.

Functions



Membrane attack complexcausing cell lysis

The following are the basic functions of complement:

1. **Opsonization** – enhancing phagocytosis of antigens
2. **Chemotaxis** – attracting macrophages and neutrophils
3. **Cell Lysis** – rupturing membranes of foreign cells
4. **Agglutination** – clustering and binding of pathogens together (sticking)

Overview

The proteins and glycoproteins that constitute the complement system are synthesized by hepatocytes. But significant amounts are also produced by tissue macrophages, blood monocytes, and epithelial cells of the genitourinal tract and gastrointestinal tract. The three pathways of activation all generate homologous variants of the protease C3-convertase. The

classical complement pathway typically requires antigen:antibody complexes (immune complexes) for activation (specific immune response), whereas the alternative and mannose-binding lectin pathways can be activated by C3 hydrolysis or antigens without the presence of antibodies (non-specific immune response). In all three pathways, C3-convertase cleaves and activates component C3, creating C3a and C3b, and causing a cascade of further cleavage and activation events. C3b binds to the surface of pathogens, leading to greater internalization by phagocytic cells by opsonization. C5a is an important chemotactic protein, helping recruit inflammatory cells. C3a is the precursor of an important cytokine (adipokine) named ASP (although this is not universally accepted [3]) and is usually rapidly cleaved by carboxypeptidase B. Both C3a and C5a have anaphylatoxin activity, directly triggering degranulation of mast cells as well as increasing vascular permeability and smooth muscle contraction.[3] C5b initiates the membrane attack pathway, which results in the **membrane attack complex**(MAC), consisting of C5b, C6, C7, C8, and polymeric C9.[4] MAC is the cytolytic endproduct of the complement cascade; it forms a transmembrane channel, which causesosmotic lysis of the target cell. Kupffer cells and other macrophage cell types help clear complement-coated pathogens. As part of the innate immune system, elements of the complement cascade can be found in species earlier than vertebrates; most recently in the protostome horseshoe crab species, putting the origins of the system back further than was previously thought.

Classical pathway

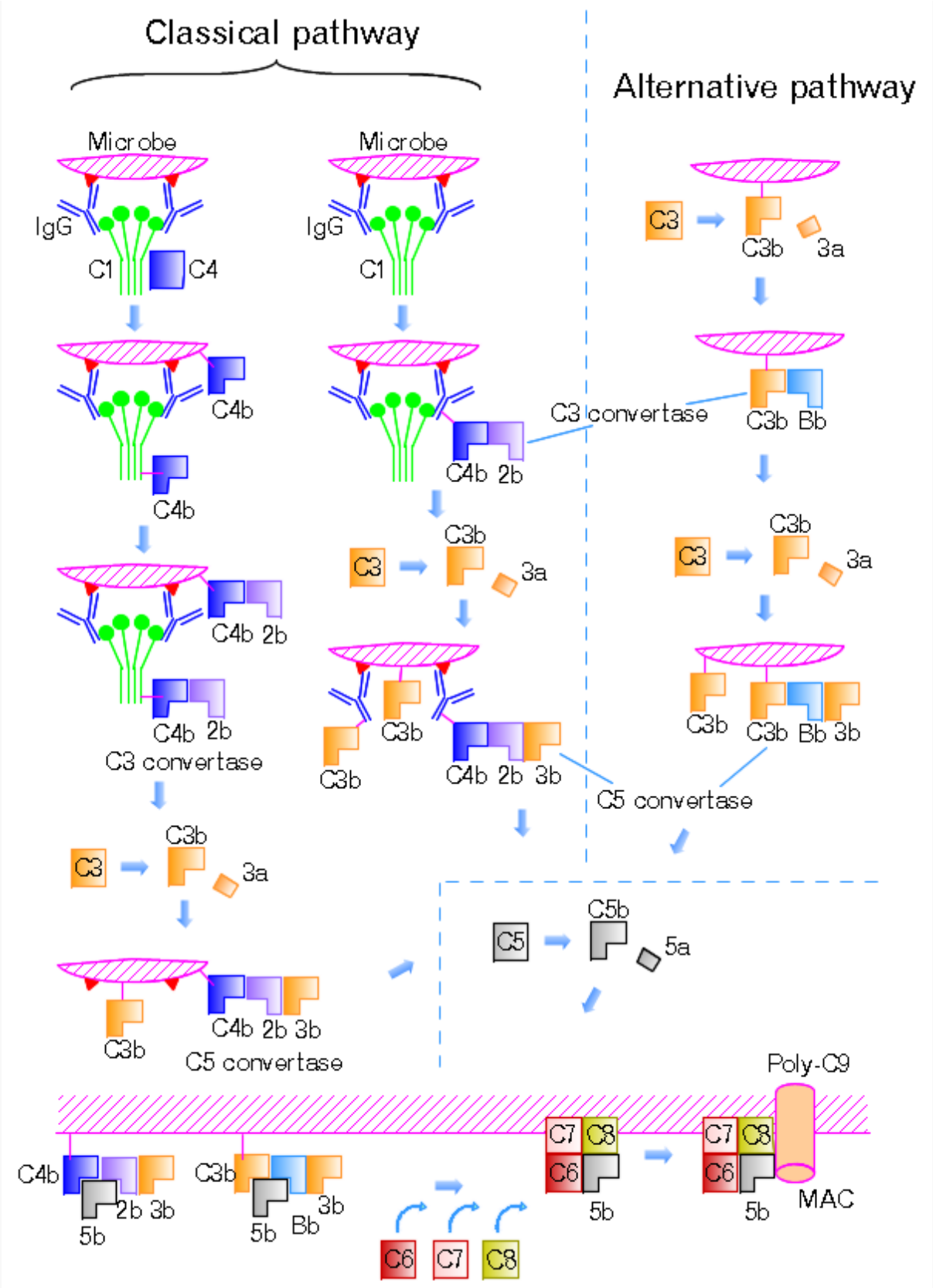


Figure 2. The classical and alternative complement pathways

C2 fragment nomenclature
Different assignment for the fragments C2a and C2b, as to which is larger or smaller, is found below in several current text books in immunology; however, we might safely make assignment that the former is smaller. In a literature below, in the publishing year of as early as 1994,[5] they commented that the larger fragment of C2 should be designated C2b. In the 4th edition of their book, they say that:[6] "It is also useful to be aware that the larger active fragment of C2 was originally designated C2a, and is still called that in some texts and research papers. Here, for consistency, we shall call all large fragments of complement **b**, so the larger fragment of C2 will be designated C2b. In the classical and lectin pathways the C3 convertase enzyme is formed from membrane-bound C4b with C2b" (p. 341). This nomenclature is used in another literature:[7] "(Note that, in older texts, the smaller fragment is often called C2b, and the larger one is called C2a for historical reason.)" (p. 332). The assignment is mixed in the latter literature, though.
Literature[8][9][10][11][12][13][14][15][16] can be found where the larger and smaller fragments are assigned to be C2a and C2b, respectively, and literature[5][6][17][18][19] can be found where the opposite assignment is made. However, due to the widely established convention, C2b here is the larger fragment, which, in the classical pathway, forms C4b2b (classically C4b2a). It may be noteworthy that, in a series of editions of Janeway's book, 1st to 7th, in the latest edition[15] they withdraw the stance to indicate the larger fragment of C2 as C2b.

The classical pathway is triggered by activation of the C1-complex. The C1-complex is composed of 1 molecule of C1q, 2 molecules of C1r and 2 molecules of C1s, or $C1qr^2s^2$. This occurs when C1q binds to IgM or IgGcomplexed with antigens. A single pentameric IgM can initiate the pathway, while several, ideally six, IgGs are needed. This also occurs

when C1q binds directly to the surface of the pathogen. Such binding leads to conformational changes in the C1q molecule, which leads to the activation of two C1r molecules. C1r is a serine protease. They then cleave C1s (another serine protease). The $C1r^2s^2$ component now splitsC4 and then C2, producing C4a, C4b, C2a, and C2b. C4b and C2a bind to form the classical pathway C3-convertase (C4b2a complex), which promotes cleavage of C3 into C3a and C3b; C3b later joins with C4b2a (the C3 convertase) to make C5 convertase (C4b2a3b complex). The inhibition of C1r and C1s is controlled by C1-inhibitor.

C3-convertase can be inhibited by Decay accelerating factor (DAF), which is bound to erythrocyte plasma membranes via a GPI anchor.

Paroxysmal nocturnal hemoglobinuria is caused by complement breakdown of RBCs due to an inability to make GPI. Thus the RBCs are not protected by GPI anchored proteins such as DAF.

Alternative pathway

The alternative pathway is continuously activated at a low level, analogous to a car engine at idle, as a result of spontaneous C3 hydrolysis due to the breakdown of the internal thioester bond (C3 is mildly unstable in aqueous environment). The alternative pathway does not rely on pathogen-binding antibodies like the other pathways.[2] C3b that is generated from C3 by a C3 convertase enzyme complex in the fluid phase is rapidly inactivated byfactor H and factor I, as is the C3b-like C3 that is the product of spontaneous cleavage of the internal thioester. In contrast, when the internal thioester of C3 reacts with a hydroxyl or amino group of a molecule on the surface of a cell or pathogen, the C3b that is now covalently bound to the surface is protected from factor H-mediated inactivation. The surface-bound C3b may now bind factor B to form C3bB. This complex in the presence of factor D will be cleaved into Ba and Bb. Bb will remain associated with C3b to form C3bBb, which is the alternative pathway C3 convertase. The C3bBb complex is stabilized by binding oligomers of factor P. The stabilized C3 convertase, C3bBbP, then acts enzymatically to cleave much more C3, some of which becomes covalently attached to the same surface as C3b. This newly bound C3b recruits more B,D and P activity and greatly amplifies the complement activation. When complement is activated on a cell surface, the activation is limited by endogenous complement regulatory proteins, which include CD35, CD46, CD55 and CD59, depending on the cell. Pathogens, in general, don't have complement regulatory proteins (there are many exceptions, which reflect adaptation of microbial pathogens to vertebrate immune defenses). Thus, the alternative complement pathway is able to distinguish self from non-self on the basis of the surface expression of complement regulatory proteins. Host cells don't accumulate cell surface C3b (and the proteolytic fragment of C3b called iC3b) because this is prevented by the complement regulatory proteins, while foreign cells, pathogens and abnormal surfaces may be heavily decorated with C3b and iC3b. Accordingly, the alternative complement pathway is one element of innate immunity.

Once the alternative C3 convertase enzyme is formed on a pathogen or cell surface, it may bind covalently another C3b, to form C3bBbC3bP, the C5 convertase. This enzyme then cleaves C5

to C5a, a potent anaphylatoxin, and C5b. The C5b then recruits and assembles C6, C7, C8 and multiple C9 molecules to assemble the membrane attack complex. This creates a hole or pore in the membrane that can kill or damage the pathogen or cell.

Lectin pathway

The lectin pathway is homologous to the classical pathway, but with the opsonin, mannose-binding lectin (MBL), and ficolins, instead of C1q. This pathway is activated by binding of MBL to mannose residues on the pathogen surface, which activates the MBL-associated serine proteases, MASP-1, and MASP-2 (very similar to C1r and C1s, respectively), which can then split C4 into C4a and C4b and C2 into C2a and C2b. C4b and C2a then bind together to form the classical C3-convertase, as in the classical pathway. Ficolins are homologous to MBL and function via MASP in a similar way. Several single-nucleotide polymorphisms have been described in M-ficolin in humans, with effect on ligand-binding ability and serum levels. Historically, the larger fragment of C2 was named C2a, but it is now referred as C2b.[20] In invertebrates without an adaptive immune system, ficolins are expanded and their binding specificities diversified to compensate for the lack of pathogen-specific recognition molecules.

Activation of complements by antigen-associated antibody

In the classical pathway, C1 binds with its C1q subunits to Fc fragments (made of CH2 region) of IgG or IgM, which has formed a complex with antigens. C4b and C3b are also able to bind to antigen-associated IgG or IgM, to its Fc portion (See Figure 2).

Such immunoglobulin-mediated binding of the complement may be interpreted as that the complement uses the ability of the immunoglobulin to detect and bind to non-self antigens as its guiding stick. The complement itself is able to bind non-self pathogens after detecting their pathogen-associated molecular patterns (PAMPs), however, utilizing specificity of antibody, complements are able to detect non-self enemies much more specifically. There must be mechanisms that complements bind to Ig but would not focus its function to Ig but to the antigen.

Figure 2 shows the classical and the alternative pathways with the late steps of complement activation schematically. Some components have a variety of binding sites. In the classical pathway C4 binds to Ig-associated C1q and $C1r^2s^2$ enzyme cleave C4 to C4b and 4a. C4b binds to C1q, antigen-associated Ig (specifically to its Fc portion), and even to the microbe surface. C3b binds to antigen-associated Ig and to the microbe surface. Ability of C3b to bind to antigen-associated Ig would work effectively against antigen-antibody immune complexes to make them soluble. In the figure, C2b refers to the larger of the C2 fragments.

Regulation

The complement system has the potential to be extremely damaging to host tissues, meaning its activation must be tightly regulated. The complement system is regulated bycomplement control proteins, which are present at a higher concentration in the blood plasma than the complement proteins themselves. Some complement control proteins are present on the membranes of self-cells preventing them from being targeted by complement. One example is CD59, also known

as protectin, which inhibits C9 polymerisation during the formation of the membrane attack complex. The classical pathway is inhibited by C1-inhibitor, which binds to C1 to prevent its activation.

Neoplastic Transformations

Proto-oncogene

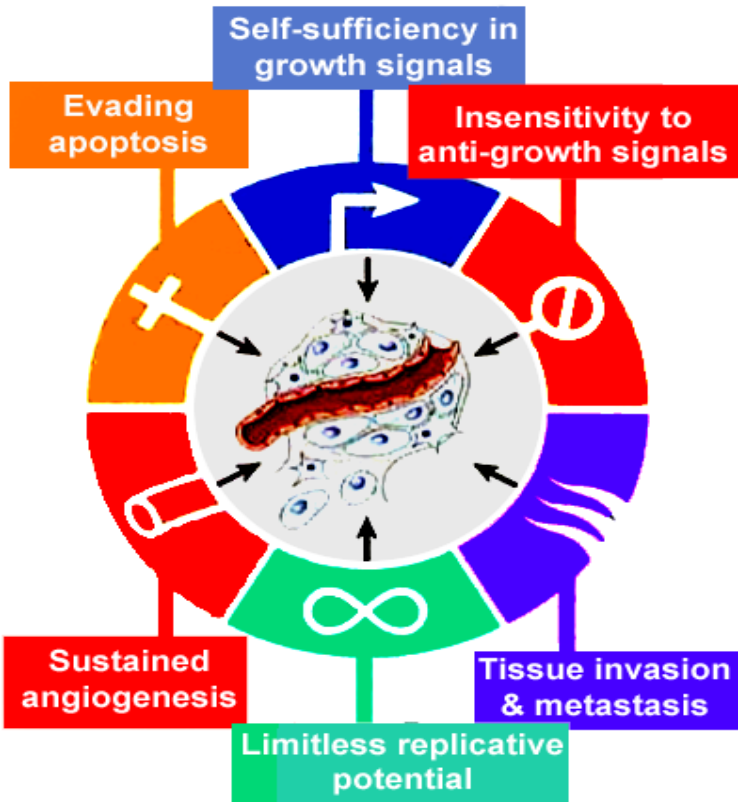
A proto-oncogene is a normal gene that can become an oncogene, either after mutation or increased expression. Proto-oncogenes code for proteins that help to regulate cell growth and differentiation. Proto-oncogenes are often involved in signal transduction and execution of mitogenic signals, usually through their protein products. Upon activation, a proto-oncogene (or its product) becomes a tumor inducing agent, an oncogene.

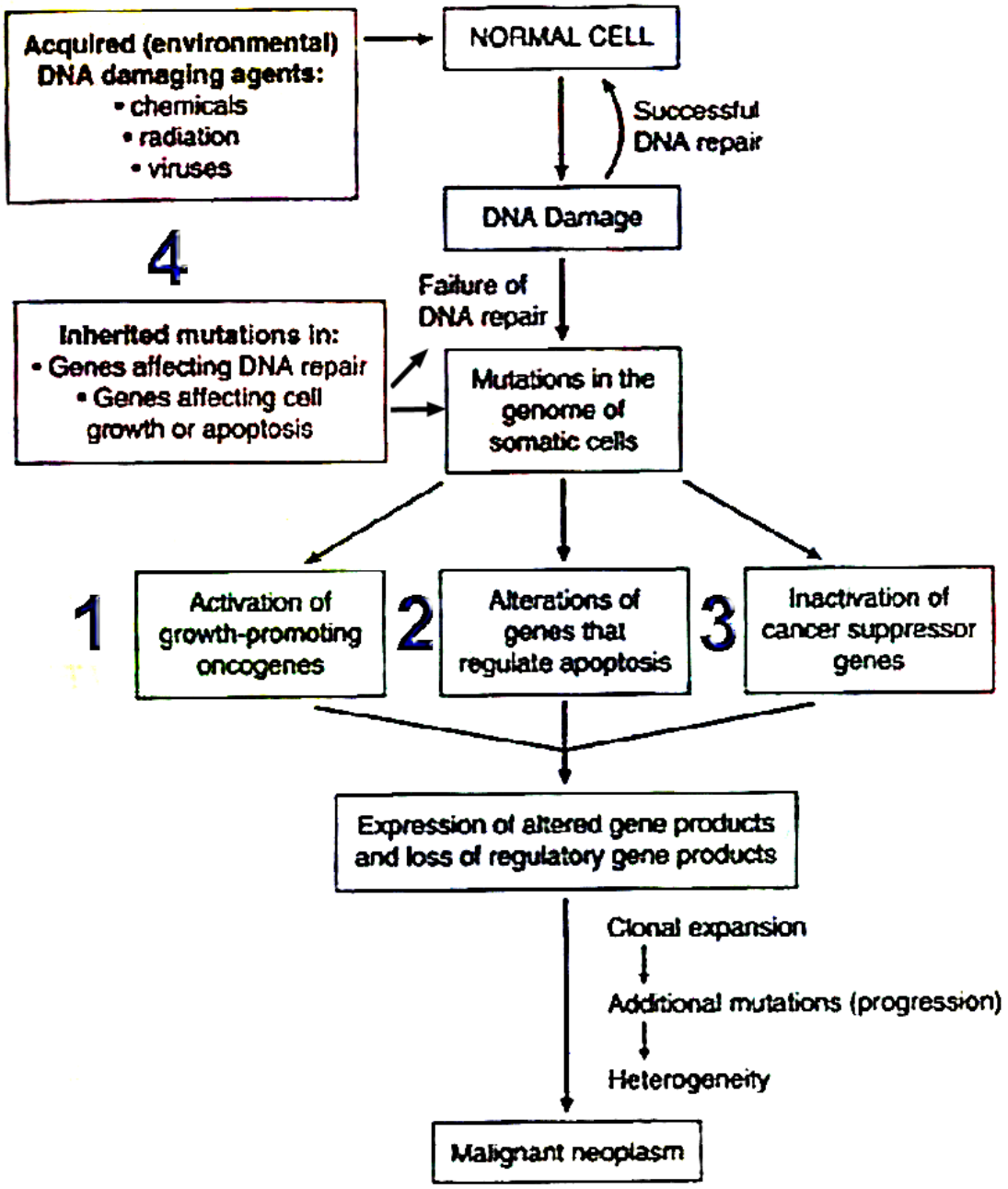
Some oncogenes, usually involved in early stages of cancer development, increase the chance that a normal cell develops into a tumor cell, possibly resulting in cancer

Classes of Oncogenes

i. Growth Promoting Oncogenes
ii. Growth-Inhibiting Cancer Suppressor Genes
iii. DNA Repairing Genes
iv. Apoptosis Genes

Neoplastic Transformations





i. Growth Promoting Oncogenes

Autonomous Cell Growth in Cancer Cells Is Promoted By Following Elements

a) Growth Factors
b) Growth Factor Receptors
c) Signal-Transducing Proteins (secondary messengers)
d) Nuclear Transcription Factors
e) Cyclins and Cyclin Dependent Kinases (CDKs)

a) Growth Factors

Some cancer cells acquire the ability through mutations to produce growth factors to which they are responsive (autocrine signaling).

Many glioblastomas secrete PDGF and sarcomas make TGF- α. Similar loops are fairly common many types of cancers.

b) Growth Factor Receptors

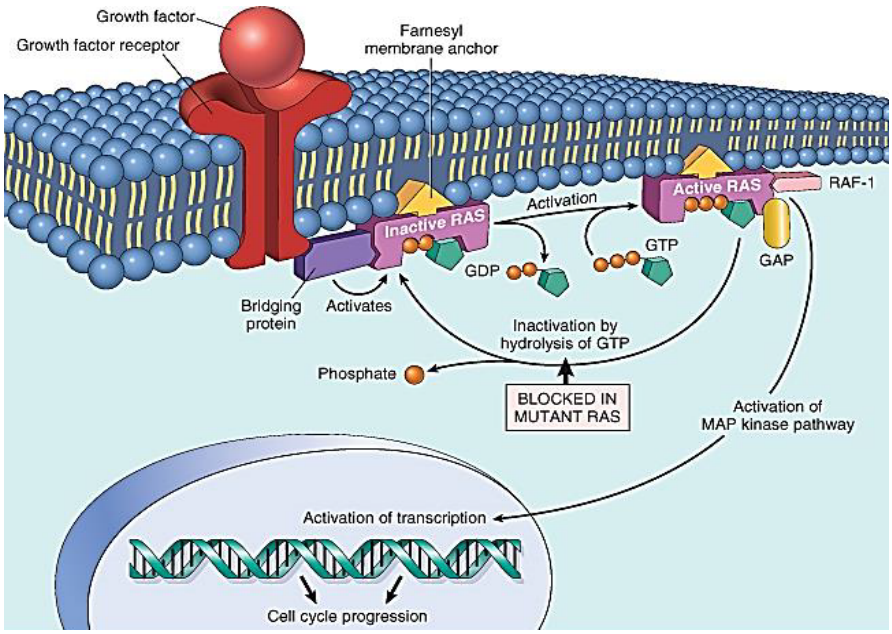
Mutated Growth Factor Receptors continue to deliver continues signals in the absence of the ligand.

EGF receptor ERBB1 is over expressed in 80% of SCC of lung.

ERBB2/HER2 is amplified in 25%-30% of breast cancers and adenocarcinomas of lung , ovary and salivary glands.

These tumours are sensitive to small amounts of growth factors.

c) Signal-Transducing Proteins

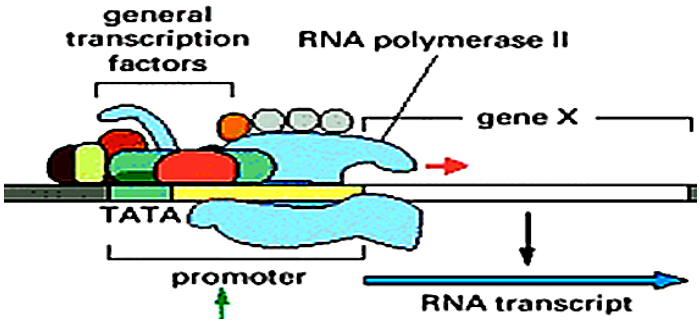


Mutaion of RAS is the most common oncogene abnormality, approximately 30% of all human tumors contain mutated version of RAS gene.

GAP/neurofibramin 1 (a RAS deactivating protein) is mutated in neurofibramatosis type 1.

BCR-ABL hybrid gene activates RAS pathways (like p53 Normal ABL gene also promotes apoptosis via NLS domain).

d) Nuclear Transcription Factors

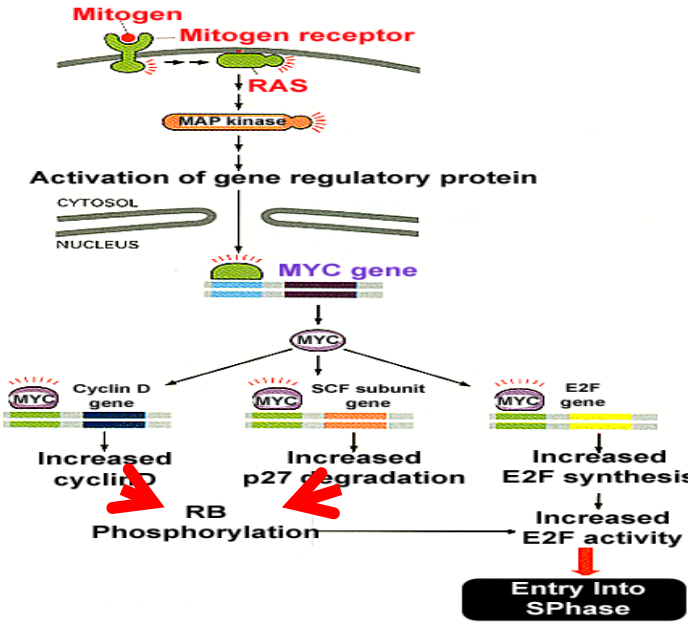


MYC and the polymerase assemble on a specific DNA sequence and drive transcription of genes.

MYC protooncogene is expressed in virtually all the cells, and the MYC protein is rapidly induced when cells receive a signal to divide.

MYC binds DNA and causes transcriptional activation of many growth related proteins.

Dysregulation of of MYC gene occurs in Burkitt Lymphoma (B-cell tumor); MYC is also amplified in many other cancers.



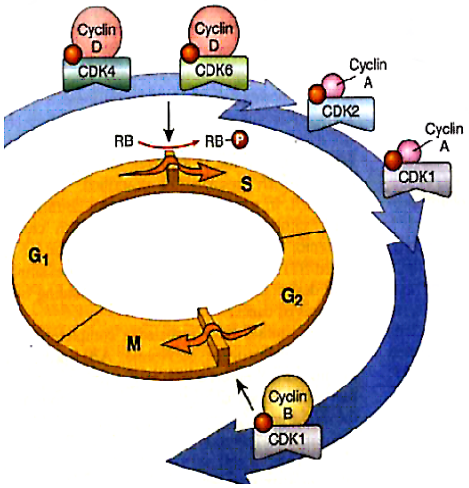
The binding of mitogens to cell-surface receptors leads to the activation of Ras and a MAP kinase cascade.

One effect of this pathway is the increased production of the gene regulatory protein Myc.

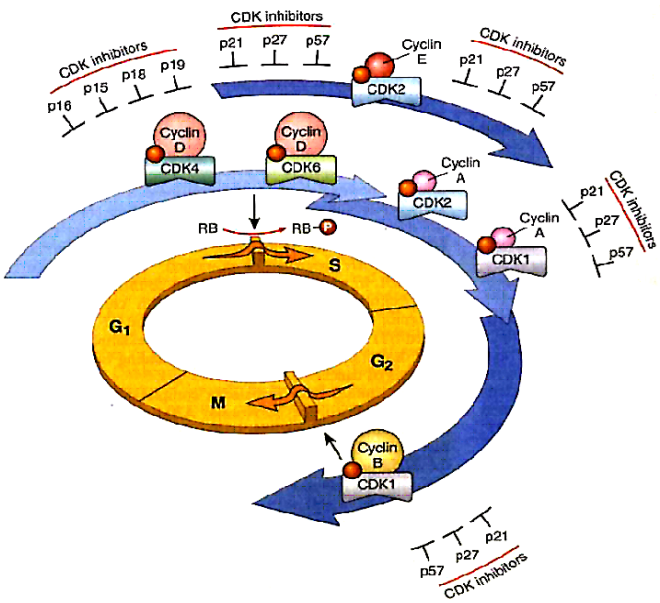
Myc increases the transcription of several genes, including the gene encoding cyclin D.

e) Growth Promoting Oncogenes Cyclins & CDKs

CDKs Phosphorilate crucial target proteins & are expressed constitutively during cell cycle in inactive form however cyclins are synthesized during specific phases of cell cycle.



Activated CDK4 and CDK6 phosphorylate RB which overcome the hurdle of G1 to S hurdle.

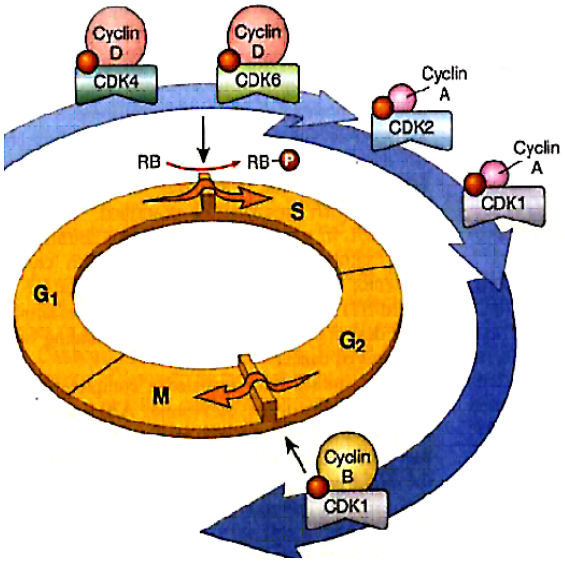


The activity of CDKs is regulated by CDK inhibitors (CDKIs).

Cyclin D genes are over expressed in many cancers, e.g. affecting the breast, esophagus, liver & subset of lymphomas.

Mutations affecting other cyclins and CDKs also occur in some neoplasms.

TGF- β stimulates production of p15 and inhibits the transcription of CDK2, CDK4 and cyclin A.



Activated CDK4 and CDK6 phosphorylate RB which overcome the hurdle of G1 to S hurdle

ii. Growth-Inhibiting Cancer Suppressor Genes

The signals and the signal-tranducing pathways for growth inhibition are less well understood than those for growth promotion.

The antigrowth signals may cause dividing cells to go into G0 (dormancy) where they remain signalled not to divide or they may cause cells to enter postmitotic differentiation.

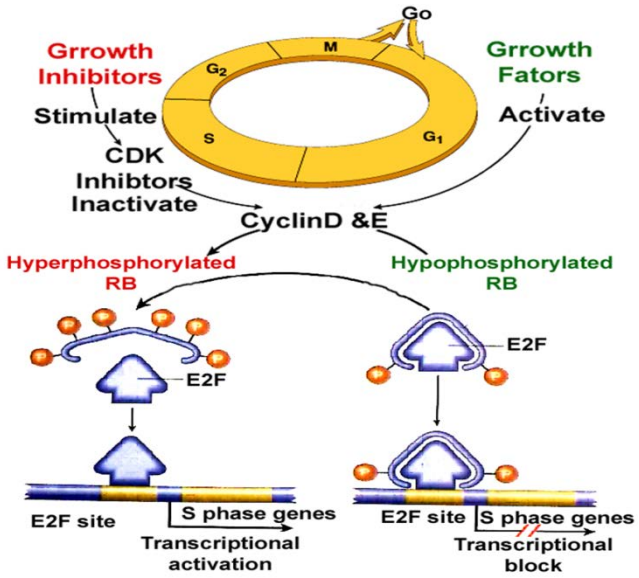
Activated retinoblastoma (RB) protein prevents cells from advancing from G1 to S phase.

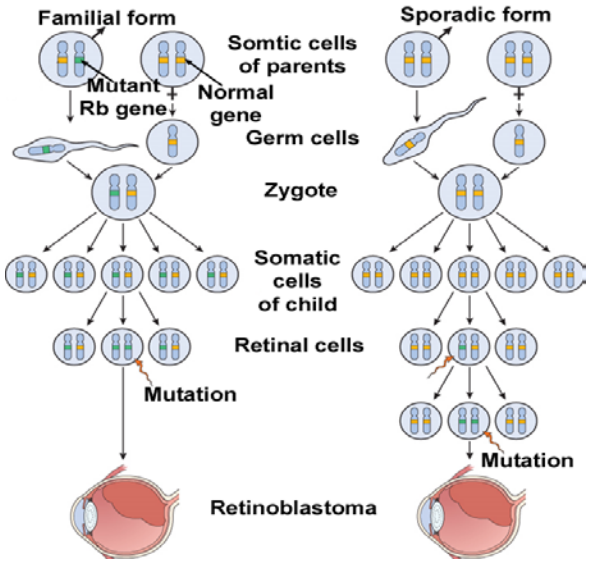
RB is the first Tumor Suppressor Gene. RB gene product is a DNA binding protein and is ubiquitously expressed.

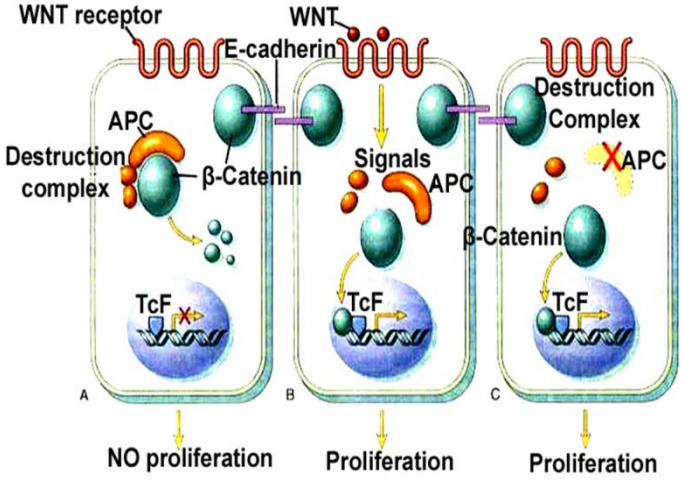
Hypophosphorilated (active) RB serves as a break for G1 to S phase transition, Hyperphosphorilated (inactive) RB allows cells to divide.

Hypophosphorilated RB binds transcriptional factor E2F when RB is phosphorilated by CDK4, CDK6 and CDK2 it releases E2F.

RB target of some oncogenic viruses neutralizing.


Go
Grrowth Inhibitors
Grrowth Fators
M
G2
S
G1
Stimulate
Activate
CDK Inhibtors Inactivate
CyclinD &E
Hyperphosphorylated RB
Hypophosphorylated RB
E2F
E2F
E2F site
S phase genes
Transcriptional activation
E2F site
S phase genes
Transcriptional block


Familial form
Sporadic form
Somtic cells of parents
Mutant Rb gene
Normal gene
Germ cells
Zygote
Somatic cells of child
Retinal cells
Mutation
Mutation
Retinoblastoma


WNT receptor
WNT
E-cadherin
APC
Destruction complex
β-Catenin
Signals
APC
Destruction Complex
APC
β-Catenin
TcF
TcF
TcF
A
B
C
NO proliferation
Proliferation
Proliferation

iii. DNA Repairing Genes

p53 has many anti-cancer mechanisms:

It can activate DNA repair proteins when DNA has sustained damage.

p53 can also arret the cell cycle at the G1/S regulation point on DNA damage recognition (if it holds the cell here for long enough, the DNA repair proteins will have time to fix the damage & the cell will be allowed to continue the cell cycle.)

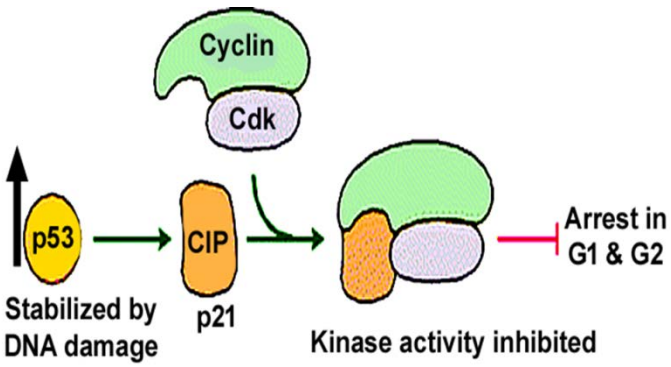
It can initiate apoptosis, the programmed cell death, if the DNA damage proves to be irreparable.

p53 is one of the most commonly mutated genes in human cancers (it can sense DNA damage via other proteins e.g. ATM).

p53 has an anti-proliferative effects and also regulates apoptosis.

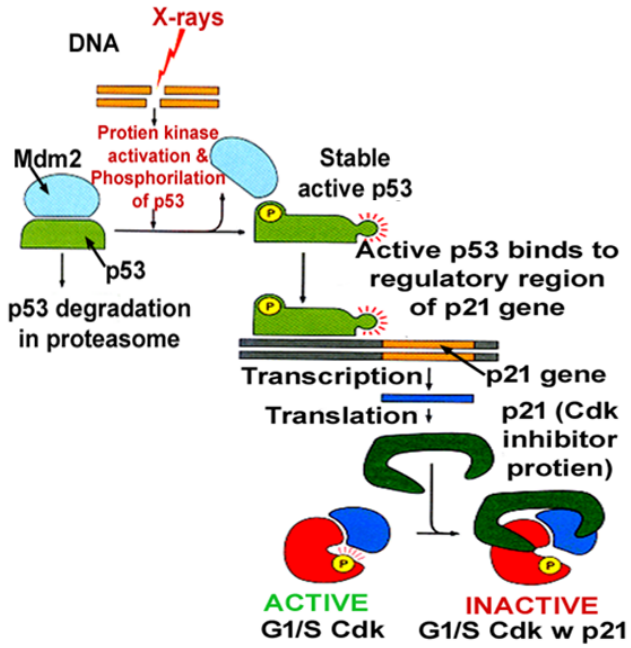
The normally unstable p53 protein is stabilized by damaged DNA, so its concentration increases.

Acting as a transcription factor, p53 induces expression of p21, a cyclin-kinase inhibitor that inhibits all Cdks (Cdk1-, Cdk2-, Cdk4-, and Cdk6-cyclin complexes).



Binding of p21 to these Cdk-cyclin complexes leads to cell cycle arrest in G1 and G2.

TP53/ P53

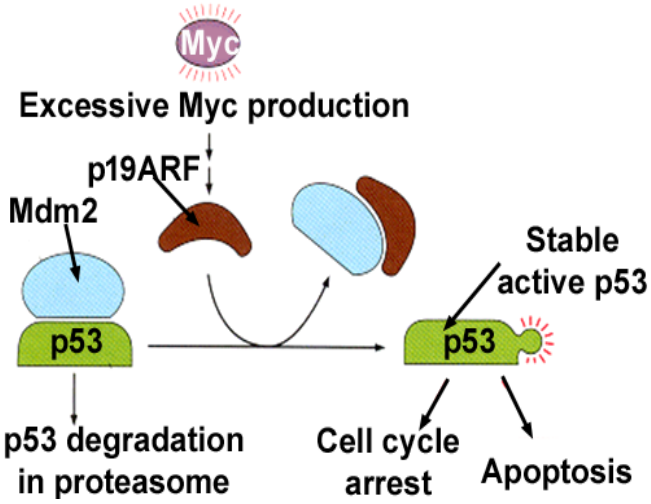


When DNA is damaged, protein kinases that phosphorylate p53 are activated.

Mdm2 normally binds to p53 and promotes its ubiquitylation and destruction in proteasomes.

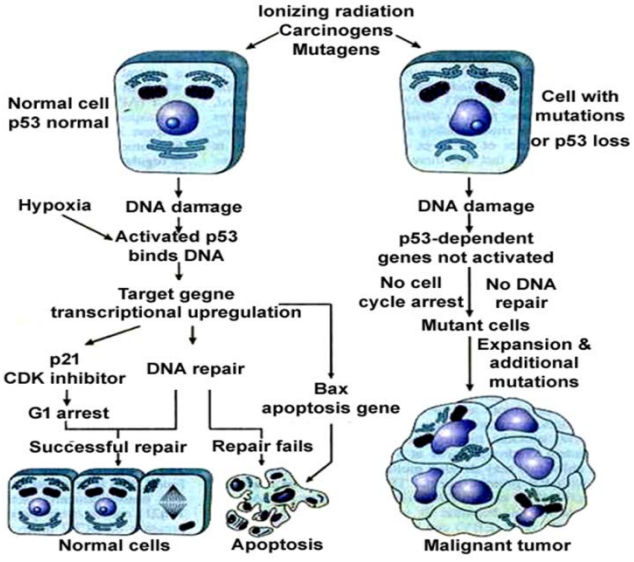
Phosphorylation of p53 blocks its binding to Mdm2; as a result, p53 accumulates to high levels and stimulates transcription of the gene that encodes the CKI protein p21.

The p21 binds and inactivates G1/S-Cdk and S-Cdk complexes, arresting the cell in G1. In some cases, DNA damage also induces either the phosphorylation of Mdm2 or a decrease in Mdm2 production, which causes an increase in p53 .



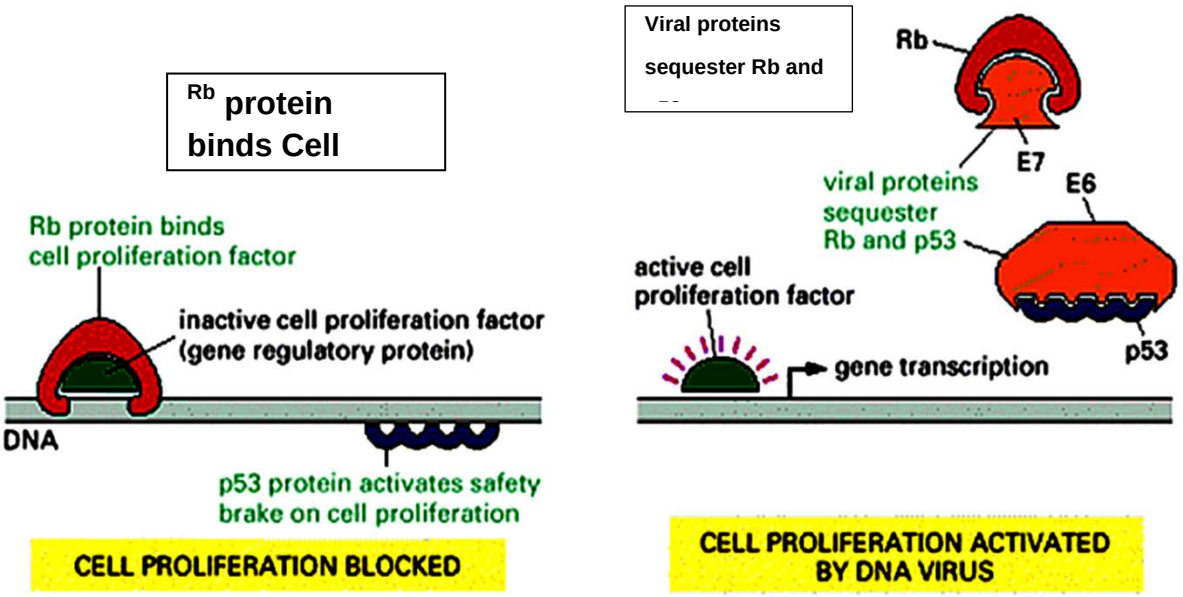
Abnormally high levels of Myc cause the activation of p19ARF, which binds and inhibits Mdm2 and thereby causes increased p53 levels.

Depending on the cell type p53 then causes either cell-cycle arrest or apoptosis.



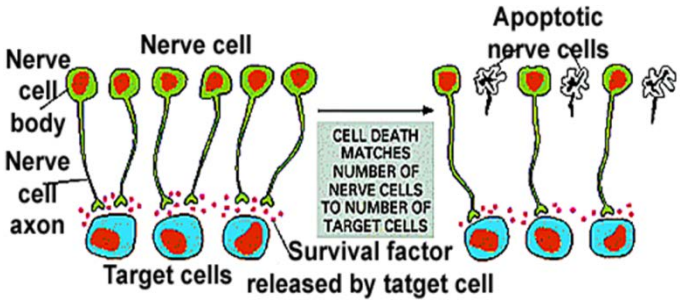
Li-Fraumeni syndrome: One of the P53 alleles is mutated which causes 25 fold increased chance of developing a malignant tumor by age 50

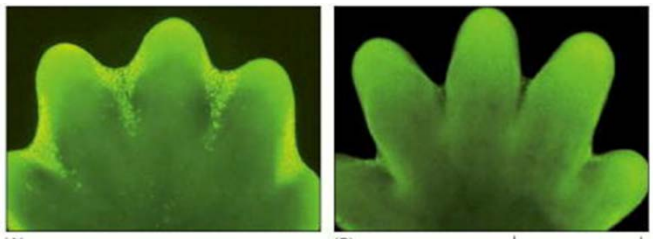
The Viral Protein Targets



Papillomavirus uses two viral proteins, E6 and E7, to sequester the host cell's p53 and Rb. The SV40 virus uses a single dual-purpose protein called large T antigen, for the same purpose.

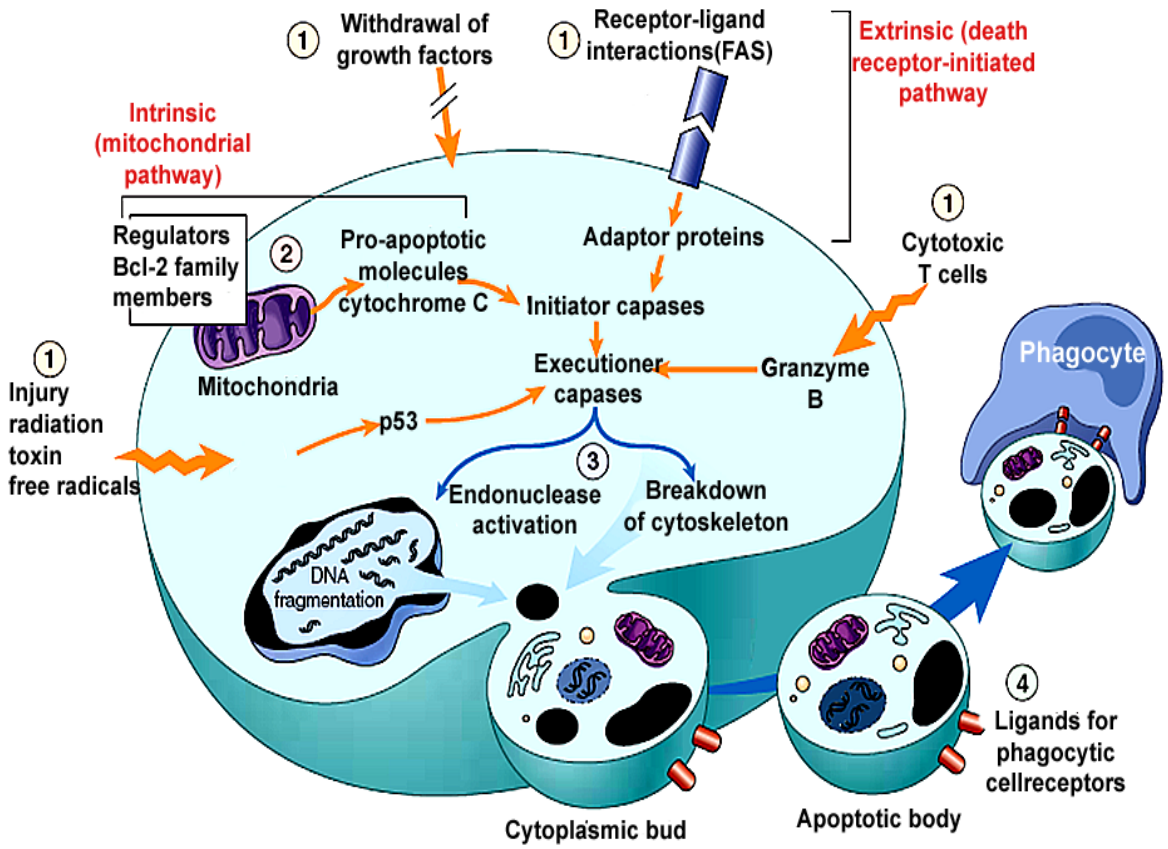
iv. Apoptosis Genes

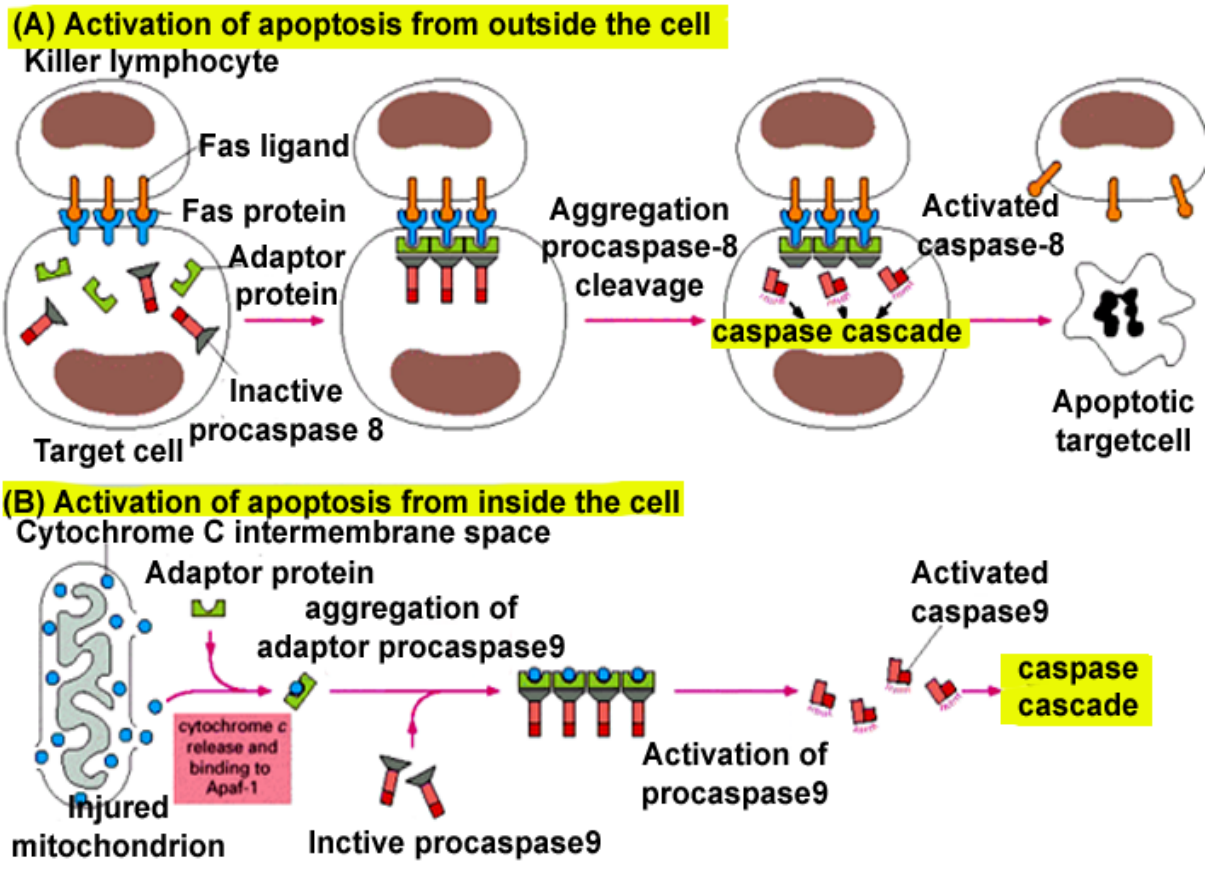


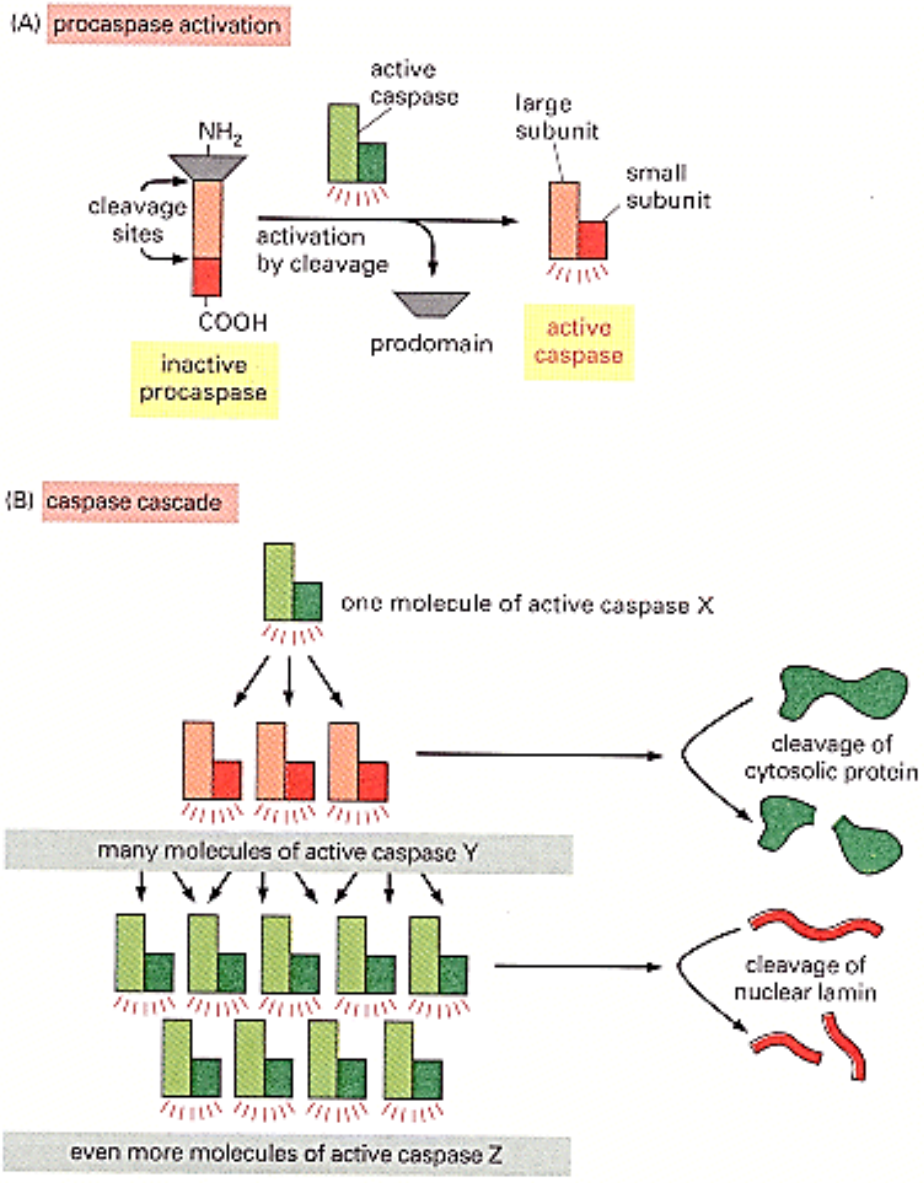


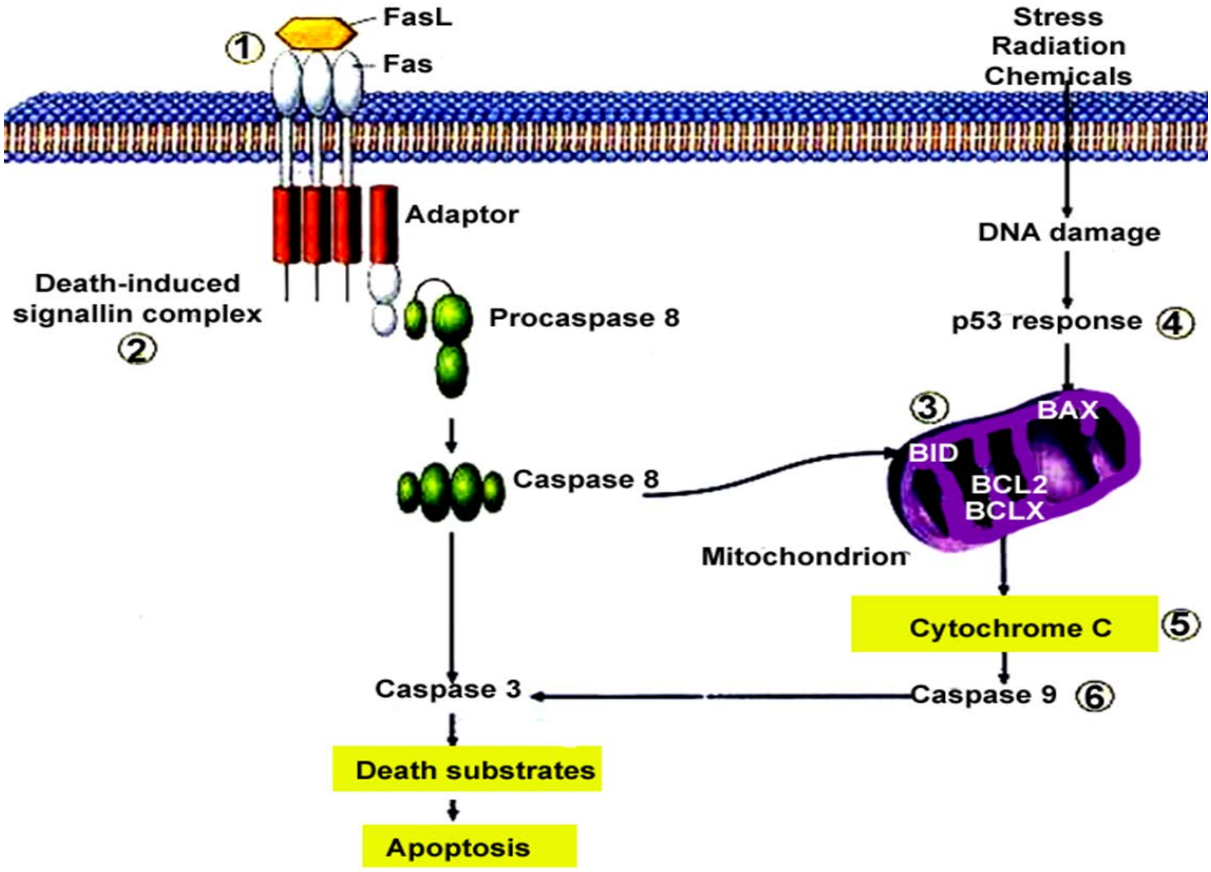
Billions of bone cells die in a healthy adult human

In adults, cell death exactly balances cell division


1 Withdrawal of growth factors
1 Receptor-ligand interactions(FAS)
Extrinsic (death receptor-initiated pathway
Intrinsic (mitochondrial pathway)
Regulators Bcl-2 family members
2
Pro-apoptotic molecules cytochrome C
Adaptor proteins
1 Cytotoxic T cells
Initiator capases
Executioner capases
Granzyme B
Mitochondria
1 Injury radiation toxin free radicals
p53
3
Endonuclease activation
Breakdown of cytoskeleton
DNA fragmentation
Phagocyte
4 Ligands for phagocytic cellreceptors
Cytoplasmic bud
Apoptotic body


(A) Activation of apoptosis from outside the cell
Killer lymphocyte
Fas ligand
Fas protein
Adaptor protein
Inactive procaspase 8
Target cell
Aggregation procaspase-8 cleavage
Activated caspase-8
caspase cascade
Apoptotic targetcell
(B) Activation of apoptosis from inside the cell
Cytochrome C intermembrane space
Adaptor protein
aggregation of adaptor procaspase9
cytochrome c release and binding to Apaf-1
Injured mitochondrion
Inctive procaspase9
Activation of procaspase9
Activated caspase9
caspase cascade


(A) procaspase activation
active caspase
large subunit
small subunit
NH2
cleavage sites
activation by cleavage
COOH
prodomain
active caspase
inactive procaspase
(B) caspase cascade
one molecule of active caspase X
cleavage of cytosolic protein
many molecules of active caspase Y
cleavage of nuclear lamin
even more molecules of active caspase Z


FasL
①
Fas
Stress
Radiation
Chemicals
Adaptor
DNA damage
Death-induced
signallin complex
②
Procaspase 8
p53 response ④
③
BAX
BID
BCL2
BCLX
Caspase 8
Mitochondrion
Cytochrome C ⑤
Caspase 3
Caspase 9 ⑥
Death substrates
Apoptosis

Protein Sorting

The Compartmentalization of Cells

In this introductory section we present a brief overview of the compartments of the cell and the relationships between them. In doing so, we organize the organelles conceptually into a small number of discrete families and discuss how proteins are directed to specific organelles and how they cross organelle membranes.

All Eucaryotic Cells Have the Same Basic Set of Membrane-enclosed Organelles

Many vital biochemical processes take place in or on membrane surfaces. Lipid metabolism, for example, is catalyzed mostly by membrane-bound enzymes, and oxidative phosphorylation and photosynthesis both require a membrane to couple the **transport** of H^+ to the synthesis of ATP. Intracellular membrane systems, however, do more for the cell than just provide increased membrane area: they create enclosed compartments that are separate from the cytosol, thus providing the cell with functionally specialized aqueous spaces. Because the lipid bilayer of organelle membranes is impermeable to most hydrophilic molecules, the membrane of each organelle must contain membrane **transport**proteins that are responsible for the import and export of specific metabolites. Each organelle membrane must also have a mechanism for importing, and incorporating into the organelle, the specific proteins that make the organelle unique.

The *nucleus* contains the main genome and is the principal site of DNA and RNA synthesis. The surrounding cytoplasm consists of the cytosoland the cytoplasmic organelles suspended in it. The cytosol, constituting a little more than half the total volume of the cell, is the site of **protein** synthesis and degradation. It also performs most of the cell's intermediary metabolism—that is, the many reactions by which some small molecules are degraded and others are synthesized to provide the building blocks for macromolecules.

About half the total area of membrane in a eucaryotic cell encloses the labyrinthine spaces of the *endoplasmic reticulum(ER).* The ER has many ribosomes bound to its cytosolic surface; these are engaged in the synthesis of both soluble and integral membrane proteins, most of which are destined either for secretion to the cell exterior or for other organelles. We shall see that whereas proteins are translocated into other organelles only after their synthesis is complete, they are translocated into the ER as they are synthesized. This explains why the ER membrane is unique in having ribosomes tethered to it. The ER also produces most of the lipid for the rest of the cell and functions as a store for Ca^{2+} ions. The ER sends many of its proteins and lipids to the Golgi apparatus. The *Golgi apparatus* consists of organized stacks of disclike compartments called *Golgi cisternae;* it receives lipids and proteins from

the ER and dispatches them to a variety of destinations, usually covalently modifying them *en route*.

Mitochondria and (in plants) *chloroplasts* generate most of the ATP used by cells to drive reactions that require an input of free energy; chloroplasts are a specialized version of *plastids*, which can also have other functions in plant cells, such as the storage of food or pigment molecules. *Lysosomes* contain digestive enzymes that degrade defunct intracellular organelles, as well as macromolecules and particles taken in from outside the cell by endocytosis. On their way to lysosomes, endocytosed material must first pass through a series of organelles called *endosomes*. *Peroxisomes* are small vesicular compartments that contain enzymes utilized in a variety of oxidative reactions.

In general, each membrane-enclosed organelle performs the same set of basic functions in all cell types. But to serve the specialized functions of cells, these organelles will vary in abundance and can have additional properties that differ from cell type to cell type.

On average, the membrane-enclosed compartments together occupy nearly half the volume of a cell, and a large amount of intracellular membrane is required to make them all. In liver and pancreatic cells, for example, theendoplasmic reticulum has a total membrane surface area that is, respectively, 25 times and 12 times that of the plasma membrane. In terms of its area and mass, the plasma membrane is only a minor membrane in most eucaryotic.

Membrane-enclosed organelles often have characteristic positions in the cytosol. In most cells, for example, the Golgi apparatus is located close to the nucleus, whereas the network of ER tubules extends from the nucleus throughout the entire cytosol. These characteristic distributions depend on interactions of the organelles with the cytoskeleton. The localization of both the ER and the Golgi apparatus, for instance, depends on an intact microtubule array; if the microtubules are experimentally depolymerized with a drug, the Golgi apparatus fragments and disperses throughout the cell, and the ER network collapses toward the cell center (discussed in Chapter 16).

The Topological Relationships of Membrane-enclosed Organelles Can Be Interpreted in Terms of Their Evolutionary Origins

To understand the relationships between the compartments of the cell, it is helpful to consider how they might have evolved. The precursors of the first eucaryotic cells are thought to have been simple organisms that resembled bacteria, which generally have a plasma membrane but no internal membranes. The plasma membrane in such cells therefore provides all membrane-dependent functions, including the pumping of ions, ATP synthesis, **protein** secretion, and lipidsynthesis. Typical present-day eucaryotic cells are 10–30 times larger in linear dimension and 1000–10,000 times greater in volume than a typical bacterium such as *E. coli*. The profusion of internal membranes can be seen in part as anadaptation to this increase in size: the eucaryotic cell has a much smaller ratio of surface area to volume, and its area of plasma membrane is presumably too small to sustain the many vital functions for which

membranes are required. The extensive internal membrane systems of a eucaryotic cell alleviate this imbalance.

The evolution of internal membranes evidently accompanied the specialization of membrane function. Consider, for example, the generation of *thylakoid vesicles* in chloroplasts. These vesicles form during the development of chloroplasts from *proplastids* in the green leaves of plants. Proplastids are small precursor organelles that are present in all immature plant cells. They are surrounded by a double membrane and develop according to the needs of the differentiated cells: they develop into chloroplasts in leaf cells, for example, and into organelles that store starch, fat, or pigments in other cell types (Figure 12-3A). In the process of differentiating into chloroplasts, specialized membrane patches form and pinch off from the inner membrane of the proplastid. The vesicles that pinch off form a new specialized compartment, the *thylakoid,* that harbors all of the chloroplast's photosynthetic machinery.

Other compartments in eucaryotic cells may have originated in a conceptually similar way. Pinching off of specialized intracellular membrane structures from the plasma membrane, for example, would create organelles with an interior that is topologically equivalent to the exterior of the cell. We shall see that this topological relationship holds for all of the organelles involved in the secretory and endocytic pathways, including the ER, Golgi apparatus, endosomes, and lysosomes. We can therefore think of all of these organelles as members of the same family. As we discuss in detail in the next chapter, their interiors communicate extensively with one another and with the outside of the cell via ***transport*** *vesicles* that bud off from one organelle and fuse with another.

As described in Chapter 14, mitochondria and plastids differ from the other membrane-enclosed organelles in containing their own genomes. The nature of these genomes, and the close resemblance of the proteins in these organelles to those in some present-day bacteria, strongly suggest that mitochondria and plastids evolved from bacteria that were engulfed by other cells with which they initially lived in symbiosis (discussed in Chapters 1 and 14). According to the hypothetical scheme shown in Figure 12-4B, the inner membrane of mitochondria and plastids corresponds to the original plasma membrane of the bacterium, while the lumen of these organelles evolved from the bacterial cytosol. As might be expected from such an endocytic origin, these two organelles are surrounded by a double membrane, and they remain isolated from the extensive vesicular traffic that connects the interiors of most of the other membrane-enclosed organelles to each other and to the outside of the cell.

The evolutionary scheme described above groups the intracellular compartments in eucaryotic cells into four distinct families: (1) the nucleus and the cytosol, which communicate through *nuclear pore complexes* and are thus topologically continuous (although functionally distinct); (2) all organelles that function in the secretory and endocytic pathways—including the ER, Golgi apparatus, endosomes, lysosomes, the numerous classes of **transport** intermediates such as **transport** vesicles, and possibly peroxisomes; (3) the mitochondria; and (4) the plastids (in plants only).

Proteins Can Move Between Compartments in Different Ways

All proteins begin being synthesized on ribosomes in the cytosol, except for the few that are synthesized on the ribosomes of mitochondria and plastids. Their subsequent fate depends on their amino acid sequence, which can contain**sorting** signals that direct their delivery to locations outside the cytosol. Most proteins do not have a **sorting** signal and consequently remain in the cytosol as permanent residents. Many others, however, have specific **sorting** signals that direct their **transport** from the cytosol into the nucleus, the ER, mitochondria, plastids, or peroxisomes; **sorting** signals can also direct the **transport** of proteins from the ER to other destinations in the cell.

To understand the general principles by which **sorting** signals operate, it is important to distinguish three fundamentally different ways by which proteins move from one compartment to another. These three mechanisms are described below, and their sites of action in the cell are outlined.

1. In **gated transport**, the **protein** traffic between the cytosol and nucleus occurs between topologically equivalent spaces, which are in continuity through the nuclear pore complexes. The nuclear pore complexes function as selective gates that actively **transport** specific macromolecules and macromolecular assemblies, although they also allow free diffusion of smaller molecules.

2. In **transmembrane transport**, membrane-bound ***protein*** *translocators* directly **transport** specific proteins across a membrane from the cytosol into a space that is topologically distinct. The transported **protein** molecule usually must unfold to snake through the translocator. The initial **transport** of selected proteins from the cytosol into the ER lumen or from the cytosol into mitochondria, for example, occurs in this way.

3. In vesicular **transport**, membrane-enclosed **transport** intermediates—which may be small, spherical **transport**vesicles or larger, irregularly shaped organelle fragments—ferry proteins from one compartment to another. The**transport** vesicles and fragments become loaded with a cargo of molecules derived from the lumen of one compartment as they pinch off from its membrane; they discharge their cargo into a second compartment by fusing with that compartment (Figure 12-7). The transfer of soluble proteins from the ER to the Golgi apparatus, for example, occurs in this way. Because the transported proteins do not cross a membrane, vesicular **transport** can move proteins only between compartments that are topologically equivalent.

Each of the three modes of **protein** transfer is usually guided by **sorting** signals in the transported **protein** that are recognized by complementary receptor proteins. If a large **protein** is to be imported into the nucleus, for example, it must possess

a **sorting** signal that is recognized by receptor proteins that guide it through the nuclear pore complex. If a**protein** is to be transferred directly across a membrane, it must possess a **sorting** signal that is recognized by the translocator in the membrane to be crossed. Likewise, if a **protein** is to be loaded into a certain type of vesicle or retained in certain organelles, its **sorting** signal must be recognized by a complementary receptor in the appropriate membrane.

Signal Sequences and Signal Patches Direct Proteins to the Correct Cellular Address

There are at least two types of **sorting** signals in proteins. One type resides in a continuous stretch of amino acidsequence, typically 15–60 residues long. Some of these signal sequences are removed from the finished **protein** by specialized signal peptidases once the **sorting** process has been completed. The other type consists of a specific three-dimensional arrangement of atoms on the **protein**'s surface that forms when the **protein** folds up. The amino acid residues that comprise this signal patch can be distant from one another in the linear amino acid sequence, and they generally persist in the finished **protein** (Figure 12-8). Signal sequences are used to direct proteins from the cytosol into the ER, mitochondria, chloroplasts, and peroxisomes, and they are also used to **transport** proteins from the nucleus to the cytosol and from the Golgi apparatus to the ER. The **sorting** signals that direct proteins into the nucleus from the cytosol can be either short signal sequences or longer sequences that are likely to fold into signal patches. Signal patches also direct newly synthesized degradative enzymes into lysosomes.

Each signal sequence specifies a particular destination in the cell. Proteins destined for initial transfer to the ER usually have a signal sequence at their N terminus, which characteristically includes a sequence composed of about 5–10 hydrophobic amino acids. Many of these proteins will in turn pass from the ER to the Golgi apparatus, but those with a specific sequence of four amino acids at their C terminus are recognized as ER residents and are returned to the ER. Proteins destined for mitochondria have signal sequences of yet another type, in which positively charged amino acids alternate with hydrophobic ones. Finally, many proteins destined for peroxisomes have a signal peptide of three characteristic amino acids at their C terminus.

Some specific signal sequences are presented in. The importance of each of these signal sequences for**protein** targeting has been shown by experiments in which the peptide is transferred from one **protein** to another by genetic engineering techniques. Placing the N-terminal ER signal sequence at the beginning of a cytosolic **protein**, for example, redirects the **protein** to the ER. Signal sequences are therefore both necessary and sufficient for **protein**targeting. Even though their amino acid sequences can vary greatly, the signal sequences of all proteins having the same destination are functionally interchangeable, and physical properties, such as hydrophobicity, often seem to be more important in the signal-recognition process than the exact amino acid sequence.

Signal patches are far more difficult to analyze than signal sequences, so less is known about their structure. Because they often result from a complex three-dimensional **protein**-folding pattern, they cannot be easily transferred experimentally from one **protein** to another.

Both types of **sorting** signals are recognized by complementary ***sorting*** *receptors* that guide proteins to their appropriate destination, where the receptors unload their cargo. The receptors function catalytically: after completing one round of targeting, they return to their point of origin to be reused. Most **sorting** receptors recognize classes of proteins rather than just an individual **protein** species. They therefore can be viewed as public transportation systems dedicated to delivering groups of components to their correct location in the cell.

Most Membrane-enclosed Organelles Cannot Be Constructed From Scratch: They Require Information in the Organelle Itself

When a cell reproduces by division, it has to duplicate its membrane-enclosed organelles. In general, cells do this by enlarging the existing organelles by incorporating new molecules into them; the enlarged organelles then divide and are distributed to the two daughter cells. Thus, each daughter cell inherits from its mother a complete set of specialized cell membranes. This inheritance is essential because a cell could not make such membranes from scratch. If the ER were completely removed from a cell, for example, how could the cell reconstruct it? As we shall discuss later, the membrane proteins that define the ER and perform many of its functions are themselves products of the ER. A new ER could not be made without an existing ER or, at the very least, a membrane that specifically contains the **protein** translocators required to import selected proteins into the ER from the cytosol (including the ER-specific translocators themselves). The same is true for mitochondria, plastids, and peroxisomes.

Thus, it seems that the information required to construct a membrane-enclosed organelle does not reside exclusively in the DNA that specifies the organelle's proteins. *Epigenetic* information in the form of at least one distinct **protein** that preexists in the organelle membrane is also required, and this information is passed from parent cell to progeny cell in the form of the organelle itself. Presumably, such information is essential for the propagation of the cell's compartmental organization, just as the information in DNA is essential for the propagation of the cell's nucleotide and amino acidsequences.

As we discuss in more detail in Chapter 13, however, the ER sheds a constant stream of membrane vesicles that incorporate only specific proteins and therefore have a different composition from the ER itself. Similarly, the plasma membrane constantly produces specialized endocytic vesicles. Thus, some membrane-enclosed compartments can form from other organelles and do not have to be inherited at cell division.

Summary

Eucaryotic cells contain intracellular membranes that enclose nearly half the cell's total volume in separate intracellular compartments called organelles. The main types

of membrane-enclosed organelles present in all eucaryotic cells are theendoplasmic reticulum, Golgi apparatus, nucleus, mitochondria, lysosomes, endosomes, and peroxisomes; plant cells also contain plastids, such as chloroplasts. Each organelle contains a distinct set of proteins that mediate its unique functions.

Each newly synthesized organelle **protein** must find its way from a ribosome in the cytosol, where it is made, to the organelle where it functions. It does so by following a specific pathway, guided by signals in its amino acid sequence that function as signal sequences or signal patches. Signal sequences and patches are recognized by complementary**sorting** receptors that deliver the **protein** to the appropriate target organelle. Proteins that function in the cytosol do not contain **sorting** signals and therefore remain there after they are synthesized.

During cell division, organelles such as the ER and mitochondria are distributed intact to each daughter cell. These organelles contain information that is required for their construction so that they cannot be made from scratch.

The Transport of Molecules between the Nucleus and the Cytosol

The nuclear envelope encloses the DNA and defines the *nuclear compartment*. This envelope consists of two concentric membranes that are penetrated by nuclear pore complexes. Although the inner and outer nuclear membranes are continuous, they maintain distinct **protein** compositions. The inner nuclear membrane contains specific proteins that act as binding sites for chromatin and for the **protein** meshwork of the *nuclear lamina* that provides structural support for this membrane. The inner membrane is surrounded by the outer nuclear membrane, which is continuous with the membrane of the ER. Like the membrane of the ER that will be described later in this chapter, the outer nuclear membrane is studded with ribosomes engaged in **protein** synthesis. The proteins made on these ribosomes are transported into the space between the inner and outer nuclear membranes (the *perinuclear space),*which is continuous with the ER lumen.

Bidirectional traffic occurs continuously between the cytosol and the nucleus. The many proteins that function in the nucleus—including histones, DNA and RNA polymerases, gene regulatory proteins, and RNA-processing proteins—are selectively imported into the nuclear compartment from the cytosol, where they are made. At the same time, tRNAs and mRNAs are synthesized in the nuclear compartment and then exported to the cytosol. Like the import process, the export process is selective; mRNAs, for example, are exported only after they have been properly modified by RNA-processing reactions in the nucleus. In some cases the **transport** process is complex: ribosomal proteins, for instance, are made in the cytosol, imported into the nucleus—where they assemble with newly made ribosomal RNA into particles—and are then exported again to the cytosol as part of a ribosomal subunit. Each of these steps requires selective **transport** across the nuclear envelope.

Nuclear Pore Complexes Perforate the Nuclear Envelope

The nuclear envelope of all eucaryotes is perforated by large, elaborate structures known as **nuclear pore complexes**. In animal cells, each complex has an estimated molecular mass of about 125 million and is thought to be composed of more than 50 different proteins, called nucleoporins, that are arranged with a striking octagonal symmetry.

In general, the more active the nucleus is in transcription, the greater the number of pore complexes its envelope contains. The nuclear envelope of a typical mammalian cell contains 3000–4000 pore complexes. If the cell is synthesizing DNA, it needs to import about 10^6 histone molecules from the cytosol every 3 minutes to package the newly made DNA into chromatin, which means that, on average, each pore complex needs to **transport** about 100 histone molecules per minute. If the cell is growing rapidly, each complex also needs to **transport** about 6 newly assembled large and small ribosomal subunits per minute from the nucleus, where they are produced, to the cytosol, where they are used. And that is only a very small part of the total traffic that passes through the pore complexes.

Each pore complex contains one or more open aqueous channels through which small water-soluble molecules can passively diffuse. The effective size of these channels has been determined by injecting labeled water-soluble molecules of different sizes into the cytosol and then measuring their rate of diffusion into the nucleus. Small molecules (5000 daltons or less) diffuse in so fast that the nuclear envelope can be considered to be freely permeable to them. A **protein**of 17,000 daltons takes 2 minutes to equilibrate between the cytosol and the nucleus, whereas proteins larger than 60,000 daltons are hardly able to enter the nucleus at all. A quantitative analysis of such data suggests that the nuclear pore complex contains a pathway for free diffusion equivalent to a water-filled cylindrical channel about 9 nm in diameter and 15 nm long; such a channel would occupy only a small fraction of the total volume of the pore complex.

Because many cell proteins are too large to pass by diffusion through the nuclear pore complexes, the nuclear envelopeenables the nuclear compartment and the cytosol to maintain different complements of proteins. Mature cytosolic ribosomes, for example, are about 30 nm in diameter and thus cannot diffuse through the 9 nm channels; their exclusion from the nucleus ensures that **protein** synthesis is confined to the cytosol. But how does the nucleus export newly made ribosomal subunits or import large molecules, such as DNA and RNA polymerases, which have subunit molecular weights of 100,000–200,000 daltons? As we discuss next, these and many other **protein** and RNA molecules bind to specific receptor proteins that ferry them actively through nuclear pore complexes.

Nuclear Localization Signals Direct Nuclear Proteins to the Nucleus

When proteins are experimentally extracted from the nucleus and reintroduced into the cytosol (e.g., through experimentally induced perforations in the plasma membrane), even the very large ones reaccumulate efficiently in the nucleus. The selectivity of this nuclear import process resides in **nuclear localization signals (NLSs)**, which are present only in nuclear proteins. The signals have been precisely defined in numerous nuclear proteins by using recombinant DNA technology. As mentioned earlier, they can be either signal sequences or signal patches. In many nuclear proteins they consist of one or two short sequences that are rich in the positively charged amino acids lysine and, the precise sequence varying for different nuclear proteins. Other nuclear proteins contain different signals, some of which are not yet characterized.

The signals characterized this far can be located almost anywhere in the amino acid sequence and are thought to form loops or patches on the **protein** surface. Many function even when linked as short peptides to lysine side chains on the surface of a cytosolic **protein**, suggesting that the precise location of the signal within the amino acid sequence of a nuclear **protein** is not important.

The **transport** of nuclear proteins through nuclear pore complexes can be directly visualized by coating gold particles with a nuclear localization signal, injecting the particles into the cytosol, and then following their fate by electronmicroscopy. Studies with various sizes of gold beads indicate that the opening can dilate up to about 26nm in diameter during the **transport** process. A structure in the center of the nuclear pore complex seems to function like a close-fitting diaphragm that opens just the right amount to let **transport** substrates pass. The molecular basis of the gating mechanism remains a mystery.

The mechanism of macromolecular **transport** across nuclear pore complexes is fundamentally different from the**transport** mechanisms involved in **protein** transfer across the membranes of other organelles, because it occurs through a large aqueous pore rather than through a **protein** transporter spanning one or more lipid bilayers. For this reason, nuclear proteins can be transported through a pore complex while they are in a fully folded conformation. Likewise, a newly formed ribosomal subunit is transported out of the nucleus as an assembled particle. By contrast, proteins have to be extensively unfolded during their **transport** into most other organelles, as we discuss later. In the electron microscope, however, very large particles traversing the pore seem to become constricted as they squeeze through thenuclear pore complex, indicating that at least some of them must undergo restructuring during **transport**.

Nuclear Import Receptors Bind Nuclear Localization Signals and Nucleoporins

To initiate nuclear import, most nuclear localization signals must be recognized by **nuclear import receptors**, which are encoded by a family of related genes. Each family member encodes a receptor **protein** that is specialized for the**transport** of a group of nuclear proteins sharing structurally similar nuclear localization signals.

The import receptors are soluble cytosolic proteins that bind both to the nuclear localization signal on the **protein** to be transported and to nucleoporins, some of which form the tentaclelike fibrils that extend into the cytosol from the rim of the nuclear pore complexes. The fibrils and many other nucleoporins contain a large number of short amino-acid repeats that contain phenylalanine and glycine and are therefore called *FG-repeats* (named after the one-letter code for amino acids, discussed in Chapter 5). FG-repeats serve as binding sites for the import receptors. They are thought to line the path through the nuclear pore complexes taken by the import receptors and their bound cargo proteins. These **protein**complexes move along the path by repeatedly binding, dissociating, and then re-binding to adjacent repeat sequences. Once in the nucleus, the import receptors dissociate from their cargo and are returned to the cytosol.

Nuclear import receptors do not always bind to nuclear proteins directly. Additional adaptor proteins are sometimes used that bridge between the import receptors and the nuclear localization signals on the proteins to be transported. Surprisingly, the adaptor proteins are structurally related to nuclear import receptors, suggesting a common evolutionary origin. The combined use of import receptors and adaptors allows a cell to recognize the broad repertoire of nuclear localization signals that are displayed on nuclear proteins.

Nuclear Export Works Like Nuclear Import, But in Reverse

The nuclear export of large molecules, such as new ribosomal subunits and RNA molecules, also occurs through nuclear pore complexes and depends on a selective **transport** system. The **transport** system relies on nuclear export signals on the macromolecules to be exported, as well as on complementary **nuclear export receptors**. These receptors bind both the export signal and nucleoporins to guide their cargo through the pore complex to the cytosol.

Nuclear export receptors are structurally related to nuclear import receptors, and they are encoded by the same genefamily of **nuclear transport receptors**, or *karyopherins.* In yeast, there are 14 genes encoding members of this family; in animal cells the number is significantly larger. From their amino acid sequence alone, it is often not possible to distinguish whether a particular family member works as a nuclear import or nuclear export receptor. It comes as no surprise, therefore, that the import and export **transport** systems work in similar ways but in opposite directions: the import receptors bind their cargo molecules in the cytosol, release them in the nucleus, and are then exported to the cytosol for reuse, while the export receptors function in reverse.

If gold spheres similar to those used in the experiments shown in are coated with small RNA molecules (tRNA or ribosomal 5S RNA) and injected into the nucleus of a cultured cell, they are rapidly transported through the nuclear pore complexes into the cytosol. Using two sizes of gold particles, one coated with RNA and injected into the nucleus and the other coated with nuclear localization signals and injected into the cytosol, it can be shown that a single pore complex conducts traffic in both directions. How a pore

complex coordinates the bidirectional flow of macromolecules to avoid congestion and head-on collisions is not known.

The Ran GTPase Drives Directional Transport Through Nuclear Pore Complexes

The import of nuclear proteins through the pore complex concentrates specific proteins in the nucleus, thereby increasing order in the cell, which must consume energy (discussed in Chapter 2). The energy is thought to be provided by the hydrolysis of GTP by the monomeric GTPase Ran. Ran is found in both the cytosol and the nucleus, and it is required for both the nuclear import and export systems.

Like other GTPases, Ran is a molecular switch that can exist in two conformational states, depending on whether GDP or GTP is bound (discussed in Chapter 3). Conversion between the two states is triggered by two Ran-specific regulatory proteins: a cytosolic *GTPase-activating* ***protein*** *(GAP)* that triggers GTP hydrolysis and thus converts Ran-GTP to Ran-GDP, and a nuclear *guanine exchange factor (GEF)* that promotes the exchange of GDP for GTP and thus converts Ran-GDP to Ran-GTP. Because *Ran-GAP* is located in the cytosol and *Ran-GEF* is located in the nucleus, the cytosol primarily contains Ran-GDP, and the nucleus primarily contains Ran-GTP.

This gradient of the two conformational forms of Ran drives nuclear **transport** in the appropriate direction. Docking of nuclear import receptors to FG-repeats on the cytosolic side of the nuclear pore complex, for example, occurs only when these receptors are loaded with an appropriate cargo. The import receptors with their bound cargo then move along tracks lined by FG-repeat sequences until they reach the nuclear side of the pore complex, where Ran-GTP binding causes the import receptors to release their cargo. By favoring cargo-dependent loading of import receptors onto the FG-repeat track in the cytosol and Ran-GTP-dependent cargo release in the nucleus, the nuclear localization of Ran-GTP imposes directionality.

Having discharged its cargo in the nucleus, the empty import receptor with Ran-GTP bound is transported back through the pore complex to the cytosol. There, two cytosolic proteins, *Ran Binding* ***Protein*** and Ran-GAP collaborate to convert Ran-GTP to Ran-GDP. The Ran Binding **Protein** first displaces Ran-GTP from the import receptor, which allows Ran-GAP to trigger Ran to hydrolyze its bound GTP. The Ran-GDP then dissociates from the Ran Binding**Protein** and is reimported into the nucleus, thereby completing the cycle.

Nuclear export occurs by a similar mechanism, except that Ran-GTP in the nucleus promotes cargo binding to the export receptor and the binding of the loaded receptor to the nuclear side of the pore complex. Once in the cytosol, Ran encounters Ran-GAP and Ran Binding **Protein** and hydrolyses its bound GTP. The export receptor then releases both its cargo and Ran-GDP in the cytosol and dissociates from the pore complex, and free export receptors are returned to the nucleus to complete the cycle.

Transport Between the Nucleus and Cytosol Can Be Regulated by Controlling Access to theTransport Machinery

Some proteins, such as those that bind newly made mRNAs in the nucleus, contain both nuclear localization and nuclear export signals. These proteins continually shuttle between the nucleus and the cytosol. The steady-state localization of such *shuttling proteins* is determined by the relative rates of their import and export. If the rate of import exceeds the rate of export, a **protein** will be located primarily in the nucleus. Conversely, if the rate of export exceeds the rate of import, a **protein** will be located primarily in the cytosol. Thus, changing the rate of import, export, or both, can change the location of a **protein**.

Some shuttling proteins move continuously in and out of the nucleus. In other cases, however, the **transport** is stringently controlled. As discussed in Chapter 7, the activity of some gene regulatory proteins is controlled by keeping them out of the nuclear compartment until they are needed there. In many cases, this control depends on the regulation of nuclear localization and export signals; these can be turned on or off, often by phosphorylation of adjacent amino acids.

Other gene regulatory proteins are bound to inhibitory cytosolic proteins that either anchor them in the cytosol (through interactions with the cytoskeleton or with specific organelles), or mask their nuclear localization signals so that they are unable to interact with nuclear import receptors. When the cell receives an appropriate stimulus, the gene regulatory**protein** is released from its cytosolic anchor or mask and is transported into the nucleus. One important example is the latent gene regulatory **protein** that controls the expression of proteins involved in cholesterol metabolism. The **protein**is made and stored in an inactive form as a transmembrane **protein** in the ER. When deprived of cholesterol, the cell activates specific proteases that cleave the **protein**, releasing its cytosolic domain. This domain is then imported into the nucleus, where it activates the transcription of genes required for cholesterol import and synthesis.

Cells control the export of RNA from the nucleus in a similar way. Messenger RNAs become bound to proteins that are loaded onto the RNA as transcription and splicing proceed. These proteins contain nuclear export signals that are recognized by export receptors that guide the RNA out of the nucleus through nuclear pore complexes. Upon entry into the cytosol, the proteins coating the RNA are stripped off and rapidly returned to the nucleus. Other RNAs, such as snRNAs and tRNAs, are exported by different sets of nuclear export receptors.

Incompletely processed pre-mRNAs are actively retained in the nucleus, anchored to the nuclear transcription and splicing machinery, which releases an RNA molecule only after its processing is completed. Genetic studies in yeastshow that a mutant pre-mRNA that cannot properly engage with the splicing machinery is improperly exported as an unspliced molecule.

The Nuclear Envelope Is Disassembled During Mitosis

The nuclear lamina is a meshwork of interconnected **protein** subunits called nuclear lamins. The lamins are a special class of intermediate filament proteins (discussed in Chapter 16) that polymerize into a two-dimensional lattice (Figure 12-20). The nuclear lamina gives shape and stability to the nuclear envelope, to which it is anchored by attachment to both the nuclear pore complexes and integral membrane proteins of the inner nuclear membrane. The lamina also interacts directly with chromatin, which itself interacts with the integral membrane proteins of the inner nuclear membrane. Together with the lamina, these membrane proteins provide structural links between the DNA and the nuclear envelope.

When a nucleus disassembles during mitosis, the nuclear lamina depolymerizes. The disassembly is at least partly a consequence of direct phosphorylation of the nuclear lamins by the cyclin-dependent kinase activated at the onset of mitosis (discussed in Chapter 17). At the same time, proteins of the inner nuclear membrane are phosphorylated, and the nuclear pore complexes disassemble and disperse in the cytosol. Nuclear envelope membrane proteins—no longer tethered to the pore complexes, lamina, or chromatin—diffuse throughout the ER membrane. Together, these events break down the barriers that normally separate the nucleus and cytosol, and these nuclear proteins that are not bound to membranes or chromosomes intermix completely with the cytosol of the dividing cell.

Later in mitosis (in late anaphase), the nuclear envelope reassembles on the surface of the chromosomes, as inner nuclear membrane proteins and dephosphorylated lamins rebind to chromatin. ER membranes wrap around groups of chromosomes and continue fusing until a sealed nuclear envelope is reformed. During this process, the nuclear pore complexes also reassemble and start actively reimporting proteins that contain nuclear localization signals. Because the nuclear envelope is initially closely applied to the surface of the chromosomes, the newly formed nucleus excludes all proteins except those initially bound to the mitotic chromosomes and those that are selectively imported through nuclear pore complexes. In this way, all other large proteins are kept out of the newly assembled nucleus.

Nuclear localization signals are not cleaved off after **transport** into the nucleus. This is presumably because nuclear proteins need to be imported repeatedly, once after every cell division. In contrast, once a **protein** molecule has been imported into any of the other membrane-enclosed organelles, it is passed on from generation to generation within thatcompartment and need never be translocated again; the signal sequence on these molecules is often removed after**protein** translocation.

Summary

The nuclear envelope consists of an inner and an outer nuclear membrane. The outer membrane is continuous with theER membrane, and the space between it and the inner membrane is continuous with the ER lumen. RNA molecules, which are made in the nucleus, and ribosomal subunits, which are assembled there, are exported to the cytosol, while all the proteins that function in the nucleus are synthesized in the cytosol and are then imported. The

extensive traffic of materials between the nucleus and cytosol occurs through nuclear pore complexes, which provide a direct passageway across the nuclear envelope.

Proteins containing nuclear localization signals are actively transported inward through the nuclear pore complexes, while RNA molecules and newly made ribosomal subunits contain nuclear export signals that direct their active**transport** outward through the pore complexes. Some proteins, including nuclear import and export receptors, continually shuttle between the cytosol and nucleus. The GTPase Ran, provides directionality for nuclear **transport**. The **transport** of nuclear proteins and RNA molecules through the pore complexes can be regulated by denying these molecules access to the **transport** machinery. Because nuclear localization signals are not removed, nuclear proteins can be imported repeatedly, as is required each time that the nucleus reassembles after mitosis.

The Transport of Proteins into Mitochondria and Chloroplasts

As discussed in Chapter 14, mitochondria and chloroplasts are double-membrane-enclosed organelles. They specialize in the synthesis of ATP, using energy derived from electron **transport** and oxidative phosphorylation in mitochondria and from photosynthesis in chloroplasts. Although both organelles contain their own DNA, ribosomes, and other components required for **protein** synthesis, most of their proteins are encoded in the cell nucleus and imported from thecytosol. Moreover, each imported **protein** must reach the particular organelle subcompartment in which it functions.

There are two subcompartments in mitochondria: the internal matrix space and the intermembrane space. These compartments are formed by the two concentric mitochondrial membranes: the inner membrane, which forms extensive invaginations, the *cristae*, and encloses the matrix space, and the outer membrane, which is in contact with the cytosol(Figure 12-22A). Chloroplasts have the same two subcompartments plus an additional subcompartment, the *thylakoidspace*, which is surrounded by the *thylakoid membrane*. Each of the subcompartments in mitochondria and chloroplasts contains a distinct set of proteins.

New mitochondria and chloroplasts are produced by the growth of preexisting organelles followed by fission (discussed in Chapter 14). Their growth depends mainly on the import of proteins from the cytosol. This requires that proteins be translocated across a number of membranes in succession and end up in the appropriate place. How this occurs is the subject of this section.

Translocation into the Mitochondrial Matrix Depends on a Signal Sequence and ProteinTranslocators

Proteins imported into the matrix of **mitochondria** are usually taken up from the cytosol within seconds or minutes of their release from ribosomes. Thus, in contrast to the **protein** translocation into the ER described later, mitochondrial proteins are first fully synthesized as precursor proteins in the cytosol and then translocated into mitochondria by a*posttranslational* mechanism. Most of the mitochondrial precursor proteins have a signal sequence at their N terminusthat is rapidly removed after import by a protease (the *signal peptidase)* in the mitochondrial matrix. The signal sequences are both necessary and sufficient for import of the proteins that contain them: through the use of genetic engineering techniques, these signals can be linked to any cytosolic **protein** to direct the **protein** into the mitochondrial matrix. Sequence comparisons and physical studies of different matrix signal sequences suggest that their common feature is the propensity to fold into an amphipathic α helix, in which positively charged residues are clustered on one side of the helix, while uncharged hydrophobic residues are clustered on the opposite side. This configuration—rather than a precise amino acid sequence—is recognized by specific receptor proteins that initiate**protein** translocation.

Protein translocation across mitochondrial membranes is mediated by multi-subunit **protein** complexes that function as**protein** translocators: the TOM complex functions across the outer membrane, and two TIM complexes, the TIM23 and TIM22 complexes, function across the inner membrane (Figure 12-24). TOM and TIM stand for translocase of the outer and inner mitochondrial membranes, respectively. These complexes contain some components that act as receptors for mitochondrial precursor proteins and other components that form the translocation channel. The TOM complex is required for the import of all nucleus-encoded mitochondrial proteins. It initially transports their signal sequences into the intermembrane space and helps to insert transmembrane proteins into the outer membrane. The TIM23 complex then transports some of these proteins into the matrix space, while helping to insert transmembrane proteins into the inner membrane. The TIM22 complex mediates the insertion of a subclass of inner membrane proteins, including the carrier**protein** that transports ADP, ATP, and phosphate. A third **protein** translocator in the inner mitochondrial membrane, the **OXA complex**, mediates the insertion of inner membrane proteins that are synthesized within the mitochondria. It also helps to insert some proteins that are initially transported into the matrix by the TOM and TIM complexes.

Mitochondrial Precursor Proteins Are Imported as Unfolded Polypeptide Chains

Almost everything we know about the molecular mechanism of **protein** import into mitochondria has been learned from analyses of cell-free, reconstituted **transport** systems. Mitochondria are first purified by differential centrifugation of homogenized cells and are then incubated with radiolabeled mitochondrial precursor proteins, which are generally taken up rapidly and efficiently. By changing the incubation conditions, it is possible to establish the biochemical requirements for the **transport**.

Mitochondrial precursor proteins do not fold into their native structures after they are synthesized; instead, they remain unfolded through interactions with other proteins in the cytosol. Some of these interacting proteins are general*chaperone proteins* belonging to the *hsp70 family* (discussed in Chapter 6), whereas others are dedicated to mitochondrial precursor proteins and bind directly to their signal sequences. All these interacting proteins help to prevent the precursor proteins from aggregating or folding up spontaneously before they engage with the TOM complex in the outer mitochondrial membrane. As a first step in the import process, the mitochondrial precursor proteins bind to importreceptor proteins of the TOM complex, which recognize the mitochondrial signal sequences. The interacting proteins are then stripped off, and the unfolded polypeptide chain is fed—signal sequence first—into the translocation channel.

Mitochondrial Precursor Proteins Are Imported into the Matrix at Contact Sites That Join the Inner and Outer Membranes

In principle, a **protein** could reach the mitochondrial matrix by crossing the two membranes one at a time, or it could pass through both at once. To distinguish between these possibilities, a cell-free mitochondrial import system was cooled to a low temperature, arresting the proteins at an intermediate step in the translocation process. The proteins that accumulated at this step already had their N-terminal signal sequence removed by the matrix signal peptidase, indicating that their N terminus must be in the matrix space. Yet, the bulk of the **protein** could still be attacked from outside the mitochondria by externally added proteolytic enzymes (Figure 12-25). This result demonstrates that the precursor proteins can pass through both mitochondrial membranes at once to enter the matrix. It is thought that theTOM complex first transports the mitochondrial targeting signal across the outer membrane. Once it reaches in theintermembrane space, the targeting signal binds to a TIM complex, opening the channel in the complex through which the polypeptide chain either enters the matrix or inserts into the inner membrane. Electron microscopists have noted numerous *contact sites* at which the inner and outer mitochondrial membranes are closely apposed, and it seems likely that translocation occurs at or near these sites.

Although the functions of the TOM and TIM complexes are usually coupled to allow **protein transport** across both membranes at the same time, both **protein** types of translocator can work independently. The TOM complex in isolated outer membranes, for example, can translocate the signal sequence of precursor proteins across the membrane. Similarly, mitochondria with experimentally disrupted outer membranes, and therefore with the TIM23 complex exposed on their surface, efficiently import precursor proteins into the matrix space. Despite the independent functional roles of the TOM and TIM translocators, the two mitochondrial membranes at contact sites may be permanently held together by the TIM23 complex, which spans both membranes.

ATP Hydrolysis and a H^+ Gradient are Used to Drive Protein Import into Mitochondria

Directional **transport** requires energy. In most biological systems, energy is supplied by ATP hydrolysis. Mitochondrial**protein** import is fueled by ATP hydrolysis at two discrete sites, one outside the mitochondria and one in the matrix (Figure 12-27). In addition, another energy source is required: an electrochemical H^+ gradient across the inner mitochondrial membrane.

The first requirement for energy occurs at the initial stage of the translocation process, when the unfolded precursor**protein**, associated with chaperone proteins, interacts with the mitochondrial import receptors. As discussed in Chapter 6, the release of newly synthesized polypeptides from the hsp70 family of chaperone proteins requires ATP hydrolysis. Experimentally, the requirement for hsp70 and ATP in the cytosol can be bypassed if the precursor **protein** is artificially unfolded prior to adding it to purified mitochondria.

Once the signal sequence has passed through the TOM complex and has become bound to either TIM complex, furthertranslocation through the TIM requires an electrochemical H^+ gradient across the inner membrane. The electrochemical gradient is maintained by the pumping of H^+ from the matrix to the intermembrane space, driven by electron **transport**processes in the inner membrane. By contrast, the outer mitochondrial membrane, like that of Gram-negative bacteria, contains a pore-forming **protein** called porin and is thus freely permeable to inorganic ions and metabolites (but not to most proteins), so that ion gradients cannot be maintained across it. The energy in the electrochemical H^+ gradient across the inner membrane is not only used to help drive most of the cell's ATP synthesis; it is also used to drive the translocation of the targeting signals through the TIM complexes. The precise mechanism by which this occurs is not known, but it is possible that the electrical components of the gradient helps to drive the positively charged signal sequence into the matrix by electrophoresis.

Hsp70 chaperone proteins in the matrix space also have a role in the translocation process, and they are the third point in the import process at which ATP is consumed, as we discuss next.

Repeated Cycles of ATP Hydrolysis by Mitochondrial Hsp70 Complete the Import Process

We know that **mitochondrial hsp70** is crucial to the import process, because mitochondria containing mutant forms of the **protein** fail to import precursor proteins. Like its cytosolic cousin, mitochondrial hsp70 has a high affinity for unfolded polypeptide chains and it binds tightly to an imported **protein** as soon as it emerges from the translocator in the matrix. The hsp70 then releases the **protein** in an ATP-dependent step. This energy-driven cycle of binding and subsequent release is thought to provide the final driving force needed to complete **protein** import after a **protein** has initially inserted into the TIM23 complex.

Two models have been proposed to explain how ATP hydrolysis by mitochondrial hsp70 drives **protein** import. In both models, hsp70 proteins are closely associated with the TIM23 complex, which deposits them onto the translocatingpolypeptide chain as it emerges

into the matrix. In the *thermal ratchet model*, the emerging chain slides back and forth in the TIM23 translocation channel by thermal motion. Each time a sufficiently long portion of the chain is exposed in the matrix, an hsp70 molecule binds to it, preventing further backsliding and thereby making the movement directional. Thus, a hand-over-hand binding of multiple hsp70 proteins translocates the polypeptide chain into the matrix. In the *cross-bridge ratchet model* (Figure 12-28B), the hsp70 proteins that bind to the emerging polypeptide chain undergo a conformational change, driven by ATP hydrolysis, that actively pulls a segment of the polypeptide chain into the matrix. A new hsp70 molecule can then bind to the segment just pulled in and repeat the cycle. In both models, therefore, hsp70 functions as a ratchet that prevents backsliding of the emerging polypeptide chain.

After the initial interaction with mitochondrial hsp70, many imported proteins are passed on to another chaperone**protein**, *mitochondrial hsp60*. As discussed in Chapter 6, hsp60 provides a chamber for the unfolded polypeptide chain that facilitates its folding by binding and releasing it through cycles of ATP hydrolysis.

Protein Transport into the Inner Mitochondrial Membrane and the Intermembrane Space Requires Two Signal Sequences

Proteins that are integrated into the inner mitochondrial membrane or that operate in the intermembrane space are initially transported from the cytosol by the same mechanism that transports proteins into the matrix. In some cases they are first transferred into the matrix (see Figure 12-26). A hydrophobic amino acid sequence, however, is strategically placed after the N-terminal signal sequence that guides import into the matrix. Once the N-terminal signal sequence has been removed by the matrix signal peptidase, the hydrophobic sequence functions as a new N-terminal signal sequence to translocate the **protein** from the matrix into or across the inner membrane, using the OXA complex as the translocator. The OXA complex is also used to insert proteins encoded in the mitochondrion into the inner membrane. Closely related translocators are found in the plasma membranes of bacteria and in the thylakoids of chloroplasts, where they are thought to help to insert membrane proteins by a similar mechanism.

An alternative route to the inner membrane avoids excursion into the matrix altogether. In this case, the TIM23 translocator in the inner membrane binds to the hydrophobic sequence that follows the N-terminal signal sequence and initiates import, causing it to act as a *stop-transfer sequence* that prevents further translocation across the inner membrane. After the N-terminal signal sequence has been cleaved off, the remainder of the **protein** is pulled through the TOM complex in the outer membrane into the intermembrane space. Different proteins use one or the other of these two pathways to the inner membrane or intermembrane space.

Proteins destined for the intermembrane space are first inserted via their hydrophobic signal sequence into the inner membrane, and then cleaved by a signal peptidase in the intermembrane space to release the mature polypeptide chain as a soluble **protein**. Many of these proteins attach as peripheral membrane proteins to the outer surface of the inner

membrane, where they form subunits of **protein** complexes that also contain transmembrane proteins.

Mitochondria are the principal site of ATP synthesis in the cell, but they also contain many metabolic enzymes, such as those of the citric acid cycle. Thus, in addition to proteins, mitochondria must also **transport** small metabolites across their membranes. While the outer membrane contains porins that make the membrane freely permeable to small molecules, the inner membrane does not. Instead, the **transport** of a vast number of small molecules across the inner membrane is mediated by a family of metabolite-specific carrier proteins. In yeast cells, these proteins comprise a family of 35 different proteins, of which the most abundant are those that **transport** ADP and ATP, or phosphate. These carrier proteins in the inner membrane are multipass transmembrane proteins, which do not have cleavable signal sequences at their N termini but, instead, contain internal signal sequences. These proteins cross the TOM complex in the outer membrane and are inserted into the inner membrane by the TIM22 complex. Their integration into the inner membrane requires the electrochemical H^+ gradient, but not mitochondrial hsp70 or ATP. The energetically favorable partitioning of the hydrophobic transmembrane regions into the inner membrane is likely to help drive integration.

Two Signal Sequences Are Required to Direct Proteins to the Thylakoid Membrane in Chloroplasts

Protein transport into chloroplasts resembles **transport** into mitochondria in many respects. Both processes occur posttranslationally, use separate translocation complexes in each membrane, occur at contact sites, require energy, and use amphipathic N-terminal signal sequences that are removed after use. With the exception of some of the chaperone molecules, however, the **protein** components that form the translocation complexes are different. Moreover, whereas mitochondria harness the electrochemical H^+ gradient across their inner membrane to drive **transport**, chloroplasts, which have an electrochemical H^+ gradient across their thylakoid membrane but not their inner membrane, use the hydrolysis of GTP and ATP to power import across their double membrane. The functional similarities may thus result from convergent evolution, reflecting the common requirements for translocation across a double-membrane system.

Although the signal sequences for import into chloroplasts superficially resemble those for import into mitochondria, mitochondria and chloroplasts are both present in the same plant cells, so proteins must choose appropriately between them. In plants, for example, a bacterial enzyme can be directed specifically to mitochondria if it is experimentally joined to an N-terminal signal sequence of a mitochondrial **protein**; the same enzyme joined to an N-terminal signal sequence of a chloroplast **protein** ends up in chloroplasts. The different signal sequences can therefore be distinguished by the import receptors on each organelle.

Chloroplasts have an extra membrane-enclosed compartment, the thylakoid. Many chloroplast proteins, including the**protein** subunits of the photosynthetic system and of

the ATP synthase (discussed in Chapter 14), are embedded in thethylakoid membrane. Like the precursors of some mitochondrial proteins, these proteins are transported from the cytosolto their final destination in two steps. First, they pass across the double membrane at contact sites into the matrix spaceof the chloroplast, called the stroma, and then they are translocated into the thylakoid membrane (or across this membrane into the thylakoid space) (Figure 12-30A). The precursors of these proteins have a hydrophobic thylakoidsignal sequence following the N-terminal chloroplast signal sequence. After the N-terminal signal sequence has been used to import the **protein** into the stroma, it is removed by a stromal signal peptidase (analogous to the matrix signal peptidase in mitochondria). This cleavage unmasks the thylakoid signal sequence, which then initiates **transport** across the thylakoid membrane. There are at least four routes for proteins to cross or become integrated into the thylakoid membrane, distinguished by their need for different stromal chaperones and energy sources.

Summary

Although mitochondria and chloroplasts have their own genetic systems, they produce only a small proportion of their own proteins. Instead, the two organelles import most of their proteins from the cytosol, using similar mechanisms. In both cases, proteins are imported in an unfolded state. Proteins are translocated into the mitochondrial matrix space by passing through the TOM and TIM complexes at sites of adhesion between the outer and inner membranes known as contact sites. Translocation into mitochondria is driven by both ATP hydrolysis and an electrochemical H^+gradient across the inner membrane, whereas translocation into chloroplasts is driven solely by the hydrolysis of GTP and ATP.

Chaperone proteins of the cytosolic hsp70 family maintain the precursor proteins in an unfolded, translocation-competent state. A second set of hsp70 proteins in the matrix or stroma bind to the incoming polypeptide chain to pull it into the organelle. Only proteins that contain a specific signal sequence are translocated into mitochondria or chloroplasts. The signal sequence is usually located at the N terminus and is cleaved off after import. Some imported proteins also contain an internal signal sequence that guides their further **transport**. **Transport** across or into the inner membrane can occur as a second step if a hydrophobic signal sequence is unmasked when the first signal sequence is removed. In chloroplasts, import from the stroma into the thylakoid likewise requires a second signal sequence and can occur by one of several routes.

The Endoplasmic Reticulum

All eucaryotic cells have an **endoplasmic reticulum (ER)**. Its membrane typically constitutes more than half of the total membrane of an average animal cell. The ER is organized into a netlike labyrinth of branching tubules and flattened sacs extending throughout the cytosol (Figure 12-35). The tubules and sacs are all thought to interconnect, so that the ER membrane forms a continuous sheet enclosing a single internal space. This highly convoluted space is called the ER lumen or the *ER cisternal space,* and it often occupies

more than 10% of the total cell volume (see Table 12-1). The ER membrane separates the ER lumen from the cytosol, and it mediates the selective transfer of molecules between these two compartments.

The ER has a central role in lipid and protein biosynthesis. Its membrane is the site of production of all the transmembrane proteins and lipids for most of the cell's organelles, including the ER itself, the Golgi apparatus, lysosomes, endosomes, secretory vesicles, and the plasma membrane. The ER membrane makes a major contribution to mitochondrial and peroxisomal membranes by producing most of their lipids. In addition, almost all of the proteins that will be secreted to the cell exterior—plus those destined for the lumen of the ER, Golgi apparatus, or lysosomes—are initially delivered to the ER lumen.

Membrane-bound Ribosomes Define the Rough ER

The ER captures selected proteins from the cytosol as they are being synthesized. These proteins are of two types:*transmembrane proteins,* which are only partly translocated across the ER membrane and become embedded in it, and*water-soluble proteins,* which are fully translocated across the ER membrane and are released into the ER lumen. Some of the transmembrane proteins function in the ER, but many are destined to reside in the plasma membrane or the membrane of another organelle. The water-soluble proteins are destined either for the lumen of an organelle or for secretion. All of these proteins, regardless of their subsequent fate, are directed to the ER membrane by the same kind ofsignal sequence and are translocated across it by similar mechanisms.

In mammalian cells, the import of proteins into the ER begins before the polypeptide chain is completely synthesized—that is, import is a co-translational process. This distinguishes the process from the import of proteins into mitochondria, chloroplasts, nuclei, and peroxisomes, which are posttranslational processes. Since one end of the protein is usually translocated into the ER as the rest of the polypeptide chain is being made, the protein is never released into the cytosoland therefore is never in danger of folding up before reaching the translocator in the ER membrane. Thus, in contrast to the posttranslational import of proteins into mitochondria and chloroplasts, chaperone proteins are not required to keep the protein unfolded. The ribosome that is synthesizing the protein is directly attached to the ER membrane. These membrane-bound ribosomes coat the surface of the ER, creating regions termed rough endoplasmic reticulum, or **rough ER**.

There are therefore two spatially separate populations of ribosomes in the cytosol. Membrane-bound ribosomes, attached to the cytosolic side of the ER membrane, are engaged in the synthesis of proteins that are being concurrently translocated into the ER. Free ribosomes, unattached to any membrane, synthesize all other proteins encoded by the nuclear genome. Membrane-bound and free ribosomes are structurally and functionally identical. They differ only in the proteins they are making at any given time. When a ribosome happens to be making a protein with an ER signal sequence, the signal directs the ribosome to the ER membrane.

Since many ribosomes can bind to a single mRNA molecule, a polyribosome is usually formed, which becomes attached to the ER membrane, directed there by the signal sequences on multiple growing polypeptide chains. The individual ribosomes associated with such an mRNA molecule can return to the cytosol when they finish translation near the 3′ end of the mRNA molecule. The mRNA itself, however, remains attached to the ER membrane by a changing population of ribosomes, each transiently held at the membrane by the translocator. In contrast, if an mRNA molecule encodes a protein that lacks an ER signal sequence, the polyribosome that forms remains free in the cytosol, and its protein product is discharged there. Therefore, only those mRNA molecules that encode proteins with an ER signal sequence bind to rough ER membranes; those mRNA molecules that encode all other proteins remain free in the cytosol. Individual ribosomal subunits are thought to move randomly between these two segregated populations of mRNA molecules.

Smooth ER Is Abundant in Some Specialized Cells

Regions of ER that lack bound ribosomes are called smooth endoplasmic reticulum, or **smooth ER**. In the great majority of cells, such regions are scanty and are often partly smooth and partly rough. They are sometimes called*transitional ER* because they contain *ER exit sites* from which transport vesicles carrying newly synthesized proteins and lipids bud off for transport to the Golgi apparatus. In certain specialized cells, however, the smooth ER is abundant and has additional functions. In particular, it is usually prominent in cells that specialize in lipid metabolism. Cells that synthesize steroid hormones from cholesterol, for example, have an expanded smooth ER compartment to accommodate the enzymes needed to make cholesterol and to modify it to form the hormones.

The main cell type in the liver, the *hepatocyte*, is another cell with an abundant smooth ER. It is the principal site of production of lipoprotein particles, which carry lipids via the bloodstream to other parts of the body. The enzymes that synthesize the lipid components of lipoproteins are located in the membrane of the smooth ER, which also contains enzymes that catalyze a series of reactions to detoxify both lipid-soluble drugs and various harmful compounds produced by metabolism. The most extensively studied of these *detoxification reactions* are carried out by the*cytochrome P450* family of enzymes, which catalyze a series of reactions in which water-insoluble drugs or metabolites that would otherwise accumulate to toxic levels in cell membranes are rendered sufficiently water-soluble to leave the cell and be excreted in the urine. Because the rough ER alone cannot house enough of these and other necessary enzymes, a major portion of the membrane in a hepatocyte normally consists of smooth ER.

When large quantities of certain compounds, such as the drug phenobarbital, enter the circulation, detoxification enzymes are synthesized in the liver in unusually large amounts, and the smooth ER doubles in surface area within a few days. Once the drug has disappeared, the excess smooth ER membrane is specifically and rapidly removed by a lysosome-dependent process called *autophagocytosis* (discussed in Chapter 13). It is not known how these dramatic changes are regulated.

Another function of the ER in most eucaryotic cells is to sequester Ca^{2+} from the cytosol. The release of Ca^{2+} into the cytosol from the ER, and its subsequent reuptake, is involved in many rapid responses to extracellular signals, as discussed in Chapter 15. The storage of Ca^{2+} in the ER lumen is facilitated by the high concentrations of Ca^{2+}-binding proteins there. In some cell types, and perhaps in most, specific regions of the ER are specialized for Ca^{2+} storage. Muscle cells, for example, have an abundant specialized smooth ER, called the *sarcoplasmic reticulum*, which sequesters Ca^{2+} from the cytosol by means of a Ca^{2+}-ATPase that pumps in Ca^{2+} into its lumen. The release and reuptake of Ca^{2+} by the sarcoplasmic reticulum trigger the contraction and relaxation, respectively, of the myofibrils during each round of muscle contraction.

Rough and Smooth Regions of ER Can Be Separated by Centrifugation

To study the functions and biochemistry of the ER, it is necessary to isolate the ER membrane. This may seem like a hopeless task because the ER is intricately interleaved with other components of the cytosol. Fortunately, when tissues or cells are disrupted by homogenization, the ER breaks into fragments and reseals into many small (~100–200 nm in diameter) closed vesicles called microsomes, which are relatively easy to purify. Microsomes derived from rough ER are studded with ribosomes and are called *rough microsomes*. The ribosomes are always found on the outside surface, so the interior of the microsome is biochemically equivalent to the lumenal space of the ER. Because they can be readily purified in functional form, rough microsomes are especially useful for studying the many processes performed by the rough ER. To the biochemist they represent small authentic versions of the rough ER, still capable ofprotein synthesis, protein glycosylation, Ca^{2+} uptake, and lipid synthesis.

Many vesicles of a size similar to that of rough microsomes, but lacking attached ribosomes, are also found in these homogenates. Such *smooth microsomes* are derived in part from smooth portions of the ER and in part from vesiculated fragments of the plasma membrane, Golgi apparatus, endosomes, and mitochondria (the ratio depending on the tissue). Thus, whereas rough microsomes are derived from rough portions of ER, the origins of smooth microsomes cannot be as easily assigned. The microsomes of the liver are an exception. Because of the unusually large quantities of smooth ER in hepatocytes, most of the smooth microsomes in liver homogenates are derived from smooth ER.

The ribosomes attached to rough microsomes make them more dense than smooth microsomes. As a result, the rough and smooth microsomes can be separated from each other by equilibrium centrifugation. When the separated rough and smooth microsomes of liver are compared with regard to such properties asenzyme activity or polypeptide composition, they are very similar, although not identical: apparently most of the components of the ER membrane can diffuse freely between the rough and smooth regions, as would be expected for a continuous, fluid membrane. The rough microsomes, however, contain more than 20 proteins that are not present in smooth microsomes, showing that some separation mechanism must exist for a subset of ER membrane proteins. Some of the proteins in this

subset help to bind ribosomes to the rough ER, while others presumably produce the flattened shape of this part of the ER. It is not clear whether these membrane proteins are confined to the rough ER by forming large two-dimensional assemblies in the lipid bilayer, or whether they are instead held in place by interactions with a network of structural proteins on one or the other face of the rough ER membrane.

Signal Sequences Were First Discovered in Proteins Imported into the Rough ER

Signal sequences (and the signal sequence strategy of protein sorting) were first discovered in the early 1970s in secreted proteins that are translocated across the ER membrane as a first step toward their eventual discharge from the cell. In the key experiment, the mRNA encoding a secreted protein was translated by ribosomes *in vitro*. When microsomes were omitted from this cell-free system, the protein synthesized was slightly larger than the normal secreted protein, the extra length being the N-terminal leader peptide. In the presence of microsomes derived from the rough ER, however, a protein of the correct size was produced. These results were explained by the *signal hypothesis*, which postulated that the leader serves as an ER signal sequence that directs the secreted protein to the ER membrane and is then cleaved off by a *signal peptidase* in the ER membrane before the polypeptide chain has been completed.

According to the signal hypothesis, the secreted protein should be extruded into the lumen of the microsome during its synthesis *in vitro*. This can be demonstrated by treatment with a protease: a newly synthesized protein made in the absence of microsomes is degraded when the protease is added to the medium, whereas the same protein made in the presence of microsomes remains intact because it is protected by the microsomal membrane. When proteins without ERsignal sequences are similarly synthesized *in vitro*, they are not imported into microsomes and are therefore degraded by protease treatment.

The signal hypothesis has been thoroughly tested by genetic and biochemical experiments and is found to apply to both plant and animal cells, as well as to protein translocation across the bacterial plasma membrane and, as we have seen, the membranes of mitochondria, chloroplasts, and peroxisomes. N-terminal ER signal sequences guide not only soluble secreted proteins, but also the precursors of all other proteins made by ribosomes bound to the rough ER membrane, including membrane proteins. The signaling function of these peptides has been demonstrated directly by usingrecombinant DNA techniques to attach ER signal sequences to proteins that do not normally have them; the resulting fusion proteins are directed to the ER.

Cell-free systems in which proteins are imported into microsomes have provided powerful assay procedures for identifying, purifying, and studying the various components of the molecular machinery responsible for the ER import process.

A Signal-Recognition Particle (SRP) Directs ER Signal Sequences to a Specific Receptor in the Rough ER Membrane

The ER signal sequence is guided to the ER membrane by at least two components: a **signal-recognition particle(SRP)**, which cycles between the ER membrane and the cytosol and binds to the signal sequence, and an **SRPreceptor** in the ER membrane. The SRP is a complex particle consisting of six different polypeptide chains bound to a single small RNA molecule (Figure 12-41A). Homologs of the SRP and its receptor are found in all organisms that have been studied, indicating that this protein-targeting mechanism arose early in evolution and has been conserved.

ER signal sequences vary greatly in amino acid sequence, but each has eight or more nonpolar amino acids at its center. How can the SRP bind specifically to so many different sequences? The answer has come from the crystal structure of the SRP protein, which shows that the signal-sequence-binding site is a large hydrophobic pocket lined by methionines. Because methionines have an unbranched, flexible side chains, the pocket is sufficiently plastic to accommodate hydrophobic signal sequences of different sequences and shapes.

The SRP binds to the ER signal sequence as soon as the peptide has emerged from the ribosome. This causes a pause inprotein synthesis, the pause presumably gives the ribosome enough time to bind to the ER membrane before the synthesis of the polypeptide chain is completed, thereby ensuring that the protein is not released into the cytosol. This safety device may be especially important for secreted and lysosomal hydrolases that could wreak havoc in the cytosol; however, cells that secrete large amounts of hydrolases take the added precaution of having high concentrations of hydrolase inhibitors in their cytosol.

Once formed, the SRP-ribosome complex binds to the SRP receptor, which is an integral membrane protein exposed only on the cytosolic surface of the rough ER membrane. This interaction brings the SRP-ribosome complex to a protein translocator. The SRP and SRP receptor are then released, and the growing polypeptide chain is transferred across the membrane.

The Polypeptide Chain Passes Through an Aqueous Pore in the Translocator

It has long been debated whether polypeptide chains are transferred across the ER membrane in direct contact with thelipid bilayer or through a pore in a protein translocator. The debate ended with the purification of the protein translocator, which was shown to form a water-filled pore in the membrane through which the polypeptide chain traverses the membrane. The translocator, called the **Sec61 complex**, consists of three or four protein complexes, each composed of three transmembrane proteins, that assemble into a donutlike structure.

When a ribosome binds, the central hole in the translocator lines up with a tunnel in the large ribosomal subunit through which the growing polypeptide chain exits from the ribosome. The bound ribosome forms a tight seal with the translocator, such that the space inside the ribosome is continuous with the lumen of the ER and no molecules can escape from the ER. The pore in the translocator cannot be open permanently, however; if it were, Ca^{2+} would leak

out of the ER when the ribosome detaches. It is thought that a lumenal ER protein serves as a plug or that the translocator itself can rearrange to close the pore when no ribosome is bound. Thus, the pore is a dynamic structure that opens only transiently when a ribosome with a growing polypeptide chain attaches to the ER membrane.

The signal sequence in the growing polypeptide chain is thought to trigger the opening of the pore: after the signal sequence is released from the SRP and the growing chain has reached a sufficient length, the signal sequence binds to a specific site inside the pore itself, thereby opening the pore. An ER signal sequence is therefore recognized twice: first, by an SRP in the cytosol, and then by a binding site in the ER protein translocator. This may help to ensure that only appropriate proteins enter the lumen of the ER.

Translocation Across the ER Membrane Does Not Always Require Ongoing Polypeptide Chain Elongation

As we have seen, translocation of proteins into mitochondria, chloroplasts, and peroxisomes occurs posttranslationally, after the protein has been made and released into the cytosol, whereas translocation across the ER membrane usually occurs during translation (co-translationally). This explains why ribosomes are bound to the ER but usually not to other organelles.

Some proteins, however, are imported into the ER after their synthesis has been completed, demonstrating thattranslocation does not always require ongoing translation. Posttranslational protein translocation is especially common across the ER membrane in yeast cells and across the bacterial plasma membrane (which is thought to be evolutionarily related to the ER; see Figure 12-4). To function in posttranslational translocation, the translocator needs accessory proteins that feed the polypeptide chain into the pore and drive translocation. In bacteria, a translocationmotor protein, the *SecA ATPase*, attaches to the cytosolic side of the translocator, where it undergoes cyclic conformational changes driven by ATP hydrolysis. Each time an ATP is hydrolyzed, a portion of the SecA protein inserts into the pore of the translocator, pushing a short segment of the passenger protein with it. As a result of this ratchet mechanism, the SecA protein pushes the polypeptide chain of the transported protein across the membrane.

Eucaryotic cells use a different set of accessory proteins that associate with the Sec61 complex. These proteins span theER membrane and use a small domain on the lumenal side of the ER membrane to deposit an hsp70-like chaperoneprotein (called *BiP*, for *b*inding *p*rotein) onto the polypeptide chain as it emerges from the pore into the ER lumen. Unidirectional translocation is driven by cycles of BiP binding and release, as described earlier for the mitochondrial hsp70 proteins that pull proteins across mitochondrial membranes.

Proteins that are transported into the ER by a posttranslational mechanism are first released into the cytosol, where they are prevented from folding up by binding to chaperone proteins, as discussed earlier for proteins destined for mitochondria and chloroplasts. In all of these

cases where translocation occurs without a ribosome sealing the pore, it remains a mystery how the polypeptide chain can slide through the pore in the translocator without allowing ions and other molecules to pass through.

The ER Signal Sequence Is Removed from Most Soluble Proteins After Translocation

We have seen that in chloroplasts and mitochondria, the signal sequence is cleaved from precursor proteins once it has crossed the membrane. Similarly, N-terminal ER signal sequences are removed by a signal peptidase on the lumenal side of the ER membrane. The signal sequence by itself, however, is not sufficient for signal cleavage by the peptidase; this requires an adjacent cleavage site that is specifically recognized by the peptidase. We shall see below that ER signal sequences that occur within the polypeptide chain—rather than at the N-terminus—do not have these recognition sites and are never cleaved; instead, they can serve to retain transmembrane proteins in the lipid bilayer after the translocationprocess has been completed.

The N-terminal ER signal sequence of a soluble protein has two signaling functions. It directs the protein to the ERmembrane, and it serves as a start-transfer signal (or start-transfer peptide) that opens the pore. Even after it is cleaved off by signal peptidase, the signal sequence is thought to remain bound to the translocator while the rest of the protein is threaded continuously through the membrane as a large loop. Once the C-terminus of the protein has passed through the membrane, the translocated protein is released into the ER lumen. The signal sequence is released from the pore and rapidly degraded to amino acids by other proteases in the ER.

While bound in the translocation pore, signal sequences are in contact not only with the Sec61 complex, which forms the walls of the pore, but also with the hydrophobic lipid core of the membrane. This was shown in chemical cross-linking experiments in which signal sequences and the hydrocarbon chains of lipids could be covalently linked together. To release the signal sequence into the membrane, the translocator has to open laterally. The translocator is therefore gated in two directions: it can open to form a pore across the membrane to let the hydrophilic portions of proteins cross the lipid bilayer, and it can open laterally within the membrane to let hydrophobic portions of proteins partition into the bilayer. This lateral gating mechanism is crucial for the insertion of transmembrane proteins into the lipid bilayer, as we discuss next.

In Single-Pass Transmembrane Proteins, a Single Internal ER Signal Sequence Remains in the Lipid Bilayer as a Membrane-spanning α Helix

The translocation process for proteins destined to remain in the membrane is more complex than it is for soluble proteins, as some parts of the polypeptide chain are translocated across the lipid bilayer whereas others are not. Nevertheless, all modes of insertion of membrane proteins can be considered as variants of the sequence of events just

described for transferring a soluble protein into the lumen of the ER. We begin by describing the three ways in whichsingle-pass transmembrane proteins become inserted into the ER.

In the simplest case, an N-terminal signal sequence initiates translocation, just as for a soluble protein, but an additional hydrophobic segment in the polypeptide chain stops the transfer process before the entire polypeptide chain is translocated. This stop-transfer signal anchors the protein in the membrane after the ER signal sequence (the start-transfer signal) has been released from the translocator and has been cleaved off. The stop-transfer sequence is transferred into the bilayer by the lateral gating mechanism, and it remains there as a single α-helical membrane-spanning segment, with the N-terminus of the protein on the lumenal side of the membrane and the C-terminus on the cytosolic side.

In the other two cases, the signal sequence is internal, rather than at the N-terminal end of the protein. Like the N-terminal ER signal sequences, the internal signal sequence is recognized by an SRP, which brings the ribosome making the protein to the ER membrane and serves as a start-transfer signal that initiates the translocation of the protein. After release from the translocator, the internal start-transfer sequence remains in the lipid bilayer as a single membrane-spanning α helix.

Internal start-transfer sequences, can bind to the translocation apparatus in either of two orientations, and the orientation of the inserted start-transfer sequence, in turn, determines which protein segment (the one preceding or the one following the start-transfer sequence) is moved across the membrane into the ER lumen. In one case, the resultingmembrane protein has its C-terminus on the lumenal side, while in the other, it has its N-terminus on the lumenal side. The orientation of the start-transfer sequence depends on the distribution of nearby charged amino acids, as described in the figure legend.

Combinations of Start-Transfer and Stop-Transfer Signals Determine the Topology of Multipass Transmembrane Proteins

In multipass transmembrane proteins, the polypeptide chain passes back and forth repeatedly across the lipid bilayer (seeFigure 10-17). It is thought that an internal signal sequence serves as a start-transfer signal in these proteins to initiatetranslocation, which continues until a stop-transfer sequence is reached. In double-pass transmembrane proteins, for example, the polypeptide can then be released into the bilayer. In more complex multipass proteins, in which many hydrophobic α helices span the bilayer, a second start-transfer sequence reinitiates translocation further down the polypeptide chain until the next stop-transfer sequence causes polypeptide release, and so on for subsequent start-transfer and stop-transfer sequences.

Whether a given hydrophobic signal sequence functions as a start-transfer or stop-transfer sequence must depend on its location in a polypeptide chain, since its function can be switched by changing its location in the protein usingrecombinant DNA techniques. Thus, the distinction between start-transfer and stop-transfer sequences results mostly from their relative order in the growing polypeptide chain. It seems that the SRP begins scanning an

unfolded polypeptide chain for hydrophobic segments at its N-terminus and proceeds toward the C-terminus, in the direction that the protein is synthesized. By recognizing the first appropriate hydrophobic segment to emerge from the ribosome, the SRP sets the "reading frame": if translocation is initiated, the next appropriate hydrophobic segment is recognized as a stop-transfer sequence, causing the region of the polypeptide chain in between to be threaded across the membrane. A similar scanning process continues until all of the hydrophobic regions in the protein have been inserted into the membrane.

Because membrane proteins are always inserted from the cytosolic side of the ER in this programmed manner, all copies of the same polypeptide chain will have the same orientation in the lipid bilayer. This generates an asymmetrical ER membrane in which the protein domains exposed on one side are different from those domains exposed on the other. This asymmetry is maintained during the many membrane budding and fusion events that transport the proteins made in the ER to other cell membranes. Thus, the way in which a newly synthesized protein is inserted into the ER membrane determines the orientation of the protein in all of the other membranes as well.

When proteins are dissociated from a membrane and are then reconstituted into artificial lipid vesicles, a random mixture of right-side-out and inside-out protein orientations usually results. Thus, the protein asymmetry observed in cell membranes seems not to be an inherent property of the protein, but instead results solely from the process by which proteins are inserted into the ER membrane from the cytosol.

Intracellular Vesicular Traffic

Every cell must eat, and it must communicate with the world around it. In a procaryotic cell, all the eating and communicating takes place across the plasma membrane. The cell secretes digestive enzymes, for example, across the plasma membrane to the cell exterior. It then transports the small metabolites generated by digestion in the extracellular space across the same membrane into the cytosol. Eucaryotic cells, by contrast, have evolved an elaborate internal membrane system that allows them to take up macromolecules by the process of *endocytosis* and deliver them to digestive enzymes stored in lysosomes inside the cell. As a consequence, metabolites generated by the digestion of macromolecules are delivered directly from the lysosomes to the cytosol as they are produced. In addition to allowing the ingestion of macromolecules by the *endocytic pathway*, the internal membrane system allows eucaryotic cells to regulate the delivery of newly synthesized proteins, carbohydrates, and lipids to the cell exterior. The *biosynthetic-secretory pathway* allows the cell to modify the molecules it produces in a series of steps, store them until needed, and then deliver them to the exterior through a specific cell-surface domain by a process called *exocytosis*. An outline of the endocytic and biosynthetic-secretory pathways, which

ultimately connect the plasma membrane to the endoplasmic reticulum (ER) deep within the cell, is shown in.

The interior space, or *lumen*, of each membrane-enclosed compartment along the biosynthetic-secretory and endocytic pathways is topologically equivalent to the lumen of every other compartment. Moreover, these compartments are in constant communication, with molecules being passed from a donor compartment to a target compartment by means of numerous membrane-enclosed *transport packages*. Some of these packages are small spherical vesicles, while others are larger irregular vesicles or fragments of the donor compartment. We shall use the term *transport vesicle* to apply to all forms of these packages.

Vesicles continually bud off from one membrane and fuse with another, carrying membrane components and soluble molecules referred to as *cargo*. This membrane **traffic** flows along highly organized, directional routes, which allows the cell to secrete and eat. The biosynthetic-secretory pathway leads outward from the ER toward the Golgi apparatus and cell surface, with a side route leading to lysosomes, while the endocytic pathway leads inward from the plasma membrane. In each case, the flow of membrane between compartments is balanced, with retrieval pathways balancing the flow in the opposite direction, bringing membrane and selected proteins back to thecompartment of origin.

To perform its function, each transport vesicle that buds from a compartment must be selective. It must take up only the appropriate proteins and must fuse only with the appropriate target membrane. A vesicle carrying cargo from the Golgi apparatus to the plasma membrane, for example, must exclude proteins that are to stay in the Golgi apparatus, and it must fuse only with the plasma membrane and not with any other organelle.

We begin this chapter by considering the molecular mechanisms of budding and fusion that underlie all transport. We then discuss the fundamental problem of how, in the face of this transport, the differences between the compartments are maintained. Finally, we consider the function of the Golgi apparatus, lysosomes, secretory vesicles, and endosomes, as we trace the pathways that connect these organelles.

The Molecular Mechanisms of Membrane Transport and the Maintenance of Compartmental Diversity

Transport processes mediate a continual exchange of components between the ten or more chemically distinct,membrane-enclosed compartments that collectively comprise the biosynthetic-secretory and endocytic pathways. In the presence of this massive exchange, how can each compartment maintain its specialized character? To answer this question, we must first consider what defines the character of a compartment. Above all, it is the composition of the enclosing membrane: molecular markers displayed on the cytosolic surface of this membrane serve as guidance cues for incoming **traffic** and ensure that transport vesicles fuse only with the correct compartment, thereby dictating the pattern of **traffic** between one compartment and another. Many membrane markers, however, are

found on more than oneorganelle, and thus it is the specific combination of marker molecules that gives each organelle its unique molecular address.

How are these membrane markers kept at high concentration on one compartment and at low concentration on another? To answer this question, we need to consider how patches of membrane, enriched or depleted in specific components, bud off from one compartment and transfer to another. In this section we describe how this is achieved.

We begin by discussing the sorting events that underlie the segregation of proteins into separate membrane domains. This sorting process depends on the assembly of a special protein coat on the cytosolic face of the donor membrane. We shall therefore consider how coats form, what they are made of, and how they enable specific components of a membrane to be extracted and delivered to another membrane. Finally, we discuss how transport vesicles dock at the appropriate target membrane and fuse with it to deliver the contents to their target organelle.

There Are Various Types of Coated Vesicles

Most transport vesicles form from specialized, coated regions of membranes. They bud off as coated vesicles that have a distinctive cage of proteins covering their cytosolic surface. Before the vesicle fuses with a target membrane, the coat is discarded, as is required to allow the two cytosolic membrane surfaces to interact directly and fuse.

The coat is thought to perform two principal functions. First, it concentrates specific membrane proteins in a specialized membrane patch that then gives rise to the vesicle membrane. It thus helps select the appropriate molecules for transport. Second, the assembly of the coat proteins into curved, basketlike lattices deforms the membrane patch and thereby molds the forming vesicles, which explains why vesicles with the same type of coat have a relatively uniform size.

There are three well-characterized types of coated vesicles, which differ in their coat proteins: *clathrin-coated, COPI-coated,* and *COPII-coated* vesicles. Each type is used for different transport steps in the cell. Clathrin-coated vesicles, for example, mediate transport from the Golgi apparatus and from the plasma membrane, whereas COPI- and COPII-coated vesicles most commonly mediate transport from the ER and the Golgi. There is, however, much more variety than this short list suggests. As we discuss below, there are at least three types of clathrin-coated vesicles, each specialized for a different transport step, and the COPI-coated vesicles may be similarly diverse. Moreover, still other coats have been seen in the electron microscope, whose molecular compositions and functions are not yet known.

The Assembly of a Clathrin Coat Drives Vesicle Formation

Clathrin-coated vesicles were the first coated vesicles discovered and have been the most thoroughly studied. They provide a good example of how vesicles form.

The major protein component of clathrin-coated vesicles is clathrin itself. Each clathrin subunit consists of three large and three small polypeptide chains that together form a three-legged structure called a *triskelion.* Clathrin triskelions assemble into a basketlike convex framework of hexagons and pentagons to form coated pits on the cytosolic surface of membranes (Figure 13-6). Under appropriate conditions, isolated triskelions spontaneously self-assemble into typical polyhedral cages in a test tube, even in the absence of the membrane vesicles that these baskets normally enclose. Thus, the geometry of the clathrin cage is determined by the clathrin triskelion alone.

A second major coat protein in clathrin-coated vesicles is a multisubunit complex called adaptin. It is required both to bind the clathrin coat to the membrane and to trap various transmembrane proteins, including transmembrane receptors that capture soluble cargo molecules inside the vesicle—so-called *cargo receptors.* In this way, a selected set of membrane proteins and the soluble proteins that interact with them are packaged into each newly formed clathrin-coated transport vesicle.

There are at least four types of adaptins, each specific for a different set of cargo receptors. Clathrin-coated vesicles budding from different membranes use different adaptins and thus package different receptors and cargo molecules. The formation of a clathrin-coated pit is driven by forces generated by the successive assembly of adaptins and the clathrin coat on the cytosolic surface of the membrane. The lateral interactions between adaptins and between clathrin molecules then aid in bud formation.

Both The Pinching-off and Uncoating of Coated Vesicles Are Regulated Processes

As a clathrin-coated bud grows, soluble cytoplasmic proteins, including dynamin, assemble as a ring around the neck of each bud. Dynamin is a GTPase, which regulates the rate with which vesicles pinch off from the membrane. In the pinching-off process, the two noncytosolic leaflets of the membrane are brought into close proximity and fuse, sealing off the forming vesicle. To perform this task, dynamin recruits other proteins to the neck of the budding vesicle, which together with dynamin help to bend the membrane, either by directly distorting the bilayer structure locally or by changing the lipid composition, or both. A local change in lipid composition may result from the action of lipid-modifying enzymes that are recruited into the dynamin complex.

Once the vesicle is released from the membrane, the clathrin coat is rapidly lost. A chaperone protein of the hsp70 family functions as an uncoating ATPase, using the energy of ATP hydrolysis to peel off the coat. Another protein called *auxillin,* which is attached to the vesicle, is believed to activate the ATPase. Because the coated bud persists much longer than the coat on the vesicle, additional control mechanisms must somehow prevent the coat from being removed before it has formed a vesicle.

Although there are many similarities in vesicle budding at various locations in the cell, each cell membrane poses its own special challenges. The plasma membrane, for example, is

comparatively flat and stiff, owing to its cholesterol-richlipid composition and underlying cortical cytoskeleton. Thus, clathrin coats have to produce considerable force to introduce curvature, especially at the neck of the bud where dynamin and its associated proteins facilitate the sharp bends required for the pinching-off of the vesicle. In contrast, vesicle budding from many **intracellular** membranes occurs preferentially at regions where the membranes are already curved, such as the rims of Golgi cisternae or membrane tubules.

COPI-coated vesicles and **COPII-coated vesicles** transport material early in the secretory pathway: COPII-coated packages bud from the ER, and COPI-coated packages bud from pre-Golgi compartments and Golgi cisternae (seeFigure 13-5). The coats of COPI and COPII vesicles consist, in part, of large protein complexes that are composed of seven individual coat-protein subunits for COPI and four for COPII coats. Some COPI coat-protein subunits show sequence similarity to adaptins, suggesting a common evolutionary origin.

Not All Transport Vesicles are Spherical

Transport vesicles occur in various sizes and shapes. When living cells that have been genetically engineered to express fluorescent membrane components are observed under the microscope, endosomes and the *trans* Golgi network are seen to continually send out long tubules. Coat proteins assemble onto the tubules and help recruit specific cargo. The tubules then either withdraw, or they pinch off with the help of dynamin-like proteins and thus can serve as transport vesicles. Depending on the relative efficiencies of membrane tubulation and severing, differently sized portions of a donororganelle can pinch off.

Tubules have a much higher surface-to-volume ratio than the organelles from which they form. They are therefore relatively enriched in membrane proteins compared with soluble cargo proteins. As we discuss later, this property of tubules is used for sorting proteins in endosomes. Thus, **vesicular** transport does not necessarily occur only through uniformly sized spherical vesicles, but can involve larger portions of a donor organelle.

Monomeric GTPases Control Coat Assembly

The **vesicular** transport performed by both clathrin-coated and COP-coated vesicles depends on a variety of GTP-binding proteins that control both the spatial and the temporal aspects of membrane exchange. As discussed in Chapter 3, large families of GTP-binding proteins regulate diverse processes within cells. These proteins act as molecular switches that flip between an active state with GTP bound and an inactive state with GDP bound. Two classes of proteins regulate the flipping: *guanine-nucleotide-exchange factors (GEFs)* activate the proteins by catalyzing the exchange of GDP for GTP, and *GTPase-activating proteins (GAPs)* inactivate the proteins by triggering the hydrolysis of the bound GTP to GDP (see Figure 3-72). Although both monomeric GTP-binding proteins (monomeric GTPases) and trimeric GTP-binding proteins (G proteins) have essential roles in **vesicular** transport, the roles of the monomeric GTPases are better understood, and we focus our discussion on them.

To ensure that membrane **traffic** to and from an organelle is balanced, coat proteins must assemble only when and where they are needed. *Coat-recruitment GTPases,* which are members of a family of monomeric GTPases, usually serve this function. They include the ARF proteins, which are responsible for both COPI coat assembly and clathrincoat assembly at Golgi membranes, and the **Sar1 protein**, which is responsible for COPII coat assembly at the ERmembrane. Clathrin coat assembly at the plasma membrane is also thought to involve a GTPase, but its identity is unknown.

Coat-recruitment GTPases are usually found in high concentration in the cytosol in an inactive, GDP-bound state. When a COPII-coated vesicle is to bud from the ER membrane, a specific GEF embedded in the ER membrane binds to cytosolic Sar1, causing the Sar1 to release its GDP and bind GTP in its place (recall that GTP is present in much higher concentration in the cytosol than GDP and therefore will spontaneously bind after GDP is released). In its GTP-bound state, Sar1 exposes a hydrophobic tail, which inserts into the lipid bilayer of the ER membrane. The tightly bound Sar1 now recruits coat protein subunits to the ER membrane to initiate budding. Other GEFs and coat-recruitment GTPases operate in a similar way on other membranes.

Some coat protein subunits also interact, albeit more weakly, with the head groups of certain lipid molecules, in particular phosphatidic acid and phosphoinositides, as well as with the cytoplasmic tails of some of the membraneproteins they recruit into the bud. Activated coat-recruitment GTPases at sites of bud formation can locally activate phospholipase D, which converts some phospholipids to phosphatidic acid, thereby enhancing the binding of coat proteins. Together, these protein-protein and protein-lipid interactions tightly bind the coat to the membrane, causing the membrane to deform into a bud, which then pinches off as a coated vesicle.

The coat-recruitment GTPases also have a role in coat disassembly. The hydrolysis of bound GTP to GDP causes theGTPase to change its conformation so that its hydrophobic tail pops out of the membrane, causing the vesicle's coat to disassemble. Although it is not known what triggers the GTP hydrolysis process, it has been proposed that the GTPases work like timers, which hydrolyze GTP at a slow but predictable rate. COPII coats, for example, accelerate GTP hydrolysis by Sar1, thereby triggering coat disassembly at a certain time after coat assembly has begun. Thus, a fully formed vesicle will be produced only when bud formation occurs faster than the timed disassembly process; otherwise, disassembly will be triggered before a vesicle pinches off, and the process will have to start again at a more appropriate time and place. Completion of coating or contact with the target membrane may also trigger coat disassembly.

SNARE Proteins and Targeting GTPases Guide Membrane Transport

To ensure that membrane **traffic** proceeds in an orderly way, transport vesicles must be highly selective in recognizing the correct target membrane with which to fuse. Because of the diversity of membrane systems, a vesicle is likely to encounter many potential target

membranes before it finds the correct one. Specificity in targeting is ensured because all transport vesicles display surface markers that identify them according to their origin and type of cargo, while target membranes display complementary receptors that recognize the appropriate markers. This crucial recognition step is thought to be controlled mainly by two classes of proteins: *SNAREs* and targeting GTPases called *Rabs*. SNARE proteins seem to have a central role both in providing specificity and in catalyzing the fusion of vesicles with the target membrane. Rabs seem to work together with other proteins to regulate the initial docking and tethering of the vesicle to the target membrane.

There are at least 20 different SNAREs in an animal cell, each associated with a particular membrane-enclosedorganelle involved in the biosynthetic-secretory or endocytic pathway. These transmembrane proteins exist ascomplementary sets—vesicle membrane SNAREs, called **v-SNAREs**, and target membrane SNAREs, called **t-SNAREs**, (Figures 13-11 and 13-12). v-SNARESs and t-SNAREs have characteristic helical domains. When a v-SNARE interacts with a t-SNARE, the helical domains of one wrap around the helical domains of the other to form stable *trans-SNARE complexes*, which lock the two membranes together. We discuss later how the trans-SNAREcomplex is thought to contribute to membrane fusion. The specificity with which SNAREs interact determines the specificity of vesicle docking and fusion. In this way SNAREs specify compartment identity and govern the orderly transfer of material during **vesicular** transport.

SNAREs have been best characterized in nerve cells, where they mediate the docking and fusion of synaptic vesicles at the nerve terminal plasma membrane (see Figure 13-12). The SNARE complexes at neuron terminals are the targets of powerful neurotoxins that are secreted by the bacteria that cause tetanus and botulism. These toxins are highly specific proteases that enter specific neurons, cleave SNARE proteins in the nerve terminals and thereby block synaptic transmissions, often fatally.

Interacting SNAREs Need To Be Pried Apart Before They Can Function Again

Most SNARE proteins in cells have already participated in multiple rounds of membrane targeting and are sometimes present in a membrane as stable complexes with one or two partner SNAREs (see Figure 13-11). The complexes have to be disassembled before the SNAREs can mediate new rounds of transport. A crucial protein called NSF cycles between membranes and the cytosol and cata-lyzes the disassembly process. It is an ATPase that structurally resembles a minor class of cytosolic chaperone proteins that use the energy of ATP hydrolysis to solubilize and help refold denatured proteins. Similarly, NSF uses ATP to unravel the coiled-coil interaction between the helical domains of SNARE proteins, using several adaptor proteins to bind to the SNAREs.

The requirement for SNARE complex disassembly may help explain why membranes do not fuse indiscriminately in cells. If the t-SNAREs in a target membrane were always active, then any membrane containing an appropriate v-SNARE would fuse whenever the two membranes made contact. The requirement for NSF-mediated reactivation of SNAREs

allows the cell to control when and where membranes fuse. In addition, t-SNAREs in target membranes are often associated with inhibitory proteins that must be released before the t-SNARE can function. This release step may be controlled by the targeting GTPases, as we discuss next.

Rab Proteins Help Ensure the Specificity of Vesicle Docking

Rab proteins make an important contribution to the specificity of **vesicular** transport. They are monomeric GTPases, and with over 30 known members, they are the largest subfamily of these GTPases. Like the SNAREs, each Rab protein has a characteristic distribution on cell membranes and every organelle has at least one Rab protein on its cytosolic. Rab proteins are thought to facilitate and regulate the rate of vesicle docking and the matching of v-SNAREs and t-SNAREs, as required for membrane fusion.

Like the coat-recruitment GTPases discussed earlier (see Figure 13-10), Rab proteins cycle between a membrane and the cytosol. In their GDP-bound state they are inactive and in the cytosol, and in their GTP-bound state they are active and associated with the membrane of an organelle or transport vesicle. Many transport vesicles only form if a proper complement of SNARE and Rab proteins are included in the membrane, so as to allow the vesicle to dock and fuse appropriately.

The amino acid sequences of Rab proteins are most dissimilar near their C-terminal tails. Tail-swapping experiments indicate that the tail determines the **intracellular** location of each family member, presumably by enabling the protein to bind to complementary proteins, including GEFs, on the surface of the appropriate organelle. Once in its GTP-bound state and membrane-bound through a lipid anchor, a Rab protein is thought to bind to other proteins (called Rab effectors) that facilitate the docking process.

In contrast to the highly conserved structure of Rab proteins, the structures of **Rab effectors** vary greatly from one Rab protein to the next. One Rab effector, for example, is a large protein complex that serves to direct vesicles to specific sites on the plasma membrane for exocytosis. Vesicle fusion is limited to the region where this complex resides, even though the required t-SNAREs are uniformly distributed in the membrane. Some Rab effectors are long, filamentous, tethering proteins, which may restrict the movement of vesicles between adjacent Golgi cisternae. Others bind to Rab proteins in their active GTP-bound state and prevent premature GTP hydrolysis. Yet others are motor proteins that propel vesicles along actin filaments or microtubules to their proper target.

Although the Rab proteins and their effectors use widely different molecular mechanisms to influence **vesicular**transport, they have a common function. They help concentrate and tether vesicles near their target site and trigger the release of SNARE control proteins. In this way Rab proteins speed up the process by which appropriate SNARE proteins in two membranes find each other. Some Rab proteins function on the vesicle, whereas others function on the target membrane. The pairing of v-SNAREs and t-SNAREs then locks the docked vesicle onto the target membrane, readying it for fusion, which we discuss next. After fusion,

the Rab protein hydrolyzes its bound GTP and the inactive GDP-bound protein returns to the cytosol to participate in another cycle of transport.

SNAREs May Mediate Membrane Fusion

Once a transport vesicle has recognized its target membrane and docked there, it unloads its cargo by membrane fusion. Fusion does not always follow immediately, however. As we discuss later, in the process of regulated exocytosis, fusion is delayed until it is triggered by a specific extracellular signal.

Thus, docking and fusion are two distinct and separable processes. Docking requires only that the two membranes come close enough for proteins protruding from the lipid bilayers to interact and adhere. Fusion requires a much closer approach, bringing the lipid bilayers to within 1.5 nm of each other so that they can join. When the membranes are in such close apposition, lipids can flow from one bilayer to the other. For this close approach, water must be displaced from the hydrophilic surface of the membrane—a process that is energetically highly unfavorable. It seems likely that all membrane fusions in cells are catalyzed by specialized fusion proteins that provide a way to overcome this energy barrier. We have already discussed the role of dynamin in a related task during clathrin-coated vesicle budding.

SNAREs are thought to have a central role in membrane fusion. The formation of the SNARE complex may work like a winch, using the energy that is freed when the interacting helices wrap around each other to pull the membrane faces together, while simultaneously squeezing out water molecules from the interface (Figure 13-15). When liposomes containing purified v-SNAREs are mixed with liposomes containing matching t-SNAREs, their membranes fuse, albeit slowly. In the cell, other proteins recruited to the fusion site presumably cooperate with SNAREs to initiate fusion. Moreover, inhibitory proteins may have to be released to allow the complete zipping-up of SNARE pairs. In some cases, such as in regulated exocytosis, a localized influx of Ca^{2+} triggers the fusion process.

Viral Fusion Proteins and SNAREs May Use Similar Strategies

Membrane fusion is important in other processes beside **vesicular** transport. Examples are the fusion of the plasma membranes of sperm and egg that occurs at fertilization (discussed in Chapter 20) and the fusion of myoblasts during muscle cell development . All cell membrane fusions require special proteins and are subject to tight controls, which ensure that only appropriate membranes fuse. The controls are crucial for maintaining both the identity of cells and the individuality of each type of **intracellular** compartment.

The membrane fusions catalyzed by viral fusion proteins are the best understood. These proteins have a crucial role in permitting the entry of enveloped viruses (which have a lipid-bilayer-based membrane coat) into the cells that they. For example, viruses—such as human immunodeficiency virus (HIV), which causes AIDS—bind to cell-surface receptors, and then the viral and plasma membranes fuse. This fusion event allows the viral nucleic acid to enter the cytosol, where it replicates. Other viruses, such as influenza virus, first enter the cell

by receptor-mediated endocytosis (discussed later) and are delivered to endosomes. In this case, the lowpH in endosomes activates a fusion protein in the viral envelope that catalyzes the fusion of the viral and endosomal membranes. This likewise releases the viral nucleic acid into the cytosol.

The three-dimensional structures of the fusion proteins of influenza virus and HIV provide valuable insights into the molecular mechanism of the membrane fusion catalyzed by these proteins. An exposure of the influenza fusion proteinto low pH, or an exposure of HIV fusion protein to receptors on the target cell membrane, uncovers previously buried hydrophobic regions. These regions, called fusion peptides, are observed to then insert directly into the hydrophobic core of lipid bilayer of the target membrane. Thus, the viral fusion proteins are, for a moment, integral membrane proteins in two separate lipid bilayers. Structural rearrangements in the fusion proteins then bring the two lipid bilayers into very close apposition and destabilize them so that the bilayers fuse. For viral fusion, the fusion proteins are the only components required, supporting the possibility that SNAREs are also the central players in the process of bilayer fusion in cells.

Summary

The differences between the many different membrane-enclosed compartments in a eucaryotic cell are maintained by directed, selective transport of particular membrane components from one compartment to another. Transport vesicles, which can be spherical or tubular, bud from specialized coated regions of the donor membrane. The assembly of the coat helps to collect specific membrane and soluble cargo molecules for transport and to drive the formation of thevesicle.

Of the various types of coated vesicles, the best characterized are clathrin-coated vesicles, which mediate transport from the plasma membrane and the trans Golgi network, and COPI- and COPII-coated vesicles, which mediate transport between the ER and the Golgi apparatus and between Golgi cisternae. In clathrin-coated vesicles, adaptins link the clathrin to the vesicle membrane and also trap specific cargo molecules for packaging into the vesicle. The coat is shed rapidly after budding, which is necessary for a vesicle to fuse with its appropriate target membrane.

Monomeric GTPases help regulate various steps in **vesicular** transport, including both vesicle budding and docking. The coat-recruitment GTPases, including Sar1 and the ARF proteins, regulate coat assembly and disassembly. A family of Rab proteins functions as vesicle targeting GTPases. Being incorporated with v-SNAREs into budding transport vesicles, the Rab proteins help ensure that the vesicles deliver their contents only to the appropriate membrane-enclosedcompartment: the one that displays complementary t-SNARE proteins. Complementary v-SNARE and t-SNARE proteins form stable trans-SNARE complexes, thereby bringing their membrane bilayers into close apposition for fusion.

Transport from the ER through the Golgi Apparatus

Newly synthesized proteins enter the biosynthetic- secretory pathway in the ER by crossing the ER membrane from the cytosol. During their subsequent transport, from the ER to the Golgi apparatus and from the Golgi apparatus to the cell surface and elsewhere, these proteins pass through a series of compartments, where they are successively modified. Transfer from one compartment to the next involves a delicate balance between forward and backward (retrieval) transport pathways. Some transport vesicles select cargo molecules and move them to the next compartment in the pathway, while others retrieve escaped proteins and return them to a previous compartment where they normally function. Thus, the pathway from the ER to the cell surface involves many sorting steps, which continually select membrane and soluble lumenal proteins for packaging and transport—in vesicles or organellefragments that bud from the ER and Golgi apparatus.

In this section we focus mainly on the Golgi apparatus (also called the **Golgi complex).** It is a major site of carbohydratesynthesis, as well as a sorting and dispatching station for the products of the ER. Many of the cell's polysaccharides are made in the Golgi apparatus, including the pectin and hemicellulose of the cell wall in plants and most of the glycosaminoglycans of the extracellular matrix in animals. But the Golgi apparatus also lies on the exit route from the ER, and a large proportion of the carbohydrates that it makes are attached as oligosaccharide side chains to the many proteins and lipids that the ER sends to it. A subset of these oligosaccharide groups serve as tags to direct specific proteins into vesicles that then transport them to lysosomes. But most proteins and lipids, once they have acquired their appropriate oligosaccharides in the Golgi apparatus, are recognized in other ways for targeting into the transport vesicles going to other destinations.

Proteins Leave the ER in COPII-coated Transport Vesicles

To initiate their journey along the biosynthetic-secretory pathway, proteins that have entered the ER and are destined for the Golgi apparatus or beyond are first packaged into small COPII-coated transport vesicles. These transport vesicles bud from specialized regions of the ER called *ER exit sites,* whose membrane lacks bound ribosomes. In most animal cells, ER exit sites seem to be randomly dispersed throughout the ER network.

Originally it was thought that all proteins that are not tethered in the ER enter transport vesicles by default. However, it is now clear that packaging into vesicles that leave the ER can also be a selective process. Some cargo proteins are actively recruited into such vesicles, where they become concentrated. It is thought that these cargo proteins display exit (transport) signals on their surface that are recognized by complementary receptor proteins that become trapped in the budding vesicle by interacting with components of the COPII coat. At a much lower rate, proteins without such exit signals can also get packaged in vesicles, so that even proteins that normally function in the ER (so-called *ER resident proteins)* slowly leak out of the ER. Similarly, secretory proteins that are made in high concentrations may leave the ER without the help of sorting receptors.

The exit signals that direct proteins out of the ER for transport to the Golgi and beyond are mostly not understood. There is one exception, however. The ERGIC53 protein seems to serve as a receptor for packaging some secretory proteins into COPII-coated vesicles. Its role in protein transport was identified because humans who lack it owing to an inherited mutation have lowered serum levels of two secreted blood-clotting factors (Factor V and Factor VIII) and therefore bleed excessively. The ERGIC53 protein is a lectin that binds mannose and is thought to recognize this sugaron Factor V and Factor VIII proteins, thereby packaging the proteins into transport vesicles in the ER.

Only Proteins That Are Properly Folded and Assembled Can Leave the ER

To exit from the ER, proteins must be properly folded and, if they are subunits of multimeric protein complexes, they may need to be completely assembled. Those that are misfolded or incompletely assembled are retained in the ER, where they are bound to chaperone proteins (see Chapter 6), such as *BiP* or *calnexin*. The chaperones may cover up the exit signals or somehow anchor the proteins in the ER (Figure 13-18). Such failed proteins are eventually transported back into the cytosol where they are degraded by proteasomes (discussed in Chapter 12). This quality-control step is important, as misfolded or misassembled proteins could potentially interfere with the functions of normal proteins if they were transported onward. The amount of corrective action is surprisingly large. More than 90% of the newly synthesized subunits of the T cell receptor and of the acetylcholine receptor (discussed in Chapter 11), for example, are normally degraded in the cell without ever reaching the cell surface, where they function. Thus, cells must make a large excess of many protein molecules from which to select the few that fold and assemble properly.

Sometimes, however, this quality-control mechanism is detrimental. The predominant mutations that cause cystic fibrosis, a common inherited disease, produce a plasma membrane protein important for Cl^- transport that is only slightly misfolded. Although the mutant protein would function perfectly normally if it reached the plasma membrane, it is retained in the ER. The devastating disease thus results not because the mutation inactivates the protein, but because the active protein is discarded before it reaches the plasma membrane.

Transport from the ER to the Golgi Apparatus Is Mediated by Vesicular Tubular Clusters

After transport vesicles have budded from an ER exit site and have shed their coat, they begin to fuse with one another. This fusion of membranes from the same compartment is called *homotypic fusion,* to distinguish it from *heterotypic fusion,* in which a membrane from one compartment fuses with the membrane of a different compartment. As with heterotypic fusion, homotypic fusion requires a set of matching SNAREs. In this case, however, the interaction is symmetrical, with v-SNAREs and t-SNAREs contributed by both membranes.

The structures formed when ER-derived vesicles fuse with one another are called ***vesicular*** *tubular clusters,* on the basis of their convoluted appearance in the electron

microscope (Figure 13-20A). These clusters constitute a newcompartment that is separate from the ER and lacks many of the proteins that function in the ER. They are generated continually and function as transport packages that bring material from the ER to the Golgi apparatus. The clusters are relatively short-lived because they quickly move along microtubules to the Golgi apparatus, where they fuse and deliver their contents.

As soon as **vesicular** tubular clusters form, they begin budding off vesicles of their own. Unlike the COPII-coated vesicles that bud from the ER, these vesicles are COPI-coated. They carry back to the ER resident proteins that have escaped, as well as proteins that participated in the ER budding reaction and are being returned. This retrieval process demonstrates the exquisite control mechanisms that regulate coat assembly reactions. The COPI coat assembly begins only seconds after the COPII coats have been shed. It remains a mystery how this switchover in coat assembly is controlled.

The *retrieval* (or *retrograde*) *transport* continues as the **vesicular** tubular clusters move to the Golgi apparatus. Thus, the clusters continuously mature, gradually changing their composition as selected proteins are returned to the ER. A similar retrieval process continues from the Golgi apparatus, after the **vesicular** tubular clusters have delivered their cargo.

The Retrieval Pathway to the ER Uses Sorting Signals

The retrieval pathway for returning escaped proteins back to the ER depends on *ER retrieval signals*. Resident ERmembrane proteins, for example, contain signals that bind directly to COPI coats and are thus packaged into COPI-coated transport vesicles for retrograde delivery to the ER. The best-characterized signal of this type consists of two lysines, followed by any two other amino acids, at the extreme C-terminal end of the ER membrane protein. It is called a*KKXX sequence*, based on the single-letter amino acid code.

Soluble ER resident proteins, such as BiP, also contain a short retrieval signal at their C-terminal end, but it is different: it consists of a Lys-Asp-Glu-Leu or similar sequence. If this signal (called the *KDEL sequence*) is removed from BiP by genetic engineering, the protein is slowly secreted from the cell. If the signal is transferred to a protein that is normally secreted, the protein is now efficiently returned to the ER, where it accumulates.

Unlike the retrieval signals on ER membrane proteins that can interact directly with the COPI coat, soluble ER resident proteins must bind to specialized receptor proteins such as the *KDEL receptor*—a multipass transmembrane protein that binds to the KDEL sequence and packages any protein displaying it into COPI-coated retrograde transport vesicles. To accomplish this task, the KDEL receptor itself must cycle between the ER and the Golgi apparatus, and its affinity for the KDEL sequence must be different in these two compartments. The receptor must have a high affinity for the KDEL sequence in **vesicular** tubular clusters and the Golgi apparatus, so as to capture escaped ER resident proteins that are present there at low concentration. It must have a low affinity for the KDEL sequence in the ER, however, to unload its cargo in spite of the very high concentration of KDEL-containing resident proteins in the ER.

How can the affinity of the KDEL receptor change depending on the compartment in which it resides? The answer may be related to the different pH values established in the different compartments, regulated by H^+ pumps in theirmembrane. The KDEL receptor could bind the KDEL sequence under the slightly acidic conditions in **vesicular** tubular clusters and the Golgi compartment but release it at the neutral pH in the ER. As we discuss later, such pH-sensitiveprotein-protein interactions form the basis for many of the sorting steps in the cell.

Most membrane proteins that function at the interface between the ER and Golgi apparatus, including v- and t-SNARES and some cargo receptors, enter the retrieval pathway to the ER. Whereas the recycling of some of these proteins is signal-mediated as just described, for others no specific signal seems to be required. Thus, while retrieval signals increase the efficiency of the retrieval process, some proteins—including some Golgi enzymes—randomly enter budding vesicles destined for the ER and are returned to the ER at a slower rate. Such Golgi enzymes cycle constantly between the ER and the Golgi, but their rate of return to the ER is slow enough for most of the protein to be found in the Golgi apparatus.

Many Proteins are Selectively Retained in the Compartments in which they Function

The KDEL retrieval pathway only partly explains how ER resident proteins are maintained in the ER. As expected, cells that express genetically modified ER resident proteins, from which the KDEL sequence has been experimentally removed, secrete these proteins. But secretion occurs at a much slower rate than for a normal secretory protein. It seems that ER resident proteins are anchored in the ER by a mechanism that is independent of their KDEL signal and that only those proteins that escape retention are captured and returned via the KDEL receptor. A suggested mechanism of retention is that ER resident proteins bind to one another, thus forming complexes that are too big to enter transport vesicles. Because ER resident proteins are present in the ER at very high concentrations (estimated to be millimolar), relatively low-affinity interactions would suffice to have most of the proteins tied up in such complexes.

Aggregation of proteins that function in the same compartment—called *kin recognition*—is a general mechanism that compartments use to organize and retain their resident proteins. Golgi enzymes that function together, for example, also bind to each other and are thereby restrained from entering transport vesicles.

The Length of the Transmembrane Region of Golgi Enzymes Determines their Location in The Cell

Vesicles that leave the Golgi apparatus of animal cells destined for the plasma membrane are rich in cholesterol. The cholesterol fills the space between the kinked hydrocarbon chains of the lipids in the bilayer, forcing them into tighter alignment and increasing the separation between the lipid head groups of the two leaflets of the bilayer. Thus the lipid bilayer of the cholesterol-derived vesicles is thicker than that of the Golgi membrane itself. Transmembrane proteins must have sufficiently long transmembrane segments to span this thickness if they

are to enter the cholesterol-rich transport vesicle budding from the Golgi apparatus destined for the plasma membrane. Proteins with shorter transmembrane segments are excluded.

This exclusion is thought to explain why membrane proteins that normally reside in the Golgi and the ER have shorter transmembrane segments (around 15 amino acids) than do plasma membrane proteins (around 20–25 amino acids). When the transmembrane segments of Golgi proteins are extended by recombinant DNA techniques, the proteins are no longer efficiently retained in the Golgi apparatus and are transported to the plasma membrane instead. Thus, at least some Golgi proteins seem to be retained in the Golgi apparatus mainly because they cannot enter transport vesicles heading for the plasma membrane.

The Golgi Apparatus Consists of an Ordered Series of Compartments

Because of its large and regular structure, the Golgi apparatus was one of the first organelles described by early light microscopists. It consists of a collection of flattened, membrane-enclosed *cisternae,* somewhat resembling a stack of pancakes. Each of these Golgi stacks usually consists of four to six cisternae, although some unicellular flagellates can have up to 60. In animal cells, many stacks are linked by tubular connections between corresponding cisternae, thus forming a single complex, which is usually located near the cell nucleus and close to the centrosome. This localization depends on microtubules. If microtubules are experimentally depolymerized, the Golgi apparatus reorganizes into individual stacks that are found throughout the **cytoplasm**, adjacent to ER exit sites. In some cells, including most plant cells, hundreds of individual Golgi stacks are normally dispersed throughout the **cytoplasm** .

During their passage through the Golgi apparatus, transported molecules undergo an ordered series of covalent modifications. Each Golgi stack has two distinct faces: a ***cis* face** (or entry face) and a ***trans* face** (or exit face). Both *cis*and *trans* faces are closely associated with special compartments, each composed of a network of interconnected tubular and cisternal structures: the ***cis* Golgi network (CGN)** (also called the *intermediate compartment)* and the ***trans* Golgi network (TGN)**, respectively. Proteins and lipids enter the *cis* Golgi network in **vesicular** tubular clusters arriving from the ER and exit from the *trans* Golgi network bound for the cell surface or another compartment. Both networks are thought to be important for protein sorting. As we have seen, proteins entering the CGN can either move onward in the Golgi apparatus or be returned to the ER. Similarly, proteins exiting from the TGN can either move onward and be sorted according to whether they are destined for lysosomes, secretory vesicles, or the cell surface, or be returned to an earlier compartment.

The Golgi apparatus is especially prominent in cells that are specialized for secretion, such as the goblet cells of the intestinal epithelium, which secrete large amounts of polysaccharide-rich mucus into the gut. In such cells, unusually large vesicles are found on the *trans* side of the Golgi apparatus, which faces the plasma membranedomain where secretion occurs.

Oligosaccharide Chains Are Processed in the Golgi Apparatus

As described in Chapter 12, a single species of ***N* -linked oligosaccharide** is attached *en bloc* to many proteins in theER and then trimmed while the protein is still in the ER. Further modifications and additions occur in the Golgi apparatus, depending on the protein. The outcome is that two broad classes of *N*-linked oligosaccharides, the *complexoligosaccharides* and the *high-mannose oligosaccharides*, are found attached to mammalian glycoproteins (Figure 13-25). Sometimes both types are attached (in different places) to the same polypeptide chain.

Complex oligosaccharides are generated by a combination of trimming the original *N*-linked oligosaccharide added in the ER and the addition of further sugars. By contrast, high-mannose oligosaccharides have no new sugars added to them in the Golgi apparatus. They contain just two *N*-acetylglucosamines and many mannose residues, often approaching the number originally present in the lipid-linked oligosaccharide precursor added in the ER. Complex oligosaccharides can contain more than the original two *N*-acetylglucosamines as well as a variable number of galactose and sialic acid residues and, in some cases, fucose. Sialic acid is of special importance because it is the only sugar in glycoproteins that bears a negative charge. Whether a given oligosaccharide remains high-mannose or is processed isdetermined largely by its position on the protein. If the oligosaccharide is accessible to the processing enzymes in the Golgi apparatus, it is likely to be converted to a complex form; if it is inaccessible because its sugars are tightly held to the protein's surface, it is likely to remain in a high-mannose form. The processing that generates complex oligosaccharide chains follows the highly ordered pathway shown.

Proteoglycans Are Assembled in the Golgi Apparatus

It is not only the *N*-linked oligosaccharide chains on proteins that are altered as the proteins pass through the Golgi cisternae *en route* from the ER to their final destinations; many proteins are also modified in other ways. Some proteins have sugars added to the OH groups of selected serine or threonine side chains. This ***O* -linked glycosylation**, like the extension of *N*-linked oligosaccharide chains, is catalyzed by a series of glycosyl transferase enzymes that use the sugarnucleotides in the lumen of the Golgi apparatus to add sugar residues to a protein one at a time. Usually, *N*-acetylgalactosamine is added first, followed by a variable number of additional sugar residues, ranging from just a few to 10 or more.

The Golgi apparatus confers the heaviest glycosylation of all on *proteoglycan core proteins*, which it modifies to produce proteoglycans. As discussed in Chapter 19, this process involves the polymerization of one or more*glycosaminoglycan chains* (long unbranched polymers composed of repeating disaccharide units) via a xylose link onto serines on the core protein. Many proteoglycans are secreted and become components of the extracellular matrix, while others remain anchored to the plasma membrane. Still others form a major component of slimy materials, such as the mucus that is secreted to form a protective coating over many epithelia.

The sugars incorporated into glycosaminoglycans are heavily sulfated in the Golgi apparatus immediately after these polymers are made, thus adding a significant portion of their characteristically large negative charge. Some tyrosine residues in proteins also become sulfated shortly before they exit from the Golgi apparatus. In both cases, the sulfation depends on the sulfate donor 3′-phosphoadenosine-5′-phosphosulfate, or PAPS, that is transported from the cytosol into the lumen of the *trans* Golgi network.

What Is the Purpose of Glycosylation?

There is an important difference between the construction of an oligosaccharide and the synthesis of other macromolecules such as DNA, RNA, and protein. Whereas nucleic acids and proteins are copied from a template in a repeated series of identical steps using the same enzyme or set of enzymes, complex carbohydrates require a different enzyme at each step, each product being recognized as the exclusive substrate for the next enzyme in the series. Given the complicated pathways that have evolved to synthesize them, it seems likely that the oligosaccharides on glycoproteins and glycosphingolipids have important functions, but for the most part these functions are not known.

N-linked glycosylation, for example, is prevalent in all eucaryotes, including yeasts, but is absent from procaryotes. Because one or more *N*-linked oligosaccharides are present on most proteins transported through the ER and Golgi apparatus—a pathway that is unique to eucaryotic cells—one might suspect that they function to aid folding and the transport process. We have already discussed a number of instances for which this is so—the use of a carbohydrate as a marker during protein folding in the ER, for example, and the use of carbohydrate-binding lectins in guiding ER-to-Golgi transport. As we discuss later, lectins also participate in protein sorting in the *trans* Golgi network.

Because chains of sugars have limited flexibility, even a small *N*-linked oligosaccharide protrudes from the surface of aglycoprotein and can thus limit the approach of other macromolecules to the protein surface. In this way, for example, the presence of oligosaccharides tends to make a glycoprotein more resistant to digestion by proteases. It may be that the oligosaccharides on cell-surface proteins originally provided an ancestral eucaryotic cell with a protective coat that, unlike the rigid bacterial cell wall, left the cell with the freedom to change shape and move. But if so, these sugar chains have since become modified to serve other purposes as well. The oligosaccharides attached to some cell-surface proteins, for example, are recognized by transmembrane lectins called *selectins*, which function in cell-cell adhesion processes.

Glycosylation can also have important regulatory roles. Signaling through the cell-surface signaling receptor Notch, for example, is important for proper cell fate determination in development. Notch is a transmembrane protein that is *O*-glycosylated by addition of a single fucose to some serines, threonines, and hydroxylysines. Some cell types express an additional glycosyltransferase that adds an *N*-acetylglucosamine to each of these fucoses in the Golgi apparatus. This addition sensitizes the Notch receptor, and thus allows these cells to

respond selectively to activating stimuli. In this way, glycosylation has become important to the establishment of spatial boundaries in developing tissues.

The Golgi Cisternae Are Organized as a Series of Processing Compartments

Proteins exported from the ER enter the first of the Golgi processing compartments (the *cis* Golgi compartment), after having passed through the *cis* Golgi network. They then move to the next compartment (the *medial* compartment, consisting of the central cisternae of the stack) and finally to the *trans* compartment, where glycosylation is completed. The lumen of the *trans* compartment is thought to be continuous with the *trans* Golgi network, where proteins are segregated into different transport packages and dispatched to their final destinations—the plasma membrane, lysosomes, or secretory vesicles.

The oligosaccharide processing steps occur in a correspondingly organized sequence in the Golgi stack, with each cisterna containing a characteristic abundance of processing enzymes. Proteins are modified in successive stages as they move from cisterna to cisterna across the stack, so that the stack forms a multistage processing unit. This compartmentalization might seem unnecessary, since each oligosaccharide processing enzyme can accept a glycoproteinas a substrate only after it has been properly processed by the preceding enzyme. Nonetheless, it is clear that processing occurs in a spatial as well as a biochemical sequence: enzymes catalyzing early processing steps are concentrated in the cisternae toward the *cis* face of the Golgi stack, whereas enzymes catalyzing later processing steps are concentrated in the cisternae toward the *trans* face.

The functional differences between the *cis*, *medial*, and *trans* subdivisions of the Golgi apparatus were discovered by localizing the enzymes involved in processing *N*-linked oligosaccharides in distinct regions of the organelle, both by physical fractionation of the organelle and by labeling the enzymes in electron microscope sections with antibodies. The removal of mannose residues and the addition of *N*-acetylglucosamine, for example, were shown to occur in the *medial*compartment, while the addition of galactose and sialic acid was found to occur in the *trans* compartment and the *trans*Golgi network.

The functional and structural divisions of the Golgi stack pose two important questions. How are molecules transported from one Golgi cisterna to the next, and how are Golgi resident proteins retained in their appropriate places?

Transport Through the Golgi Apparatus May Occur by Vesicular Transport or Cisternal Maturation

It is still uncertain how the Golgi apparatus achieves and maintains its polarized structure and how molecules move from one cisterna to another. Functional evidence from *in vitro* transport assays and the finding of abundant transport vesicles in the vicinity of Golgi cisternae initially led to the view that these vesicles transport proteins between the cisternae, budding from one cisterna and fusing with the next. According to this **vesicular transport model**, the Golgi apparatus is a relatively static structure, with its

enzymes held in place, while the molecules in transit are moved through the cisternae in sequence, carried by transport vesicles. Retrograde flow retrieves escaped ER and Golgi proteins and returns them to preceding compartments. Directional flow is achieved as forward-moving cargo molecules are selectively packaged into forward-moving vesicles, whereas proteins to be retrieved are selectively packaged into retrograde vesicles. Although both types of vesicles are likely to be COPI-coated, the coats may contain different adaptor proteins to confer selectivity on the packaging of cargo molecules. Alternatively, transport vesicles that shuttle between Golgi cisternae may not be directional at all, transporting cargo material randomly back and forth; directional flow would then occur because of the continual input at the *cis* cisterna and output at the *trans* cisterna. In either case, the movement of vesicles from each cisterna to an adjacent one is helped by a neat trick: the budding vesicles remain tethered by filamentous proteins that restrict their movement, so that their fusion with the correct target membrane is facilitated.

According to an alternative hypothesis, called the **cisternal maturation model**, the Golgi is viewed as a dynamic structure in which the cisternae themselves move through the Golgi stack. The **vesicular** tubular clusters that arrive from the ER fuse with one another to become a *cis* Golgi network, and this network then progressively matures to become a*cis* cisterna, then a *medial* cisterna, and so on. Thus, at the *cis* face of a Golgi stack, new *cis* cisternae would continually form and then migrate through the stack as they mature. This model is supported by microscopic observations demonstrating that large structures such as collagen rods in fibroblasts and scales in certain algae—which are much too large to fit into classical transport vesicles—move progressively through the Golgi stack.

In the maturation model, the characteristic distribution of Golgi enzymes is explained by retrograde flow. Everything moves continuously forward with the maturing cisterna, including the processing enzymes that belong in the early Golgi apparatus. But budding COPI-coated vesicles continually collect the appropriate enzymes, almost all of which aremembrane proteins, and carry them back to the earlier cisterna where they function. A newly formed *cis* cisterna would therefore receive its normal complement of resident enzymes primarily from the cisterna just ahead of it and would later pass them back to the next *cis* cisterna that forms.

As we discuss later, when a cisterna finally moves up to become part of the *trans* Golgi network, various types of coated vesicles bud off of it until this network disappears, to be replaced by a maturing cisterna just behind. At the same time, other transport vesicles are continually retrieving membrane from post-Golgi compartments and returning this membrane to the *trans* Golgi network.

The **vesicular** transport and the cisternal maturation model are not mutually exclusive. Indeed, evidence suggests that transport may occur by a combination of the two mechanisms, in which some cargo is moved forward rapidly in transport vesicles, whereas other cargo is moved forward more slowly as the Golgi apparatus constantly renews itself through cisternal maturation.

Matrix Proteins Form a Dynamic Scaffold That Helps Organize the Apparatus

The unique architecture of the Golgi apparatus depends on both the microtubule cytoskeleton, as already discussed, and cytoplasmic Golgi matrix proteins, which form a scaffold between adjacent cisternae and give the Golgi stack its structural integrity. Some of the matrix proteins form long, filamentous tethers that are thought to help retain Golgi transport vesicles close to the organelle. When the cell prepares to divide, mitotic protein kinases phosphorylate the Golgi matrix proteins, causing the Golgi apparatus to fragment and disperse throughout the cytosol. During disassembly, Golgi enzymes are returned in vesicles to the ER, while other Golgi fragments are distributed to the two daughter cells. There, the matrix proteins are dephosphorylated, leading to the reassembly of the Golgi apparatus.

Remarkably, the Golgi matrix proteins can assemble into appropriately localized stacks near the centrosome even when Golgi membrane proteins are experimentally prevented from leaving the ER. This observation suggests that the matrix proteins are largely responsible for both the structure and location of the Golgi apparatus.

Summary

Correctly folded and assembled proteins in the ER are packaged into COPII-coated transport vesicles that pinch off from the ER membrane. Shortly thereafter the coat is shed and the vesicles fuse with one another to form **vesicular**tubular clusters, which move on microtubule tracks to the Golgi apparatus. Many resident ER proteins slowly escape, but they are returned to the ER from the **vesicular** tubular clusters and the Golgi apparatus by retrograde transport in COPI-coated vesicles.

The Golgi apparatus, unlike the ER, contains many sugar nucleotides, which are used by a variety of glycosyl transferase enzymes to perform glycosylation reactions on lipid and protein molecules as they pass through the Golgi apparatus. The N-linked oligosaccharides that are added to proteins in the ER are often initially trimmed by the removal of mannoses, and further sugars are added. Moreover, the Golgi is the site where *O*-linked glycosylation occurs and where glycosaminoglycan chains are added to core proteins to form proteoglycans. Sulfation of the sugars in proteoglycans and of selected tyrosines on proteins also occurs in a late Golgi compartment.

The Golgi apparatus distributes the many proteins and lipids that it receives from the ER and then modifies the plasma membrane, lysosomes, and secretory vesicles. It is a polarized structure consisting of one or more stacks of disc-shaped cisternae, each stack organized as a series of at least three functionally distinct compartments, termed cis, medial, and trans cisternae. The cis and trans cisternae are both connected to special sorting stations, called the cis Golgi networkand the trans Golgi network, respectively. Proteins and lipids move through the Golgi stack in the cis-to-trans direction. This movement may occur by **vesicular** transport, by progressive maturation of the cis cisternae that migrate continuously through the stack, or by a combination of these two mechanisms. The enzymes that function in each particular region of the stack are thought to be kept there by continual

retrograde **vesicular** transport from more distal cisternae. The finished new proteins end up in the trans Golgi network, which packages them in transport vesicles and dispatches them to their specific destinations in the cell.

Transport into the Cell from the Plasma Membrane: Endocytosis

The routes that lead inward from the cell surface to lysosomes start with the process of endocytosis, by which cells take up macromolecules, particulate substances, and, in specialized cases, even other cells. In this process, the material to be ingested is progressively enclosed by a small portion of the plasma membrane, which first invaginates and then pinches off to form an *endocytic vesicle* containing the ingested substance or particle. Two main types of endocytosis are distinguished on the basis of the size of the endocytic vesicles formed. One type is called *phagocytosis* ("cellular eating"), which involves the ingestion of large particles, such as microorganisms or dead cells via large vesicles called*phagosomes* (generally >250 nm in diameter). The other type is *pinocytosis* ("cellular drinking"), which involves the ingestion of fluid and solutes via small pinocytic vesicles (about 100 nm in diameter). Most eucaryotic cells are continually ingesting fluid and solutes by pinocytosis; large particles are most efficiently ingested by specialized phagocytic cells.

Specialized Phagocytic Cells Can Ingest Large Particles

Phagocytosis is a special form of endocytosis in which large particles such as microorganisms and dead cells are ingested via large endocytic vesicles called phagosomes. In protozoa, phagocytosis is a form of feeding: large particles taken up into phagosomes end up in lysosomes, and the products of the subsequent digestive processes pass into thecytosol to be utilized as food. However, few cells in multicellular organisms are able to ingest such large particles efficiently. In the gut of animals, for example, the particles of food are broken down extracellularly and their hydrolysis products are imported into cells.

Phagocytosis is important in most animals for purposes other than nutrition, and it is mainly carried out by specialized cells—so-called *professional phagocytes.* In mammals, three classes of white blood cells act as professional phagocytes—macrophages, neutrophils, and dendritic cells. These cells all develop from hemopoietic stem cells (discussed in Chapter 22), and they defend us against infection by ingesting invading microorganisms. Macrophages also have an important role in scavenging senescent cells and cells that have died by apoptosis. In quantitative terms, the latter function is by far the most important: our macrophages phagocytose more than 10^{11}senescent red blood cells in each of us every day, for example.

Whereas the endocytic vesicles involved in pinocytosis are small and uniform, phagosomes have diameters that aredetermined by the size of the ingested particle, and they can be almost as large as the phagocytic cell itself. The phagosomes fuse with lysosomes inside the cell, and the ingested material is then degraded. Any indigestible substances will remain in lysosomes, forming *residual bodies.* Some of the

internalized plasma membrane components never reach the lysosome, because they are retrieved from the phagosome in transport vesicles and returned to the plasma membrane.

To be phagocytosed, particles must first bind to the surface of the phagocyte. However, not all particles that bind are ingested. Phagocytes have a variety of specialized surface receptors that are functionally linked to the phagocytic machinery of the cell. Unlike pinocytosis, which is a constitutive process that occurs continuously, phagocytosis is a triggered process, requiring that receptors be activated that transmit signals to the cell interior and initiate the response. The best-characterized triggers are antibodies, which protect us by binding to the surface of infectious microorganisms to form a coat in which the tail region of each antibody molecule, called the Fc region, is exposed on the exterior (discussed in Chapter 24). This antibody coat is recognized by specific *Fc receptors* on the surface of macrophages and neutrophils, whose binding induces the phagocytic cell to extend pseudopods that engulf the particle and fuse at their tips to form a phagosome.

Several other classes of receptors that promote phagocytosis have been characterized. Some recognize *complement*components, which collaborate with antibodies in targeting microbes for destruction (discussed in Chapter 25). Others directly recognize oligosaccharides on the surface of certain microorganisms. Still others recognize cells that have died by apoptosis. Apoptotic cells lose the asymmetric distribution of phospholipids in their plasma membrane. As a consequence, negatively charged phosphatidylserine, which is normally confined to the cytosolic leaflet of the lipid bilayer, is now exposed on the outside of the cell, where it triggers the phagocytosis of the dead cell.

Remarkably, macrophages will also phagocytose a variety of inanimate particles—such as glass, latex beads, or asbestos fibers—yet they do not phagocytose live animal cells. It seems that living animal cells display “don't-eat-me” signals in the form of cell-surface proteins that bind to inhibiting receptors on the surface of macrophages. The inhibitory receptors recruit tyrosine phosphatases that antagonize the **intracellular** signaling events required to initiate phagocytosis, thereby locally inhibiting the phagocytic process. Thus phagocytosis, like many other cell processes, depends on a balance between positive signals that activate the process and negative signals that inhibit it.

Pinocytic Vesicles Form from Coated Pits in the Plasma Membrane

Virtually all eucaryotic cells continually ingest bits of their plasma membrane in the form of small pinocytic (endocytic) vesicles, which are later returned to the cell surface. The rate at which plasma membrane is internalized in this process ofpinocytosis varies between cell types, but it is usually surprisingly large. A macrophage, for example, ingests 25% of its own volume of fluid each hour. This means that it must ingest 3% of its plasma membrane each minute, or 100% in about half an hour. Fibroblasts endocytose at a somewhat lower rate (1% per minute), whereas some amoebae ingest their plasma membrane even more rapidly. Since a cell's surface area and volume remain unchanged during this process, it is clear that the same amount of membrane that is being removed by endocytosis is being added to the cell

surface by*exocytosis*, the converse process, as we discuss later. In this sense, endocytosis and exocytosis are linked processes that can be considered to constitute an *endocytic-exocytic cycle*.

The endocytic part of the cycle often begins at clathrin-coated pits. These specialized regions typically occupy about 2% of the total plasma membrane area. The lifetime of a clathrin-coated pit is short: within a minute or so of being formed, it invaginates into the cell and pinches off to form a clathrin-coated vesicle (Figure 13-41). It has been estimated that about 2500 clathrin-coated vesicles leave the plasma membrane of a cultured fibroblast every minute. The coated vesicles are even more transient than the coated pits: within seconds of being formed, they shed their coat and are able to fuse with early endosomes. Since extracellular fluid is trapped in clathrin-coated pits as they invaginate to form coated vesicles, any substance dissolved in the extracellular fluid is internalized—a process called *fluid-phase endocytosis*.

Not All Pinocytic Vesicles Are Clathrin-coated

In addition to clathrin-coated pits and vesicles, there are other, less well-understood mechanisms by which cells can form pinocytic vesicles. One of these pathways initiates at **caveolae** (from the Latin for "little cavities"), originally recognized by their ability to transport molecules across endothelial cells, which form the inner lining of blood vessels. Caveolae are present in the plasma membrane of most cell types, and in some of these they are seen as deeply invaginated flasks in the electron microscope . They are thought to form from *lipid rafts*, which are patches of the plasma membrane that are especially rich in cholesterol, glycosphingolipids, and GPI-anchored membrane proteins. The major structural protein in caveolae is **caveolin**, a multipass integral membrane protein that is a member of a heterogeneous protein family.

In contrast to clathrin-coated and COPI- or COPII-coated vesicles, caveolae are thought to invaginate and collect cargo proteins by virtue of the lipid composition of the calveolar membrane, rather than by the assembly of a cytosolic proteincoat. Caveolae pinch off from the plasma membrane and can deliver their contents either to endosome-like compartments or (in a process called transcytosis, which is discussed later) to the plasma membrane on the opposite side of a polarized cell. Some animal viruses also enter cells in vesicles derived from caveolae. The viruses are first delivered to an endosome-like compartment, from where they are moved to the ER. In the ER, they extrude their genome into thecytosol to start their infectious cycle. It remains a mystery how material endocytosed in caveolae-derived vesicles can end up in so many different locations in the cell.

Cells Import Selected Extracellular Macromolecules by Receptor-mediated Endocytosis

In most animal cells, clathrin-coated pits and vesicles provide an efficient pathway for taking up specific macromolecules from the extracellular fluid. In this process, called receptor-

mediated endocytosis, the macromolecules bind to complementary transmembrane receptor proteins, accumulate in coated pits, and then enter the cell as receptor-macromolecule complexes in clathrin-coated vesicles. Receptor-mediated endocytosis provides a selective concentrating mechanism that increases the efficiency of internalization of particular ligands more than a hundredfold, so that even minor components of the extracellular fluid can be specifically taken up in large amounts without taking in a correspondingly large volume of extracellular fluid. A particularly well-understood and physiologically important example is the process whereby mammalian cells take up cholesterol.

Many animals cells take up cholesterol through receptor-mediated endocytosis and, in this way, acquire most of the cholesterol they require to make new membrane. If the uptake is blocked, cholesterol accumulates in the blood and can contribute to the formation in blood vessel walls of *atherosclerotic plaques*, deposits of lipid and fibrous tissue that can cause strokes and heart attacks by blocking blood flow. In fact, it was through a study of humans with a strong genetic predisposition for *atherosclerosis* that the mechanism of receptor-mediated endocytosis was first clearly revealed.

Most cholesterol is transported in the blood as cholesteryl esters in the form of lipid-protein particles known as **low-density lipoproteins (LDL)**. When a cell needs cholesterol for membrane synthesis, it makes transmembrane receptor proteins for LDL and inserts them into its plasma membrane. Once in the plasma membrane, the LDL receptors diffuse until they associate with clathrin-coated pits that are in the process of forming. Since coated pits constantly pinch off to form coated vesicles, any LDL particles bound to LDL receptors in the coated pits are rapidly internalized in coated vesicles. After shedding their clathrin coats, the vesicles deliver their contents to early endosomes, which are located near the cell periphery. Once the LDL and LDL receptors encounter the low pH in the endosomes, LDL is released from its receptor and is delivered via late endosomes to lysosomes. There the cholesteryl esters in the LDL particles are hydrolyzed to free cholesterol, which is now available to the cell for new membrane synthesis. If too much free cholesterol accumulates in a cell, the cell shuts off both its own cholesterol synthesis and the synthesis of LDL receptor proteins, so that it ceases either to make or to take up cholesterol.

This regulated pathway for the uptake of cholesterol is disrupted in individuals who inherit defective genes encodingLDL receptor proteins. The resulting high levels of blood cholesterol predispose these individuals to develop atherosclerosis prematurely, and many die at an early age of heart attacks resulting from coronary artery disease. In some cases, the receptor is lacking altogether. In others, the receptors are defective—in either the extracellular binding site for LDL or the **intracellular** binding site that attaches the receptor to the coat of a clathrin-coated pit . In the latter case, normal numbers of LDL-binding receptor proteins are present, but they fail to become localized in the clathrin-coated regions of the plasma membrane. Although LDL binds to the surface of these mutant cells, it is not

internalized, directly demonstrating the importance of clathrin-coated pits in the receptor-mediated endocytosis of cholesterol.

More than 25 different receptors are known to participate in receptor-mediated endocytosis of different types of molecules, and they all apparently use the same clathrin-coated-pit pathway. Many of these receptors, like the LDLreceptor, enter coated pits irrespective of whether they have bound their specific ligands. Others enter preferentially when bound to a specific ligand, suggesting that a ligand-induced conformational change is required for them to activate the signal sequence that guides them into the pits. Since most plasma membrane proteins fail to become concentrated in clathrin-coated pits, the pits must function as molecular filters, preferentially collecting certain plasma membrane proteins (receptors) over others.

Signal peptides guide transmembrane proteins into clathrin-coated pits by binding to the adaptins. Despite a common function, their amino acid sequences vary. A common endocytosis signal consists of only four amino acids Y-X-X-Ψ, where Y is tyrosine, X any polar amino acid, and Ψ a hydrophobic amino acid. This short peptide, which is shared by many receptors, binds directly to one of the adaptins in clathrin-coated pits. By contrast, the cytosolic tail of the LDLreceptor contains a unique signal (Asn-Pro-Val-Tyr) that apparently binds to the same adaptin protein.

Electron-microscope studies of cultured cells exposed simultaneously to different labeled ligands demonstrate that many kinds of receptors can cluster in the same coated pit. The plasma membrane of one clathrin-coated pit can probably accommodate up to 1000 receptors of assorted varieties. Although all of the receptor-ligand complexes that use this endocytic pathway are apparently delivered to the same endosomal compartment, the subsequent fates of the endocytosed molecules vary, as we discuss next.

Endocytosed Materials That Are Not Retrieved From Endosomes End Up in Lysosomes

The endosomal compartments of a cell can be complex. They can be made visible in the electron microscope by adding a readily detectable tracer molecule, such as the enzyme peroxidase, to the extracellular medium and leaving the cells for various lengths of time to take it up by endocytosis. The distribution of the molecule after its uptake reveals the endosomal compartments as a set of heterogeneous, membrane-enclosed tubes extending from the periphery of the cell to the perinuclear region, where it is often close to the Golgi apparatus. Two sequential sets of endosomes can be readily distinguished in such labeling experiments. The tracer molecule appears within a minute or so in early endosomes, just beneath the plasma membrane. After 5–15 minutes, it moves to late endosomes, close to the Golgi apparatus and near the nucleus. Early and late endosomes differ in their protein compositions; they are associated with different Rab proteins, for example.

As mentioned earlier, the interior of the endosomal compartment is kept acidic (pH ~6) by a vacuolar H^+ ATPase in the endosomal membrane that pumps H^+ into the lumen from

the cytosol. In general, later endosomes are more acidic than early endosomes. This acidic environment has a crucial role in the function of these organelles.

We have already seen how endocytosed materials that reach the late endosomes become mixed with newly synthesizedacid hydrolases and end up being degraded in lysosomes. Many molecules, however, are specifically diverted from this journey to destruction. They are recycled instead from the early endosomes back to the plasma membrane via transport vesicles. Only molecules that are not retrieved from endosomes in this way are delivered to lysosomes for degradation.

Specific Proteins Are Removed from Early Endosomes and Returned to the Plasma Membrane

The **early endosomes** form a compartment that acts as the main sorting station in the endocytic pathway, just as the *cis*and *trans* Golgi networks serve this function in the biosynthetic-secretory pathway. In the acidic environment of the early endosome, many internalized receptor proteins change their conformation and release their ligand, just as the M6Preceptors unload their cargo of acid hydrolases in the even more acidic late endosomes. Those endocytosed ligands that dissociate from their receptors in the early endosome are usually doomed to destruction in lysosomes, along with the other soluble contents of the endosome. Some other endocytosed ligands, however, remain bound to their receptors, and thereby share the fate of the receptors.

The fates of the receptor proteins—and of any ligands remaining bound to them—vary according to the specific type of receptor. (1) Most receptors are recycled and return to the same plasma membrane domain from which they came; (2) some proceed to a different domain of the plasma membrane, thereby mediating a process called *transcytosis*; and (3) some progress to lysosomes, where they are degraded.

The LDL receptor follows the first pathway. It dissociates from its ligand LDL in the early endosome and is recycled to the plasma membrane for reuse, leaving the discharged LDL to be carried to lysosomes. The recycling vesicles bud from long, narrow tubules that extend from the early endosomes. It is likely that the geometry of these tubules helps the sorting process. Because tubules have a large membrane area enclosing a small volume, membrane proteins tend to accumulate there. Transport vesicles that return material to the plasma membrane begin budding from the tubules, but tubular portions of the early endosome also pinch off and fuse with one another to form *recycling endosomes*, a way-station for the **traffic** between the early endosomes and the plasma membrane. During this process, the tubules and then the recycling endosome continuously shed vesicles that return to the plasma membrane.

The **transferrin receptor** follows a similar recycling pathway, but it also recycles its ligand. Transferrin is a solubleprotein that carries iron in the blood. Cell-surface transferrin receptors deliver transferrin with its bound iron to early endosomes by receptor-mediated endocytosis. The low pH in the endosome induces transferrin to release its bound iron, but the iron-free

transferrin itself (called apotransferrin) remains bound to its receptor. The receptor-apotransferrincomplex enters the tubular extensions of the early endosome and from there is recycled back to the plasma membrane. When the apotransferrin returns to the neutral pH of the extracellular fluid, it dissociates from the receptor and is thereby freed to pick up more iron and begin the cycle again. Thus, transferrin shuttles back and forth between the extracellular fluid and the endosomal compartment, avoiding lysosomes and delivering the iron that cells need to grow to the cell interior.

The second pathway that endocytosed receptors can follow from endosomes is taken both by opioid receptors (seeFigure 13-47) and by the receptor that binds *epidermal growth factor (EGF).* EGF is a small, extracellular signal proteinthat stimulates epidermal and various other cells to divide. Unlike LDL receptors, EGF receptors accumulate in clathrin-coated pits only after binding EGF, and most of them do not recycle but are degraded in lysosomes, along with the ingested EGF. EGF binding therefore first activates **intracellular** signaling pathways and then leads to a decrease in the concentration of EGF receptors on the cell surface, a process called *receptor down-regulation* that reduces the cell's subsequent sensitivity to EGF (discussed in Chapter 15).

Multivesicular Bodies Form on the Pathway to Late Endosomes

It is still uncertain how endocytosed molecules move from the early to the late endosomal compartment so as to end up in lysosomes. A current view is that portions of the early endosomes migrate slowly along microtubules toward the cell interior, shedding tubules of material to be recycled to the plasma membrane. The migrating endosomes enclose large amounts of invaginated membrane and internally pinched-off vesicles and are therefore called **multivesicular bodies**. It is unknown whether multivesicular bodies eventually fuse with a late endosomal compartment or if they fuse instead with each other to become late endosomes. At the end of this pathway, the late endosomes are converted to lysosomes as a result of their fusion with hydrolase-bearing transport vesicles from the *trans* Golgi network and their increased acidification.

The multivesicular bodies carry specific endocytosed membrane proteins that are to be degraded but exclude others that are to be recycled. As part of the protein-sorting process, specific proteins—for example, the occupied EGF receptordescribed previously—selectively partition to the invaginating membrane of the multivesicular bodies (Figure 13-50). In this way, the receptors, as well as any signaling proteins strongly bound to them, are rendered fully accessible to the digestive enzymes that will degrade them (see Figure 13-50).

Membrane proteins that are sorted into the internal membrane vesicles of a multivesicular body are first covalently modified with the small protein ubiquitin. Unlike multi-ubiquitylation which typically targets substrate proteins for degradation in proteasomes (discussed in Chapter 6), ubiquitin tagging for sorting into the internal membrane vesicles of a multivesicular body requires the addition of only a single ubiquitin molecule that is added to activated receptors while still at the plasma membrane. The ubiquitin tag facilitates the

uptake of the receptors into endocytic vesicles and is then recognized again by proteins that mediate the sorting process into the internal membrane vesicles of multivesicular bodies. In addition, membrane invagination in multivesicular bodies is regulated by a lipid kinase that phosphorylatesphosphatidylinositol. The phosphorylated head groups of these lipids are thought to serve as docking sites for the proteins that mediate the invagination process. Local modification of lipid molecules is thus another way in which specific membrane patches can be induced to change shape and destiny.

In addition to endocytosed membrane proteins, multivesicular bodies also contain most of the soluble content of early endosomes destined for digestion in lysosomes.

Macromolecules Can Be Transferred Across Epithelial Cell Sheets by Transcytosis

Some receptors on the surface of polarized epithelial cells transfer specific macromolecules from one extracellular space to another by transcytosis . These receptors are endocytosed and then follow a pathway from endosomes to a different plasma membrane. A newborn rat, for example, obtains antibodies from its mother's milk (which help protect it against infection) by transporting them across the epithelium of its gut. The lumen of the gut is acidic, and, at this low pH, the antibodies in the milk bind to specific receptors on the apical (absorptive) surface of the gut epithelial cells. The receptor-antibody complexes are internalized via clathrin-coated pits and vesicles and are delivered to early endosomes. The complexes remain intact and are retrieved in transport vesicles that bud from the early endosome and subsequently fuse with the basolateral domain of the plasma membrane. On exposure to the neutral pH of the extracellular fluid that bathes the basolateral surface of the cells, the antibodies dissociate from their receptors and eventually enter the newborn's bloodstream.

The transcytotic pathway from the early endosome to the plasma membrane is not direct. The receptors first move from the early endosome to an intermediate endosomal compartment, the **recycling endosome** described previously. The variety of pathways that different receptors follow from endosomes implies that, in addition to binding sites for their ligands and binding sites for coated pits, many receptors also possess sorting signals that guide them into the appropriate type of transport vesicle leaving the endosome and thereby to the appropriate target membrane in the cell.

A unique property of a recycling endosomes is that the exit of membrane proteins from the compartment can be regulated. Thus, cells can adjust the flux of proteins through the transcytotic pathway according to need. Although the mechanism of regulation is uncertain, it allows recycling endosomes an important role in adjusting the concentration of specific plasma membrane proteins. Fat cells and muscle cells, for example, contain large **intracellular** pools of theglucose transporters that are responsible for the uptake of glucose across the plasma membrane. These proteins are stored in specialized recycling endosomes until the cell is stimulated by the hormone insulin to increase its rate of glucose uptake. Then transport vesicles bud from the recycling endosome and deliver large numbers

of glucose transporters to the plasma membrane, thereby greatly increasing the rate of glucose uptake into the cell.

Epithelial Cells Have Two Distinct Early Endosomal Compartments But a Common Late Endosomal Compartment

In polarized epithelial cells, endocytosis occurs from both the *basolateral domain* and the *apical domain* of the plasma membrane. Material endocytosed from either domain first enters an early endosomal compartment that is unique to that domain. This arrangement allows endocytosed receptors to be recycled back to their original membrane domain, unless they contain signals that mark them for transcytosis to the other domain. Molecules endocytosed from either plasma membrane domain that are not retrieved from the early endosomes end up in a common late endosomal compartment near the cell center and are eventually degraded in lysosomes.

Whether cells contain a few connected or many unconnected endosomal compartments seems to depend on the cell type and the physiological state of the cell. Like many other membrane-enclosed organelles, endosomes of the same type can readily fuse with one another (an example of homotypic fusion, discussed earlier) to create large continuous endosomes.

Summary

Cells ingest fluid, molecules, and particles by endocytosis, in which localized regions of the plasma membraneinvaginate and pinch off to form endocytic vesicles. Many of the endocytosed molecules and particles end up in lysosomes, where they are degraded. Endocytosis occurs both constitutively and as a triggered response to extracellular signals. Endocytosis is so extensive in many cells that a large fraction of the plasma membrane is internalized every hour. To make this possible, most of the plasma membrane components (proteins and lipid) that are endocytosed are continually returned to the cell surface by exocytosis. This large-scale endocytic-exocytic cycle is mediated largely byclathrin-coated pits and vesicles.

Many cell-surface receptors that bind specific extracellular macromolecules become localized in clathrin-coated pits. As a result, they and their ligands are efficiently internalized in clathrin-coated vesicles, a process called receptor-mediated endocytosis. The coated endocytic vesicles rapidly shed their clathrin coats and fuse with early endosomes.

Most of the ligands dissociate from their receptors in the acidic environment of the endosome and eventually end up in lysosomes, while most of the receptors are recycled via transport vesicles back to the cell surface for reuse. But receptor-ligand complexes can follow other pathways from the endosomal compartment. In some cases, both the receptor and the ligand end up being degraded in lysosomes, resulting in receptor down-regulation. In other cases, both are transferred to a different plasma membrane domain, and the ligand is thereby released by exocytosis at a surface of the cell different from that where it originated,

a process called transcytosis. The transcytosis pathway includes recycling endosomes, where endocytosed plasma membrane proteins can be stored until they are needed.

Tissue Replacement

A Natural Phenomenon

Bones regenerate every 10 years
Skin regenerates every two weeks
Macrophages scavenge senescent RBCs (10^{11}/day)
Tissue repair in response to minor injury
All organs have limited ability to regenerate (1 cm)

Every 30 Seconds a Person Dies from a Disease that can be Treated with Tissue Replacement

Human Intervention History

Blood Transfusion

- 1492: Pope Innocent VIII
- 1901: Human blood groups discovery Austrian Karl Landsteiner (1939 Rh)
- 1925: First transfusion center Alexander Bogdanov Moscow

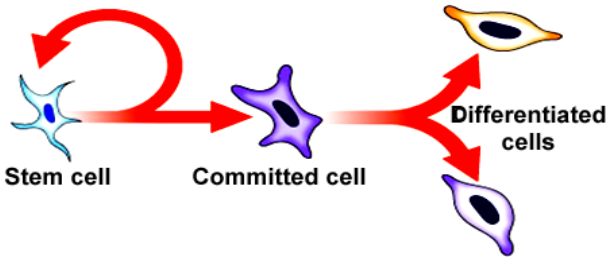
Solid Tissue

- 1905: Cornea transplant Eduard Zirm
- 1954: Kidney transplant Joseph Murray
- 1998: Hand transplant France
- 2005: Partial face transplant France

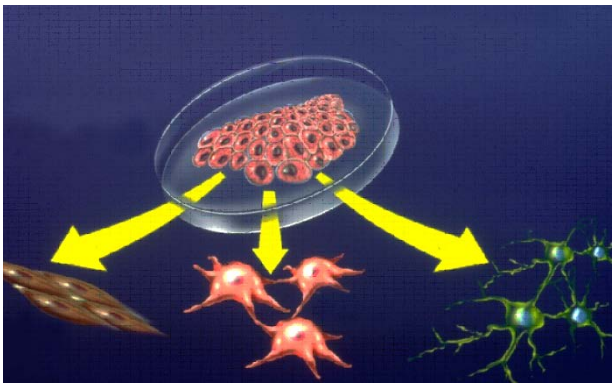
Current Approaches

i. **Organ Repair via Stem Cells**
ii. **Cytokine Induce Stem Cell Proliferation & Differentiation**
iii. **In-vitro Organ/Tissue Fabrication**

i. Organ Repair via Stem Cells



Stem cells have two remarkable qualities that make them a good choice for repairing tissue damage



1. **Stem cells are unspecialized cells that renew themselves.**
2. **Under certain physiologic or experimental conditions, Stem cells can be induced to become cells with special functions.**

There are two types of stem cells Embryonic Stem Cells and Adult Stem Cells.

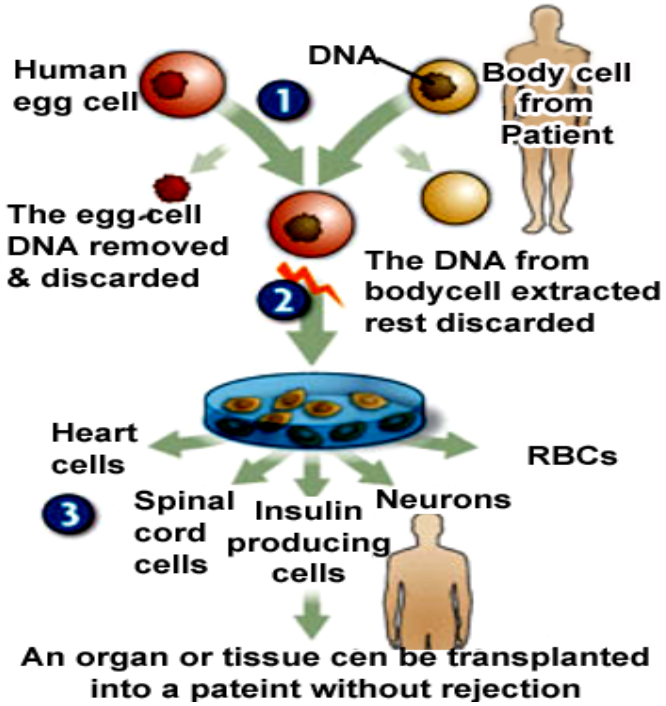
Embryonic Stem Cells

Are derived from the embryos (ethical dilemma: embryos destroyed when embryonic stem cells are isolated from the embryos) Blastomere Biopsy can be used to circumvent this issue.

Have the ability to differentiate into any cell type in the body.

Have unlimited ability to renew themselves.

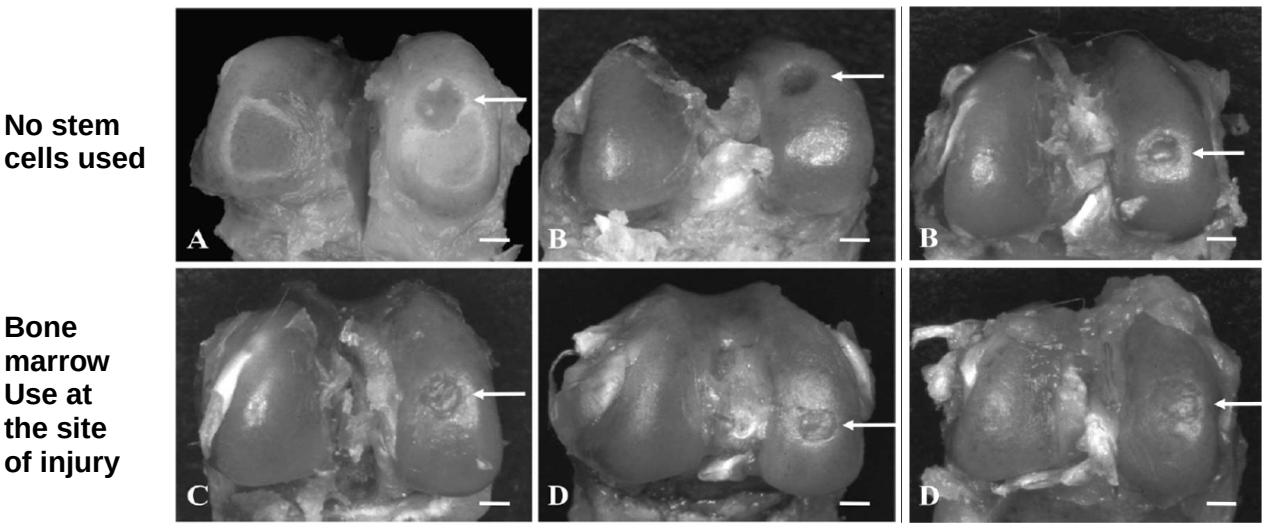
Being foreign in origin, when transplanted can cause immune rejection. This issue can be resolve by generating customized embryonic stem cells (genetic makeup identical to the recipient) by using nuclear transfer technique.



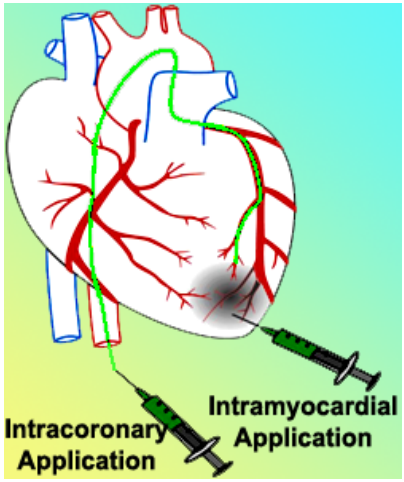
Adult Stem cells

Are harvest from there niches from a (postnatal) individual.
No ethical dilemma for their use in research, since no potential person is destroy in obtaining these cells.
Have a limited ability to renew themselves and differentiate in to different cell types.
Do not cause an immune rejection (if donor and the recipient are the same person).

Adult stem cells used to repair damaged rat bone

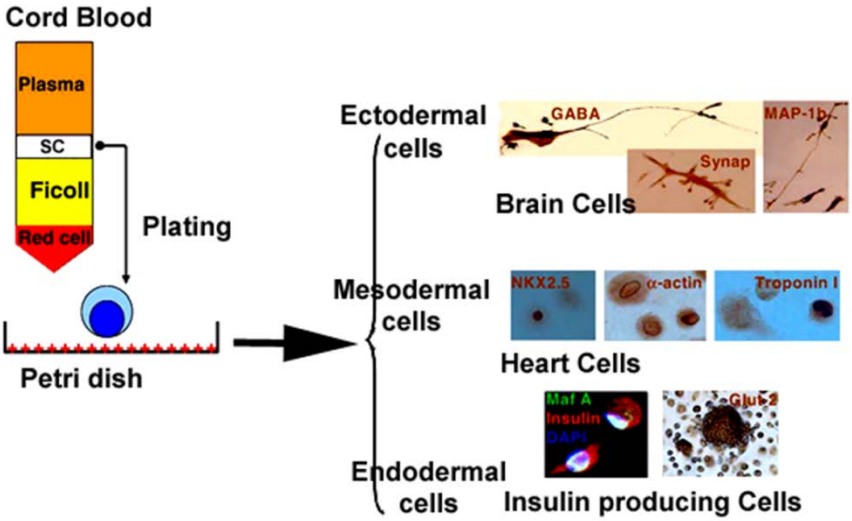


Repair of Damaged Heart Tissue after Heart Attack



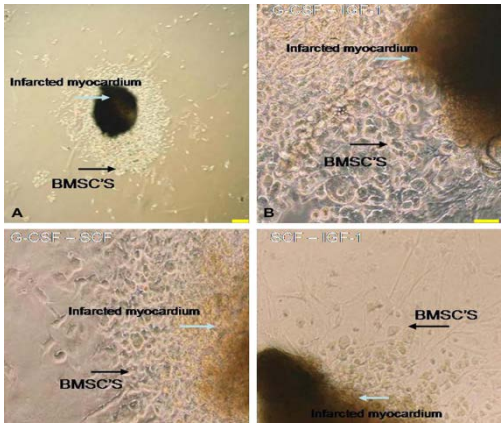
Düsseldorf group performed clinical stem cell application for the first time in 2001, treating acute myocardial infarction by intracoronary cell transfer. Currently, the optimal time point for cardiac stem cell application in an ischemic setting seems to lie within the first month after myocardial infarction. "There is growing evidence that bone marrow-derived stem/progenitor cells improve cardiac function" (Leri 2009).

Use of the Cord Blood cells may offer the best of both worlds (embryonic and Adult)



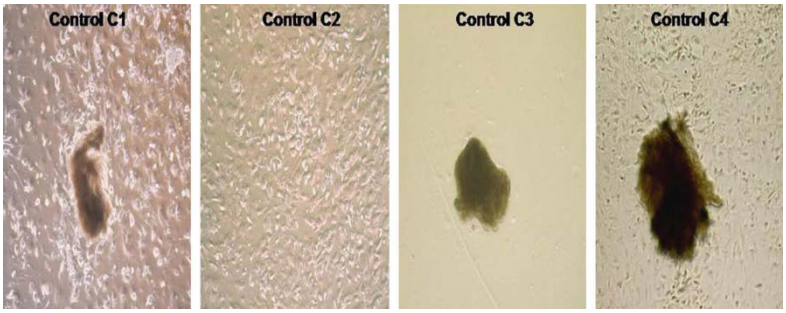
Cytokine Induced Stem Cell Proliferation & Differentiation

Almost all the tissues have a reservoir of stem cells which if appropriated stimulated can also be employed to repair damaged tissues.



IGF-1 and G-CSF complement each other in BMSC migration towards infarcted myocardium in a novel in vitro model

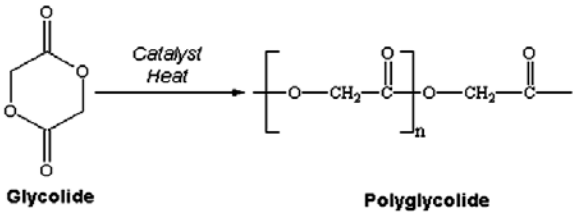
Mohsin Khan, Sobia Manzoor, Sadia Mohsin, Shaheen N. Khan1, Fridoon Jawad Ahmad*,1



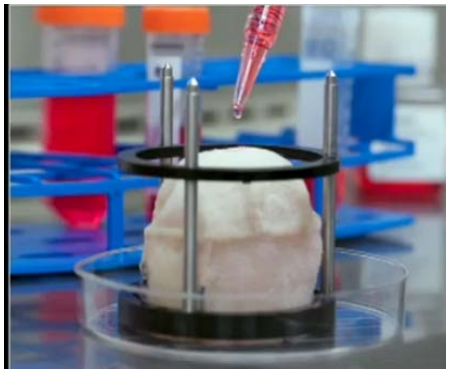
In-vitro Organ/Tissue Fabrication

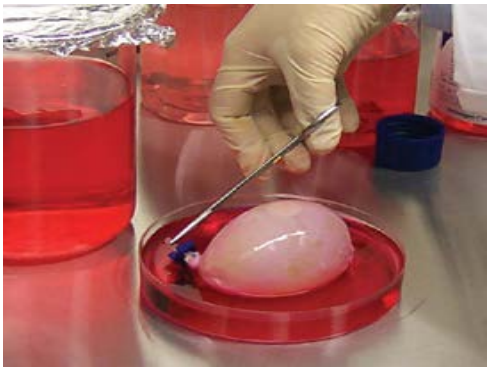
Artificial Scaffolds

Scaffolds are required to fabricate solid organs since these organs have a specific shape that is important for their proper functioning.



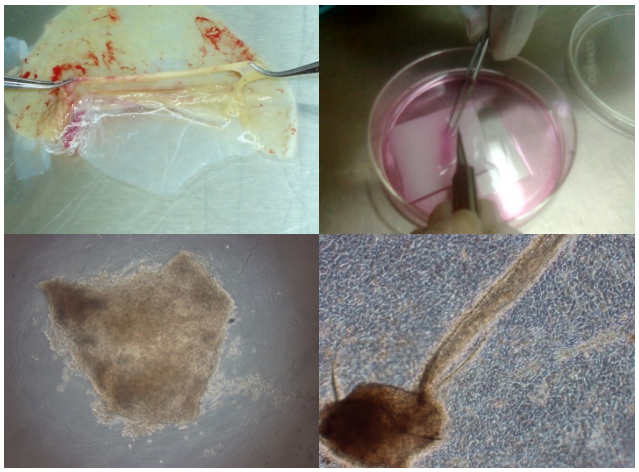
can be used to make different scaffolds that can be seeded with cells in order to make complex three dimensional solid organs.



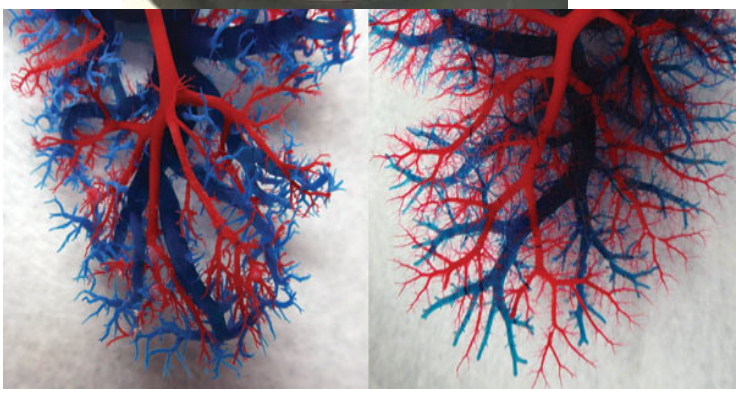


Scaffolds are biodegradable and once they are degraded only the cellular tissue remains.

Natural Scaffolds



De-cellularized amniotic membranes can be used to make two dimensional structures like cornea.



Solid organ with complex structures/vasculature (e.g. liver) can be de-cellularized and their extracellular matrix can be used to make such organs.