VU Medical Zone

Admin: Amaan Khan

Introduction to Bioinformatics

BIF101 Final Merge ppt

Lecture 96 to 150

Protein Databases

Introduction

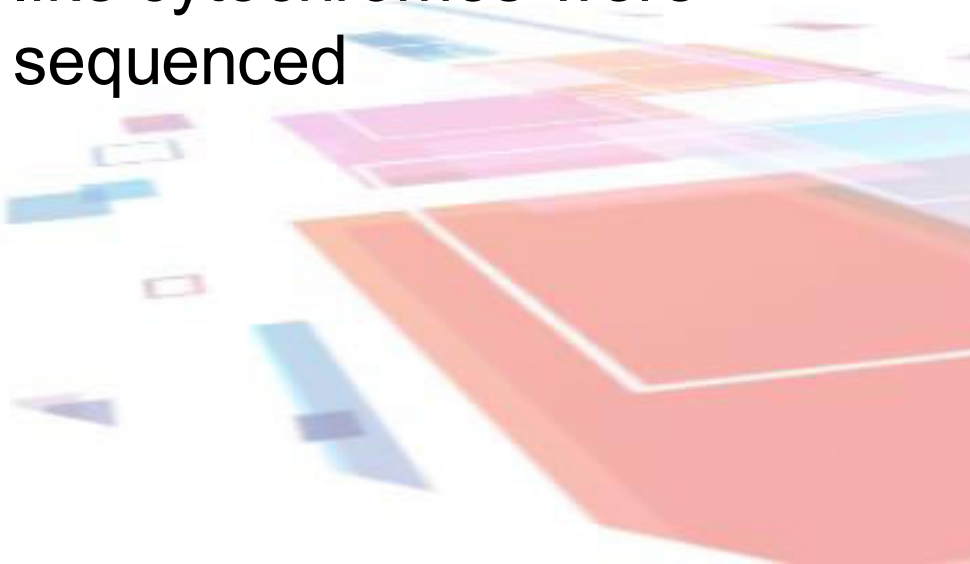
Protein databases store

- **Protein sequences**
- **Motif**
- **Structure**
- **Structure alignments**

Protein Databases

Origin

First sequences to be collected were Proteins using **Sanger and Tupy's** methods (1951) where Common protein families like cytochromes were sequenced



Protein Databases

Origin

Atlas of protein sequences (mainly cytochromes) was assembled by Margret Dayhoff and her collaborators at **N**ational **B**iomedical **R**esearch **F**oundation (NBRF) in 1960s

Protein Databases

PIR ((Protein Information Resource)

The collection of Dayhoff and co became **PIR**

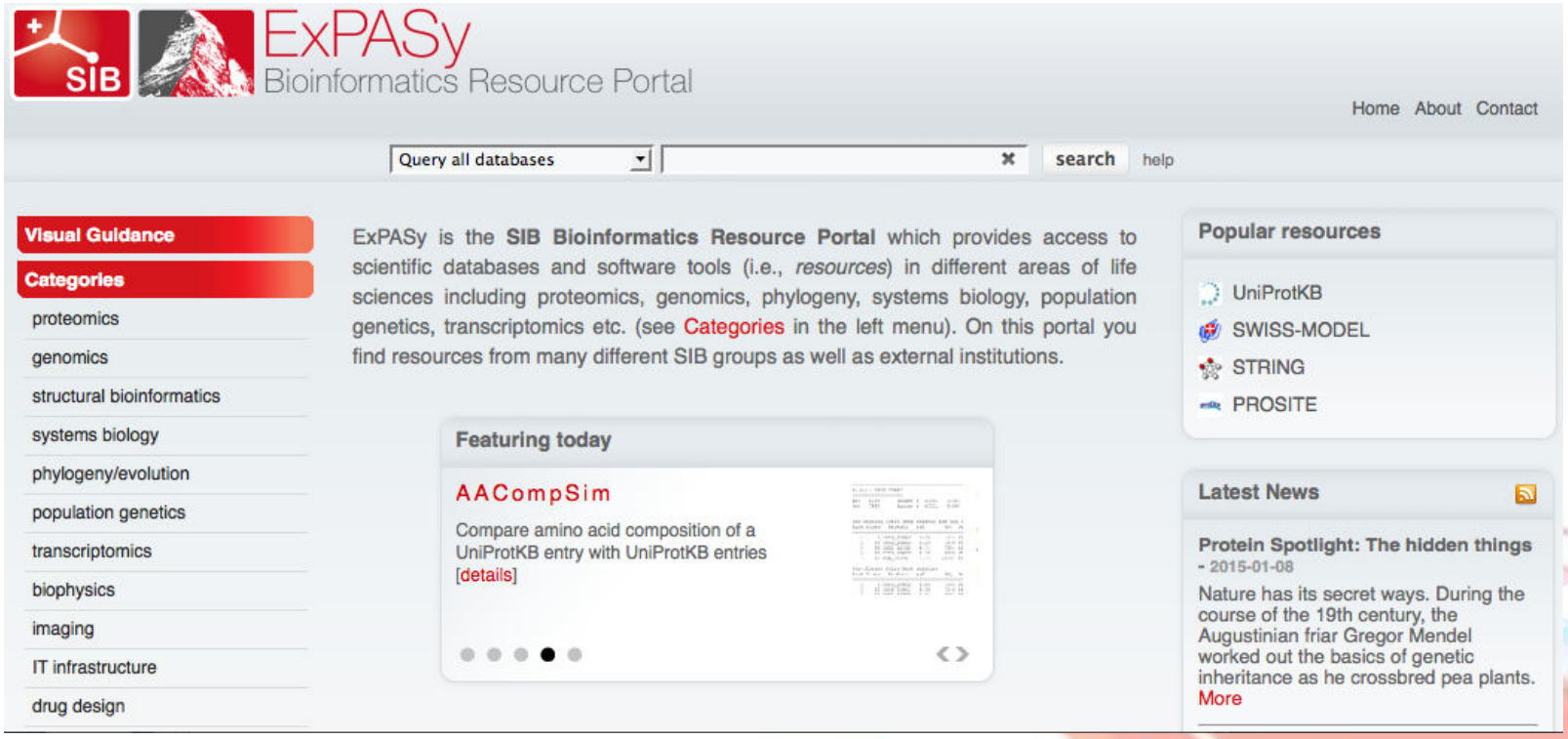
which is now a collaboration of **NBRF**, **M**unich **C**enter for **P**rotein **S**equences (**MIPS**) and **J**apan **I**nternational **P**rotein **I**nformation **D**atabase (**JIPID**)

Protein Databases

Protein Sequences

- **Swiss-Prot** is Collaboration between the SIB and EBI
- Weekly releases from about 50 servers across the world, the main source being ExPASy in Geneva

Protein Databases



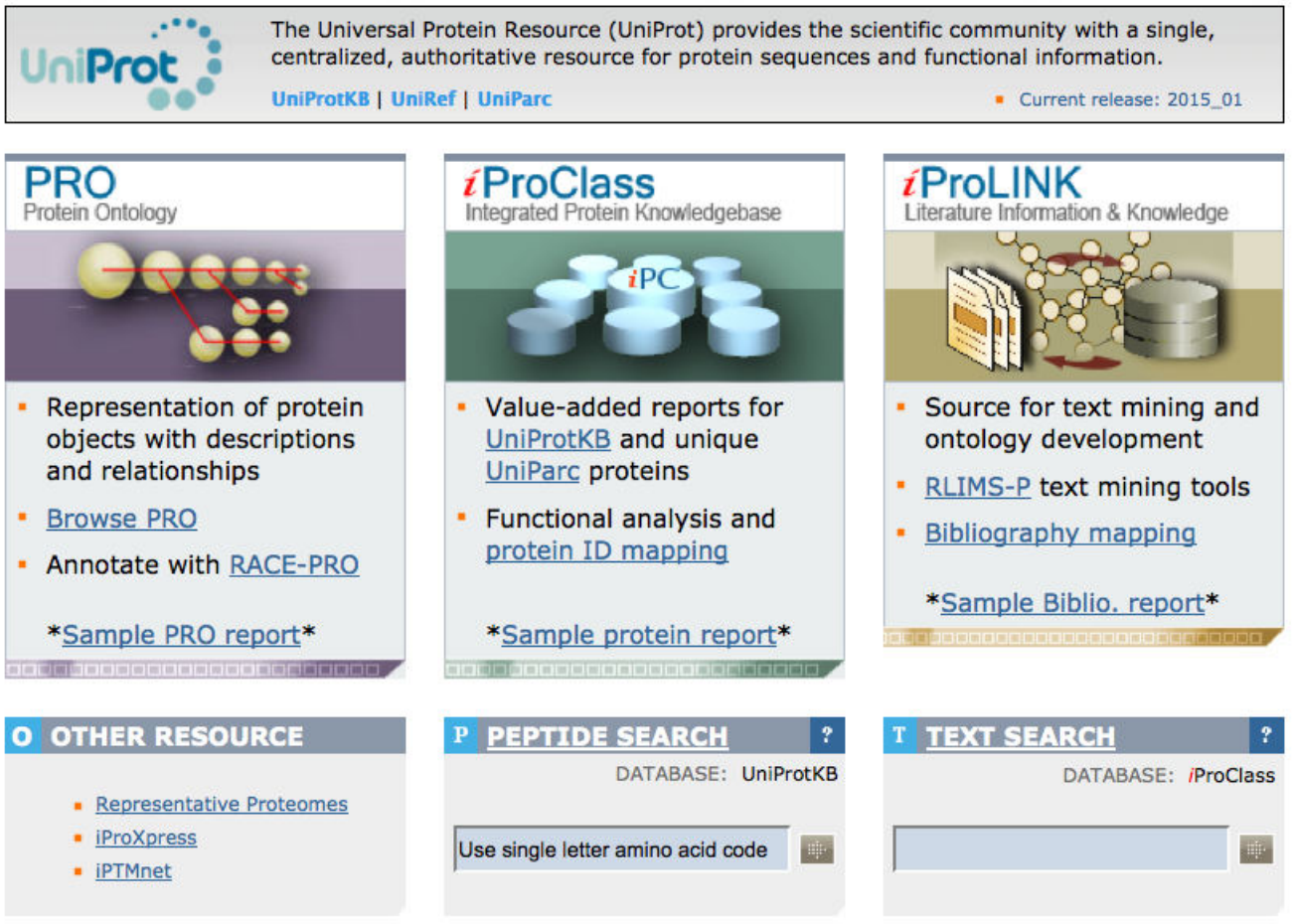
http://www.expasy.org/

Protein Databases

Protein Sequences

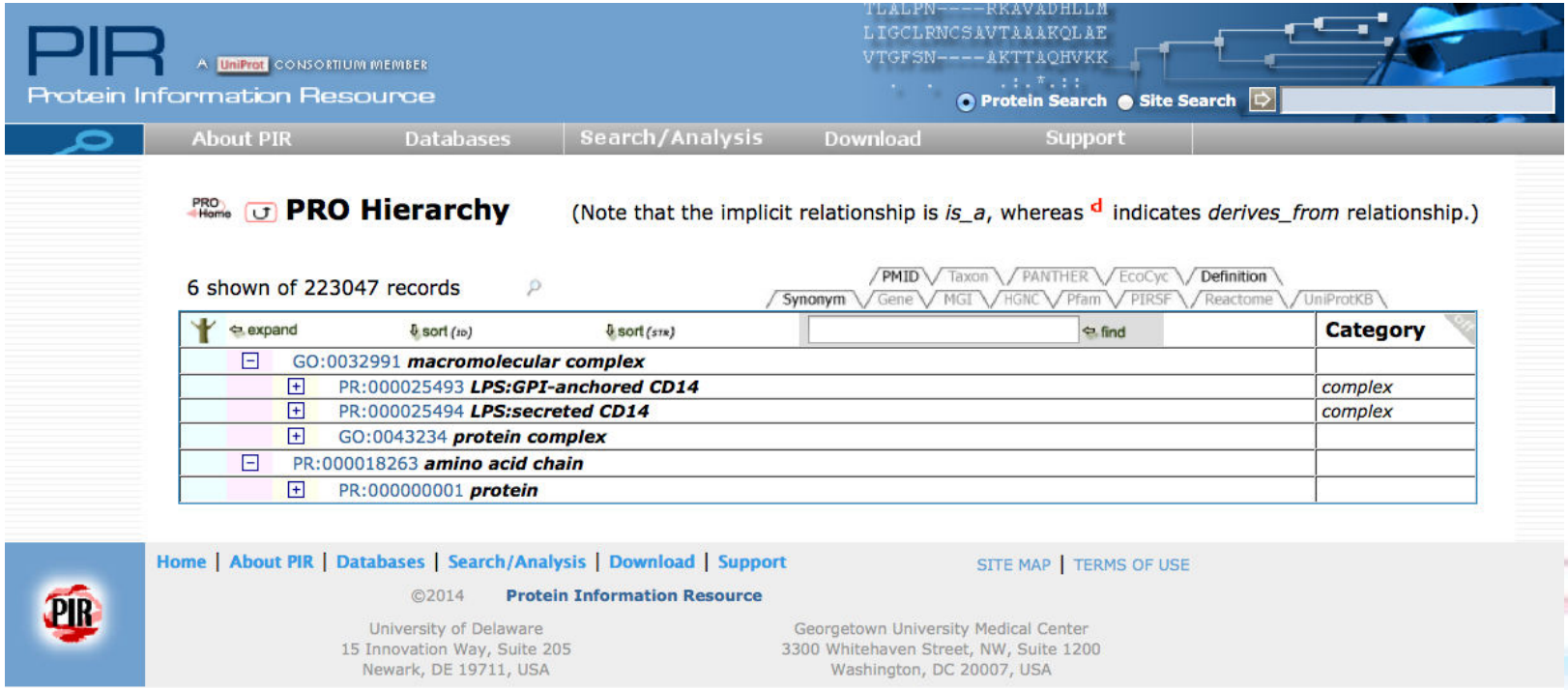
- International partnership between PIR, EBI and SIB
- Created **UniProt**, by unifying the PIR-PSD, Swiss-Prot, and TrEMBL databases

Protein Databases



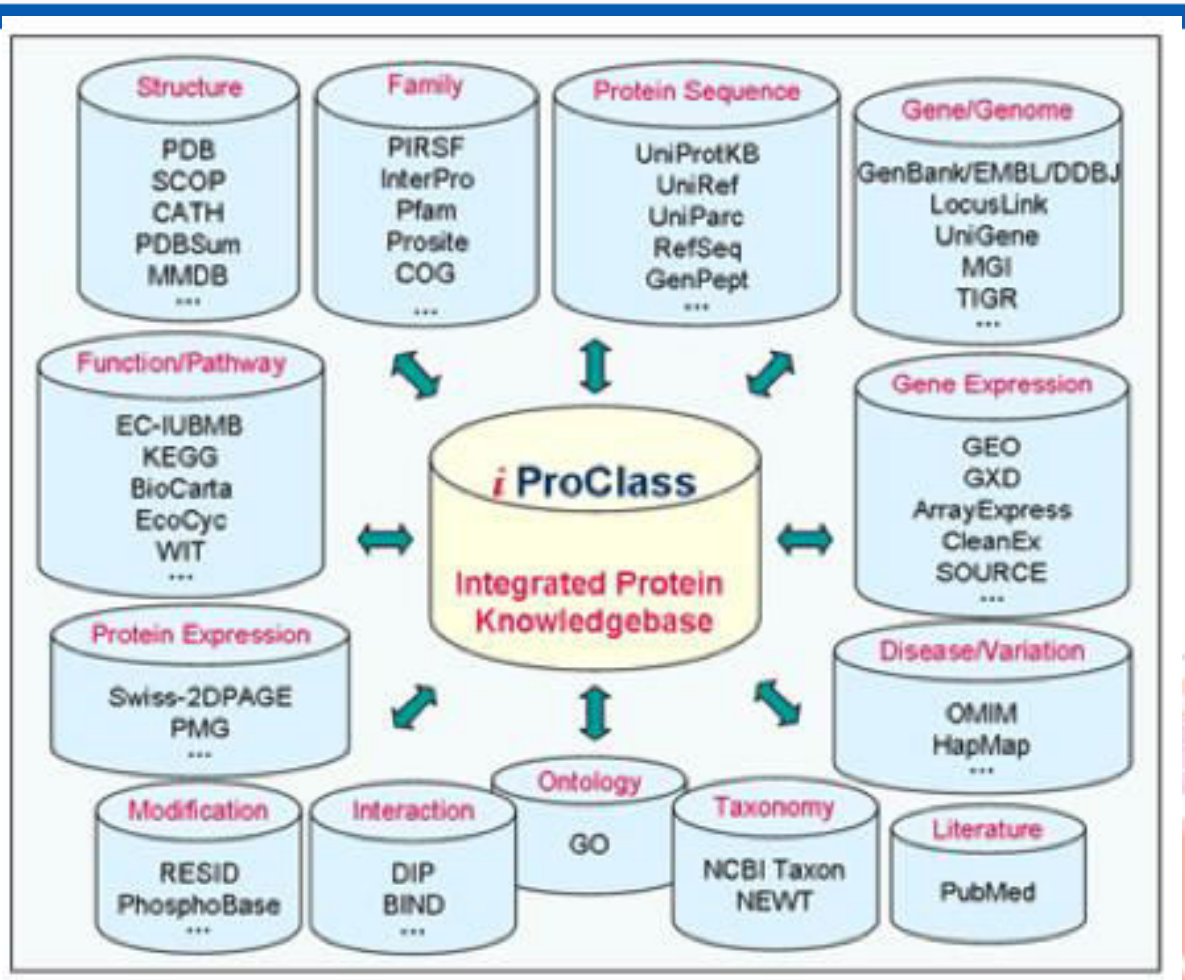
http://www.uniprot.org/

Protein Databases



http://pir.georgetown.edu/

Protein Databases

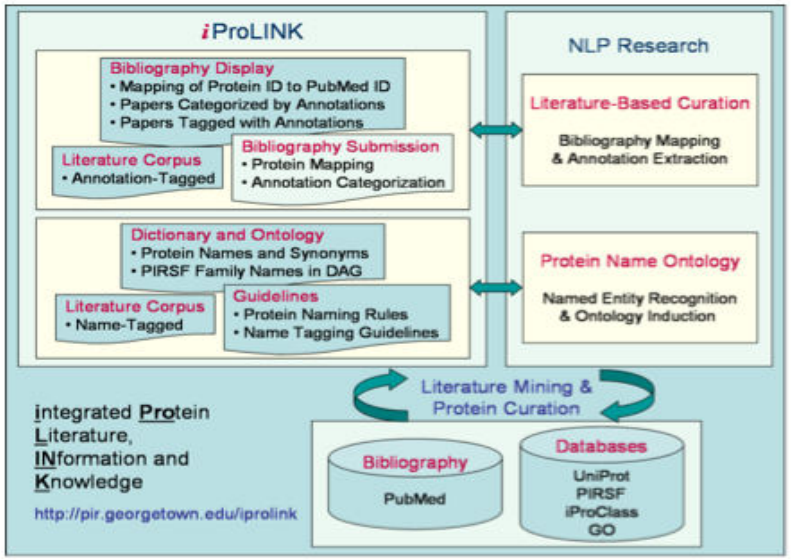


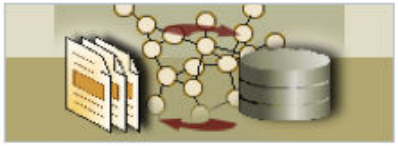
http://www.uniprot.org/

Protein Databases

iProLINK Tutorial

iProLINK (integrated Protein Literature, INformation and Knowledge) has been developed as a resource to facilitate text mining in the area of literature-based database curation, named entity recognition, and protein ontology development. The collection of data sources can be utilized by computational and biological researchers to explore literature information on proteins and their features or properties (Hu *et al.*, 2004).





Bibliography Mapping/ Annotation Extraction

- Bibliography Display/Submission
- Annotation-Tagged Corpora
- RLIMS-P Text Mining Tool
- eFIP Text Mining Tool
- eGIFT Text Mining Tool
- iSimp Sentence Simplification System

Entity Recognition/ Ontology Development

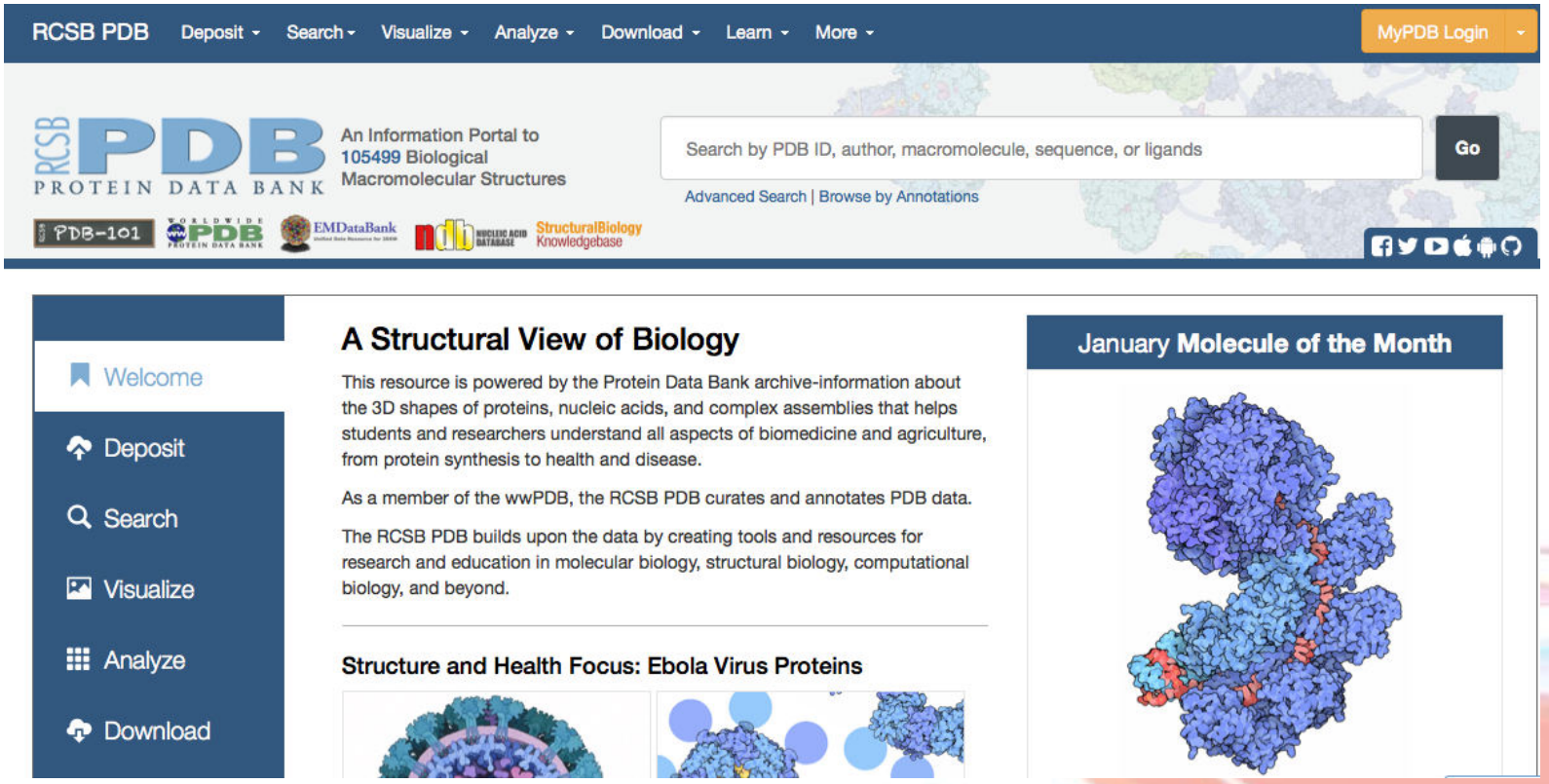
- Name Tagging Guidelines/Corpora
- Protein Ontology Development

Collaborators

***i*ProLINK Paper**

http://www.uniprot.org/

Protein Databases



http://www.rcsb.org/pdb/home/home.do

Protein Databases

SCOPe: Structural Classification of Proteins — extended. Release 2.04 (July 2014, new entries added 2014-12-18)

Browse | Stats & History | ASTRAL Subsets | Downloads | Related Resources | References | Help | About

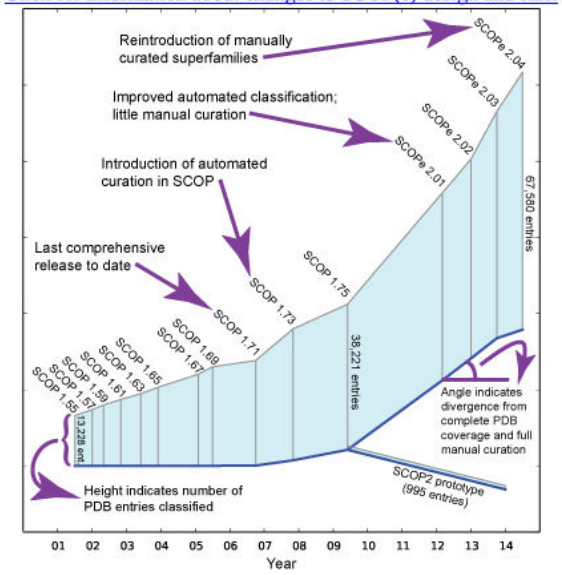
Welcome to SCOPe!
SCOPe is a database developed at the Berkeley Lab and UC Berkeley to extend the development and maintenance of SCOP.
SCOP was conceived at the MRC Laboratory of Molecular Biology, and developed in collaboration with researchers in Berkeley.
Work on SCOP (version 1) concluded in June 2009 with the release of SCOP 1.75.
SCOPe classifies many newer structures through a combination of automation and manual curation, and corrects some errors in SCOP,
aiming to have the same accuracy as the hand-curated SCOP releases. **SCOPe** also incorporates and updates the ASTRAL database.
For prior releases, click on the Stats & History tab above. For more info, click on the About tab above.
New PDB entries were last added on **2014-12-18**; for more info on periodic updates click on the Help tab above.

Search **SCOPe** (example): Search

Classes in SCOPe 2.04:

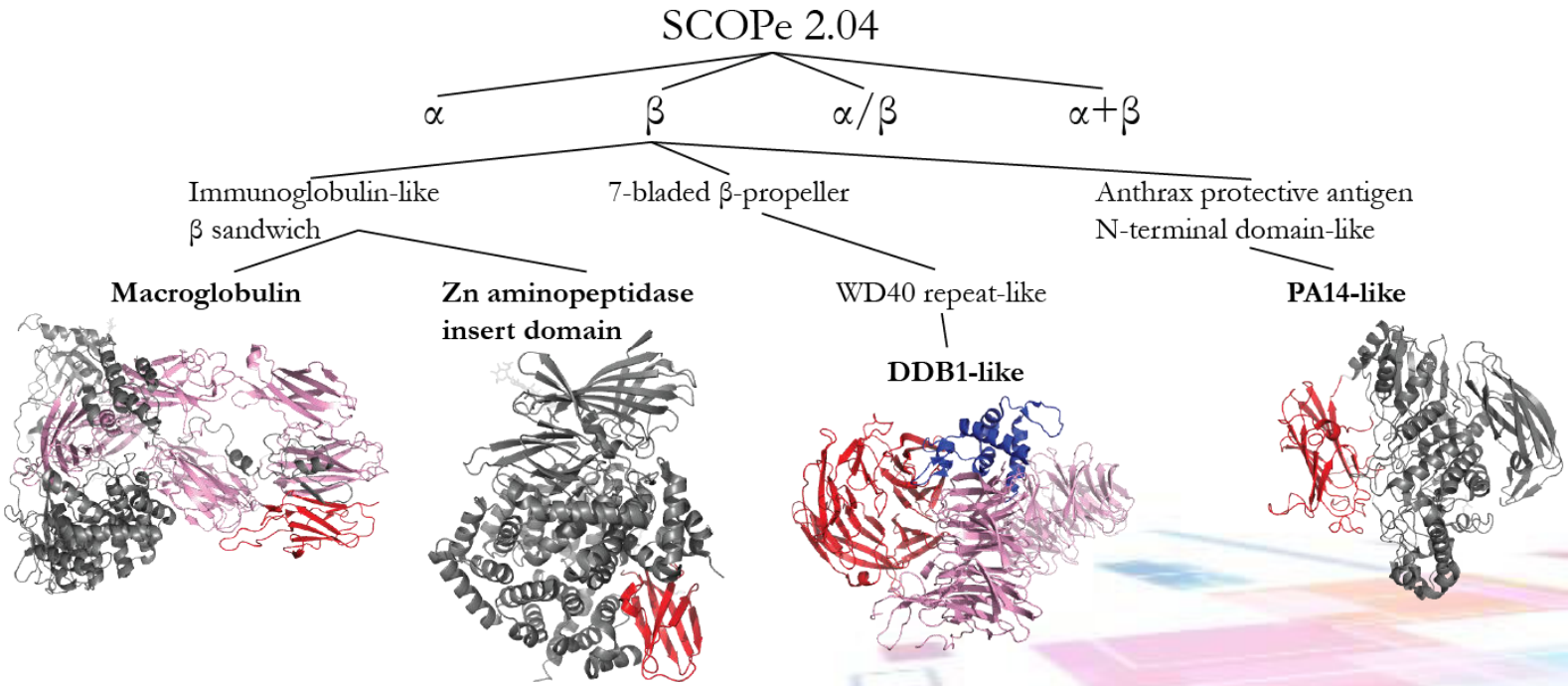
1. a: All alpha proteins [46456] (285 folds)
2. b: All beta proteins [48724] (176 folds)
3. c: Alpha and beta proteins (a/b) [51349] (148 folds)
4. d: Alpha and beta proteins (a+b) [53931] (380 folds)
5. e: Multi-domain proteins (alpha and beta) [56572] (68 folds)
6. f: Membrane and cell surface proteins and peptides [56835] (57 folds)
7. g: Small proteins [56992] (91 folds)
8. h: Coiled coil proteins [57942] (7 folds)
9. i: Low resolution protein structures [58117] (25 folds)

Click for information about changes to SCOP(e) design and size.



http://scop.mrc-lmb.cam.ac.uk/scop/

Protein Databases



http://scop.mrc-lmb.cam.ac.uk/scop/

Protein Databases

Conclusions

- First sequences to be collected were Protein sequences
- Protein databases are classified on the basis of sequences, motifs and structures
- Growth of Sequence in Databases is exponential

Genome & Organism Specific Databases

Origin

- First attempt to sequence free living organism was launched in late 1990's
  - (Blattner et al. 1997)
- Viruses had already been sequenced

Genome & Organism Specific Databases

Origin

- ***Haemophilus influenzae*** was the first published genome
  - (Fleischmann et al. 1995)
- The project initiated at **T**he **I**nstitute of **G**enome **R**esearch **(TIGR)** under leadership of **Craig Ventor**

Genome & Organism Specific Databases

Origin

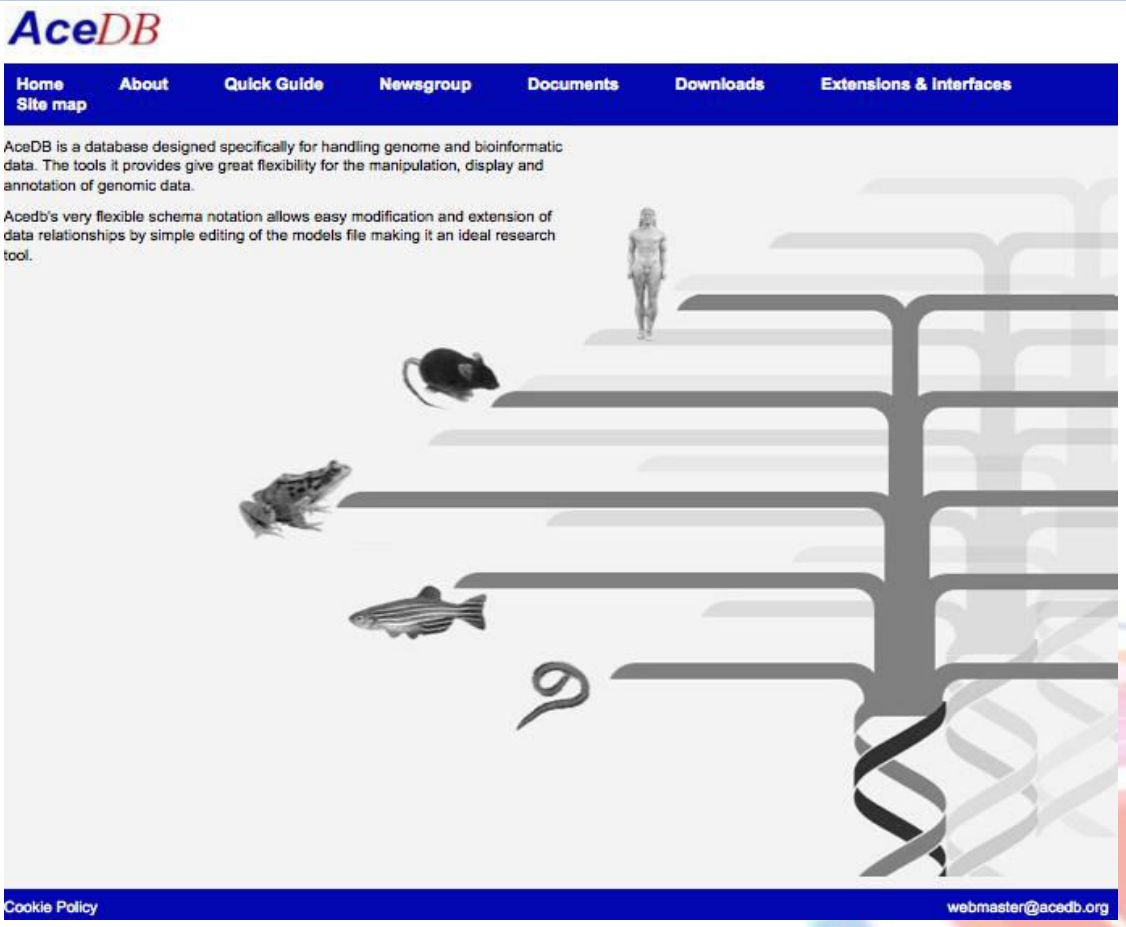
- Use of **shotgun sequencing method** speeded up the process
- 1.8 million bp
- Took 9-months and cost was 1 million
- Paved the way for sequencing of many other organisms

Genome & Organism Specific Databases

Examples

- **AceDB** (**A** ***C. e****legans* **D**ata**B**ase) was the first genome database developed in 1989
  - by **Richard durbin** and **Thierry-Miegi**

Genome & Organism Specific Databases

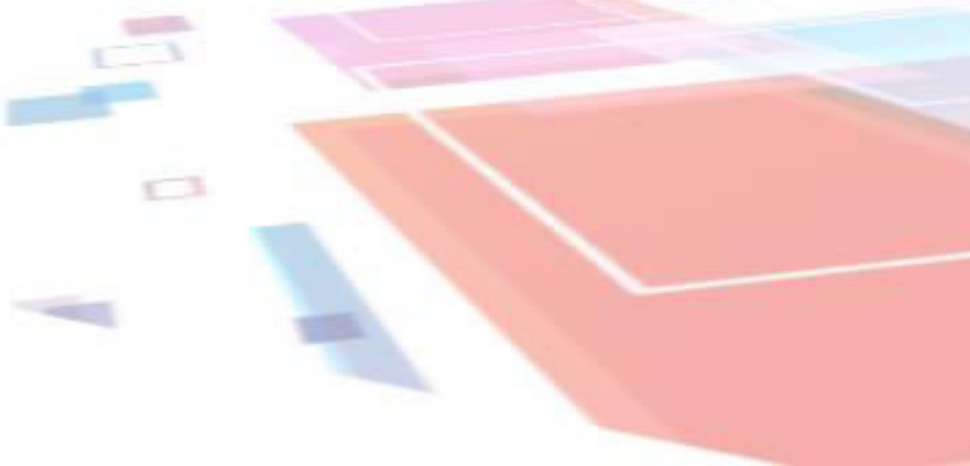


http://www.acedb.org/

Genome & Organism Specific Databases

Examples

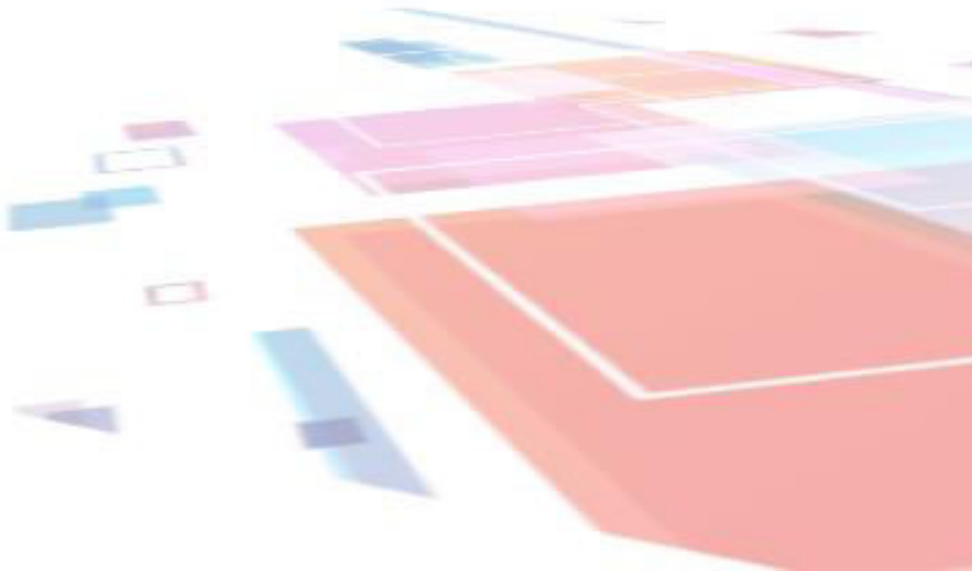
- **TAIR** (**T**he **A**rabidopsis **I**nformation **R**esource) http://www.arabidopsis.org/ and **SGB** (**S**accharomyces **G**enome **D**atabase) http://www.yeastgenome.org/ utilized **AceDB** system



Genome & Organism Specific Databases

Human Genome Project

- Pilot project, the **Human Genome Initiative** begun by **D**epartment **O**f **E**nergy (**DOE**) in 1986

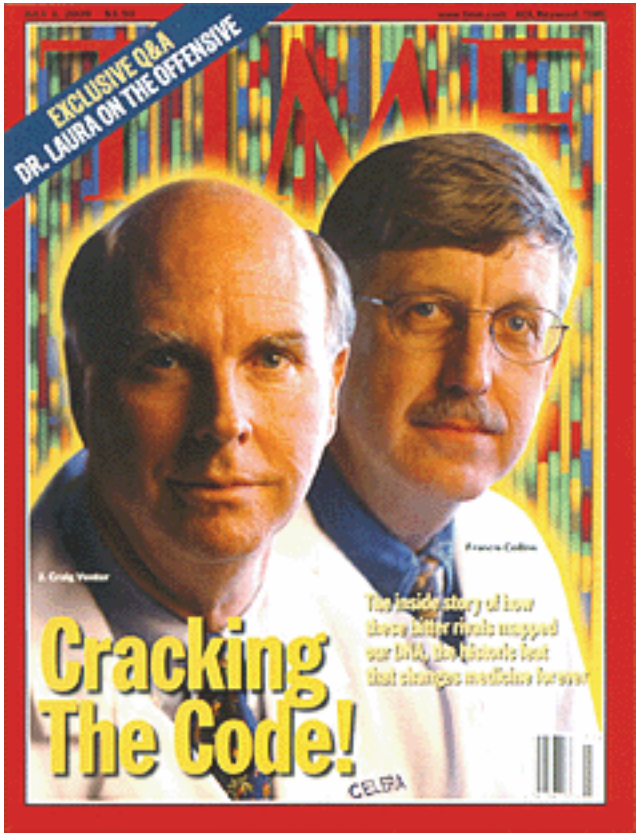


Genome & Organism Specific Databases

Human Genome Project

- **N**ational **H**uman **G**enome **R**esearch **I**nitiative (**NHGRI**), a federally funded organization (NIH) started in 1988 **(Francis Collin)**
- Celera, a commercial vendor **(Craig Venter)** joined the venture in 1998

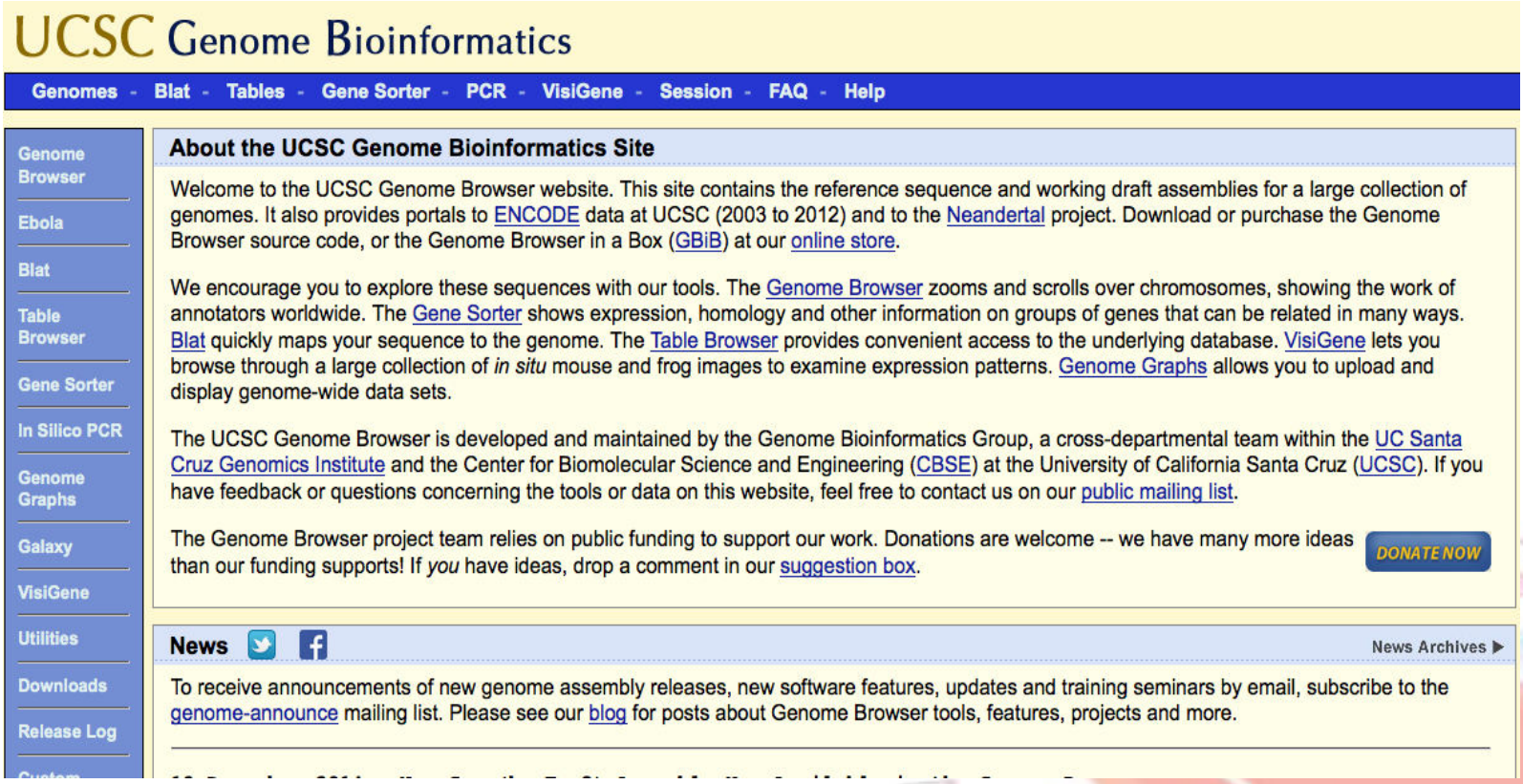
Genome & Organism Specific Databases



Human Genome Project

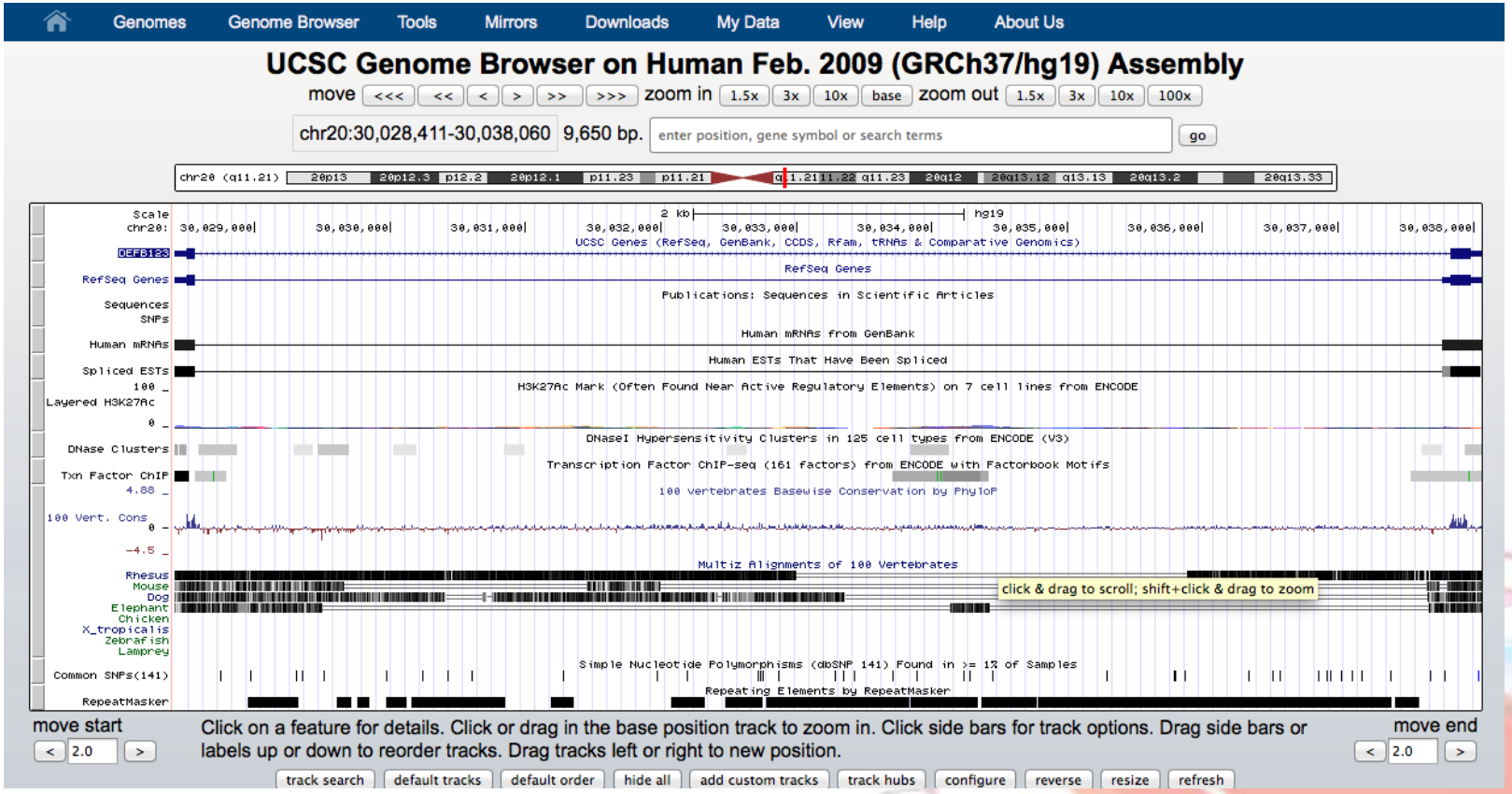
- Both simultaneously announced the completion of the project in 2000
- Total 3.4 billion bases sequenced at a cost of $1/base

Genome & Organism Specific Databases



http://genome.ucsc.edu/

Genome & Organism Specific Databases



http://genome.ucsc.edu/

Genome & Organism Specific Databases

Conclusion

- Success of ***Haemophilus influenzae*** paved the way for other genome sequencing projects
- Human Genome Project was accomplished by NHGRI and Celera
- Genome browsers help viewing Genome features

Gene Expression Databases

Gene Expression Omnibus (GEO)

- A public repository for the archiving and distribution of gene expression data submitted by the scientific community

Gene Expression Databases

Gene Expression Omnibus (GEO)

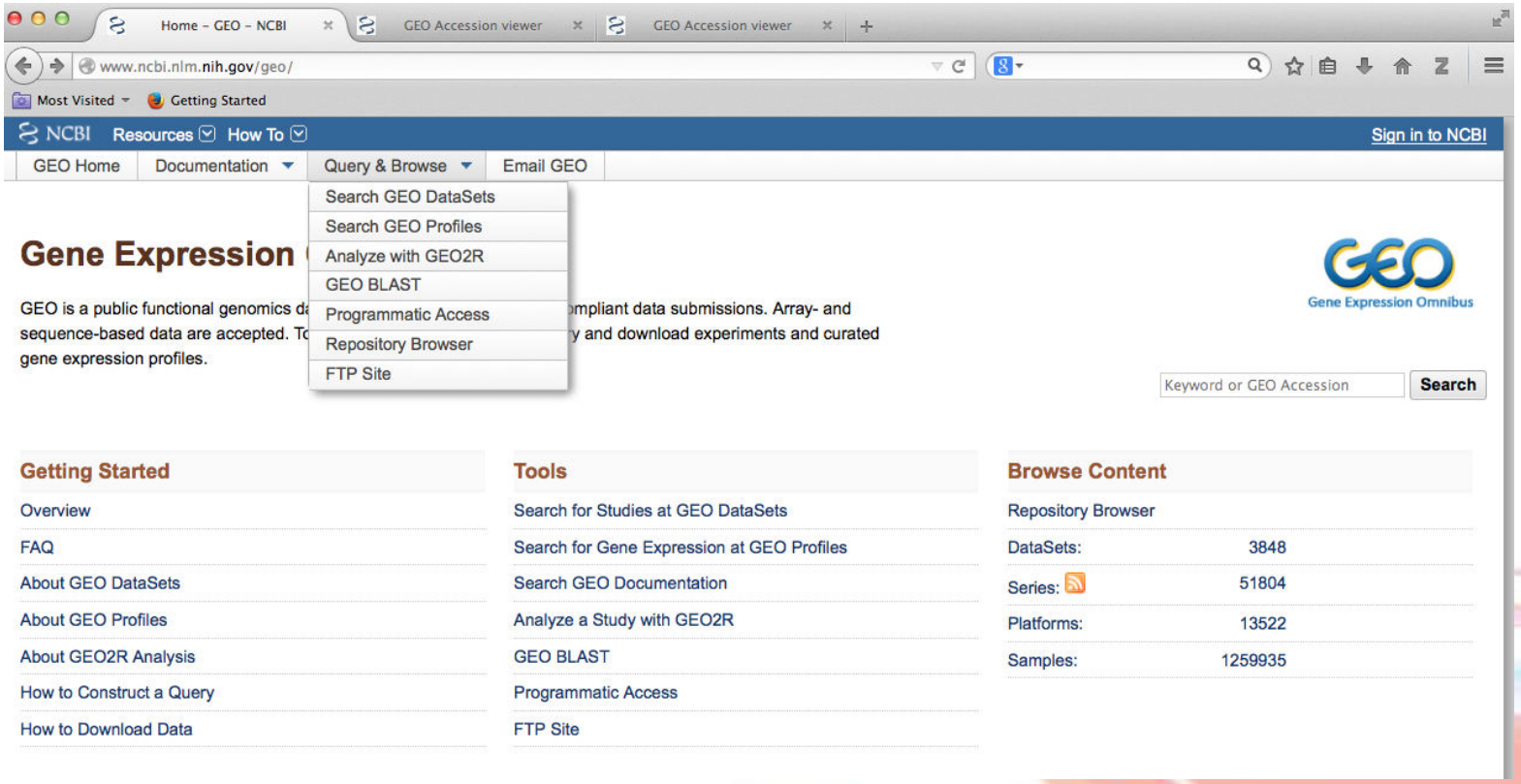
- MIAME compliant data
  - Minimum Information About a Microarray Experiment http://www.mged.org/Workgroups/MIAME/miame.html

Gene Expression Databases

Gene Expression Omnibus (GEO)

- Convenient for deposition of gene expression data, as required by funding agencies and journals
- Curated resource for gene expression data
- Browsing, querying, analysis and retrieval

Gene Expression Databases



http://www.ncbi.nlm.nih.gov/geo/

Gene Expression Databases

Gene Architecture

GEO has four kinds of records

- **Sample (GSM)**
  preparation and description of the samples
- **Platform (GPL)**
  technology used and the features detected
  - microarray or RNASeq

Gene Expression Databases

Gene Architecture

GEO has four kinds of records

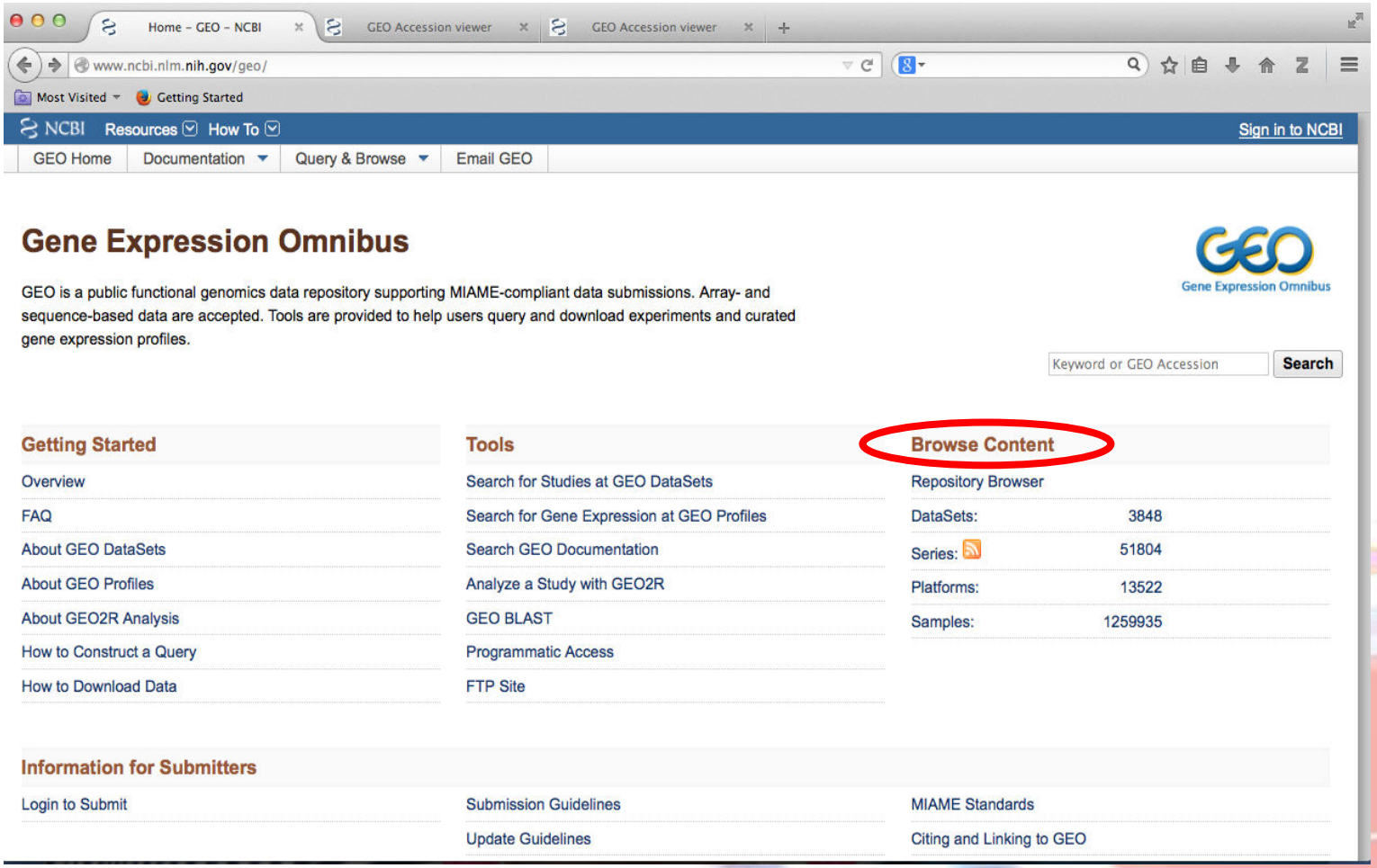
- **Series (GSE)**

  defines a set of samples and how they are related

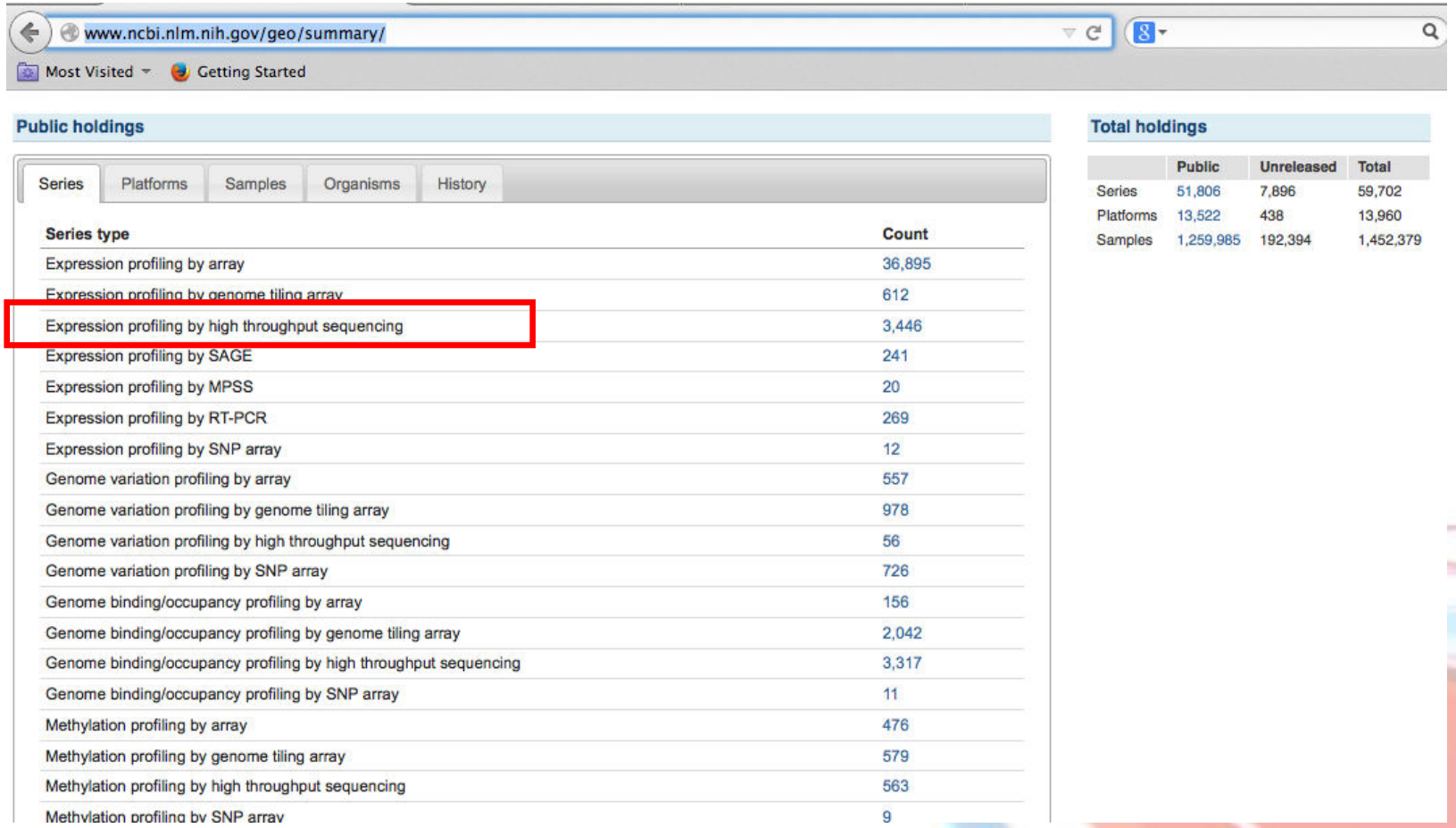
- **Datasets (GDS)**

  sample data collections assembled by GEO

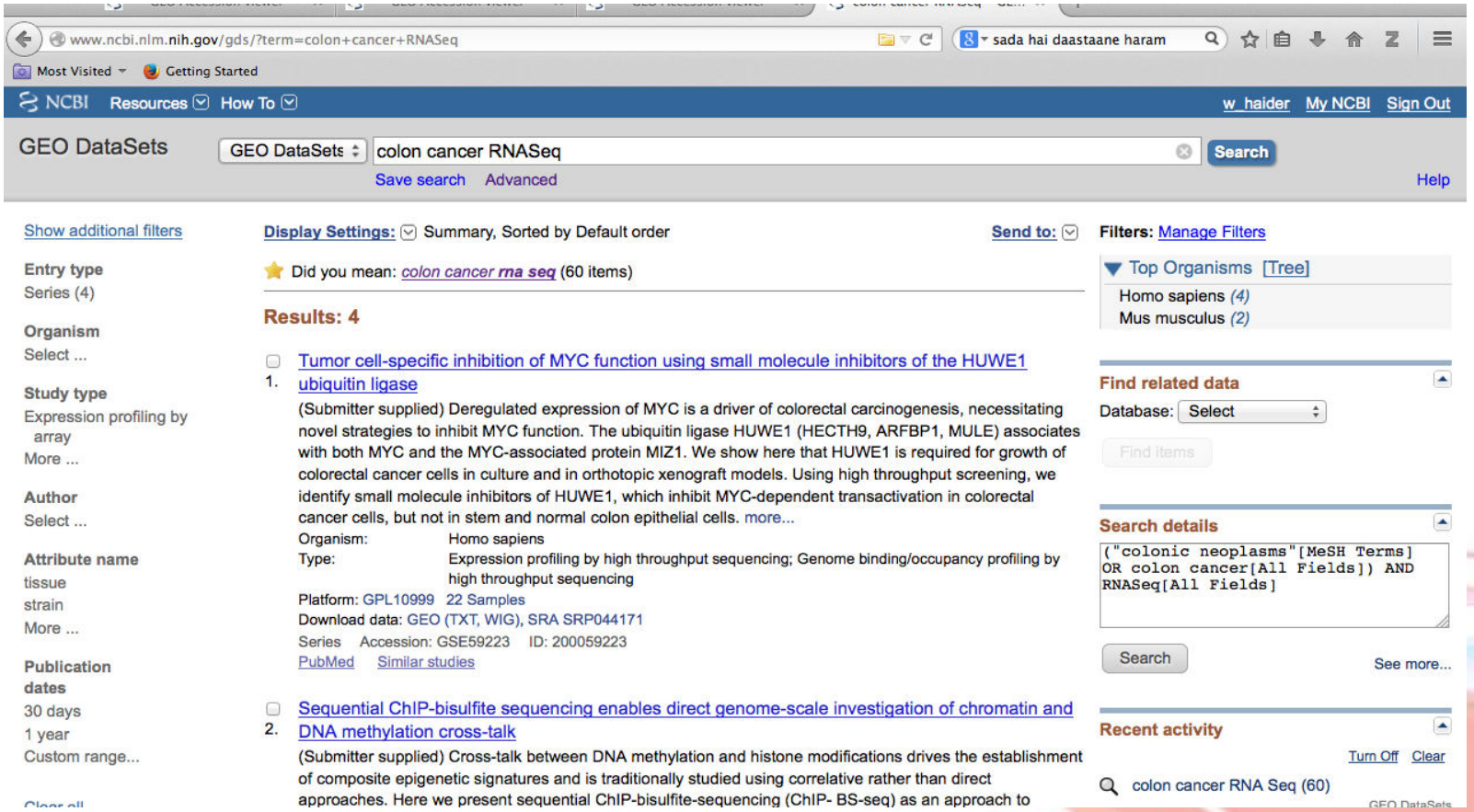
Gene Expression Databases



Gene Expression Databases



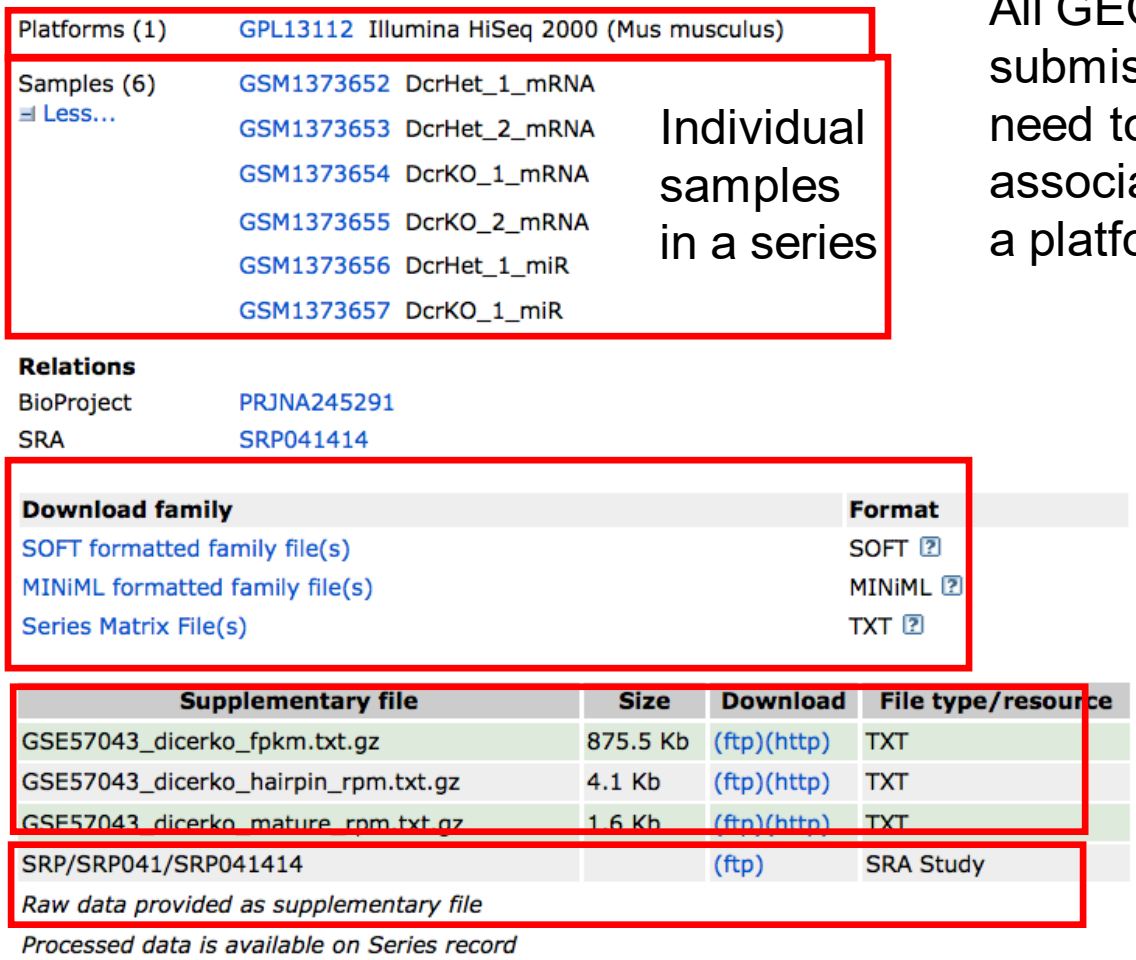
Gene Expression Databases



Gene Expression Databases

Data is submitted to GEO as a Series, which represents the experiment design

Gene Expression Databases



Platforms (1)	GPL13112 Illumina HiSeq 2000 (Mus musculus)
Samples (6)	GSM1373652 DcrHet_1_mRNA
Less...	GSM1373653 DcrHet_2_mRNA
	GSM1373654 DcrKO_1_mRNA
	GSM1373655 DcrKO_2_mRNA
	GSM1373656 DcrHet_1_miR
	GSM1373657 DcrKO_1_miR

Individual samples in a series

Relations

BioProject	PRJNA245291
SRA	SRP041414

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE57043_dicerko_fpkm.txt.gz	875.5 Kb	(ftp)(http)	TXT
GSE57043_dicerko_hairpin_rpm.txt.gz	4.1 Kb	(ftp)(http)	TXT
GSE57043_dicerko_mature_rpm.txt.gz	1.6 Kb	(ftp)(http)	TXT
SRP/SRP041/SRP041414		(ftp)	SRA Study

Raw data provided as supplementary file

Processed data is available on Series record

All GEO submissions need to be associated with a platform file.

Click ? One by one

Normalized counts

Raw Reads

Gene Expression Databases

PNAS

An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival

Lance D. Miller*†, Johanna Smeds‡, Joshy George*, Vinsensius B. Vega*, Liza Vergara*, Alexander Ploner§, Yudi Pawitan§, Per Hall§, Sigrid Klaar‡, Edison T. Liu*†, and Jonas Bergh‡

*Genome Institute of Singapore, 60 Biopolis Street, #02-01, Singapore 138672; ‡Department of Oncology and Pathology, Radiumhemmet, Karolinska Institute and Hospital, S-17176 Stockholm, Sweden; and §Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, 17177 Stockholm, Sweden

GENETICS. For the article "An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival," by Lance D. Miller, Johanna Smeds, Joshy George, Vinsensius B. Vega, Liza Vergara, Alexander Ploner, Yudi Pawitan, Per Hall, Sigrid Klaar, Edison T. Liu, and Jonas Bergh, which appeared in issue 38, September 20, 2005, of *Proc. Natl. Acad. Sci. USA* (**102,** 13550-13555; first published September 2, 2005; 10.1073/pnas.0506230102), the breast cancer microarray data discussed in this publication have been deposited in the National Center for Biotechnology Information's Gene Expression Omnibus database (GEO, www.ncbi.nlm.nih.gov/geo/) and are accessible through GEO Series accession no. GSE3494 [NCBI GEO] .

Gene Expression Databases

Conclusion

- **GEO** is a public repository for the archiving and distribution of gene expression data
- Best resource to get microarray and Next Generation Sequencing (RNASeq) data

Medical Databases

Introduction

Informatics in health care may be called as health informatics

- It deals with the resources, devices, and methods required to optimize the acquisition, storage, retrieval, and use of information in health and biomedicine.

(wiki)

Medical Databases

Introduction

Medical databases store and provide medical information

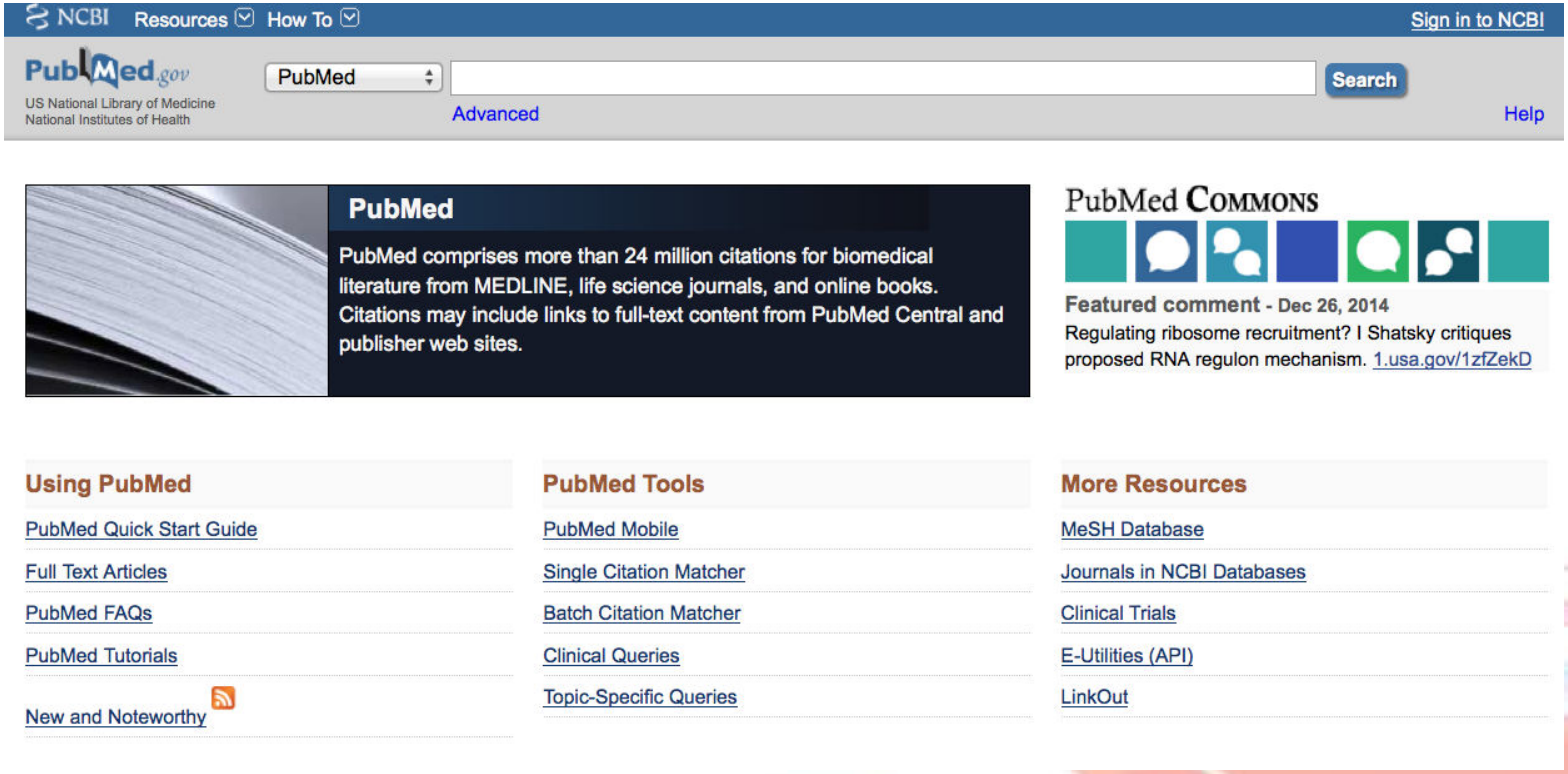
- The premier database for biomedical literature is the National Library of Medicine (NLM)'s MEDLINE, accessible through **PubMed**

Medical Databases

PUBMED

- Comprises of more than 24 million citations for biomedical literature from MEDLINE, life science journals, and online books
- Citations may include links to full-text content from PubMed Central and publisher web sites

Medical Databases



Medical Databases

MEDLINE

- MEDLINE is the primary resource for biomedical journal articles
- Millions of citations to articles in biomedical journals
- MEDLINE uses the MeSH vocabulary

Medical Databases

Other Databases

MEDLINE is the primary resource, but other databases may also be helpful

- Academic OneFile
- CINAHL (Cumulated Index of Nursing and Allied Health Literature)
- PsycINFO
- Web of Knowledge

Medical Databases

Academic OneFile

- Academic OneFile lists articles from journals covering a broad range of subjects
- While it does not primarily focus on medical topics, useful articles can still be found here

Medical Databases

EBSCO Health

Request Information

About EBSCO Health | Products | Who We Serve | Benefits | Success Stories | Contact Us

« All Products

Allied Health / Nursing

- CINAHL Complete »
- CINAHL Core eBook Packages »
- CINAHL Database »
- CINAHL Plus »
- CINAHL Plus with Full Text »
- CINAHL with Full Text »
- DynaMed »
- EDS Discovery Health »
- Health Source: Nursing/Academic Edition »
- Isabel »
- Medcom »
- Nursing & Allied Health Collection: Basic Edition »
- Nursing Reference Center »
- Nursing Reference Center Plus »
- Patient Education Reference Center »
- Rehabilitation Reference Center »
- Social Work Reference Center »

CINAHL Database

Cumulative Index to Nursing and Allied Health Literature

Nurses, allied health professionals, researchers, nurse educators and students depend on the *CINAHL Database* to research their subject area from this authoritative index of nursing and allied health journals which includes over 3 million records dating back to 1960.

Interested in learning more about CINAHL Database?

Request a Free Trial

Easy Access to the Most Authoritative Nursing and Allied Health Literature Available

CINAHL provides indexing of the top nursing and allied health literature available including nursing journals and publications from the National League for Nursing, and the American Nurses' Association. Literature covers a wide range of topics including nursing, biomedicine, health sciences librarianship, alternative/complementary medicine, consumer health and 17 allied health disciplines.

In addition, *CINAHL* provides access to health care books, nursing dissertations, selected

Content Lists

Coverage List:
PDF » | Excel » | HTML »

Content Includes

- More than 3 Million Records
- Indexing for more than 3,000 Journals
- 70 Full-Text Journals

Also Contains

- Author Affiliations
- Searchable Cited References for more than 1,250 journals

Additional Resources:

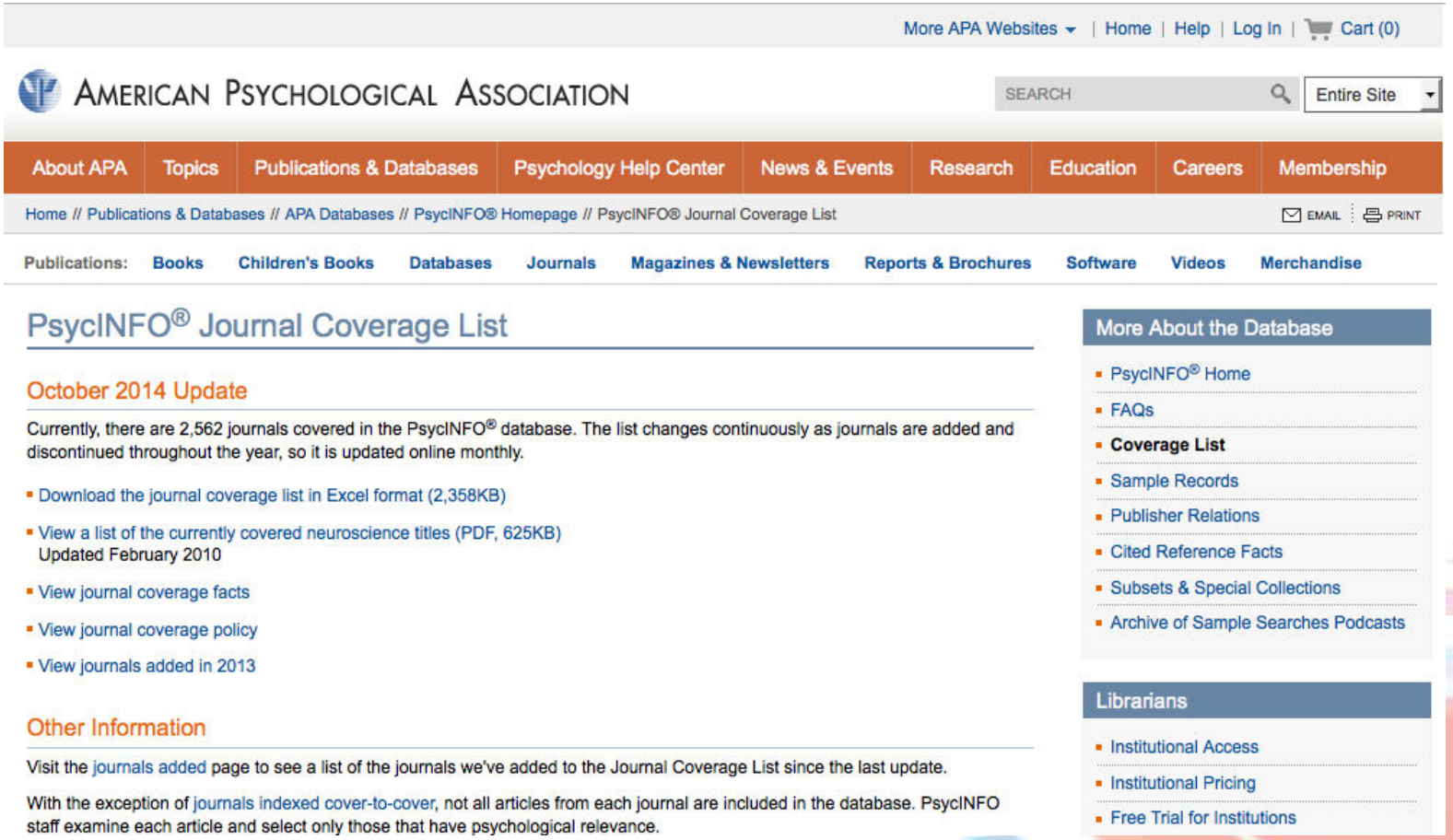
Medical Databases

PsycINFO

- PsycINFO searches the psychological literature
- While it does not primarily focus on medical topics, useful articles can still be found here

http://www.apa.org/pubs/databases/psycinfo/coverage.aspx

Medical Databases



http://www.apa.org/pubs/databases/psycinfo/coverage.aspx

Medical Databases

Web of Science

- Major source for articles in a wide range of fields, including the sciences, social sciences, and humanities.
- Excellent place to find articles from scientific journals that may not be included in MEDLINE

Medical Databases

Conclusions

Informatics in health care may be called as health informatics

- Medical databases deal with the acquisition, storage, retrieval, and use of information in health and biomedicine.

Sequence Submission

Introduction

- Sequences are submitted to the databases in order to share them with the scientific community
- Generally sequences are submitted at the time of publication and are reviewed by peers

Sequence Submission

Caution

- It is important to ensure that sequence files do not contain any special characters
- Control characters in addition to standard ASCII characters must be removed else they might mess up the analysis or data transfer

Sequence Submission

Table 2.1. *Base–nucleic acid codes*

Symbol	Meaning	Explanation
G	G	Guanine
A	A	Adenine
T	T	Thymine
C	C	Cytosine
R	A or G	puRine
Y	C or T	pYrimidine
M	A or C	aMino
K	G or T	Keto
S	C or G	Strong interactions 3 h bonds
W	A or T	Weak interactions 2 h bonds
H	A, C or T not G	H follows G in alphabet
B	C, G or T not A	B follows A in alphabet
V	A, C or G not T (not U)	V follows U in alphabet
D	A, G or T not C	D follows C in alphabet
N	A,C,G or T	Any base

Adapted from NC-IUB (1984).

Mount, pg 28

- Nomenclature committee of International Union of Biochemistry (IUB) has established standard codes to represent ambiguous bases or aminoacids

Sequence Submission

Table 2.2. *Table of standard amino acid code letters*

1-letter code	3-letter code	Amino acid
A[a]	Ala	alanine
C	Cys	cysteine
D	Asp	aspartic acid
E	Glu	glutamic acid
F	Phe	phenylalanine
G	Gly	glycine
H	His	histidine
I	Ile	isoleucine
K	Lys	lysine
L	Leu	leucine
M	Met	methionine
N	Asn	asparagine
P	Pro	proline
Q	Gln	glutamine
R	Arg	arginine
S	Ser	serine
T	Thr	threonine
V	Val	valine
W	Trp	tryptophan
X	Xxx	undetermined amino acid
Y	Tyr	tyrosine
Z[b]	Glx	either glutamic acid or glutamine

Adapted from IUPAC-IUB (1969, 1972, 1983).

[a] Letters not shown are not commonly used.

[b] Note that sometimes when computer programs translate DNA sequences, they will put a "Z" at the end to indicate the termination codon. This character should be deleted from the sequence.

Mount, pg 28

Sequence Submission

NCBI

NCBI has two options

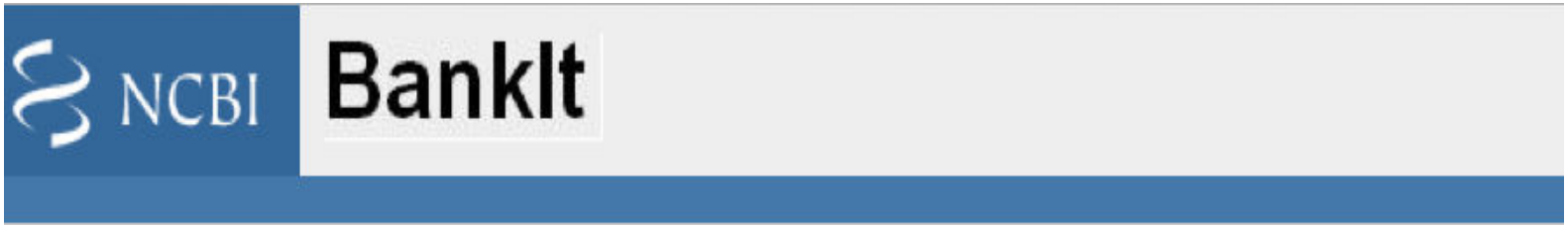
- **BANKIt**
  - For simple sequences and annotations
  - For submission through web
  - Do not require advanced tools

Sequence Submission

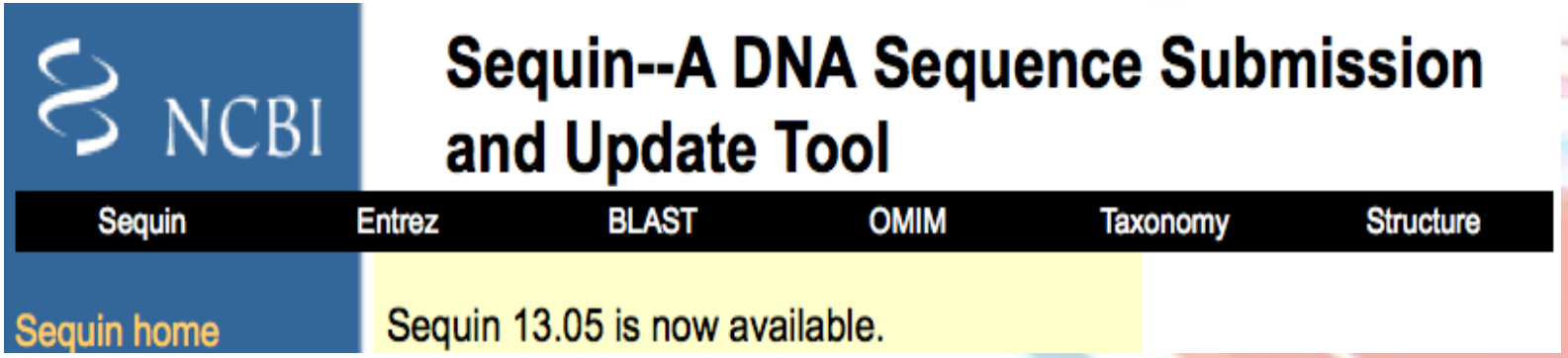
NCBI

- **Sequin**
  - For Complex sequences and annotations
  - For off-line submissions
  - Do require advanced tools and graphical reports

Sequence Submission



http://www.ncbi.nlm.nih.gov/WebSub/?tool=genbank



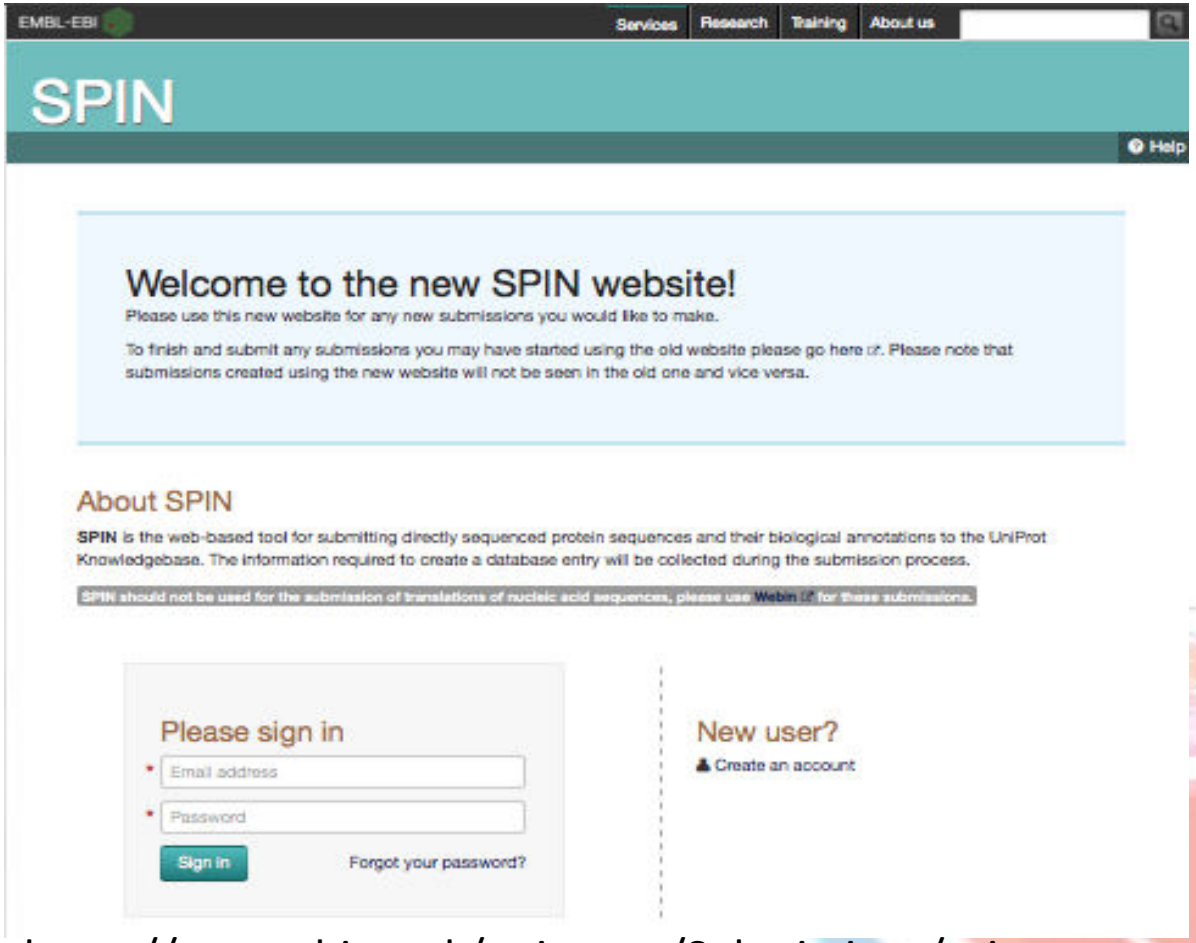
Sequence Submission

UniProt

SPIN web-based tool for submitting

- directly sequenced protein sequences
- biological annotations to the knowledgebase

Sequence Submission

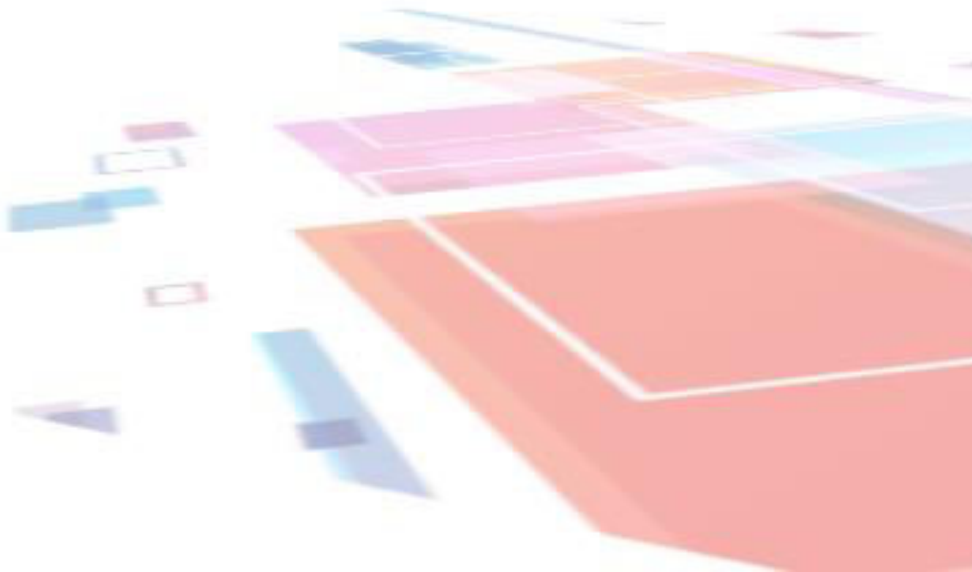


https://www.ebi.ac.uk/swissprot/Submissions/spin

Sequence Submission

Conclusion

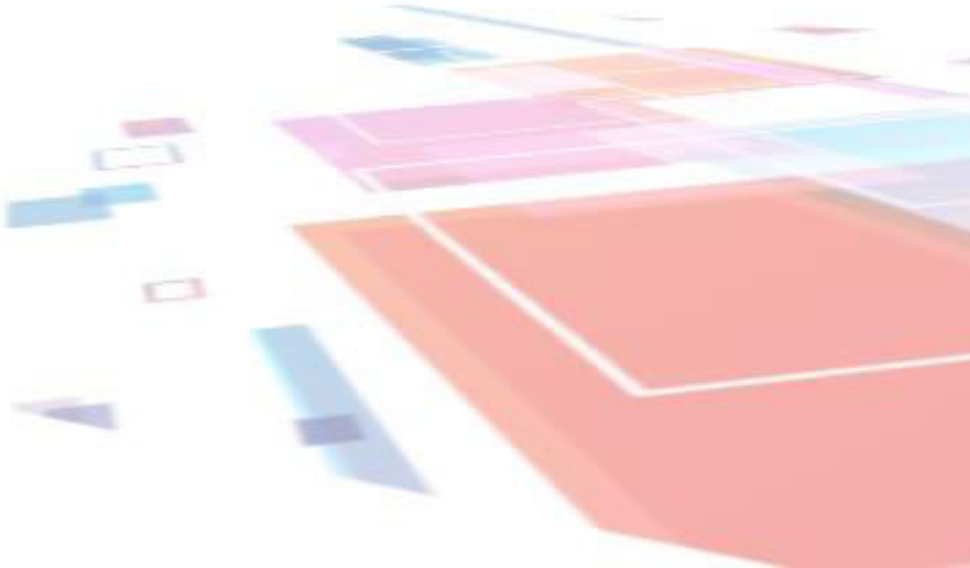
- Sequences are stored in databases in specific format



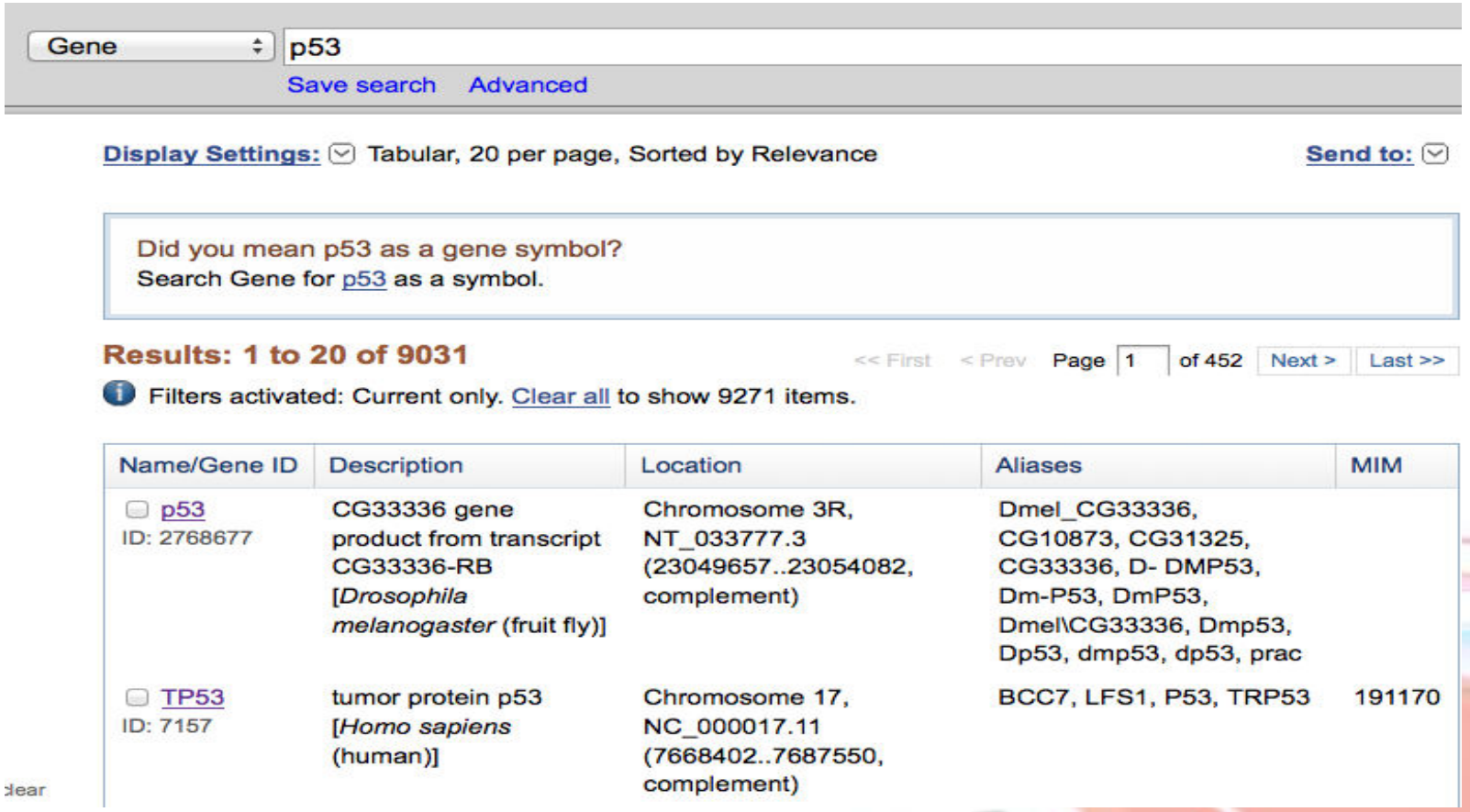
DNA Sequence Retrieval

Introduction

- Databases not merely collect and organize data but allow intelligent data retrieval
- And/or analysis



DNA Sequence Retrieval



http://www.ncbi.nlm.nih.gov/

DNA Sequence Retrieval

p53 [*Drosophila melanogaster* (fruit fly)]

Gene ID: 2768677, updated on 4-Jan-2015

Summary

Official Symbol	p53 provided by FlyBase
Primary source	FLYBASE:FBgn0039044
Locus tag	Dmel_CG33336
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Drosophila melanogaster (old-lineage: Eukaryota; Metazoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha; Ephydroidea; Drosophilidae; Drosophila; Sophophora)
Lineage	Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha; Ephydroidea; Drosophilidae; Drosophila; Sophophora
Also known as	CG10873; CG31325; CG33336; D-p53; Dm-P53; Dmel\CG33336; dmp53; Dmp53; DmP53; DMP53; dp53; Dp53; prac

http://www.ncbi.nlm.nih.gov/

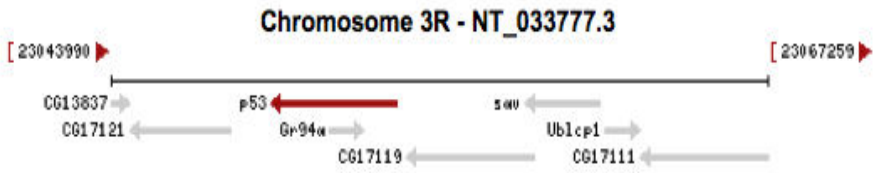
DNA Sequence Retrieval

Genomic context

Location: 94D10-94D10 See p53 in Epigenomics, MapViewer

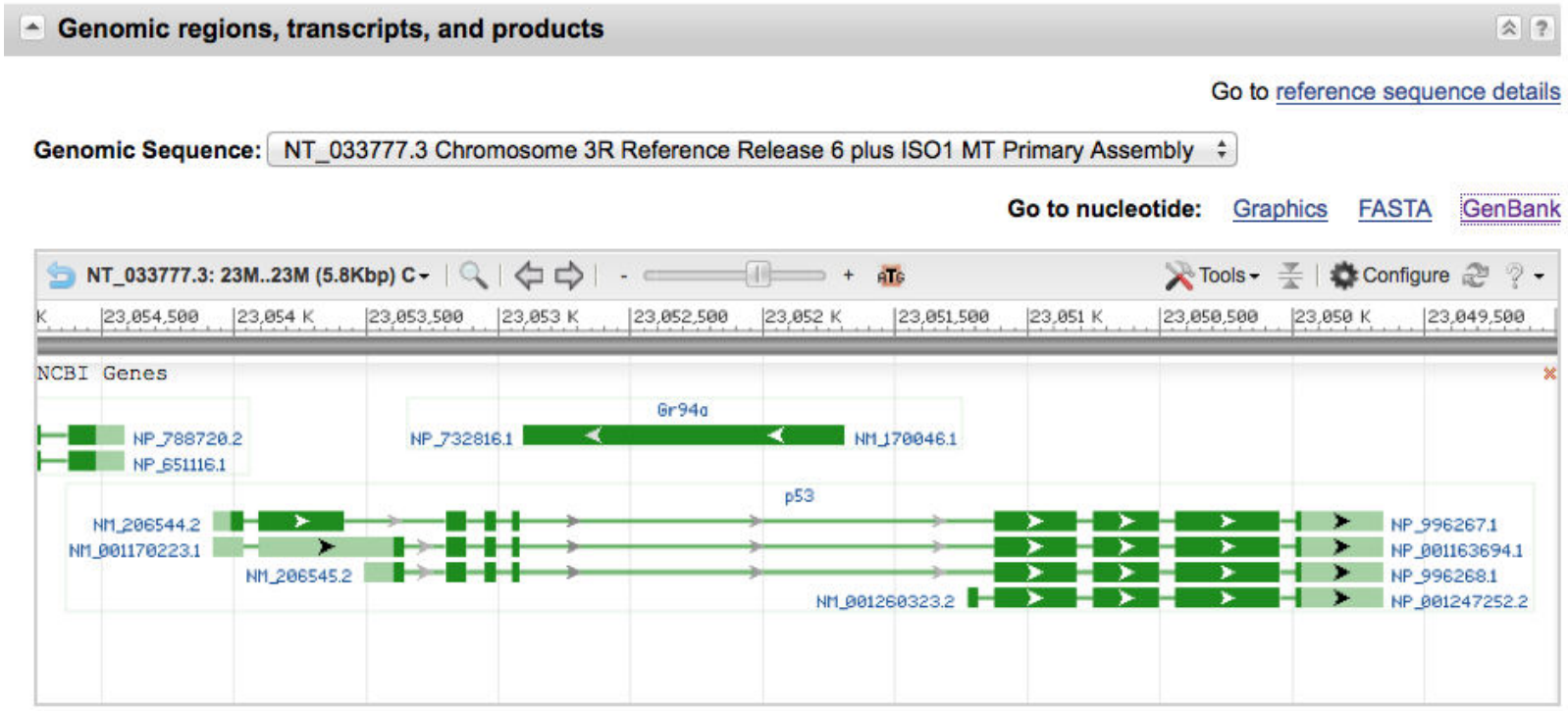
Exon count: 10

Annotation release	Status	Assembly	Chr	Location
Release 6.01	current	Release 6 plus ISO1 MT (GCF_000001215.4)	3R	NT_033777.3 (23049657..23054082, complement)
Release 5.57	previous assembly	Release 5 (GCF_000001215.2)	3R	NT_033777.2 (18875379..18879804, complement)



http://www.ncbi.nlm.nih.gov/

DNA Sequence Retrieval



http://www.ncbi.nlm.nih.gov/

DNA Sequence Retrieval

Drosophila melanogaster chromosome 3R

NCBI Reference Sequence: NT_033777.3

FASTA Graphics

```
LOCUS       NT_033777               4426 bp    DNA     linear   INV 05-AUG-2014
DEFINITION  Drosophila melanogaster chromosome 3R.
ACCESSION   NT_033777 REGION: complement(23049657..23054082)
VERSION     NT_033777.3  GI:671162122
DBLINK      BioProject: PRJNA164
            BioSample: SAMN02803731
KEYWORDS    RefSeq.
SOURCE      Drosophila melanogaster (fruit fly)
  ORGANISM  Drosophila melanogaster
            Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta;
            Pterygota; Neoptera; Endopterygota; Diptera; Brachycera;
            Muscomorpha; Ephydroidea; Drosophilidae; Drosophila; Sophophora.
REFERENCE   1  (bases 1 to 4426)
  AUTHORS   Hoskins,R.A., Carlson,J.W., Kennedy,C., Acevedo,D., Evans-Holm,M.,
            Frise,E., Wan,K.H., Park,S., Mendez-Lago,M., Rossi,F.,
            Villasante,A., Dimitri,P., Karpen,G.H. and Celniker,S.E.
  TITLE     Sequence finishing and mapping of Drosophila melanogaster
            heterochromatin
  JOURNAL   Science 316 (5831), 1625-1628 (2007)
   PUBMED   17569867
```

http://www.ncbi.nlm.nih.gov/

DNA Sequence Retrieval

```
FEATURES             Location/Qualifiers
     source          1..4426
                     /organism="Drosophila melanogaster"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:7227"
                     /chromosome="3R"
                     /genotype="y[1]; Gr22b[1] Gr22d[1] cn[1] CG33964[R4.2]
                     bw[1] sp[1]; LysC[1] MstProx[1] GstD5[1] Rh6[1]"
     gene            1..4426
                     /gene="p53"
                     /locus_tag="Dmel_CG33336"
                     /gene_synonym="CG10873; CG31325; CG33336; D-p53; Dm-P53;
                     Dmel\CG33336; dmp53; Dmp53; DmP53; DMP53; dp53; Dp53;
                     prac"
                     /map="94D10-94D10"
                     /db_xref="FLYBASE:FBgn0039044"
                     /db_xref="GeneID:2768677"
     mRNA            join(1..118,178..501,884..964,1035..1071,1135..1161,
                     2959..3268,3333..3579,3642..4036,4096..4426)
                     /gene="p53"
                     /locus_tag="Dmel_CG33336"
                     /gene_synonym="CG10873; CG31325; CG33336; D-p53; Dm-P53;
                     Dmel\CG33336; dmp53; Dmp53; DmP53; DMP53; dp53; Dp53;
                     prac"
                     /product="p53, transcript variant B"
                     /note="p53-RB; Dmel\p53-RB; CG33336-RB; Dmel\CG33336-RB"
                     /transcript_id="NM_206544.2"
                     /db_xref="GI:281362333"
                     /db_xref="FLYBASE:FBtr0084360"
                     /db_xref="FLYBASE:FBgn0039044"
                     /db_xref="GeneID:2768677"
```

http://www.ncbi.nlm.nih.gov/

DNA Sequence Retrieval

```
CDS             join(75..118,178..501,884..964,1035..1071,1135..1161,
                2959..3268,3333..3579,3642..4036,4096..4118)
                /gene="p53"
                /locus_tag="Dmel_CG33336"
                /gene_synonym="CG10873; CG31325; CG33336; D-p53; Dm-P53;
                Dmel\CG33336; dmp53; Dmp53; DmP53; DMP53; dp53; Dp53;
                prac"
                /note="CG33336 gene product from transcript CG33336-RB;
                CG33336-PB; p53-PB; p53-like regulator of apoptosis and
                cell cycle; Dmp53; protein 53; drosophila p53"
                /codon_start=1
                /product="p53, isoform B"
                /protein_id="NP_996267.1"
                /db_xref="GI:45553461"
                /db_xref="FLYBASE:FBpp0083753"
                /db_xref="FLYBASE:FBgn0039044"
                /db_xref="GeneID:2768677"
                /translation="MSLHKSASFSLTFNQNTSIVSRSNSRTIFEAFKEFLDFWDIGNE
                VSAESAVRVSSNGAFNLPQSFGNESNEYAHLATPVDPAYGGNNTNNMMQFTNNLEILA
                NNNSDGNNKINACNKFVCHKGTDSEDDSTEVDIKEDIPKTVEVSGSELTTEPMAFLQG
                LNSGNLMQFSQQSVLREMMLQDIQIQANTLPKLENHNIGGYCFSMVLDEPPKSLWMYS
                IPLNKLYIRMNKAFNVDVQFKSKMPIQPLNLRVFLCFSNDVSAPVVRCQNHLSVEPLT
                ANNAKMRESLLRSENPNSVYCGNAQGKGISERFSVVVPLNMSRSVTRSGLTRQTLAFK
                FVCQNSCIGRKETSLVFCLEKACGDIVGQHVIHVKICTCPKRDRIQDERQLNSKKRKS
                VPEAAEEDEPSKVRRCIAIKTEDTESNDSRDCDDSAAEWNVSRTPDGDYRLAITCPNK
                EWLLQSIEGMIKEAAAEVLRNPNQENLRRHANKLLSLKKRAYELP"
```

http://www.ncbi.nlm.nih.gov/

DNA Sequence Retrieval

```
ORIGIN
        1 cctggagcac ggaagattct tgcggacaca aatcgcaact gctaaataaa atttatttat
       61 ttgagtgcac agccatgagt cttcacaagt ccgcgtcgtt tagcttgact tttaaccagt
      121 gagcggagat attttattcg gtcttaccca acaaaataat gttgcgcctt tttgcagaaa
      181 cacttcgatt gtttcgcgta gcaatagtcg cacaattttt gaagctttca aggagttcct
      241 ggatttttgg gatatcggca acgaagtttc tgcagagtca gcagttcggg tctccagcaa
      301 cggagctttc aacttgccgc agagttttgg caacgaatcc aacgaatatg cccacctggc
      361 tacgcctgtg gatccagcct acggaggcaa caacacgaac aacatgatgc agttcacgaa
      421 caatctggaa attttggcca acaataattc cgatggcaat aacaaaatta atgcatgcaa
      481 caaattcgtc tgccacaagg ggtgagcaaa ttcaaaacac gcgctccaat cgataaacat
      541 tggctacggc gattgttcgc gctgcgtggc gaatggcaaa atccaaatag tcggtggcca
      601 ctacgattct gtagtttttt gttagcgaat ttttaatatt tagcctcctt ccccaacaag
      661 atcgcttgat cagatatagc cgactaagat gtatatatca cagccaatgt cgtggcacaa
      721 agaaaggtac agtgcggcaa caaattgatg atcgaacagt agaaaccttg catgtagcaa
       ------------------------------------
     4261 ggcatgttcg atggccgaaa agaaaacatt tttatatttt tgatagtata ctgttgttaa
     4321 ctgcagttct atgtgactac gtaacttttg tctaccacaa caaacatact ctgtacaaaa
     4381 aagccaaaag tgaatttatt aaagagttgt catattttgc aaacat
//
```

http://www.ncbi.nlm.nih.gov/

DNA Sequence Retrieval

Conclusions

- DNA Sequences are stored in DNA sequence databases in specified formats
- Genebank format is a standard format

Sequence Formats

FASTA Sequence Format

- DNA sequences are stored in specified formats
- Different softwares need sequences in different formats

Sequence Formats

FASTA Sequence Format

- FASTA is the most frequently used format to present DNA and Protein sequences
- It is recommended that all lines of text be shorter than 80 characters in length

Sequence Formats

>gi|568815581:c7687550-7668402 Homo sapiens chromosome 17, GRCh38 Primary Assembly
GATGGGATTGGGGTTTTCCCCTCCCATGTGCT
CATCTAGAGCCACCGTCCAGGGAGCAGGTAGC
TGCTGGGCTCTCCACGACGGTGACACGC--------

>gi|120407068|ref|NP_000537.3| cellular tumor antigen p53 isoform a [Homo sapiens]
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLS
PLAPPVLGFLHSGTAKSVTCTYSPALNKMFCQLA
KT--*

Sometimes a * might be present in the end

Sequence Formats

Genebank Format

- A sequence file in GenBank format can contain several sequences
- One sequence starts with a line containing the word LOCUS and a number of annotation lines

Sequence Formats

Genebank Format

- The start of the sequence is marked by a line containing "ORIGIN" and the end of the sequence is marked by two slashes ("//")

Sequence Formats

Genebank Format

LOCUS AAU03518 237 bp DNA PRI 04-FEB-1995
DEFINITION Aspergillus awamori internal transcribed spacer 1 (ITS1) and 18S rRNA and 5.8S rRNA genes, partial sequence.
ACCESSION U03518
BASE COUNT 41 a 77 c 67 g 52 t
ORIGIN
1 aacctgcgga aggatcatta ccgagtgcgg gtcctttggg cccaacctcc catccgtgtc
61 tattgtaccc tgttgcttcg gcgggcccgc cgcttgtcgg ccgccggggg ggcgcctctg
121 ccccccgggc ccgtgcccgc cggagacccc aacacgaaca ctgtctgaaa gcgtgcagtc
181 tgagttgatt gaatgcaatc agttaaaact ttcaacaatg gatctcttgg ttccggc
//

Sequence Formats

EMBLFormat

- An example sequence in EMBL format is:

```
ID   AA03518    standard; DNA; FUN; 237 BP.
XX
AC   U03518;
XX
DE   Aspergillus awamori internal transcribed spacer 1 (ITS1) and 18S
DE   rRNA and 5.8S rRNA genes, partial sequence.
XX
SQ   Sequence 237 BP; 41 A; 77 C; 67 G; 52 T; 0 other;
     aacctgcgga aggatcatta ccgagtgcgg gtcctttggg cccaacctcc catccgtgtc        60
     tattgtaccc tgttgcttcg gcgggcccgc cgcttgtcgg ccgccggggg ggcgcctctg       120
     ccccccgggc ccgtgcccgc cggagacccc aacacgaaca ctgtctgaaa gcgtgcagtc       180
     tgagttgatt gaatgcaatc agttaaaact ttcaacaatg gatctcttgg ttccggc          237
//
```

Sequence Formats

SwissProt Format

- SwissProt protein sequence format is similar to EMBL format
- There is considerably more information about physical and biochemical properties of a protein is provided

Sequence Formats

SwissProt Format

ID - Identification.

AC - Accession number(s).
DT - Date.
DE - Description.
GN - Gene name(s).
OS - Organism species.
OG - Organelle.
OC - Organism classification.
RN - Reference number.

RP - Reference position.
RC - Reference comments.
RX - Reference cross-references.
RA - Reference authors.
RL - Reference location.
CC - Comments or notes.
DR - Database cross-references.
KW - Keywords.
FT - Feature table data.
SQ - Sequence header.
// - Termination line.

Sequence Formats

XML Format

- Extensible Markup Language
- Readable by both man and machine
- Becoming standard data format for transferring genome data

Sequence Formats

XML Format

```
<xsd:annotation>
    <xsd:documentation>
      XML Schema for SBOL core data model
compatible with RDF/XML serialization.
      <dc:date>2012-01-19</dc:date>
      <dc:creator>Evren Sirin</dc:creator>
      <dc:contributor>Michal
Galdzicki</dc:contributor>
    </xsd:documentation>
  </xsd:annotation>
```

Sequence Formats

Sequence converters

- READSEQ is a useful sequence converter developed by D.G.Gilbert at Indiana University, USA
- Recognizes DNA or Protein sequence file and interconvert them

Sequence Formats

Conclusions

- Databases store sequences in specified formats
- Genebank, DDBJ and EMBL has similar formats
- Different softwares need sequences in different formats

Data Retrieval

Data Retrieval

- Nearly all biological databases are available for download as simple text (flat) files
- A local version of the database allows one greater freedom in processing the data

Data Retrieval

Entrez

- is an integrated search engine which allows users to search and retrieve different data from the NCBI
- It can be accessed from the site www.ncbi.nlm.nih.gov/Entrez/

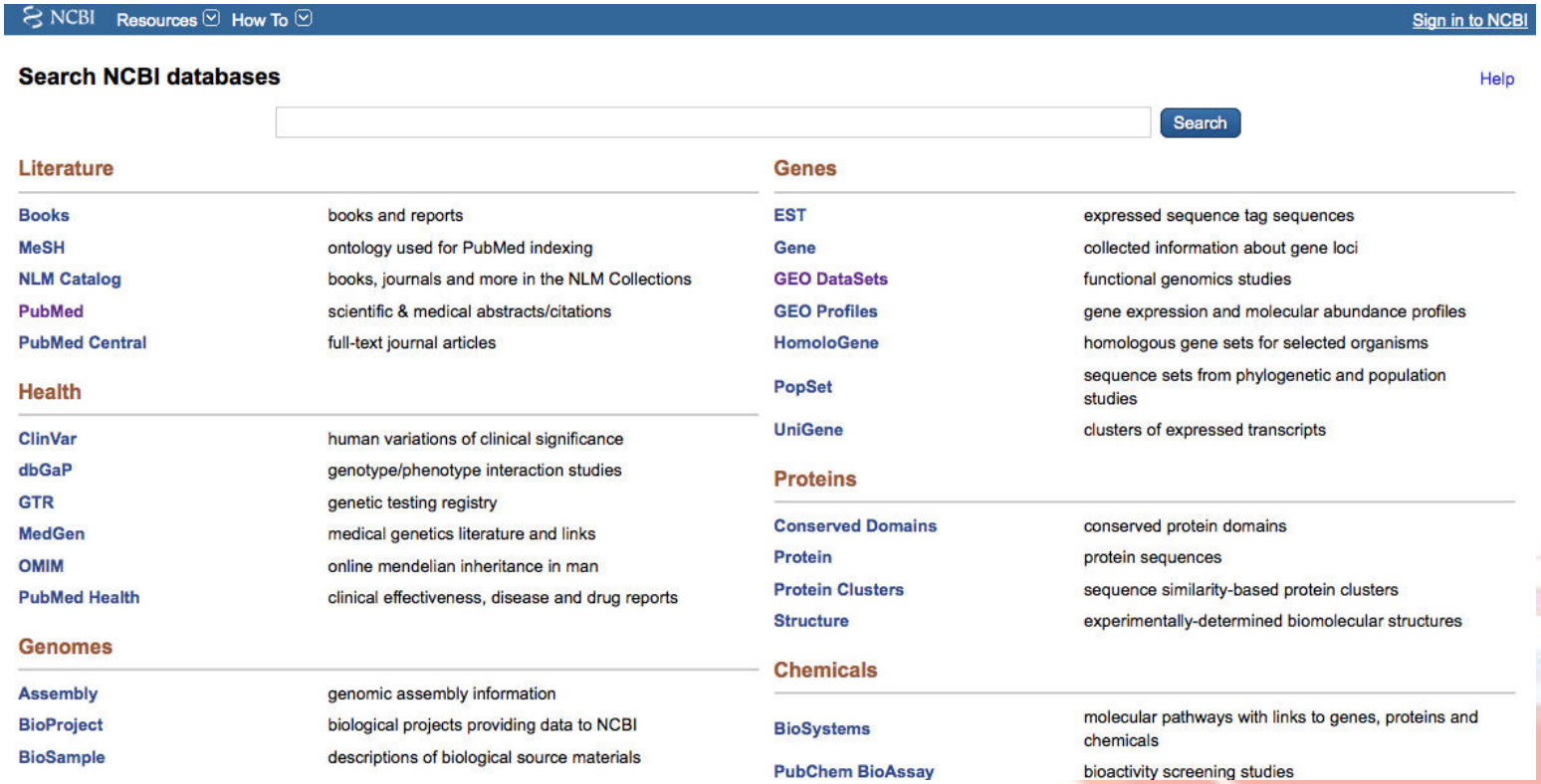
Data Retrieval

Entrez

- integrates PubMed and 39 other scientific literatures, nucleotide and protein databases
- protein domain data, population studies, expression data, pathways, genome details and taxonomic information

Data Retrieval

Data Retrieval



Data Retrieval

Bulk Data Retrieval

- The best option is to use ftp (File transfer protocol)
- The File Transfer Protocol (**FTP**) is a standard network protocol used to transfer files
- Via command line or application programs like FTP clients

Data Retrieval

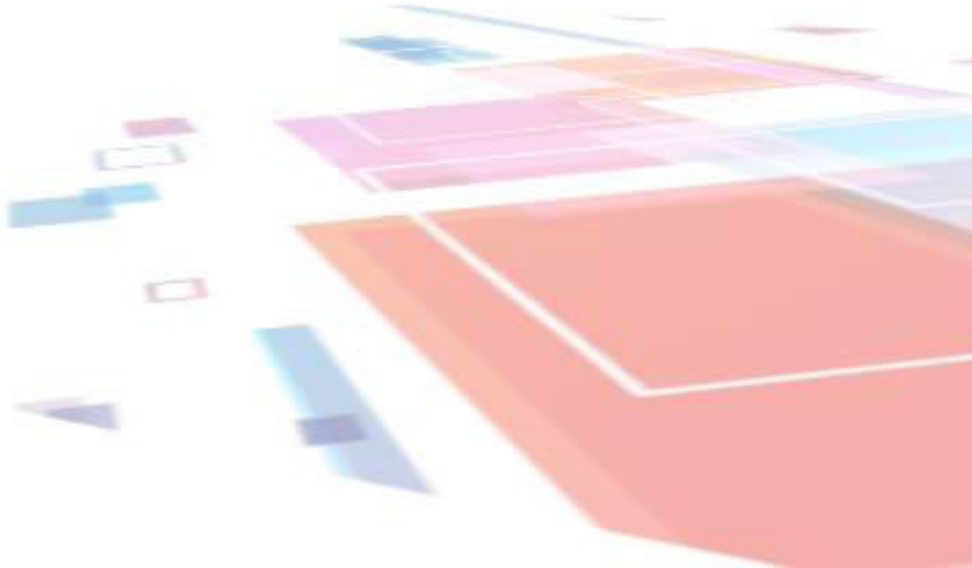
Bulk Data Retrieval

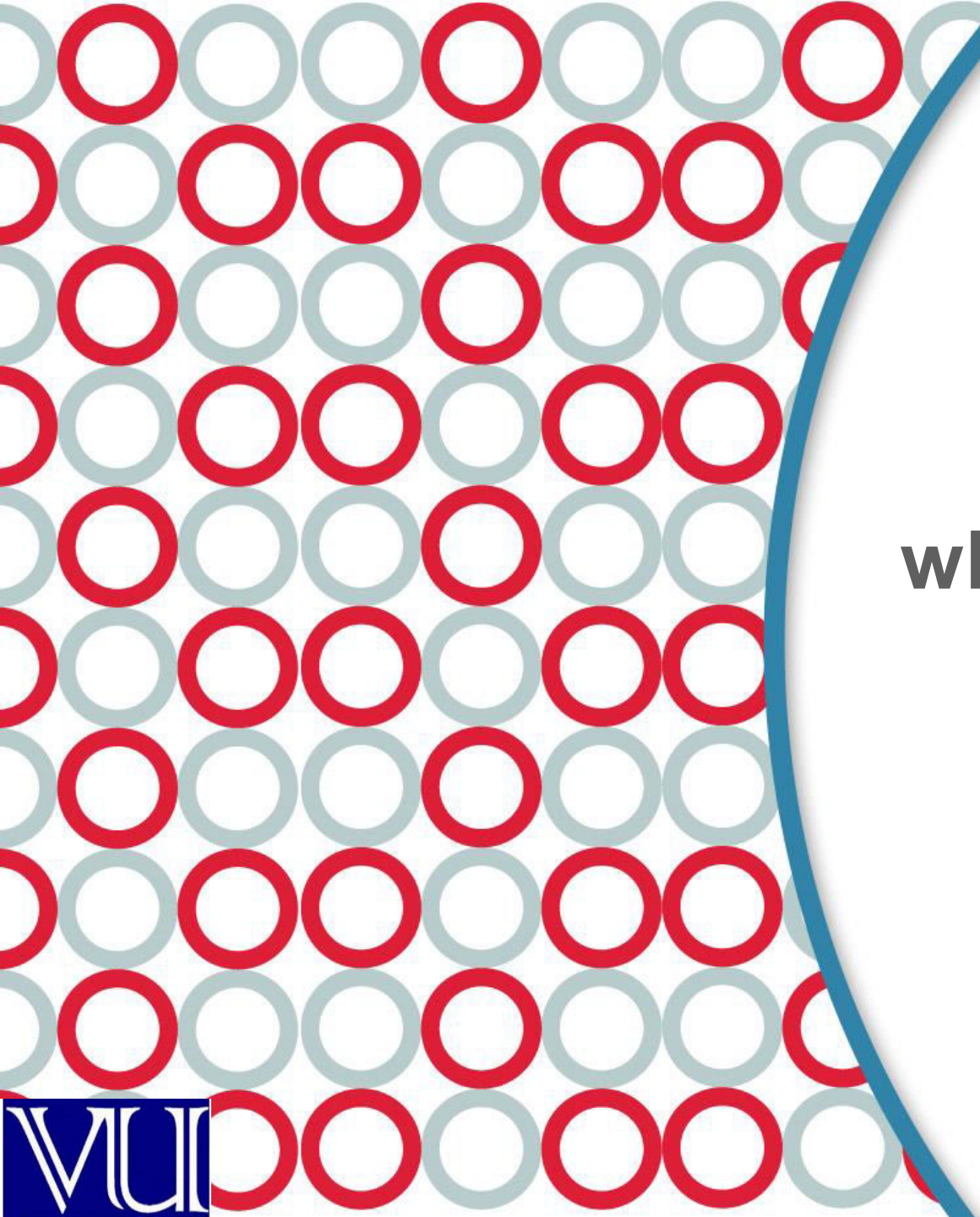
- Data needs to be transformed or processed using programming languages
- PERL and Python are good for processing Biological data

Data Retrieval

Conclusions

- Data is transferred over the internet
- Data needs to be transformed or processed



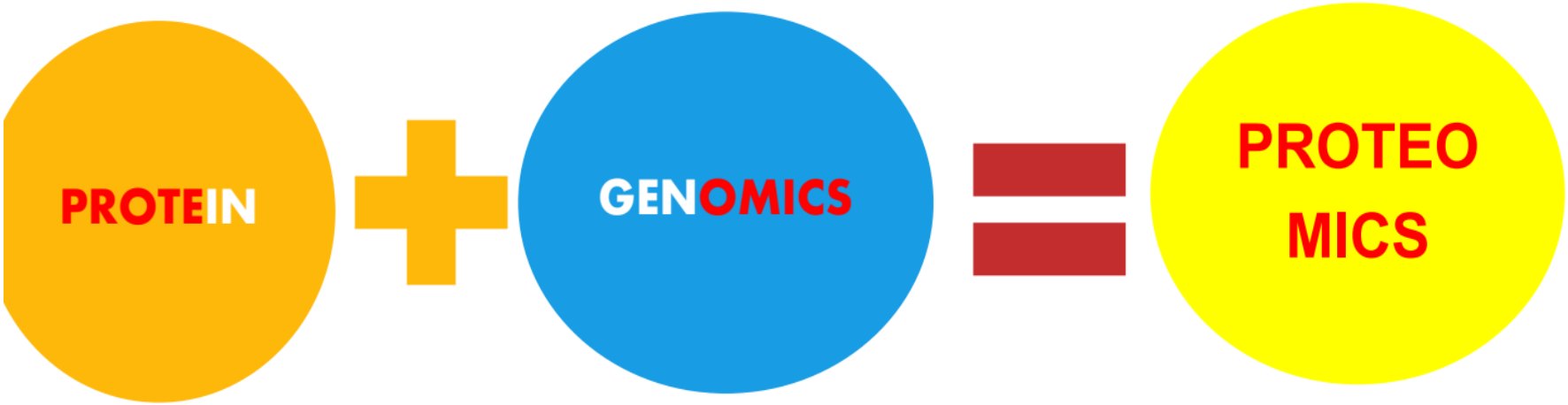


why Proteomics

Lecture-2

What do mean by proteomics?

- Proteomics is the large-scale study of proteins, usually by biochemical methods. The word proteomics has been associated
- traditionally with displaying a large number of proteins from a given
- cell line or organism on two -dimensional polyacrylamide gels



Scope of Proteomics

The identification, characterization and quantification of all proteins involved in a particular pathway, organelle, cell, tissue, organ or organism that can be studied in concert to provide accurate and comprehensive data about that system.

Or - A complete description of proteins expressed in any given cell at any given time

Why should we study Proteomics?

- directly contributes to drug development
- Verification of a gene product by proteomic methods
- Modifications of the proteins
- Protein expression level d

WHY PROTEOMICS?

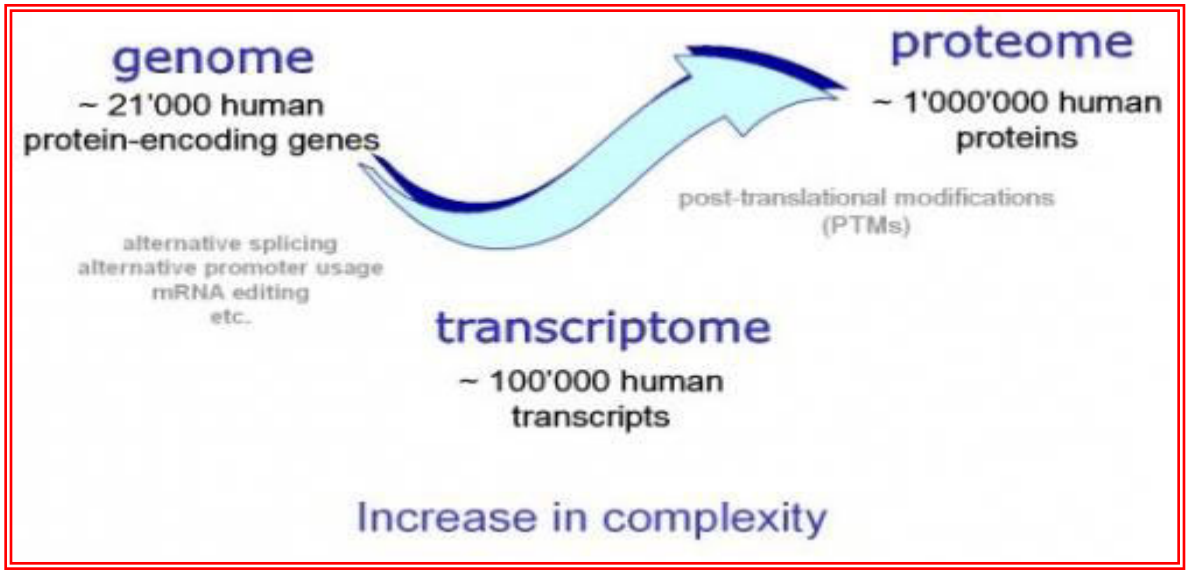
Many types of information cannot be obtained from the study of genes alone. For example, proteins, not genes, are responsible for the phenotypes of cells. It is impossible to elucidate mechanisms of disease, aging, and effects of the environment solely by studying the genome.

Advantages of study of proteomics

- Shows that genetic alterations are not the reason for all types of diseases
- Helps in determining the proper treatment of diseases
- With the help of three dimensional analysis of proteins we have found that HIV protease is the enzyme which is responsible for AIDS.
- One of the most important use of proteomics in diagnosis is the identification of biomarkers. The study of drugs in proteomics is called pharmacoproteomics.

Proteomics aims

- Genomics integrated strategies
- Study of post-translational modifications
- Identification of novel protein targets for drugs
- Analysis of tumor tissues
- Comparison between normal and diseased tissues
- Comparison between diseased and pharmacologically treated tissues


genome
~ 21'000 human
protein-encoding genes
proteome
~ 1'000'000 human
proteins
post-translational modifications
(PTMs)
alternative splicing
alternative promoter usage
mRNA editing
etc.
transcriptome
~ 100'000 human
transcripts
Increase in complexity

Limitations of Genomics Challenge of Proteomics

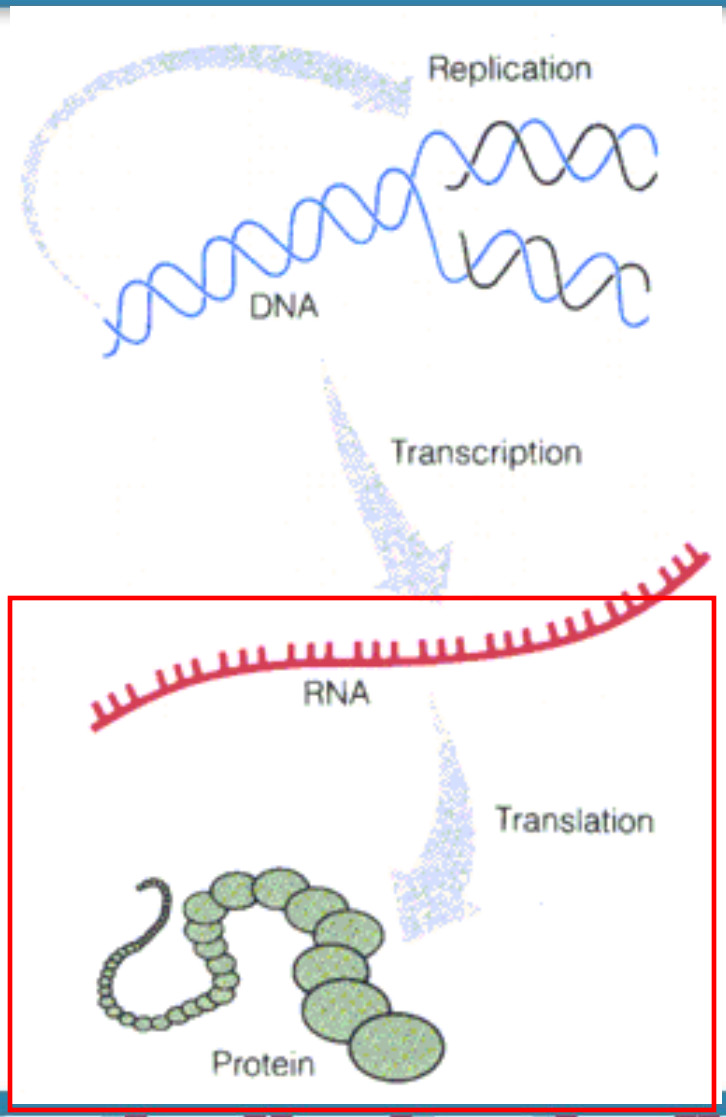
- co-translational modifications
  - differential mRNA splicing
- post-translational modifications (PTMs)
  - C-terminal GPI anchor
  - phosphorylation
  - sulfation
  - glycosylation
  - N-myristoylation
  - hydroxylation
  - N-methylation
  - carboxymethylation
  - signal peptidase site.....

Proteomics Introducción

BIO 601

Lecture-1

The Central Dogma



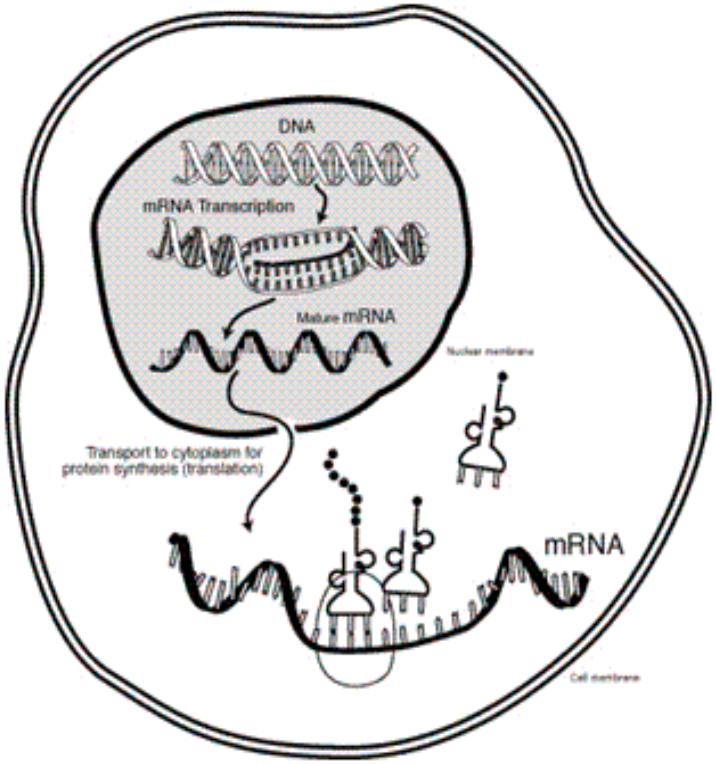
The central dogma states that information in nucleic acid can be perpetuated or transferred, but the transfer of information form into protein is irreversible.

Central Dogma of Biology

- DNA -> RNA -> Protein Synthesis
- Transcription:
  - Process of DNA serving as a template for RNA synthesis
- Translation:
  - Process of RNA serving as a template for protein synthesis

How do DNA and genes relate to proteins?

- DNA provides the genes, or genetic code, for protein synthesis
- Genes are expressed because DNA codes for RNA which then codes for ALL of our proteins



Background: 3 Types of RNA

- mRNA: Messenger RNA
  - 1st RNA's made DIRECTLY from DNA template
  - Travel from nucleus to ribosome
- rRNA: Ribosomal RNA
  - Helps form ribosomes in cytoplasm
- tRNA: Transfer RNA
  - Brings amino acids from cytoplasm to ribosome so proteins can be made

Step 1: Transcription

- INSIDE of the nucleus DNA is used to make mRNA
- DNA is unzipped then RNA polymerase makes an mRNA strand from the DNA template
- New mRNA strand then leaves the nucleus and travels into the cytoplasm
- DNA is ALWAYS left protected in the nucleus

Step 1: Transcription

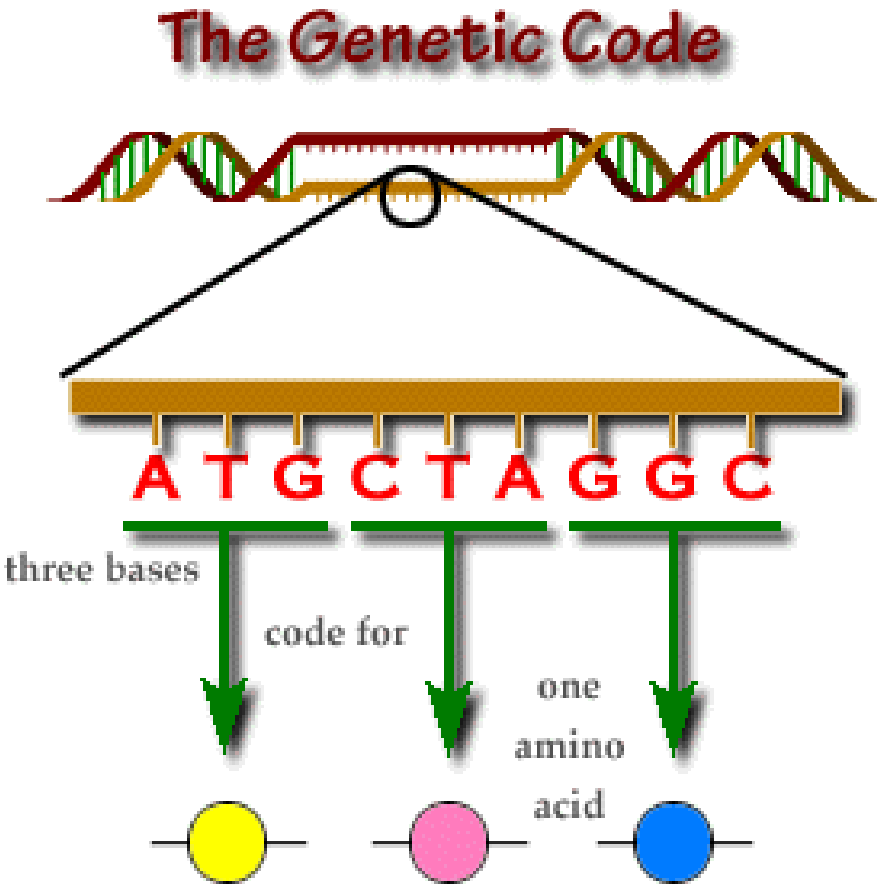
- DNA: 5’ AAA TTT GGG CCC ATC GCA 3’
- mRNA: 3’ UUU AAA CCC GGG UAG CGU 5’

- DNA: CTA GTT CCC TAA AAG GAG
- mRNA: GAU CAA GGG AUU UUC CUC

- DNA: TAC CGA GGT TTA ACT
- mRNA: AUG UGA CCA AAU UGA

Step 2: Translation

- Each nucleotide sequence serves as a code for what amino acid will be added to the protein being made
- Nucleotides read in triplets, or codons



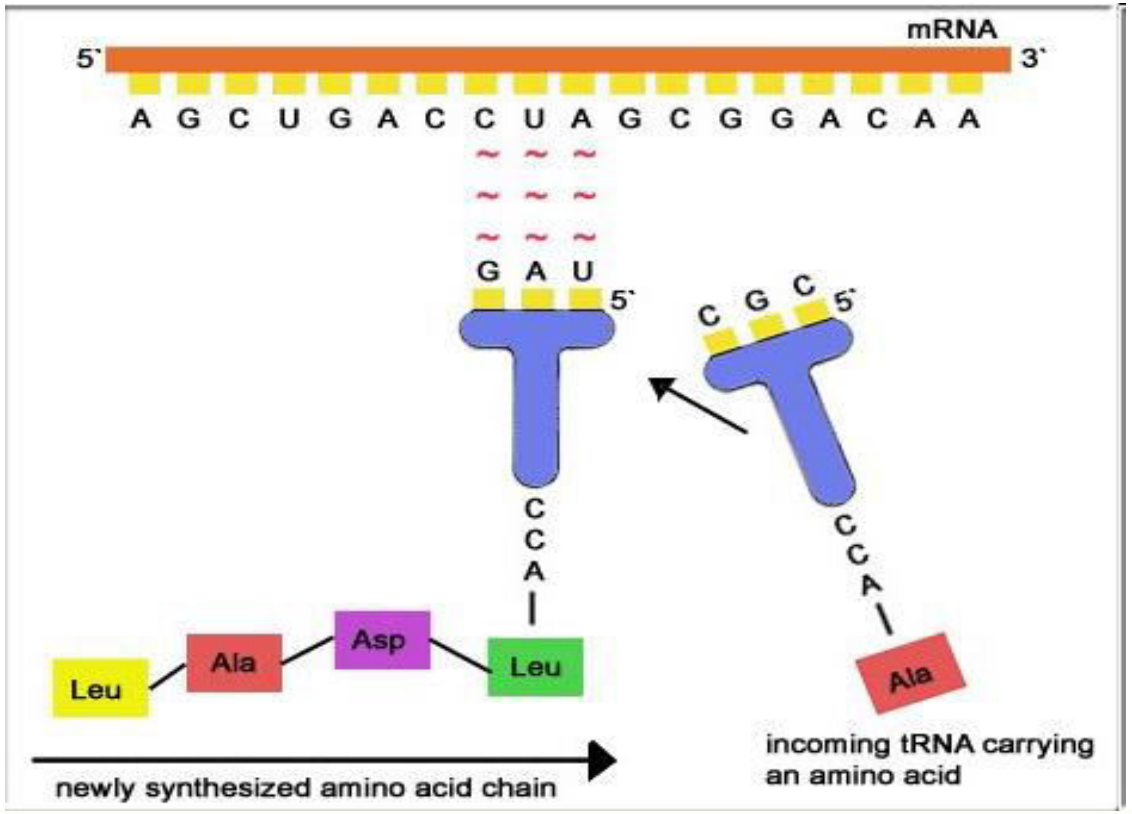
Step 2: Translation

Second base in codon

First base in codon	U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Step 2: Translation

- mRNA is now connected to the ribosome
- tRNA has a corresponding anti-codon and brings over the corresponding amino acid



The end result…

- An amino acid sequence that makes a protein
- GENES code for proteins/enzymes
- We NEED proteins to function
- The shape of the protein determines its function

TYPES OF PROTEOMICS

Lecture-3

Scope of Proteomics

- Expression proteomics
- Structural proteomics
- Functional proteomics

EXPRESSION PROTEOMICS

- Expression proteomics is used to study the qualitative and quantitative expression of total proteins under two different conditions.
  - Normal and diseased state.
  - E.g. :tumor or normal cell.
  - It studied that protein is over expressed or under expressed.
  - 2-D electrophorasis.

STRUCTURAL PROTEOMICS

- Structural proteomics helps to understand three dimensional shape and structural complexities of functional proteins.
- It determine either by amino acid sequence in protei from a gene this process is known as **homology modeling**.
- It identify all the protein present in complex system protein-protein interaction.
- Mass spectroscopy is used for structure determinatio

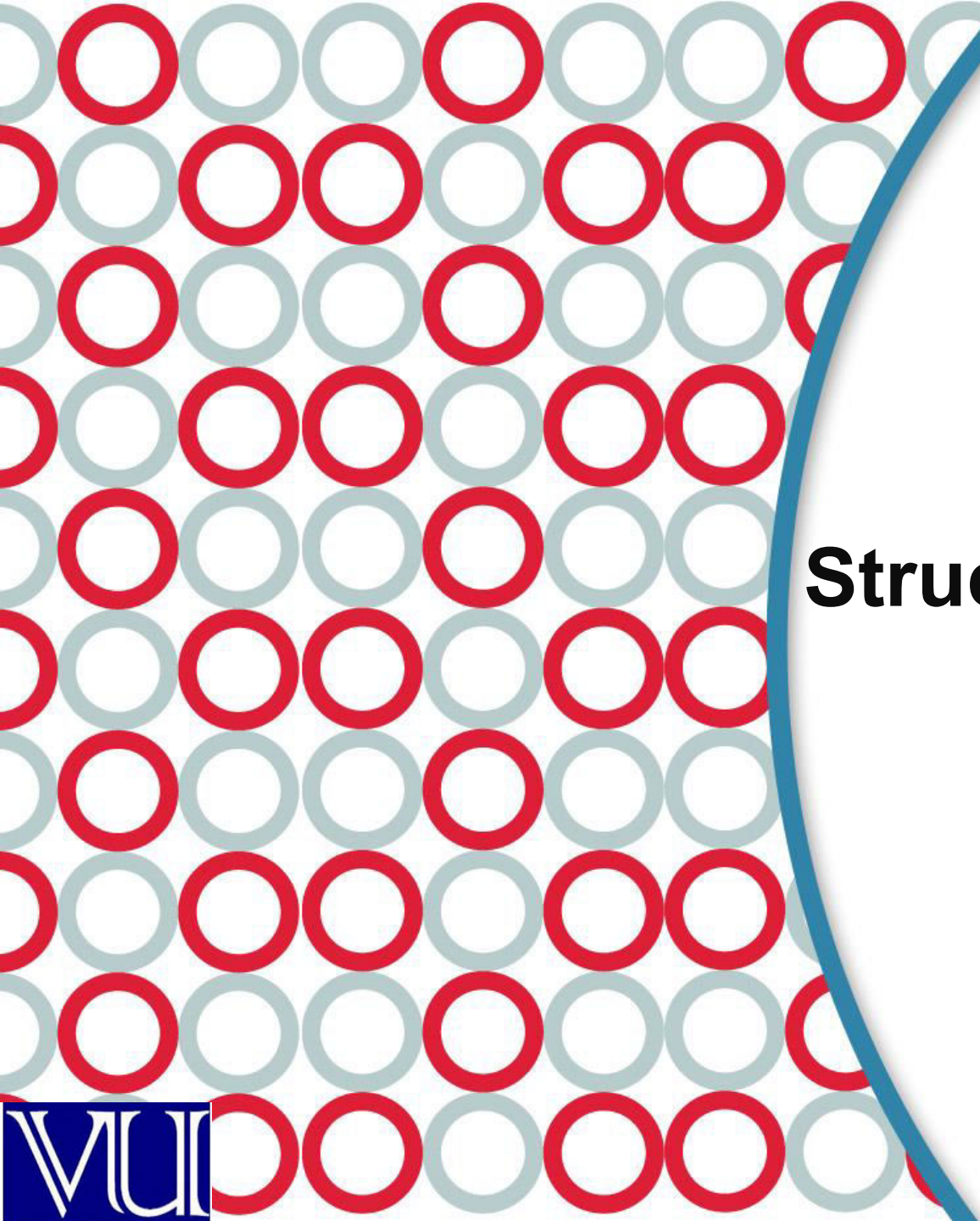
FUNCTIONAL PROTEOMICS

 Functional proteomics explains understanding the protein functions well as unrevealing molecular mechanisms within the cell that depend on identification of the interacting protein partners.

 So that detailed description of the cellular signaling pathways mi greatly benefit from the elucidation of protein- protein interacti

Limitations of Genomics Challenge of Proteomics

- co-translational modifications
  - differential mRNA splicing
- post-translational modifications (PTMs)
  - C-terminal GPI anchor
  - phosphorylation
  - sulfation
  - glycosylation
  - N-myristoylation
  - hydroxylation
  - N-methylation
  - carboxymethylation
  - signal peptidase site.....



Structural Proteomics

Lecture-1

What is structural proteomics/genomics?

- High-throughput determination of the 3D structure of proteins
- Goal: to be able to determine or predict the structure of every protein.
  - Direct determination - X-ray crystallography and nuclear magnetic resonance (NMR).
  - Prediction
    - Comparative modeling -
    - Threading/Fold recognition
    - Ab initio

What is structural proteomics/genomics?

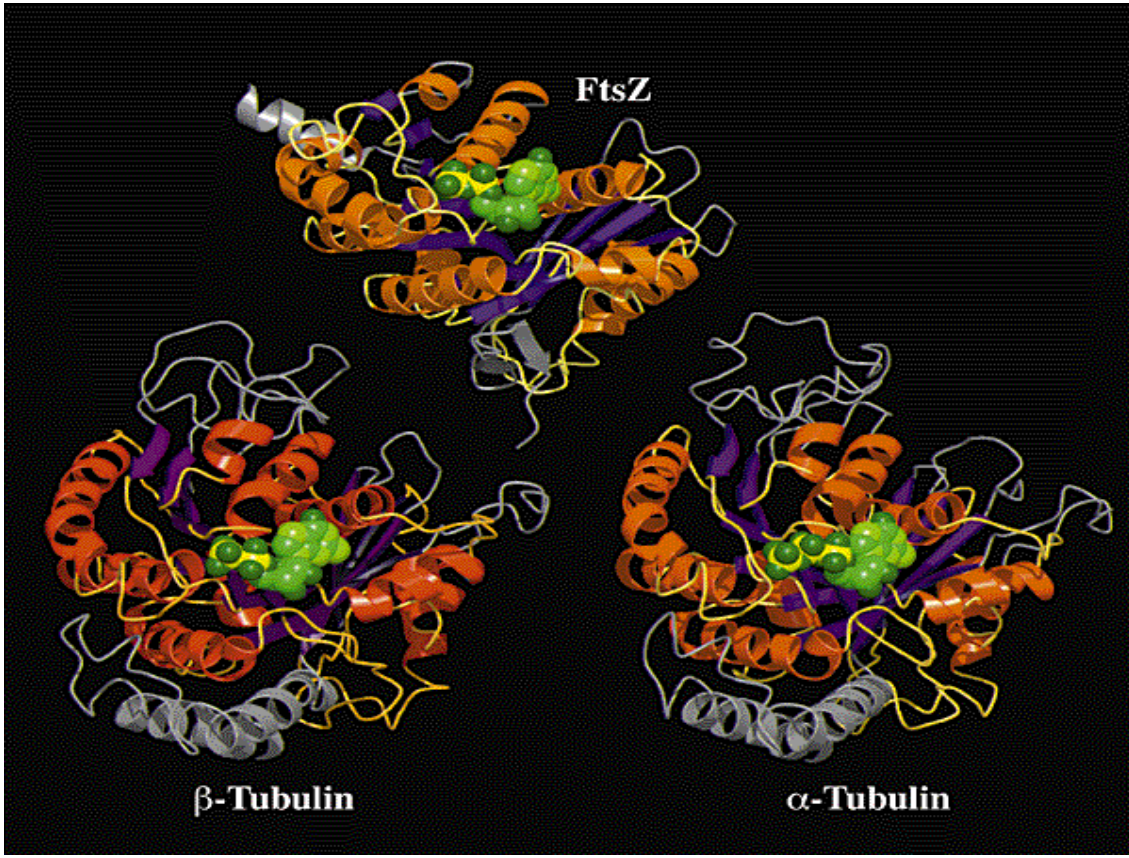
- High-throughput determination of the 3D structure of proteins
- Goal: to be able to determine or predict the structure of every protein.
  - Direct determination - X-ray crystallography and nuclear magentic resonance (NMR).
  - Prediction
    - Comparative modeling -
    - Threading/Fold recognition
    - Ab initio

Why structural proteomics?

- To study proteins in their active conformation.
  - Study protein:drug interactions
  - Protein engineering
- Proteins that show little or no similarity at the primary sequence level can have strikingly similar structures.

An example

- FtsZ - protein required for cell division in prokaryotes, mitochondria, and chloroplasts.
- Tubulin - structural component of microtubules - important for intracellular trafficking and cell division.
- FtsZ and Tubulin have limited sequence similarity and would not be identified as homologous proteins by sequence analysis.



FtsZ and tubulin have little similarity at the amino acid sequence level

Burns, R., Nature **391**:121-123
Picture from E. Nogales

Are FtsZ and tubulin homologous?

- Yes! Proteins that have conserved secondary structure can be derived from a common ancestor even if the primary sequence has diverged to the point that no similarity is detected.

Current state of structural proteomics

- As of Feb. 2002 - 16,500 structures
  - Only 1600 non-redundant structures
- To identify all possible folds - predicted another 16,000 novel sequences needed for 90% coverage.
  - Of the 2300 structures deposited in 2000, only 11% contained previously unidentified folds.
- Overall goal - directly solve enough structures directly to be able to computationally model all future proteins.

Protein domains - structure

- “clearly recognizable portion of a protein that folds into a defined structure”
  - Doesn’t have to be the same as the domains we have been investigating with CDD.
  - RbsB proteins as an example.

Main secondary structure elements

α-helix - right handed helical structure

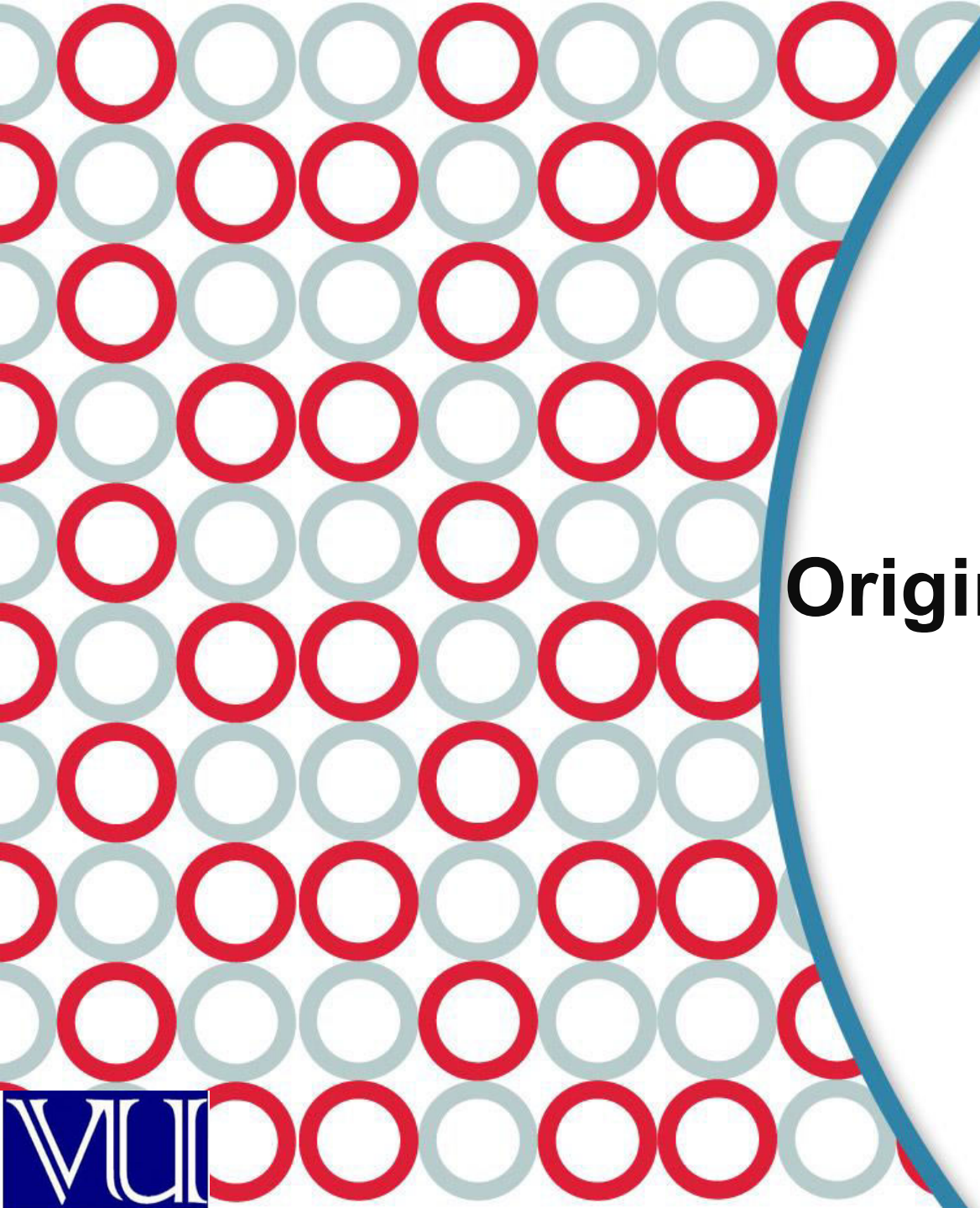
β-sheet - composed of two or more β-strands, conformation is more “zig-zag” than helical.

Folds/motifs

- How these secondary structure elements come together to form structure.
  - Helix-turn-helix
- Determining the structure of nearly all folds is the goal of structural biology

X-Ray Crystallography

- Make crystals of your protein
  - 0.3-1.0mm in size
  - Proteins must be in an ordered, repeating pattern.
- X-ray beam is aimed at crystal and data is collected.
- Structure is determined from the diffraction data.



Origin of proteomics

Lecture-1

Proteomics Origins

- In 1975, the introduction of the 2D gel by O'Farrell who began mapping proteins from *E. coli.*
- Although many proteins could be separated and visualized, they could not be identified.

Human protein index

- Despite these limitations, shortly thereafter a large-scale analysis of all human proteins was proposed.
- The goal of this project, termed the human protein index, was to use two-dimensional protein electrophoresis (2-DE) and other methods to catalog all human proteins.
- However, lack of funding and technical limitations prevented this project from continuing.

Proteomics Origins

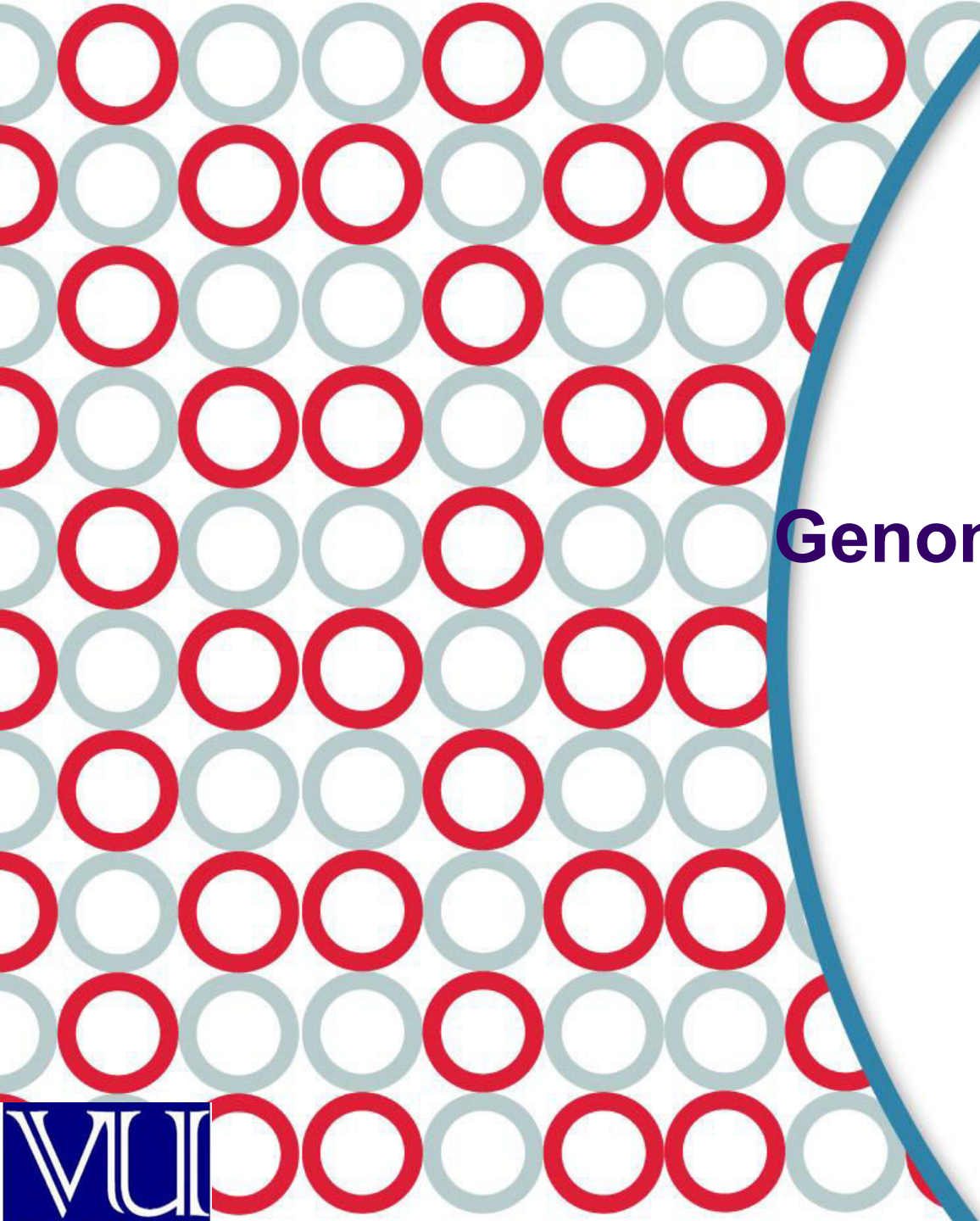
- The first major technology to emerge for the identification of proteins was the sequencing of proteins by Edman degradation.
- A major breakthrough was the development of microsequencing techniques for electroblotted proteins.

Proteomics Origins

- Microsequencing technique was used for the identification of proteins from 2-D gels to create the first 2-D databases.
- Improvements in microsequencing technology resulted in increased sensitivity of Edman sequencing in the 1990s to high picomole amounts.

Proteomics Origins

- One of the most important developments in protein identification has been the development of MS technology.
- The sensitivity of analysis and accuracy of results for protein identification by MS have increased by several orders of magnitude.
- It is now estimated that proteins in the femtomolar range can be identified in gels.
- Because MS is more sensitive, can tolerate protein mixtures, and is amenable to high-throughput operations



Genomics vs. proteomics

Lecture-1

Genomics vs. proteomics

- Genomics and proteomics are closely-related fields.
- The main difference between genomics and proteomics is that genomics is the study of the entire set of genes in the genome of a cell whereas proteomics is the study of the entire set of proteins produced by the cell.

Nature of study material

Genomics: The genome is constant. Every cell of an organism has the same set of genes.

Proteomics: The proteome is dynamic and varies. The set of proteins produced in different tissues varies according to the gene expression.

Use of High throughput techniques

Genomics: High throughput techniques are used in the genomics to map, sequence, and analyze genomes.

Proteomics: In proteomics, characterization of the 3D structure and the function of proteins is carried out by the use of high throughput methods.

Techniques involved

Genomics: The techniques involved in genomics include gene sequencing strategies such as directed gene sequencing, whole-genome shotgun sequencing, construction of expressed sequence tags (ESTs), identification of single nucleotide polymorphisms (SNPs)

Techniques involved

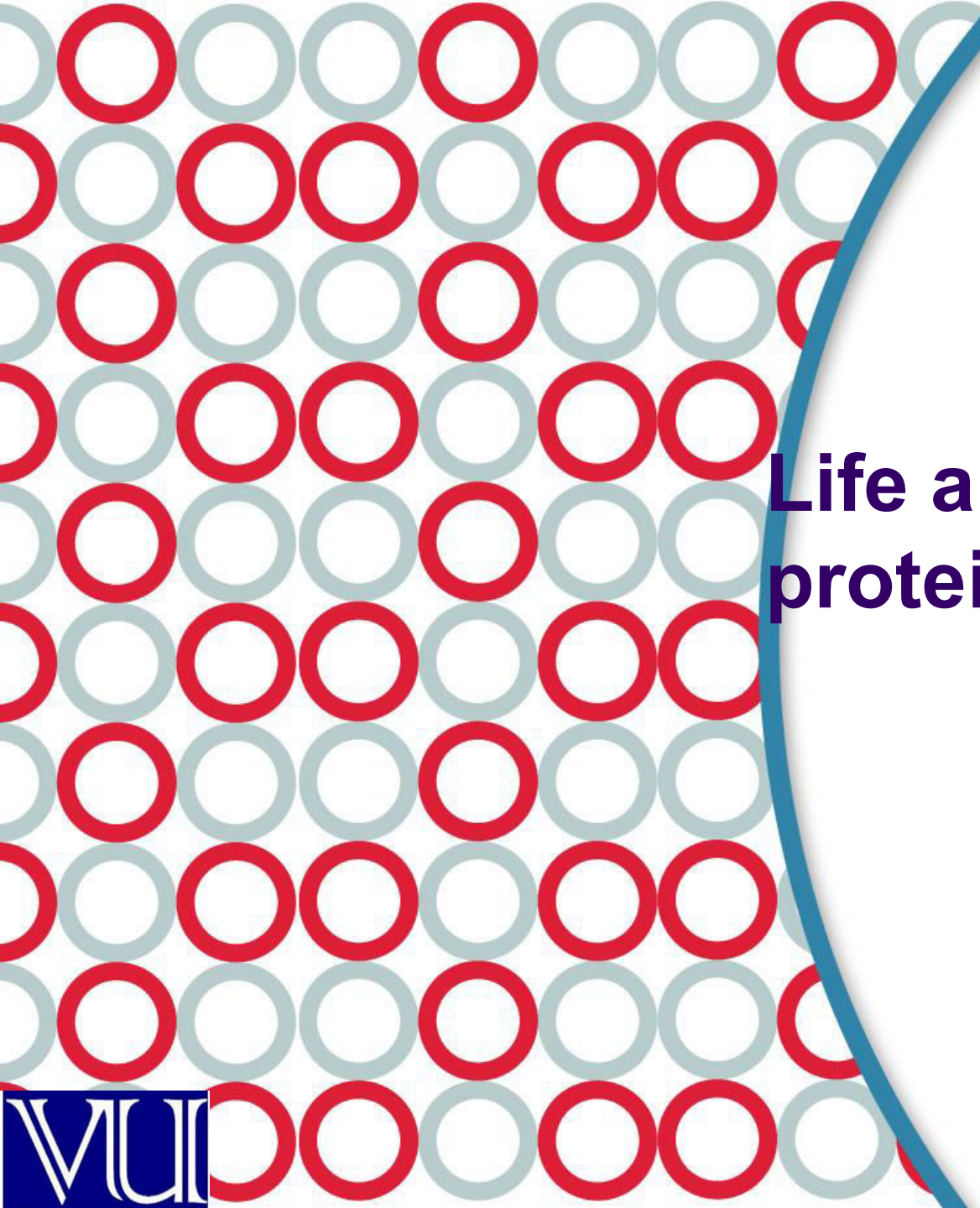
Proteomics:

- Extraction and electrophoretic separation of proteins.
- Digestion of proteins with the use of trypsin into small fragments.
- Determination of the amino acid sequence by mass spectrometry.
- Identification of proteins using the information in the protein databases.

Importance

Genomics: Genomic studies are important to understand the structure, function, location, regulation of the genes of an organism.

Proteomics: The study of the entire set of proteins produced by a cell type is done in order to understand its structure and function.



Life and death of a protein

Lecture-1

Life and Death of a Protein

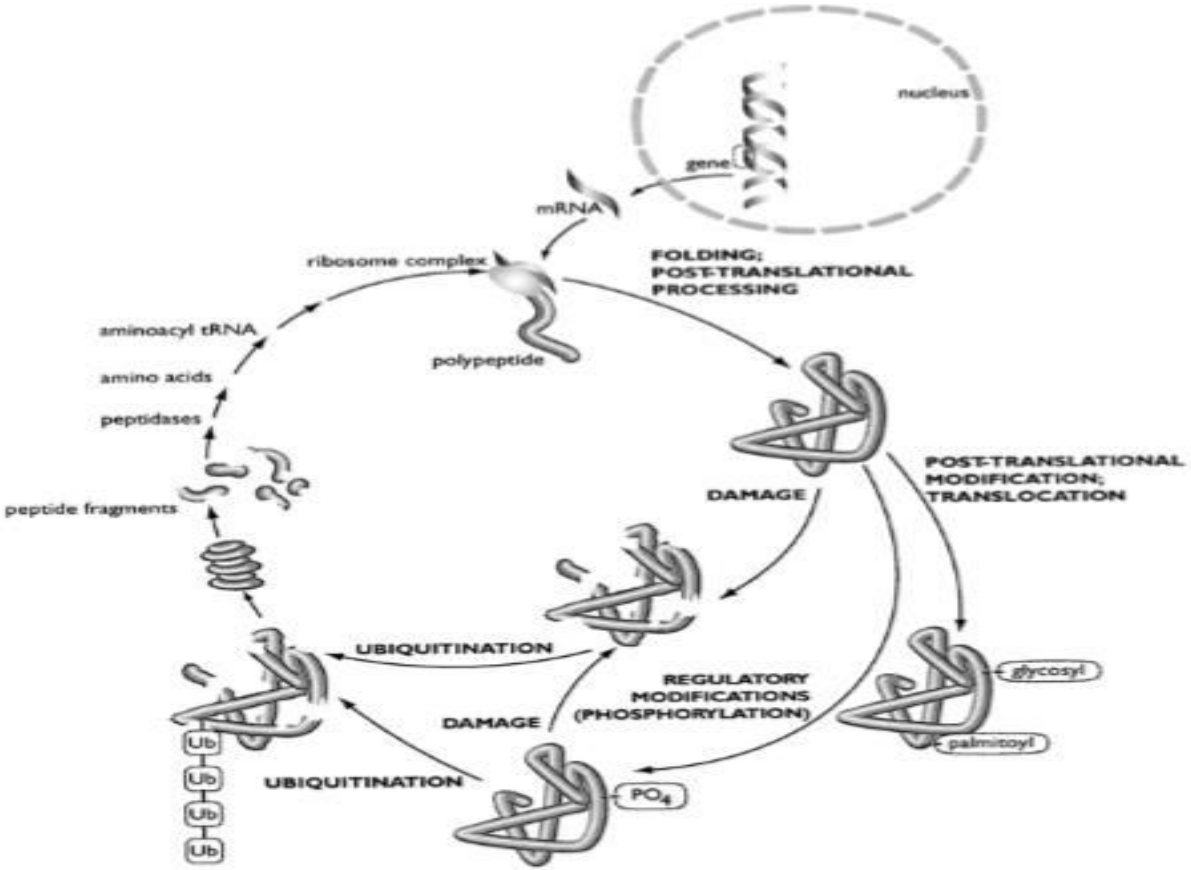
Proteins are synthesized by the translation of mRNAs into polypeptides on ribosomes.

In most cases, the initial polypeptide-translation product undergoes some type of modification before it assumes its functional role in a living system.

These changes are broadly termed “posttranslational modifications” and encompass a wide variety of reversible and irreversible chemical reactions.

Approximately 200 different types of posttranslational modifications have been reported. Some of these are summarized in Fig. 1, which depicts the life cycle of a prototypical protein.

Life Cycle of the Cell



Modifications during Protein Cycle

Modifications those occur early in the life of the protein

- Carboxylation of glutamate residues
- Removal of the N-terminal methionine
- Glycosylation
- Addition of Prosthetic groups
- Formation of multisubunit complexes
- Prenylation of cysteine residues assists anchoring of proteins in or on membranes.

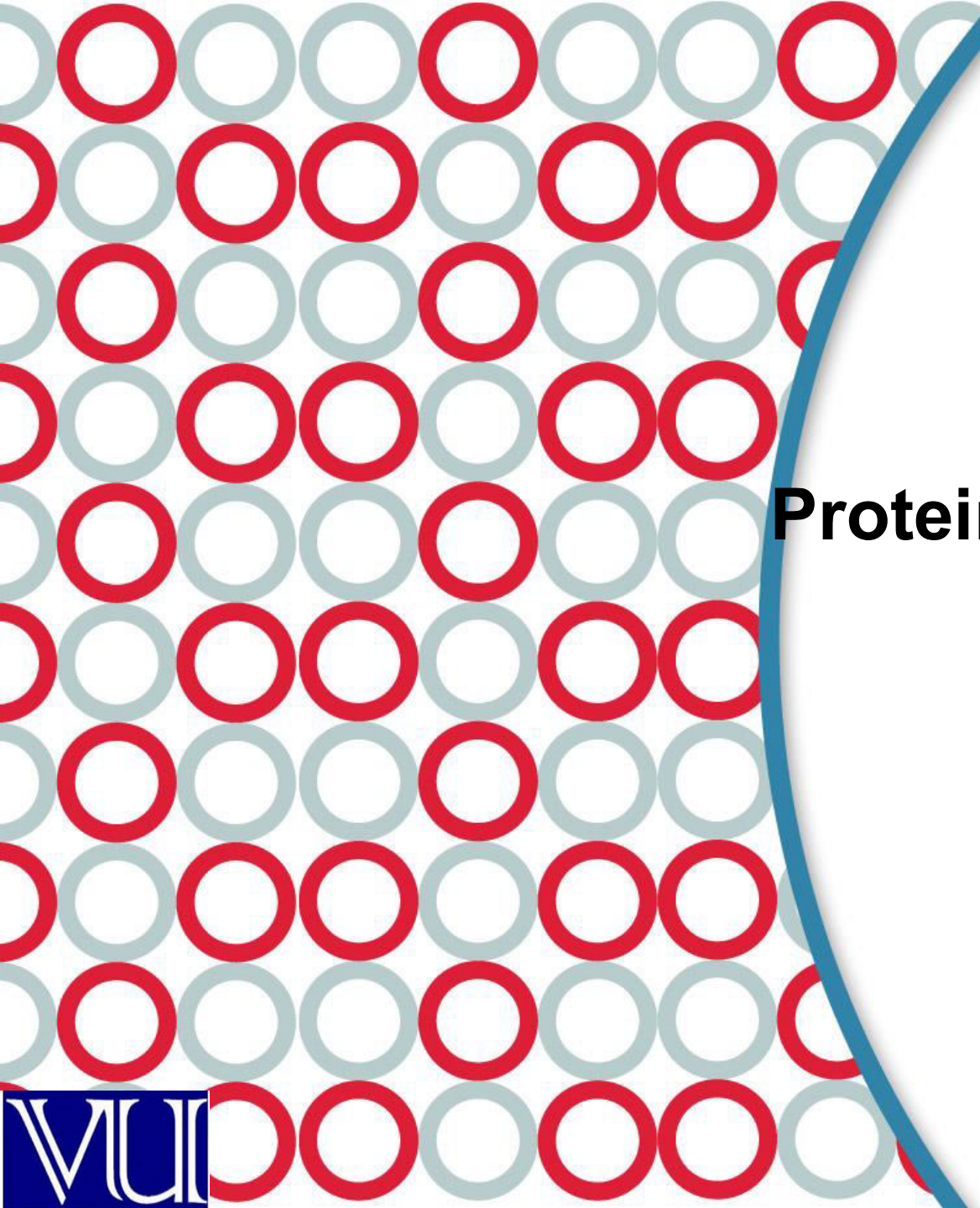
These more or less “permanent” modifications and transport ultimately result in the delivery of functional proteins to specific locations in cells.

Cont…

- The activities of many proteins are then controlled by posttranslational modifications.
- The most prominent and best-understood of these is phosphorylation of serine, threonine, or tyrosine residues.
- Phosphorylation may activate or inactivate enzymes, alter proteinprotein interactions and associations, change protein structures, and target proteins for degradation.
- Protein phosphorylation regulates protein function in diverse contexts and appears to be a key switch for rapid on-off control of signaling cascades, cell-cycle control, and other key cellular functions.

Degradation of Proteins

- Protein modifications appear to be critical to initiating processes that ultimately degrade proteins.
- Phosphorylation of some proteins is rapidly followed by conjugation with ubiquitin, which leads to degradation by the 26S proteasomal complex.
- There evidently are other stimuli for protein ubiquitination and turnover, including oxidative damage and other protein modifications.
- Proteins also undergo degradation by lysosomal enzymes.
- Any protein may be present in many forms at any one time in a cell.
- Collectively, the proteome of a cell comprises all of these many forms of all expressed proteins. This certainly makes the proteome bewilderingly complex.



Proteins as Modular Stru

BIO 601

Lecture-1

Proteins as Modular Structures

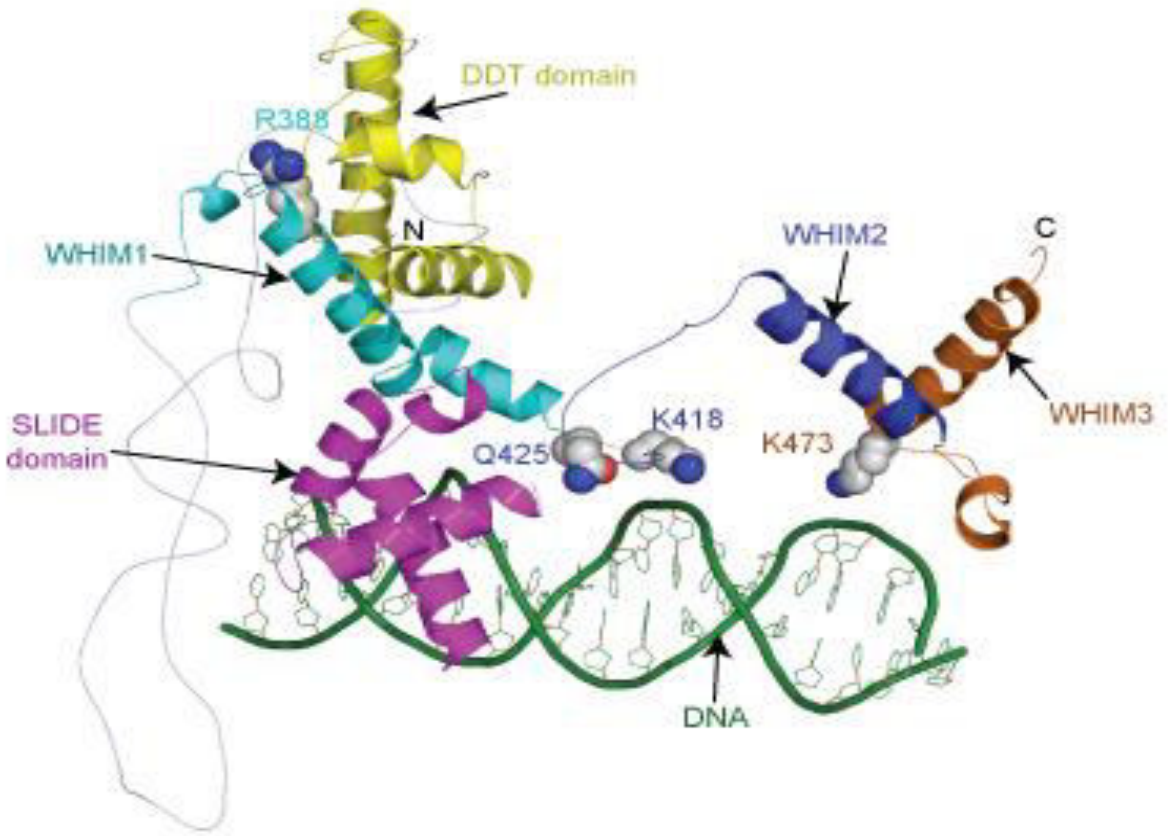
Segments of amino acid sequences can be considered as functional building blocks or modules.

The modular units in proteins that confer specific properties and functions are referred to as “motifs” or “domains”.

Motifs and domains are recognizable sequences that confer similar properties or functions when they occur in a variety of proteins.

In some cases, amino acid sequences within motifs and domains are highly conserved and do not vary from protein to protein.

In other cases, some key amino acids occur in a reproducible relationship to each other in a sequence, even though various substitutions in other amino acids occur


DDT domain
R388
N
WHIM1
WHIM2
C
WHIM3
SLIDE
domain
K418
Q425
K473
DNA

Proteins as Modular Structures

Longer amino acid sequences often form domains, which confer specific properties or functions on a protein.

Some domain structures refer simply to sequences that confer a bulk physical property to a segment of the polypeptide, such as transmembrane domains, which simply form helices that span a lipid bilayer membrane.

Other domain structures provide hydrogen bonding or other contacts for key enzyme substrates or prosthetic groups.

In many cases, domains are made up of combinations of units of secondary structure, such as helix-loop-helix domains.

Proteomics Introducción

BIO 601

Lecture-1

Protein Localization

- Any process in which a protein is transported to, or maintained in, a specific location.
- Cells are organized into many different compartments such as the cytosol, nucleus, endoplasmic reticulum (ER), and mitochondria. Almost all proteins are made in the cytosol, yet each cellular compartment requires a specific set of proteins.
- How does the cell regulate protein localization to be sure that proteins end up where they should?

Compartments of an Animal Cell

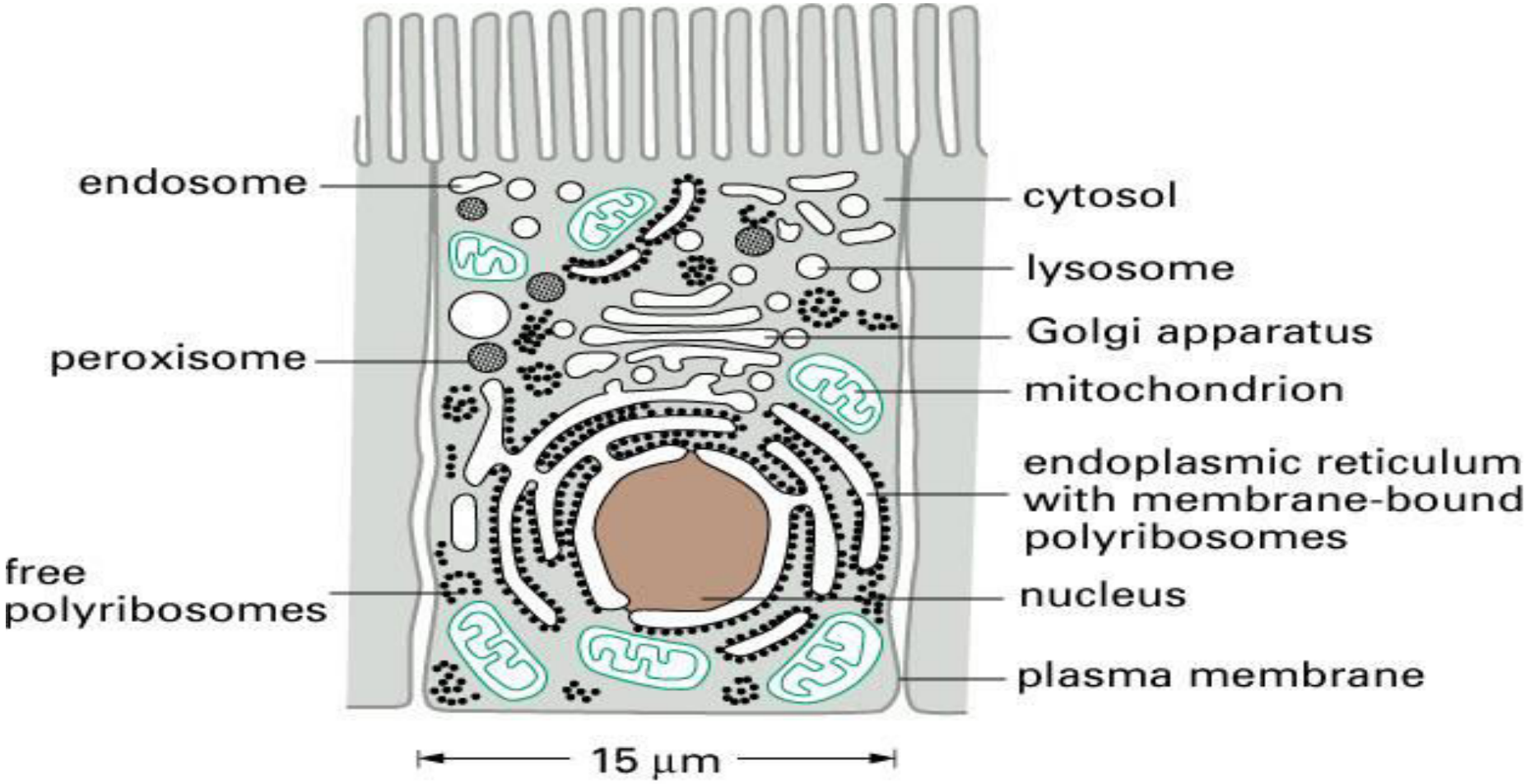


Figure 12–1. Molecular Biology of the Cell, 4th Edition.

Functions of major intracellular compartments:

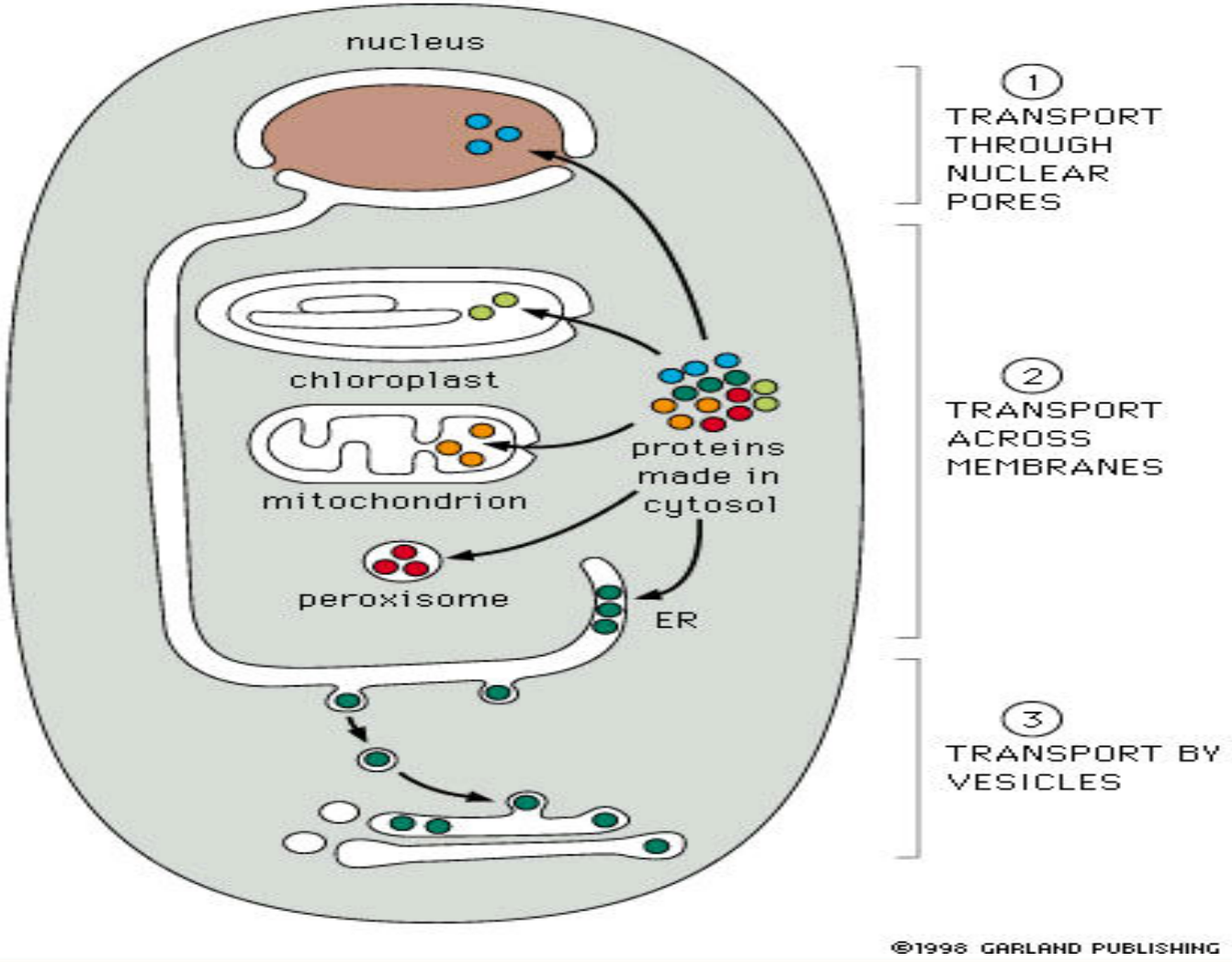
- Nucleus - contains main genome, DNA and RNA synthesis.
- Cytosol - most protein synthesis, glycolysis and metabolic pathways synthesizing amino acids, nucleotides.
- Endoplasmic reticulum - synthesis of membrane proteins, lipid synthesis.
- Golgi apparatus - covalent modification of proteins from ER, sorting of proteins for transport to other parts of the cell.

Cont…

- Mitochondria and chloroplasts (plants) - ATP synthesis.
- Lysosomes - degradation of defunct intracellular organelles and material taken in from the outside of the cell by endocytosis.
- Endosomes - sorts proteins received from both the endocytic pathway and from the Golgi apparatus.
- Peroxisomes - oxidize a variety of small molecules.

Three basic modes of protein Localization

1. Gated transport
2. Transmembrane transport
3. Vesicular transpo



Roadmap of protein traffic

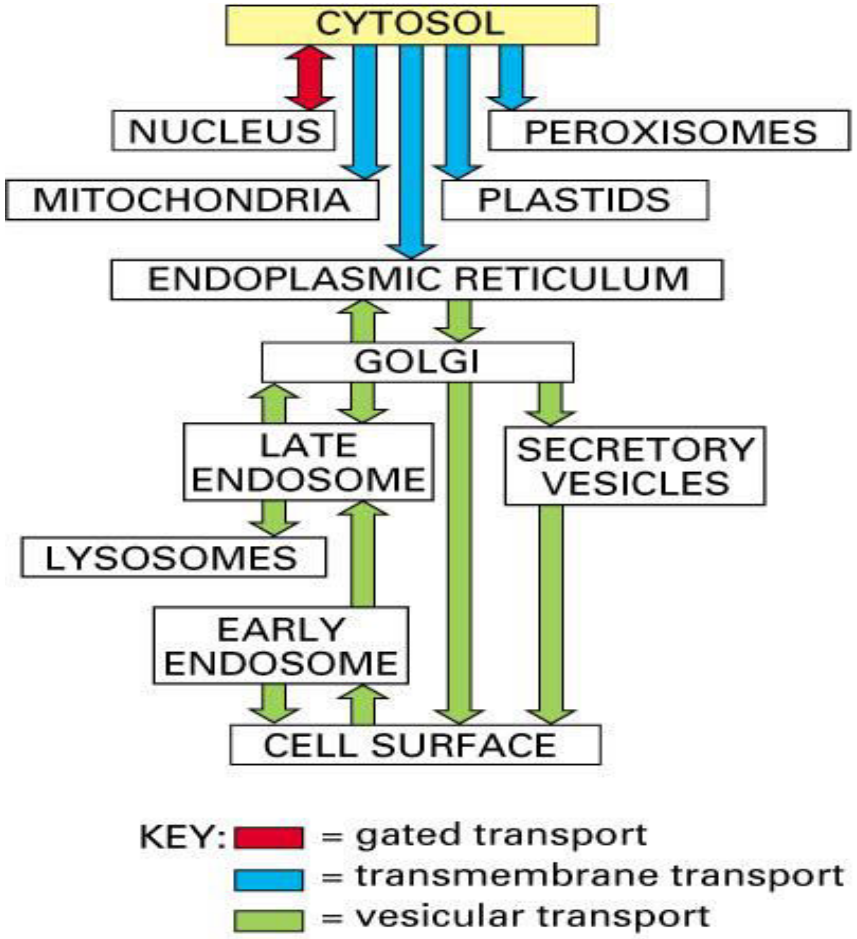


Figure 12–6. Molecular Biology of the Cell, 4th Edition.

The genesis and function of internal compartments depends on the appropriate targeting of proteins

Some of the green route illustrated.

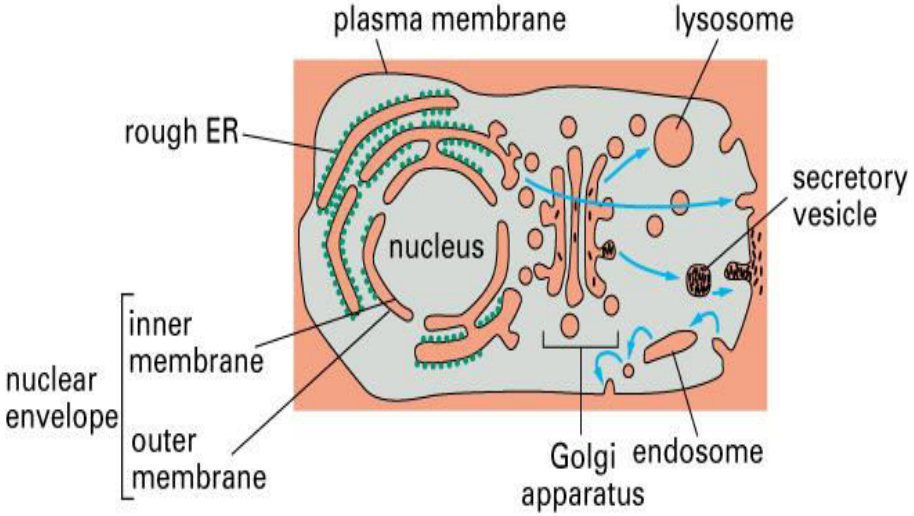
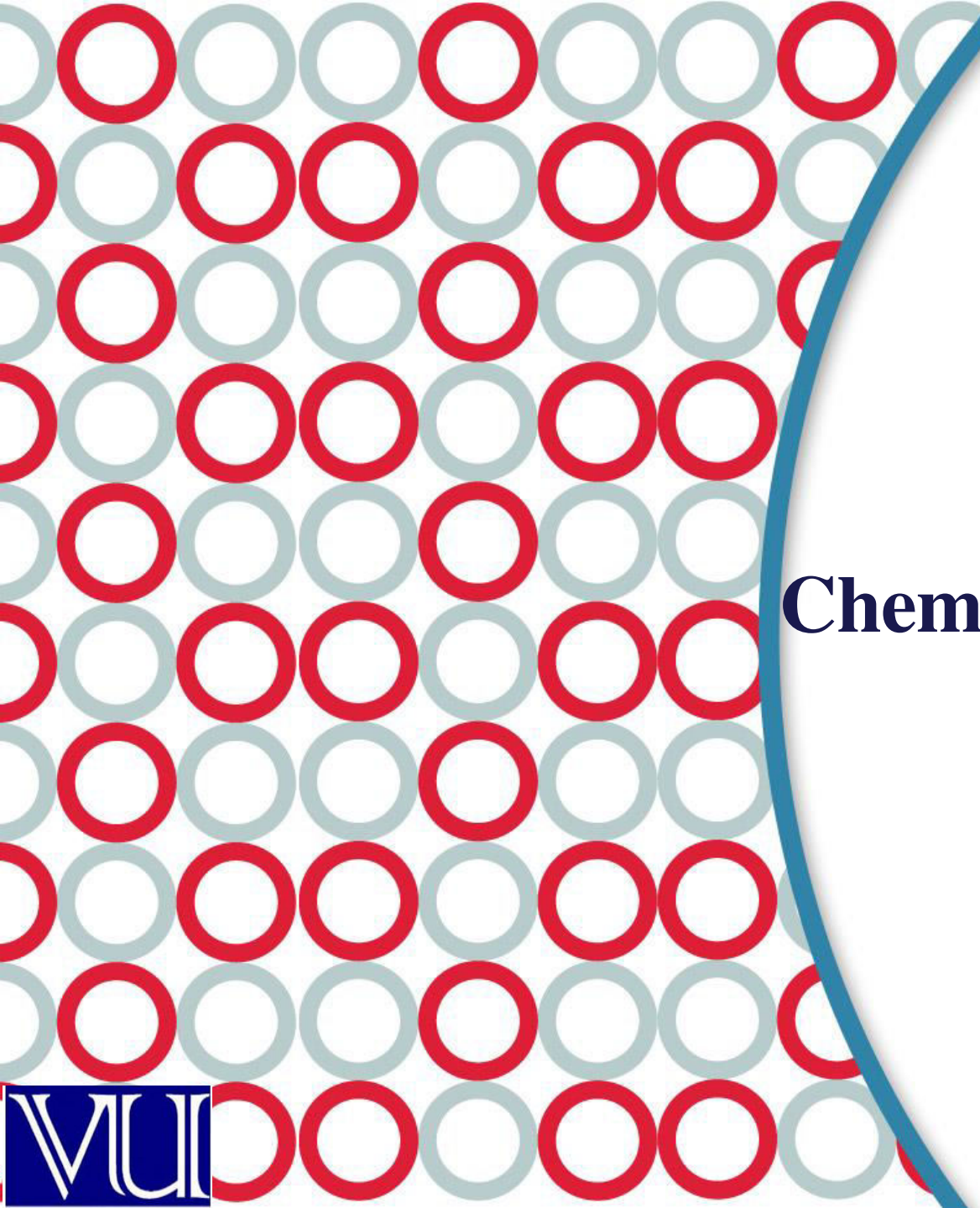


Figure 12–5. Molecular Biology of the Cell, 4th Edition.



Chemical Composition o

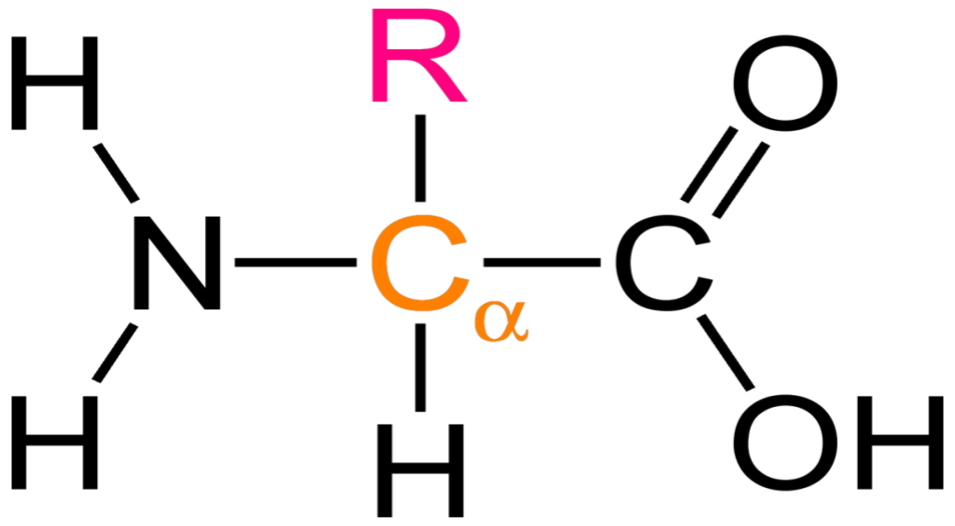
BIO 601

Lecture-

Chemical composition of proteins

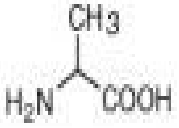
- Proteins are polymers of amino acids.
- They range in size from small to very large.
- All the proteins are made up of Twenty different types of amino acids. So these amino acids are called standard amino acids.

Chemical composition of Proteins

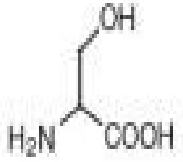


Chemical composition of proteins

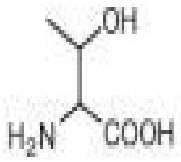
Glycine (Gly, G)



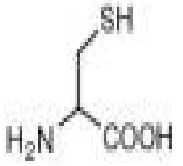
Alanine (Ala, A)



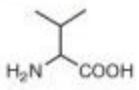
Serine (Ser, S)



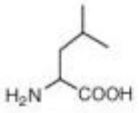
Threonine (Thr, T)



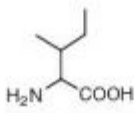
Cysteine (Cys, C)



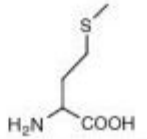
Valine (Val, V)



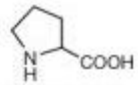
Leucine (Leu, L)



Isoleucine (Ile, I)

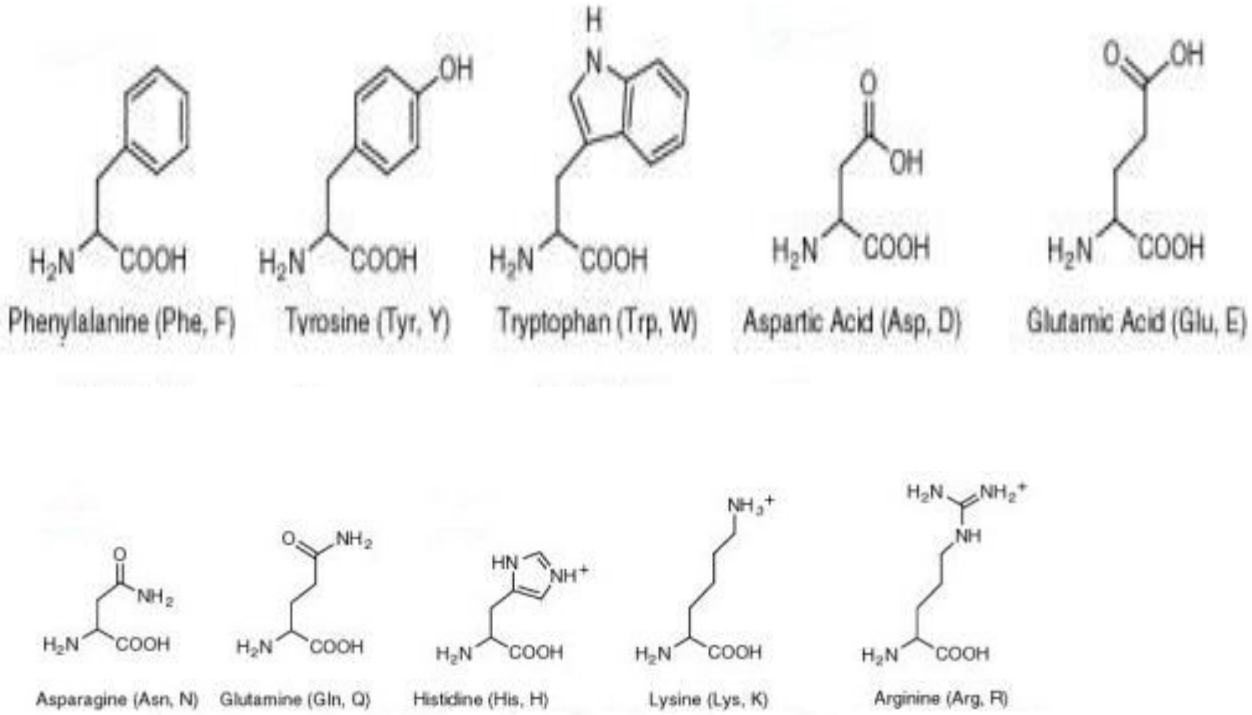


Methionine (Met, M)



Proline (Pro, P)

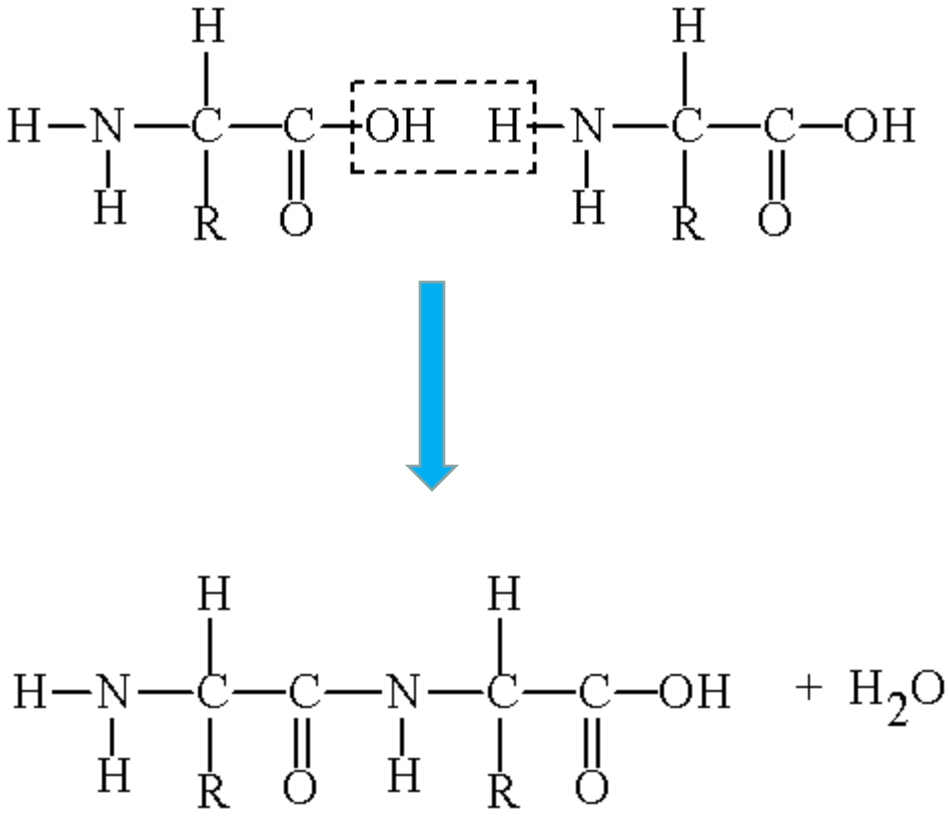
Chemical composition of proteins



Chemical composition of proteins

- In a protein molecule, each amino acid residue is joined to its neighbour by a specific type of covalent bond which is called Peptide Bond.

Chemical composition of proteins



Chemical composition of proteins

- Amino acids can successively join to form dipeptides, tripeptides, tetrapeptides, oligo peptides and polypeptides.

Chemical composition of proteins

Introduction to Homology Modelling

Introduction to Homology Modelling

Background

- Proteins are 3D molecules with their own unique structures
- Protein structure is reflective of the protein function
- Protein structure includes 1', 2', 3' and 4' structures

Introduction to Homology Modelling

Background

- 1’ structure of proteins is the sequence of proteins and can be obtained by mass spectrometry

- 2’ structures formed by proteins are the helices, beta sheets, loops and coils

Introduction to Homology Modelling

Background

- 3’ structure of proteins is the combination of 2’ structures such that the overall protein structure is formed

- 4’ protein structure is formed when two or more proteins complex together

Introduction to Homology Modelling

Background

- X-Ray Crystallography and NMR Spectroscopy are used to find the structures of proteins
- However, these methods are difficult and expensive
- Solution: Prediction of structures

Introduction to Homology Modelling

Introduction

- Protein sequence gives rise to its structure
- If another protein which has a similar sequence also has its structure known, the structure of an unknown protein can be predicted based on that similar protein

Introduction to Homology Modelling

Introduction

- So, it is then possible to identify unknown protein structures by just examining the homologous protein sequences

Introduction to Homology Modelling

Conclusions

- Sequence Identity
- Alignment Length

Which combination of identity and alignment length is suitable for best for structure prediction?

END

Homology Modelling

Homology, Paralogy and Orthology

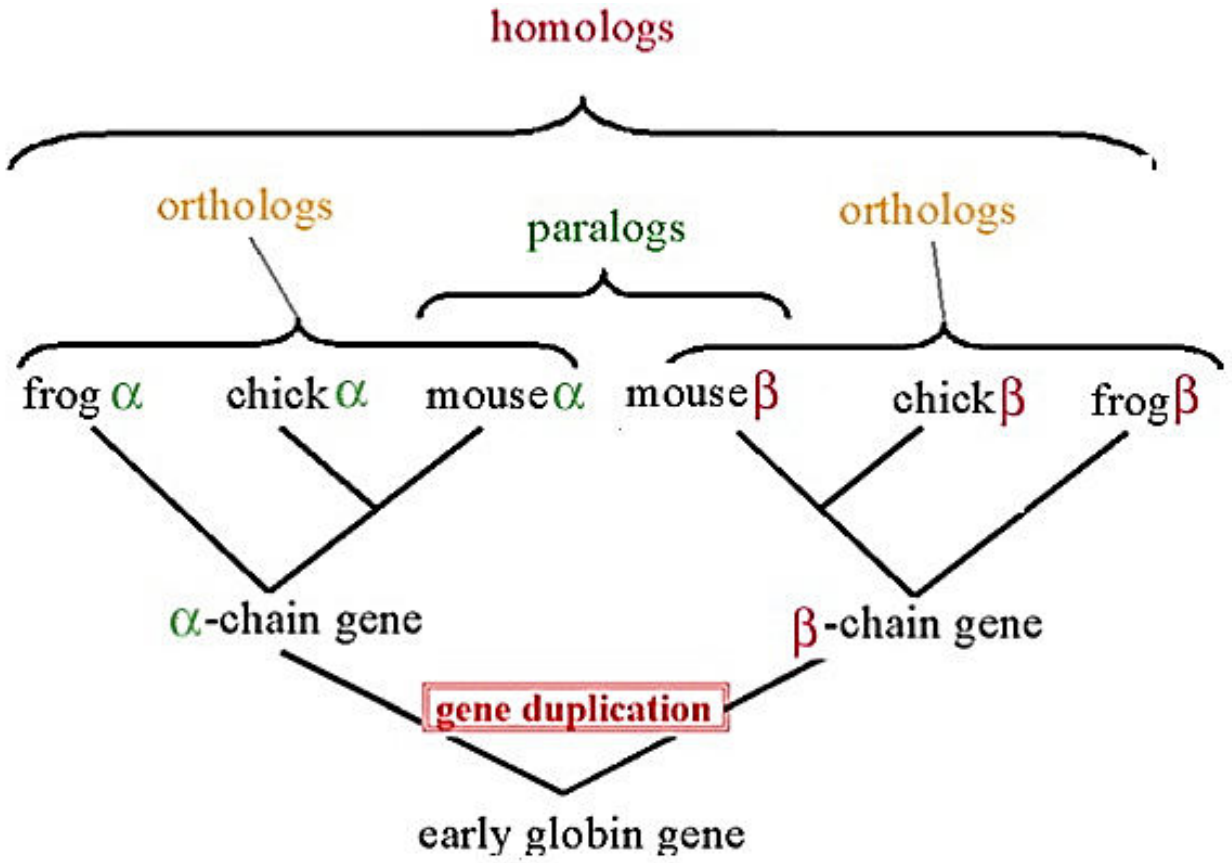
Homology, Paralogy and Orthology

Background

- In homology modelling, proteins with similar 1’ sequences are considered
- Given that one of them has its 3’ structure known, then the 3’ structure of other protein can be predicted

Homology, Paralogy and Orthology

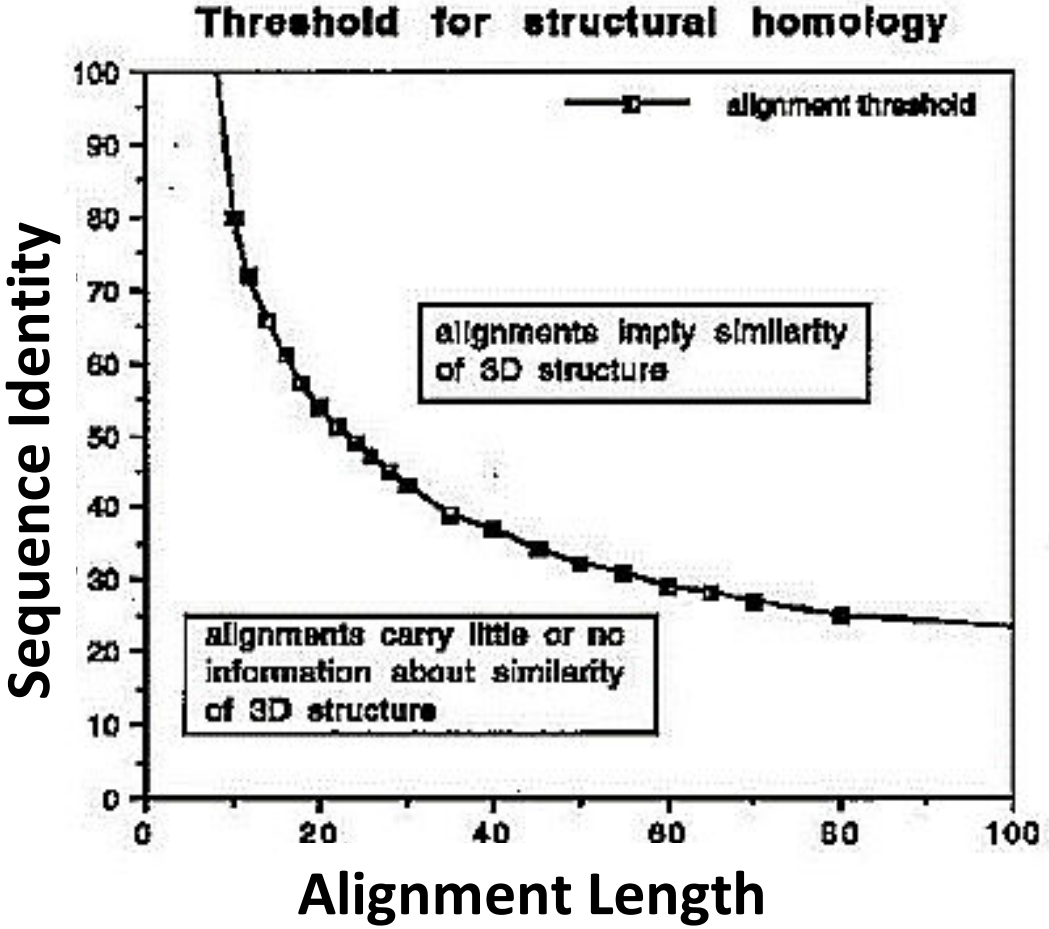
Homology: Paralogy vs. Orthology



http://bioweb.uwlax.edu/GenWeb/Molecular/Bioinformatics/Unit_4/Lab_4-2/lab_4-2.htm

Homology, Paralogy and Orthology

How much homology is required or better?



http://www.cmbi.ru.nl/gvteach/astra/lectures/homology_modelling.ppt

Homology, Paralogy and Orthology

Conclusions

- Good sequence alignment and identity ensures that homology modelling will give accurate results
- Next, what is the workflow for homology modelling?

END

Homology Modelling

Workflow of Structural Modelling

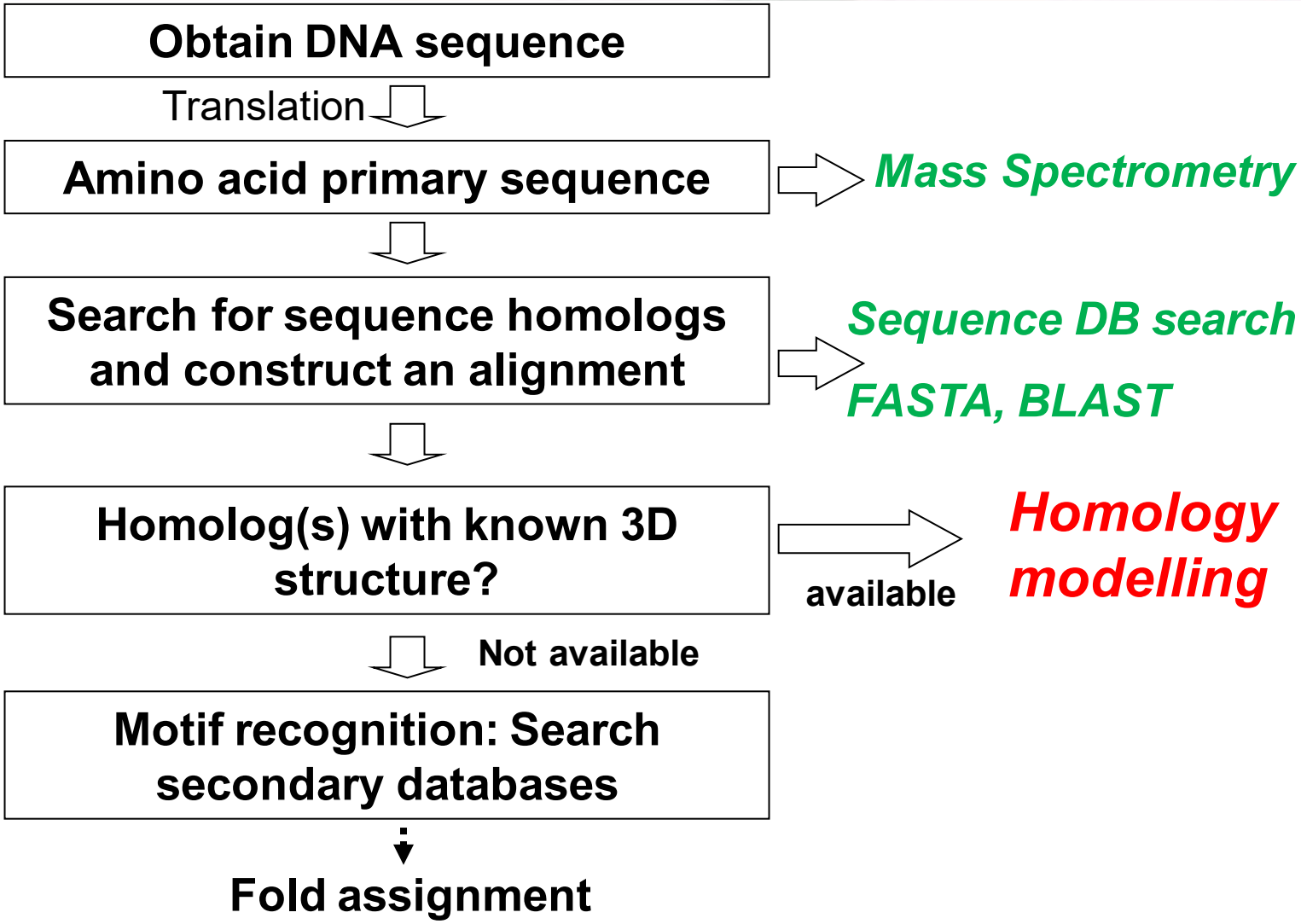
Workflow of Structural Modelling

Background

- Homology modelling is used to predict structures of proteins having high sequence similarity with other proteins with known structures!

- Let's consider the workflow of homology modelling

Workflow of Structural Modelling



Workflow of Structural Modelling

Introduction

Overall, there are three different strategies for structure prediction

1. Homology Modelling
2. Threading/Fold Recognition
3. Ab Initio Modelling

Workflow of Structural Modelling

Conclusions

- Next, we will proceed to perform homology modelling
- For that there is a seven step procedure which we will see in the next module.

END

Homology Modelling

Seven Steps to Homology Modelling – I

Seven Steps to Homology Modelling - I

Background

Protein structure can be predicted by 3 methods:

1. Homology Modelling

2. Fold Recognition / Threading

3. Ab Initio Modelling

Seven Steps to Homology Modelling - I

Introduction

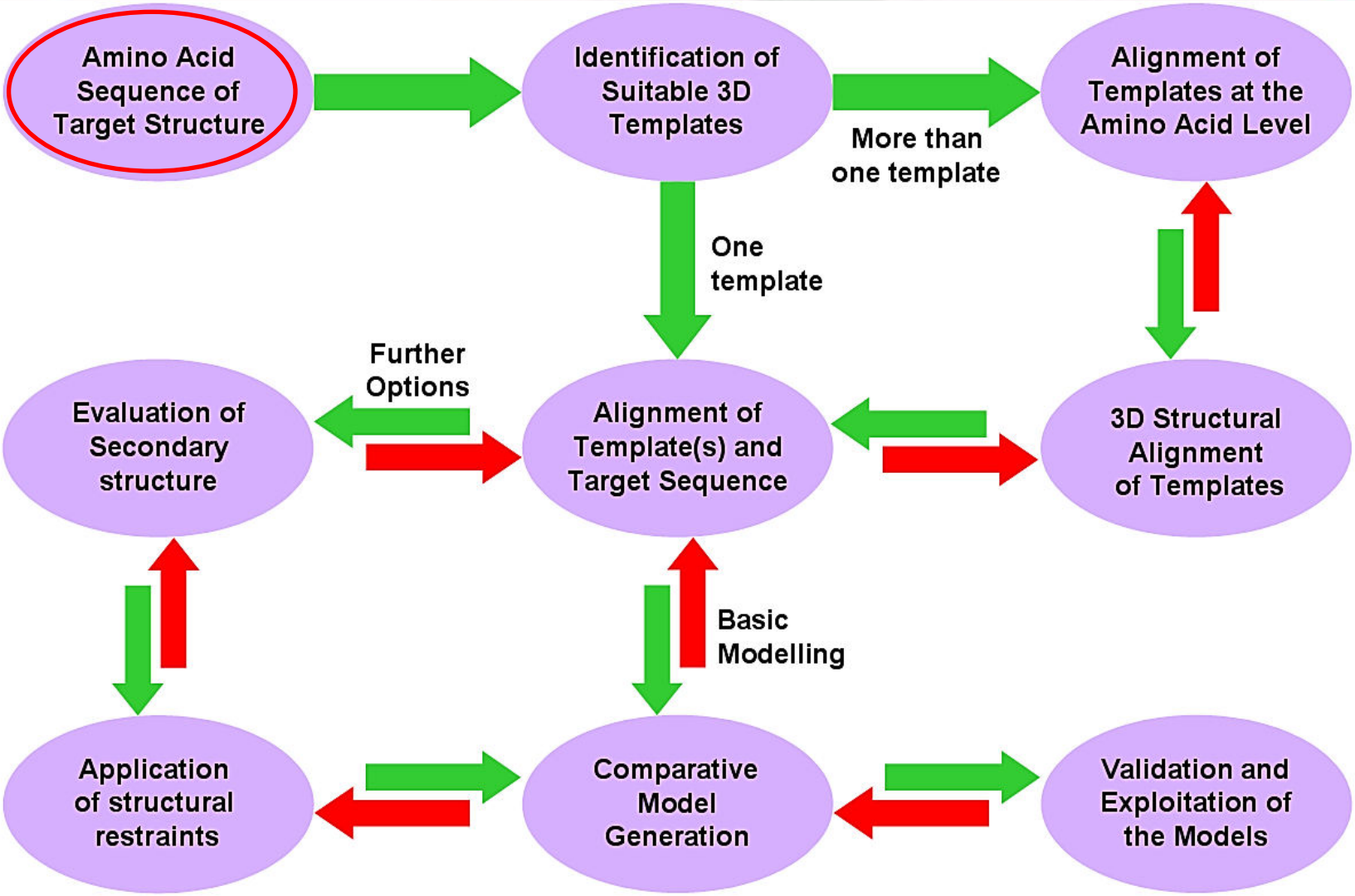
- Let's start by looking at Homology Modelling
- There are seven salient steps in any Homology Modelling pipeline
- Definition of Template (known) & Target (unknown)

Seven Steps to Homology Modelling - I

Homology modeling of the target structure can be done as follows:

1. Template recognition and initial alignment
2. Alignment correction
3. Backbone generation
4. Loop modeling
5. Side-chain modeling
6. Model optimization
7. Model validation

Seven Steps to Homology Modelling - I



Seven Steps to Homology Modelling - I

Conclusions

- Homology modelling works in seven steps
- It is a repetitive process
- Next, we will look at each step in detail!

END

Homology Modelling

Seven Steps to Homology Modelling – II

Seven Steps to Homology Modelling - II

Background

Template recognition and initial alignment

Alignment correction

Backbone generation

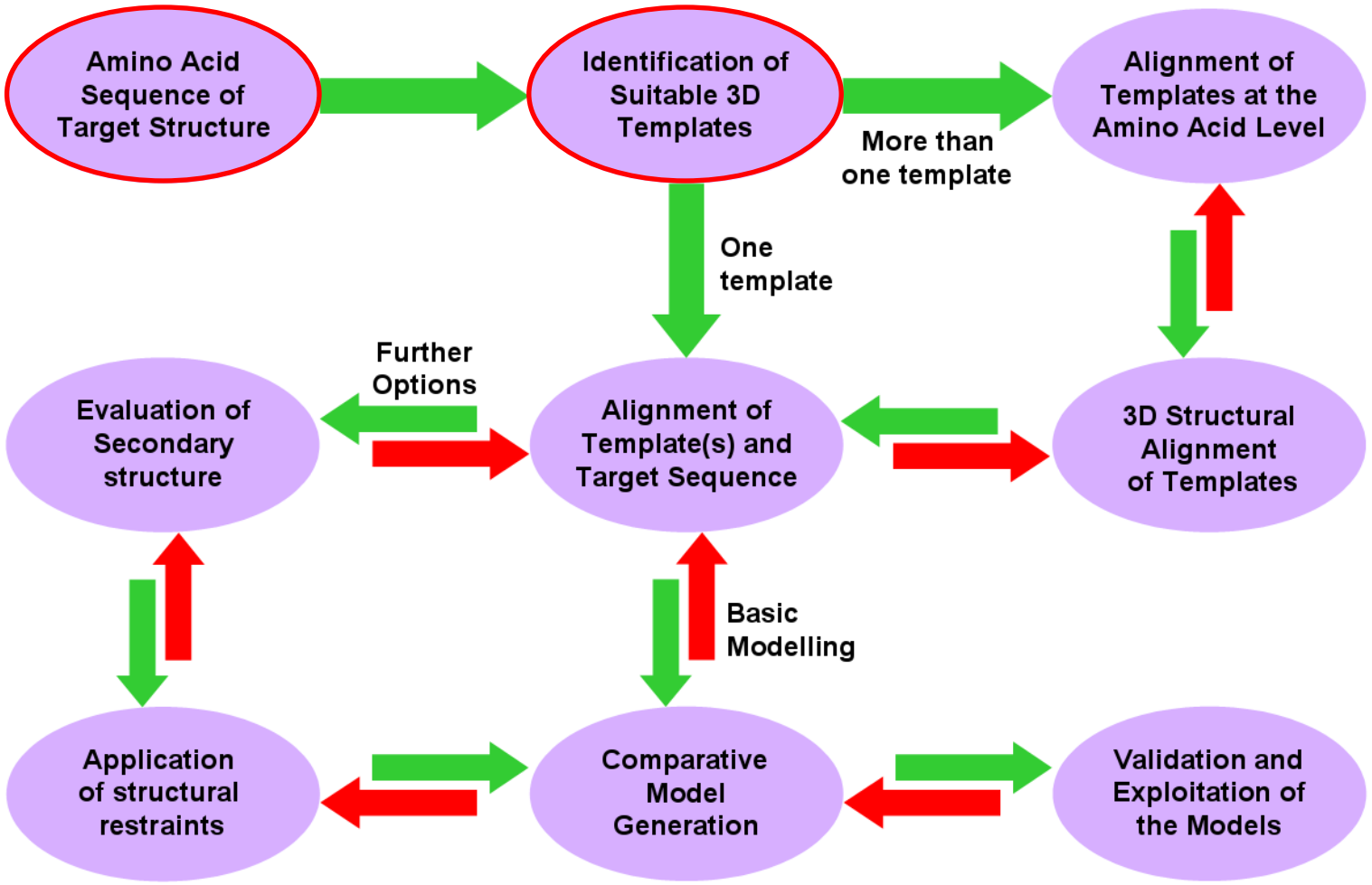
Loop modeling

Side-chain modeling

Model optimization

Model validation

Seven Steps to Homology Modelling - II



Seven Steps to Homology Modelling - II

Compare the sequence of the unknown protein with all the sequences of known structures stored in the Protein Data Bank (PDB).

↓

BLAST this sequence against PDB sequences – Obtain a list of known protein structures that match the sequence.

↓

BLAST uses a residue exchange scoring matrix.
Residues that are easily exchanged (e.g. Ile to Leu) get a better score than residues that have different properties (for example Glu to Trp).
Function specific conserved residues get the best score (e.g. Cys to Cys).

Seven Steps to Homology Modelling - II

BLAST will provide a list of possible templates for the unknown structure
To make the best initial alignment, BLAST uses an alignment-matrix based on the residue exchange matrix and adds extra penalties for opening and extension of a gap between residues.

The target-sequence is sent to a BLAST server, which searches the PDB to obtain a list of possible templates and their alignments.
The best hit has to be chosen, which is not necessarily the first one.

Seven Steps to Homology Modelling - II

Conclusions

- Now the template and target are selected
- Next, we perform fine-tuning of alignment and introduce corrections to ready the mismatches and gaps

END

Homology Modelling

Seven Steps to Homology Modelling – III

Seven Steps to Homology Modelling - III

Background

Template recognition and initial alignment

Alignment correction

Backbone generation

Loop modeling

Side-chain modeling

Model optimization

Model validation

Seven Steps to Homology Modelling - III

Fine tune and adjust the BLAST alignments

Example: Ala -> Glu is possible but unlikely in a hydrophobic core, so these residues should not be aligned.

Use MSA tools (e.g. ClustalW), to find the residues and properties that need to be conserved.

Examine the template structure to check which residues are in the core hence less likely to change than the residues at the outside.

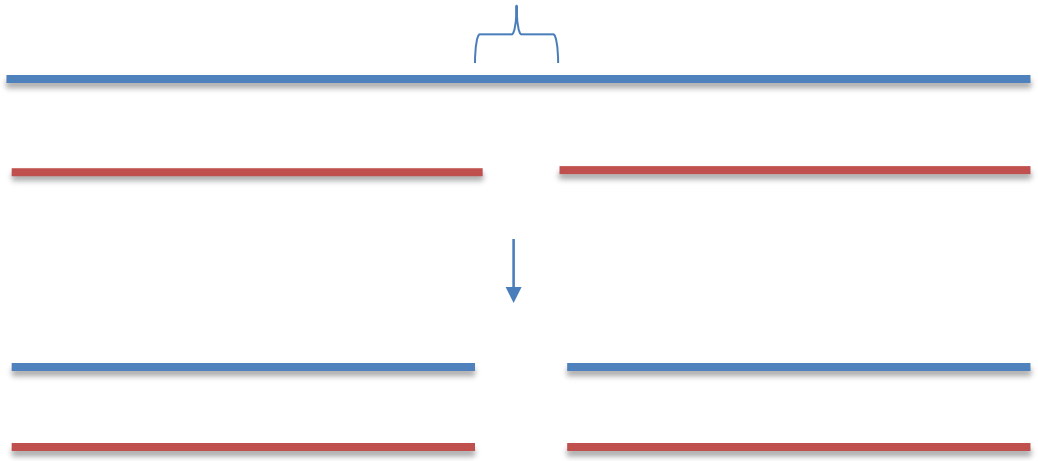
Seven Steps to Homology Modelling - III

Insertions and deletions can be made in those parts of the sequence which are highly variable .

Note: MSA can be helpful to find these places.

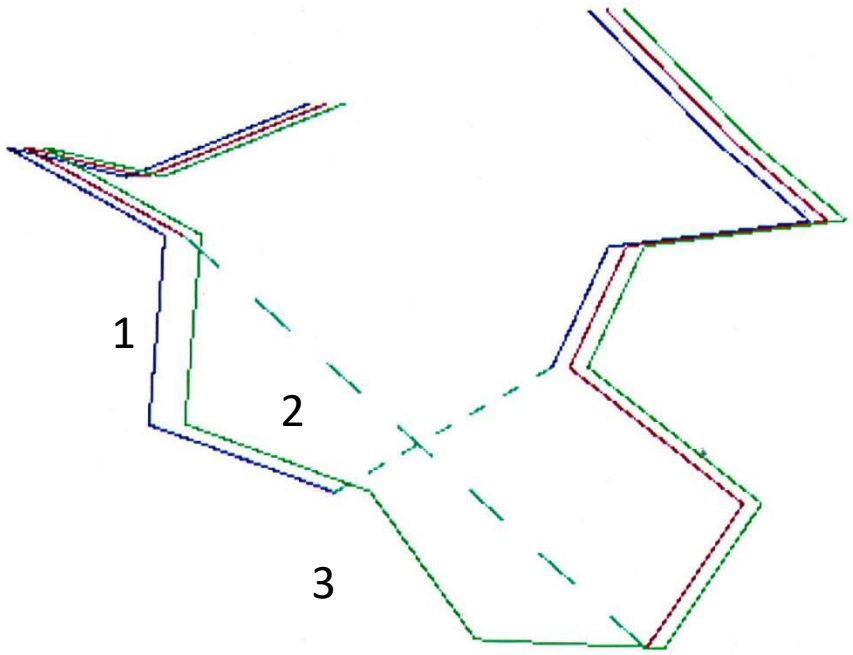
Gaps have to be shifted around until they are as small as possible.

Seven Steps to Homology Modelling - III



Note: After a deletion of 3 residues a big gap occurs in the RED structure, which was the best alignment.

Seven Steps to Homology Modelling - III



Remember that deletion of 3 residues left a big gap in the RED structure, which was the best alignment.

After shifting several residues, the gap is much smaller (blue structure) and more likely to be correct.

Seven Steps to Homology Modelling - III

Conclusions

- The alignment now stands fine tuned and corrected
- Gaps and mismatches have been evaluated and adjusted
- Next step, using this alignment, assemble the backbone

END

Homology Modelling

Seven Steps to Homology Modelling – V

Seven Steps to Homology Modelling - V

Background

Template recognition and initial alignment

Alignment correction

Backbone generation

Loop modeling

Side-chain modeling

Model optimization

Model validation

Seven Steps to Homology Modelling - V

Note that the conserved residues were already copied!
Now, we just need to place the side chains

↓

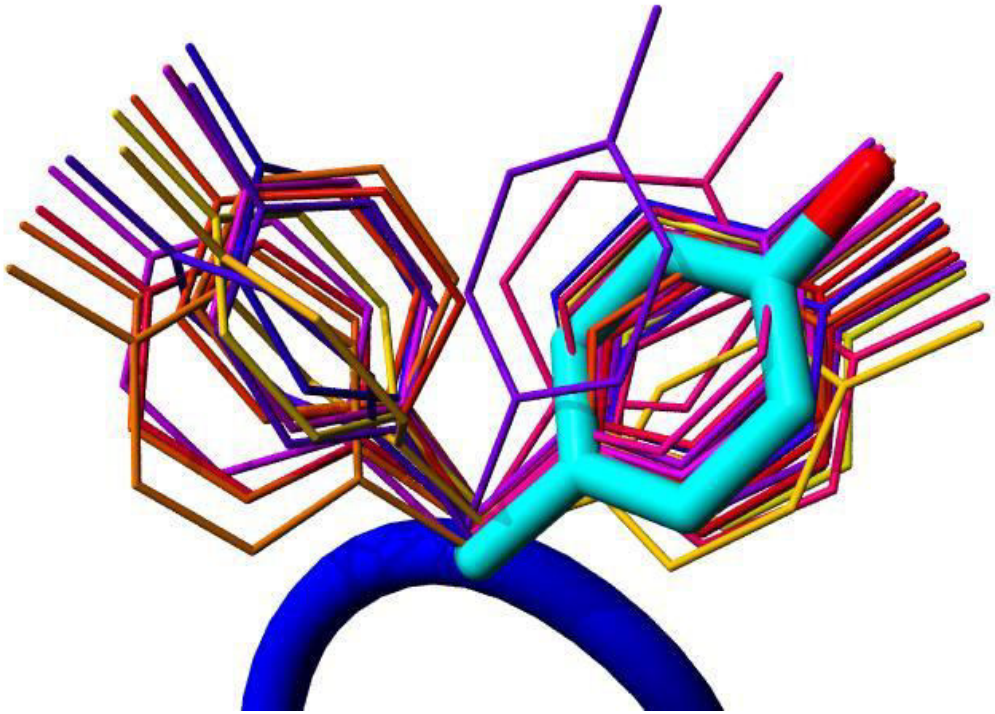
Copy the torsion angles C-alpha/beta to the target!
Rotamers tend to be conserved in homologous proteins and can be predicted as backbone configurations strongly prefer a specific rotamer.

↓

Moreover, libraries of flanking residues can also help estimate the side chain positioning.

Seven Steps to Homology Modelling - V

The backbone of tyrosine strongly prefers two rotamers and the real side-chain may fit one of them!



Seven Steps to Homology Modelling - V

Next....

Template recognition and initial alignment

Alignment correction

Backbone generation

Loop modeling

Side-chain modeling

Model optimization

Model validation

Seven Steps to Homology Modelling - V

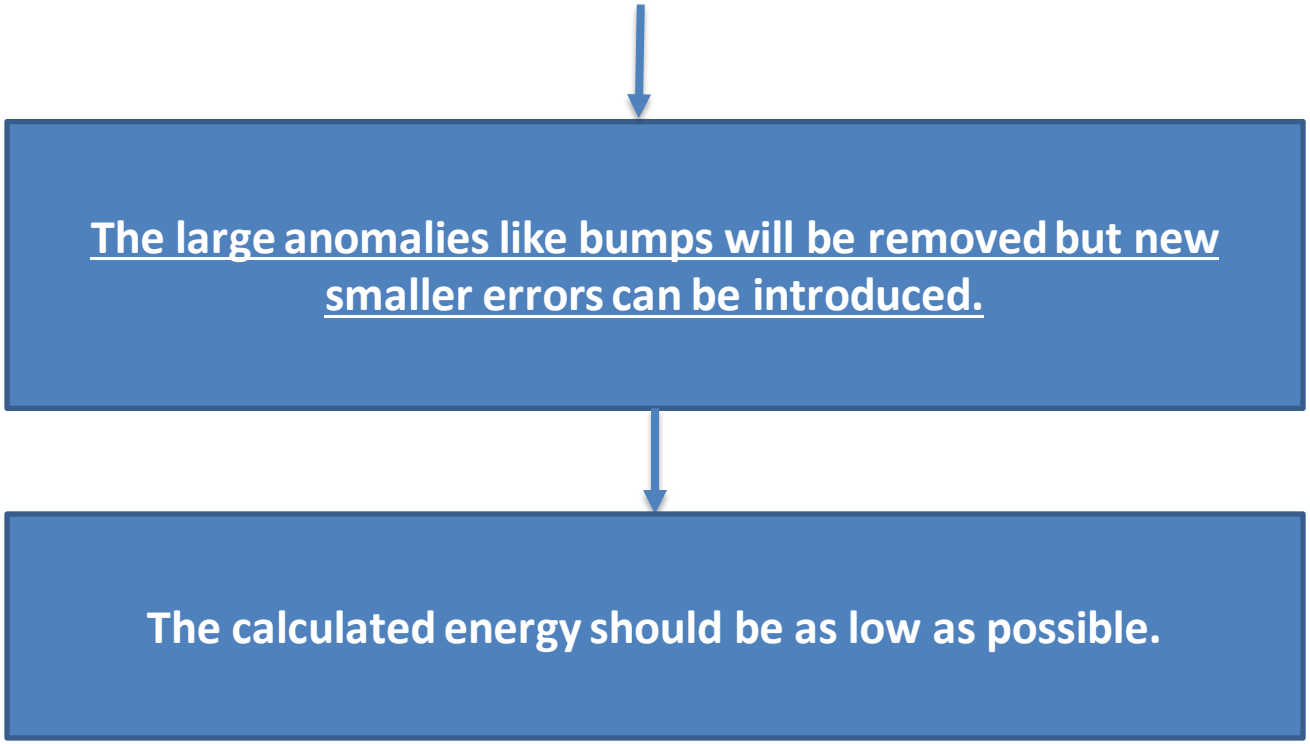
What is the need for further optimization?

Because the updated side-chains can effect the backbone, and this can effect the structure prediction.

Optimization can be done by performing refinements using Molecular Dynamics simulations of the model.

The model is placed in a force-field and the movements of the molecules are followed in time, this mimics the folding of the protein.

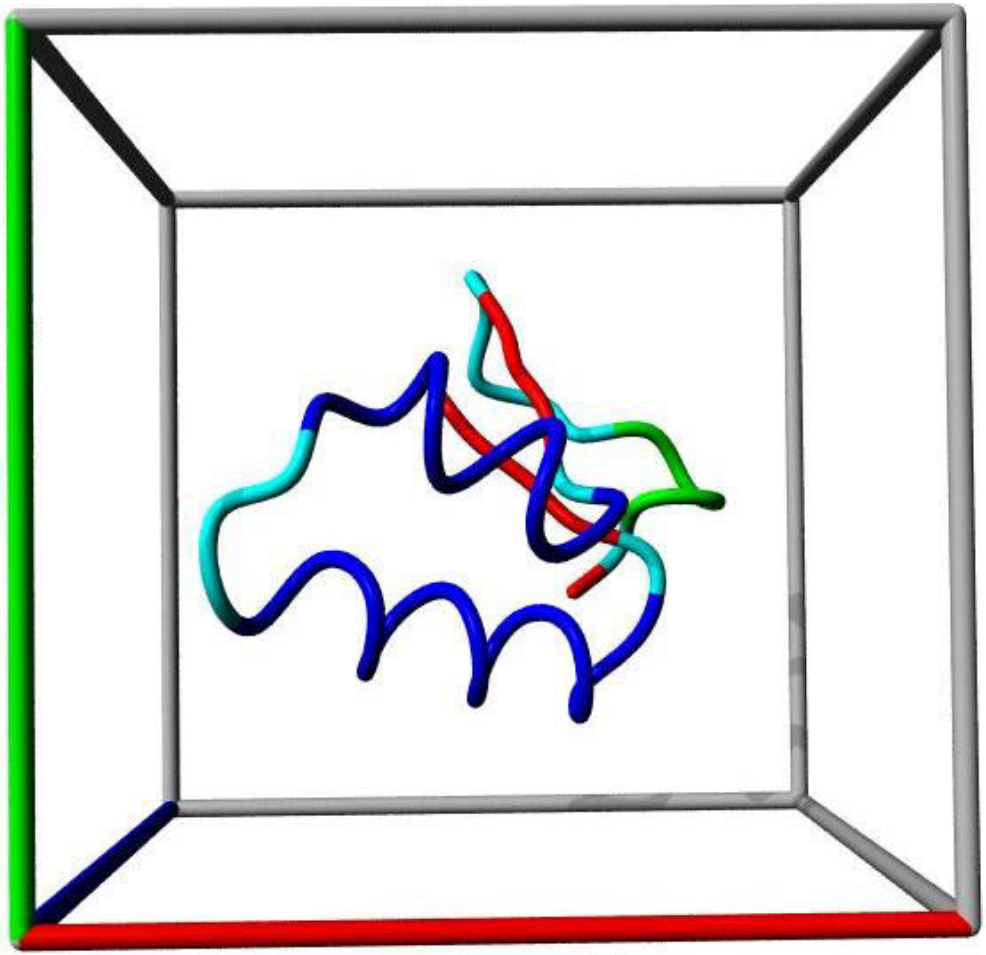
Seven Steps to Homology Modelling - V



Seven Steps to Homology Modelling - V

MD Example Sim: Crambin

(Ethiopian Cabbage Protein)



Seven Steps to Homology Modelling - V

Conclusions

- Now we have minimized large errors
- However, smaller errors may still exist
- Next step, validate the model that we have constructed!

END

Homology Modelling

Seven Steps to Homology Modelling – VI

Seven Steps to Homology Modelling - VI

Background

Template recognition and initial alignment

Alignment correction

Backbone generation

Loop modeling

Side-chain modeling

Model optimization

Model validation

Seven Steps to Homology Modelling - VI

The model with the lowest energy might still be folded completely wrong.

So, the model should be checked again for normal ranges of:
Bumps, Bond angles, Torsion angles, Bond lengths
Other properties, like the distribution of polar/apolar residues, can be compared with real structures.

Important Points:
An error occurs far away from the active site, is tolerable.
But when an error occurs in the active site, one should reconsider the template and/or alignment.

Seven Steps to Homology Modelling - VI

Limitations of Homology Modelling

Large Bias towards structure of template

Cannot study conformational changes

Cannot elicit new catalytic/binding sites

Seven Steps to Homology Modelling - VI

Conclusions

- So how can we overcome such limitations?
- Other strategies include: Threading, and Ab Initio Modelling
- We will also examine online tools for each

END

Homology Modelling

MODELLER for Homology Modelling

MODELLER for Homology Modelling

Background

Template recognition and initial alignment

Alignment correction

Backbone generation

Loop modeling

Side-chain modeling

Model optimization

Model validation

MODELLER for Homology Modelling

Background

- Modeller is a software for homology modelling
- salilab.org/modeller
- Inputs:
  Python script file,
  Sequence alignment
  & Template (PDB)

MODELLER for Homology Modelling

```
from modeller import *
from modeller.automodel import *
log.verbose()
env = environ()
env.io.atom_files_directory = './'

a = automodel(
  env,
  alnfile = 'herg.ali',
  knowns = '1q5o',
  sequence = 'herg'
)

a.starting_model= 1
a.ending_model = 1
a.make()
```

Input Python Script (*.py)

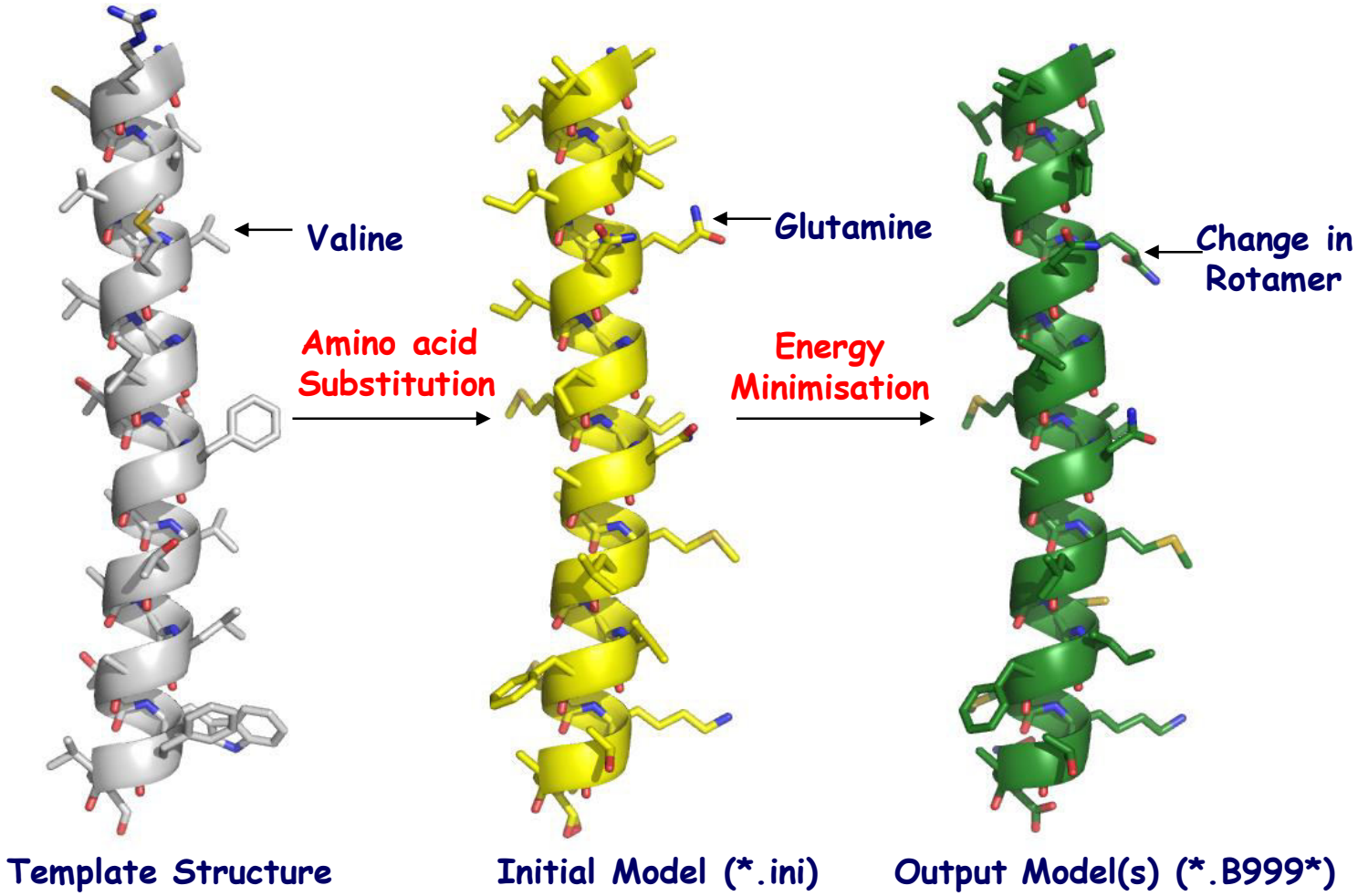
```
>P1;1q5o
structureX: 1q5o : 443 : A : 644 : A ::::
DSSRRQYQEKYKQVEQYMSFHKLPADFRQKIHDYYEHRYQ-GKMFDEDSILGELNGPLRE
EIVNFNCRKLVASMPLFANADPNFVTAMLTKLKFEVFQPGDYIIREGTIGKKMYFIQHGV
VSVLTKGNKEMKLSDGSYFGEICLL--TRGRRTASVRADTYCRLYSLSVDNFNEVLEEYP
MMRRAFETVAIDRLDRIGKKNSIL.*

>P1;herg
sequence: herg : 1 :::::::
YSGTARYHTQMLRVREFIRFHQIPNPLRQRLEEYFQHAWSYTNGIDMNAVLKGFPECLQA
DICLHLNRSLLQHCKPFRGATKGCLRALAMKFKTTHAPPGDTLVHAGDLLTALYFISRGS
IEILRGDVVVAILGKNDIFGEPLNLYARPGKSNGDVRALTYCDLHKIHRDDLLEVLDMYP
EFSDHFWSSLEITFNLRDTN-MIP.*
```

Sequence Alignment (*.ali)

```
ATOM   1  N   ASP A 443    -15.943  41.425  44.702  1.00 44.68
ATOM   2  CA  ASP A 443    -15.424  42.618  45.447  1.00 43.15
ATOM   3  C   ASP A 443    -14.310  43.306  44.686  1.00 41.81
ATOM   4  O   ASP A 443    -14.298  44.528  44.539  1.00 42.61
                          etc...
```

Template Structure (*.pdb)

MODELLER for Homology Modelling



MODELLER for Homology Modelling

- **.log** : log output from the run.
- **.B*** : model generated in the PDB format.
- **.D*** : progress of optimisation.
- **.V*** : violation profile.
- **.ini** : initial model that is generated.
- **.rsr** : restraints in user format.
- **.sch** : schedule file for the optimisation process.

MODELLER for Homology Modelling

Automated Modelling Servers

Swiss Model
http://swissmodel.expasy.org//SWISS-MODEL.html

Robetta
http://robetta.bakerlab.org/

3D Jigsaw
http://www.bmm.icnet.uk/servers/3djigsaw/

Phyre
http://www.sbg.bio.ic.ac.uk/phyre/

MODELLER for Homology Modelling

Conclusions

- Homology modelling helps predict protein structures by using prior structural information

- Several tools are available to perform homology modelling in a programmatic or automated way!

END

Homology Modelling

Fold Recognition – Threading I

Fold Recognition – Threading I

Background

Template recognition and initial alignment

Alignment correction

Backbone generation

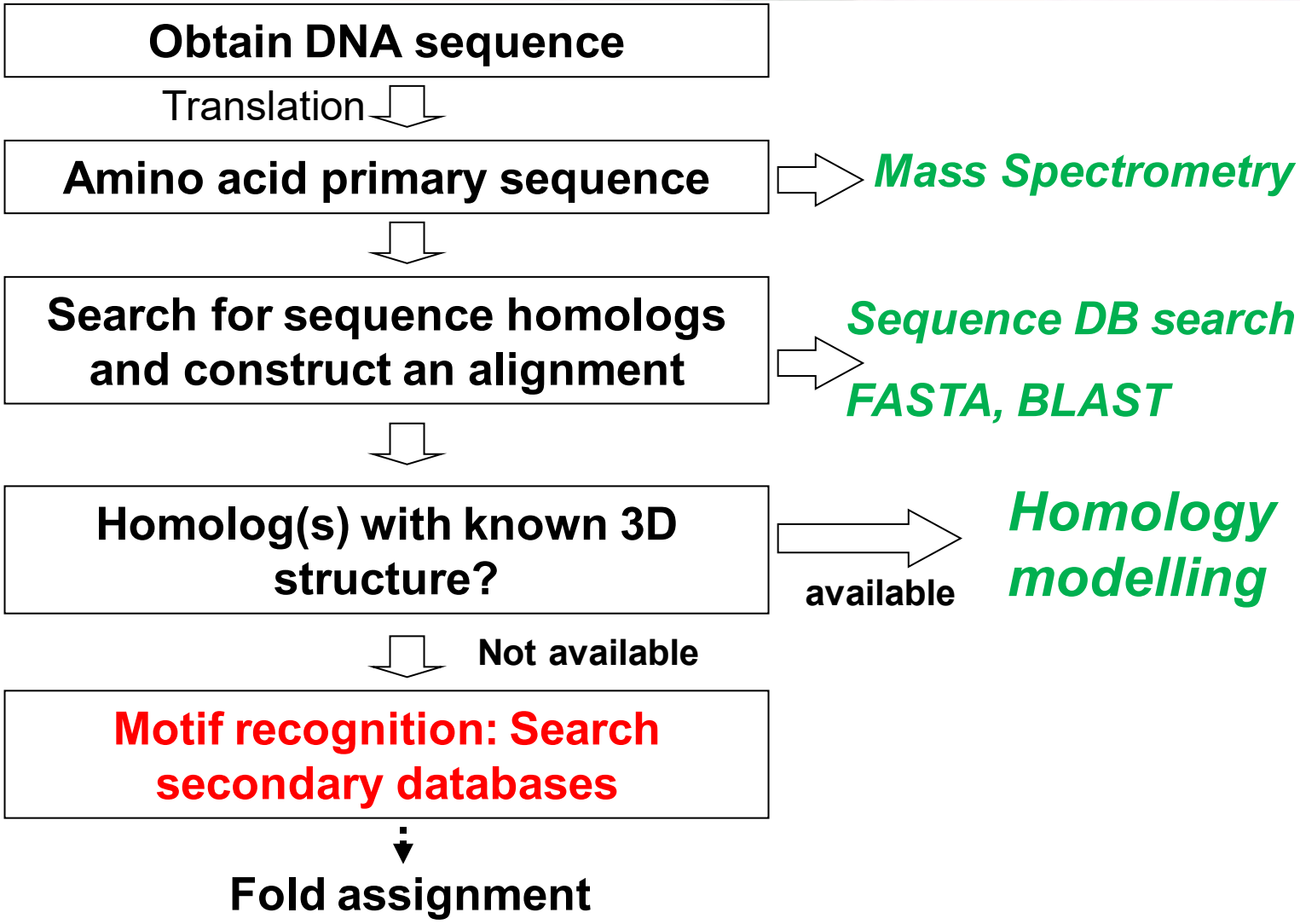
Loop modeling

Side-chain modeling

Model optimization

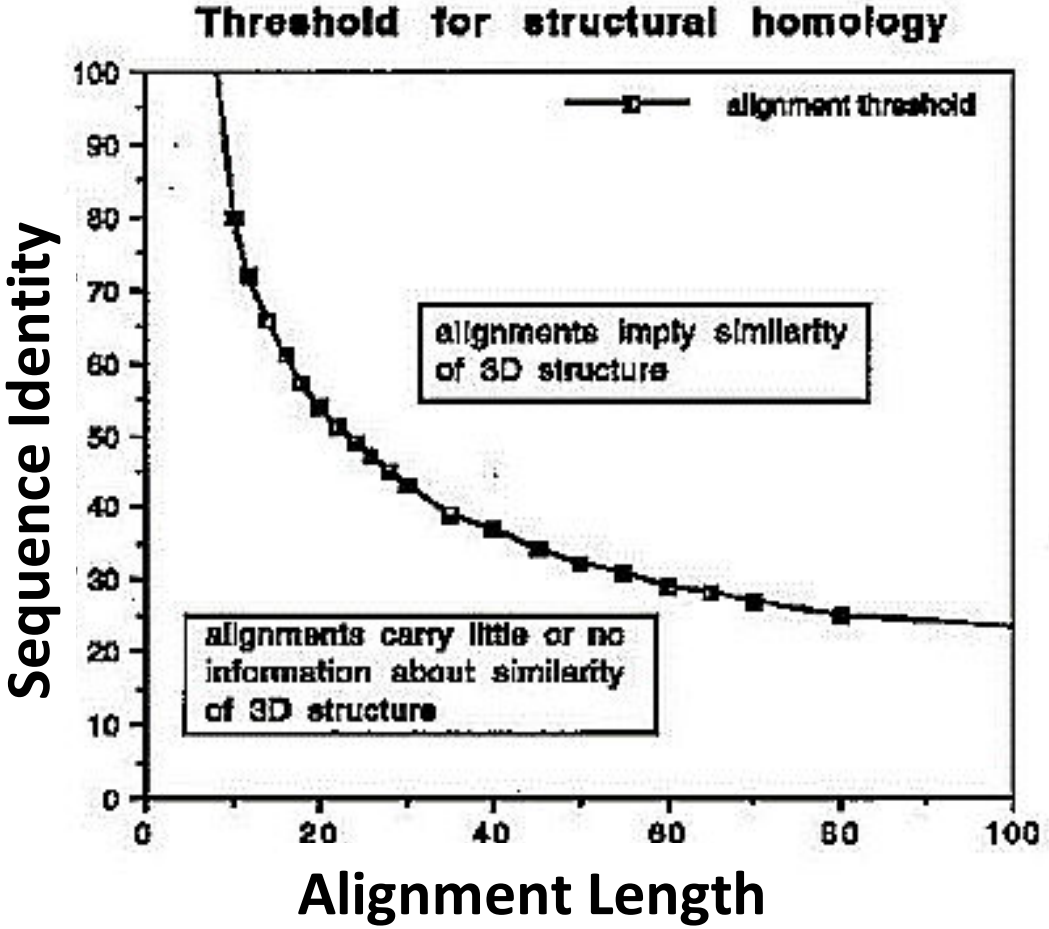
Model validation

Fold Recognition – Threading I



Fold Recognition – Threading I

When should we use Fold Recognition?



http://www.cmbi.ru.nl/gvteach/astra/lectures/homology_modelling.ppt

Fold Recognition – Threading I

Introduction

- A protein fold is defined by the way the secondary structure elements of the structure are arranged relative to each other in space.

- Common folds include 4-helix bundle and the TIM barrel.

Fold Recognition – Threading I

Introduction

- 5,000 stable folds in nature

- Fold recognition: Finding the best fit of a sequence to a set of candidate folds

Fold Recognition – Threading I

Conclusions

- Fold recognition or Threading is a technique for predicting protein structures
- It is useful in cases where homology modelling fails to predict quality structures

END

Homology Modelling

Fold Recognition – Threading II

Fold Recognition – Threading II

Background

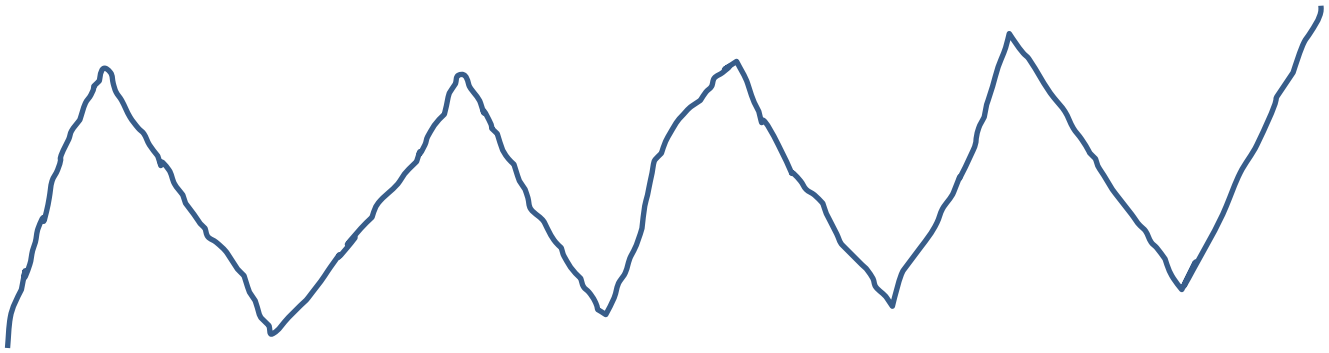
- Fold recognition is also called Threading
- Technique for predicting protein structures
- Employed when homology modelling cannot predict quality structures

Fold Recognition – Threading II

Find the best way to

“mount” the residue sequence of one protein

onto a known protein structure!



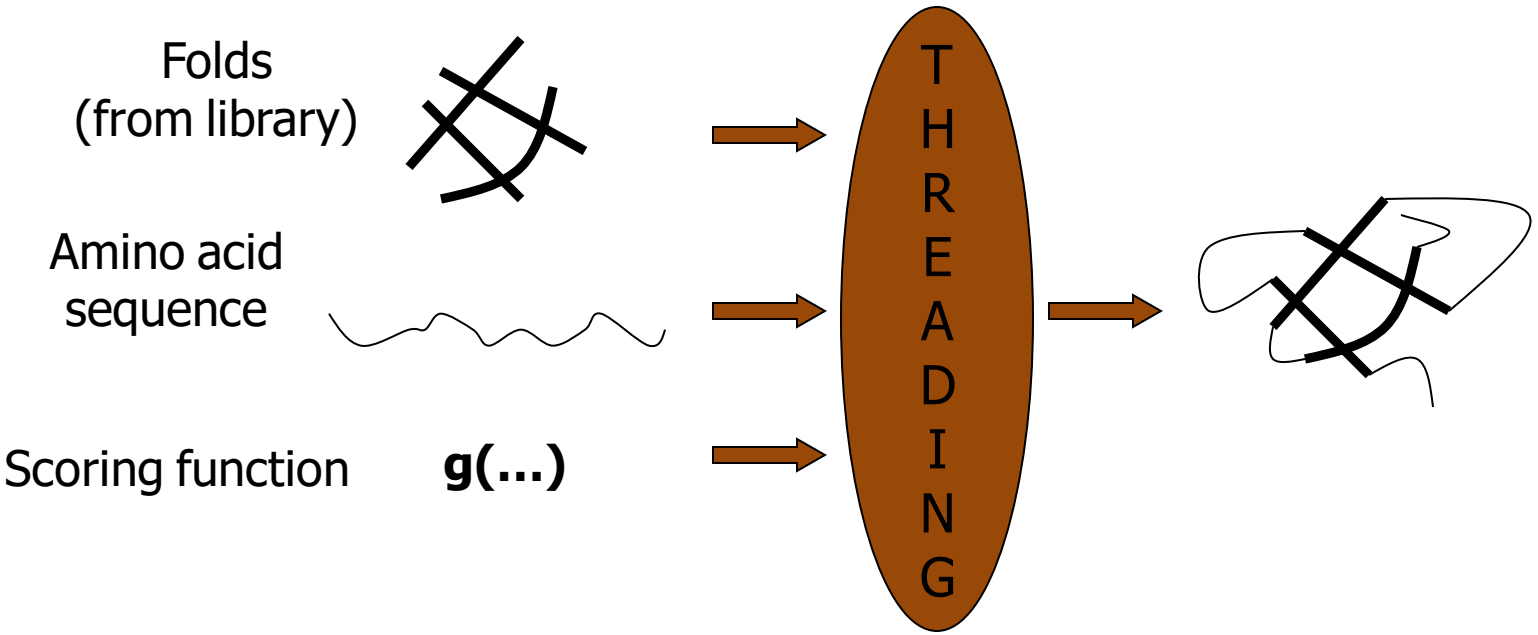
Fold Recognition – Threading II

The process of threading

- In the process of “Threading”, we mount an amino acid sequence on to the backbone of template structures in a folds library
- Each step is “drag” along the sequence (MQVKLFTY...) through each location of each template fold
- Then, for each fold, we must compute the fitness of sequence matching that fold!

Fold Recognition – Threading II

Inputs and outputs of threading



Fold Recognition – Threading II

Conclusions

- Threading involves "passing" the amino acid sequence through each fold in the database
- The best match is computed using a scoring function

END

Homology Modelling

Fold Recognition – Threading III

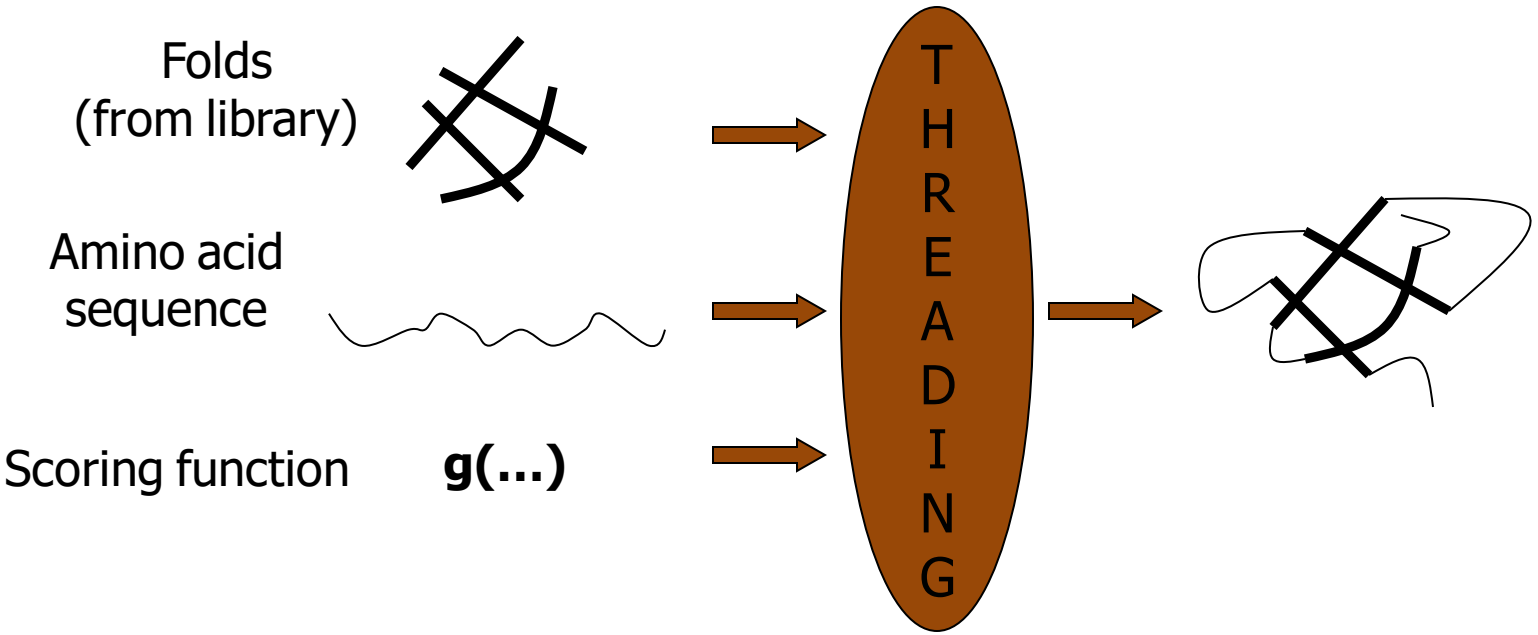
Fold Recognition – Threading III

Background

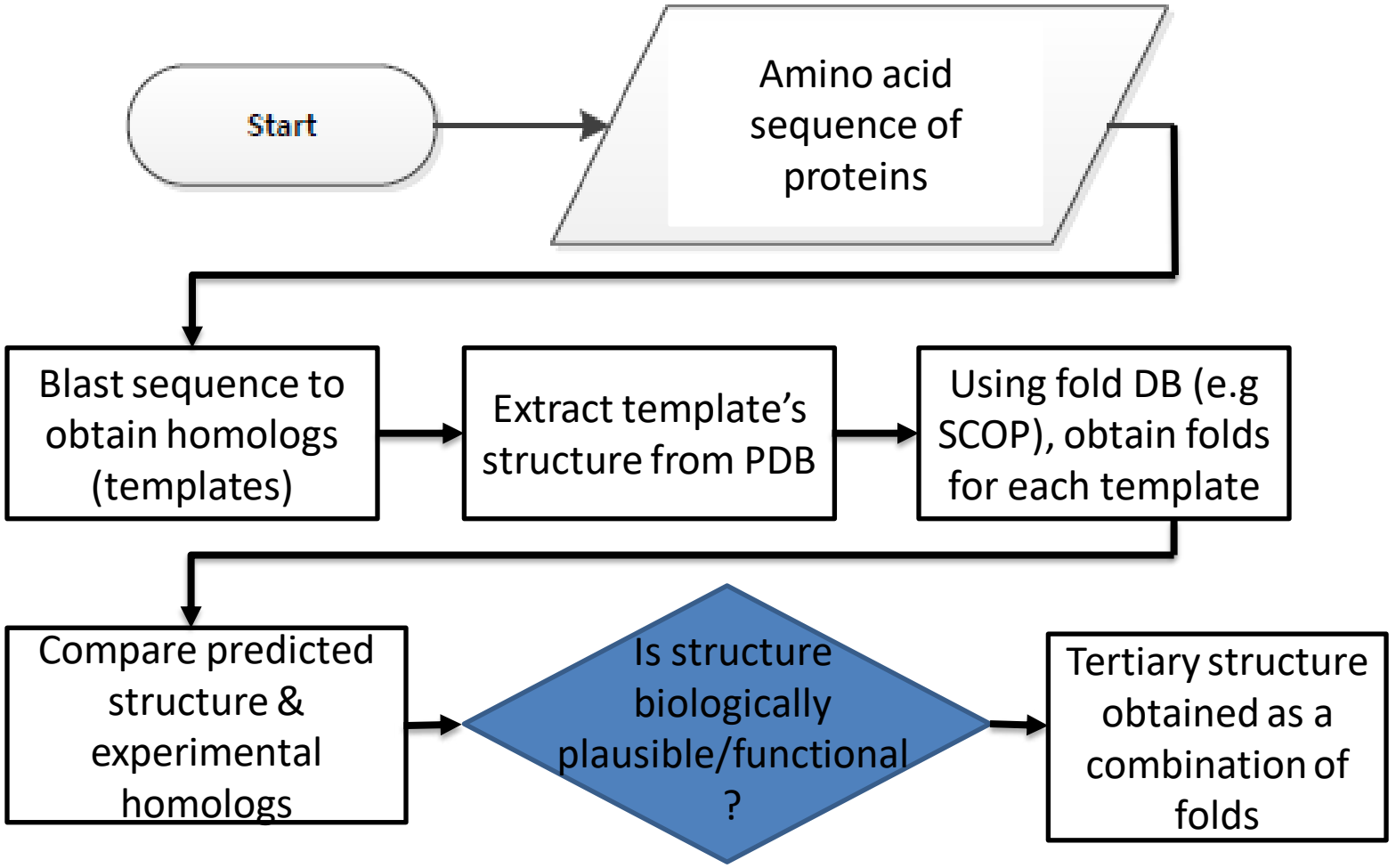
- Threading involves "passing" the amino acid sequence through each fold in the database
- The best match is computed using a scoring function
- Flowchart

Fold Recognition – Threading III

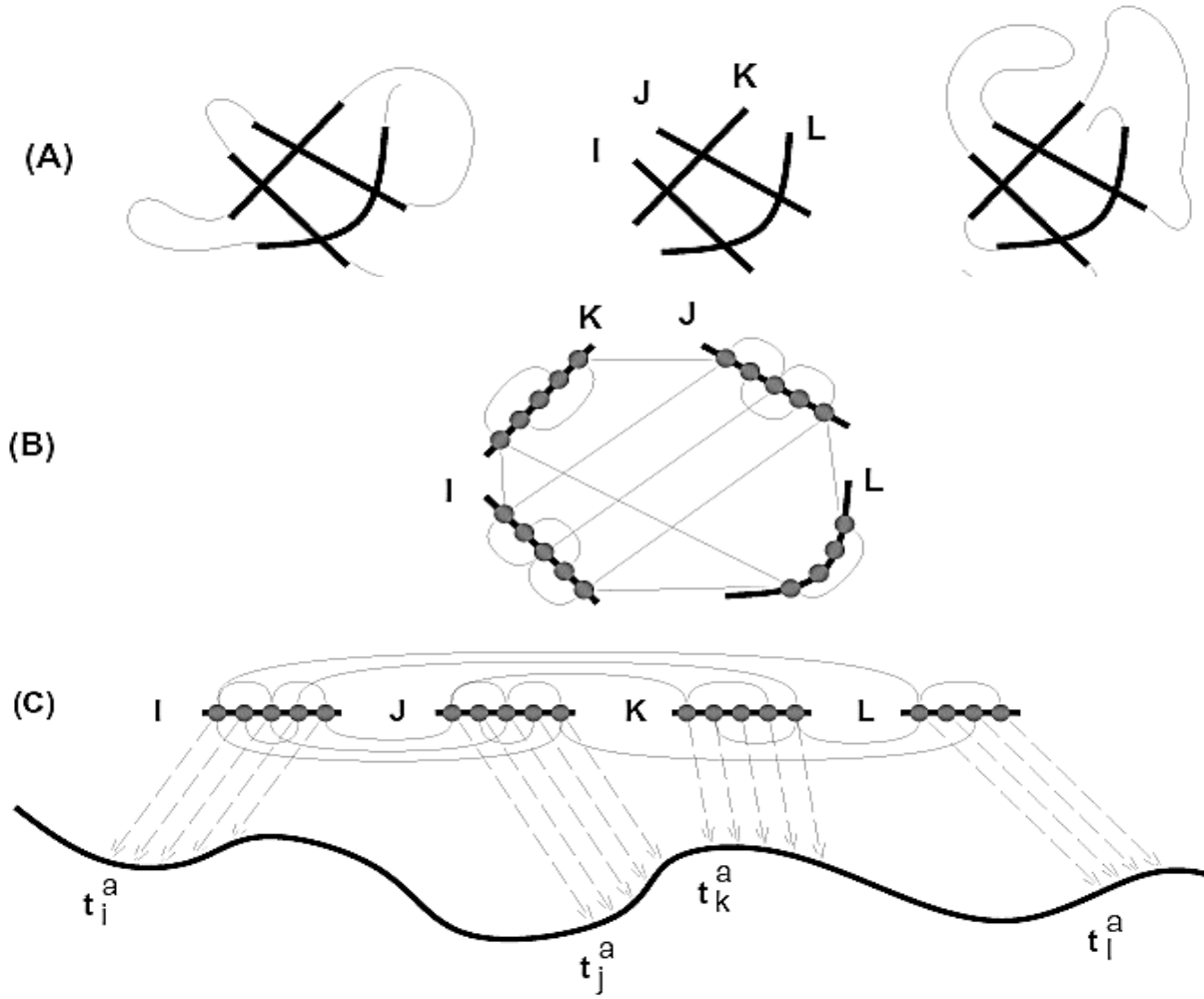
Inputs and outputs of threading



Fold Recognition – Threading III



Fold Recognition – Threading III



Two structurally similar proteins

Spatial adjacencies (interactions)

Possible threading with a sequence

Fold Recognition – Threading III

Conclusions

- Combinations of secondary structures come together to form the best prediction
- Scoring typically involves using a Z-Score function based on energy of a molecule

END

Homology Modelling

Online Tools for Threading - iTasser

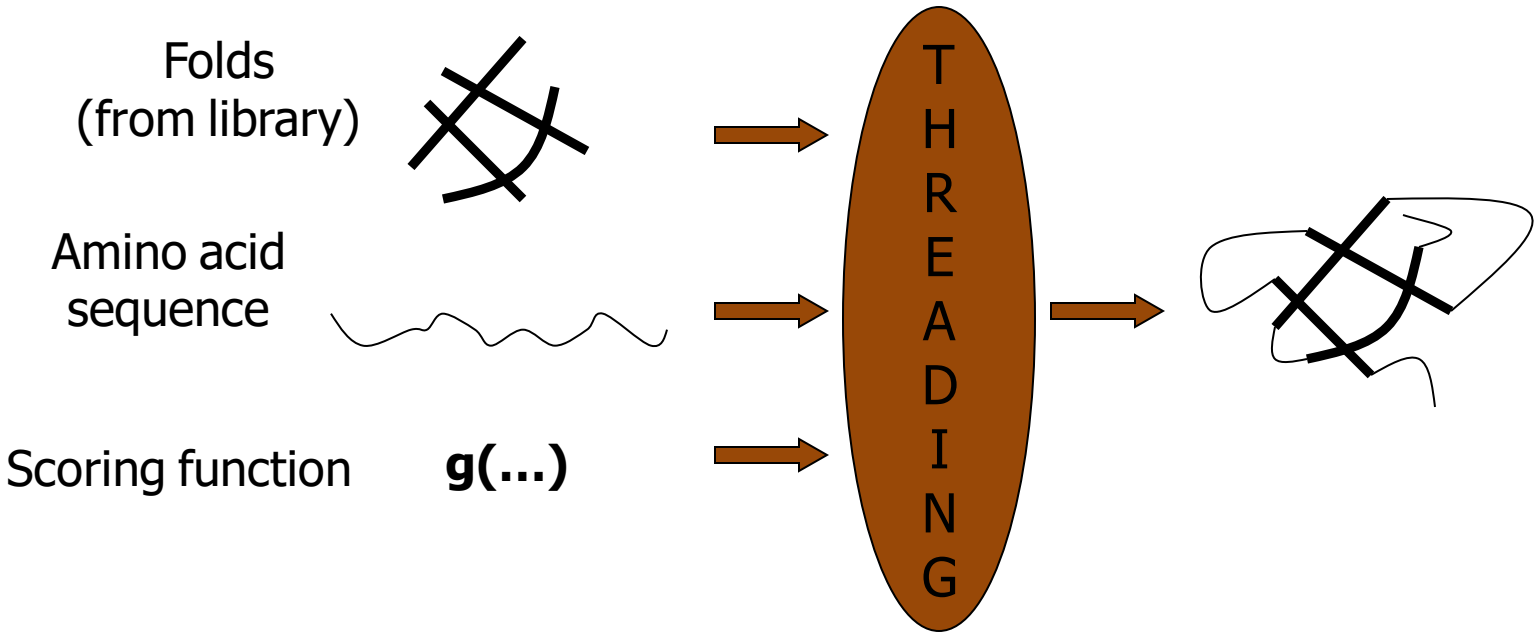
Online Tools for Threading - iTasser

Background

- Threading involves “passing” the amino acid sequence through each fold in the database
- The best match is computed using a scoring function
- iTASSER

Online Tools for Threading - iTasser

Inputs and outputs of threading



Online Tools for Threading - iTasser

Iterative threading assembly refinement (I-TASSER) server

- Software for automated protein structure &function prediction based on the sequence-to-structure-to-function.
- Steps:
  - Starts from amino acid sequence
  - i-TASSER first generates 3D atomic models from multiple threading alignments and iterative structural assembly simulations.
  - The function of the protein is then inferred by structurally matching the 3D models with other known proteins.
  - Outputs full-length secondary & tertiary structures and functional annotations on ligand-binding sites
  - An estimate of accuracy of the predictions is provided based on the confidence score of the modeling

Online Tools for Threading - iTasser

Link: http://zhanglab.ccmb.med.umich.edu/I-TASSER/

Online Tools for Threading - iTasser

Copy and paste your sequence below ([10, 1500] residues in FASTA format). Click here for a sample input:

Or upload the sequence from your local computer:

Choose File No file chosen

Email: (mandatory, where results will be sent to)

Password: (mandatory, please click here if you do not have a password)

ID: (optional, your given name of the protein)

▶ **Option I:** Assign additional restraints & templates to guide I-TASSER modeling.

▶ **Option II:** Exclude some templates from I-TASSER template library.

▶ **Option III:** Specify secondary structure for specific residues.

☑ Keep my results public (uncheck this box if you want to keep your job private. A key will be assigned for you to access tl
results)

Run I-TASSER Clear form

(Please submit a new job only after your old job is completed)

(If you want to submit multiple urgent jobs, please go to DNAStar which implements I-TASSER through cloud computing)

Online Tools for Threading - iTasser

Conclusions

- iTASSER helps thread amino acid sequences on fold and secondary structure databases
- It also helps predict function of structures output.

END

Homology Modelling

Advantages and Disadvantages of Threading

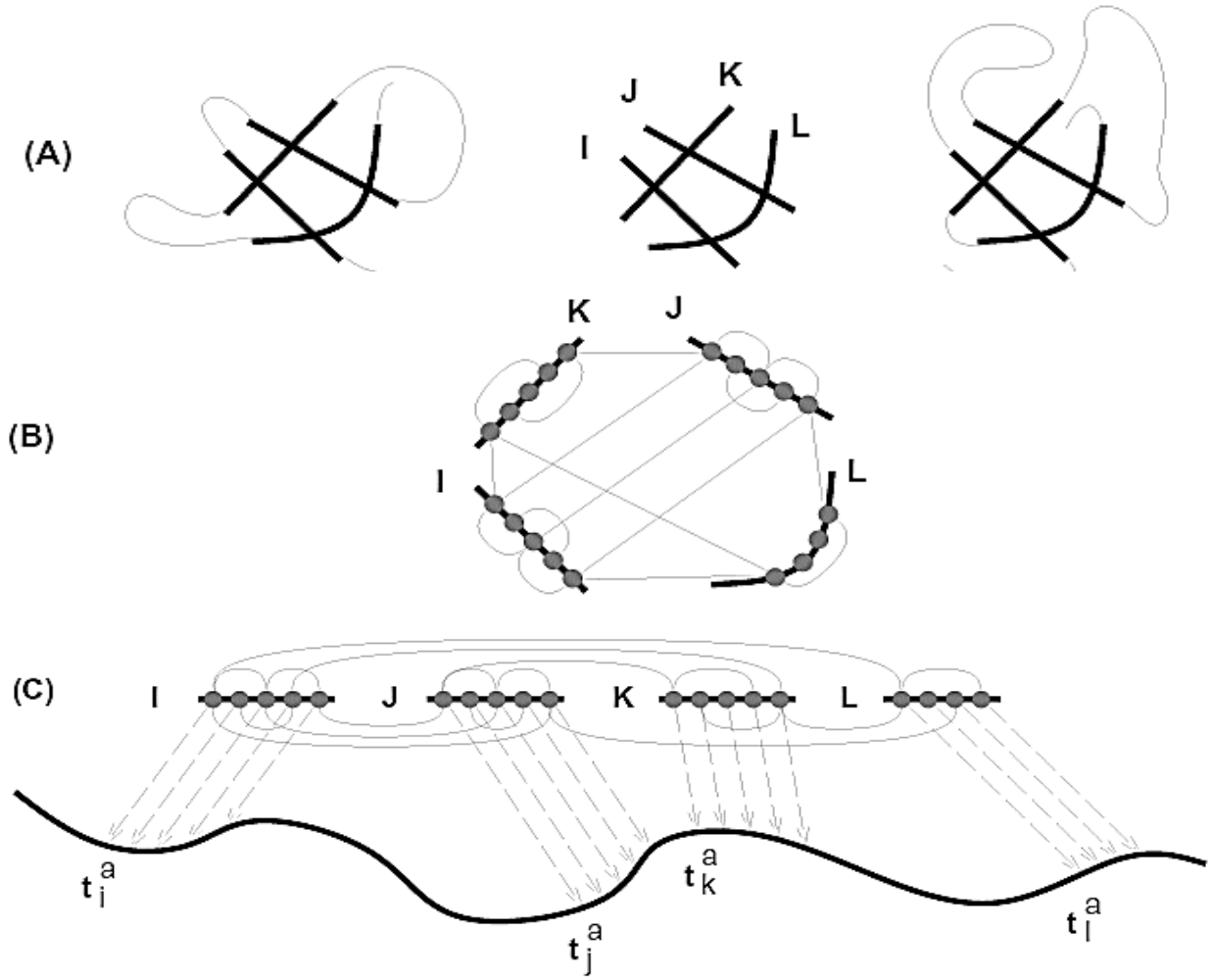
Advantages and Disadvantages of Threading

Background

- Fold recognition or Threading is a technique for predicting protein structures
- It is useful in cases where homology modelling fails to predict quality structures

Advantages and Disadvantages of Threading

Inputs and outputs of threading



Two structurally similar proteins

Spatial adjacencies (interactions)

Possible threading with a sequence

Advantages and Disadvantages of Threading

Advantages

- Threading helps predict secondary structures of proteins towards tertiary structure prediction
- For the “Twilight Zone” with low alignment quality and identity, threading is useful

Advantages and Disadvantages of Threading

Disadvantages

- Novel proteins cannot be predicted using threading
- Fewer than 30% of the predicted first hits are true remote homologues
- Validation of each result is necessary

END

Homology Modelling

3D-1D Bowie Algorithm

3D-1D Bowie Algorithm

Background

- Homology employed high alignment scores
- Threading worked by creating combinations of primary sequences and corresponding secondary structures

3D-1D Bowie Algorithm

Introduction

- Proposed by Bowie et al in 1991
- Converts 3D structure into a 1-D string profile for each structure in the fold library
- Align the target sequence to these profiles

3D-1D Bowie Algorithm

Inputs and outputs of 3D-1D

- Identify amino acids based on: protein core, side chain positioning, solubility etc. (6 in all)
- Part of secondary structure including α-helix, β-sheet etc (3 in all)
- Total of 3 x 6 = 18 distinct states

3D-1D Bowie Algorithm

Inputs and outputs of 3D-1D

$P_{a:j}$= prob. of finding amino acid (a) in environment (j)

P_a=probability of finding (a) anywhere

Maximize sum of scores for the fold:

$$s_{aj} = \log\left(\frac{P_{a:j}}{P_a}\right)$$

Advantages and Disadvantages of Threading

Conclusion

- 3D-1D methods convert structure and environment information into “profiles”
- Score for each amino acid is computed for each profile

END

Homology Modelling

Introduction to Ab Initio Modelling

Introduction to Ab Initio Modelling

Background

- Ab initio methods have Anfinsen's thermodynamic hypothesis at the center
- These methods attempt to identify the structure with minimum free energy

Introduction to Ab Initio Modelling

Need for Ab Initio Modelling

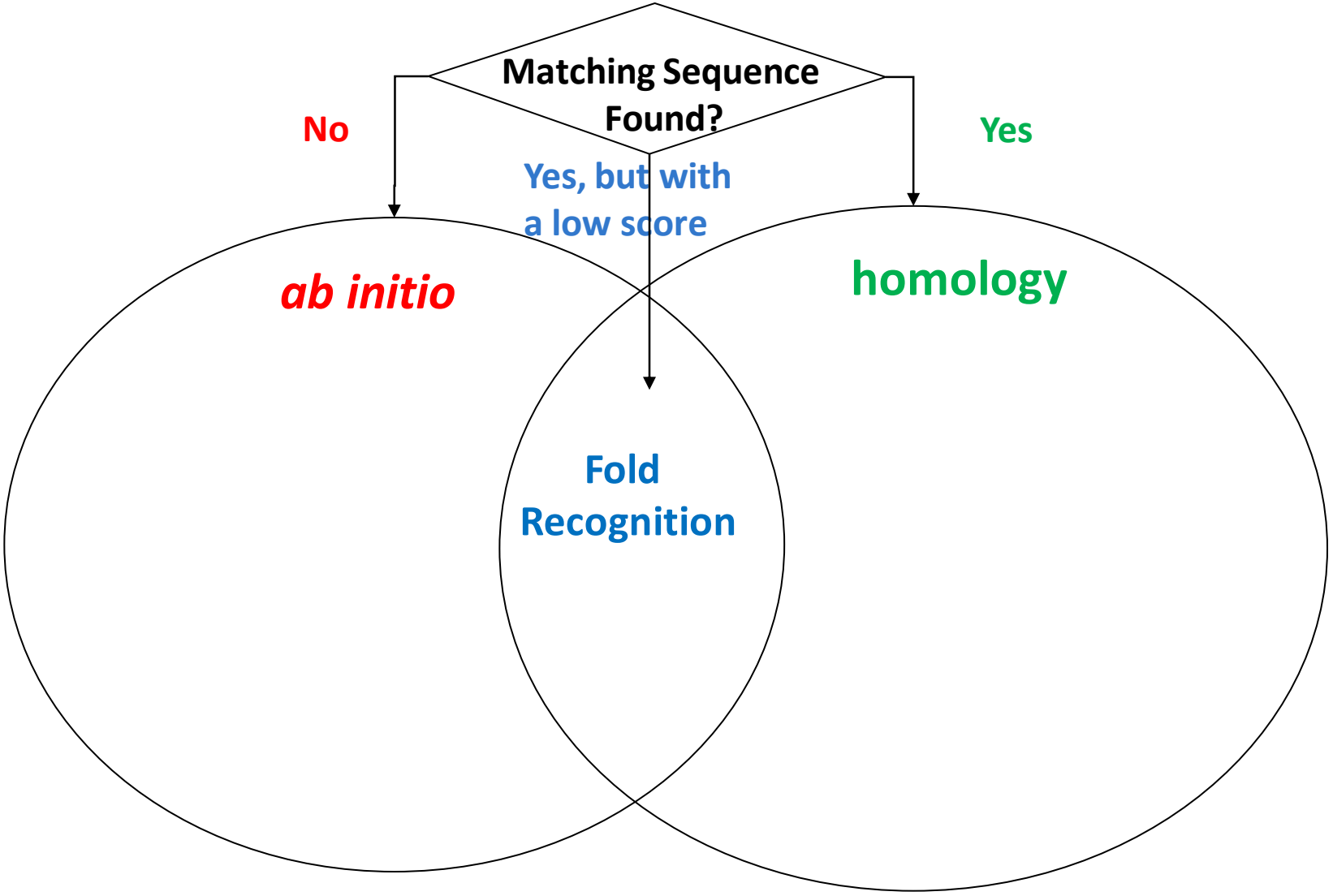
- Applicable to any sequence
- Not very accurate biologically
- Accuracy and applicability are limited by our understanding of the protein folding problem

Introduction to Ab Initio Modelling

Limitation

- Computationally expensive
- Suitable for proteins with less than 100 residues

Introduction to Ab Initio Modelling



Introduction to Ab Initio Modelling

Conclusion

- Ab initio methods rely on computing the energies of folded proteins

- The protein structures with the lowest energy are deemed as plausible predictions

END

Homology Modelling

Rationale of Ab Initio Modelling

Rationale of Ab Initio Modelling

Background

- Ab initio methods rely on computing the energies of folded proteins
- The protein structures with the lowest energy are declared as plausible predictions

Rationale of Ab Initio Modelling

Rationale

- Sometimes it so happens that even slightly homologous proteins may not be available.
- This renders homology modelling and threading/fold recognition as futile

Rationale of Ab Initio Modelling

Rationale

- Also, newer protein structures continue to be discovered every day
- These could not have been identified by methods which only rely on matching with available structures

Rationale of Ab Initio Modelling

Rationale

- Lastly, homology / fold recognition predict protein structures without computing fundamental physical/chemical properties of the mechanisms and driving forces in structure formation

Rationale of Ab Initio Modelling

Conclusion

- Ab initio methods, in contrast, base their predictions on physical models for these mechanisms
- Energy released during the folding process is computed for predicting structure

END

Homology Modelling

Strategies for Ab Initio Modelling

Strategies for Ab Initio Modelling

Background

- Ab initio methods base their predictions on physical models of folding mechanisms
- Stabilization is measured by energy released during the folding process

Strategies for Ab Initio Modelling

Energy Optimization in Ab Initio Modelling

1. Start with a rough initial model.

2. Define an energy function mapping structures to energy values. We have to minimize this later!!

3. Solve the computational problem of finding the global minimum.

Strategies for Ab Initio Modelling

Simulation of the Folding Process

1. Build an accurate initial model (including energy and forces).

2. Accurately simulate the dynamics of the protein folding process.

3. The native structure will steadily emerge.

Strategies for Ab Initio Modelling

Conclusion

- Start with an energy function
- Fold structures in order to obtain the most stable structure
- This structure will have the minimum energy

END

Homology Modelling

Energy States of Folded Proteins

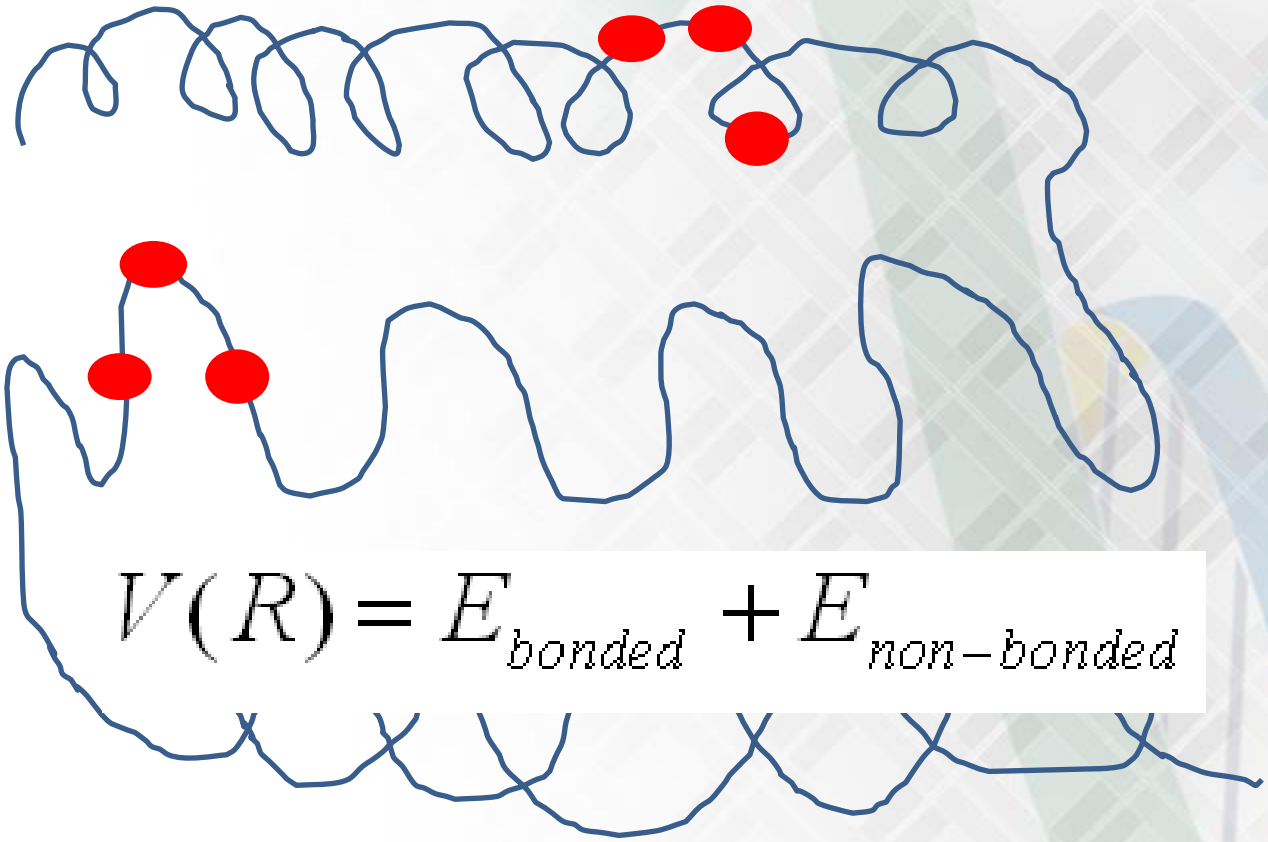
Energy States of Folded Proteins

Background

- Ab initio methods predict protein structures by folding proteins based on each constituent atom's volume, charge, mass etc.

Energy States of Folded Proteins

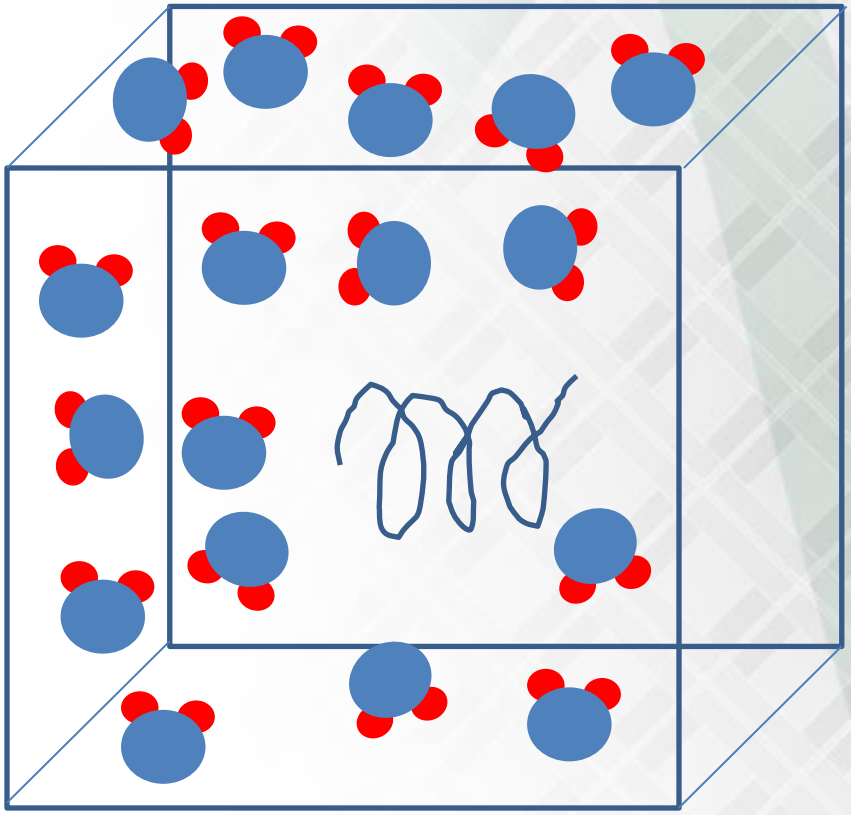
Energies of Bonded Atoms vs. Nonbonded Atoms



$$V(R) = E_{bonded} + E_{non-bonded}$$

Energy States of Folded Proteins

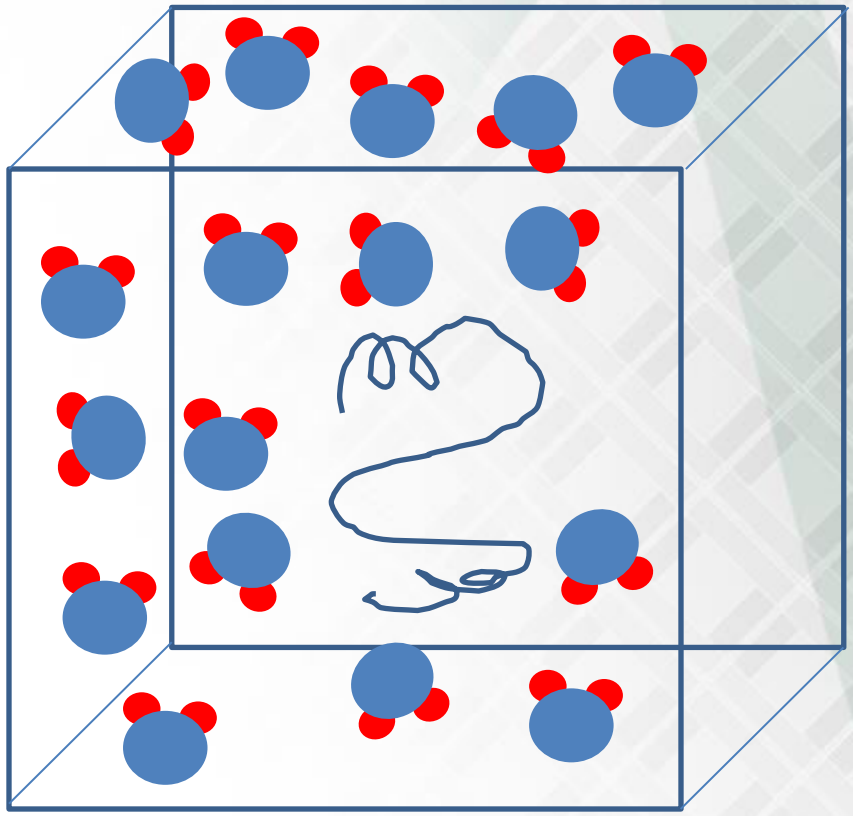
Force Fields for Energy Calculations
->Start with an initial structure



$$V(R) = E_{bonded} + E_{non-bonded}$$

Energy States of Folded Proteins

Force Fields for Energy Calculations



$$V(R) = E_{bonded} + E_{non-bonded}$$

Energy States of Folded Proteins

Conclusion

- The protein structure reporting lowest energy is selected to be the optimal structure
- How easy is it to compute the "really" lowest energy of a folded protein?

END

Homology Modelling

Local versus Global Minima

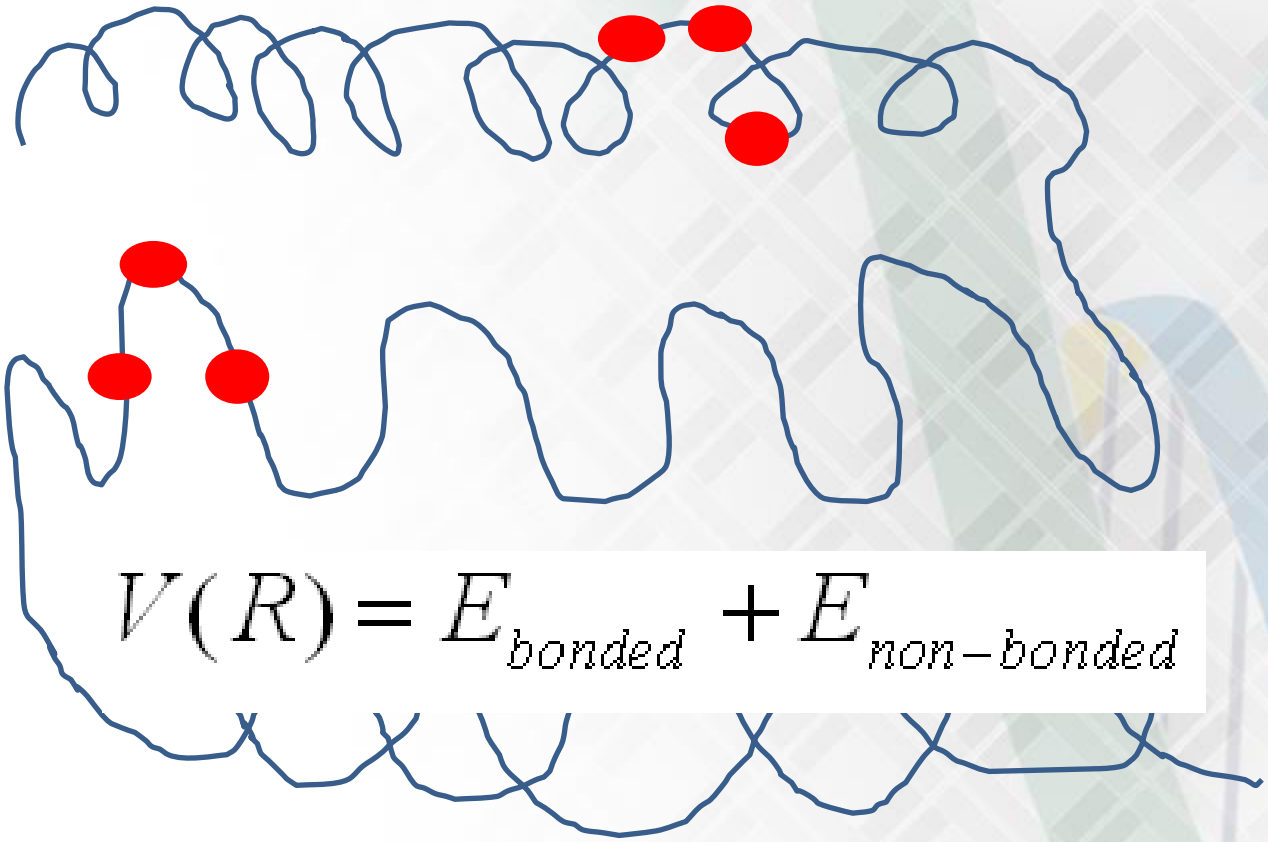
Local versus Global Minima

Background

- The protein structure reporting lowest energy is selected to be the optimal structure
- How easy is it to compute the “really” lowest energy of a folded protein?

Local versus Global Minima

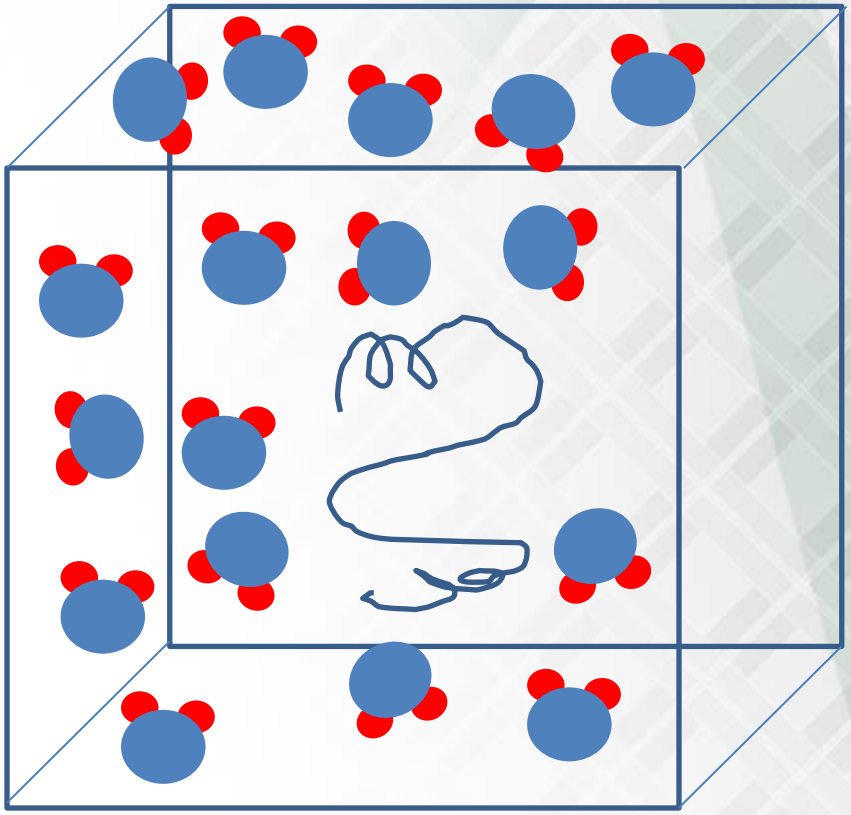
Energies of Bonded Atoms vs. Nonbonded Atoms



$$V(R) = E_{bonded} + E_{non-bonded}$$

Local versus Global Minima

Force Fields for Energy Calculations



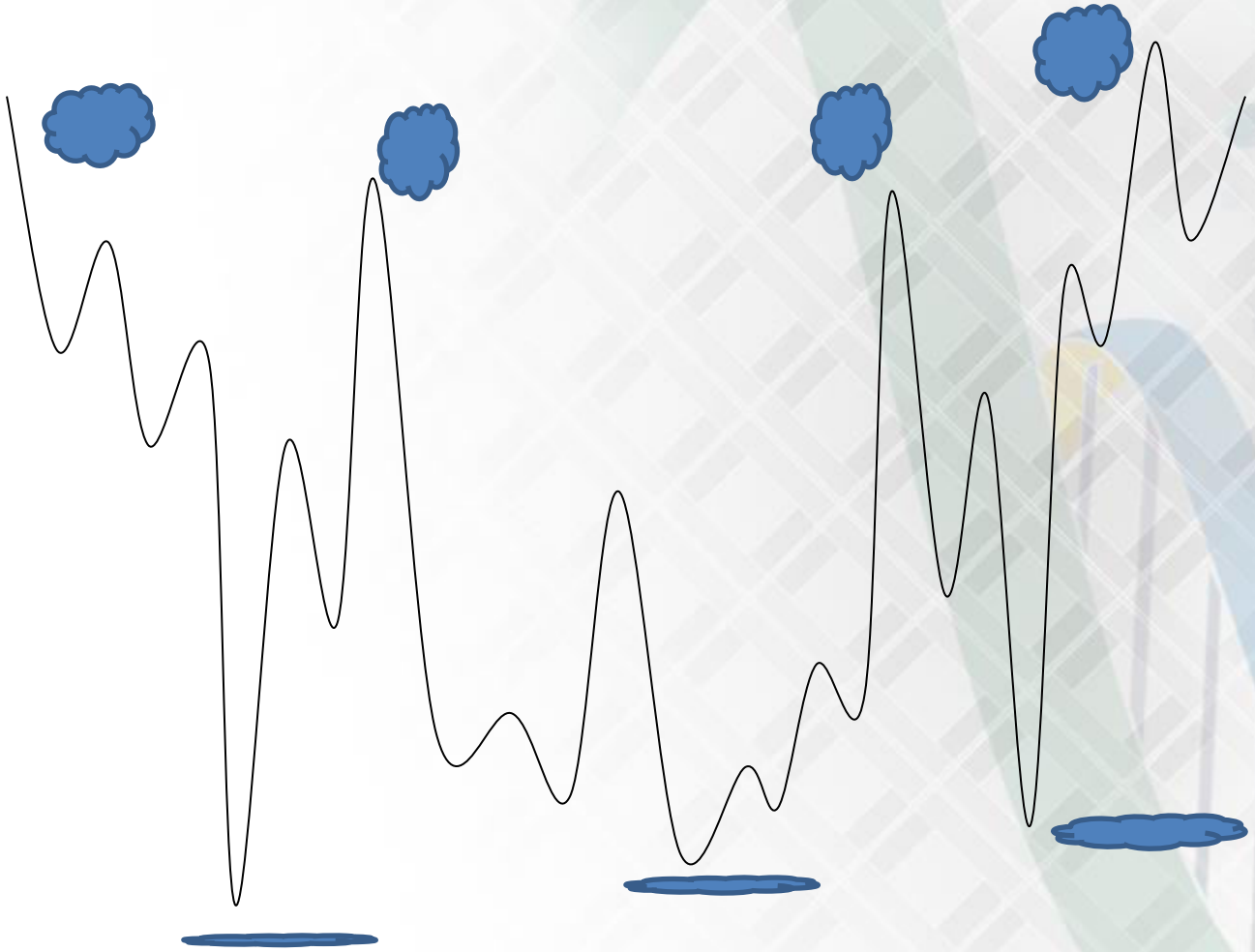
$$V(R) = E_{bonded} + E_{non-bonded}$$

Local versus Global Minima

Global Energy Optimization helps find global minimum

Local versus Global Minima

Global Energy Optimization helps find global minimum



Local versus Global Minima

Best Case Energy Function

- Clear energy minimum in the native structure
- Viable path towards this minimum
- Global optimization finds the most stable structure

Local versus Global Minima

Optimal Energy Function

- Easier to design and compute
- Native structure not always at the global minimum
- No clear way of choosing among alternative structures that are generated

END

Homology Modelling

Pros and Cons of Ab Initio Modelling

Pros and Cons of Ab Initio Modelling

Background

- Native structure not always at the global minimum
- No clear way of choosing among alternative structures that are generated

Pros and Cons of Ab Initio Modelling

Advantages

- Ab Initio methods can fold any target sequence using only physical atomic properties
- Predictions are mostly accurate and correctly describe the natural folding process

Pros and Cons of Ab Initio Modelling

Disadvantages

- Ab initio methods are the very difficult to design (energy function)
- These methods are slow due to the huge possibilities

Pros and Cons of Ab Initio Modelling

Disadvantages

- An order of 10^{12} steps are needed to simulate protein folding for medium sized protein structures

Pros and Cons of Ab Initio Modelling

Challenges in Ab Initio Modelling

- Very hard to accurately describe energy functions that can reliably discriminate native and non-native structures.
- Enormous amount of computations.

END

Homology Modelling

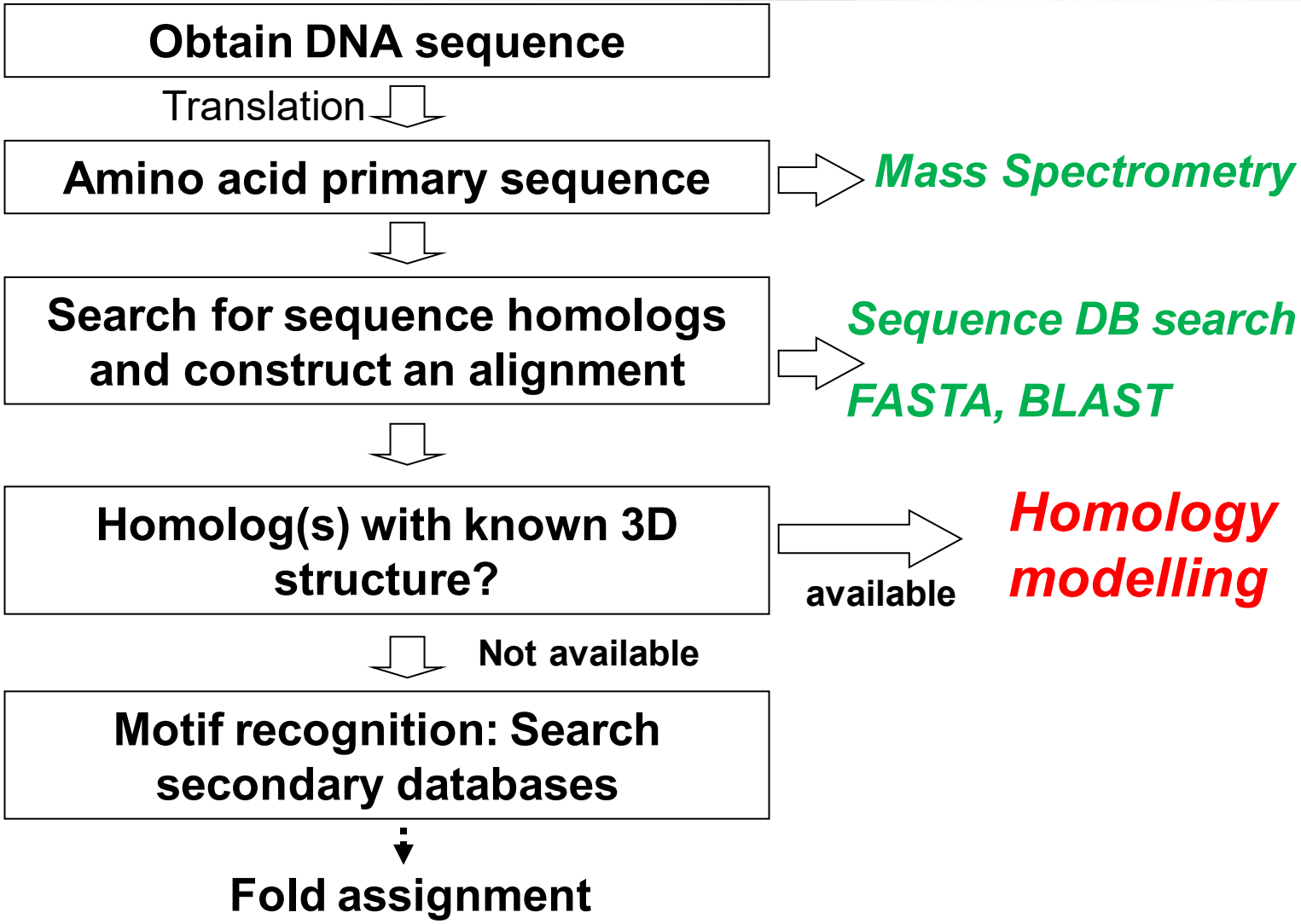
Summary of Structural Modelling – I

Summary of Structural Modelling - I

Strategies for Structural Modelling

- Homology Modelling
- Fold Recognition
- Ab Initio Modelling

Summary of Structural Modelling - I



Summary of Structural Modelling - I

Homology modeling of the target structure can be done as follows:

1. Template recognition and initial alignment
2. Alignment correction
3. Backbone generation
4. Loop modeling
5. Side-chain modeling
6. Model optimization
7. Model validation

Summary of Structural Modelling - I

Conclusion

- Homology modelling is performed in cases of high identity and alignment score
- For the “Twilight zone”, other strategies are employed

END

Homology Modelling

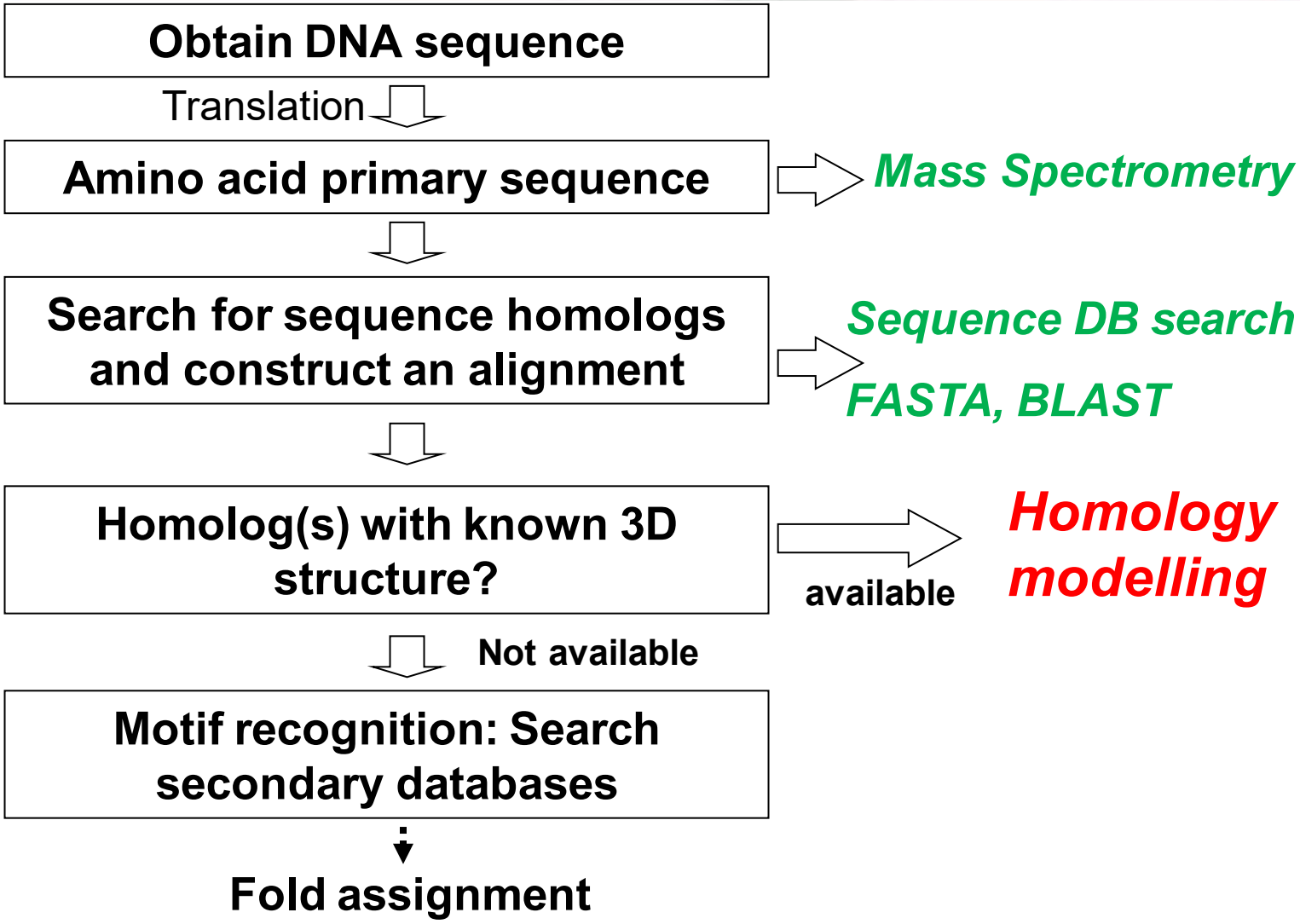
Summary of Structural Modelling – II

Summary of Structural Modelling - II

Strategies for Structural Modelling

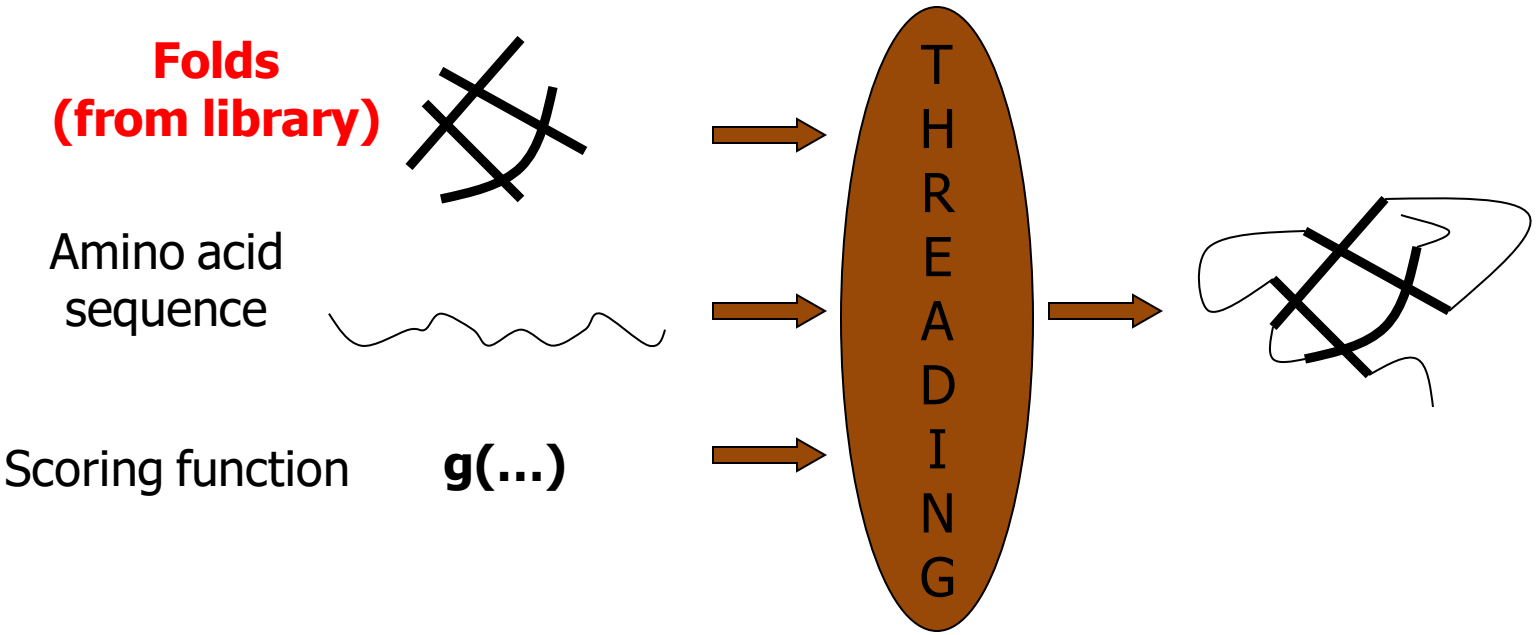
- Homology Modelling
- Fold Recognition
- Ab Initio Modelling

Summary of Structural Modelling - II

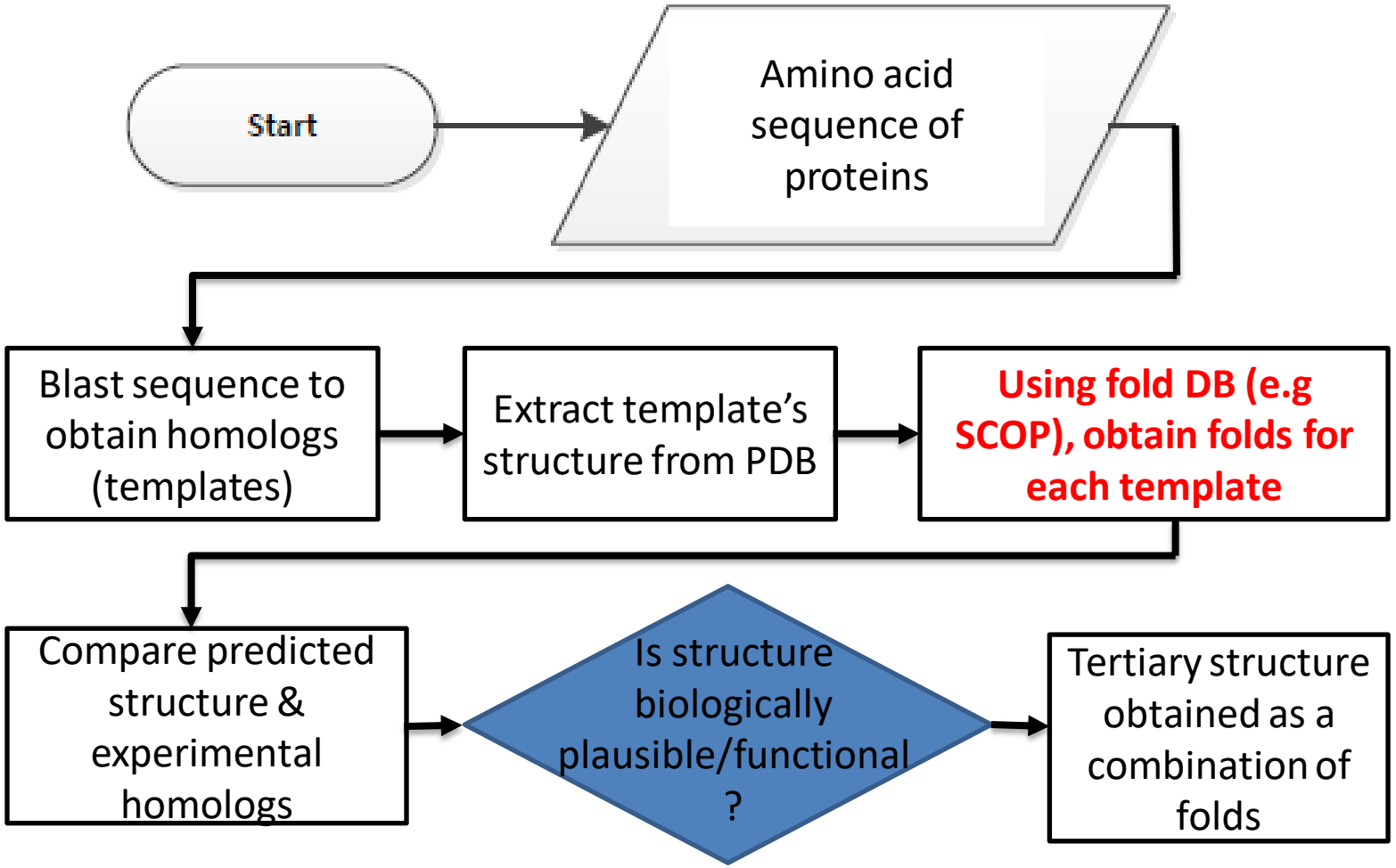


Summary of Structural Modelling - II

Inputs and outputs of threading



Summary of Structural Modelling - II



Summary of Structural Modelling - II

Conclusion

- For low identity and alignment scores, a "Twilight zone" for structure prediction exists
- Fold recognition / threading is useful in such cases

END

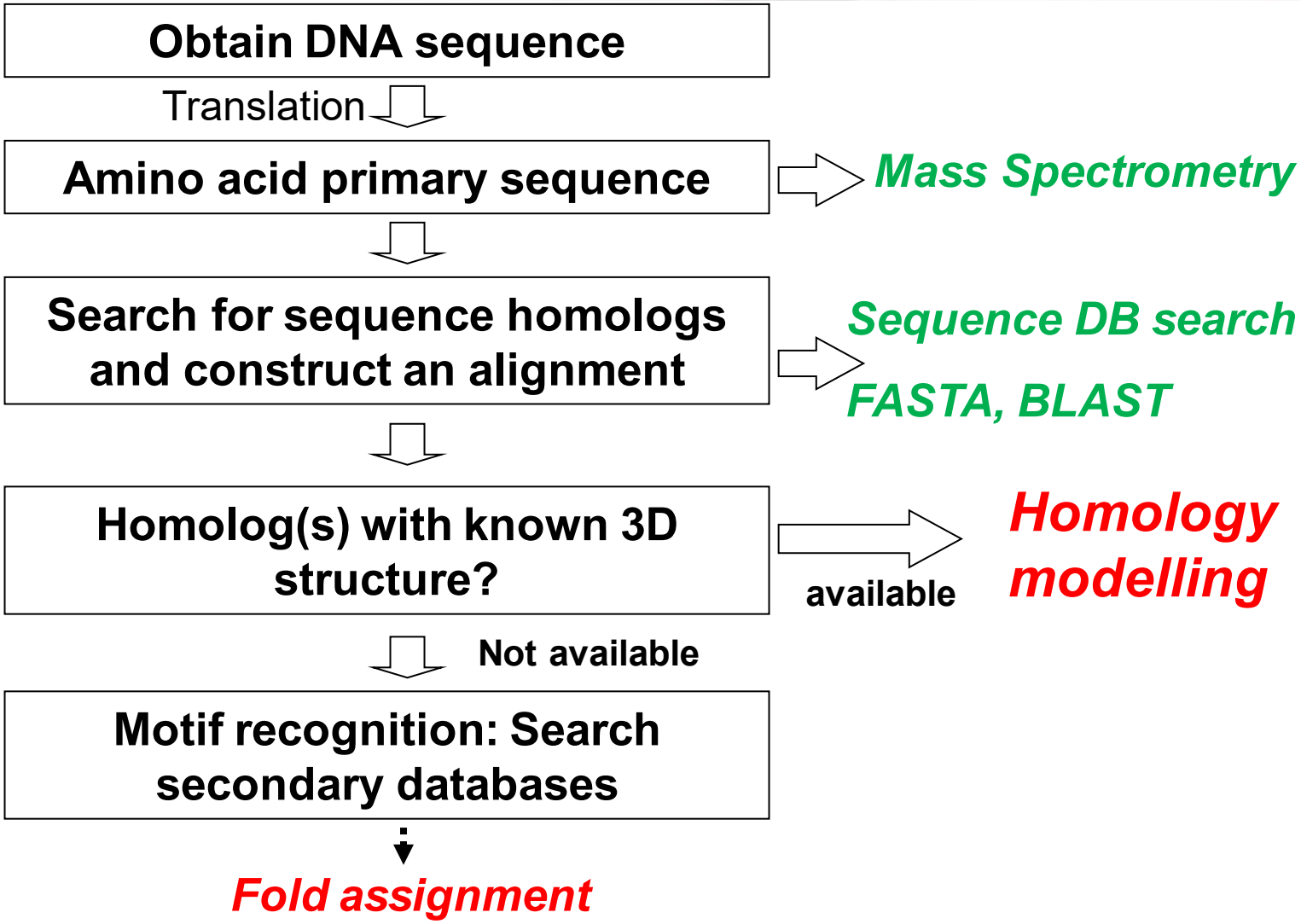
Summary of Structural Modelling – III

Summary of Structural Modelling - III

Strategies for Structural Modelling

- Homology Modelling
- Fold Recognition
- Ab Initio Modelling

Summary of Structural Modelling - III



Summary of Structural Modelling - III

Energy Optimization in Ab Initio Modelling

1. Start with a rough initial model.

2. Define an energy function mapping structures to energy values. We have to minimize this later!!

3. Solve the computational problem of finding the global minimum.

Summary of Structural Modelling - III

Simulation of the Folding Process

1. Build an accurate initial model (including energy and forces).

2. Accurately simulate the dynamics of the protein folding process.

3. The native structure will steadily emerge.

Summary of Structural Modelling - III

Conclusion

- For cases where even the fold libraries do not give any high scoring matches, Ab Initio strategies can help model the structure
- However, this is a complex and computationally expensive process

END

Conclusion of the Course

Review of Sequence Analysis

Review of Sequence Analysis

Important Concepts

How do we sequence:

- Genomes
- Proteomes

Review of Sequence Analysis

Important Concepts

How do we compare sequences:

- Pair-wise Sequence Alignment
- Multiple Sequence Alignment

Review of Sequence Analysis

Important Concepts

Types of Alignments:

Global Alignment
(Needle Wunsch)

Local Alignment
(Smith Waterman)

Review of Sequence Analysis

Important Concepts

Advanced Tools:

Fast Alignment (FASTA)

Basic Local Alignment Search Tool (BLAST)

Review of Sequence Analysis

Important Concepts

Databases:

GenBank

UniProt

Review of Sequence Analysis

Important Concepts

Online Portals:

Ensemble

Expasy

UniProtKB

END

Conclusion of the Course

Review of Phylogenetics

Review of Phylogenetics

Important Concepts

- Molecular Evolution
- Insertions
- Deletions
- Substitutions

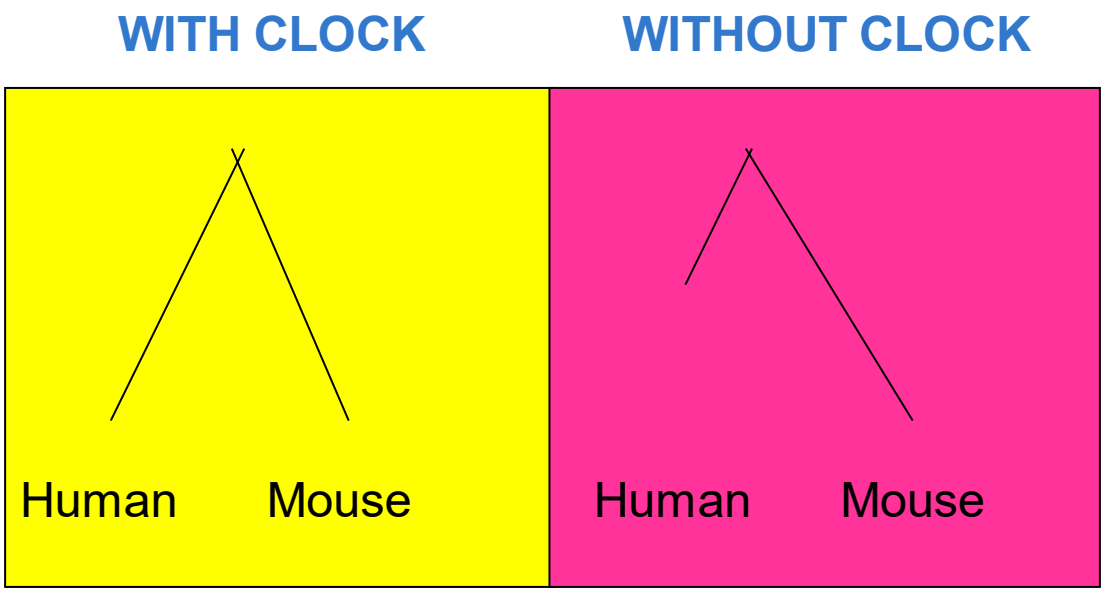
Review of Phylogenetics

Important Concepts

- Phylogenetic Trees
- Scaled Trees
- Unscaled Trees

Review of Phylogenetics

Edge length with and without a clock!



Since the rate of evolution in species is different, it needs to be considered!

Review of Phylogenetics

Important Concepts

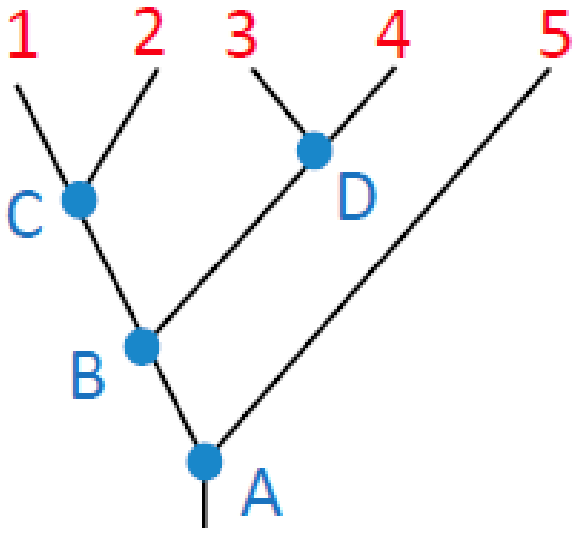
- Phylogenetic Trees
- Rooted Trees
- Unrooted Trees

Review of Phylogenetics

UPGMA: Unweighted Pair – Group Method using arithmetic Averages

- Two sequences with with the **shortest evolutionary distance** between them are considered

- These sequences will be **the last to diverge**, and **represented by the most recent internal node.**



Review of Phylogenetics

Important Concepts

Clustering Vs. Non-clustering Methods:

UPGMA is a clustering method

Maximum Parsimony etc are non-clustering methods (not included in this course).

END

Conclusion of the Course

Review of Protein Sequencing

Review of Protein Sequencing

Important Concepts

- Techniques of protein sequencing
- Edman Degradation
- Mass Spectrometry

Review of Protein Sequencing

Important Concepts

- Protein Ionization
- Mass Analysis
- Protein Fragmentation

Review of Protein Sequencing

Important Concepts

- MS1
- MS2

Review of Protein Sequencing

Important Concepts

- Estimating and scoring whole protein mass

Review of Protein Sequencing

Important Concepts

- Extracting & Scoring Peptide Sequence Tags

Review of Protein Sequencing

Important Concepts

- Searching Post-translational Modifications

Review of Protein Sequencing

Important Concepts

Composite Scoring Schemes

Online tools:

- **Mascot**
- **Sequest**
- **Prosight PC**

END

Conclusion of the Course

Review of Protein Structures

Review of Protein Structures

Important Concepts

Protein Structures are generally of four types:

- Primary
- Secondary
- Tertiary
- Quaternary

Review of Protein Structures

Important Concepts

Techniques for determining protein structures

- **X-Ray Crystallography**
- **NMR Spectroscopy**

Review of Protein Structures

Important Concepts

- Why number of known protein sequences is much larger as compared to known proteins structures?

Review of Protein Structures

Important Concepts

Types of Protein Secondary Structures

- Helices
- Beta Sheets
- Coils
- Loops

Review of Protein Structures

Important Concepts

- Foundation of structure prediction algorithms
- Propensities of certain amino acids to form specific secondary structures

Review of Protein Structures

Important Concepts

- Algorithm for predicting protein structures
- Chou Fasman Algorithm

Review of Protein Structures

Important Concepts

- Protein Structure Database - PDB
- Online tools for predicting structures by using proteins sequences

END

Conclusion of the Course

Review of Protein Structures

Review of Protein Structures

Important Concepts

Protein Structures are generally of four types:

- Primary
- Secondary
- Tertiary
- Quaternary

Review of Protein Structures

Important Concepts

Techniques for determining protein structures

- **X-Ray Crystallography**
- **NMR Spectroscopy**

Review of Protein Structures

Important Concepts

- Why number of known protein sequences is much larger as compared to known proteins structures?

Review of Protein Structures

Important Concepts

Types of Protein Secondary Structures

- Helices
- Beta Sheets
- Coils
- Loops

Review of Protein Structures

Important Concepts

- Foundation of structure prediction algorithms
- Propensities of certain amino acids to form specific secondary structures

Review of Protein Structures

Important Concepts

- Algorithm for predicting protein structures
- Chou Fasman Algorithm

Review of Protein Structures

Important Concepts

- Protein Structure Database - PDB
- Online tools for predicting structures by using proteins sequences

END

Conclusion of the Course

Review of Homology Modelling

Review of Homology Modelling

Important Concepts

Four Strata of Protein Structures

- Primary
- Secondary
- Tertiary
- Quaternary

Review of Homology Modelling

Important Concepts

- Justification for homology modelling
- Number of known protein sequences is much larger as compared to known proteins structures

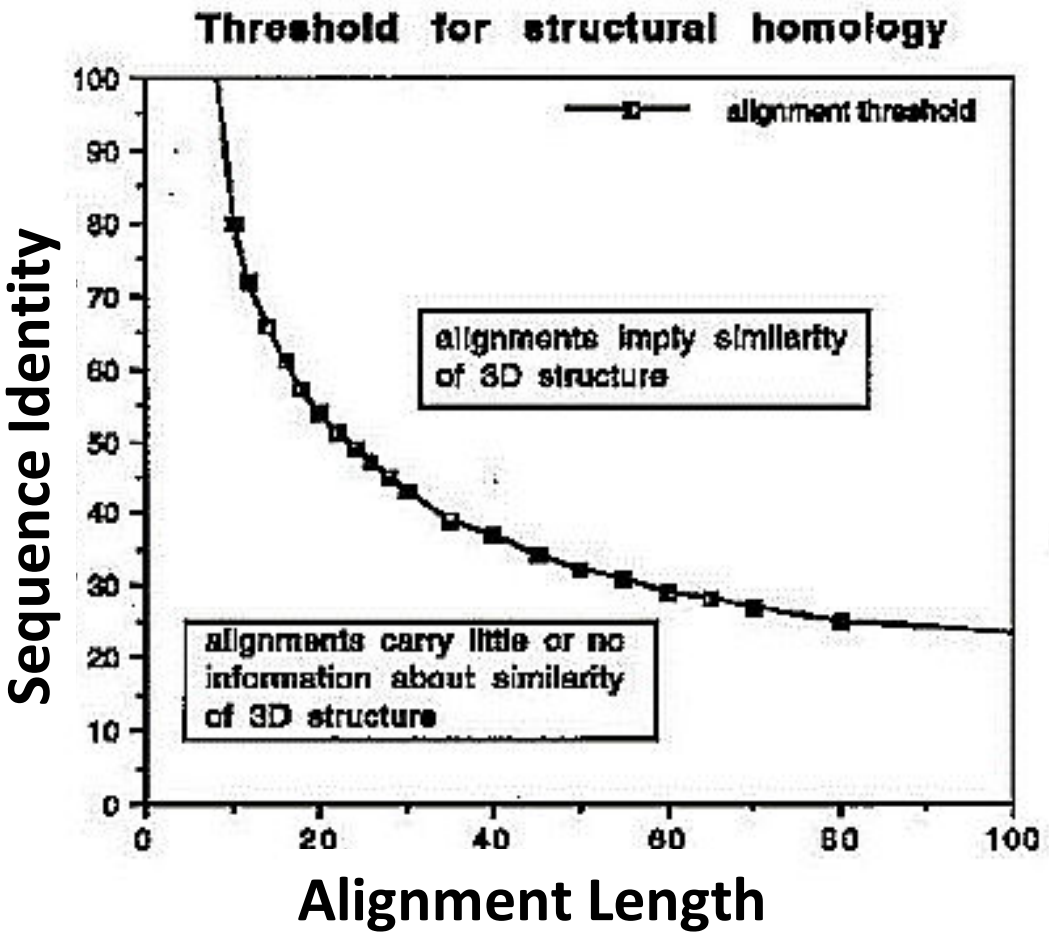
Review of Homology Modelling

Important Concepts

Three Strategies for Structure Prediction

- Homology Modelling
- Fold Recognition
- Ab Initio Modelling

Review of Homology Modelling



http://www.cmbi.ru.nl/gvteach/astra/lectures/homology_modelling.ppt

Review of Protein Structures

Important Concepts

- Protein Structure Database - PDB
- Online tools for predicting structures such as MODELLER and iTASSER

END

Conclusion of the Course

Conclusions from this Course

Conclusions from this Course

Important Concepts

- Definition of Bioinformatics
- Need for Bioinformatics
- Areas within Bioinformatics

Conclusions from this Course

Important Concepts

- Bioinformatics as an interdisciplinary area
- Need to store, process and analyze biological data
- Requirement of newer faster algorithms

Conclusions from this Course

Important Concepts

Specific areas focused were:

- Comparing sequences
- Comparing structures
- Predicting structures

Conclusions from this Course

Important Concepts

We looked at:

- Algorithms
- Databases
- Online Tools

for each topic.

Conclusions from this Course

Important Concepts

- We studied the basic algorithms for each topic
- With evolution and growth of Bioinformatics, newer and better algorithms are now also available!

END

Conclusion of the Course

Advanced Follow-up Courses

Advanced Follow-up Courses

Important Concepts

- We looked into the foundations of Bioinformatics
- However, each topic that was studied has a undergone a lot of development

Advanced Follow-up Courses

Important Concepts

- For advanced study in Genomics, you may take "Computational Genomics" course

- Topics:
  Genome Assembly, Gene Finding, Annotation, GWAS etc

Advanced Follow-up Courses

Important Concepts

- For advanced study in Proteomics, you may take "Computational Proteomics" course.

- Topics:
  Protein Sequencing, PTM search, Structure Modelling and PPI studies

Advanced Follow-up Courses

Important Concepts

- For advanced study in Integrative Biology, you may take “Systems Biology” course.

- Topics: Metabolomics, Transcriptomics, Network Biology etc

Advanced Follow-up Courses

Important Concepts

Also, now there are cutting edge courses on:

- **Nano-Bio-IT**
- **Computational Drug Design**
- **Personalized Medicine**

END

Conclusion of the Course

Careers in Bioinformatics

Careers in Bioinformatics

Background

- Pakistan as an infrastructure-limited country
- The onset of digital revolution
- Emergence of data as the most precious commodity, globally

Careers in Bioinformatics

Background

- Specifically, health data as a key commodity of the future
- Health and disease as the primordial challenge of mankind

Careers in Bioinformatics

Background

- Unique opportunity for us in Pakistan
- Bioinformatics requires two things

1. Smart mind
2. Internet connected computer

Careers in Bioinformatics

One man company

- You can take public databases and design drugs!
- One man vs. Roche?

Careers in Bioinformatics

BIGDATA

- You can make a startup company which manages and process health BIGDATA!
- All it needs is basic software development skills coupled with Bioinformatics

Careers in Bioinformatics

The next disruption

- The next Google, Facebook and Uber is going to emerge from Health and Bioinformatics
- Pharmaceutical companies are investing into bioinformatics human resource development

Careers in Bioinformatics

Jobs Market

- Pharmaceutical Giants
- Research Centers & Universities
- Hospital & Diagnostic IT departments
- Your own startup company

END

Final Term Syllabus

Effort By

Amaan Khan