SEMESTER V

16BTU502A BIOINFORMATICS 4H - 4C

Total hours/week: L:4 T:0 P:0 Marks: Internal: 40 External: 60 Total: 100

Scope: This course demonstrates the application of the information technology in data management and applications in concepts to biological problems. Subsequently, this course provides insight into the biological databases and the student should be employable by the biopharmaceutical industry as a data analyzer and/or bioinformatician.

Objective: The objective of the course is to introduce fundamental concepts in bioinformatics.

UNIT-I

History and milestone of Bioinformatics. The notion of Homology. Sequence Information Sources, EMBL, GENBANK, Entrez, Unigene, Understanding the structure of each source and using it on the web. Genome sequencing projects – Steps, Human Genome Project and other genome projects.

UNIT-II

Basic concepts of biomolecules – Protein and amino acid, DNA and RNA - Sequence, Structure and function. Protein Information Sources, PDB, SWISSPROT, TREMBL, Understanding the structure of each source and using it on the web. Introduction of Data Generating Techniques and Bioinformatics problem posed by them- Restriction Digestion, Chromatograms, Blots, PCR, Microarrays, Mass Spectrometry.

UNIT-III

Sequence and Phylogeny analysis, Detecting Open Reading Frames, Outline of sequence Assembly, Mutation/Substitution Matrices, Pairwise Alignments, Introduction to BLAST, using it on the web, Interpreting results, Multiple Sequence Alignment, Phylogenetic Analysis. Introduction to BLAST, using it on the web, Interpreting results, Multiple Sequence Alignment, Phylogenetic Analysis.

UNIT-IV

Biological databases: Types of databases, Sequence databases, Nucleic acid sequence databases - Primary (GenBank, EMBL, DDBJ), Secondary (UniGene, SGD, EMI Genomes, Genome Biology), Protein sequence database – Primary (PIR, SWISS-PROT), Secondary (PROSITE, Pfam), Structural databases (PDB, SCOP, CATH), Bibliographic databases and Organism specific databases.

UNIT-V

Searching Databases: SRS, Entrez, Sequence Similarity Searches-BLAST, FASTA, Data Submission. Genome Annotation: Pattern and repeat finding, Gene identification tools. Gene prediction: Gene prediction in prokaryote and eukaryotes. Extrinsic approaches and Ab initio approaches. Predicting the protein secondary structure (Domain, blocks, motifs), Predicting protein tertiary structure (Homology, Ab-initio, threading and fold recognition) and visualization of predicted structure.

References

1. Ghosh, Z. & Bibekanand M. (2008). *Bioinformatics: Principles and Applications*. Oxford University Press.
2. Pevsner, J. (2009). *Bioinformatics and Functional Genomics* (2nd ed.). Wiley-Blackwell.
3. Campbell ,A. M., & Heyer, L.J. (2006). *Discovering Genomics, Proteomics and Bioinformatics* (2nd ed.). Benjamin Cummings.

KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University Established Under Section 3 of UGC Act 1956)

Coimbatore – 641 021.

LECTURE PLAN

DEPARTMENT OF BIOTECHNOLOGY

STAFF NAME: Dr. BARATHKUMAR, S.
SUBJECT NAME: BIOINFORMATICS SUB.CODE:16BTU502A
SEMESTER: V CLASS: III B.Sc.

Duration Hours	Topics to be covered	Support materials
Unit I		
1 hour	Introduction to bioinformatics	T1- pg. 3
	History & Milestones	
1 hour	Notation of homology	T1- pg. 4,5
1 hour	Sequence information sources: a)EMBL b)GENBANK }Understanding –structure and usage	R1-W14
1hour	c)Entre z d)Unigene }Understanding structure and usage	W14
1 hour	**Genome projects – steps**	T1-26
1 hour	Human Genome Projects	W5,6
1 hours	Class Test for Unit I	
7 hours	Total no. of hours planned for Unit I	
References T1 - Jin Xiong (2006) Essential Bioinformatics, Cambridge University Press T2 -Attwood TK and Parry-smith DJ (2006).Introduction R1 -Bioinformatics, pearson Education Ltd. Nucleic acid research ,2008,Database issue.D25-30 W5 – http://en.wikipedia.org/wiki/Human_Genome_Project W14 – http://www.ncbi.nlm.nih.gov/books/NB/21105/pdf-genbank		
Duration hours	**Topics to be covered**	**Support materials**
Unit II		
1 hour	Concept of Biomolecules – Sequence, Stucture & functions	T1- pg. 173
	Amino acids	
1 hour	Proteins	T1- pg. 174-176

1 hour	DNA	
1 hour	RNA	T1- pg. 231-234
1 hour	Protein formation sources –structure and	T4,W15-17
	PDB	
	SWISSPROT	
	TREMB1	
1 hour	Data generating techniques by	
	Restriction digestion	
	Chromatograms	
	Blots PCR Microarrays Mass spectroscopy	T3-184 T3-288,247 W4 W5
1 hour	Unit I ,II possible questions - discussion &Revision	
1 hour	Class test for Unit II	
7 hours	Total no. of hours planned for Unit II	

References

T1 - Jin Xiong (2006) Essential Bioinformatics, Cambridge University Press

T3 -T.A.Brown (2001).Gene cloning –An introduction ,3rd

W4 –http//www.microarray /books/N&B21105/pdf

W5 –http//enwikipedia.org/wiki/mass spectrometry

Duration Hours	**Topics to be covered**	**Support materials**
	Unit III	
1 hour	Sequence &phylogeny analysis	T1- pg. 31-32
	Detection of open reading frames Outline of sequence assembly	
1 hour	Mutation /substitution matrices	T1- pg. 31-32
1 hour	Pairwise alignments BLAST –usage &Result interpretation	T1-34-40
1hour	Multiple sequence alignment	T1-63
1 hour	Phylogenetic analysis	T1-163-168
1 hour	Revision of unit III	
1 hours	Class Test for Unit III	
7 hours	Total no. of hours planned for Unit III	

Duration hours	**Topics to be covered**	**Support materials**
	Unit IV	
1 hour	Introduction of biological database-types	T1-10-18;18
	Nucleic acid sequence database	
1 hour	a)primary-genbank,EMBL,DDBJ	T1- 21
1 hour	**b**)secondary –unigene ,SGD,EM1	R4-W19
1 hour	Protein Sequence database	T2-36-43

	a)Primay-PIB,SWISSPROT	W15
	b)Secondary-Prosite,Ptron	T2-36-43
1 hour	Structural databases	}W17
	PDB	
	SCOP ; CATH	
	Bibliographic database Organism specific database	W18 W19,20
1 hour	Unit III & IV possible questions - discussion &Revision	
1 hour	Class test for Unit III & IV	
7 hours	Total no. of hours planned for Unit IV	

References

R2 -Intro .to computer database(2007),Thomson course
W15-http//www.ebi.ac.uk/fan/database/protein7.html.
W17-http//www.science.co.:1/biomedical/structure-database-ap
W19-http//bioinformatics.Igc.gulbencian.pt/resources/databases/organisms specific databases
W20-http//www.ebi.ac.uk/2com/databases/taxonomic.html

Duration hours	**Topics to be covered**	**Support materials**
	Unit V	
1 hour	Database searching: SRS Entrez	T1-18
1 hour	Sequence similarity searches – BLAST - FASTA Data submission	T1- 52-60
1 hour	Genome annotation – pattern and repeat finding gene identification tools	T1 - 99-103
1 hour	Gene prediction in prokaryotes Gene prediction in eukaryotes	T1- 103-110
1 hour	Extrinsic approaches Abinitin approaches	T1 – 113-122
1 hour	Protein secondary structure	
	Protein tertiary structure	
3 hours	Visualization of proteins – 3D structure	
	Discussion of previous year ESE QP	
	Discussion of previous year ESE QP	

Unit I

SYLLABUS

History and milestone of Bioinformatics. The notion of Homology. Sequence Information Sources, EMBL, GENBANK, Entrez, Unigene, Understanding the structure of each source and using it on the web. Genome sequencing projects – Steps, Human Genome Project and other genome projects.

Introduction to concepts of Bioinformatics

Bioinformatics is an interdisciplinary research area at the interface between computer science and biological science.

A variety of definitions exist in the literature and on the world wide web;

According to **Luscombe et al**.

- Bioinformatics is a union of biology and informatics:
- *bioinformatics* involves the technology that uses computers for storage, retrieval, manipulation, and distribution of information related to biological macromolecules such as DNA, RNA, and proteins.

Bioinformatics is the application of statistics and computer science to the field of molecular biology.

Bioinformatics differs from a related field, *computational biology*.

- **Bioinformatics** is limited to sequence, structural, and functional analysis of genes and genomes and their corresponding products and is often considered *computational molecular biology*.
- But, **computational biology** includes all biological areas that involve computation.

For example, mathematical modeling of ecosystems, population dynamics, application of the game theory in behavioral studies, and phylogenetic construction using fossil records all employ computational tools, but do not necessarily involve biological macromolecules.

Bioinformatics now entails the creation and advancement of databases, algorithms, computational and statistical techniques and theory to solve formal and practical problems arising from the management and analysis of biological data.

The term ***bioinformatics*** was coined by **Paulien Hogeweg and Ben Hesper** in **1978** for the study of informatic processes in biotic systems.

Common activities in bioinformatics include

- mapping and analyzing DNA and protein sequences,
- aligning different DNA and protein sequences to compare them and
- creating and viewing 3-D models of protein structures.

Primary goal or objective of bioinformatics is

- ✓ To better understand a living cell and how it functions at the molecular level.
- ✓ To increase the understanding of biological processes.
- ✓ To analysis and interpretation of various types of data, including nucleotide and amino acid sequences, protein domains, and protein structures.
- ✓ By analyzing raw molecular sequence and structural data, bioinformatics research can generate new insights and provide a "global" perspective of the cell.
- ✓ To understand functions of a cell by analyzing sequence data because the flow of genetic information is dictated by the "central dogma" of biology in which DNA is transcribed to RNA, which is translated to proteins.
- ✓ Cellular functions are mainly performed by proteins whose capabilities are ultimately determined by their sequences. Therefore, solving functional problems using sequence and structural approaches are important.
- ✓ To focus on developing and applying computationally intensive techniques (e.g., pattern recognition, data mining, machine learning algorithms, and visualization).

Scope of bioinformatics:

Bioinformatics consists of **two subfields** and are complementary to each other

1. the development of computational tools and databases
2. application of these tools and databases in generating biological knowledge to better understand living systems.

1. The **tool development** includes

- writing software for sequence, structural, and functional analysis,
- as well as the construction and curating of biological databases.

2. **Application of these tools** in three areas of genomic and molecular biological research:

- molecular sequence analysis,
- molecular structural analysis, and
- molecular functional analysis.

The analyses of biological data often generate new problems and challenges that in turn spur the development of new and better computational tools.

Sequence analysis includes

- sequence alignment,
- sequence database searching,
- motif and pattern discovery,
- gene and promoter finding,
- reconstruction of evolutionary relationships,
- genome assembly and comparison.

Structural analyses includes

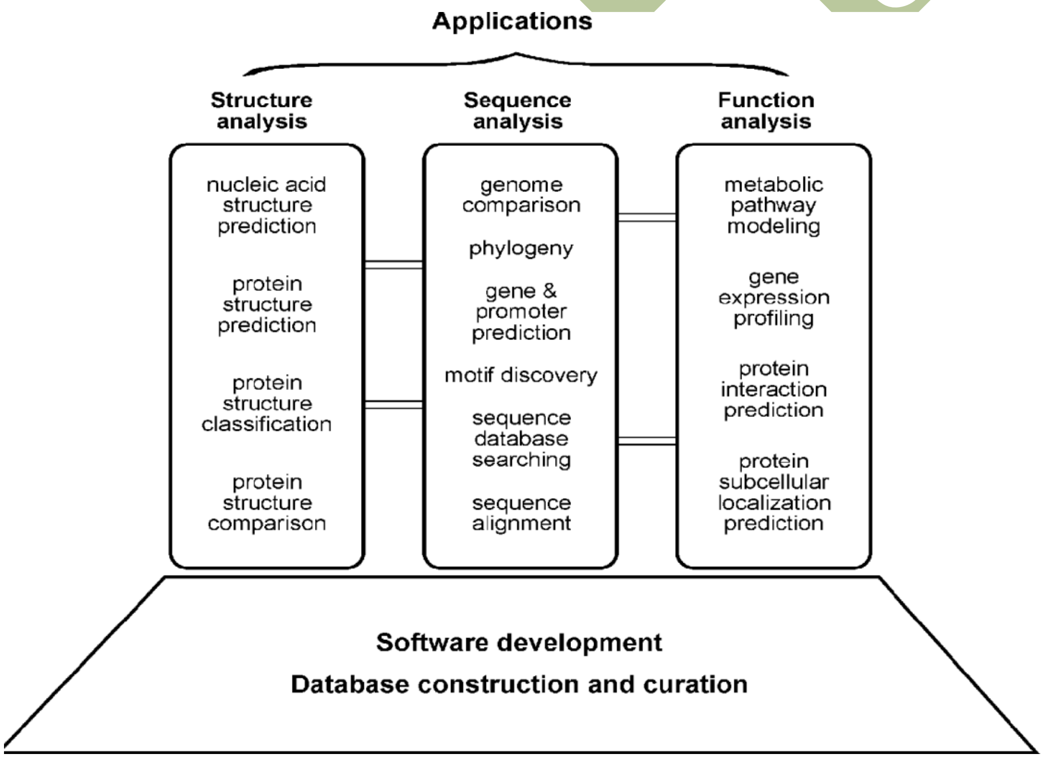
- protein and nucleic acid structure analysis,
- comparison, classification, and prediction.

Functional analyses includes

- gene expression profiling,
- protein– protein interaction prediction,
- protein subcellular localization prediction,
- metabolic pathway reconstruction, and simulation.

These three aspects of bioinformatics analysis are not isolated but often interact to produce integrated results.

Overview of various subfields of bioinformatics


Applications
Structure analysis
Sequence analysis
Function analysis
nucleic acid structure prediction
protein structure prediction
protein structure classification
protein structure comparison
genome comparison
phylogeny
gene & promoter prediction
motif discovery
sequence database searching
sequence alignment
metabolic pathway modeling
gene expression profiling
protein interaction prediction
protein subcellular localization prediction
Software development
Database construction and curation

Application of Bioinformatics in various Fields

Bioinformatics has not only become essential for basic genomic and molecular biology research, but is having a major impact on many areas of biotechnology and biomedical sciences.

Bioinformatics is being used in following fields:

- Molecular medicine
- Personalised medicine
- Preventative medicine
- Gene therapy
- Drug development
- Microbial genome applications
- Waste cleanup
- Climate change Studies
- Alternative energy sources
- Biotechnology
- Antibiotic resistance
- Forensic analysis of microbes
- Bio-weapon creation
- Evolutionary studies
- Crop improvement
- Insect resistance
- Improve nutritional quality
- Development of Drought resistance varieties
- Vetinary Science
- Forensic DNA analysis
- Knowledge- based drug design

Major research areas of bioinformatics includes:

- sequence alignment, gene finding, genome assembly,

- protein structure alignment, protein structure prediction,
- prediction of gene expression and protein-protein interactions,
- genome-wide association studies and the modeling of evolution.
- drug design, drug discovery.

1. Sequence analysis

Sequence alignment and Sequence database

- ✓ Since the Phage Φ-X174 was sequenced in 1977, the DNA sequences of thousands of organisms have been decoded and stored in databases.
- ✓ This sequence information is analyzed to determine genes that encode polypeptides (proteins), RNA genes, regulatory sequences, structural motifs, and repetitive sequences.
- ✓ A comparison of genes within a species or between different species can show similarities between protein functions, or relations between species (the use of molecular systematics to construct phylogenetic trees).
- ✓ Today, computer programs such as BLAST are used daily to search sequences from more than 260 000 organisms, containing over 190 billion nucleotides.
- ✓ Shotgun sequencing is the method of choice for virtually all genomes sequenced today, and genome assembly algorithms are a critical area of bioinformatics research.
- ✓ Another aspect of bioinformatics in sequence analysis is annotation, which involves computational gene finding to search for protein-coding genes, RNA genes, and other functional sequences within a genome.
- ✓ Bioinformatics helps to bridge the gap between genome and proteome projects — for example, in the use of DNA sequences for protein identification.

2. Genome annotation

- ✓ In the context of genomics, **annotation** is the process of marking the genes and other biological features in a DNA sequence.

- ✓ The first genome annotation software system was designed in **1995** by **Dr. Owen White**, for the first genome of a free-living organism, the bacterium *Haemophilus influenzae*. Dr. White built a software system to find the genes (places in the DNA sequence that encode a protein), the transfer RNA, and other features, and to make initial assignments of function to those genes.
- ✓ Most current genome annotation systems work similarly, but the programs available for analysis of genomic DNA are constantly changing and improving.

3. Computational evolutionary biology

Evolutionary biology is the study of the origin and descent of species, as well as their change over time.

Bioinformatics has enabled to:

- trace the evolution of a large number of organisms by measuring changes in their DNA, rather than through physical taxonomy or physiological observations alone,
- compare entire genomes, which permits the study of more complex evolutionary events, such as gene duplication, horizontal gene transfer, and the prediction of factors important in bacterial speciation,
- build complex computational models of populations to predict the outcome of the system over time
- track and share information on an increasingly large number of species and organisms

4. Analysis of gene expression

The expression of many genes can be determined by measuring mRNA levels with multiple techniques including

- microarrays,
- expressed cDNA sequence tag (EST) sequencing,

- serial analysis of gene expression (SAGE) tag sequencing,
- massively parallel signature sequencing (MPSS), or
- various applications of multiplexed in-situ hybridization.

Bioinformatics have been applied to develop statistical tools for separating signal from noise in high-throughput gene expression studies.

5. Analysis of regulation

- ✓ Regulation is the complex orchestration of events starting with an extracellular signal such as a hormone and leading to an increase or decrease in the activity of one or more proteins.
- ✓ Bioinformatics techniques have been applied to explore various steps in this process. For example, promoter analysis involves the identification and study of sequence motifs in the DNA surrounding the coding region of a gene.

6. Analysis of protein expression

- ✓ Protein microarrays and high throughput (HT) mass spectrometry (MS) can provide a snapshot of the proteins present in a biological sample.
- ✓ Bioinformatics is very much involved in protein microarray and HT MS data.

7. Analysis of mutations in cancer

- ✓ In cancer, the genomes of affected cells are rearranged in complex or even unpredictable ways.
- ✓ Massive sequencing efforts are used to identify previously unknown point mutations in a variety of genes in cancer.
- ✓ Bioinformaticians continue to produce specialized automated systems to manage the sheer volume of sequence data produced, and they create new algorithms and software to compare the sequencing results to the growing collection of human genome sequences and germline polymorphisms.

- ✓ New physical detection technologies are employed, such as oligonucleotide microarrays to identify chromosomal gains and losses (called comparative genomic hybridization), and single-nucleotide polymorphism arrays to detect known *point mutations*.
- ✓ These detection methods simultaneously measure several hundred thousand sites throughout the genome, and when used in high-throughput to measure thousands of samples, generate terabytes of data per experiment.

8. Comparative genomics

- ✓ The core of comparative genome analysis is the establishment of the correspondence between genes (orthology analysis) or other genomic features in different organisms.
- ✓ A multitude of evolutionary events acting at various organizational levels shape genome evolution.

At the lowest level, point mutations affect individual nucleotides.

At a higher level, large chromosomal segments undergo duplication, lateral transfer, inversion, transposition, deletion and insertion. Ultimately, whole genomes are involved in processes of hybridization, polyploidization and endosymbiosis, often leading to rapid speciation.

- ✓ Complexity of genome evolution helps to developers of mathematical models and algorithms, based on parsimony models to Markov Chain Monte Carlo algorithms for Bayesian analysis of problems based on probabilistic models.

9. Modeling biological systems

- ✓ Systems biology involves the use of computer simulations of cellular subsystems to both analyze and visualize the complex connections of these cellular processes (such as the networks of metabolites and enzymes which comprise metabolism, signal transduction pathways and gene regulatory networks).
- ✓ Artificial life or virtual evolution attempts to understand evolutionary processes via the

computer simulation of simple (artificial) life forms.

10. High-throughput image analysis

- ✓ Computational technologies are used to accelerate or fully automate the processing, quantification and analysis of large amounts of high-information-content biomedical imagery.
- ✓ Modern image analysis systems helps an observer's ability to make measurements from a large or complex set of images, by improving accuracy, objectivity, or speed.
- ✓ A fully developed analysis system may completely replace the observer.

Biomedical imaging is becoming more important for both diagnostics and research. Some examples are:

- high-throughput and high-fidelity quantification and sub-cellular localization (high-content screening, cytohistopathology, Bioimage informatics)
- morphometrics
- clinical image analysis and visualization
- determining the real-time air-flow patterns in breathing lungs of living animals
- quantifying occlusion size in real-time imagery from the development of and recovery during arterial injury
- making behavioral observations from extended video recordings of laboratory animals
- infrared measurements for metabolic activity determination
- inferring clone overlaps in DNA mapping, e.g. the Sulston score

11. Structural Bioinformatic Approaches

- ✓ Protein structure prediction is another important application of bioinformatics.
- ✓ The amino acid sequence of a protein, the so-called primary structure, can be easily determined from the sequence on the gene that codes for it.

- ✓ Knowledge of this structure is important in understanding the function of the protein.

Structural information is usually classified as one of

secondary, *tertiary* and *quaternary* structure.

In the genomic branch of bioinformatics,

- ✓ homology is used to predict the function of a gene:

if the sequence of gene *A*, whose function is known, is homologous to the sequence of gene *B*, whose function is unknown, one could infer that B may share A's function.

In the structural branch of bioinformatics,

- ✓ homology is used to determine which parts of a protein are important in structure formation and interaction with other proteins.
- ✓ In a technique called homology modeling, this information is used to predict the structure of a protein once the structure of a homologous protein is known.
- ✓ Other techniques for predicting protein structure include protein threading and *de novo* (from scratch) physics-based modeling.

Molecular Interaction

- ✓ Efficient software is available today for studying interactions among proteins, ligands and peptides.
- ✓ Types of interactions most often encountered in the field include - Protein-ligand (including drug), protein-protein and protein-peptide.
- ✓ Molecular dynamic simulation of movement of atoms about rotatable bonds is the fundamental principle behind computational algorithms, termed **docking algorithms** for studying molecular interactions.

Docking algorithms -Protein-protein docking

In the last two decades,

tens of thousands of protein three-dimensional structures have been determined by X-ray crystallography and Protein nuclear magnetic resonance spectroscopy (protein NMR).

A variety of methods have been developed to tackle the Protein-protein docking problem.

Software and tools

Software tools for bioinformatics range

from simple command-line tools, to more complex graphical programs and standalone web-services available from various bioinformatics companies or public institutions.

Web services in bioinformatics

SOAP and REST-based interfaces have been developed for a wide variety of bioinformatics applications allowing an application running on one computer in one part of the world to use algorithms, data and computing resources on servers in other parts of the world.

The main advantages derive from the fact that end users do not have to deal with software and database maintenance overheads.

Basic bioinformatics services are classified by the **EBI** into three categories:

- **SSS** (Sequence Search Services),
- **MSA** (Multiple Sequence Alignment) and
- **BSA** (Biological Sequence Analysis).

The availability of these service-oriented bioinformatics resources demonstrate the applicability of web based bioinformatics solutions, and range from a collection of standalone tools with a common data format under a single, standalone or web-based interface, to integrative, distributed and extensible bioinformatics workflow management systems.

History of Bioinformatics

The Modern bioinformatics is can be classified into **two** broad categories,

Biological Science and computational Science.

However, it is the 1990s when the INTERNET arrived when the full fledged bioinformatics field was born.

Here are some of the major events in bioinformatics over the last several decades.

The events listed in the list occurred long before the term, "bioinformatics", was coined.

BioInformatics Events	
1843	Richard Owen elaborated the distinction of **homology** and **analogy.**
1961	Sidney Brenner, François Jacob, Matthew Meselson, identify messenger RNA,
1965	Margaret Dayhoff's Atlas of Protein Sequences
1970	Needleman-Wunsch algorithm
1977	DNA sequencing and software to analyze it (Staden)
1981	Smith-Waterman algorithm developed
1981	The concept of a sequence motif (Doolittle)
1982	GenBank Release 3 made public
1982	Phage lambda genome sequenced
1983	Sequence database searching algorithm (Wilbur-Lipman)
1985	FASTP/FASTN: fast sequence similarity searching
1988	National Center for Biotechnology Information (NCBI)

	created at NIH/NLM
1988	EMBnet network for database distribution
1990	BLAST: fast sequence similarity searching
1991	EST: expressed sequence tag sequencing
1993	Sanger Centre, Hinxton, UK
1994	EMBL European Bioinformatics Institute, Hinxton, UK
1995	First bacterial genomes completely sequenced
1996	Yeast genome completely sequenced
1997	PSI-BLAST
1998	Worm (multicellular) genome completely sequenced
1999	Fly genome completely sequenced
2000	Jeong et al. **The large-scale organization of metabolic networks**
2000	The genome for *Pseudomonas aeruginosa* (6.3 Mbp) published
2000	The A. thaliana genome (100 Mb) secquenced
2001	The human genome (3 Giga base pairs) published

Milestones in bioinformatics:

Listed below are some of the major events in bioinformatics over the last several decades. Most of the events in the list occurred long before the term, "bioinformatics", was coined.

1962	Pauling's theory of molecular evolution
1965	Margaret Dayhoff's Atlas of Protein Sequences
1970	Needleman-Wunsch algorithm

1977	DNA sequencing and software to analyze it (Staden)
1981	Smith-Waterman algorithm developed
1981	The concept of a sequence motif (Doolittle)
1982	GenBank Release 3 made public
1982	Phage lambda genome sequenced
1983	Sequence database searching algorithm (Wilbur-Lipman)
1985	FASTP/FASTN: fast sequence similarity Searching
1988	National Center for Biotechnology Information (NCBI) created at NIH/NLM
1988	EMBnet network for database distribution
1990	BLAST: fast sequence similarity searching
1991	EST: expressed sequence tag sequencing
1993	Sanger Centre, Hinxton, UK
1994	EMBL European Bioinformatics Institute, Hinxton, UK
1995	First bacterial genomes completely sequenced
1996	Yeast genome completely sequenced
1997	PSI-BLAST, ***Escherichia coli* genome completely sequenced**

1998	Worm (multicellular) genome completely sequenced
1999	Fly genome completely sequenced
2000	First plant genome sequenced – *Arabidopsis*
2001	Draft of human genome sequence
2002	Draft of mouse genome sequence, Japanese puffer fish genome, rice genome sequence
2003	Sequence of human chromosome 14
2005	Rice genome sequence

Sequence information sources

- The ultimate goal of genome analysis is understanding the biology of each particular organism in both functional and evolutionary terms, which requires combining disparate data from a variety of sources. Reliable information resources, compiling data on sequenced genomes and linking it to the wealth of associated functional data, are indispensable for comparative genomics.
- The amount of genome-related information stored in public databases and freely available to anyone with an Internet access is enormous.

1. Nucleotide sequence databases

- Biological sequence has been deposited in one of the three major international nucleotide sequence databases: GenBank at the NCBI (Bethesda, Maryland, USA); the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database at the European Bioinformatics Institute (EBI) in Hinxton, near Cambridge, UK; and the DNA Database of Japan (DDBJ) at the National Institute of Genetics in Mishima, Japan.
- These databases form an International Nucleotide Sequence Database Collaboration and exchange updates on a daily basis, so that the DNA sequence information kept in each database is essentially the same and is arranged using common principles (see http://www.ncbi.nlm.nih.gov/projects/collab). Although data representation in

GenBank, EMBL, and DDBJ might differ slightly, each nucleotide sequence has the same accession number in all three databases.

- The information stored in these databases is available to the public by anonymous ftp and through the World Wide Web. This means that one can connect to the web site of any of the three databases, GenBank (http://www.ncbi.nlm.nih.gov/Entrez), EMBL (http://www.ebi.ac.uk), or DDBJ (http://www.ddbj.nig.ac.jp), and get the same nucleotide sequence using the same accession number. Thus, a sequence with a given GenBank accession number could have been originally submitted to EMBL or DDBJ, and vice versa.

- In everyday practice, people often refer to the public nucleotide database simply as "GenBank" when they actually mean the combination of all three public databases.

- Although the nucleotide sequence data in GenBank, EMBL, and DDBJ are the same, these three databases differ in the additional services that they offer. NCBI, for example, maintains several other databases in addition to GenBank, such as the Taxonomy database and PubMed .

- Accordingly, each nucleotide entry at the NCBI web site is hyperlinked to the corresponding journal article in PubMed (if available) and to the taxonomic entry for the source organism.

2. Protein sequence databases

- For most of the 20th century, biologists usually had at least some idea of what they were studying, and new sequences were coming from well-defined projects that investigated a particular protein or a group of proteins.

- As a result, the first protein sequence database, Atlas of Protein Sequence and Structure, created by Margaret Dayhoff in the early 1960's, contained very few uncharacterized proteins and was used mainly to document and investigate sequence diversity between homologous proteins (e.g. globins or cytochromes) from diverse organisms.

- This trend continued for a few years, even after the introduction of rapid DNA sequencing methods. However, with the rapid increase in gene sequencing rate in the early 1980's, more and more new protein sequences were derived from translation of anonymous pieces of DNA (or mRNA), first as a collateral benefit of sequencing the gene of interest and later through genome projects.

- This quantitative growth of sequence information was accompanied by a qualitative change that brought about several major problems. Although these problems could be considered just the issues of database quality control, they touch upon fundamental scientific questions.

Entrez

- Entrez is the text-based search and retrieval system used at the National Center for Biotechnology Information (NCBI) for all of the major databases, including PubMed, Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, Taxonomy, and many others. Entrez is at once an indexing and retrieval system, a collection of data from many sources, and an organizing principle for biomedical information.
- These general concepts are the focus of this chapter. Other chapters cover the details of a specific Entrez database (e.g., PubMed) or a specific source of data (e.g., GenBank).

Entrez Nodes Represent Data

- An Entrez "node" is a collection of data that is grouped together and indexed together. It is usually referred to as an Entrez database. In the first version of Entrez, there were three nodes: published articles, nucleotide sequences, and protein sequences. Each node represents specific data objects of the same type, e.g., protein sequences, which are each given a unique ID (UID) within that logical Entrez Protein's node.

- Records in a node may come from a single source (e.g., all published articles are from PubMed) or many sources (e.g., proteins are from translated GenBank sequences, SWISS-PROT, or PIR).

- Note that the UID identifies a single, well-defined object (i.e., a particular protein sequence or PubMed citation). There may be other information about objects in nodes, such as protein names or Enzyme Commission (EC) numbers, that may be used as index terms to find the record, but these pieces of information are not the central organizing principle of the node.

- Each data object represents a stable, objective observation of data as much as possible, rather than interpretations of the data, which are subject to change or confusion over time or across disciplines.

- For example, barring experimental error, a particular mRNA sequence report is not likely to change over the years; however, the given name, position on the chromosome, or function of the protein product may well change as our knowledge develops.

EMBL

- The European Bioinformatics Institute (EBI) is an Outstation of the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany. The EBI is located in the grounds of the Wellcome Trust Genome Campus near Cambridge, UK, next to the Sanger Centre and the UK Human Genome Mapping Project Resource Centre.
- The main missions of the Service Programme of the EBI centre on building, maintaining and providing biological databases and information services to support data deposition and exploitation. In this respect a number of databases are operated, namely the EMBL Nucleotide Sequence Database (EMBL-Bank), the Protein Databases (SWISS-PROT and TrEMBL), the Macromolecular Structure Database (MSD) and ArrayExpress for gene expression data plus several other databases many of which are produced in collaboration with external groups.
- The EMBL Nucleotide Sequence Database (http://www.ebi.ac.uk/embl/) is the European member of the tri-partide International Nucleotide Sequence Database Collaboration DDBJ/EMBL/GenBank. Main data sources are large-scale genome sequencing centres, individual scientists and the European Patent Office (EPO). Direct submissions to EMBL-Bank are complemented by daily data exchange with collaborating databases DDBJ (Japan) and GenBank (USA).
- The EMBL database is growing rapidly as a result of major genome sequencing efforts. Within a 12 month period the database size has increased from about 6.7 million entries comprising 8255 million nucleotides (Release 63, June 2000) to over 12 million entries and 12 820 million nucleotides (Release 67, June 2001). During the same period the number of organisms represented in the database has risen by >30% to over 75 000 species.

GenBank

- The GenBank sequence database is an open access, annotated collection of all publicly available nucleotide sequences and their protein translations.
- This database is produced at National Center for Biotechnology Information (NCBI) as part of the International Nucleotide Sequence Database Collaboration, or INSDC.
- GenBank and its collaborators receive sequences produced in laboratories throughout the world from more than 100,000 distinct organisms.
- GenBank continues to grow at an exponential rate, doubling every 10 months.

- Release 155, produced in August 2006, contained over 65 billion nucleotide bases in more than 61 million sequences.

- GenBank is built by direct submissions from individual laboratories, as well as from bulk submissions from large-scale sequencing centers.

Unigene

- UniGene is an experimental system for automatically partitioning GenBank sequences into a non-redundant set of gene-oriented clusters. Each UniGene cluster contains sequences that represent a unique gene, as well as related information such as the tissue types in which the gene has been expressed and map location.
- In addition to sequences of well-characterized genes, hundreds of thousands novel expressed sequence tag (EST) sequences have been included. Consequently, the collection may be of use to the community as a resource for gene discovery. UniGene has also been used by experimentalists to select reagents for gene mapping projects and large-scale expression analysis.
- However, it should be noted that the procedures for automated sequence clustering are still under development and the results may change from time to time as improvements are made. Feedback from users has been especially useful in identifying problems and we encourage you to report any problems you encounter.
- It should also be noted that no attempt has been made to produce contigs or consensus sequences. There are several reasons why the sequences of a set may not actually form a single contig. For example, all of the splicing variants for a gene are put into the same set. Moreover, EST-containing sets often contain 5' and 3' reads from the same cDNA clone, but these sequences do not always overlap.
- Currently, sequences from the animals human, rat, mouse, cow, zebrafish and clawed frog have been processed. Plant organisms are wheat, rice, barley, maize and cress. These species were chosen because they have the greatest amounts of EST data available and represent a variety of species.

Genome sequencing projects

- ➢ Genome projects are scientific endeavours that ultimately aim to determine the complete genome sequence of an organism (animal, plant, fungus, bacterium, archaean, protist or virus).
- ➢ The genome sequence for any organism requires the DNA sequences for each of the chromosomes in an organism to be determined.

- For bacteria, which usually have just one chromosome, a genome project will aim to map the sequence of that chromosome.
- Humans, with 22 pairs of autosomes and 2 sex chromosomes, will require 46 separate chromosome sequences in order to represent the completed genome.
- The Human Genome Project was a landmark genome project that is already having a major impact on research across the life sciences, with potential for spurring numerous medical and commercial developments.

Goal of genome projects:

sequencing a genome is to obtain information about the complete set of genes in that particular genome sequence.

Steps involved in genome projects: 2 steps

- Genome assembly
- Genome annotation

Genome assembly

- Genome assembly refers to the process of taking a large number of short DNA sequences, all of which were generated by a shotgun sequencing project, and putting them back together to create a representation of the original chromosomes from which the DNA originated.
- In a shotgun sequencing project, all the DNA from a source (usually a single organism, anything from a bacterium to a mammal) is first fractured into millions of small pieces.
- These pieces are then "read" by automated sequencing machines, which can read up to 900 nucleotides or bases at a time.
- A genome assembly algorithm works by taking all the pieces and aligning them to one another, and detecting all places where two of the short sequences, or *reads*, overlap.
- These overlapping reads can be merged together, and the process continues.

- Genome assembly is a very difficult computational problem, made more difficult because many genomes contain large numbers of identical sequences, known as *repeats*.
- These repeats can be thousands of nucleotides long, and some occur in thousands of different locations, especially in the large genomes of plants and animals.
- The resulting (draft) genome sequence is produced by combining the information sequenced contigs and then employing linking information to create scaffolds. Scaffolds are positioned along the physical map of the chromosomes creating a "golden path".

Assembly software

Large-scale DNA sequencing centers developed their own software for assembling the sequences that they produced.

An example of such assembler - *Short Oligonucleotide Analysis Package* developed by BGI for de novo assembly of human-sized genomes, alignment, SNP detection, resequencing, indel finding, and structural variation analysis.

Genome annotation

Genome annotation is the process of attaching biological information to sequences. It consists of two main steps:

1. identifying elements on the genome, a process called gene prediction, and
2. attaching biological information to these elements.

- Automatic annotation tools try to perform all this by computer analysis, as opposed to manual annotation which involves human expertise. Ideally, these approaches co-exist and complement each other in the same annotation pipeline.
- The basic level of annotation is using BLAST for finding similarities, and then annotating genomes based on that.
- However, nowadays more and more additional information is added to the annotation platform.

- Some databases use genome context information, similarity scores, experimental data, and integrations of other resources to provide genome annotations through their Subsystems approach.
- Other databases (e.g Ensembl) rely on both curated data sources as well as a range of different software tools in their automated genome annotation pipeline.

Structural annotation consists of the identification of genomic elements.

- ORFs and their localisation
- gene structure
- coding regions
- location of regulatory motifs

Functional annotation consists of attaching biological information to genomic elements.

- biochemical function
- biological function
- involved regulation and interactions
- expression

Complete genome projects - When is a genome project finished?

- When sequencing a genome, there are usually regions that are difficult to sequence (often regions with highly repetitive DNA).
- A complete genome project should include the sequences of mitochondria and (for plants) chloroplasts as these organelles have their own genomes.

Example of on-going projects relevant to genome annotation:

- ENCyclopedia Of DNA Elements (ENCODE)
- Entrez Gene
- Ensembl

- GENCODE
- Gene Ontology Consortium
- GeneRIF
- RefSeq
- Uniprot
- Vertebrate and Genome Annotation Project (Vega)

Example genome projects

List of sequenced eukaryotic genomes –

http://en.wikipedia.org/wiki/List_of_sequenced_eukaryotic_genomes),

List of sequenced archaeal genomes –

http://en.wikipedia.org/wiki/List_of_sequenced_archaeal_genomes

List of sequenced prokaryotic genomes –

http://en.wikipedia.org/wiki/List_of_sequenced_prokaryotic_genomes

Many organisms have genome projects that have either been completed or will be completed shortly, including:

- Humans, *Homo sapiens*; see Human genome project
- Palaeo-Eskimo, an ancient-human
- Neanderthal, "*Homo neanderthalensis*" (partial);
- Common Chimpanzee *Pan troglodytes*; Chimpanzee Genome Project
- Domestic Cow
- Bovine Genome
- Honey Bee Genome Sequencing Consortium
- Human microbiome project
- International Grape Genome Program

- International HapMap Project

Human Genome Project (HGP)

Human Genome Project (HGP) is an international scientific research project with a primary goal of determining the sequence of chemical base pairs which make up DNA and to identify and map the approximately 20,000–25,000 genes of the human genome from both a physical and functional standpoint.

The project began in **1990** and was initially headed by **Ari Patrinos**, U.S. Department of Energy's Office of Science.

Francis Collins directed the National Institutes of Health National Human Genome Research Institute efforts.

A **working draft of the genome** was released in 2000

A **complete draft of the genome** was released in 2003.

The Human Genome Project originally aimed to map the nucleotides contained in a human haploid reference genome (more than three billion).

Several groups have announced efforts to extend this to diploid human genomes including the International HapMap Project, Applied Biosystems, Perlegen, Illumina, JCVI, Personal Genome Project, and Roche-454.

Methods used in human genome project

- The **IHGSC** used pair-end sequencing plus whole-genome shotgun mapping of large (≈100 Kbp) plasmid clones and shotgun sequencing of smaller plasmid sub-clones plus a variety of other mapping data to orient and check the assembly of each human chromosome.

- The **Celera group** used the "whole-genome shotgun" sequencing method, relying on sequence information to orient and locate their fragments within the chromosome.
- However they used the publicly available data from HGP to assist in the assembly and orientation process, raising concerns that the Celera sequence was not independently derived.

Key Findings of Human genome project: - draft (2001) and complete (2004) genome sequences

1. There are approximately 20,500 genes in human beings, the same range as in mice and twice that of roundworms. Understanding how these genes express themselves will provide clues to how diseases are caused.

2. Between 1.1% to 1.4% of the genome's sequence codes for proteins

3. The human genome has significantly more segmental duplications (nearly identical, repeated sections of DNA) than other mammalian genomes. These sections may underlie the creation of new primate-specific genes

4. At the time when the draft sequence was published less than 7% of protein families appeared to be vertebrate specific

Advantages of Human Genome Project:

1. Knowledge of the effects of variation of DNA among individuals can revolutionize the ways to diagnose, treat and even prevent a number of diseases that affects the human beings.
2. It provides clues to the understanding of human biology.

Review Questions

Short Answer Questions (2 Marks)

1. Define bioinformatics
2. List out the objectives of bioinformatics?
3. What are the fields of bioinformatics scope?
4. List out the sub field of sequence analysis?
5. List out the sub field of structural analysis.
6. List out the sub field of functional analysis.
7. Differentiate bioinformatics and computational biology?
8. What are the major research areas of bioinformatics?
9. List out the different field of bioinformatics.
10. Define genome?
11. Define human genome project.
12. Define genome assembly in genome project.
13. Define genome annotation.
14. Define genome project.
15. Describe the method used in HGP?
16. List the findings of HGP?
17. Describe the advantages of HGP?

Essay Answer Questions (6 & 8 Marks)

1. Discuss the objectives and scope of Bioinformatics?
2. Describe the various applications of bioinformatics.
3. Give a detailed account on history and milestones of bioinformatics.
4. Describe Genome project and steps involved in genome project.
5. Discuss the Human genome project.
6. Give a account on other organism genome projects.

Further Readings:

Jin Xiong (2006) Essential Bioinformatics, Cambridge University Press.

Applications of bioinformatics – en.wikipedia.org/wiki/**Bioinformatics**

- ✓ www.roseindia.net/**bioinformatics**/**applications**.shtml

History – www.roseindia.net/**bioinformatics**/**history_of_bioinformatics**.shtml

Milestones –

- ✓ http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/milestones.html
- ✓ http://zion.ugent.be/BioZendium/index.php/Milestones_in_bioinformatics
- ✓ http://arxiv.org/ftp/arxiv/papers/0911/0911.4230.pdf

Genome projects –

- ✓ Attwood TK and Parry- Smith DJ (2006) Introduction to Bioinformatics. Pearson Education Ltd.
- ✓ http://en.wikipedia.org/wiki/Genome_project

Unit II

SYLLABUS

Basic concepts of biomolecules – Protein and amino acid, DNA and RNA - Sequence, Structure and function. Protein Information Sources, PDB, SWISSPROT, TREMBL, Understanding the structure of each source and using it on the web. Introduction of Data Generating Techniques and Bioinformatics problem posed by them- Restriction Digestion, Chromatograms, Blots, PCR, Microarrays, Mass Spectrometry.

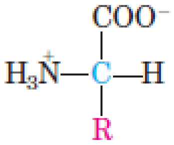
Introduction to Biomolecules

Amino acids

- Proteins are polymers of amino acids, with each amino acid residue joined to its neighbor by a specific type of covalent bond. (The term "residue" reflects the loss of the elements of water when one amino acid is joined to another.)
- Proteins can be broken down (hydrolyzed) to their constituent amino acids by a variety of methods.
- Twenty different amino acids are commonly found in proteins. The first to be discovered was asparagine, in 1806. The last of the 20 to be found, threonine, was not identified until 1938.
- Amino acids have trivial or common names, in some cases derived from the source from which they were first isolated. Asparagine was first found in asparagus, and glutamate in wheat gluten; tyrosine was first isolated from cheese (its name is derived from the Greek *tyros*, "cheese"); and glycine (Greek *glykos*, "sweet") was so named because of its sweet taste.

Amino Acids Share Common Structural Features

- All 20 of the common amino acids are α-amino acids.
- They have a carboxyl group and an amino group bonded to the same carbon atom (the α carbon).

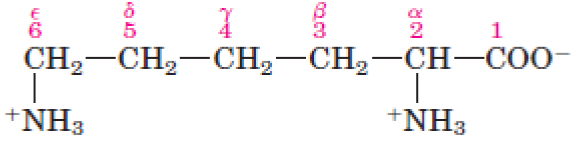


- They differ from each other in their side chains, or **R groups,** which vary in structure, size, and electric charge, and which influence the solubility of the amino acids in water.
- In addition to these 20 amino acids there are many less common ones. Some are residues modified after a protein has been synthesized; others are amino acids present in living organisms but not as constituents of proteins.
- The common amino acids of proteins have been assigned three-letter abbreviations and one-letter symbols (Table), which are used as shorthand to indicate the composition and sequence of amino acids polymerized in proteins.

Identifying the carbons in an amino acid

- The additional carbons in an R group are commonly designated β, γ, δ, ε and so forth, proceeding out from the α carbon.
- Within this latter convention, the carboxyl carbon of an amino acid would be C-1 and the α carbon would be C-2. In some cases, such as amino acids with heterocyclic R groups, the Greek lettering system is ambiguous and the numbering convention is therefore used.

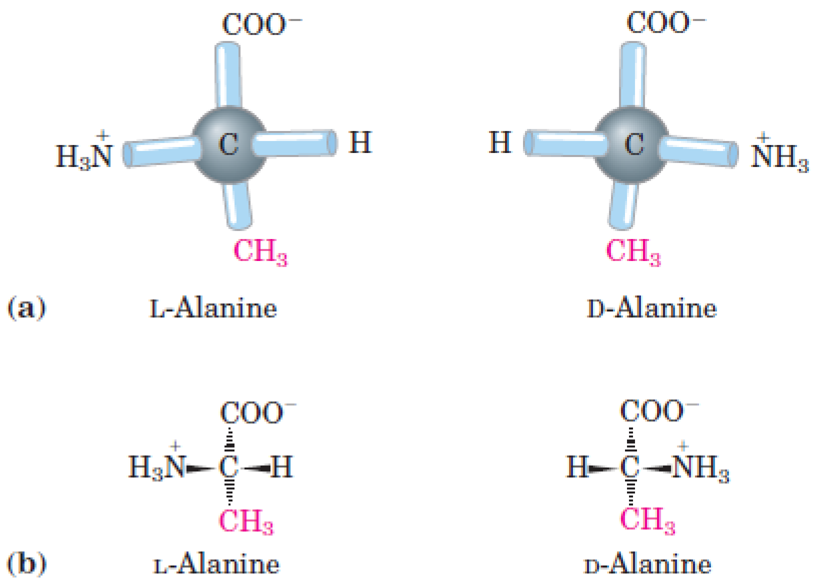
Example



Lysine

- For all the common amino acids except glycine, the α carbon is bonded to four different groups: a carboxyl group, an amino group, an R group, and a hydrogen atom in glycine, the R group is another hydrogen atom. The α-carbon atom is thus a **chiral center**.

- Because of the tetrahedral arrangement of the bonding orbitals around the α-carbon atom, the four different groups can occupy two unique spatial arrangements, and thus amino acids have two possible stereoisomers.
- Since they are nonsuperimposable mirror images of each other , the two forms represent a class of stereoisomers called **enantiomers.** All molecules with a chiral center are also **optically active,** that is, they rotate plane-polarized light.



(a) L-Alanine D-Alanine

(b) L-Alanine D-Alanine

Amino acid	Three letter code	One letter code
Alanine	ala	A
Arginine	arg	R
Asparagines	asn	N
aspartic acid	asp	D
asparagine or aspartic acid	asx	B
Cysteine	cys	C

glutamic acid	glu	E
Glutamine	gln	Q
glutamine or glutamic acid	glx	Z
Glycine	gly	G
Histidine	his	H
Isoleucine	ile	I
Leucine	leu	L
Lysine	lys	K
Methionine	met	M
Phenylalanine	phe	F
Proline	pro	P
Serine	ser	S
Threonine	thr	T
tryptophan	trp	W
tyrosine	tyr	Y
Valine	val	V

Classification of amino acids

Based on the requirement

Amino acids are classified into two types called essential and non-essential amino acids.

Essential amino acids

An essential amino acid or indispensable amino acid is an amino acid that cannot be synthesized *de novo(from scratch)* by the organism being considered, and therefore must be supplied in its diet.

Examples: phenylalanine, valine, threonine, tryptophan, isoleucine, methionine, leucine, lysine, and histidine. Additionally, cysteine (or sulphur-containing amino acids), tyrosine (or aromatic amino acids), and arginine are required by infants and growing children.

Non-essential amino acids

Those of the naturally occurring amino acids, that the human body can synthesize for itself, and so need not be provided by dietary protein.

Examples: alanine, asparagines, aspartic acid, cysteine, glutamic acid, glutamine, glycine, proline, tyrosine and serine.

Classification based on R Group

Nonpolar, Aliphatic R Groups

- The R groups in this class of amino acids are nonpolar and hydrophobic.
- The side chains of alanine, valine, leucine, and isoleucine tend to cluster together within proteins, stabilizing protein structure by means of hydrophobic interactions.
- Glycine has the simplest structure. Although it is formally nonpolar, its very small side chain makes no real contribution to hydrophobic interactions.
- Methionine, one of the two sulfur-containing amino acids, has a nonpolar thioether group in its side chain. Proline has an aliphatic side chain with a distinctive cyclic structure. The
- Secondary amino (imino) group of proline residues is held in a rigid conformation that reduces the structural flexibility of polypeptide regions containing proline.

Aromatic R Groups

- Phenylalanine, tyrosine, and tryptophan, with their aromatic side chains, are relatively nonpolar (hydrophobic). All can participate in hydrophobic interactions.
- The hydroxyl group of tyrosine can form hydrogen bonds, and it is an important functional group in some enzymes. Tyrosine and tryptophan are significantly more polar than phenylalanine, because of the tyrosine hydroxyl group and the nitrogen of the tryptophan

indole ring. Tryptophan and tyrosine, and to a much lesser extent phenylalanine, absorb ultraviolet light.

- This accounts for the characteristic strong absorbance of light by most proteins at a wavelength of 280 nm, a property exploited by researchers in the characterization of proteins.

Polar, Uncharged R Groups

- The R groups of these amino acids are more soluble in water, or more hydrophilic, than those of the nonpolar amino acids, because they contain functional groups that form hydrogen bonds with water.
- This class of amino acids includes serine, threonine, cysteine, asparagine, and glutamine. The polarity of serine and threonine is contributed by their hydroxyl groups; that of cysteine by its sulfhydryl group; and that of asparagine and glutamine by their amide groups.
- Asparagine and glutamine are the amides of two other amino acids also found in proteins, aspartate and glutamate, respectively, to which asparagine and glutamine are easily hydrolyzed by acid or base.
- Cysteine is readily oxidized to form a covalently linked dimeric amino acid called cystine, in which two cysteine molecules or residues are joined by a disulfide bond. The disulfide-linked residues are strongly hydrophobic (nonpolar).
- Disulfide bonds play a special role in the structures of many proteins by forming covalent links between parts of a protein molecule or between two different polypeptide chains.

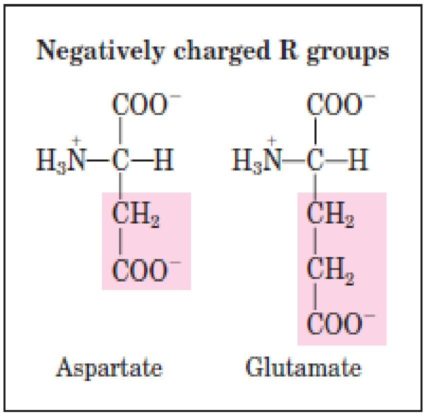
Positively Charged (Basic) R Groups

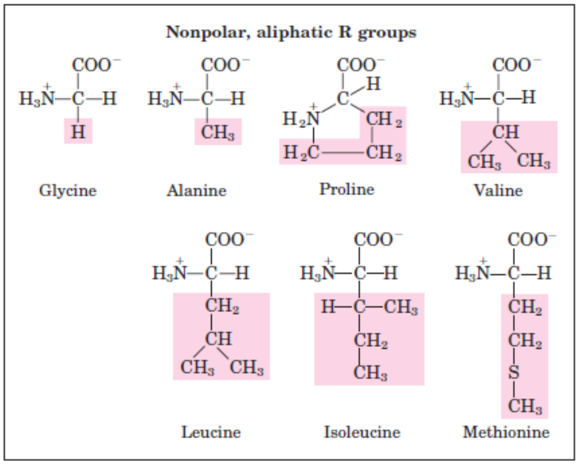
- The most hydrophilic R groups are those that are either positively or negatively charged.
- The amino acids in which the R groups have significant positive charge at pH 7.0 are lysine, which has a second primary amino group at the ε position on its aliphatic chain; arginine, which has a positively charged guanidino group; and histidine, which has an imidazole group. Histidine is the only common amino acid having an ionizable side chain with a p*K*a near neutrality.

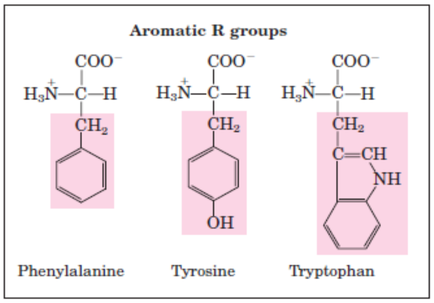
- In many enzyme-catalyzed reactions, a His residue facilitates the reaction by serving as a proton donor/acceptor.

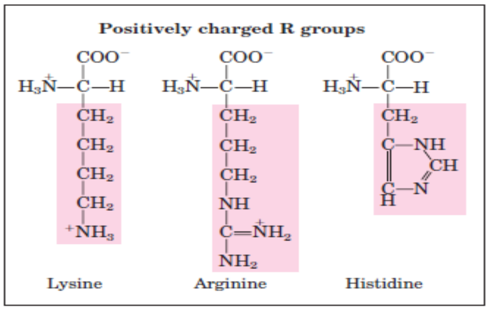
Negatively Charged (Acidic) R Groups

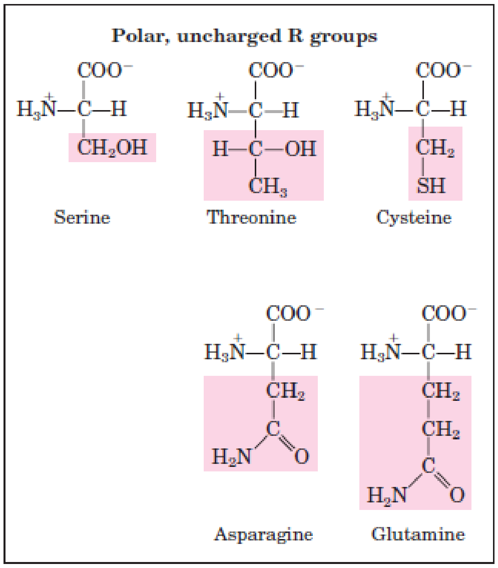
- The two amino acids having R groups with a net negative charge at pH 7.0 are aspartate and glutamate, each of which has a second carboxyl group.





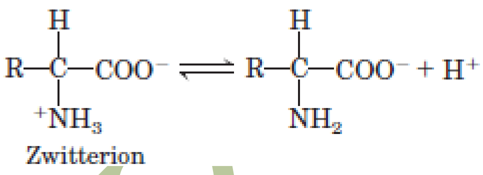


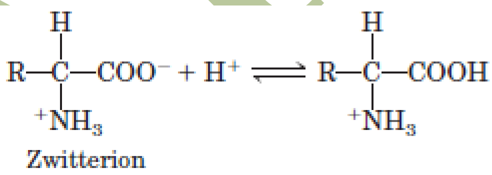




Amino Acids Can Act as Acids and Bases

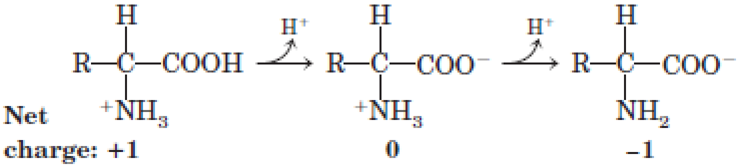
- When an amino acid is dissolved in water, it exists in solution as the dipolar ion, or zwitterion. A zwitterion can act as either an acid (proton donor) or a base (proton acceptor).





Example : an acid A base

- Substances having this dual nature are **amphoteric** and are often called **ampholytes.**
- A simple monoamino monocarboxylic *α*- amino acid, such as alanine, is a diprotic acid when fully protonated, it has two groups, the COOH group and the NH3 group, that can yield protons.

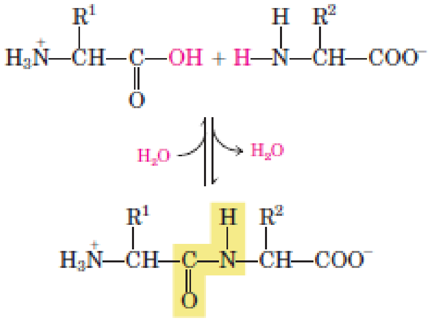


Proteins/ Peptides

- Polymers of amino acids are known as the peptides and proteins. Biologically occurring polypeptides range in size from small to very large, consisting of two or three to thousands of linked amino acid residues.
- Two amino acid molecules can be covalently joined through a substituted amide linkage, termed a **peptide bond,** to yield a dipeptide. Such a linkage is formed by removal of the elements of water (dehydration) from the α-carboxyl group of one amino acid and the α-amino group of another. Peptide bond formation is an example of a condensation reaction.
- Three amino acids can be joined by two peptide bonds to form a tripeptide; similarly, amino acids can be linked to form tetrapeptides, pentapeptides, and so forth. When a few amino acids are joined in this fashion, the structure is called an oligopeptide. When many amino acids are joined, the product is called a polypeptide.
- Proteins may have thousands of amino acid residues. Although the terms "protein" and "polypeptide" are sometimes used interchangeably, molecules referred to as polypeptides generally have molecular weights below 10,000, and those called proteins have higher molecular weights.

Formation of a peptide bond by condensation.

The α-amino group of one amino acid (with R2 group) acts as a nucleophile to displace the hydroxyl group of another amino acid (with R1 group), forming a peptide bond (shaded).



- In a peptide, the amino acid residue at the end with a free α-amino group is the

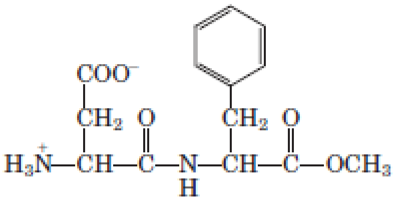
amino-terminal (or *N*-terminal) residue; the residue at the other end, which has a free carboxyl group, is the **carboxyl-terminal** (*C*-terminal) residue.

- Hydrolysis of a peptide bond is an exergonic reaction, it occurs slowly because of its high activation energy. As a result, the peptide bonds in proteins are quite stable, with an average half-life (*t*1/2) of about 7 years under most intracellular conditions.
- Like free amino acids, peptides have characteristic titration curves and a characteristic isoelectric pH (pI) at which they do not move in an electric field. These properties are exploited in some of the techniques used to separate peptides and proteins

Biologically Active Peptides and Polypeptides Occur in a Vast Range of Sizes

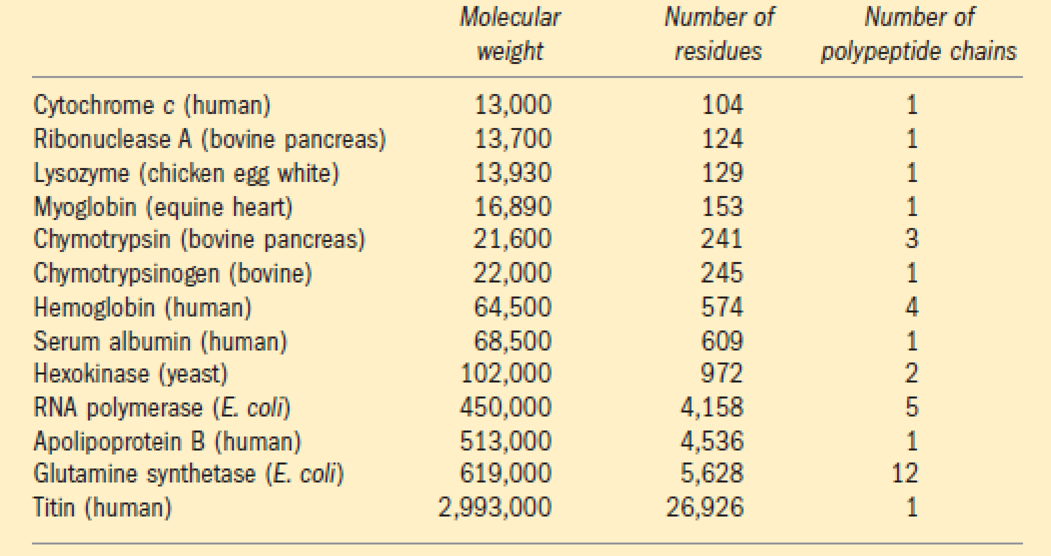
No generalizations can be made about the molecular weights of biologically active peptides and proteins in relation to their functions. Naturally occurring peptides range in length from two to many thousands of amino acid residues. Even the smallest peptides can have biologically important effects.

Example: Consider the commercially synthesized dipeptide L-aspartyl-L-phenylalanine methyl ester, the artificial sweetener better known as aspartameor NutraSweet.



L-Aspartyl-L-phenylalanine methyl ester
(aspartame)

Molecular Data on Some Proteins



	Molecular weight	Number of residues	Number of polypeptide chains
Cytochrome *c* (human)	13,000	104	1
Ribonuclease A (bovine pancreas)	13,700	124	1
Lysozyme (chicken egg white)	13,930	129	1
Myoglobin (equine heart)	16,890	153	1
Chymotrypsin (bovine pancreas)	21,600	241	3
Chymotrypsinogen (bovine)	22,000	245	1
Hemoglobin (human)	64,500	574	4
Serum albumin (human)	68,500	609	1
Hexokinase (yeast)	102,000	972	2
RNA polymerase (*E. coli*)	450,000	4,158	5
Apolipoprotein B (human)	513,000	4,536	1
Glutamine synthetase (*E. coli*)	619,000	5,628	12
Titin (human)	2,993,000	26,926	1

Types of protein

1. Simple proteins

Many proteins contain only amino acid residues and no other chemical constituents; these are considered simple proteins. It is further classified into following groups.

Example: The enzymes ribonuclease A and chymotrypsinogen,

Based on solubility

a) Albumin – water soluble
b) Globulin – Water insoluble

Based on overall shape

a) Gloubular protein - have axial ratio less than 10, are characterized by compactly folded and coiled polypeptide chains. Example- insulin
b) Fibrous protein – have axial ratio greater than 10, are characterized by group of polypeptide chains coiled in a spiral or helix and cross-linked covalently or by hydrogen bonds. Example – keratin

Based on function

Function	Protein
Catalytic role	Enzymes
Contraction	Actin, myosin
Gene regulation	Histones, non -histone nuclear proteins
Hormone role	Insulin
Protection	Fibrin, interferon
Regulatory role	Calmodulin
Structural role	Collagen, elastin
Transport	Albumin, fatty acids

2. Conjugated proteins

Some proteins contain permanently associated chemical components in addition to amino acids; these are called conjugated proteins.

The non–amino acid part of a conjugated protein is usually called its prosthetic group.

Conjugated proteins are classified into different types on the basis of the nature of prosthetic group attached. They are as follow in the table below.

Class	Prosthetic group	Example
Nucleoprotein	Nucleic acid	Constituents of chromatin
phosphoprotein	Phosphoric acid	caesin
Glycoprotein	Carbohydrates(less than 4% hexosamine)	Immunoglobulin G
Lipoprotein	Lipids	B_1-Lipoprotein
Mucoprotein	Carbohydrates(more than 4%	Ovomucin

	hexosamine)	
Chromoprotein	Heterocyclic compounds like porphyrins	Hemoglobin,melanoprotein
Metalloprotein	metals	Ferritin(iron),ceruloplasm(copper)
Flavoprotein	Flavin nucleotides	Succinate dehydrogenase

3. Derived proteins

These are derived from simple proteins or conjugated proteins by the action of acids, alkalies or enzymes. They are product resulted from partial to complete hydrolysis of the protein. They are two types

a) Primary derived proteins

They are metaproteins, derived from denaturation by the action of heat, acids and alkalies.

b) Secondary derived proteins

They are obtained at a later stage of hydrolysis. Example Proteases, Peptones, Peptides and diketopiperazines.

Structure of proteins

- A major requirement for understanding protein structure is a large database of three-dimensional structures. This is particularly important for the comparative method of structure prediction.
- There are two methods by which protein structures can be determined: X-ray crystallography and NMR
- The character of a protein is determined by the amino acid sequence and composition of the polypeptide chain.
- Four levels of protein structure are commonly defined. They are as follows,

1. Primary structure

The **primary structure** of a protein is the sequence of amino acids.

The amino acids are arranged in a linear form.

Example : Lysozyme is an enzyme that attacks bacterial cell walls. It is found in secretions such as tears and in the white of eggs. Lysozyme has the following primary structure:

(NH2)KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWWCDNGRTPGSRNLCNIPCSALLSSDITASVNCAKKIVSDGDGMNAWVAWRNRCKGTDVQAWIRGCRL(COOH).

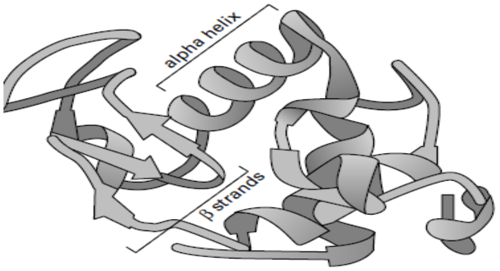
2. **Secondary structure**

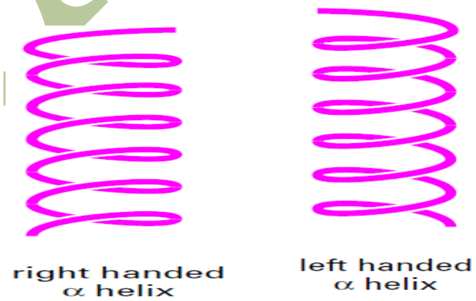
Two types of protein backbone organization are common to many proteins.

These are named the ***α*** **helix** and the ***β*** **sheet.** Collectivity of these two repeating patterns are known as Secondary structure of proteins.

α **helix**

- In an *α* helix the polypeptide chain twists around in a spiral, each turn of the helix taking 3.6 amino acid residues.
- This allows the nitrogen atom in each peptide bond to form a hydrogen bond with the oxygen four residues ahead of it in the polypeptide chain.
- All the peptide bonds in the helix are able to form such hydrogen bonds, producing a rod in which the amino acid side chains point outward. Because it introduces a kink into the polypeptide chain, proline cannot participate in α helix.

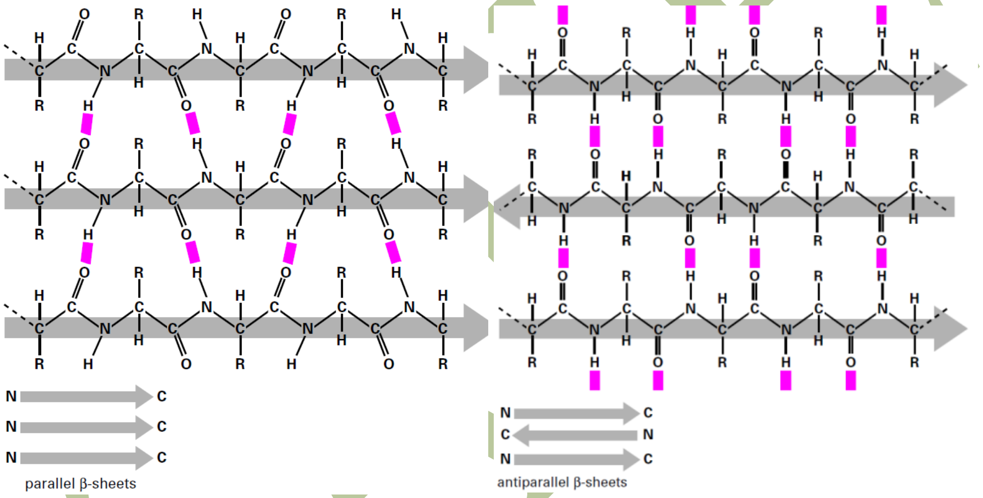




β **sheet**

- In a *β* sheet lengths of polypeptide run alongside each other, and hydrogen bonds form between the peptide bonds of the strands.

- This generates a sheet that has the side chains protruding above and below it.
- Along a single strand the side chains alternate up then down, up then down. Because the actual geometry prevents them from being completely flat, they are sometimes called *β* pleated sheets.
- A polypeptide chain can form two types of *β* sheet: Either all of the strands in the *β* sheet are running in the same direction forming a **parallel *β* sheet** or they can alternate in direction making an **antiparallel*β* sheet.**
- The polypeptide chains in *β* sheets are fully extended unlike the chain in an *α* helix.



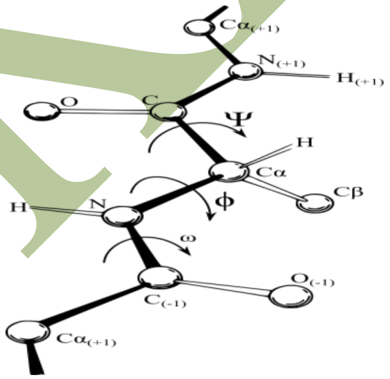
Ramachandran plot

- A Ramachandran plot (also known as a Ramachandran diagram or a [φ,ψ] plot), originally developed in 1963 by G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, is a way to visualize backbone dihedral angles ψ against φ of amino acid residues in protein structure.
- The figure at left illustrates the definition of the φ and ψ backbone dihedral angles (called φ and φ' by Ramachandran). The ω angle at the peptide bond is normally 180°, since the partial-double-bond character keeps the peptide planar.

- The figure at top right shows the allowed φ,ψ backbone conformational regions from the Ramachandran et al. 1963 and 1968 hard-sphere calculations: full radius in solid outline, reduced radius in dashed, and relaxed tau (N-Calpha-C) angle in dotted lines.
- Because dihedral angle values are circular and 0° is the same as 360°, the edges of the Ramachandran plot "wrap" right-to-left and bottom-to-top. For instance, the small strip of allowed values along the lower-left edge of the plot are a continuation of the large, extended-chain region at upper left.

Uses of Ramachandran plot

- A Ramachandran plot can be used in two somewhat different ways. One is to show in theory which values, or conformations, of the ψ and φ angles are possible for an amino-acid residue in a protein (as at top right).
- A second is to show the empirical distribution of datapoints observed in a single structure (as at right, here) in usage for structure validation, or else in a database of many structures (as in the lower 3 plots at left). Either case is usually shown against outlines for the theoretically favored regions.



3. Tertiary structure

- The three-dimensional protein structure often has protrusions, clefts, or grooves on the surface where particular amino acids are positioned to form sites that bind ligands.
- In the case of enzymes, catalyze reactions within or between ligands.
- The whole three dimensional arrangement of the amino acids in the protein is called the tertiary structure.
- A tertiary structure is unique to a particular protein. However, common patterns or motifs occur in tertiary structures.

Determinants of tertiary proteins

- Globular proteins have a core of hydrophobic amino acid residues and a surface region of water-exposed, charged, hydrophilic residues. This arrangement may stabilise interactions within the tertiary structure.
- For example, in secreted proteins, which are not bathed in cytoplasm, disulfide bonds between cysteine residues help to maintain the tertiary structure. There is a commonality of stable tertiary structures seen in proteins of diverse function and diverse evolution.
- For example, the TIM barrel, named for the enzyme triose phosphate isomerase, is a common tertiary structure as is the highly stable, dimeric, coiled coil structure. Hence, proteins may be classified by the structures they hold.
- Databases of proteins which use such a classification include *SCOP* and*CATH*.

Stability of native states

- The native state or native conformation of a protein is its most typical conformation in a cellular environment.
- **Chaperone proteins -** It is commonly assumed that the native state of a protein is also the most thermodynamically stable and that a protein will reach its native state, given its chemical kinetics, before it is translated.

- Protein chaperones within the cytoplasm of a cell assist a newly synthesised polypeptide to attain its native state.
- Some chaperone proteins are highly specific in their function, for example,protein disulfide isomerase; others are general in their function and may assist most globular proteins, for example, the prokaryotic GroEL/GroES system of proteins and the homologous eukaryotic heat shock proteins (the Hsp60/Hsp10 system).

Kinetic traps

- Folding kinetics may trap a protein in a high-energy conformation. The high-enery conformation may contribute to the function of the protein.
- For example, the Influenza hemagglutinin protein is a single polypeptide chain which when activated, is proteolytically cleaved to form two polypeptide chains. The two chains are held in a high-energy conformation.
- When the local pH drops, the protein undergoes an energetically favorable conformational rearrangement that enables it to penetrate the host cell membrane.

Metastability

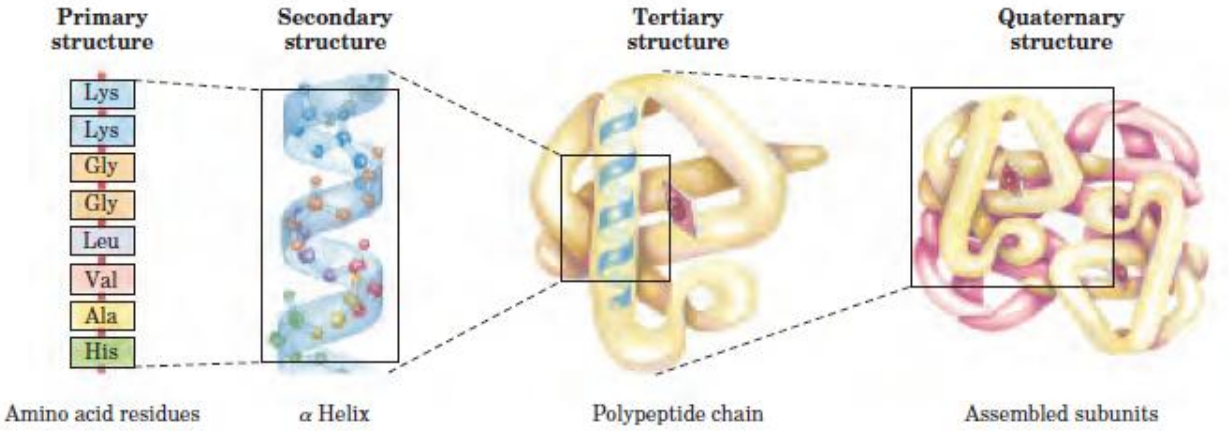
- Some tertiary protein structures may exist in long-lived states which are less than the expected most stable state. For example, many serpins (serine protease inhibitors) show this metastability.
- They undergo a conformational change when a loop of the protein is cut by a protease.

Cytoplasmic environment

4. Quaternary structure

- Many globular proteins have further level of organization called quaternary structure, which describes the association of protein units to produce an aggregate protein with a definite functional property.
- The complex catalytic function of isoenzymes is depending upon their quaternary structure.

- The bonds involved are covalent and mainly hydrophobic between nonpolar regions on the surface on the molecules concerned.
- For example hemoglobin is composed of four polypeptide chains, normally in two identical pairs, forming hemoglobin tetramer; it is more effective in oxygen transfer than in monomeric form.
- In addition to the four protein chains, hemoglobin also incorporates an iron porphyrin, which facilitates the binding of oxygen.



Nucleic Acids

DNA

- Deoxyribonucleic acids (DNAs) are polymeric molecules consisting of nucleotide building blocks. They are genetic material of almost all organisms except some RNA viruses.
- In prokaryotes, DNA is not separated from the rest of the cellular contents.In eukaryotes, however, DNA is located in the nucleus, where it is separated fromthe rest of the cell by the nuclear envelope.
- Eukaryotic DNA is bound to proteins, forming a complex called chromatin. During interphase (when cells are not dividing), some of the chromatin is diffuse (euchromatin) and some isdense (heterochromatin), but no distinct structures can be observed.

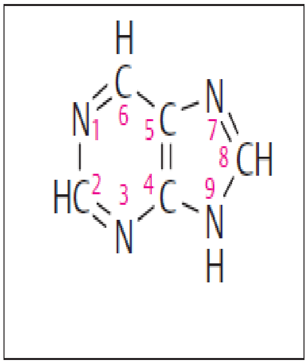
- However,before mitosis (when cells divide), the DNA is replicated, resulting in two identical chromosomes called sister chromatids. During metaphase (a period in mitosis), these condense into discrete, visible chromosomes.
- Less than 0.1% of the total DNA in a cell is present in mitochondria. The genetic information in a mitochondrion is encoded in less than 20,000 base pairs of DNA.

Composition of DNA

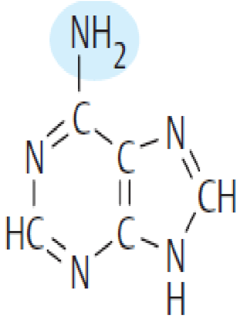
All nucleic acids are made up from nucleotide components, which in turn consist of a base, a sugar, and a phosphate residue. (phosphoric acid)

Bases

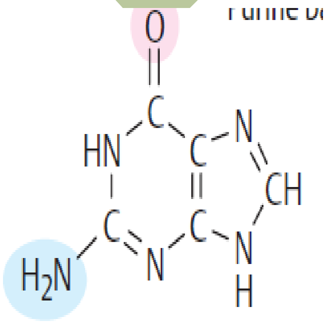
- They are aromatic heterocyclic compounds derived from either **pyrimidine** or **purine**.
- The purine bases **adenine** and **guanine** and the pyrimidine base **thymine** and **cytosine** are present in DNA.



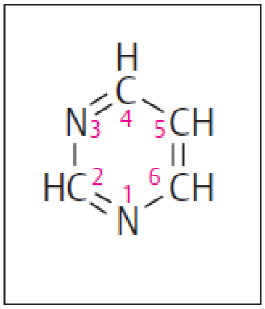
Purine



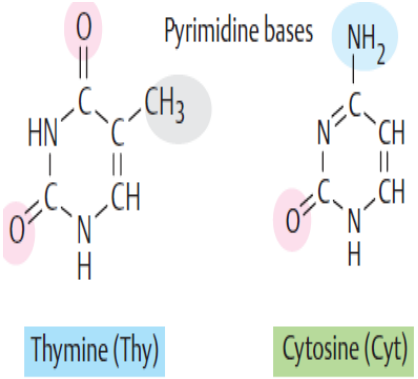
Adenine (Ade)



Guanine (Gua)

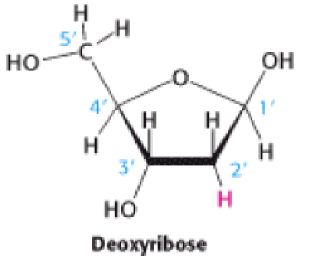


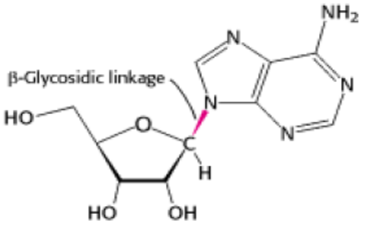
Pyrimidine



Sugars

- DNA contain 2’- deoxy-D-ribose, the pentose residues are present in the furanose form.
- The sugars and bases are linked by an *N*-glycosidic bond between the C-1 of the sugar and either the N-9 of the purine ring or N-1 of the pyrimidine ring. This bond always adopts the β-configuration.





Glycosidic linkage in a nucleoside.

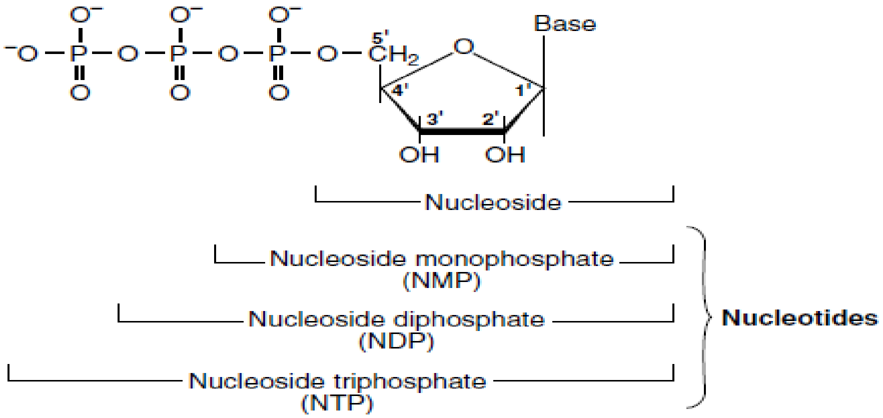
Nucleosides

- In nucleosides, the nitrogenous base is linked by an *N*-glycosidic bond to the anomeric carbon of the sugar, i.e., deoxyribose.
- When a nucleic acid base is N-glycosidically linked to ribose or 2-deoxyribose, it yields a nucleoside. (Base + Sugar)

Nucleotides

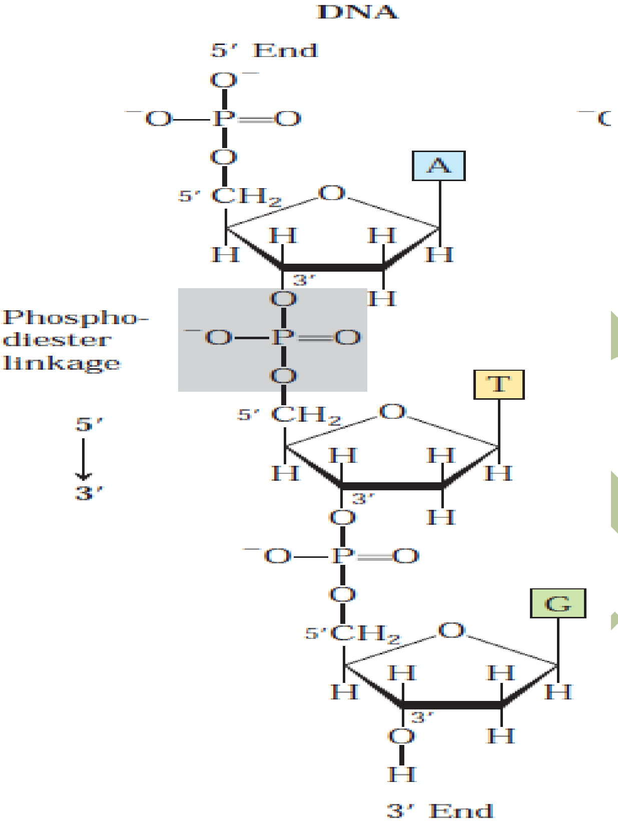
- A nucleotide is a nucleoside with an inorganic phosphate attached to a 5’-hydroxyl group of the sugar in ester linkage. The names and abbreviations of nucleotides specify the base, the sugar, and the number of phosphates attached (MP, **mono****p**hosphate; DP, **di****p**hosphate; TP, **tri****p**hosphate).

- In deoxynucleotides, the prefix "d" precedes the abbreviation. For example, GDP is guanosine diphosphate (the base guanine attached to a ribose that has two phosphate groups) and dATP is deoxyadenosine triphosphate (the base adenine attached to a deoxyribose with three phosphate groups).


Base
5'
4'
1'
3'
2'
OH
OH
Nucleoside
Nucleoside monophosphate (NMP)
Nucleoside diphosphate (NDP)
Nucleoside triphosphate (NTP)
Nucleotides

Phosphodiester Bonds

- The successive nucleotides of DNA are covalently linked through phosphate-group "bridges," in which the 5_-phosphate group of one nucleotide unit is joined to the 3_-hydroxyl group of the next nucleotide, creating a **phosphodiester linkage**
- Thus the covalent backbones of nucleic acids consist of alternating phosphate and pentose residues, and the nitrogenous bases may be regarded as side groups joined to the backbone at regular intervals. The backbones of DNA are hydrophilic.
- All the phosphodiester linkages have the same orientation along the chain (Fig. 8–7), giving each linear nucleic acid strand a specific polarity and distinct 5' and 3' ends. By definition, the **5' end** lacks a nucleotide at the 5' position and the **3' end** lacks a nucleotide at the 3' position. Other groups (most often one or more phosphates) may be present on one or both ends.



Base paring

Concept of Base-Pairing was proposed by Chargaff in 1950. His proposal is called as Chargaff's rule of Base pairing.

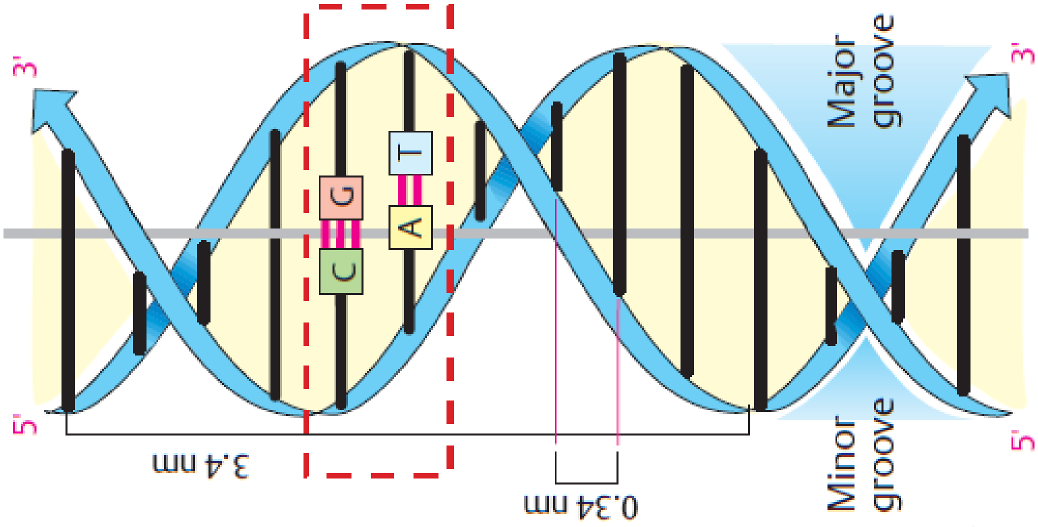
According to Chargaff's rule of Base pairing,

1) Total amount of purines equaled the total amount of pyrimidines. (A+G=T+C)
2) Adenine always pairs with thymine (A+T) and guanine always pairs with cytosine (G+C)
3) The amount of adinine equaled the amount of thymine (A=T), likewise, amount of guanine equaled the amount of cytosine (G=C).
4) Two hydrogen bonds are formed between adenine and thymine and three hydrogen bonds are formed between guanine and cytosine.

Watson and Crick model of DNA

Based on x-ray analysis, Watson and Crick proposed the structure of DNA, according to him,

- The two strands are complementary to each other.
- The two complementary strands of DNA run in opposite directions. On one strand, the 5'-carbon of the sugar is above the 3'-carbon. This strand is said to run in a 5_ to 3_ direction. On the other strand, the 3'-carbon is above the 5'-carbon. This strand is said to run in a 3' to 5' direction.
- Thus, the strands are antiparallel (that is, they run in opposite directions.)
- The two strands are wrapped helically around each other,with sugar-phosphate chain on the outside(forming ribbon like backbone of double helix) and purines and pyrimidines on the inside of the helix(projecting between two sugar phosphate backbones as transverse bars).
- Both polynucleotide strands remains separated by 20A° distance.
- The coiling of double helix is right handed and a complete turn occurs every 34A°. Since each nucleotide occupies 3.4A° distance along the length of a polynucleotide strand, ten mononucleotides occur per complete turn.(10 base pair per turn of the helix)
- The offset pairing of the two strands creates a **major groove** and **minor groove** on the surface of the duplex.



Polymorphism of DNA helix / Alternative forms of DNA double helix

For about 20 years after discovery of DNA double helix in 1953, some experiments shown that DNA is much more polymorphic. Thus DNA has following Types,

B-form/B DNA

- Biologically important form of DNA, naturally found in most living systems. Watson and Crick model DNA.

A-form/A DNA

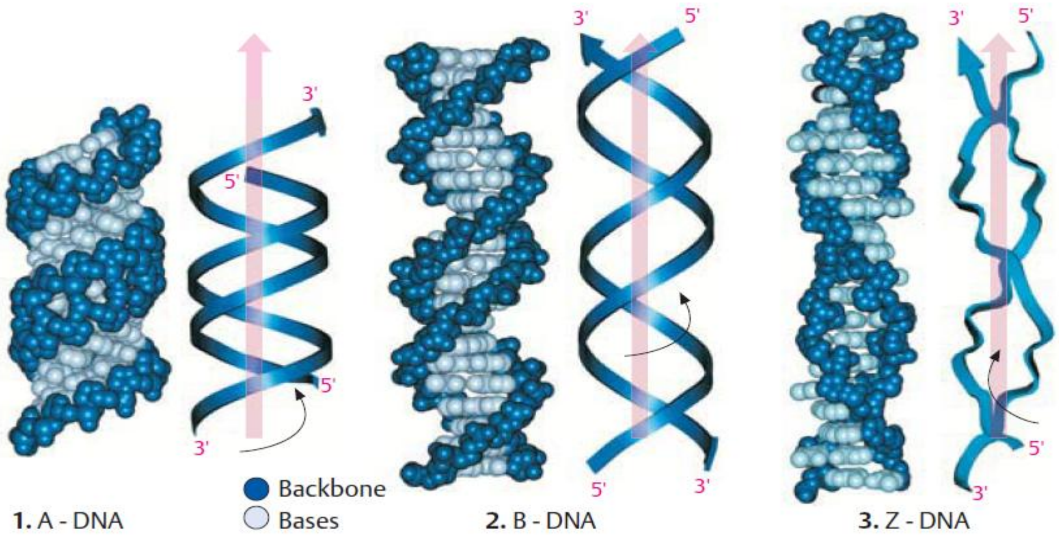
- It is right handled but less hydrated than B-form DNA.
- It is more compact with 11 base pair per turn of the helix.
- The double helix is 23A° in diameter.
- The bases are tilted more in relation to the axis of the helix than in the B-DNA.

Z-form/Z DNA

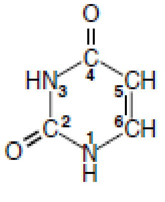
- It is observed by crystallographic studies. It reveals that synthetic nucleotides consists of alternating purines and pyrimidines such as GCGCGCGCGCGC.
- They are called as Z DNA because of its zigzag nature.
- They are left handed DNA, with 12 base pair per turn of helix.
- It is found in solutions of high-ionic strength, Example- 2M NaCl.
- The double helix is 18A° in diameter.

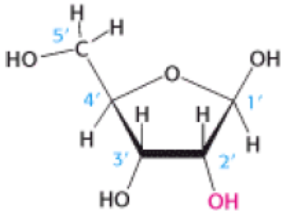
Other forms of DNA

- C form DNA found at 66 percent relative humidity with Li^+ ions.
- D form and E form DNA are found as rare extreme variants and has only 8 and 7.5 base pair per turn respectively. These DNAs lack guanine.



RNA

- RNA is the genetic material of some of viruses.
- RNA is similar to DNA. Like DNA, it is composed of nucleotides joined by 3'- to 5' phosphodiester bonds, the purine bases adenine and guanine, and the pyrimidine base cytosine. However, its other pyrimidine base is uracil rather than thymine.
- Uracil and thymine are identical bases except that thymine has a methyl group at position 5 of the ring. In RNA, the sugar is ribose, which contains a hydroxyl group on the 2´-carbon.
- 

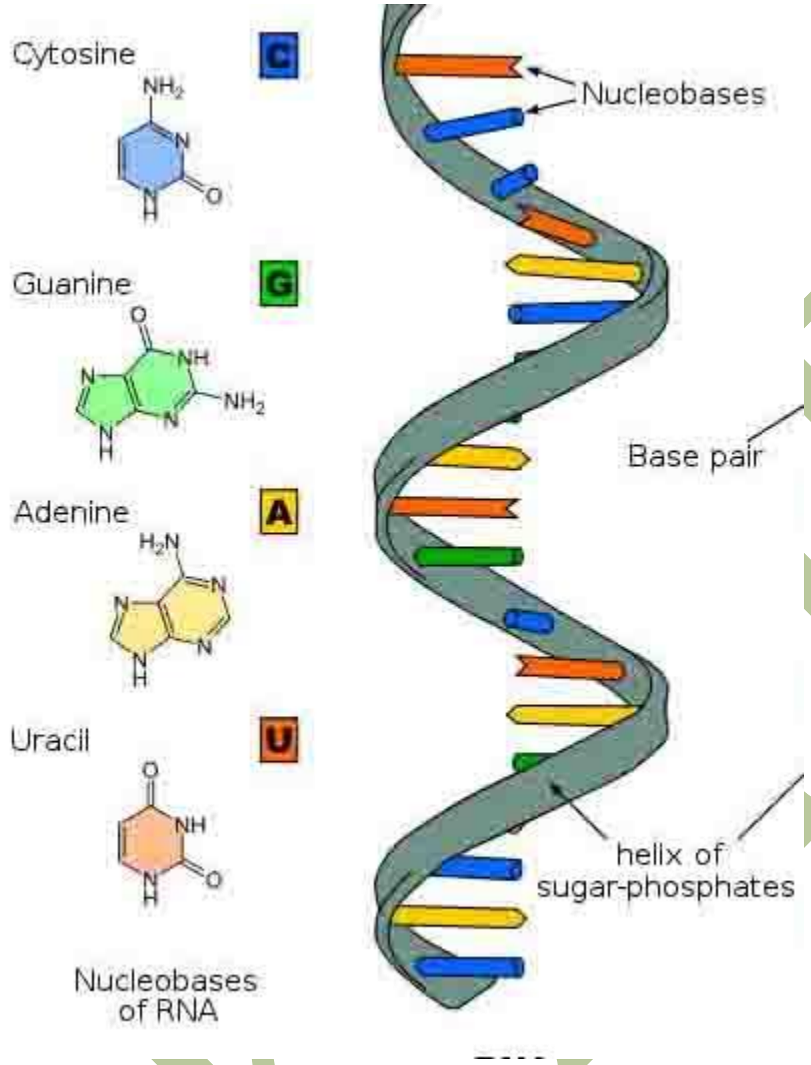
Uracil (U)


Ribose
- RNA chains are usually single-stranded and lack the continuous helical structure of double stranded DNA.

- However, RNA still has considerable secondary and tertiary structure because base pairs can form in regions where the strand loops back on itself. As in DNA, pairing between the bases is complementary and antiparallel.
- But in RNA, adenine pairs with uracil rather than thymine. Basepairing in RNA can be extensive, and the irregular looped structures generated are important for the binding of molecules, such as enzymes, that interact with specific regions of the RNA.

Types of RNA

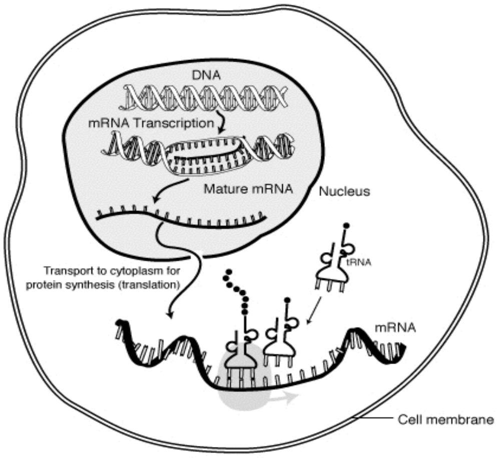
- The three major types of RNA (mRNA, rRNA, and tRNA) participate directly in the process of protein synthesis. Other less abundant RNAs are involved in replication or in the processing of RNA, that is, in the conversion of RNA precursors to their mature forms.



Structure of mRNA

- ➢ Each mRNA molecule contains a nucleotide sequence that is converted into the amino acid sequence of a polypeptide chain in the process of translation.
- ➢ In eukaryotes, messenger RNA (mRNA) is transcribed from protein-coding genes as a long primary transcript that is processed in the nucleus to form mRNA.
- ➢ The various processing intermediates, which are mRNA precursors, are called pre-mRNA or hnRNA (heterogenous nuclear RNA).

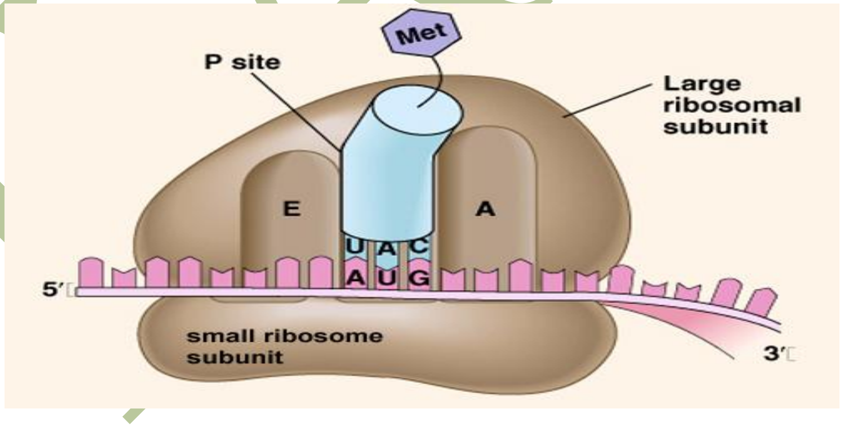
- mRNA travels through nuclear pores to the cytoplasm, where it binds to ribosomes and tRNAs and directs the sequential insertion of the appropriate amino acids into a polypeptide chain. Eukaryotic mRNA consists of a leader sequence at the 5´ end, a coding region, and a trailer sequence at the 3' end.
- The leader sequence begins with a guanosine cap structure at its 5' end. The coding region begins with a trinucleotide start codon that signals the beginning of translation, followed by the trinucleotide codons for amino acids, and ends at a termination signal.
- The trailer terminates at its 5' end with a poly (A) tail that may be up to 200 nucleotides long. Most of the leader sequence, all of the coding region, and most of the trailer are formed by transcription of the complementary nucleotide sequence in DNA.
- However, the terminal guanosine in the cap structure and the poly(A) tail do not have complementary sequences; they are added posttranscriptionally.



Structure of rRNA

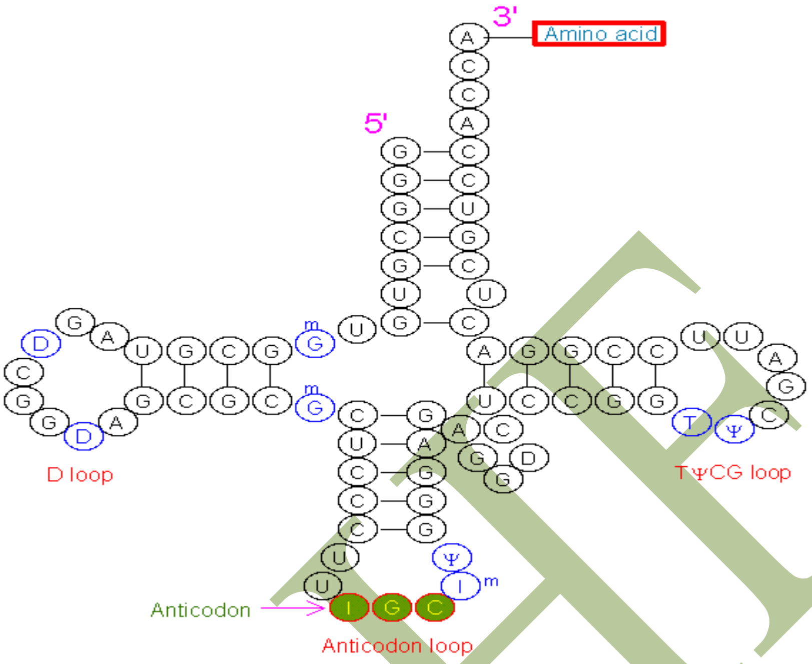
- Ribosomes are subcellular ribonucleoprotein complexes on which protein synthesis occurs. Different types of ribosomes are found in prokaryotes and in the cytoplasm
- and mitochondria of eukaryotic cells.
- Prokaryotic ribosomes contain three types of rRNA molecules with sedimentation coefficients of 16, 23, and 5S.

- The 30S ribosomal subunit contains the 16S rRNA complexed with proteins, and the 50S ribosomal subunit contains the 23S and 5S rRNAs complexed with proteins.
- The 30S and 50S ribosomal subunits join to form the 70S ribosome, which participates in protein synthesis.
- Cytoplasmic ribosomes in eukaryotes contain four types of rRNA molecules of 18, 28, 5, and 5.8S.
- The 40S ribosomal subunit contains the 18S rRNA complexed with proteins, and the 60S ribosomal subunit contains the 28, 5, and 5.8S rRNAs complexed with proteins. In the cytoplasm, the 40S and 60S ribosomal subunits combine to form the 80S ribosomes that participate in protein synthesis.
- Mitochondrial ribosomes, with a sedimentation coefficient of 55S, are smaller than cytoplasmic ribosomes. Their properties are similar to those of the 70S ribosomes of bacteria.
- rRNAs contain many loops and exhibit extensive base-pairing in the regions between the loops. The sequences of the rRNAs of the smaller ribosomal subunits exhibit secondary structures that are common to many different genera.



`Structure of tRNA

- During protein synthesis, tRNA molecules carry amino acids to ribosomes and ensure that they are incorporated into the appropriate positions in the growing polypeptide chain.
- This is done through base-pairing of three bases of the tRNA (the anticodon) with the three base codons within the coding region of the mRNA.
- Therefore, cells contain at least 20 different tRNA molecules that differ somewhat in nucleotide sequence, one for each of the amino acids found in proteins.
- Many amino acids have more than one tRNA. tRNA molecules contain not only the usual nucleotides, but also derivatives of these nucleotides that are produced by posttranscriptional modifications.
- In eukaryotic cells, 10 to 20% of the nucleotides of tRNA are modified. Most tRNA molecules contain ribothymidine (T), in which a methyl group is added to uridine to form ribothymidine. They also contain dihydrouridine (D), in which one of the double bonds of the base is reduced; and pseudouridine (Ψ), in which uracil is attached to ribose by a carbon–carbon bond rather than a nitrogen–carbon bond.
- The base at the 5'-end of the anticodon of tRNA is frequently modified. tRNA molecules are rather small compared with both mRNA and the large rRNA molecules. On average, tRNA molecules contain approximately 80 nucleotides and have a sedimentation coefficient of 4S.
- Because of their small size and high content of modified nucleotides, tRNAs were the first nucleic acids to be sequenced. Since 1965 when Robert Holley deduced the structure of the first tRNA, the nucleotide sequences of many different tRNAs have been determined. Although their primary sequences differ, all tRNA molecules can form a structure resembling a cloverleaf.



Protein Information Sources

- The Protein Information Resource (PIR) is an integrated public resource of protein informatics that supports genomic and proteomic research and scientific discovery. PIR maintains the Protein Sequence Database (PSD), an annotated protein database containing over 283 000 sequences covering the entire taxonomic range.

- Family classification is used for sensitive identification, consistent annotation, and detection of annotation errors. The superfamily curation defines signature domain architecture and categorizes memberships to improve automated classification. To increase the amount of experimental annotation, the PIR has developed a bibliography system for literature searching, mapping, and user submission, and has conducted retrospective attribution of citations for experimental features.

- PIR also maintains NREF, a non-redundant reference database, and iProClass, an integrated database of protein family, function, and structure information. PIR-NREF provides a timely and comprehensive collection of protein sequences, currently consisting of more than 1 000 000 entries from PIR-PSD, SWISS-PROT, TrEMBL, RefSeq, GenPept, and PDB. The PIR web site (http://pir.georgetown.edu) connects data analysis tools to underlying databases for information retrieval and knowledge discovery, with functionalities for interactive queries, combinations of sequence and text searches, and sorting and visual exploration of search results.

The Protein Data Bank (PDB)

- The Protein Data Bank (PDB) is a crystallographic database for the three-dimensional structural data of large biological molecules, such as proteins and nucleic acids. The data, typically obtained by X-ray crystallography, NMR spectroscopy, or, increasingly, cryo-electron microscopy, and submitted by biologists and biochemists from around the world, are freely accessible on the Internet via the websites of its member organisations (PDBe, PDBj, and RCSB). The PDB is overseen by an organization called the Worldwide Protein Data Bank, wwPDB.
- The PDB is a key resource in areas of structural biology, such as structural genomics. Most major scientific journals, and some funding agencies, now require scientists to submit their structure data to the PDB. Many other databases use protein structures deposited in the PDB. For example, SCOP and CATH classify protein structures, while PDBsum provides a graphic overview of PDB entries using information from other sources, such as Gene ontology.

TrEMBL

- UniProtKB/TrEMBL is a computer-annotated protein sequence database complementing the UniProtKB/Swiss-Prot Protein Knowledgebase.
- UniProtKB/TrEMBL contains the translations of all coding sequences (CDS) present in the EMBL/GenBank/DDBJ Nucleotide Sequence Databases and also protein sequences extracted from the literature or submitted to UniProtKB/Swiss-Prot.
- The database is enriched with automated classification and annotation.

The UniProtKB/TrEMBL group is headed by: Rolf Apweiler.

SWISS-PROT

- ➢ Which provides detailed sequence annotation that includes structure, function, and protein family assignment.
- ➢ The sequence data are mainly derived from TrEMBL, a database of translated nucleic acid sequences stored in the EMBL database.

- The annotation of each entry is carefully curated with good quality by human experts. The protein annotation includes function, domain structure, catalytic sites, cofactor binding, post translational modification, metabolic pathway information, disease association, and similarity with other sequences.
- Much of this information is obtained from scientific literature and entered by database curators.
- The annotation provides significant added value to each original sequence record.
- The data record also provides cross referencing links to other online resources of interest. Other features such as very low redundancy and high level of integration with other primary and secondary databases make SWISS-PROT very popular among biologists.

Restriction digestion

- Restriction digestion is a process in which DNA is cut at specific sites, dictated by the surrounding DNA sequence. Restriction digestion is accomplished by incubation of the target DNA molecule with restriction enzymes - enzymes that recognize and bind specific DNA sequences and cleave at specific nucleotides either within the recognition sequence or outside of the recognition sequence.

- Restriction digestion can result in the production of blunt ends (ends of a DNA molecule that end with a base pair) or sticky ends (ends of a DNA molecule that end with a nucleotide overhang). Restriction digestion is usually used to prepare a DNA fragment for subsequence molecular cloning, as the procedure allows fragments of DNA to be pieced together like building blocks via ligation.

- The results of a restriction digestion can be evaluated by gel electrophoresis, in which the products of the digestion are separated by molecule length (based on the negative charge of DNA molecules) in a polymer gel to which an electric field has been applied.

- The components of a typical restriction digestion reaction include the DNA template, the restriction enzyme of choice, a buffer and sometimes BSA protein. The reaction is incubated at a specific temperature required for optimal activity of the restriction enzyme and terminated by heat.

Chromatogram

- A chromatogram (sometimes also called electropherogram) is the visual representation of a DNA sample produced by a sequencing machine (such as Applied Biosystems ABI PRISM 7700 Sequence Detection System).

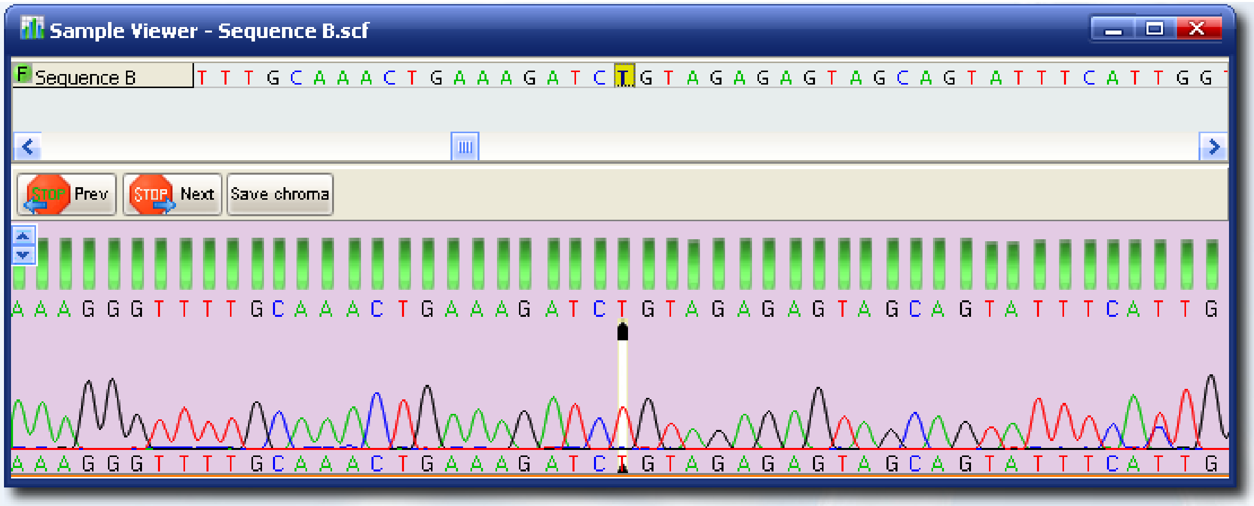

Sample Viewer - Sequence B.scf
F Sequence B
TTTGCAAACTGAAAGATCTGTAGAGAGTAGCAGTATTTCATTGG
Prev
Next
Save chroma
AAAGGGTTTTGCAAACTGAAAGATCTGTAGAGAGTAGCAGTATTTCATTG
AAAGGGTTTTGCAAACTGAAAGATCTGTAGAGAGTAGCAGTATTTCATTG

Fig. 1 - Example of chromatogram. The green bars above chromatogram peaks high confidence scores.

- With DNA Baser it is easy to make the difference between a good (trusted) chromatogram and a poor one. Just look at the green bars displayed above each base. These bars represents the confidence scores. A high bar means that the base call can be trusted. A low bar means that the base call cannot be trusted.
- Usually, poor chromatograms enter unwanted ambiguities in your contig. However, if you assemble a good chromatogram and a poor one, DNA Baser will know how to extract the useful information from he good chromatogram and give less trust to the bad chromatogram.

Blot

- A **blot**, in molecular biology and genetics, is a method of transferring proteins, DNA or RNA, onto a carrier (for example, a nitrocellulose, polyvinylidene fluoride (PVDF) or nylonmembrane).
- In many instances, this is done after a gel electrophoresis, transferring the molecules from the gel onto the blotting membrane, and other times adding the samples directly onto the

membrane. After the blotting, the transferred proteins, DNA or RNA are then visualized by colorant staining (for example, silver staining of proteins), autoradiographic visualization of radioactive labelled molecules (performed before the blot), or specific labelling of some proteins or nucleic acids.

- The latter is done with antibodies or hybridization probes that bind only to some molecules of the blot and have an enzyme joined to them.
- After proper washing, this enzymatic activity (and so, the molecules we search in the blot) is visualized by incubation with proper reactive, rendering either a colored deposit on the blot or a chemiluminiscent reaction which is registered by photographic film.

Southern Blot

A Southern blot is a method routinely used in molecular biology for detection of a specific DNA sequence in DNA samples. Southern blotting combines transfer of electrophoresis-separated DNA fragments to a filter membrane and subsequent fragment detection by probe hybridization.

Western Blot

A Western blot is used for the detection of specific proteins in complex samples. Proteins are first separated by size using electrophoresis before being transferred to an appropriate blotting matrix (usually PVDF or nitrocellulose) and subsequent detection with antibodies.

Northern Blot

A Southern blot is a method routinely used in molecular biology for detection of a specific RNA sequence in RNA samples. Southern blotting combines transfer of electrophoresis-separated RNA fragments to a filter membrane and subsequent fragment detection by probe hybridization.

Microarray

- A typical microarray experiment involves the hybridization of an mRNA molecule to the DNA template from which it is originated. Many DNA samples are used to construct an array. The amount of mRNA bound to each site on the array indicates the expression level of the various genes. This number may run in thousands. All the data is collected and a profile is generated for gene expression in the cell.

- An array is an orderly arrangement of samples where matching of known and unknown DNA samples is done based on base pairing rules. An array experiment makes use of common assay systems such as microplates or standard blotting membranes. The sample spot sizes are typically less than 200 microns in diameter usually contain thousands of spots.

- Thousands of spotted samples known as probes (with known identity) are immobilized on a solid support (a microscope glass slides or silicon chips or nylon membrane). The spots can be DNA, cDNA, or oligonucleotides. These are used to determine complementary binding of the unknown sequences thus allowing parallel analysis for gene expression and gene discovery. An experiment with a single DNA chip can provide information on thousands of genes simultaneously. An orderly arrangement of the probes on the support is important as the location of each spot on the array is used for the identification of a gene.

Mass Spectrometry

Mass spectrometry is a powerful analytical technique used to quantify known materials, to identify unknown compounds within a sample, and to elucidate the structure and chemical properties of different molecules. The complete process involves the conversion of the sample into gaseous ions, with or without fragmentation, which are then characterized by their mass to charge ratios (m/z) and relative abundances.

This technique basically studies the effect of ionizing energy on molecules. It depends upon chemical reactions in the gas phase in which sample molecules are consumed during the formation of ionic and neutral species.

Basic Principle

A mass spectrometer generates multiple ions from the sample under investigation, it then separates them according to their specific mass-to-charge ratio (m/z), and then records the relative abundance of each ion type.

The first step in the mass spectrometric analysis of compounds is the production of gas phase ions of the compound, basically by electron ionization. This molecular ion undergoes fragmentation. Each primary product ion derived from the molecular ion, in turn, undergoes fragmentation, and so on. The ions are separated in the mass spectrometer according to their mass-to-charge ratio, and are detected in proportion to their abundance. A mass spectrum of the molecule is thus produced. It displays the result in the form of a plot of ion abundance versus mass-to-charge ratio. Ions provide information concerning the nature and the structure of their precursor molecule. In the spectrum of a pure compound, the molecular ion, if present, appears at the highest value of m/z (followed by ions containing heavier isotopes) and gives the molecular mass of the compound.

Components

The instrument consists of three major components:

1. Ion Source: For producing gaseous ions from the substance being studied.
2. Analyzer: For resolving the ions into their characteristics mass components according to their mass-to-charge ratio.
3. Detector System: For detecting the ions and recording the relative abundance of each of the resolved ionic species.

In addition, a sample introduction system is necessary to admit the samples to be studied to the ion source while maintaining the high vacuum requirements (~10-6 to 10-8 mm of mercury) of the technique; and a computer is required to control the instrument, acquire and manipulate data, and compare spectra to reference libraries.

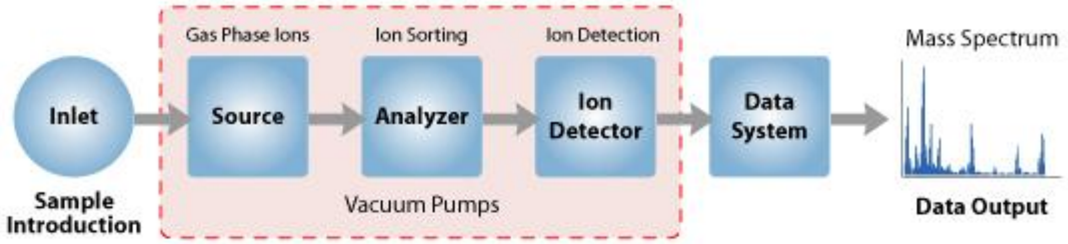


Figure: Components of a Mass Spectrometer

With all the above components, a mass spectrometer should always perform the following processes:

1. Produce ions from the sample in the ionization source.
2. Separate these ions according to their mass-to-charge ratio in the mass analyzer.
3. Eventually, fragment the selected ions and analyze the fragments in a second analyzer.
4. Detect the ions emerging from the last analyzer and measure their abundance with the detector that converts the ions into electrical signals.
5. Process the signals from the detector that are transmitted to the computer and control the instrument using feedback.

Polymerase Chain Reaction

- Polymerase chain reaction (PCR) is a common laboratory technique used to make many copies (millions or billions!) of a particular region of DNA. This DNA region can be anything the experimenter is interested in. For example, it might be a gene whose function a researcher wants to understand, or a genetic marker used by forensic scientists to match crime scene DNA with suspects.
- Typically, the goal of PCR is to make enough of the target DNA region that it can be analyzed or used in some other way. For instance, DNA amplified by

PCR may be sent for sequencing, visualized by gel electrophoresis, or cloned into a plasmid for further experiments.

- PCR is used in many areas of biology and medicine, including molecular biology research, medical diagnostics, and even some branches of ecology.

The key ingredients of a PCR reaction are *Taq* polymerase, primers, template DNA, and nucleotides (DNA building blocks). The ingredients are assembled in a tube, along with cofactors needed by the enzyme, and are put through repeated cycles of heating and cooling that allow DNA to be synthesized.

The basic steps are:

1. Denaturation (96°C): Heat the reaction strongly to separate, or denature, the DNA strands. This provides single-stranded template for the next step.
2. Annealing (55 - 65°C): Cool the reaction so the primers can bind to their complementary sequences on the single-stranded template DNA.
3. Extension (72°C): Raise the reaction temperatures so *Taq* polymerase extends the primers, synthesizing new strands of DNA.

Review Questions

Short Answer Questions (2 Marks)

1. Define amino acid.
2. List out the type of amino acids.
3. Define peptide bond.
4. Define isoelectric point.
5. Explain the structure of amino acid.
6. Define protein.
7. Define primary structure of proteins.
8. Define primary secondary of proteins.

9. Define tertiary structure of proteins.
10. Define quarternary structure of proteins.
11. What is nucleic acid?
12. What is DNA?
13. What are the components of DNA?
14. Define nucleotide with structure.
15. What are nitrogenous bases?
16. What is phospodiester bond?
17. Define Watson and Crick base pairing rule.
18. What is the function of DNA?
19. Define B form of DNA.
20. Explain the types of RNA.
21. Define nucleic acid bloting.
22. List out the types of microarrays.
23. Give the principles of mass spectrometry.

Essay Answer Questions **(6 & 8 Marks)**

1. Describe about amino acids and proteins.
2. Describe in detail about protein structure.
3. Differentiate the types of DNA.
4. Describe the structure of DNA with diagram.
5. Explain about RNA and its types.
6. Give a detailed note on chromatogram.
7. Explain in detail about restriction digestion.
8. Discuss in detail about PCR.

Further Readings

http://en.wikipedia.org/wiki/Computer

http://en.wikipedia.org/wiki/Operating_system

http://en.wikipedia.org/wiki/Computer_network

http://en.wikipedia.org/wiki/Web_browser

http://en.wikipedia.org/wiki/Web_search_engine

http://en.wikipedia.org/wiki/Email

http://en.wikipedia.org/wiki/Database

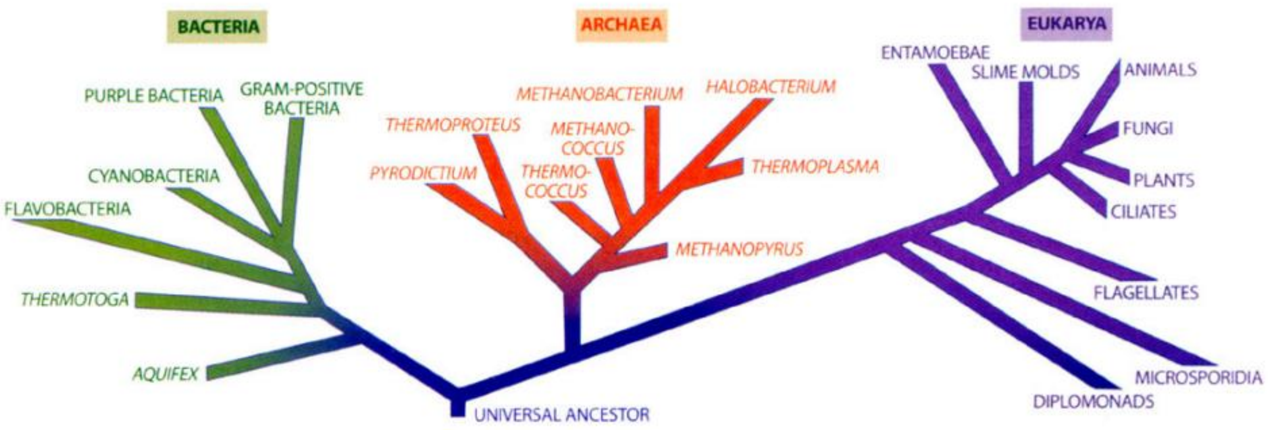
Unit III

SYLLABUS

Sequence and Phylogeny analysis, Detecting Open Reading Frames, Outline of sequence Assembly, Mutation/Substitution Matrices, Pairwise Alignments, Introduction to BLAST, using it on the web, Interpreting results, Multiple Sequence Alignment, Phylogenetic Analysis. Introduction to BLAST, using it on the web, Interpreting results, Multiple Sequence Alignment, Phylogenetic Analysis.

Sequence analysis

In bioinformatics, sequence analysis is the process of subjecting a DNA, RNA or peptide sequence to any of a wide range of analytical methods to understand its features, function, structure, or evolution. Methodologies used include sequence alignment, searches against biological databases, and others. Since the development of methods of high-throughput production of gene and protein sequences, the rate of addition of new sequences to the databases increased exponentially. Such a collection of sequences does not, by itself, increase the scientist's understanding of the biology of organisms. However, comparing these new sequences to those with known functions is a key way of understanding the biology of an organism from which the new sequence comes. Thus, sequence analysis can be used to assign function to genes and proteins by the study of the similarities between the compared sequences. Nowadays, there are many tools and techniques that provide the sequence comparisons (sequence alignment) and analyze the alignment product to understand its biology.



Sequence analysis in molecular biology includes a very wide range of relevant topics:

1. The comparison of sequences in order to find similarity, often to infer if they are related (homologous)
2. Identification of intrinsic features of the sequence such as active sites, post translational modification sites, gene-structures, reading frames, distributions of introns and exons and regulatory elements
3. Identification of sequence differences and variations such as point mutations and single nucleotide polymorphism (SNP) in order to get the genetic marker.
4. Revealing the evolution and genetic diversity of sequences and organisms
5. Identification of molecular structure from sequence alone

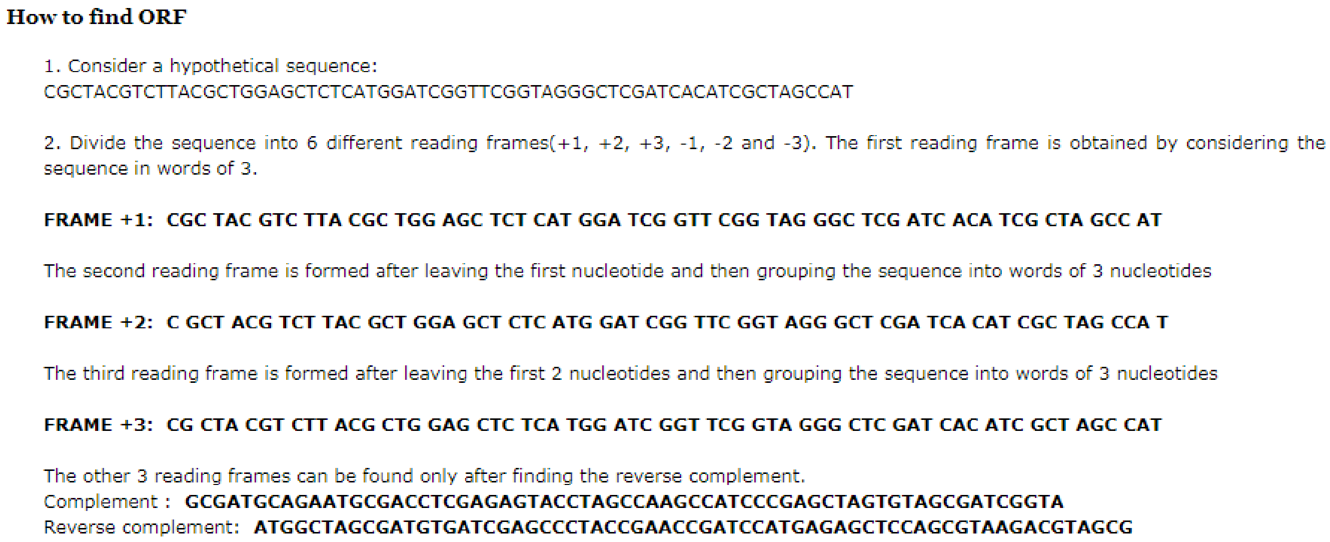
Detection of Open Reading Frame (ORF)

- In molecular genetics, an open reading frame (ORF) is the part of a reading frame that has the ability to be translated. An ORF is a continuous stretch of codons that contain a start codon (usually AUG) and a stop codon (usually UAA, UAG or UGA).

- DNA (Deoxyribonucleic acid) is the genetic material that contains all the genetic information in a living organisms. The information is stored as genetic codes using adenine (A), guanine (G), cytosine(C) and thymine (T). During the transcription process, DNA is transcribed to mRNA. Each of these base pairs will bond with a sugar and phosphate molecule to form a nucleotide.
- Three nucleotides that codes for a particular amino acid during translation is called as a codon. The region of a nucleotide that starts from an initiation codon and ends with a stop codon is called an **Open Reading Frame (ORF)**. Proteins are formed from ORF. By analyzing the ORF we can predict the possible amino acids that might be produced during translation.
- The ORF finder is a program available at NCBI website. It identifies all ORF or possible protein coding region from six different reading frame.
- DNA (Deoxyribonucleic acid) is the genetic material that contains the genetic information for development and helps in maintaining all the functions in a living organisms.The information is stored as genetic codes using four different bases.
- They are adenine (A), guanine (G), cytosine(C) and thymine (T). In two strands of DNA, adenine always pair with thymine and guanine pair with cytosine. Each of these base pairs

will bond with a sugar and phosphate molecule to form a nucleotide. The base pairing of DNA will result in a ladder shape structure of these strands which is called a double helix. RNA is differs from DNA only in 1 base pair i.e. in RNA it is uracil (U) instead of thymine(T).

- mRNA (messenger RNA) is a type of RNA which is formed from DNA transcription. During the transcription process, DNA is transcribed to mRNA in the nucleus and moves to the cytoplasm through the nuclear pores. This mRNA is translated to protein in the cytoplasm with the help of ribosomes.
- In mRNA, 3 nucleotides are considered at a time since a set of 3 nucleaotides (refered to as codon) codes for an amino acid. The region of a nucleotide that starts from an initiation codon and ends with a stop codon is called an Open Reading Frame (ORF).
- An initiation codon is the triplet codon that codes for the first amino acid in the translation process. The translation process will start only with the initiation codon, ATG which codes for the amino acid methionine. The translation process stops when it comes across a stop codon. There are three stop codons: TAA ("ochre"), TAG ("amber") and TGA ("opal" or "umber"). Any of these codons can stop the translation. Genetic codon can form 64 triplets(4^3) from the 4 nucleotides that codes for amino acids. Protein is formed from the ORF.



How to find ORF

1. Consider a hypothetical sequence:
CGCTACGTCTTACGCTGGAGCTCTCATGGATCGGTTCGGTAGGGCTCGATCACATCGCTAGCCAT

2. Divide the sequence into 6 different reading frames(+1, +2, +3, -1, -2 and -3). The first reading frame is obtained by considering the sequence in words of 3.

FRAME +1: CGC TAC GTC TTA CGC TGG AGC TCT CAT GGA TCG GTT CGG TAG GGC TCG ATC ACA TCG CTA GCC AT

The second reading frame is formed after leaving the first nucleotide and then grouping the sequence into words of 3 nucleotides

FRAME +2: C GCT ACG TCT TAC GCT GGA GCT CTC ATG GAT CGG TTC GGT AGG GCT CGA TCA CAT CGC TAG CCA T

The third reading frame is formed after leaving the first 2 nucleotides and then grouping the sequence into words of 3 nucleotides

FRAME +3: CG CTA CGT CTT ACG CTG GAG CTC TCA TGG ATC GGT TCG GTA GGG CTC GAT CAC ATC GCT AGC CAT

The other 3 reading frames can be found only after finding the reverse complement.
Complement : **GCGATGCAGAATGCGACCTCGAGAGTACCTAGCCAAGCCATCCCGAGCTAGTGTAGCGATCGGTA**
Reverse complement: **ATGGCTAGCGATGTGATCGAGCCCTACCGAACCGATCCATGAGAGCTCCAGCGTAAGACGTAGCG**

Now same process as that of +1, +2 and +3 strands is repeated for -1, -2 and -3 strands with reverse complement sequence

FRAME -1: ATG GCT AGC GAT GTG ATC GAG CCC TAC CGA ACC GAT CCA TGA GAG CTC CAG CGT AAG ACG TAG CG
FRAME -2: A TGG CTA GCG ATG TGA TCG AGC CCT ACC GAA CCG ATC CAT GAG AGC TCC AGC GTA AGA CGT AGC G
FRAME -3: AT GGC TAG CGA TGT GAT CGA GCC CTA CCG AAC CGA TCC ATG AGA GCT CCA GCG TAA GAC GTA GCG

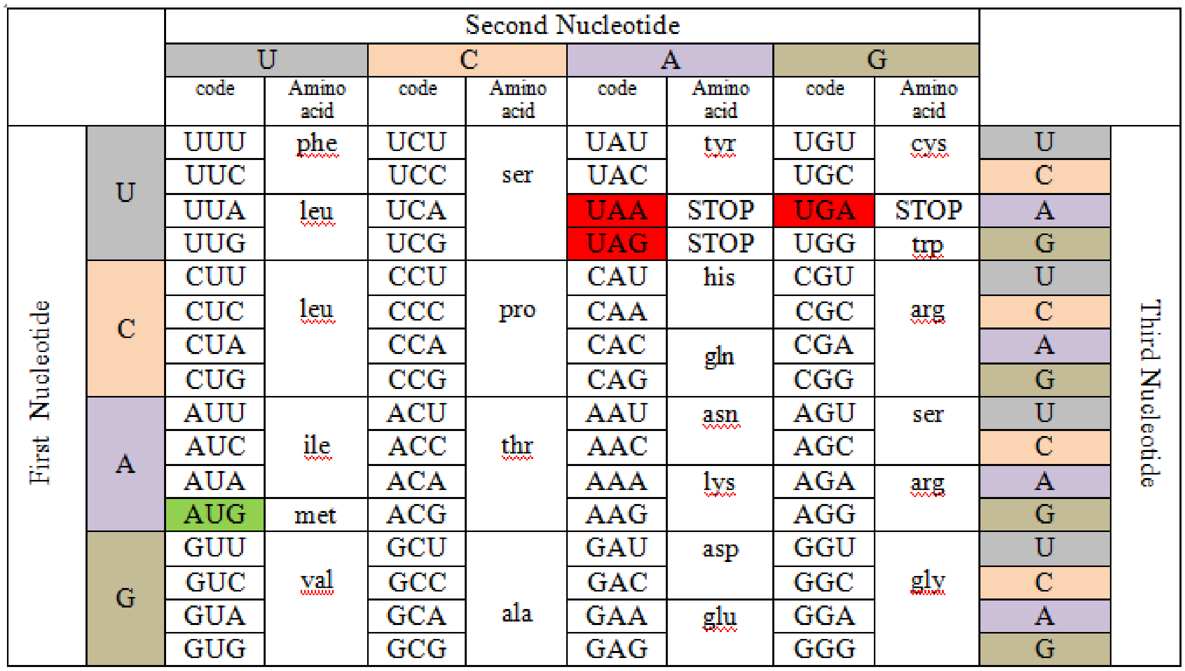
3. Now mark the start codon and stop codons in the reading frames

FRAME +1: CGC TAC GTC TTA CGC TGG AGC TCT CAT GGA TCG GTT CGG TAG GGC TCG ATC ACA TCG CTA GCC AT
FRAME +2: C GCT ACG TCT TAC GCT GGA GCT CTC ATG GAT CGG TTC GGT AGG GCT CGA TCA CAT CGC TAG CCA T
FRAME +3: CG CTA CGT CTT ACG CTG GAG CTC TCA TGG ATC GGT TCG GTA GGG CTC GAT CAC ATC GCT AGC CAT
FRAME -1: ATG GCT AGC GAT GTG ATC GAG CCC TAC CGA ACC GAT CCA TGA GAG CTC CAG CGT AAG ACG TAG CG
FRAME -2: A TGG CTA GCG ATG TGA TCG AGC CCT ACC GAA CCG ATC CAT GAG AGC TCC AGC GTA AGA CGT AGC G
FRAME -3: AT GGC TAG CGA TGT GAT CGA GCC CTA CCG AAC CGA TCC ATG AGA GCT CCA GCG TAA GAC GTA GCG

4. Identify the open reading frame (ORF) - sequence stretch begining with a start codon and ending in a stop codon.

FRAME +2: ATG GAT CGG TTC GGT AGG GCT CGA TCA CAT CGC TAG
FRAME -1: ATG GCT AGC GAT GTG ATC GAG CCC TAC CGA ACC GAT CCA TGA
FRAME -3: ATG AGA GCT CCA GCG TAA

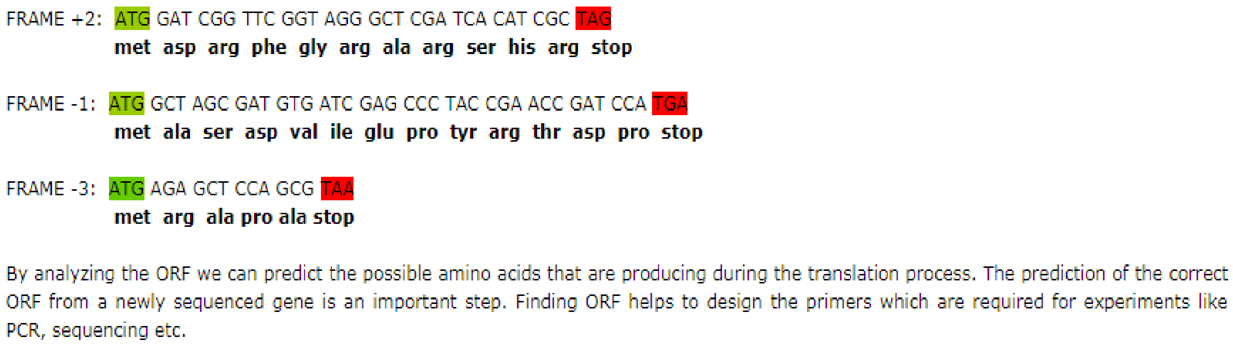
5. Based on the amino acid table the peptide sequence is found



		Second Nucleotide									
		U		C		A		G			
		code	Amino acid	code	Amino acid	code	Amino acid	code	Amino acid		
First Nucleotide	U	UUU	phe	UCU	ser	UAU	tyr	UGU	cys	U	Third Nucleotide
		UUC		UCC		UAC		UGC		C	
		UUA	leu	UCA		UAA	STOP	UGA	STOP	A	
		UUG		UCG		UAG	STOP	UGG	trp	G	
	C	CUU	leu	CCU	pro	CAU	his	CGU	arg	U	
		CUC		CCC		CAA		CGC		C	
		CUA		CCA		CAC	gln	CGA		A	
		CUG		CCG		CAG		CGG		G	
	A	AUU	ile	ACU	thr	AAU	asn	AGU	ser	U	
		AUC		ACC		AAC		AGC		C	
		AUA		ACA		AAA	lys	AGA	arg	A	
		AUG	met	ACG		AAG		AGG		G	
	G	GUU	val	GCU	ala	GAU	asp	GGU	gly	U	
		GUC		GCC		GAC		GGC		C	
		GUA		GCA		GAA	glu	GGA		A	
		GUG		GCG		GAG		GGG		G	

Figure 1: Amino Acid Table

Sequence Assembly



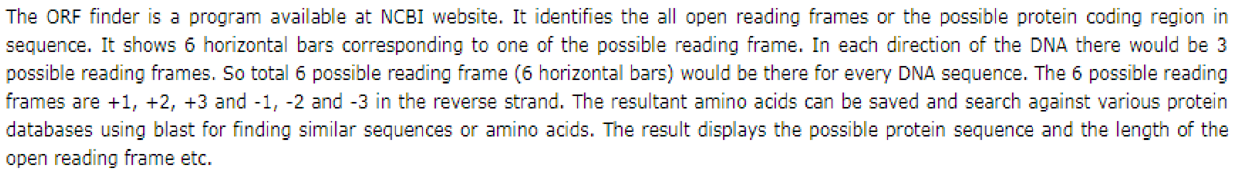
```
FRAME +2:  ATG GAT CGG TTC GGT AGG GCT CGA TCA CAT CGC TAG
           met asp arg phe gly arg ala arg ser his arg stop

FRAME -1:  ATG GCT AGC GAT GTG ATC GAG CCC TAC CGA ACC GAT CCA TGA
           met ala ser asp val ile glu pro tyr arg thr asp pro stop

FRAME -3:  ATG AGA GCT CCA GCG TAA
           met arg ala pro ala stop
```

By analyzing the ORF we can predict the possible amino acids that are producing during the translation process. The prediction of the correct ORF from a newly sequenced gene is an important step. Finding ORF helps to design the primers which are required for experiments like PCR, sequencing etc.

ORF Finder:



The ORF finder is a program available at NCBI website. It identifies the all open reading frames or the possible protein coding region in sequence. It shows 6 horizontal bars corresponding to one of the possible reading frame. In each direction of the DNA there would be 3 possible reading frames. So total 6 possible reading frame (6 horizontal bars) would be there for every DNA sequence. The 6 possible reading frames are +1, +2, +3 and -1, -2 and -3 in the reverse strand. The resultant amino acids can be saved and search against various protein databases using blast for finding similar sequences or amino acids. The result displays the possible protein sequence and the length of the open reading frame etc.

- Sequence assembly refers to aligning and merging fragments from a longer DNA sequence in order to reconstruct the original sequence. This is needed as DNA sequencing technology cannot read whole genomes in one go, but rather reads small pieces of between 20 and 30000 bases, depending on the technology used. Typically the short fragments, called reads, result from shotgun sequencing genomic DNA, or gene transcript (ESTs).

- The problem of sequence assembly can be compared to taking many copies of a book, passing each of them through a shredder with a different cutter, and piecing the text of the book back together just by looking at the shredded pieces.

- Besides the obvious difficulty of this task, there are some extra practical issues: the original may have many repeated paragraphs, and some shreds may be modified during shredding to have typos. Excerpts from another book may also be added in, and some shreds may be completely unrecognizable.

Mutation/Substitution Matrices

(http://steipe.biochemistry.utoronto.ca/abc/index.php/Mutation_Data_Matrices)

- In bioinformatics and evolutionary biology, a substitution matrix describes the rate at which one character in a sequence changes to other character states over time.
- Substitution matrices are usually seen in the context of amino acid or DNA sequence alignments, where the similarity between sequences depends on their divergence time and the substitution rates as represented in the matrix.

Phylogenetic analysis

- Similarities and divergence among related biological sequences revealed by sequence alignment often have to be rationalized and visualized as phylogenetic trees.

- Thus, molecular phylogenetics is a fundamental aspect of bioinformatics.

MOLECULAR EVOLUTION AND MOLECULAR PHYLOGENETICS

“What is evolution?”

Evolution can be defined in various ways under different contexts.

In the biological context,

Evolution can be defined as the development of a biological form from other preexisting forms or its origin to the current existing form through natural selections and modifications.

The driving force behind evolution is natural selection in which “unfit” forms are eliminated through changes of environmental conditions or sexual selection so that only the fittest are selected.

The underlying mechanism of evolution is genetic mutations that occur spontaneously.

The mutations on the genetic material provide the biological diversity within a population; hence, the variability of individuals within the population to survive successfully in a given environment.

Genetic diversity thus provides the source of raw material for the natural selection to act on.

<u>Phylogenetics</u>

is the study of the evolutionary history of living organisms using tree like diagrams to represent pedigrees of these organisms.

The tree branching patterns representing the evolutionary divergence are referred to as *phylogeny*.

Phylogenetics can be studied in various ways.

- studied using **fossil records**, which contain morphological information about ancestors of current species and the timeline of divergence.

- studied using **molecular data** that are in the form of DNA or protein sequences can also provide very useful evolutionary perspectives of existing organisms because, as organisms evolve, the genetic materials accumulate mutations over time causing phenotypic changes.

Because genes are the medium for recording the accumulated mutations, they can serve as ***molecular fossils***. Through comparative analysis of the molecular fossils from a number of related organisms, the evolutionary history of the genes and even the organisms can be revealed.

Advantage of using molecular data:

- Molecular data are more numerous than fossil records and easier to obtain.
- There is no sampling bias involved, whichhelps to mend the gaps in real fossil records.
- More clear-cut and robust phylogenetic trees can be constructed with the molecular data.
- Therefore,they have become favorite and sometimes the only information available for researchers to reconstruct evolutionary history.

The advent of the genomic era with tremendous amounts of molecular sequence data has led to the rapid development of molecular phylogenetics.

Molecular phylogenetics can be defined as the study of evolutionary relationships of genes and other biological macromolecules by analyzing mutations at various positions in their sequences and developing hypotheses about the evolutionary relatedness of the biomolecules.

Based on the sequence similarity of the molecules, evolutionary relationships between the organisms can be inferred.

Major Assumptions:

To use molecular data to reconstruct evolutionary history requires number of reasonable assumptions.

[1] molecular sequences used in phylogenetic construction are homologous, meaning that they share a common origin and subsequently diverged through time.

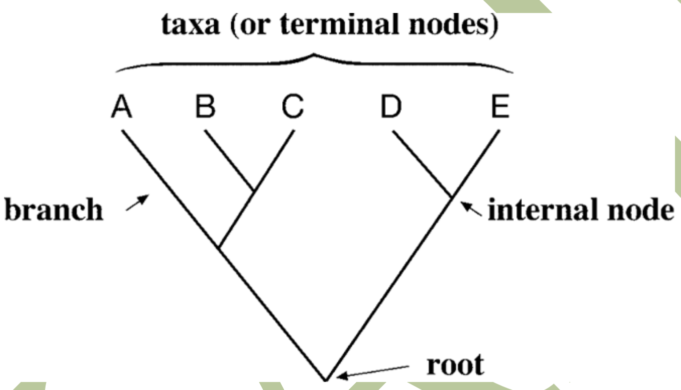
Phylogenetic divergence is assumed to be bifurcating, meaning that a parent branch splits into two daughter branches at any given point.

[2] each position in a sequence evolved independently.

[3] The variability among sequences is sufficiently informative for constructing unambiguous phylogenetic trees.

TERMINOLOGY USED IN PHYLOGENETIC TREE

A typical bifurcating phylogenetic tree is a graph shown in Figure.



- ✓ The lines in the tree are called *branches.*
- ✓ At the tips of the branches are present-day species or sequences known as ***taxa*** (the singular form is *taxon*) or operational taxonomic units.
- ✓ The connecting point where two adjacent branches join is called a ***node***, which represents an inferred ancestor of extant taxa.
- ✓ The bifurcating point at the very bottom of the tree is the ***root node***, which represents the common ancestor of all members of the tree.
- ✓ A group of taxa descended from a single common ancestor is defined as a ***clade*** or ***monophyletic*** *group*.

In a monophyletic group,

- ✓ two taxa share a unique common ancestor not shared by any other taxa.
- ✓ They are also referred to as ***sister taxa*** to each other (e.g., taxa B and C).
- ✓ The branch path depicting an ancestor–descendant relationship on a tree is called a ***lineage***, which is often synonymous with a tree branch leading to a defined monophyletic group.
- ✓ When a number of taxa share more than one closest common ancestors, they do not fit the definition of a clade. In this case, they are referred to as ***paraphyletic*** (e.g., taxa B, C, and D).
- ✓ The branching pattern in a tree is called ***tree topology***.
- ✓ When all branches bifurcate on a phylogenetic tree, it is referred to as ***dichotomy***.
- ✓ In this case, each ancestor divides and gives rise to two descendants.
- ✓ Sometimes, a branch point on a phylogenetic tree may have more than two descendents, resulting in a ***multifurcating node***.
- ✓ The phylogeny with multifurcating branches is called ***polytomy*** (as shown in the following Figure).

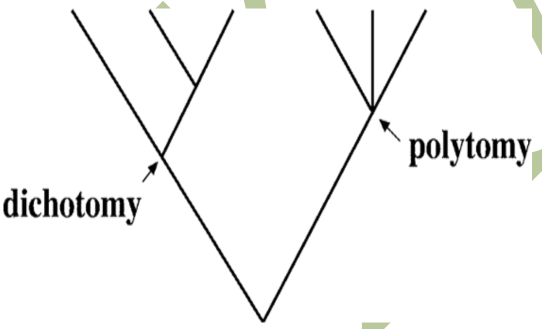


Figure showing an example of bifurcation and multifurcation.

Multifurcation is normally a result of insufficient evidence to fully resolve the tree or a result of an evolutionary process known as

- ✓ A polytomy can be a result of either an ancestral taxon giving rise to more than two immediate descendants simultaneously during evolution, a process known as ***radiation,*** or an unresolved phylogeny in which the exact order of bifurcations cannot be determined precisely.

- ✓ A phylogenetic tree can be either rooted or unrooted (as shown the following Figure).

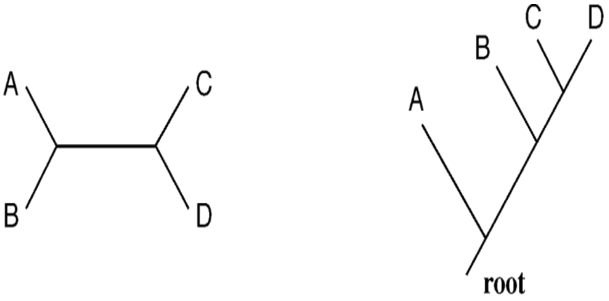


Figure shows rooted versus unrooted trees.

A phylogenetic tree without definition of a root is unrooted (*left*).

- ✓ An ***unrooted phylogenetic tree*** does not assume knowledge of a common ancestor, but only positions the taxa to show their relative relationships.

Because there is no indication of which node represents an ancestor, there is no direction of an evolutionary path in an unrooted tree.

- ✓ To define the direction of an evolution path, a tree must be rooted.
- ✓ In a ***rooted tree***, all the sequences under study have a common ancestor or root node from which a unique evolutionary path leads to all other nodes.
- ✓ Rooted tree is more informative than an unrooted one.
- ✓ To convert an unrooted tree to a rooted tree,one needs to first determine where the root is.
- ✓ Strictly speaking, the root of the tree is not known; the common ancestor is already extinct.

There are two ways to define the root of a tree.

1. ***outgroup approach***, - which is a sequence that is homologous to the sequences under consideration, but separated from those sequences at an early evolutionary time.

- Outgroups are generally determined from independent sources of information.

For example, a bird sequence can be used as a root for the phylogenetic analysis of mammals based on multiple evidence that indicate that birds branched off prior to all mammalian taxa in the ingroup.

- Outgroups are required to be distinct from the in group sequences, but not too distant from the in group.

- Using too divergent sequences as an outgroup can lead to errors in tree construction.

2. ***midpoint rooting approach***

In the absence of a good outgroup, a tree can be rooted using the *midpoint rooting approach*, in which the mid point of the two most divergent groups judged by overall branch lengths is assigned as the root.

This type of rooting assumes that divergence from root to tips for both branches is equal and follows the "molecular clock" hypothesis.

Molecular clock is an assumption by which molecular sequences evolve at constant rates so that the amount of accumulated mutations is proportional to evolutionary time.

Based on this hypothesis, branch lengths on a tree can be used to estimate divergence time.

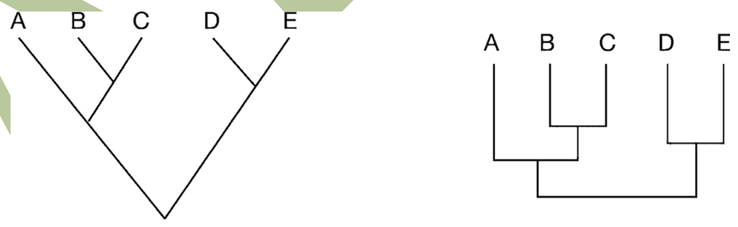
FORMS OF TREE REPRESENTATION

The topology of branches in a tree defines the relationships between the taxa.

The trees can be drawn in different ways as shown the following figure

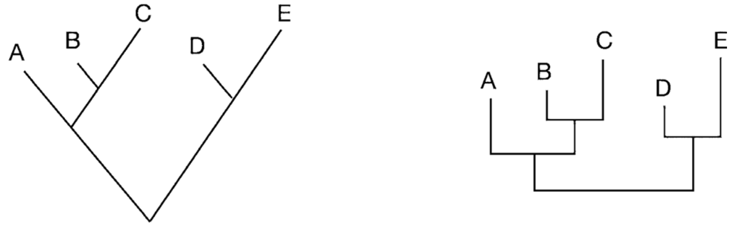
1. cladogram

2. phylogram



The branch lengths are unscaled in the cladograms

Cladogram



The branch lengths scaled in the phylograms.

Phylogram

In each of these tree representations, the branches of a tree can freely rotate without changing the relationships among the taxa.

In a *phylogram*,

- ✓ the branch lengths represent the amount of evolutionary divergence. Such trees are said to be scaled.
- ✓ The scaled trees have the advantage of showing both the evolutionary relationships and information about the relative divergence time of the branches.

In a *cladogram*,

- ✓ the external taxa line up neatly in a row or column.
- ✓ Their branch lengths are not proportional to the number of evolutionary changes and thus have no phylogenetic meaning.
- ✓ In these unscaled tree, only the topology of the tree matters, which shows the relative ordering of the taxa.

Procedure of phylogenetic tree:

Molecular phylogenetic tree construction can be divided into five steps:

(1) choosing molecular markers;

(2) performing multiple sequence alignment;

(3) choosing a model of evolution;

(4) determining a tree building method;

(5) assessing tree reliability.

(1) Choice of Molecular Markers

- For constructing molecular phylogenetic trees, one can use either nucleotide or protein sequence data.
- The choice of molecular markers is an important factor because it can make a major difference in obtaining a correct tree.

- The decision to use nucleotide or protein sequences depends on the properties of the sequences and the purposes of the study.
- For studying very closely related organisms, nucleotide sequences, which evolve more rapidly than proteins, can be used.

For example,

- ✓ For evolutionary analysis of different individuals within a population, non coding regions of mitochondrial DNA are often used.
- ✓ For studying the evolution of more widely divergent groups of organisms, choose either slowly evolving nucleotide sequences, such as ribosomal RNA or protein sequences.
    - ▪ If the phylogenetic relationships to be delineated are at the deepest level, such as between bacteria and eukaryotes, using conserved protein sequences makes more sense than using nucleotide sequences.

(2) Alignment

- ▪ The second step in phylogenetic analysis is to construct sequence alignment.
- ▪ This is the most critical step in the procedure because it establishes positional correspondence in evolution.
- ▪ Only the correct alignment produces correct phylogenetic inference because aligned positions are assumed to be genealogically related.
- ▪ Incorrect alignment leads to systematic errors in the final tree or even a completely wrong tree.
- ▪ For that reason, it is essential that the sequences are correctly aligned.
- ▪ Multiple state-of-the-art alignment programs such as T-Coffee should be used.
- ▪ The alignment results from multiple sources should be inspected and compared carefully to identify the most reasonable one.
- ▪ Automatic sequence alignments almost always contain errors and should be further edited or refined if necessary.

- Manual editing is often critical in ensuring alignment quality.
- It is also often necessary to decide whether to use the full alignment or to extract parts of it. Truly ambiguously aligned regions have to be removed from consideration prior to phylogenetic analysis.
- Which part of the alignment to remove is often at the discretion of the researcher. It is a rather subjective process.
- In extreme cases, someresearchers like to remove all insertions and deletions (indels) and only use positions that are shared by all sequences in the dataset.
- The clear drawback of this practice is that many phylogenetic signals are lost.
- In fact, gap regions often belong to *signature indels* unique to identification of a subgroup of sequences and should to be retained for treeing purposes.
- In addition, there is an automatic approach in improving alignment quality.

Rascal and NorMD - can help to improve alignment by correcting alignment errors and removing potentially unrelated or highly divergent sequences.

Gblocks (http://woody.embl-heidelberg.de/phylo/) - can help to detect and eliminate the poorly aligned positions and divergent regions so to make the alignment more suitable for phylogenetic analysis.

Multiple Substitutions

- A simple measure of the divergence between two sequences is to count the number of substitutions in an alignment.
- The proportion of substitutions defines the observed distance between the two sequences.
- However, the observed number of substitutions may not represent the true evolutionary events that actually occurred.
- When a mutation is observed as A replaced by C, the nucleotide may have actually undergone a number of intermediate steps to become C, such as A→T→G→C.

- Similarly, a back mutation could have occurred when a mutated nucleotide reverted back to the original nucleotide. This means that when the same nucleotide is observed, mutations like G→C→G may have actually occurred.
- Moreover, an identical nucleotide observed in the alignment could be due to parallel mutations when both sequences mutate into T, for instance.
- Such multiple substitutions and convergence at individual positions obscure the estimation of the true evolutionary distances between sequences. This effect is known as ***homoplasy***, which, if not corrected, can lead to the generation of incorrect trees.
- To correct homoplasy, statistical models are needed to infer the true evolutionary distances between sequences.

(3) Choosing Substitution Models

The statistical models used to correct homoplasy are called ***substitution models*** **or** ***evolutionary models.***

For constructing DNA phylogenies, there are a number of nucleotide substitution models available.
- These models differ in how multiple substitutions of each nucleotide are treated.

1. Jukes–Cantor Model

The simplest nucleotide substitution model is the Jukes–Cantor model, which assumes that all nucleotides are substituted with equal probability.

A formula for deriving evolutionary distances that include hidden changes is introduced by using alogarithmic function.

$d\mathrm{AB} = -(3/4)\ \ln[1 - (4/3)p\mathrm{AB}]$

where dAB is the evolutionary distance between sequences A and B and

pAB is the observed sequence distance measured by the proportion of substitutions over the entire length of the alignment.

2. KimuraModel

Another model to correct evolutionary distances is called the Kimura two-parameter model. This is a more sophisticated model in which mutation rates for transitions and transversion are assumed to be different, which is more realistic.

According tothis model, transitions occur more frequently than transversions, which, therefore,provides a more realistic estimate of evolutionary distances.

The Kimura model uses the following formula:

$d\text{AB} = -(1/2)\ln(1 - 2p\text{ti} - p\text{tv}) - (1/4)\ln(1 - 2p\text{tv})$

where dAB is the evolutionary distance between sequences A and B,

$p$ti is the observed frequency for transition,

$p$tv the frequency of transversion.

(4) Phylogenetic Tree Construction Methodsand Programs

There are currently two main categories of tree-building methods, each having advantages and limitations.

1. Distance based method

is based on distance, which is the amount of dissimilarity between pairs of sequences, computed on the basis of sequence alignment.

The distance-based methods assume that all sequences involved are homologous and that tree branches are additive, meaning that the distance between two taxa equals the sum of all branch lengths connecting them.

2. Character based method

is based on discrete characters, which are molecular sequences from individual taxa. The basic assumption is that characters at corresponding positions in a multiple sequence alignment are homologous among the sequences involved. Therefore, the character states of the common ancestor can be traced from this dataset. Another assumption is that each character evolves independently and is therefore treated as an individual evolutionary unit.

DISTANCE-BASED METHODS

- True evolutionary distances between sequences can be calculated from observed distances after correction using a variety of evolutionary models.
- The computed evolutionary distances can be used to construct a matrix of distances between all individual pairs of taxa.
- Based on the pairwise distance scores in the matrix, a phylogenetic tree can be constructed for all the taxa involved.
- The algorithms for the distance-based tree-building method can be subdivided into either clustering based or
    - optimality based.

Clustering-type algorithms

- compute a tree based on a distance matrix starting from the most similar sequence pairs.

- These algorithms includes

(i) Unweighted pair group method using arithmetic average (UPGMA) algorithm

(ii) neighbor joining algorithm.

Optimality-based algorithms

- compare many alternative tree topologies and select one that has the best fit between estimated distances in the tree and the actual evolutionary distances.

- This category includes

(i) Fitch–Margoliash algorithms

(ii) minimum evolution algorithms.

Clustering-Based Methods

(i) Unweighted Pair Group Method Using Arithmetic Average (UPGMA) algorithm

✓ The simplest clustering method is UPGMA, which builds a tree by a sequential clustering method.

- ✓ Given a distance matrix, it starts by grouping two taxa with the smallest pairwise distance in the distance matrix.
- ✓ A node is placed at the midpoint or half distance between them.
- ✓ It then creates a reduced matrix by treating the new cluster as a single taxon.
- ✓ The distances between this new composite taxon and all remaining taxa are calculated to create a reduced matrix.
- ✓ The same grouping process is repeated and another newly reduced matrix is created.
- ✓ The iteration continues until all taxa are placed on the tree.
- ✓ The last taxon added is considered the outgroup producing a rooted tree.
- ✓ The basic assumption of the UPGMA method is that all taxa evolve at a constant rate and that they are equally distant from the root, implying that a molecular clock is in effect.
- ✓ However, real data rarely meet this assumption. Thus, UPGMA often produces erroneous tree topologies.
- ✓ However, owing to its fast speed of calculation, it has found extensive usage in clustering analysis of DNA microarray data.

(ii) Neighbor Joining algorithm

- ✓ The UPGMA method uses unweighted distances and assumes that all taxa have constant evolutionary rates.
- ✓ Since this molecular clock assumption is often not met in biological sequences, to build a more accurate phylogenetic trees, the neighbor joining (NJ) method can be used, which is somewhat similar to UPGMA.
- ✓ It builds a tree by using stepwise reduced distance matrices.
- ✓ the NJ method does not assume the taxa to be equidistant from the root.
- ✓ It corrects for unequalevolutionary rates between sequences by using a conversion step.
- ✓ This conversion requires the calculations of "*r*-values" and "transformed *r*-values" using the following formula:

$d\_\mathrm{AB} = d\mathrm{AB} - 1/2 \times (r\mathrm{A} + r\mathrm{B})$

where *d*_AB is the converted distance between A and B and

*d*AB is the actual evolutionary distance between A and B.

The value of *r*A (or *r*B) is the sum of distances of A (or B) to all other taxa.

Optimality-Based Methods

optimality-based methods have a well-defined algorithm to compare all possible tree topologies and select a tree that best fits the actual evolutionary distance matrix.

Based on the differences in optimality criteria, there are two types of algorithms,

(i) Fitch–Margoliash algorithms

(ii) minimum evolution algorithms

(i) Fitch–Margoliash algorithms

- ✓ The Fitch–Margoliash (FM) method selects a best tree among all possible trees based on minimal deviation between the distances calculated in the overall branches in the tree and the distances in the original dataset.
- ✓ It starts by randomly clustering two taxa in a node and creating three equations to describe the distances, and then solving the three algebraic equations for unknown branch lengths.
- ✓ The clustering of the two taxa helps to create a newly reduced matrix.
- ✓ This process is iterated until a tree is completely resolved.
- ✓ The method searches for all tree topologies and selects the one that has the lowest squared deviation of actual distances and calculated tree branch lengths.

(ii) Minimum Evolution algorithm

Minimum evolution (ME) constructs a tree with a similar procedure, but uses a different optimality criterion that finds a tree among all possible trees with a minimum overall branch length.

Searching for the minimum total branch length isan indirect approach to achieving the best fit of the branch lengths with the original dataset.

Pros and Cons

The most frequently used distance methods are clustering based.

Major advantage is that

- they are computationally fast and are therefore capable of handling datasets that are deemed to be too large for any other phylogenetic method.
- The overall advantage of all distance-based methods is the ability to make use of a large number of substitution models to correct distances.

Drawback is that

- The actual sequence information is lost when all the sequence variation is reduced to a single value.
- ancestral sequences at internal nodes cannot be inferred.

CHARACTER-BASED METHODS

Character-based methods (also called *discrete methods*) are

- based directly on the sequence characters rather than on pairwise distances.

- They count mutational events accumulated on the sequences and may therefore avoid the loss of information when characters are converted to distances.

- This preservation of character information means that evolutionary dynamics of each character can be studied.

- Ancestral sequences can also be inferred.

- The two most popular character-based approaches are

(i) maximum parsimony (MP) mehod

(ii) maximum likelihood (ML) method.

(i) Maximum Parsimony method

- ✓ The parsimony method chooses a tree that has the fewest evolutionary changes or shortest overall branch lengths.
- ✓ It is based on a principle related to a medieval philosophy called *Occam's razor*.
- ✓ The theory was formulated by William of Occam in the thirteenth century
- ✓ For phylogenetic analysis, parsimony seems a good assumption.

By this principle,

- ✓ a tree with the least number of substitutions is probably the best to explain the differences among the taxa under study.
- ✓ This view is justified by the fact that evolutionary changes are relatively rare within a reasonably short time frame.
- ✓ This implies that a tree with minimal changes is likely to be a good estimate of the true tree.
- ✓ By minimizing the changes, the method minimizes the phylogenetic noise owing to homoplasy and independent evolution.

(ii) Maximum Likelihood Method

- ✓ uses probabilistic models to choose a best tree that has the highest probability or likelihood of reproducing the observed data.
- ✓ It finds a tree that most likely reflects the actual evolutionary process.
- ✓ ML is an exhaustive method that searches every possible tree topology and considers every position in an alignment, not just informative sites.
- ✓ By employing a particular substitution model that has probability values of residue substitutions, ML calculates the total likelihood of ancestral sequences evolving to internal nodes and eventually to existing sequences.
- ✓ ML works by calculating the probability of a given evolutionary path for a particular extant sequence.
- ✓ The probability values are determined by a substitution model (either for nucleotides or amino acids).

(5) PHYLOGENETIC TREE EVALUATION

After phylogenetic tree construction, the next step is to statistically evaluate the reliability of the inferred phylogeny.

There are two questions that need to be addressed.

(i) how reliable the tree or a portion of the tree is;

(ii) whether this tree is significantly better than another tree.

To **answer the first question**,

We need to use analytical resampling strategies such as **bootstrapping** and **jackknifing**, which repeatedly resample data from the original dataset.

For the second question,

conventional statistical tests are needed.

What Is Bootstrapping?

- *Bootstrapping* is a statistical technique that tests the sampling errors of a phylogenetic tree.

- The rationale for bootstrapping is that a newly constructed tree is possibly biased owing to incorrect alignment or chance fluctuations of distance measurements.

To determine the robustness or reproducibility of the current tree,

- trees are repeatedly constructed with slightly perturbed alignments that have some random fluctuations introduced.

A truly robust phylogenetic relationship should have enough characters to support the relationship even if the dataset is perturbed in such away.

Otherwise, the noise introduced in the resampling process is sufficient to generate different trees, indicating that the original topology may be derived from weak phylogenetic signals. Thus, this type of analysis gives an idea of the statistical confidence of the tree topology.

Parametric and Nonparametric Bootstrapping

- Bootstrap resampling relies on perturbation of original sequence datasets.

There are two perturbation strategies.

Nonparametric bootstrapping - is through random replacement of sites.

parametric bootstrapping - new datasets can be generated based on a particular sequence

distribution.

Both types of bootstrapping can be applied to the distance, parsimony, and likelihood tree construction methods.

In nonparametric bootstrapping,

a new multiple sequence alignment of the same length is generated with random duplication of some of the sites (i.e., the columns in an alignment) at the expense of some other sites.

- In other words, certain sites are randomly replaced by other existing sites.

- This process is repeated 100 to 1,000 times to create 100 to 1,000 new alignments that are used to reconstruct phylogenetic trees using the same method as the originally inferred tree.

- The new datasets with altered the nucleotide or amino acid composition and rate heterogeneity may result in certain parts of the tree having a different topology from the original inferred tree.

- All the bootstrapped trees are summarized into a consensus tree based on a majority rule.

- The most supported branching patterns shown at each node are labeled with bootstrap values, which are the percentage of appearance of a particular clade.

Thus,

the bootstrap test provides a measure for evaluating the confidence levels of the tree topology. Analysis has shown that a bootstrap value of 70% approximately corresponds to 95% statistical confidence.

In parametric bootstrapping

- uses altered datasets with random sequences confined within a particular sequence distribution according to a given substitution model.

- The parametric bootstrapping method may help avoid the problem of certain sites being repeated too many times as in nonparametric bootstrapping resulting in skewed sequence distribution.

- If a correct nucleotide/amino acid distributionmodel is used, parametric bootstrapping generates more reasonable replicates than random replicates. Thus, this procedure is considered more robust than nonparametric bootstrapping.

Jackknifing

- In addition to bootstrapping, another often used resampling technique is jackknifing.

- In jackknifing, one half of the sites in a dataset are randomly deleted, creating datasets half as long as the original.

- Each new dataset is subjected to phylogenetic tree construction using the same method as the original.

Advantage of jackknifing

- is that sites are not duplicated relative to the original dataset and that computing time

is much shortened because of shorter sequences.

Disadvantage of this approach

- is that the size of datasets has been changed into one half and that the datasets are no

longer considered replicates. Thus, the results may not be comparable with that from bootstrapping.

PHYLOGENETIC PROGRAMS

There are numerous phylogenetic programs available,

For a list of hundreds of phylogenetic software programs, available in Felsenstein’s collection at: http://evolution.genetics.washington.edu/phylip/software.html.

Most of these programsare freely available. Some are comprehensive packages; others are more specialized to perform a single task.

PAUP* (Phylogenetic analysis using parsimony and other methods, by David Swofford, http://paup.csit.fsu.edu/)

- is a commercial phylogenetic package.
- It is probably one of the most widely used phylogenetic programs available from Sinauer Publishers.
- It is a Macintosh program (UNIX version available in the GCG package) with a very user-friendly graphical interface.
- PAUP was originally developed as a parsimony program, but expanded to a comprehensive package that is capable of performing distance, parsimony, and likelihood analyses.
- The distance options include NJ, ME, FM, andUPGMA.
- PAUP is also able to perform nonparametric bootstrapping, jackknifing, KH testing, and SH testing.

Phylip (Phylogenetic inference package; by Joe Felsenstein) at http:// bioweb.pasteur.fr/seqanal/phylogeny/phylip-uk.html.

- is a free multi platform comprehensive package containing thirty-five subprograms for performing distance, parsimony, and likelihood analysis, as well as bootstrapping for both nucleotide and amino acid sequences.
- It is command-line based, but relatively easy to use for each single program.

PHYML (http://atgc.lirmm.fr/phyml/)

- is a web-based phylogenetic program using the GA.
- It first builds an NJ tree and uses it as a starting tree for subsequent iterative refinement through subtree swapping.
- Branch lengths are simultaneously optimized during this process.
- The tree searching stops when the total ML score no longer increases.
- The main advantage of this program is the ability to build trees from verylarge datasets with hundreds of taxa and to complete tree searching within a relatively short time frame.

Database similarity searching

- A main application of pairwise alignment is retrieving biological sequences in databases based on similarity.
- This process involves submission of a query sequence and performing a pairwise comparison of the query sequence with all individual sequences in a database.
- Thus, database similarity searching is pairwise alignment on a large scale.
- This type of searching is one of the most effective ways to assign putative functions to newly determined sequences.

UNIQUE REQUIREMENTS OF DATABASE SEARCHING

There are unique requirements for implementing algorithms for sequence database searching.

1. ***sensitivity***, which refers to the ability to find as many correct hits as possible.

These correct hits are considered "true positives" in the database searching exercise.

2. ***selectivity***, also called *specificity*, which refers to the ability to exclude incorrect hits. These incorrect hits are unrelated sequences mistakenly identified in database searching and are considered "false positives."

3. ***speed***, which is the time it takes to get results from database searches.

In database searching, there are two fundamental types of algorithms.

1. ***exhaustive type***, which uses a rigorous algorithm to find the best or exact solution for a particular problem by examining all mathematical combinations. Dynamic programming is an example of the exhaustive method and is computationally very intensive.

2. ***heuristic type***, which is a computational strategy to find an empirical or near optimal solution by using rules of thumb. Essentially, this type of algorithms take shortcuts by reducing the search space according to some criteria.

Sequence alignment

- It is an important first step toward structural and functional analysis of newly determined sequences.

- As new biological sequences are being generated at exponential rates, sequence comparison is becoming increasingly important to draw functional and evolutionary inference of a new protein with proteins already existing in the database.
- Sequence comparison lies at the heart of bioinformatics analysis.

The most fundamental process in this type of comparison is sequence alignment.

- This is the process by which sequences are compared by searching for common character patterns and establishing residue–residue correspondence among related sequences.

Pairwise sequence alignment

- is the process of aligning two sequences
- is the basis of database similarity searching and multiple sequence alignment.

EVOLUTIONARY BASIS of sequence alignment

- DNA and proteins are products of evolution.
- The building blocks of these biological macromolecules, nucleotide bases, and amino acids form linear sequences that determine the primary structure of the molecules.
- These molecules can be considered molecular fossils that encode the history of millions of years of evolution.
- During this time period, the molecular sequences undergo random changes, some of which are selected during the process of evolution.
- As the selected sequences gradually accumulate mutations and diverge over time, traces of evolution may still remain in certain portions of the sequences to allow identification of the common ancestry.
- The presence of evolutionary traces is because some of the residues that perform key functional and structural roles tend to be preserved by natural selection; other residues that may be less crucial for structure and function tend to mutate more frequently.

For example, active site residues of an enzyme family tend to be conserved because they are responsible for catalytic functions. **Therefore, by comparing sequences through alignment, patterns of conservation and variation can be identified**.

When a sequence alignment is generated correctly,

- it reflects the evolutionary relationship of the two sequences:

- regions that are aligned but not identical represent residue substitutions;

- regions where residues from one sequence correspond to nothing in the other represent insertions or deletions that have taken place on one of the sequences during evolution.

The **degree of sequence conservation** in the alignment reveals evolutionary relatedness of different sequences, whereas the **variation** between sequences reflects the changes that have occurred during evolution in the form of substitutions, insertions, and deletions.

- Identifying the evolutionary relationships between sequences helps to characterize the function of unknown sequences.

When a sequence alignment reveals *significant* similarity among a group of sequences, they can be considered as belonging to the **same family**.

If one member within the family has a known structure and function, then that information can be transferred to those that have not yet been experimentally characterized.

- Therefore, **sequence alignment can be used as basis** for prediction of structure and function of uncharacterized sequences.

- Sequence alignment provides inference for the relatedness of two sequences under study.

If the two sequences share significant similarity, meaning that the two sequences must have derived from a common evolutionary origin.

SEQUENCE HOMOLOGY VERSUS SEQUENCE SIMILARITY

An important concept in sequence analysis is sequence homology.

Sequence homology

- is an inference or a conclusion about a common ancestral relationship drawn from sequence similarity comparison when the two sequences share a high enough degree of similarity.

- homology is a qualitative statement.

Sequence similarity

- is the percentage of aligned residues that are similar in physiochemical properties such as size, charge, and hydrophobicity.

- is a direct result of observation from the sequence alignment.

- Sequence similarity can be quantified using percentages;

For example, two sequences share 40% similarity.

Generally, the sequence similarity level depends on

- ✓ the type of sequences being examined
- ✓ sequence lengths.

Nucleotide sequences consist of only four characters,

therefore, unrelated sequences have at least a 25% chance of being identical.

For protein sequences, there are twenty possible amino acid residues,

therefore, two unrelated sequences can match up 5% of the residues by random chance.

If gaps are allowed, the percentage could increase to 10–20%.

Sequence length is also a crucial factor.

shorter the sequence, the higher the chance that some alignment by random chance.

longer the sequence, the higher the chance that some alignment by random chance.

For determining a homology relationship of two protein sequences, for example, if both sequences are aligned at full length, which is 100 residues long, an identity of 30% or higher can be safely regarded as having close homology. They are referred to as **“safe zone”**

If their identity level falls between 20% and 30%, determination of homologous relationships in this range becomes less certain. This is the area regarded as the **“twilight zone”**.

Below 20% identity, where high proportions of nonrelated sequences are present, homologous relationships cannot be reliably determined and this area regarded as **“midnight zone”**.

SEQUENCE SIMILARITY VERSUS SEQUENCE IDENTITY

- Sequence similarity and sequence identity are synonymous for nucleotide sequences.
- For protein sequences, however, the two concepts are very different.

In a protein sequence alignment,

Sequence identity refers to the percentage of matches of the same amino acid residues between two aligned sequences.

Sequence similarity refers to the percentage of aligned residues that have similar physicochemical characteristics and can be more readily substituted for each other.

Methods of Pair-wise sequence alignment

The overall goal of pair-wise sequence alignment is to find the best pairing of two sequences, such that there is maximum correspondence among residues.

To achieve this goal, one sequence needs to be shifted relative to the other to find the position where maximum matches are found.

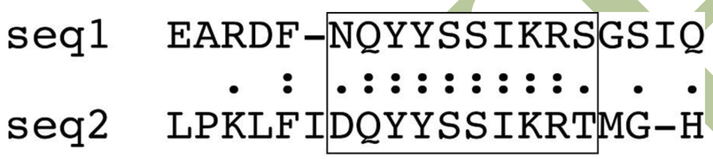
There are two different alignment strategies that are often used:

- global alignment
- local alignment.

Global Alignment

- In *global alignment*, two sequences to be aligned are assumed to be generally similar over their entire length.

- Alignment is carried out from beginning to end of both sequences to find the best possible alignment across the entire length between the two sequences.
- This method is more applicable for aligning two closely related sequences of roughly the same length.
- For divergent sequences and sequences of variable lengths, this method may not be able to generate optimal results because it fails to recognize highly similar local regions between the two sequences.

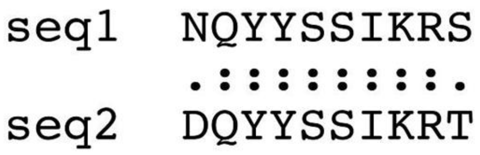


```
seq1   EARDF-NQYYSSIKRSGSIQ
           . :  .::::::::.  .  .
seq2   LPKLFIDQYYSSIKRTMG-H
```

In this figure, the region with the highest similarity is highlighted in a box.

Local alignment

- does not assume that the two sequences in question have similarity over the entire length.
- It only finds local regions with the highest level of similarity between the two sequences and aligns these regions without regard for the alignment of the rest of the sequence regions.
- This approach can be used for aligning more divergent sequences with the goal of searching for conserved patterns in DNA or protein sequences.
- The two sequences to be aligned can be of different lengths.
- This approach is more appropriate for aligning divergent biological sequences containing only modules that are similar, which are referred to as *domains* or *motifs*.



```
seq1   NQYYSSIKRS
       .::::::::.
seq2   DQYYSSIKRT
```

In the line between the two sequences, “:” indicates identical residue matches and “.”indicates similar residue matches.

Alignment Algorithms

Alignment algorithms, both global and local, are fundamentally similar and only differ in the optimization strategy used in aligning similar residues.

Both types of algorithms can be based on one of the three methods:

1. dot matrix method,
2. dynamic programming method,
3. word method.

Dot Matrix Method

The most basic sequence alignment method is the dot matrix method, also known as the *dot plot method.*

- It is a graphical way of comparing two sequences in a two dimensional matrix.

In a dot matrix,

- two sequences to be compared are written in the horizontal and vertical axes of the matrix.

- The comparison is done by scanning each residue of one sequence for similarity with all residues in the other sequence.

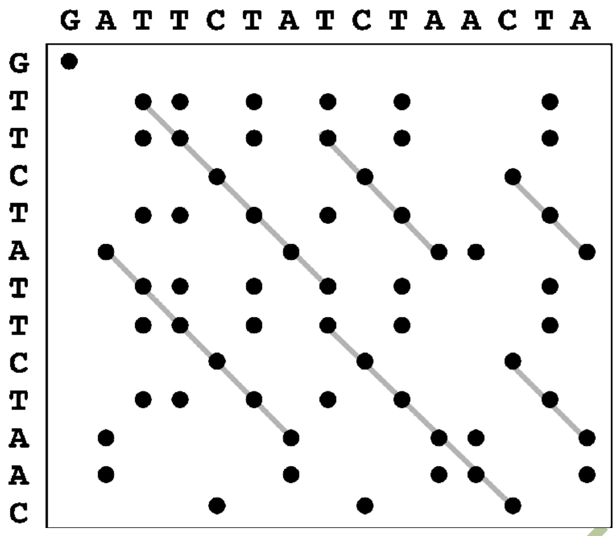
- If a residue match is found, a dot is placed within the graph. Otherwise, the matrix positions are left blank.

- When the two sequences have substantial regions of similarity, many dots line up to form contiguous diagonal lines, which reveal the sequence alignment.

- If there are interruptions in the middle of a diagonal line, they indicate insertions or deletions.

- Parallel diagonal lines within the matrix represent repetitive regions of the sequences (Figure).

Figure : Example of comparing two sequences using dot plots. Lines linking the dots in diagonals indicate sequence alignment. Diagonal lines above or below the main diagonal represent internal repeats of either sequence.



A problem exists when comparing large sequences using the dot matrix method due the high noise level.

For DNA sequences,

- the problem is particularly acute because there are only four possible characters in DNA and each residue therefore has a one-in-four chance of matching a residue in another sequence.
- There are many variations of using the dot plot method.

For example, a sequence can be aligned with itself to identify internal repeat elements.

- In the self comparison, there is a main diagonal for perfect matching of each residue.

- If repeats are present, short parallel lines are observed above and below the main diagonal.

- Self complementarity of DNA sequences (also called *inverted repeats*) – for example, those that form the stems of a hairpin structure – can also be identified using a dot plot.

- In this case, a DNA sequence is compared with its reverse-complemented sequence.

- Parallel diagonals represent the inverted repeats.

For comparing protein sequences,

- a weighting scheme has to be used to account for similarities of physicochemical properties of amino acid residues.
- The dot matrix method gives a direct visual statement of the relationship between two sequences and helps easy identification of the regions of greatest similarities.

Advantage of this method is

- Identification of sequence repeat regions based on the presence of parallel diagonals of the same size vertically or horizontally in the matrix.
- The method has some applications in genomics.
- It is useful in identifying chromosomal repeats
- comparing gene order conservation between two closely related genomes.
- used in identifying nucleic acid secondary structures through detecting self-complementarity of a sequence.

Limitation of this visual analysis method

- it lacks statistical rigor in assessing the quality of the alignment.
- The method is also restricted to pairwise alignment.
- It is difficult for the method to scale up to multiple alignment.

The following are examples of webservers that provide pairwise sequence comparison using dot plots.

- ✓ Dotmatcher (bioweb.pasteur.fr/seqanal/interfaces/dotmatcher.html)
- ✓ Dottup (bioweb.pasteur.fr/seqanal/interfaces/dottup.html)
- ✓ Dothelix (www.genebee.msu.su/services/dhm/advanced.html)
- ✓ MatrixPlot (www.cbs.dtu.dk/services/MatrixPlot/)

Dotmatcher

- aligns and displays dot plots of two input sequences (DNA or proteins) in FASTA format.

- A window of specified length and a scoring scheme are used.

- Diagonal lines are only plotted over the position of the windows if the similarity is above a certain threshold.

Dottup

- aligns sequences using the word method

- capable of handling genome-length sequences.

- Diagonal lines are only drawn if exact matches of words of specified length are found.

Dothelix

- is a dot matrix program for DNA or protein sequences.

- The program has a number of options for length threshold (similar to window size) and implements scoring matrices for protein sequences.

- In addition to drawing diagonal lines with similarity scores above a certain threshold,

- program displays actual pairwise alignment.

MatrixPlot

- is a more sophisticated matrix plot program for alignment of protein and nucleic acid sequences.

- The user has the option of adding information such as sequence logo profiles and distance matrices from known three-dimensional structures of proteins or nucleic acids.

- Instead of using dots and lines, the program uses colored grids to indicate alignment or other user-defined information.

Dynamic Programming Method

Dynamic programming is a method that determines optimal alignment by matching two sequences for all possible pairs of characters between the two sequences.

It is fundamentally similar to the dot matrix method in that it also creates a two dimensional alignment grid.

- it finds alignment in a more quantitative way by converting a dot matrix into a scoring matrix to account for matches and mismatches between sequences.

- By searching for the set of highest scores in this matrix, the best alignment can be accurately obtained.

- Dynamic programming works by first constructing a two-dimensional matrix

whose axes are the two sequences to be compared.

- The residue matching is according to a particular scoring matrix.

- The scores are calculated one row at a time.

- This starts with the first row of one sequence, which is used to scan through the entire length of the other sequence, followed by scanning of the second row. The matching scores are calculated.

- The scanning of the second row takes into account the scores already obtained in the first round. The best score is put into the bottom right corner of an intermediate matrix. This process is iterated until values for all the cells are filled.

- Thus, the scores are accumulated along the diagonal going from the upper left corner to the lower right corner.

- Once the scores have been accumulated in matrix, the next step is to find the path that represents the optimal alignment.

- This is done by tracing back through the matrix in reverse order from the lower right-hand corner of the matrix toward the origin of the matrix in the upper left-hand corner.

- The best matching path is the one that has the maximum total score.

If two or more paths reach the same highest score, one is chosen arbitrarily to represent the best alignment.

Most commonly used pairwise alignment web servers apply the local alignment strategy, which include SIM, SSEARCH, and LALIGN.

SCORING MATRICES

SIM (http://bioinformatics.iastate.edu/aat/align/align.html)

SSEARCH (http://pir.georgetown.edu/pirwww/search/pairwise.html)

LALIGN (www.ch.embnet.org/software/LALIGN form.html)

SIM

- is a web-based program for pairwise alignment using the Smith–Waterman algorithm that finds the best scored nonoverlapping local alignments between two sequences.

- It is able to handle tens of kilobases of genomic sequence.

- The user has the option to set a scoring matrix and gap penalty scores.

- A specified number of best scored alignments are produced.

SSEARCH

- is a simple web-based programs that uses the Smith–Waterman algorithm for pairwise alignment of sequences.

- Only one best scored alignment is given.

- There is no option for scoring matrices or gap penalty scores.

LALIGN

- is a web-based program that uses a variant of the Smith–Waterman algorithm to align two sequences.

- gives a specified number of best scored alignments.

- The user has the option to set the scoring matrix and gap penalty scores.

- The same web interface also provides an option for global alignment performed by the ALIGN program.

Multiple sequence alignment

- Multiple sequence alignment is an essential technique in many bioinformatics applications.

- A natural extension of pairwise alignment is multiple sequence alignment, which is to align multiple related sequences to achieve optimal matching of the sequences.
- Related sequences are identified through the database similarity searching.
- It is theoretically possible to use dynamic programming to align any number of sequences as for pairwise alignment. However, the amount of computing time and memory it requires increases exponentially as the number of sequences increases.
- As a consequence, full dynamic programming cannot be applied for datasets of more than ten sequences. In practice, heuristic approaches are most often used.
- As the process generates multiple matching sequence pairs, it is often necessary to convert the numerous pairwise alignments into a single alignment, which arranges sequences in such a way that evolutionarily equivalent positions across all sequences are matched.

Advantages of multiple sequence alignment

- it reveals more biological information than many pairwise alignments can.

- it allows the identification of conserved sequence patterns and motifs in the whole sequence family, which are not obvious to detect by comparing only two sequences.

- Many conserved and functionally critical amino acid residues can be identified in a protein multiple alignment.

- essential prerequisite to carrying out phylogenetic analysis of sequence families and prediction of protein secondary and tertiary structures.

- has applications in designing degenerate polymerase chain reaction (PCR) primers based on multiple related sequences.

SCORING FUNCTION

- Multiple sequence alignment is to arrange sequences in such a way that a maximum number of residues from each sequence are matched up according to a particular scoring function.
- The scoring function for multiple sequence alignment is based on the concept of sum of pairs (SP).

it is the sum of the scores of all possible pairs of sequences in a multiple alignment based on a particular scoring matrix.

- In calculating the SP scores, each column is scored by summing the scores for all possible pairwise matches, mismatches and gap costs.
- The score of the entire alignment is the sum of all of the column scores.
- The purpose of most multiple sequence alignment algorithms is to achieve maximum SP scores.

Many algorithms have been developed to achieve optimal alignment.

- Some programs are exhaustive in nature; some are heuristic.
- Because exhaustive programs are not feasible in most cases, heuristic programs are commonly used.

These include

- progressive,
- iterative,
- block-based approaches.

Progressive method

- is a stepwise assembly of multiple alignment according to pairwise similarity.

Example is **Clustal**, - which is characterized by adjustable scoring matrices and gap penalties as well as by the application of weighting schemes.

T-Coffee and **DbClustal** have been developed that combine both global and local alignment to generate more sensitive alignment.

Praline is profile based and has the capacity to restrict alignment based on protein structure information and is thus much more accurate than Clustal.

Iterative approach

- works by repetitive refinement of suboptimal alignments.

Block-based method - focuses on identifying regional similarities.

EXHAUSTIVE ALGORITHMS

The exhaustive alignment method involves examining all possible aligned positions simultaneously.

Similar to dynamic programming in pair-wise alignment, which involves the use of a two-dimensional matrix to search for an optimal alignment,

To use dynamic programming for multiple sequence alignment,

extra dimensions are needed to take all possible ways of sequence matching into consideration.

This means to establish a multidimensional search matrix.

For example, for three sequences, a three-dimensional matrix is required to account for all possible alignment scores.

For aligning *N* sequences, an *N*-dimensional matrix is needed to be filled with alignment scores.

- As the amount of computational time and memory space required increases exponentially with the number of sequences, it makes the method computationally difficult to use for a large data set.

For this reason, full dynamic programming is limited to small datasets of less than ten short sequences.

DCA (Divide-and-Conquer Alignment, http://bibiserv.techfak.uni-bielefeld.de/dca/)

- is a web-based program that is in fact semi-exhaustive because certain steps of computation are reduced to heuristics.

- It works by breaking each of the sequences into two smaller sections.

- The breaking points are determined based on regional similarity of the sequences.

- If the sections are not short enough, further divisions are carried out.

- When the lengths of the sequences reach a predefined threshold, dynamic programming is applied for aligning each set of subsequences.

- The resulting short alignments are joined together head to tail to yield a multiple alignment of the entire length of all sequences.

HEURISTIC ALGORITHMS

Because the use of dynamic programming is not feasible for routine multiple sequence alignment, faster and heuristic algorithms have been developed.

The heuristic algorithms fall into three categories:

- ✓ progressive alignment type,
- ✓ iterative alignment type,
- ✓ block-based alignment type.

Progressive Alignment Method

- depends on the stepwise assembly of multiple alignment and is heuristic in nature.

- It speeds up the alignment of multiple sequences through a multistep process.

- It first conducts pairwise alignments for each possible pair of sequences using the Needleman–Wunsch global alignment method and records these similarity scores from the pairwise comparisons.

- The scores can either be percent identity or similarity scores based on a particular substitution matrix.

- Both scores correlate with the evolutionary distances between sequences.

- The scores are then converted into evolutionary distances to generate a distance matrix for all the sequences involved.

- A simple phylogenetic analysis is then performed based on the distance matrix to group sequences based on pair-wise distance scores.

As a result,

- a phylogenetic tree is generated using the neighbor-joining method.

- The tree reflects evolutionary proximity among all the sequences.

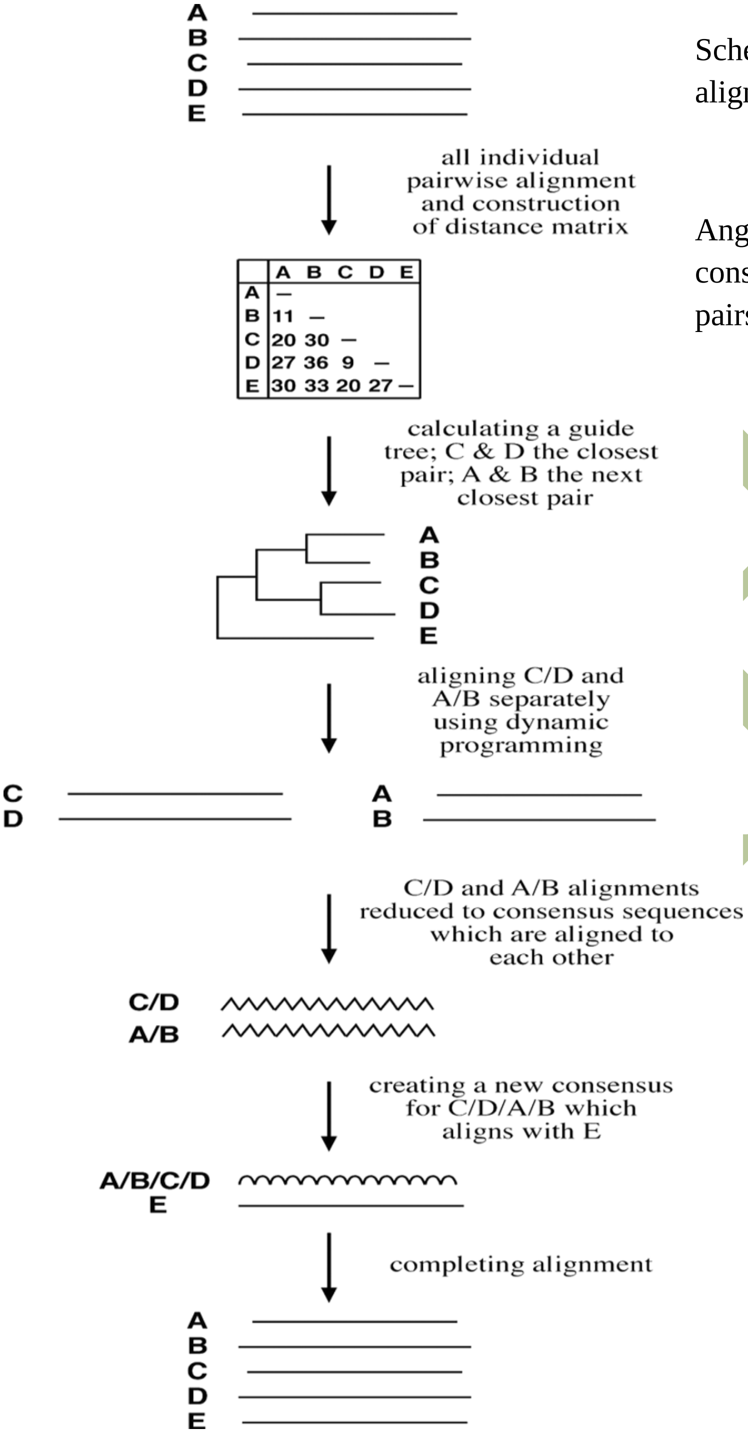
- the resulting tree is an approximate tree and the tree can be used as a guide for directing realignment of the sequences called as a *guide tree.*

According to the guide tree,

- The two most closely related sequences are first re-aligned using the Needleman–Wunsch algorithm.
- To align additional sequences, the two already aligned sequences are converted to a consensus sequence with gap positions fixed.
- The consensus is then treated as a single sequence in the subsequent step.

In the next step,

- the next closest sequence based on the guide tree is aligned with the consensus sequence using dynamic programming.
- More distant sequences or sequence profiles are subsequently added one at a time in accordance with their relative positions on the guide tree.
- After realignment with a new sequence using dynamic programming, a new consensus is derived, which is then used for the next round of alignment.
- The process is repeated until all the sequences are aligned (as shown in the Figure)


A
B
C
D
E
all individual pairwise alignment and construction of distance matrix
calculating a guide tree; C & D the closest pair; A & B the next closest pair
A
B
C
D
E
aligning C/D and A/B separately using dynamic programming
C
D
A
B
C/D and A/B alignments reduced to consensus sequences which are aligned to each other
C/D
A/B
creating a new consensus for C/D/A/B which aligns with E
A/B/C/D
E
completing alignment
A
B
C
D
E

	A	B	C	D	E
A	—				
B	11	—			
C	20	30	—		
D	27	36	9	—	
E	30	33	20	27	—

Schematic of a typical progressive alignment procedure (e.g., Clustal).

Angled wavy lines represent consensus sequences for sequence pairs A/B and C/D.

Clustal - (www.ebi.ac.uk/clustalw/)

is a progressive multiple alignment program available either as a stand-alone or on-line program.

The stand-alone program, which runs on UNIX and Macintosh, has two variants, ClustalW and ClustalX.

The W version provides a simple text-based interface and the X version provides user-friendly graphical interface.

Features of Clustal:

1) this program is the flexibility of using substitution matrices.

2) Clustal does not rely on a single substitution matrix.

Instead, it applies different scoring matrices when aligning sequences, depending on degrees of similarity.

3) The choice of a matrix depends on the evolutionary distances measured from the guide tree. For example,

for closely related sequences that are aligned in the initial steps,

Clustal automatically uses the BLOSUM62 or PAM120 matrix.

When more divergent sequences are aligned in later steps of the progressive alignment,

BLOSUM45 or PAM250 matrices may be used.

4) Clustal is the use of adjustable gap penalties that allow more insertions and deletions in regions that are outside the conserved domains, but fewer in conserved regions.

For example, a gap near a series of hydrophobic residues carries more penalties than the series of hydrophilic or glycine residues, which are common in loop regions.

In addition, gaps that are too close to one another can be penalized more than gaps occurring in isolated loci.

5) The program also applies a weighting scheme to increase the reliability of aligning divergent sequences (sequences with less than 25% identity).

This is done by down weightingredundant and closely related groups of sequences in the alignment by a certain factor.

6) This scheme is useful in preventing similar sequences from dominating the alignment.

The weight factor for each sequence is determined by its branch length on the guide tree.

The branch lengths are normalized by how many times sequences share a basal branch from the root of the tree. The obtained value for each sequence is subsequently used to multiply the raw alignment scores of residues from that sequence so to achieve the goal of decreasing the matching scores of frequent characters in a multiple alignment and thereby increasing the ones of infrequent characters.

DbClustal (http://igbmc.u-strasbg.fr:8080/DbClustal/dbclustal.html)

is a Clustalbased database search algorithm for protein sequences that combines local and global alignment features.

It first performs a BLASTP search for a query sequence.

The resulting sequence alignment pairs above a certain threshold are analyzed to obtain *anchorpoints*, which are common conserved regions, by using a program called Ballast.

A global alignment is subsequently generated by Clustal, which is weighted toward the anchor points.

Poa (Partial order alignments,www.bioinformatics.ucla.edu/poa/)

is a progressive alignment program that does not rely on guide trees.

Instead, the multiple alignment is assembled by adding sequences in the order they are given.

HEURISTIC DATABASE SEARCHING

Currently, there are two major heuristic algorithms for performing database searches:

BLAST and **FASTA**.

- These methods are not guaranteed to find the optimal alignment or true homologs, but are 50–100 times faster than dynamic programming.

- Both BLAST and FASTA use a heuristic *word method* for fast pairwise sequence alignment. This is the third method of pairwise sequence alignment.

- It works by finding short stretches of identical or nearly identical letters in two sequences. - - These short strings of characters are called *words*, which are similar to the windows used in the dot matrix method.

- Once regions of high sequence similarity are found, adjacent high-scoring regions can be joined into a full alignment.

BASIC LOCAL ALIGNMENT SEARCH TOOL (BLAST)

- was developed by Stephen Altschul of NCBI in 1990

- become one of the most popular programs for sequence analysis.
- BLAST uses heuristics to align a query sequence with all sequences in a database.

The objective is

- to find high-scoring ungapped segments among related sequences.

- helps to discriminate related sequences from unrelated sequences in a database.

BLAST performs sequence alignment through the following steps.

The **first step**

Is to create a list of words fromthe query sequence.
Each word is typically three residues for protein sequences and eleven residues for DNA sequences.
The list includes every possible word extracted from the query sequence.

This step is also called *seeding*.

The second step

is to search a sequence database for the occurrence of these words.

This step is to identify database sequences containing the matching words.

The matching of the words is scored by a given substitution matrix.

A word is considered a match if it is above a threshold.

The fourth step

involves pairwise alignment by extending from the words in both directions while counting the alignment score using the same substitution matrix.

The extension continues until the score of the alignment drops below a threshold due to mismatches (the drop threshold is twenty-two for proteins and twenty for DNA).

The resulting contiguous aligned segment pair without gaps is called *high-scoring segment pair* (HSP).

In the original version of BLAST, the highest scored HSPs are presented as the final report.

They are also called maximum scoring pairs.

A recent improvement in the implementation of BLAST

is the ability to provide gapped alignment.

- In gapped BLAST, the highest scored segment is chosen to be extended in both directions using dynamic programming where gaps may be introduced.

Variants of BLAST

BLAST is a family of programs that includes

BLASTN,

BLASTP,

BLASTX

TBLASTN,

TBLASTX.

BLASTN - queries nucleotide sequences with a nucleotide sequence database.

BLASTP - uses protein sequences as queries to search against a protein sequence database. **BLASTX** - uses nucleotide sequences as queries and translates them in all six reading frames to produce translated protein sequences, which are used to query a protein sequence database. **TBLASTN** - queries protein sequences to a nucleotide sequence database with the sequences translated in all six reading frames.

TBLASTX - uses nucleotide sequences, which are translated in all six frames, to search against a nucleotide sequence database that has all the sequences translated in six frames.

BLASTweb server(www.ncbi.nlm.nih.gov/BLAST/)

- ✓ has been designed in such a way as to simplify the task of program selection.
- ✓ The programs are organized based on the type of query sequences, protein sequences, nucleotide sequences, or nucleotide sequence to be translated.
- ✓ In addition, programs for special purposes are grouped separately;

For example,

bl2seq, immunoglobulinBLAST, and VecScreen,

- ✓ The BLAST programs specially designed for searching individual genome databases are also listed in a separate category.
- ✓ The choice of the type of sequences also influences the sensitivity of the search.

Review Questions:

Short Answer Questions **(2 Marks)**

1. Define alignment?
2. Differentiate pairwise and multiple Alignment?
3. Differentiate local and global alignment?
4. Define sequence homology?
5. Define sequence similarity?

6. Name any two methods for local alignment?
7. Name any two tools for alignment?
8. What are scoring matrices?
9. What are the algorithms developed for optimizing MSA?
10. Define phylogeny?
11. What is evolution?
12. Define taxa?
13. Define monophyletic group?
14. Differentiate dichotomy and polytomy?
15. Differentiate rooted tree and unrooted tree?
16. Differentiate cladogram and dendrogram?
17. What are substitution models? What is its application in predicting phylogeny?
18. Define bootstrapping?
19. What is meant by sensitivity of database searching?
20. Define BLAST?

Essay Answer Questions (6 & 8 Marks)

1. Describe dot matrix method of sequence alignment?
2. Write notes on the role of scoring matrices in sequence alignment?
3. Describe exhaustive algorithm of MSA?
4. Describe heuristic algorithm of MSA?
5. Write notes on ClustalW?
6. Define phylogenetic tree? What are the different ways of representing a tree?
7. Write notes on the unique requirements of database searching?
8. Describe BLAST.

Further Readings:

Jin Xiong (2006) Essential Bioinformatics, Cambridge University Press. (Pages: 31 to 47; 63 to 71; 127 to 167; 51 to 52).

Unit IV

SYLLABUS

Biological databases: Types of databases, Sequence databases, Nucleic acid sequence databases - Primary (GenBank, EMBL, DDBJ), Secondary (UniGene, SGD, EMI Genomes, Genome Biology), Protein sequence database – Primary (PIR, SWISS-PROT), Secondary (PROSITE, Pfam), Structural databases (PDB, SCOP, CATH), Bibliographic databases and Organism specific databases.

Introduction to Biological Databases

The very first challenge in the genomics is to store and handle the accumulating volume of raw sequence data and informations through the establishment and use of computer databases.

The development of databases to handle the large amount of molecular biological data is a fundamental task of bioinformatics.

WHAT IS A DATABASE?

A *database* is a computerized archive used to store and organize data in such a way that information can be retrieved easily by different search criteria.

Databases are composed of computer hardware and software for data management.

Objective of the database development

- To organize data in a set of structured records to enable easy retrieval of information.
- Each record, also called as *entry*, should contain a number of fields that hold the actual data items, for example, fields for names, phone numbers, addresses, dates.
- To retrieve a particular record from the database, a user can specify a particular piece of information, called *value*, to be found in a particular field and expect the computer to retrieve the whole data record. This process is called *making a query*.
- Biological databases have a higher level of requirement, known as *knowledge discovery*, which refers to the identification of connections between pieces of information that were not known when the information was first entered.

For example, databases containing raw sequence information can perform extra computational tasks to identify sequence homology or conserved motifs.

- These features facilitate the discovery of new biological insights from raw data.

Types of databases:

- Originally, databases all used a flat file format, which is a long text file that contains many entries separated by a *delimiter*, a special character such as a vertical bar (|).
- Within each entry are a number of fields separated by tabs or commas.
- To facilitate the access and retrieval of data, sophisticated computer software programs for organizing, searching, and accessing data have been developed.
- They are called as ***database management systems***.
- These systems contain not only raw data records but also operational instructions to help identify hidden connections among data records.
- The purpose of establishing a data structure is for easy execution of the searches and to combine different records to form final search reports.

Depending on the types of data structures, these database management systems can be classified into two types:

1. Relational database management systems
2. Object-oriented database management systems

Databases employing these management systems are known as

Relational databases

Object-oriented databases.

Relational Databases

- Instead of using a single table as in a flat file database, relational databases use a set of tables to organize data.

- Each table, also called a *relation*, is made up of columns and rows.
- Columns represent individual fields.
- Rows represent values in the fields of records.
- The columns in a table are indexed according to a common feature called an *attribute*, so they can be cross-referenced in other tables.
- Relational databases can be created using a special programming language called *structured query language* (SQL).

Object-Oriented Databases

- Object-oriented databases have been developed that store data as objects.
- In an object-oriented programming language, an object can be considered as a unit that combines data and mathematical routines that act on the data.
- Programming languages like C++ are used to create object-oriented databases.
- The object-oriented database system is more flexible; data can be structured based on hierarchical relationships.

BIOLOGICAL DATABASES

Current biological databases use all three types of database structures:

- flat files,
- relational,
- object oriented.

Based on their contents, biological databases can be divided into **three categories**:

1. primary databases,
2. secondary databases,
3. specialized databases.

Primary databases contain

- Original biological data.

- They are archives of raw sequence or structural data submitted by the scientific community.
- Examples: GenBank and Protein Data Bank (PDB).
- There are three major public sequence databases that store raw nucleic acid sequence data produced and submitted by researchers worldwide which are freely available on internet.
- GenBank,
- European Molecular Biology Laboratory (EMBL) database
- DNA Data Bank of Japan (DDBJ).
- Most of the data in the databases are contributed directly by authors with a minimal level of annotation.
- A small number of sequences, especially those published in the 1980s, were entered manually from published literature by database management staff.
- Presently, sequence submission to either GenBank, EMBL, or DDBJ is a precondition for publication in most scientific journals to ensure the fundamental molecular data to be made freely available.
- These three public databases closely collaborate and exchange new data daily.
- They together constitute the **International Nucleotide Sequence Database Collaboration.**
- This means that by connecting to any one of the three databases, one should have access to the same nucleotide sequence data.
- **PDB** - is a only one centralized database for the three-dimensional structures of biological macromolecules.
- This database archives atomic coordinates of macromolecules (both proteins and nucleic acids) determined by x-ray crystallography and NMR.
- It uses a flat file format to represent protein name, authors, experimental details, secondary structure, cofactors, and atomic coordinates.

- The web interface of PDB also provides viewing tools for simple image manipulation.

Secondary databases contain

- To turn the raw sequence information into more sophisticated biological knowledge, much post processing of the sequence information is needed.
- Thus secondary databases contains computationally processed sequence information derived from the primary databases.
- The amount of computational processing work varies greatly among the secondary databases;

some are simple archives of translated sequence data from identified open reading

frames in DNA,

whereas others provide additional annotation and information related to higher levels of information regarding structure and functions.

Example of secondary databases is

SWISS-PROT,

- which provides detailed sequence annotation that includes structure, function, and protein family assignment.
- The sequence data are mainly derived from TrEMBL, a database of translated nucleic acid sequences stored in the EMBL database.
- The annotation of each entry is carefully curated with good quality by human experts. The protein annotation includes function, domain structure, catalytic sites, cofactor binding, post translational modification, metabolic pathway information, disease association, and similarity with other sequences.
- Much of this information is obtained from scientific literature and entered by database curators.
- The annotation provides significant added value to each original sequence record.

- The data record also provides cross referencing links to other online resources of interest. Other features such as very low redundancy and high level of integration with other primary and secondary databases make SWISS-PROT very popular among biologists.

UniProt database

- A recent effort to combine **SWISS-PROT, TrEMBL, and PIR** led to the creation of UniProt database
- It has larger coverage than any one of the three databases while at the same time maintaining the original SWISS-PROT feature of low redundancy, cross-references, and a high quality of annotation.

Pfam and Blocks databases

- contain aligned protein sequence information, motifs and patterns, which can be used for classification of protein families and inference of protein functions.

DALI database

- is a protein secondary structure database that is vital for protein structure classification and threading analysis to identify distant evolutionary relationships among proteins.

Specialized databases contain

- Serve a specific research community or focus on a particular organism.
- The content of these databases may be sequences or other types of information.
- The sequences in these databases may overlap with a primary database, but may also have new data submitted directly by authors.
- Because they are often curated by experts in the field, they may have unique organizations and additional annotations associated with the sequences.
- Many genome databases that are taxonomic specific fall within this category.

- **Examples:** Flybase, WormBase, AceDB, TAIR, GenBank EST database and Microarray Gene Expression Database.
- Some of these deal with particular classes of sequence:

RDP - the 'Ribosomal Database Project' provides ribosome related data services to the scientific community, including online data analysis, rRNA derived phylogenetic trees, and aligned and annotated rRNA sequences.

HIV-SD - the 'HIV Sequence Database' collects, curates and annotates HIV and SIV sequence data and provides various tools for analysing this data.

IMGT - the 'ImMunoGeneTics database' is a database specialising in Immunoglobulins, T cell receptors and the Major Histocompatibility Complex (MHC) of all vertebrate species.

- Others nucleotide sequence databases are focussing on particular features such as:

TRANSFAC - contains sequence information on transcription factors and transcription factor binding sites.

EPD - the 'Eukaryotic Promoter Database' is an annotated non-redundant collection of eukaryotic POL II promoters, for which the transcription start site has been determined experimentally.

REBASE - for restriction enzymes and restriction enzyme sites.

GOBASE - is a specialised database of organelle genomes.

INFORMATION RETRIEVAL FROM BIOLOGICAL DATABASES

There are a number of retrieval systems for biological data.

The most popular retrieval systems for biological databases are

- Entrez
- Sequence Retrieval Systems (SRS)

To perform complex queries in a database often requires the use of Boolean operators.

Most search engines of public biological databases use some form of this Boolean logic.

This is to join a series of keywords using logical terms such as AND, OR, and NOT to indicate relationships between the keywords used in a search.

AND means that the search result must contain both words;

OR means to search for results containing either word or both;

NOT excludes results containing either one of the words.

Parentheses ( ) to define a concept if multiple words and relationships are involved, so that the computer knows which part of the search to execute first. Items contained within parentheses are executed first.

Quotes can be used to specify a phrase.

Entrez

- The NCBI developed and maintains Entrez, a biological database retrieval system.
- It is a gateway that allows text-based searches for a wide variety of data, including annotated genetic sequence information, structural information, as well as citations and abstracts, full papers, and taxonomic data.
- The key feature of Entrez is its ability to integrate information, which comes from cross-referencing between NCBI databases based on preexisting and logical relationships between individual entries.
- This is highly convenient: users do not have to visit multiple databases located in disparate places.

For example, in a nucleotide sequence page,

one may find cross-referencing links to the translated protein sequence, genome mapping data, or to the related PubMed literature information, and to protein structures if available.

There are **several options common to all NCBI databases** that help to narrow the search.

One option is **“Limits,”** which helps to restrict the search to a subset of a particular database (e.g., the field for author or publication date) or a particular type of data (e.g., chloroplast DNA/RNA).

Another option is **“Preview/Index,”** which connects different searches with the Boolean operators and uses a string of logically connected keywords to perform a new search.

“History” option provides a record of the previous searches so that the user can review, revise, or combine the results of earlier searches.

“Clipboard” that stores search results for later viewing for a limited time.

To store information in the Clipboard, the **“Send to Clipboard”** function should be used.

One of the databases accessible from Entrez is a biomedical literature database known as **PubMed,** which contains abstracts and in some cases the full text articles from nearly 4,000 journals.

An important feature of PubMed is the retrieval of information based on medical subject headings (MeSH) terms.

The MeSH system consists of a collection of more than 20,000 controlled and standardized vocabulary terms used for indexing articles.

PubMed uses a word weight algorithm to identify related articles with similar words in the titles, abstracts, and MeSH. By using this feature, articles on the same topic that were missed in the original search can be retrieved.

GenBank

is the most complete collection of annotated nucleic acid sequence data for almost every organism.

The content includes genomic DNA, mRNA, cDNA, ESTs, high throughput raw sequence data, and sequence polymorphisms.

GenBank is a relational database.

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences.

There are approximately 106,533,156,756 bases in 108,431,692 sequence records in the traditional GenBank divisions and 148,165,117,763 bases in 48,443,067 sequence records in the WGS division as of August 2009.

There is also a **GenPept database** for protein sequences, the majority of which are conceptual translations from DNA sequences, and amino acid sequences derived using peptide sequencing techniques.

There are two ways to search for sequences in GenBank.

Text-based keywords similar to a PubMed search

Molecular sequences to search by sequence similarity using BLAST.

The following are the informations obtainable from the GenBank

1. Submissions to GenBank

Many journals require submission of sequence information to a database prior to publication so that an accession number may appear in the paper.

There are several options for submitting data to GenBank:

- **BankIt**, a WWW-based submission tool for convenient and quick submission of sequence data
- **Sequin**, NCBI's stand-alone submission software for MAC, PC, and UNIX platforms, is available by FTP. When using Sequin, the output files for direct submission should be sent to GenBank by e-mail.
- **tbl2asn**, a command-line program, automates the creation of sequence records for submission to GenBank using many of the same functions as Sequin. It is used primarily for submission of complete genomes and large batches of sequences.

- **Barcode Submission Tool**, a WWW-based tool for the submission of GenBank sequences and trace data for Barcode of Life projects.

Currently, only mitochondrial cytochrome c oxidase subunit I (COI) genes are being accepted with this tool.

There are specialized, streamlined procedures for batch submissions of sequences, such as EST, STS, and GSS sequences.

2. Submissions of Sequence Reads

- Reads of Sanger-style sequencing can be submitted to the Trace Archive.
- Runs of next-generation sequencing, for example 454 or Solexa, can be submitted to the Short Read Archive (SRA).

3. Updating or Revising a GenBank Sequence

Revisions or updates to GenBank entries can be made by the submitters at any time and can be accepted through the Update option on the BankIt page.

4. Access to GenBank

There are several ways to search and retrieve data from GenBank.

- Search GenBank for sequence identifiers and annotations with Entrez Nucleotide, which is divided into three divisions:

CoreNucleotide (the main collection),

dbEST (Expressed Sequence Tags),

dbGSS (Genome Survey Sequences).

- Search and align GenBank sequences to a query sequence using **BLAST** (Basic Local Alignment Search Tool).
- Search, link, and download sequences programatically using NCBI e-utilities.

5. GenBank Data Usage

The GenBank database is designed to provide and encourage access within the scientific community to the most up to date and comprehensive DNA sequence information. Therefore, NCBI places no restrictions on the use or distribution of the GenBank data.

Structure of GenBank - Sequence Format

- ✓ To search GenBank effectively using the text-based method requires an understanding of the GenBank sequence format.
- ✓ Search output for sequence files is produced as flat files for easy reading.
- ✓ The resulting flat files contain three sections –

Header, Features, and Sequence entry.

LOCUS	The LOCUS field contains a number of different data elements, including locus name, sequence length, molecule type, GenBank division, and modification date. Each element is described below.
• Locus Name	The locus name in this example is SCU49845. The locus name was originally designed to help group entries with similar sequences: the first three characters usually designated the organism; the fourth and fifth characters were used to show other group designations, such as gene product; for segmented entries, the last character was one of a series of sequential integers. the locus name is usually the first letter of the genus and species names, followed by the accession number.
• Sequence Length	Number of nucleotide base pairs (or amino acid residues) in the sequence record. Example, the sequence length is 5028 bp. There is no maximum limit for sequence size that can be submitted to GenBank.

	The minimum length required for submission is 50 bp.
• Molecule Type	The type of molecule that was sequenced. Example, the molecule type is DNA or RNA or protein The various molecule types are described in the Sequin documentation and can include genomic DNA, genomic RNA, precursor RNA, mRNA (cDNA), ribosomal RNA, transfer RNA, small nuclear RNA, and small cytoplasmic RNA.
• GenBank Division	The GenBank division to which a record belongs is indicated with a three letter abbreviation. In this example, GenBank division is PLN. The GenBank database is divided into 18 divisions: 1. PRI - primate sequences 2. ROD - rodent sequences 3. MAM - other mammalian sequences 4. VRT - other vertebrate sequences 5. INV - invertebrate sequences 6. PLN - plant, fungal, and algal sequences 7. BCT - bacterial sequences 8. VRL - viral sequences 9. PHG - bacteriophage sequences 10. SYN - synthetic sequences 11. UNA - unannotated sequences 12. EST - EST sequences (expressed sequence tags) 13. PAT - patent sequences

	14. STS - STS sequences (sequence tagged sites) 15. GSS - GSS sequences (genome survey sequences) 16. HTG - HTG sequences (high-throughput genomic sequences) 17. HTC - unfinished high-throughput cDNA sequencing 18. ENV - environmental sampling sequences
• Modification Date	The date in the LOCUS field is the date of last modification. The sample record shown here was last modified on 21-JUN-1999. In some cases, the modification date might correspond to the release date.
DEFINITION	Brief description of sequence; includes information such as source organism, gene name/protein name, or some description of the sequence's function (if the sequence is non-coding). If the sequence has a coding region (CDS), description may be followed by a completeness qualifier, such as "complete cds".
ACCESSION	The unique identifier for a sequence record. An accession number applies to the complete record and is usually a combination of a letter(s) and numbers, such as a single letter followed by five digits (e.g., U12345) or two letters followed by six digits (e.g., AF123456). Accession numbers do not change, even if information in the record is changed at the author's request. Records from the RefSeq database of reference sequences have a different accession number format that begins with two letters followed by an underscore bar and six or more digits, for example: NT_123456 constructed genomic contigs

	NM_123456 mRNAs NP_123456 proteins NC_123456 chromosomes
VERSION	A nucleotide sequence identification number that represents a single, specific sequence in the GenBank database. If there is any change to the sequence data (even a single base), the version number will be increased, e.g., U12345.1 → U12345.2, but the accession portion will remain stable. The accession.version system of sequence identifiers runs parallel to the GI number system, i.e., when any change is made to a sequence, it receives a new GI number AND an increase to its version number.
• GI	"GenInfo Identifier" sequence identification number, in this case, for the nucleotide sequence. If a sequence changes in any way, a new GI number will be assigned. A separate GI number is also assigned to each protein translation within a nucleotide sequence record, and a new GI is assigned if the protein translation changes in any way (see below). GI sequence identifiers run parallel to the new accession.
KEYWORDS	Word or phrase describing the sequence. The Keywords field is present in sequence records primarily for historical reasons, and is not based on a controlled vocabulary.
SOURCE	Free-format information including an abbreviated form of the organism name, sometimes followed by a molecule type.
• Organism	The formal scientific name for the source organism (genus and species)

	and its lineage, based on the phylogenetic classification scheme used in the NCBI Taxonomy Database.
REFERENCE	Publications by the authors of the sequence that discuss the data reported in the record. References are automatically sorted within the record based on date of publication, showing the oldest references first. Some sequences have not been reported in papers and show a status of "unpublished" or "in press". Various classes of publication can be present in the References field, including journal article, book chapter, book, thesis/monograph, proceedings chapter, proceedings from a meeting, and patent. The last citation in the REFERENCE field usually contains information about the submitter of the sequence, rather than a literature citation. It is therefore called the "submitter block" and shows the words "Direct Submission" instead of an article title. The various subfields under References are searchable in the Entrez search fields noted below.
• AUTHORS	List of authors in the order in which they appear in the cited article. Enter author names in the form: Lastname AB (without periods after the initials).
• TITLE	Title of the published work or tentative title of an unpublished work. Sometimes the words "Direct Submission" instead of an article title.
• JOURNAL	MEDLINE abbreviation of the journal name. (Full spellings can be obtained from the Entrez Journals Database.)

• PUBMED	PubMed Identifier (PMID). References that include PubMed IDs contain links from the sequence record to the corresponding PubMed record.
• Direct Submission	Contact information of the submitter, such as institute/department and postal address.
FEATURES	Information about genes and gene products, as well as regions of biological significance reported in the sequence. These can include regions of the sequence that code for proteins and RNA molecules, as well as a number of other features.
• source	Mandatory feature in each record that summarizes the length of the sequence, scientific name of the source organism, and Taxon ID number. Can also include other information such as map location, strain, clone, tissue type, etc., if provided by submitter.
Taxon	A stable unique identification number for the taxon of the source oganism. A taxonomy ID number is assigned to each taxon (species, genus, family, etc.) in the NCBI Taxonomy Database.
• CDS	Coding sequence; region of nucleotides that corresponds with the sequence of amino acids in a protein (location includes start and stop codons). The CDS feature includes an amino acid translation. Submitters are also encouraged to annotate the mRNA feature, which includes the 5' untranslated region (5'UTR), coding sequences (CDS, exon), and 3' untranslated region (3'UTR).
• <1..206	Base span of the biological feature indicated to the left, in this case, a CDS feature. (The CDS feature is described above, and its base span includes the start and stop codons.) Features can be complete, partial on the 5' end,

	partial on the 3' end, and/or on the complementary strand. Examples: 1. complete feature is simply written as *n..m* Example: 687..3158 The feature extends from base 687 through base 3158 in the sequence shown. 2. < indicates partial on the 5' end Example: <1..206 The feature extends from base 1 through base 206 in the sequence shown, and is partial on the 5' end 3. > indicates partial on the 3' end Example: 4821..5028> The feature extends from base 4821 through base 5028 and is partial on the 3' end. 4. (complement) indicates that the feature is on the complementary strand Example: complement(3300..4037) The feature extends from base 3300 through base 4037 but is actually on the complementary strand.
protein_id	A protein sequence identification number, similar to the Version number of a nucleotide sequence. Protein IDs consist of three letters followed by five digits, a dot, and a version number. If there is any change to the sequence data (even a single amino acid), the version number will be increased, but the accession portion will remain stable (e.g., AAA98665.1 will change to AAA98665.2).

GI	"GenInfo Identifier" sequence identification number, in this case, for the protein translation. The GI system of sequence identifiers runs parallel to the accession.version system, which was implemented by GenBank, EMBL, and DDBJ in February 1999.
translation	The amino acid translation corresponding to the nucleotide coding sequence (CDS). In many cases, the translations are conceptual and can indicate whether the CDS is based on experimental or non-experimental evidence.
• gene	A region of biological interest identified as a gene and for which a name has been assigned. The base span for the gene feature is dependent on the furthest 5' and 3' features.
complement	Indicates that the feature is located on the complementary strand.
• Other Features	Examples of other records that show a variety of biological features; a graphic format is also available for each sequence record and visually represents the annotated features: • AF165912 (gene, promoter, TATA signal, mRNA, 5'UTR, CDS, 3'UTR) GenBank flat file • AF090832 (protein bind, gene, 5'UTR, mRNA, CDS, 3'UTR) GenBank flat file • L00727 (alternatively spliced mRNAs) GenBank flat file A complete list of features is available from the resources noted above.
ORIGIN	The ORIGIN may be left blank, may appear as "Unreported," or may give a local pointer to the sequence start, usually involving an experimentally determined restriction cleavage site or the genetic locus (if available). The sequence data begin on the line immediately below ORIGIN. To

	view/save the sequence data only, display the record in FASTA format.

Protein sequence database: -

Primary protein sequence databases:

1. PIR - Protein Information Resource

2. MIPS – Martinsried Institute for protein sequences

3. SWISS – PROT

4. TrEMBL – Translated EMBL

5. NRDB – Non-Redundant Database

6. OWL

7. SWISS-PROT + TrEMBL

Secondary protein sequence databases:

1. PROSITE

2. Profiles

3. PRINTS

4. Pfam

5. BLOCKS

6. IDENTIFY

Primary protein sequence databases:

1. PIR –

- Developed by Margaret Dayhoff during 1960s at National Biomedical Research Foundation (NBRF).

- Maintained by PIR-International consortium
- This consortium includes

- Protein Information Resource (PIR) at NBRF

- International Protein Information Database of Japan (JIPID)

- Martinsried Institute for Protein Sequences (MIPS)

- Based on data quality and annotation level, PIR database divided into four sections:

1. PIR 1 – contains fully classified and annotated entries

2. PIR 2 – contains preliminary entries, not completely reviewed and may contain redundancy

3. PIR 3 – contains unverified entries

4. PIR 4 - four catagories

(i) Conceptual translations of artefactual sequences

(ii) Conceptual translations of sequences that are not transcribed or translated

(iii) Conceptual translations of genetically engineered

(iv) sequences that are not genetically encoded and not produced on ribosomes.

SWISS-PROT

- established by the Department of Medical Biochemistry at the University of Geneva and EMBL during 1986.
- Now this database is maintained collaboratively by SIB (Swiss Institute of Bioinformatics) and EBI/EMBL.
- Minimally redundant database
- Interlinked to many other resources.
- This database provides high-level annotations including the descriptions of function of protein, structure of its domains, its post-translational modifications, varients etc.

- Now contains .... entries from more than ...... different species.

Structure of SWISS –PROT

The quality of annotations and structure of database made SWISS-PROT as choice of most research purposes than the other databases.

- Each entry in this database consists of following:
- Each line is flagged with a two-letter code – which helps to present the information in a structured way.
- Entries begins with an Identification line (ID)

Ends with a // terminator.

ID line – informs the entry name, length of the protein name.

Contains ID code – designed to be informative and people-friendly in the form of PROTEIN_SOURCE – indicates the organism name,

PROTEIN part of the code denotes the type of protein.

AC line – denotes Accession number – remain static between database releases.

DT line – provide information about the date of entry of the sequence to the database

And details of when it was last modified.

DE line – informs the name by which the protein is known.

GN line – gives the gene name

OS line – organism species name

OC line – Organism classification within the biological kingdoms.

The next section of the database provides a list of supporting references.

Following the **references comment lines** (CC) are present and divided into themes which tells about

- FUNCTION of protein

- Its post-transcriptional modifications (PTM),

- Its TISSUE SPECIFICITY, SUB CELLULAR LOCATION.

Database **cross-reference lines (DR)** follow the comment field – provides links to other databases including primary sequence, secondary databases, specialized databases etc.

KEYWORD line (KW) - provides the keyword related the entries

Feature Table line (FT) – highlights regions of interest in the sequences, including

local secondary structure (transmembrane domains), Ligand binding sites, post- translational modifications.

The final section of the database entry includes

SQ line (SQ) – sequence information in single letter amino acid code,

each line contains 60 residues.

Sequence data in SWISS-PROT, contains precursor form of protein, therefore informations related to size, molecular weight, region of signal sequence (SIGNAL), transit (TRANSIT) or pro-peptide (PROPEP) respectively. The keys CHAIN and PEPTIDE are used to denote the location of the mature form.

Secondary protein databases:

- These secondary protein sequence databases have become important tools for identifying distant relationships in novel sequences and for inferring protein function.
- These databases have developed by using signature-recognition methods to address different sequence analysis problems, resulting in rather different and independent databases.
- To perform a comprehensive analysis, a user therefore has to know several important things. For example,

- what are the resources and where can they be found?

- What is the difference between them in terms of diagnostic performance and

family coverage?

- What do the different search outputs mean?

- Is it sufficient to use just one of the databases, and if so, which one?

The sequence of an unknown protein is too distantly related to any protein of known structure to detect its resemblance by overall sequence alignment,

but it can be identified by the occurrence in its sequence of a particular cluster of residue types which is commonly known as a pattern, motif, signature, or fingerprint.

These motifs arise because of particular requirements on the structure of specific region(s) of a protein, which may be important, for example, for their binding properties or for their enzymatic activity.

There are a few databases available, which use different methodology and a varying degree of biological information on the characterised protein families, domains and sites.

A brief description of some of specialised protein sequence databases:

Secondary database	**Primary source**	**Stored information**
PROSITE	SWISS-PROT	Regular expressions (patterns)
Profiles	SWISS-PROT	Weighted matrics (profiles)
PRINTS	OWL	Aligned motifs (fingerprints)
Pfam	SWISS-PROT	Hidden Markov Models (HMMs)
BLOCKS	PROSITE/PRINTS	Aligned motifs (Blocks)
IDENTIFY	BLOCKS/PRINTS	Fuzzy regular expressions (patterns)

Examples of secondary protein databases include:

- **PROSITE** –

First secondary database developed and maintained by Swiss Institute of Bioinformatics.

is the extensive documentation on many protein families, as defined by sequence domains or motifs.

PROSITE contains biologically significant sites and patterns using computational tools and it can rapidly and reliably identify to which family of proteins the new sequence belongs.

The profile structure used in PROSITE is similar to but slightly more general than the one introduced by Gribskov and co-workers (Gribskov et al.,1987). Generalised profiles are remarkably similar to the specific type of Hidden Markov Models (HMMs) used in Pfam.

Structure of PROSITE:

ID (IDentification) line - is always the first line of an entry.

The general form of the ID line is:

ID ENTRY_NAME; ENTRY_TYPE.

The first item on the ID line is the entry name. This name is a useful means of identifying an entry.

The entry name consists of from 2 to 21 uppercase alphanumeric characters. The characters that are allowed in an entry name are: A-Z, 0-9, and the underscore character "_".

The second item on the ID line indicates the type of PROSITE entry. Currently this can be one the following:

PATTERN

MATRIX

RULE

AC (ACcession number) line –

- It is always the second line of an entry.

- lists the accession number associated with an entry.

- Accession numbers provide a stable way of identifying entries from release to release.

- Accession numbers allow unambiguous citation of database entries.

- Researchers who wish to cite a PROSITE entry in their publications should always cite the accession number of that entry in order to ensure that readers can find the relevant data in a subsequent release.

The format of the AC line is:

AC PSnnnnn;

Where 'PS' stands for PROSITE and 'nnnnn' is a five digit number.

DT (DaTe) line –

- It is always the third line of an entry.

- shows the date of entry or last modification of the entry.

The format of the DT line is:

DT MMM-YYYY (CREATED); MMM-YYYY (DATA UPDATE); MMM-YYYY (INFO UPDATE).

where:

MMM is the month and YYYY the year.

First date indicates when the entry first appeared in the database.

Second date indicates when the 'primary' data of the entry was last modified.

Third date indicates when any data other then the 'primary' data has been modified.

Example:

DT APR-1990 (CREATED); JUL-1990 (DATA UPDATE); JUL-1998 (INFO UPDATE).

DE (DEscription) line -

- It is always the fourth line of an entry.

- provides descriptive information about the content of the entry.

The format of the DE line is:

DE Description.

The description is given in ordinary English and is free-format.
Examples:

DE Myb DNA-binding domain repeat signature 1.

PA (PAttern) lines –

- contains the definition of a PROSITE pattern.

The patterns are described using the following conventions:

- The standard IUPAC one-letter codes for the amino acids are used.
- The symbol 'x' is used for a position where any amino acid is accepted.
- Ambiguities are indicated by listing the acceptable amino acids for a given position, between square parentheses '[ ]'. For example: [ALT] stands for Ala or Leu or Thr.
- Ambiguities are also indicated by listing between a pair of curly brackets '{ }' the amino acids that are not accepted at a given position.

  For example: {AM} stands for any amino acid except Ala and Met.
- Each element in a pattern is separated from its neighbor by a '-'.
- Repetition of an element of the pattern can be indicated by following that element with a numerical value or a numerical range between parenthesis. Examples: x(3) corresponds to x-x-x, x(2,4) corresponds to x-x or x-x-x or x-x-x-x.
- When a pattern is restricted to either the N- or C-terminal of a sequence, that pattern either starts with a '<' symbol or respectively ends with a '>' symbol. In some rare cases (e.g. PS00267 or PS00539), '>' can also occur inside square brackets for the C-terminal element. 'F-[GSTV]-P-R-L-[G>]' means that either 'F-[GSTV]-P-R-L-G' or 'F-[GSTV]-P-R-L>' are considered.

A period ends the pattern.

Examples:

PA [AC]-x-V-x(4)-{ED}.

This pattern is translated as: [Ala or Cys]-any-Val-any-any-any-any-{any but Glu or Asp}

PA <A-x-[ST](2)-x(0,1)-V.

This pattern, which must be in the N-terminal of the sequence ('<'), is translated as: Ala-any-[Ser or Thr]-[Ser or Thr]-(any or none)-Val

MA (MAtrix) lines –

contain the definition of a PROSITE profile (or matrix) entry.

PP line –

PROSITE profiles normally use two cut-off levels,

a reliable cut-off (LEVEL=0) and

a low confidence cut-off (LEVEL=-1).

The low level cut-off usually covers the twilight zone where few true positives, that cannot be separated from false positives, might be present.

The output of the *pfsearch* and the *pfscan* programs indicate strong matches (level 0) with '!' and weak matches (level -1) with '?'.

This specific tagging in the match list can be used in post-processing, to validate some true positives present in the twilight zone or to eliminate some false positives detected with significant score.

NR (Numerical Results) lines –

contain information relevant to the results of the scan with a pattern on the complete Swiss-Prot knowledgebase.

The format of the NR line is:

NR /QUALIFIER=data; /QUALIFIER=data; ....

The qualifiers that are currently defined are:

/RELEASE Swiss-Prot release number and total number of sequence entries in that release.

/TOTAL Total number of hits in Swiss-Prot.

/POSITIVE Number of hits on proteins that are known to belong to the set in consideration.

/UNKNOWN Number of hits on proteins that could possibly belong to the set in consideration.

/FALSE_POS Number of false hits (on unrelated proteins).

/FALSE_NEG Number of known missed hits.

/PARTIAL Number of partial sequences which belong to the set in consideration, but which are not hit by the pattern or profile because they are partial (fragment) sequences.

CC (Comments) lines –

contains various types of comments.

The format of the CC line is:

CC /QUALIFIER=data; /QUALIFIER=data; ....

The qualifiers that are currently defined are:

/TAXO_RANGE Taxonomic range.

/MAX-REPEAT Maximum known number of repetitions of the pattern or profile in a single protein.

/SITE	Indication of an `interesting' site in a pattern.
/SKIP-FLAG	Indication of an entry that can be, in some cases, ignored by a program (because it is too unspecific).
/VERSION	The version number of a pattern or a profile.

There are 5 qualifiers specific to profile entries:

/MATRIX_TYPE	Describes the region of the protein identified by the profile.
/SCALING_DB	Scaling database used to calibrate the profile.
/AUTHOR	Author of the profile.
/FT_KEY	Feature key to describe the region covered by the profile.
/FT_DESC	Feature description of the region covered by the profile.

- **PRINTS** - A different approach to pattern recognition, termed "fingerprinting" is used by this database. Within a sequence alignment, it is usual to find several motifs that characterise the aligned family. The ability to tolerate mismatches, both at the level of residues within individual motifs, and at the level of motifs within the fingerprint as a whole, renders fingerprinting a powerful diagnostic technique.
- **Pfam** - Another important secondary protein database is Pfam. The methodology used by Pfam to create protein family or domain signatures is Hidden Markov Models (HMMs). HMMs are closely related to profiles, but are based on probability theory methods. These allow a direct statistical approach to identifying and scoring matches, and also to combining information from a multiple alignment with prior knowledge. These databases are useful for analysing multidomain proteins. The biggest drawback of Pfam is its lack of biological information (annotation) of the protein families.

- **BLOCKS** - Blocks are multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins. The blocks for the Blocks Database are made automatically by looking for the most highly conserved regions in groups of proteins documented in InterPro.
- **SBASE** - This is a protein domain library sequences database that contains annotated structural, functional, ligand-binding and topogenic segments of proteins, cross-referenced to all major sequence databases and sequence pattern collections.

Structural databases:

Structure database is a database that is modeled around the various experimentally determined macromolecular structures.

- This Database contains Structures of Protein, DNA, and RNA Molecules. Most coordinates were obtained from X-Ray or NMR studies. The Database is maintained at the Brookhaven National Laboratory.
- The aim of most protein structure databases is to organize and annotate the protein structures, providing the biological community access to the experimental data in a useful way.
- The number of known protein structures is increasing very rapidly and these are available through the Protein Data Bank (PDB).
- The Nucleic Acid Database (NDB) is the database for structural information about nucleic acid molecules.
- The Cambridge Crystallographic Data Centre (CCDC) provides a database of structures of ‘small molecules', of interest to biologists concerned with protein-ligand interactions.

Examples of MACROMOLECULAR 3D STRUCTURE DATABASES

Protein Data Bank (PDB):

- The Protein Data Bank (PDB) was established in 1971 as the central archive of all experimentally determined protein structure data.
- Today the PDB is maintained by an international consortia collectively known as the Worldwide Protein Data Bank (wwPDB).
- Aim of the wwPDB is to maintain a single archive of macromolecular structural data that is freely and publicly available to the global community.

RCSB PDB : (http://www.rcsb.org/pdb/home/)

- The RCSB PDB contains 3-D biological macromolecular structure data from X-ray crystallography, NMR, and Cryo-EM.
- It is operated by Rutgers, The State University of New Jersey and the San Diego Supercomputer Center at the University of California, San Diego.
- The RCSB PDB provides a variety of tools and resources for studying the structures of biological macromolecules and their relationships to sequence, function, and disease.

MMDB: http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml

- NCBI's structure database is called MMDB (Molecular Modeling DataBase),
- it is a subset of three-dimensional structures obtained from the Protein Data Bank (PDB) excluding theoretical models.
- MMDB is a database of ASN.1-formatted records.
- It was designed for flexibility, and as such, is capable of archiving conventional structural data as well as future descriptions of biomolecules, such as those generated by electron microscopy (surface models).

EBI structure databases: (http://www.ebi.ac.uk/Databases/structure.html)

1) **MSD (macromolecular structure databases)**:

The Macromolecular Structure Database is a European project for the collection, management and distribution of data about macromolecular structures. It is responsible for the deposition and validation of new protein structures. It includes PDB search tools.

2) **CSA(Catalytic Site Atlas)**:

The Catalytic Site Atlas is a resource of catalytic sites and residues identified in enzymes using structural data.

3) **DSSP**:

The DSSP database is a database of secondary structure assignments (and much more) for all of the entries in the Protein Data Bank (PDB).

4) **HSSP(homology-derived structures of proteins)**:

HSSP is a derived database merging structural (2-Dimensional and 3-Dimensional) and sequence information (1-Dimensional).

5) **PDBsum**:

PDBsum is a pictorial database providing an at-a-glance overview of every macromolecular structure (nucleic acids and proteins) deposited in the Protein Data Bank (PDB).

PSdB: (http://www.daviddeerfield.com/PSdb/)

- The Protein Structure Database (PSdb), is a protein database, derived from the information available in the Protein Databank and NRL-3D database,
- It relates secondary (e.g. Helix, Sheet, Turn, Random Coil) and tertiary information (e.g. Solvent accessibility, internal relative distances, and ligand interactions) to the primary structure.

CATH: (http://www.cathdb.info/)

- CATH is a hierarchical classification of protein domain structures, which clusters proteins at four major levels :

Class (C), Architecture (A), Topology (T) and Homologous superfamily (H).

- The boundaries and assignments for each protein domain are determined using a combination of automated and manual procedures which include computational techniques, empirical and statistical evidence, literature review and expert analysis.

SCOP: (http://Scopes-lmb.cam.ac.uk/scop/)

- The SCOP database, created by manual inspection and by a battery of automated methods, aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known.
- It provides a broad survey of all known protein folds, detailed information about the close relatives of any particular protein, and a framework for future research and classification.

SWISS-3D IMAGE: (http://expasy.org/sw3d/)

- It is an image database which strives to provide high quality pictures of biological macromolecules with known three-dimensional structure.
- The database contains mostly images of experimentally elucidated structures,
- but also provides views of well accepted theoretical protein models.

SWISS-MODEL: (http://swissmodell.expasy.org//SWISSMODEL.html)

- SWISS-MODEL is a fully automated protein structure homology-modeling server, accessible via the ExPASy web server, or from the program DeepView (Swiss Pdb-Viewer).

ModBase: (http://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi)

- It is a database of annotated comparative protein structure models, and associated resources. MODBASE contains theoretically calculated models, not experimentally determined structures.

Bibliographic databases

- Services that produced abstracts of scientific literature began to make their data available in machine-readable form in the early 1960's.
- A **bibliographic database** is a database of bibliographic records, an organized digital collection of references to published literature, including journal and newspaper articles, conference proceedings, reports, government and legal publications, patents, books, etc.
- In contrast to library catalogue entries, a large proportion of the bibliographic records in bibliographic databases describe analytics (articles, conference papers, etc.) rather than complete monographs,
- and they generally contain subject descriptions in the form of keywords, subject classification terms, or abstracts.
- A bibliographic database may be general in scope or cover a specific academic discipline.
- A significant number of bibliographic databases are still proprietary, available by licensing agreement from vendors, or directly from the abstracting and indexing services that create them.
- Many bibliographic databases evolve into digital libraries, providing the full-text of the indexed contents.
- Others converge with non-bibliographic scholarly databases to create more complete disciplinary search engine systems, such as Chemical Abstracts or Entrez.
- The tools that enable scientificaly coherent and efficient use of the resources described below are an important activity in the field of bioinformatics called Text Mining, which is described by Dr. Dietrich Rebholz-Schuhmann, from the EBI.

The best known bibliographic databases:

1. MEDLINE - accessible through EBI's SRS.

2. PUBMED -accessible through NCBI's ENTREZ.

EMBASE is a commercial product for the medical literature.

BIOSIS, the inheritor of the old Biological Abstracts, covers a broad biological field; the Zoological Record indexes the zoological literature.

CAB International maintains abstract databases in the fields of agriculture and parasitic diseases.

AGRICOLA is for the agricultural field what MEDLINE is for the medical field.

The bibliographical databases are with the exception of MEDLINE/PUBMED only available through commercial database vendors.

Organism specific databases:

- There are countless organism-specific databases present.

Below list includes only organisms that are of direct interest to researchers.

Virus

organism	**database**	**description**
Virus	VIDA	Organizes open reading frames ORFs) from viral genomic sequences into virus-specific homologous protein families.

Bacteria and Archaea

organism	**database**	**description**
Microbial	CMR	The Comprehensive Microbial Resource displays information on all of the publicly available, complete prokaryotic genomes.
Escherichia coli	EcoCyc	A database for the bacterium Escherichia coli K-12 MG1655.
Escherichia coli	EcoGene	Contains updated information about the E. coli K-12 genome and proteome sequences, including extensive gene bibliographies. A major EcoGene focus has been the re-

		evaluation of translation start sites.
Bacillus subtilis	SubtiList	A reference database for the Bacillus subtilis genome.
Bacillus subtilis	DBTBS	A database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information.

Protists

organism	**database**	**description**
Plasmodium	PlasmoDB	The Plasmodium Genome Resource hosts genomic and proteomic data (and more) for different species of the parasitic eukaryote Plasmodium. It brings together data provided by numerous laboratories worldwide, and adds its own data analysis.
Plasmodium falciparum	GeneDB P.falcipatum	The GeneDB is a project of the Sanger Institute Pathogen Sequencing Unit's and aims to provide reliable access to the latest sequence data and annotation/curation for the whole range of organisms sequenced by the Unit.
Tetrahymena thermophila	TGD	Provides information on the genome, genes, and proteins of Tetrahymena collected from the scientific literature, research community, and many other sources.

Fungi

organism	**database**	**Description**
Saccharomyces cerevisiae	SGD	Saccharomyces Genome Database is a scientific database of the molecular biology and genetics of the yeast Saccharomyces cerevisiae, which is commonly known as baker's or budding yeast.

Schizosaccharomyces pombe(*fission yeast*)	S.pombe GeneDB	Contains all S. pombe known and predicted protein coding genes, pseudogenes, transposons, tRNAs, rRNAs, snRNAs, snoRNAs and other known and predicted non-coding RNAs.
Neirospora crassa	MNCDB	The MIPS Neurospora crassa Genome Database aims to present information on the molecular structure and functional network of the entirely sequenced, filamentous fungus Neurospora crassa.

Animals - Invertebrates

organism	**database**	**description**
Drosophila melanogaster	FlyBase	A comprehensive database for information on the genetics and molecular biology of Drosophila. It includes data from the Drosophila Genome Projects and data curated from the literature.
Caenorhabditis elegans	Wormbase	Repository of mapping, sequencing and phenotypic information about the C. elegans and some related nematodes.

Animals - Vertebrates

organism	**database**	**description**
Homo sapiens	GDB	The Human Genome Database, GDB is the official central repository for genomic mapping data resulting from the Human Genome Initiative. Holds data on human gene loci, polymorphisms, mutations,probes, genetic maps, GenBank, citations and contacts.
Homo sapiens	HPRD	Human Protein Reference Database - is a comprehensive collection of protein features, post-translational modifications (PTMs, protein-protein interactions and disease association for each protein in

		the human proteome.
Homo sapiens	mtDB	Human Mitochondrial Genome Database provides a comprehensive database of complete human mitochondrial genomes.
Mus musculus	MGI	Mouse Genome Informatics provides integrated access to data on the genetics, genomics, and biology of the laboratory mouse.
Rattus	RGD	The Rat Genome Database curates and integrates rat genetic and genomic data and provides access to this data to support research using the rat as a genetic model for the study of human disease.

Plants

organism	**database**	**description**
Arabidopsis thaliana	TAIR	The Arabidopsis Information Resource maintains a database of genetic and molecular biology data that includes the complete genome sequence along with gene structure, gene product information, metabolism, gene expression, DNA and seed stocks, genome maps, genetic and physical markers, publications, and information about the Arabidopsis research community.
Arabidopsis thaliana	MATDB	MIPS Arabidopsis thaliana database is the www access to data of Arabidopsis sequences and annotation produced by the Arabidopsis Genome Initiative, plus the mitochondrial and chloroplast genomes.

Tree of Life

database	**description**

Tree of Life	Provides identification keys, figures, phylogenetic trees, and other systematic information for a group of organisms; and provides information about the evolutionary history and characteristics of creatures, from frogs and flowers to dinosaurs and protists. The project presents the evolutionary tree of life as an integrated whole.

Review Questions:

Short Answer Questions **(2 Marks)**

1. Define a database.
2. What is database management system?
3. Differentiate primary and secondary databases?
4. What is a relational database?
5. Define specialized databases?
6. Expand PDB and NCBI.
7. What is Uniprot database? What is its special feature?
8. What is Entrez?
9. Name any two options to submit data to Genbank?
10. Define Accession number? What is its role in sequence retrieval?
11. Expand PIR and MIPS.
12. What are the categories of PIR?
13. What are structural databases?
14. What is the difference between PDB and NDB?
15. Differentiate SCOP and CATH?
16. Define bibliographic database?
17. Give two examples of specialized database?
18. What is ModBase?
19. Expand MMDB? What is its application?
20. Expand PRINTS and BLOCKS?

Essay Answer Questions **(6 & 8 Marks)**

1. Describe in detail the classification of biological databases based on their contents?
2. Write notes on the different methods of information retrieval from biological databases?
3. Write notes on primary protein sequence databases?
4. What is secondary protein Sequence database? Describe with examples.
5. Give an account on 3D structure databases?
6. Write notes on Bibliographic databases?

Further Readings:

Attwood, TK., Parry-Smith, DJ. Introduction to Bioinformatics, Pearson Education, 2006.

Nucleic Acids Research, 2008 Database issue:D25-30

http://www.ncbi.nlm.nih.gov/books/NBK21105/pdf/ch1.pdf - GenBank book ref

http://www.ebi.ac.uk/2can/databases/protein7.html

http://www.science.co.il/Biomedical/Structure-Databases.asp

http://www.roseindia.net/bioinformatics/biologicaldatabases.shtml

http://www.ebi.ac.uk/2can/databases/bib.html

http://www.ebi.ac.uk/2can/databases/taxonomic.html

http://bioinformatics.igc.gulbenkian.pt/resources/databases/organismspecificdatabases/

http://www.expasy.ch/prosite/prosuser.html

Unit V

SYLLABUS

Searching Databases: SRS, Entrez, Sequence Similarity Searches-BLAST, FASTA, Data Submission. Genome Annotation: Pattern and repeat finding, Gene identification tools. Gene prediction: Gene prediction in prokaryote and eukaryotes. Extrinsic approaches and Ab initio approaches. Predicting the protein secondary structure (Domain, blocks, motifs), Predicting protein tertiary structure (Homology, Ab-initio, threading and fold recognition) and visualization of predicted structure.

Searching Databases

- The most obvious first stage in the analysis of any new sequence is to perform comparisons with sequence databases to find homologues. These searches can now be performed just about anywhere and on just about any computer. In addition, there are numerous web servers for doing searches, where one can post or paste a sequence into the server and receive the results interactively:
- There are many methods for sequence searching. By far the most well known are the BLAST suite of programs. One can easily obtain versions to run locally (either at NCBI or Washington University), and there are many web pages that permit one to compare a protein or DNA sequence against a multitude of gene and protein sequence databases.

Sequence Retrival System

- SRS (Sequence Retrieval System) is an information indexing and retrieval system designed for libraries with a flat file format such as the EMBL nucleotide sequence databank, the SwissProt protein sequence databank or the Prosite library of protein subsequence consensus patterns. SRS supports the data structure of these libraries by providing special indices for inzplemenzing lists of subenfities (e.g. feature tables) or hierarchically structured data–fields (e.g. taxonomic classification).

- A language (ODD) has been designed for the convenient specification of library format and organization, representation of individual data–fields within the system (design of indices) and structuring other data needed during retrieval. This ensures flexibility required for coping with different library formats, which are subject to continuous change. Queries and inspection of retrieved entries can be performed from a user interface with pull–down menus and windows. SRS supports rious input and output formats but is particularly well adapted to the GCG programs.

- SRS is a homogeneous interface to over 80 biological databases that had been developed at the European Bioinformatics Institute (EBI) at Hinxton, UK. It includes databases of sequences, metabolic pathways, transcription factors, application results (like BLAST, SSEARCH, FASTA), protein 3-D structures, genomes, mappings, mutations, and locus specific mutations.
- The web page listing all the databases contains a link to a description page about the database including the date on which it was last updated. You select one or more of the databases to search before entering your query. After getting results you choose an alignment algorithm (like CLUSTALW, PHYLIP) enter parameters, and run it. The SRS is highly recommended for use.

Sequence Similarity Searching

This is a method of searching sequence databases by using alignment to a query sequence. By statistically assessing how well database and query sequences match one can infer homology and transfer information to the query sequence.

The tools can be launched with different form pre-sets using the links - these can be changed on the tool page as well.

FASTA

FASTA is another commonly used sequence similarity search tool which uses heuristics for fast local alignment searching.

Protein Nucleotide Genomes Whole Genome Shotgun

SSEARCH

SSEARCH is an optimal (as opposed to heuristics-based) local alignment search tool using the Smith-Waterman algorithm. Optimal searches guarantee you find the best alignment score for your given parameters.

Protein Nucleotide Genomes Whole Genome Shotgun

NCBI BLAST

NCBI BLAST is the most commonly used sequence similarity search tool. It uses heuristics to perform fast **local** alignment searches.

Protein Nucleotide Vectors

PSI-BLAST

PSI-BLAST allows users to construct and perform a BLAST search with a custom, position-specific, scoring matrix which can help find distant evolutionary relationships. PHI-BLAST functionality is also available to restrict results using patterns.

Protein

Gene Prediction

With the rapid accumulation of genomic sequence information,

- there is a need to use computational approaches to accurately predict gene structure. Computational

- Gene prediction is a prerequisite for detailed functional annotation of genes and genomes.

- The process includes detection of the location of open reading frames (ORFs) and delineation of the structures of introns as well as exons if the genes of interest are of eukaryotic origin.

The ultimate goal is to describe all the genes computationally with near 100% accuracy.

The ability to accurately predict genes can significantly reduce the amount of experimental verification work required.

A number of gene prediction algorithms for prokaryotic genomes have been developed with varying degrees of success.

Algorithms for eukaryotic gene prediction, however, are still yet to reach satisfactory results.

CATEGORIES OF GENE PREDICTION PROGRAMS

The current gene prediction methods can be classified into two major categories,

1. ab initio–based approaches
2. homology-based approaches.

Ab initio–based approach

- predicts genes based on the given sequence alone.

- It depends on two major features associated with genes.

1) the existence of gene signals, which include start and stop codons, intron splice signals, transcription factor binding sites, ribosomal binding sites, and polyadenylation (poly-A) sites.

In addition, the triplet codon structure limits the coding frame length to multiples of three, which can be used as a condition for gene prediction.

2) gene content, which is statistical description of coding regions. It has been observed that nucleotide composition and statistical patterns of the coding regions tend to vary significantly from those of the non coding regions.

- These unique features can be detected by employing probabilistic models such as Markov models or hidden Markov models (HMMs) to help distinguish coding from noncoding regions.

Homology-based method

- makes predictions based on significant matches of the query sequence with sequences of known genes.

For instance, if a translated DNA sequence is found to be similar to a known protein or protein family from a database search, this can be strong evidence that the region codes for a protein.

Alternatively, when possible exons of a genomic DNA region match a sequenced cDNA, this also provides experimental evidence for the existence of a coding region.

GENE PREDICTION IN PROKARYOTES

Prokaryotes, which include bacteria and Archaea,

- have relatively small genomes with sizes ranging from 0.5 to 10Mbp (1Mbp=106 bp).

- The gene density in the genomes is high, with more than 90% of a genome sequence containing coding sequence.

- There are very few repetitive sequences.

- Each prokaryotic gene is composed of a single contiguous stretch of ORF coding for a single protein or RNA with no interruptions within a gene.

In bacteria,

- the majority of genes have a start codon ATG (or AUG in mRNA; because prediction is done at the DNA level, T is used in place of U), which codes for methionine.

- Occasionally, GTG and TTG are used as alternative start codons, but methionine is still the actual amino acid inserted at the first position.

Because there may be multiple ATG, GTG, or TGT codons in a frame, the presence of these codons at the beginning of the frame does not necessarily give a clear indication of the translation initiation site.

So, to help identify this initiation codon, other features associated with translation are used.

Those features are

1) ribosomal binding site,

also called the *Shine-Delgarno sequence*, which is a stretch of purine-rich sequence complementary to 16S rRNA in the ribosome. It is located immediately downstream of the transcription initiation site and slightly upstream of the translation start codon.

In many bacteria, it has a consensus motif of AGGAGGT. Identification of the ribosome binding site can help locate the start codon.

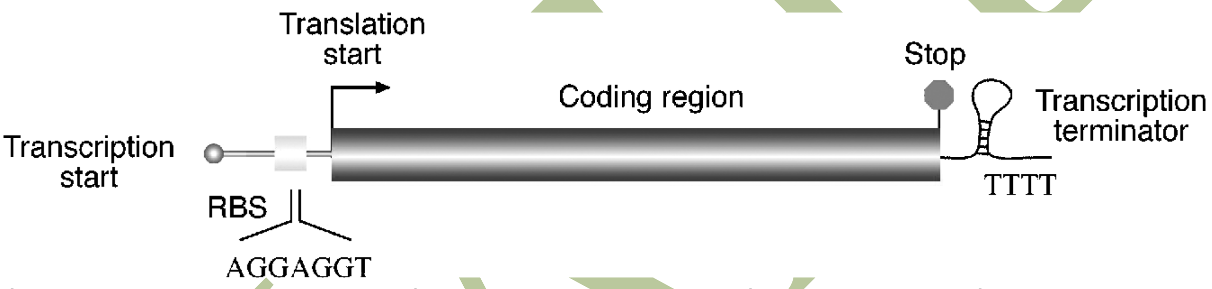
2) At the end of the protein coding region is a stop codon that causes translation to stop. There are three possible stop codons, identification of which is straightforward.

3) Many prokaryotic genes are transcribed together as one operon.

The end of the operon is characterized by a transcription termination signal called *ρ-independent terminator*. The terminator sequence has a distinct stem-loop secondary structure followed by a string of Ts.

4) Identification of the terminator site, in conjunction with promoter site identification can sometimes help in gene prediction.

Structure of a typical prokaryotic gene structure (RBS, ribosome binding site)



The following describes a number of HMM/IMM-based gene finding programs for prokaryotic organisms.

1) **GeneMark** (http://opal.biology.gatech.edu/GeneMark/)

is a suite of gene prediction programs based on the fifth-order HMMs.

The main program–GeneMark.hmm – is trained on a number of complete microbial genomes. If the sequence to be predicted is from a non listed organism, the most closely related organism can be chosen as the basis for computation.

Another option for predicting genes from a new organism is to use a self-trained program GeneMarkS as long as the user can provide at least 100 kbp of sequence on which to train the model.

2) **Glimmer** (Gene Locator and Interpolated Markov Modeler, www.tigr.org/softlab/glimmer /glimmer.html)

is a UNIX program from TIGR that uses the IMM algorithm to predict potential coding regions.

The computation consists of two steps, namely model building and gene prediction. The model building involves training by the input sequence, which optimizes the parameters of the model.

In an actual gene prediction, the overlapping frames are "flagged" to alert the user for further inspection.

Glimmer also has a variant, GlimmerM, for eukaryotic gene prediction.

3) **FGENESB** (www.softberry.com/berry.phtml?topic=gfindb)

is a web-based program that is also based on fifth-order HMMs for detecting coding regions.

The program is specifically trained for bacterial sequences.

It uses the Vertibi algorithm to find an optimal match for the query sequence with the intrinsic model.

4) **RBSfinder** (ftp://ftp.tigr.org/pub/software/RBSfinder/)

Is a UNIX program that uses the prediction output from Glimmer and searches for the Shine–Delgarno sequences in the vicinity of predicted start sites.

If a high-scoring site is found by the intrinsic probabilistic model, a start codon is

confirmed; otherwise the program moves to otherputative translation start sites and

repeats the process.

GENE PREDICTION IN EUKARYOTES

- Eukaryotic nuclear genomes are much larger than prokaryotic ones, with sizes ranging from 10 Mbp to 670 Gbp (1 Gbp = 109 bp).

- They tend to have a very low gene density.

Most importantly, eukaryotic genomes are characterized by a mosaic organization in which a gene is split into pieces (called *exons*) by intervening noncoding sequences (called *introns*)

- The nascent transcript from a eukaryotic gene is modified in three different ways before becoming a mature mRNA for protein translation.

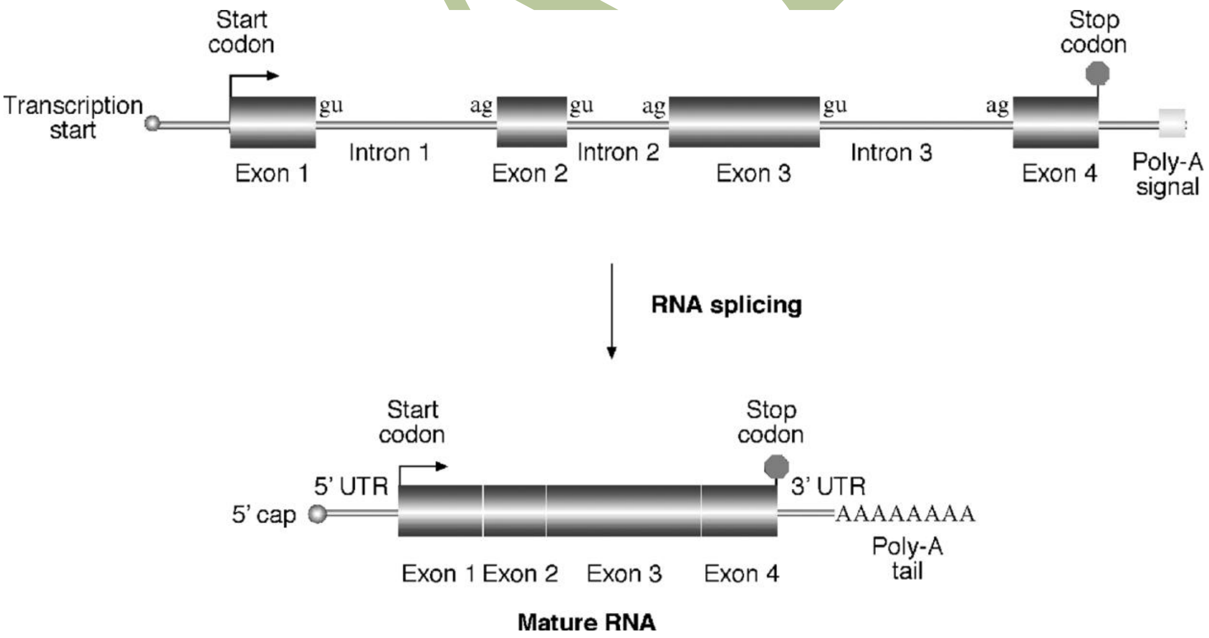
1. The first is capping at the 5' end of the transcript, which involves methylation at the initial residue of the RNA.

2. The second event is splicing, which is the process of removing introns and joining exons. The splicing process involves a large RNA-protein complex called spliceosome. The reaction requires intermolecular interactions between a pair of nucleotides at each end of an intron and the RNA component of the spliceosome.

3. The third modification is polyadenylation, which is the addition of a stretch of As (~250) at the 3' end of the RNA. This process is controlled by a poly-A signal, a conserved motif slightly downstream of a coding region with a consensus CAATAAA(T/C).

Structure of a typical eukaryotic RNA as primary transcript from genomic DNA and as mature RNA after posttranscriptional processing.

Abbreviations: UTR, untranslated region; poly-A, polyadenylation.



Protein structural bioinformatics

Protein Structure Basics

Proteins perform most essential biological and chemical functions in a cell.

They play important roles in structural,enzymatic, transport, and regulatory functions.

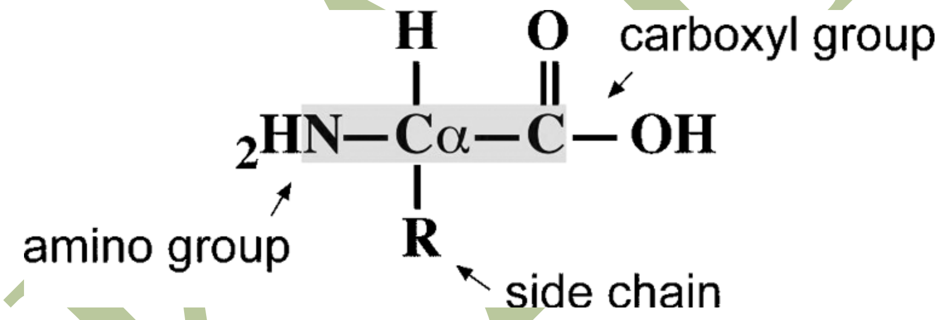
The protein functions are strictly determined by their structures.

Therefore, protein structural bioinformatics is an essential element of bioinformatics.

AMINO ACIDS

The building blocks of proteins are twenty naturally occurring amino acids, small molecules that contain a free amino group (NH2) and a free carboxyl group (COOH).

Both of these groups are linked to a central carbon (C*α*), which is attached to a hydrogen and a side chain group (R) (Fig. 12.1). Amino acids differ only by the side chain R group.



The chemical reactivities of the R groups determine the specific properties of the amino acids.

Amino acids can be grouped into several categories based on the chemical and physical properties of the side chains, such as size and affinity for water.

According to these properties, the side chain groups can be divided into small, large, hydrophobic, and hydrophilic categories.

Within the hydrophobic set of amino acids, they can be further divided into aliphatic and aromatic.

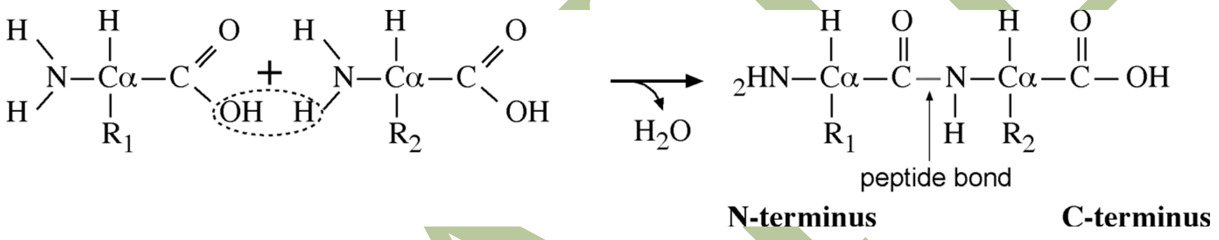
Aliphatic side chains are linear hydrocarbon chains and *aromatic side chains* are cyclic rings.

Within the hydrophilic set, amino acids can be subdivided into polar and charged.

Charged amino acids can be either positively charged (basic) or negatively charged (acidic).

PEPTIDE FORMATION

The peptide formation involves two amino acids covalently joined together between the carboxyl group of one amino acid and the amino group of another (shown in Figure).



This reaction is a condensation reaction involving removal of elements of water from the two molecules. The resulting product is called a *dipeptide.*

The newly formed covalent bond connecting the two amino acids is called a *peptide bond.* Once an amino acid is incorporated into a peptide, it becomes an amino acid residue. Multiple amino acids can be joined together to forma longer chain of amino acid polymer.

A linear polymer of more than fifty amino acid residues is referred to as a *polypeptide.*

A polypeptide, also called a protein, has a well-defined three-dimensional arrangement.

On the other hand, a polymer with fewer than fifty residues is usually called a peptide without a well-defined three-dimensional structure.

The residues in a peptide or polypeptide are numbered beginning with the residue containing the amino group, referred to as the *N*-terminus, and ending with the residue containing the carboxyl group, known as the *C*-terminus.

DIHEDRAL ANGLES

A peptide bond is actually a partial double bond owing to shared electrons between O=C–N atoms.

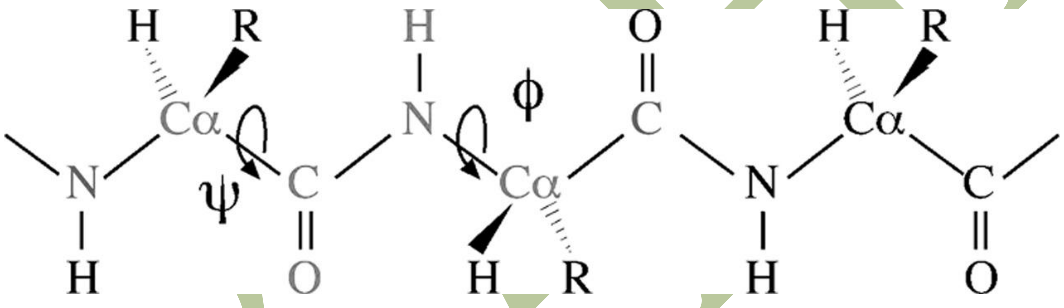
The rigid double bond structure forces atoms associated with the peptide bond to lie in the same plane, called the *peptide plane*.

Because of the planar nature of the peptide bond and the size of the R groups, there are considerable restrictions on the rotational freedom by the two bonded pairs of atoms around the peptide bond.

The angle of rotation about the bond is referred to as the *dihedral angle* (also called the *tortional angle*).

For a peptide unit, the atoms linked to the peptide bond can be moved to a certain extent by the rotation of two bonds flanking the peptide bond.

This is measured by two dihedral angles (as shown in Figure).



One is the dihedral angle along the N–C*α* bond, which is defined as phi (*φ*); and the other is the angle along the C*α*–C bond, which is called psi (*ψ*).

Various combinations of *φ* and *ψ* angles allow the proteins to fold in many different ways.

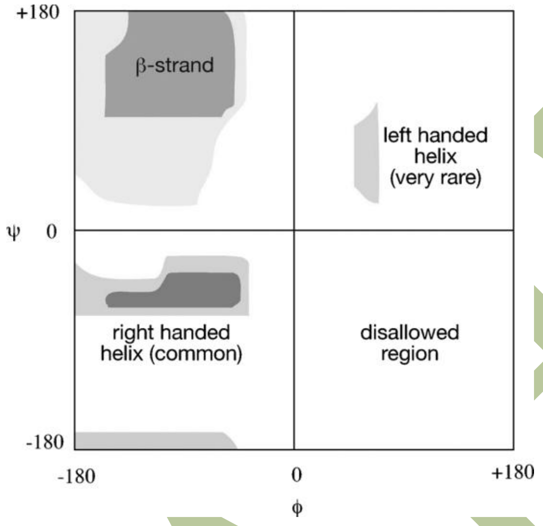
Ramachandran Plot

The rotation of *φ* and *ψ* is not completely free because of the planar nature of the peptide bond and the steric hindrance from the side chain R group.

Consequently, there is only a limited range of peptide conformation.

When φ and ψ angles of amino acids of a particular protein are plotted against each other, the resulting diagram is called a Ramachandran plot.

This plot maps the entire conformational space of a peptide and shows sterically allowed and disallowed regions (as shown in Figure).



It can be very useful in evaluating the quality of protein models.

Protein structures

can be organized into four levels of hierarchies with increasing complexity.

These levels are

[1] primary structure,

[2] secondary structure,

[3] tertiary structure,

[4] quaternary structure.

(1) Primary structure:

A linear amino acid sequence of a protein is the primary structure.

This is the simplest level with amino acid residues linked together through peptide bonds.

(2) Secondary structure:

defined as the local conformation of a peptide chain.

The secondary structure is characterized by highly regular and repeated arrangement of amino acid residues stabilized by hydrogen bondsbetween main chain atoms of the C=Ogroup and the NH group of different residues.

(3) Tertiary structure:

which is the three dimensional arrangement of various secondary structural elements and connecting regions.

The tertiary structure can be described as the complete three-dimensional assembly of all amino acids of a single polypeptide chain.

(4) Quaternary structure:

which refers to the association of several polypeptide chains into a protein complex, which is maintained by noncovalent interactions.

In such a complex, individual polypeptide chains are called *monomers* or *subunits*. Intermediate between secondary and tertiary structures, a level of supersecondary structure is

often used, which is defined as two or three secondary structural elements forming a unique functional domain, a recurring structural pattern conserved in evolution.

SECONDARY STRUCTURES

As mentioned, local structures of a protein with regular conformations are known as secondary structures.

They are stabilized by hydrogen bonds formed between carbonyl oxygen and amino hydrogen of different amino acids.

Chief elements ofsecondary structures are α-helices and β-sheets.

Useful rules in guiding the prediction of protein secondary structures.

For α-Helices

- ✓ An α-helix has a main chain backbone conformation that resembles a cork screw.
- ✓ Nearly all known α-helices are right handed, exhibiting a rightward spiral form.

- ✓ In such a helix, there are 3.6 amino acids per helical turn.
- ✓ The structure is stabilized by hydrogen bonds formed between the main chain atoms of residues *i* and *i + 4*.
- ✓ The hydrogen bonds are nearly parallel with the helical axis.
- ✓ The average *φ* and *ψ* angles are 60◦ and 45◦, respectively, and are distributed in a narrowly defined region in the lower left region of a Ramachandran plot.
- ✓ Hydrophobic residues of the helix tend to face inside and hydrophilic residues of the helix face outside. Thus, every third residue along the helix tends to be a hydrophobic residue. Ala, Gln, Leu, and Met are commonly found in an *α*-helix, but not Pro, Gly, and Tyr.

For *β*-Sheets

- ✓ A *β*-sheet is a fully extended configuration built up from several spatially adjacent regions of a polypeptide chain.
- ✓ Each region involved in forming the *β*-sheet is a *β*-strand.
- ✓ The *β*-strand conformation is pleated with main chain backbone zigzagging and side chains positioned alternately on opposite sides of the sheet.
- ✓ *β*-Strands are stabilizedby hydrogen bonds between residues of adjacent strands.
- ✓ *β*-strands near the surface of the protein tend to show an alternating pattern of hydrophobic and hydrophilic regions, whereas strands buried at the core of a protein are nearly all hydrophobic.
- ✓ The *β*-strands can run in the same direction to form a parallel sheet or can run every other chain in reverse orientation to form an antiparallel sheet, or a mixture of both.
- ✓ The hydrogen bonding patterns are different in each configurations.
- ✓ The *φ* and *ψ* angles are also widely distributed in the upper left region in a Ramachandran plot. Because of the long-range nature of residuesinvolved in this type of conformation, it is more difficult to predict *β*-sheets than *α*- helices.

Coils and Loops

- ✓ There are also local structures that do not belong to regular secondary structures (α-helices and β-strands).
- ✓ The irregular structures are coils or loops.
- ✓ The loops are often characterized by sharp turns or hairpin-like structures.
- ✓ If the connecting regions are completely irregular, they belong to random coils.
- ✓ Residues in the loop or coil regions tend to be charged and polar and located on the surface of the protein structure.
- ✓ They are often the evolutionarily variable regions where mutations, deletions, and insertions frequently occur.
- ✓ They can be functionally significant because these locations are often the active sites of proteins.

Coiled Coils

- ✓ Coiled coils are a special type of super secondary structure characterized by a bundle of two or more α-helices wrapping around each other.
- ✓ The helices forming coiled coils have a unique pattern of hydrophobicity, which repeats every seven residues (five hydrophobic and two hydrophilic).

Reference:

Jin Xiong (2006) Essential Bioinformatics, Cambridge University Press.**(page – 173 to 180).**

Protein Secondary Structure Prediction (page 200 to 212)

- Protein secondary structures are stable local conformations of a polypeptide chain.

- They are critically important in maintaining a protein three-dimensional structure.

- The highly regular and repeated structural elements include α-helices and β-sheets.

- It has been estimated that nearly 50% of residues of a protein fold into either α-helices and β-strands.

Protein secondary structure prediction

refers to the prediction of the conformational state of each amino acid residue of a protein sequence as one of the three possible states, namely, helices, strands, or coils, denoted as H, E, and C, respectively.

The prediction is based on the fact that secondary structures have a regular arrangement of amino acids, stabilized by hydrogen bonding patterns.

The structural regularity serves the foundation for prediction algorithms.

Applications of predicting protein secondary structures:

- It can be useful for the classification of proteins and for the separation of protein domains and functional motifs.
- Secondary structures are much more conserved than sequences during evolution. As a result, correctly identifying secondary structure elements (SSE) can help to guide sequence alignment or improve existing sequence alignment of distantly related sequences.
- In addition, secondary structure prediction is an intermediate step in tertiary structure prediction as in threading analysis.

SECONDARY STRUCTURE PREDICTION FOR GLOBULAR PROTEINS

The formation of α-helices is determined by short-range interactions, whereas the formation of β-strands is strongly influenced by long-range interactions.

Prediction for long-range interactions is theoretically difficult.

After more than three decades of effort, prediction accuracies have only been improved from about 50% to about 75%.

The secondary structure prediction methods can be either

(1) ab initio based- which make use of single sequence information only,

(2) homology based - which make use of multiple sequence alignment information.

Ab initio methods,

which belong to early generation methods, predict secondary structures based on statistical calculations of the residues of a single query sequence.

Homology-based methods

do not rely on statistics of residues of a single sequence, but on common secondary structural patterns conserved among multiple homologous sequences.

Ab Initio–Based Methods

This type of method predicts the secondary structure based on a single query sequence.

It measures the relative propensity of each amino acid belonging to a certain secondary structure element.

The propensity scores are derived from known crystal structures.

Examples of ab initio prediction are the

(i) Chou–Fasman algorithum

(ii) Garnier, Osguthorpe, Robson (GOR) methods.

The ab initio methods were developed in the 1970s when protein structural data were very limited.

(i) Chou–Fasman algorithm (http://fasta.bioch.virginia.edu/fasta/chofas.htm)

determines the propensity or intrinsic tendency of each residue to be in the helix, strand, and β-turn conformation using observed frequencies found in protein crystal structures (conformational values for coils are not considered).

(ii) The GOR method (http://fasta.bioch.virginia.edu/fasta www/garnier.htm)

is also based on the “propensity” of each residue to be in one of the four conformational states, helix (H), strand(E), turn(T),andcoil (C).

However, instead of using the propensity value from a single residue to predict a conformational state, it takes short-range interactions of neighboring residues into account.

It examines a window of every seventeen residues and sums up propensity scores for all residues for each of the four states resulting in four summed values.

The highest scored state defines the conformational state for the center residue in the window (ninth position).

The GOR method has been shown to be more accurate than Chou–Fasman because it takes the neighboring effect of residues into consideration.

Homology-Based Methods

The third generation of algorithms were developed in the late 1990s by making use of evolutionary information.

This type of method combines the ab initio secondary structure prediction of individual sequences and alignment information from multiple homologous sequences (>35% identity).

The idea behind this approach is that close protein homologs should adopt the same secondary and tertiary structure.

When each individual sequence is predicted for secondary structure using a method similar to the GOR method, errors and variations may occur.

However, evolutionary conservation dictates that there should be no major variations for their secondary structure elements.

Therefore, by aligning multiple sequences, information of positional conservation is revealed. Because residues in the same aligned position are assumed to have the same secondary structure, any inconsistencies or errors in prediction of individual sequences can be corrected using a majority rule. This homology based method has helped improve the prediction accuracy by another 10% over the second-generation methods.

The following lists several frequently used third generation prediction algorithms available as web servers.

PHD (Profile network from Heidelberg; http://dodo.bioc.columbia.edu/predictprotein/submit def.html)

- is a web-based program that combines neural network with multiple sequence alignment.
- It first performs a BLASTP of the query sequence against a nonredundant protein sequence database to find a set of homologous sequences, which are aligned with the MAXHOM program (a weighted dynamic programming algorithm performing global alignment).

PSIPRED (http://bioinf.cs.ucl.ac.uk/psiform.html)

- is a web-based program that predicts protein secondary structures using a combination of evolutionary information and neural networks.
- The multiple sequence alignment is derived from a PSI-BLAST database search.
- A profile is extracted from the multiple sequence alignment generated from three rounds of automated PSI-BLAST.

SSpro (http://promoter.ics.uci.edu/BRNN-PRED/)

- is a web-based program thatcombines PSI-BLAST profiles with an advanced neural network, known as *bidirectional recurrent neural networks* (BRNNs).

PROF (Protein forecasting; www.aber.ac.uk/~phiwww/prof/)

- is an algorithm that combines PSI-BLAST profiles and a multistaged neural network, similar to that in PHD.
- In addition, it uses a linear discriminant function to discriminate between the three states.

HMMSTR (Hidden Markov model [HMM] for protein STRuctures; www.bioinfo.rpi.edu/~bystrc/hmmstr/server.php)

- uses a branched and cyclic HMM to predict secondary structures.
- It first breaks down the query sequence into many very short segments (three to nine residues, called I-sites) and builds profiles based on a library of known structure motifs.
- It then assembles these local motifs into a super secondary structure.

- It further uses an HMM with a unique topology linking many smaller HMMs into a highly branched multicyclic form

Jpred (www.compbio.dundee.ac.uk/~www-jpred/)

- combines the analysis results from six prediction algorithms, including PHD, PREDATOR, DSC, NNSSP, Jnet, and ZPred.
- The query sequence is first used to search databases with PSI-BLAST for three iterations. Redundant sequence hits are removed.
- The resulting sequence homologs are used to build a multiple alignment from which a profile is extracted.
- The profile information is submitted to the six prediction programs.
- If there is sufficient agreement among the prediction programs, the majority of the prediction is taken as the structure.

PredictProtein (www.embl-heidelberg.de/predictprotein/predictprotein.html)

- Is another multiple prediction server that uses Jpred, PHD, PROF, and PSIPRED, among others.
- The difference is that the server does not run the individual programs but sends the query to other servers which e-mail the results to the user separately.
- It does not generate a consensus.
- It is up to the user to combine multiple prediction results and derive a consensus.

Protein Tertiary Structure Prediction

Structural prediction is a powerful tool to understand the functions of biological macromolecules at the atomic level.

DNA structure, a double helix, is invariable regardless of sequence variations.

Necessary for protein structure prediction:

- protein structures vary depending on the sequences.
- much slower rate of structure determination by x-ray crystallography or NMR spectroscopy compared to gene sequence generation from genomic studies.
- Consequently, the gap between protein sequence information and protein structural information is increasing rapidly. Protein structure prediction aims to reduce this sequence–structure gap.
- In contrast to sequencing techniques, experimental methods to determine protein structures are time consuming and limited.Currently, it takes 1 to 3 years to solve a protein structure.
- Certain proteins, especially membrane proteins, are extremely difficult to solve by x-ray or NMR techniques.
- There are many important proteins for which the sequence information is available, but their three-dimensional structures remain unknown.
- The full understanding of the biological roles of these proteins requires knowledge of their structures.

Therefore, it is often necessary to obtain approximate protein structures through computer modeling.

<u>Methods of protein three-dimensional structure prediction:</u>

There are three computational approaches to protein three-dimensional structural modeling and prediction.

- Homology modeling - knowledge-based methods
- Threading - knowledge-based methods
- Ab initio prediction.

In Knowledge-based methods - predict protein structures based on knowledge of existing protein structural information in databases.

Homology modeling

builds an atomic model based on an experimentally determined structure that is closely related at the sequence level.

Threading

identifies proteins that are structurally similar, with or without detectable sequence similarities.

Ab initio approach

is simulation based and predicts structures based on physicochemical principles governing protein folding without the use of structural templates.

Homology modelling or comparative modelling

Homology modeling predicts protein structures based on sequence homology with known structures. It is also known as *comparative modeling*.

Principle

if two proteins share a high enough sequence similarity, they are likely to have very similar three-dimensional structures. If one of the protein sequences has a known structure, then the structure can be copied to the unknown protein with a high degree of confidence.

Homology modeling produces an all-atom model based on alignment with template proteins.

Homology modeling procedure consists of six steps:

First step - is template selection, which involves identification of homologous sequences in the protein structure database to be used as templates for modeling.

Second step - is alignment of the target and template sequences.

Third step - is to build a framework structure for the target protein consisting of main chain atoms.

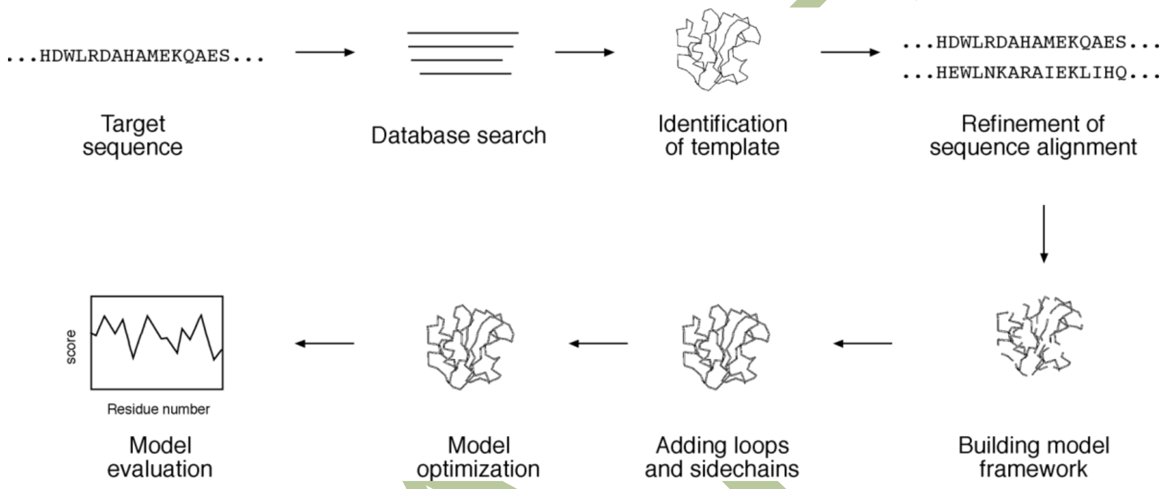
Fourth step - of model building includes the addition and optimization of side chain atoms and loops.

Fifth step - is to refine and optimize the entire model according to energy criteria.

Sixth step - involves evaluating of the overall quality of the model obtained.

If necessary, alignment and model building are repeated until a satisfactory result is obtained.

Flow chart showing steps involved in homology modelling:



Template Selection

- The first step in protein structural modeling is to select appropriate structural templates.
- This forms the foundation for rest of the modeling process.
- The template selection involves searching the Protein Data Bank (PDB) for homologous proteins with determined structures.
- The search can be performed using a heuristic pair-wise alignment search program such as BLAST or FASTA.

As a rule of thumb,

a database protein should have at least 30% sequence identity with the query sequence to be selected as template.

Thus it is recommended that the structure(s) with the highest percentage identity, highest resolution, and the most appropriate cofactors is selected as a template.

If, no highly similar sequences can be found in the structure database,

- either a more sensitive profile-based PSI-BLAST method or
- a fold recognition method such threading can be used to identify distant homologs.

Modeling can therefore only be done with the aligned domains of the target protein.

Sequence Alignment

Once the structure with the highest sequence similarity is identified as a template,

- the full-length sequences of the template and target proteins need to be realigned using refined alignment algorithms to obtain optimal alignment.
- This realignment is the most critical step in homology modeling, which directly affects the quality of the final model.
- Errors made in the alignment step cannot be corrected in the following modeling steps.
- Therefore, the best possible multiple alignment algorithms, such as Praline and T-Coffee should be used for this purpose.
- Even alignment using the best alignment program may not be error free and should be visually inspected to ensure that conserved key residues are correctly aligned.
- If necessary, manual refinement of the alignment should be carried out to improve alignment quality.

Backbone Model Building

Once optimal alignment is achieved,

residues in the aligned regions of the target protein can assume a similar structure as the template proteins, meaning that the coordinates of the corresponding residues of the template proteins can be simply copied onto the target protein.

- If the two aligned residues are identical, coordinates of the side chain atoms are copied along with the main chain atoms.
- If the two residues differ, only the backbone atoms can be copied.
- The side chain atoms are rebuilt in a subsequent procedure.

- In backbone modeling, it is simplest to use only one template structure.
- As mentioned, the structure with the best quality and highest resolution is normally chosen if multiple options are available. This structure tends to carry the fewest errors.
- Occasionally, multiple template structures are available for modeling.
- In this situation, the template structures have to be optimally aligned and superimposed before being used as templates in model building.
- One can either choose to use average coordinate values of the templates or the best parts from each of the templates to model.

Loop Modeling

- In the sequence alignment for modeling, there are often regions caused by insertions and deletions producing gaps in sequence alignment.
- The gaps cannot be directly modeled, creating “holes” in the model.
- Closing the gaps requires loop modeling, which is a very difficult problem in homology modeling and is also a major source of error.
- Loop modeling can be considered a mini–protein modeling problem by itself.
- Unfortunately, there are no methods available that can model loops reliably.

Currently, there are **two main techniques** used to approach the problem:

(1) database searching method and (2) *ab initio* method.

(1) **database searching method**

The best loop can be selected based on sequence similarity as well as minimal steric clashes with the neighboring parts of the structure. The conformation of the best matching fragments is then copied onto the anchoring points of the stems.

(2) ***ab initio* method**

The ab initio method generates many random loops and searches for the one that does

not clash with nearby side chains and also has reasonably low energy and φ and ψ

angles in the allowable regions in the Ramachandran plot.

If the loops are relatively short (three to five residues), reasonably correct models can be built using either of the two methods. If the loops are longer, it is very difficult to achieve a reliable model.

The following are specialized programs for loop modeling:

FREAD (www-cryst.bioc.cam.ac.uk/cgi-bin/coda/fread.cgi) is a web server that models loops using the database approach.

PETRA (www-cryst.bioc.cam.ac.uk/cgi-bin/coda/pet.cgi) is a web server that uses the ab initio method to model loops.

CODA (www-cryst.bioc.cam.ac.uk/∼charlotte/Coda/search coda.html) is a web server that uses a consensus method based on the prediction results from FREAD and PETRA.

For loops of three to eight residues, it uses consensus conformation of both methods and for nine to thirty residues, it uses FREAD prediction only.

Side Chain Refinement

- Once main chain atoms are built, the positions of side chains that are not modeled must be determined.
- Modeling side chain geometry is very important in evaluating protein–ligand interactions at active sites and protein–protein interactions at the contact interface.
- A side chain can be built by searching every possible conformation at every torsion angle of the side chain to select the one that has the lowest interaction energy with neighboring atoms. Most current side chain prediction programs use the concept of *rotamers*, which are favored side chain torsion angles extracted from known protein crystal structures.
- A collection of preferred side chain conformations is a rotamer library in which the rotamers are ranked by their frequency of occurrence.
- Having a rotamer library reduces the computational time significantly because only a small number of favored torsion angles are examined.
- In prediction of side chain conformation, only the possible rotamers with the lowest interaction energy with nearby atoms are selected.

- In many cases, even applying the rotamer library for every residue can be computationally too expensive.
- To reduce search time further, backbone conformation can be taken into account.
- It has been observed that there is a correlation of backbone conformations with certain rotamers.
- By using such correlations, many possible rotamers can be eliminated and the speed of conformational search can be much improved.
- After adding the most frequently occurring rotamers, the conformations have to be further optimized to minimize steric overlaps with the rest of the model structure.
- Most modeling packages incorporate the side chain refinement function.
- A specialized side chain modeling program that has reasonably good performance is SCWRL (sidechain placement with a rotamer library; www.fccc.edu/research/labs/dunbrack/scwrl/),
- It removes rotamers that have steric clashes with main chain atoms.
- The final, selected set of rotamers has minimal clashes with main chain atoms and other side chains.

Model Evaluation

- The final homology model has to be evaluated to make sure that the structural features of the model are consistent with the physicochemical rules.
- This involves checking anomalies in φ–ψ angles, bond lengths, close contacts.
- Another way of checking the quality of a protein model is to implicitly take these stereochemical properties into account.
- This is a method that detects errors by compiling statistical profiles of spatial features and interaction energy from experimentally determined structures.
- By comparing the statistical parameters with the constructed model, the method reveals which regions of a sequence appear to be folded normally and which regions do not. If

structural irregularities are found, the region is considered to have errors and has to be further refined.

Procheck (www.biochem.ucl.ac.uk/~roman/procheck/procheck.html) is a UNIX program that is able to check general physicochemical parameters such as φ–ψ angles, chirality, bond lengths, bond angles.

The parameters of the model are used to compare with those compiled from well-defined, high-resolution structures.

If the program detects unusual features, it highlights the regions that should be checked or refined further.

WHAT IF (www.cmbi.kun.nl:1100/WHATIF) is a comprehensive protein analysis server that validates a protein model for chemical correctness. It has many functions, including checking of planarity, collisions with symmetry axes (close contacts), proline puckering, anomalous bond angles, and bond lengths. It also allows the generation of Ramachandran plots as an assessment of the quality of the model.

ANOLEA (Atomic Non-Local Environment Assessment; http://protein.bio.puc.cl/cardex/servers/anolea/index.html) is a web server that uses the statistical evaluation approach. It performs energy calculations for atomic interactions in a protein chain and compares these interaction energy values with those compiled from a database of protein x-ray structures.

Verify3D (www.doe-mbi.ucla.edu/Services/Verify 3D/) is another server using the statistical approach. It uses a precomputed database containing eighteen environmental profiles based on secondary structures and solvent exposure, compiled from high-resolution protein structures.

Comprehensive Modeling Programs

A number of comprehensive modeling programs are able to perform the complete procedure of homology modeling in an automated fashion.

Some freely available protein modeling programs and servers are listed.

Modeller (http://bioserv.cbs.cnrs.fr/HTML BIO/frame mod.html) is a web server for homology modeling.

Swiss-Model (www.expasy.ch/swissmod/SWISS-MODEL.html) is an automated modeling server that allows a user to submit a sequence and to get back a structure automatically.

3D-JIGSAW (www.bmm.icnet.uk/servers/3djigsaw/) is a modeling server that works in either the automatic mode or the interactive mode. Its loop modeling relies on the database method.

Homology Model Databases

The availability of automated modeling algorithms has allowed several research groups to use the fully automated procedure to carry out large-scale modeling projects.

ModBase (http://alto.compbio.ucsf.edu/modbase-cgi/index.cgi) is a database of protein models generated by the Modeller program. For most sequences that have been modeled, only partial sequences or domains that share strong similarities with templates are actually modeled.

3Dcrunch (www.expasy.ch/swissmod/SWISS-MODEL.html) is another database archiving results of large-scale homology modeling projects. Models of partial sequences from the Swiss-Prot database are derived using the Swiss-Model program.

THREADING AND FOLD RECOGNITION

threading or *structural fold recognition* predicts the structural fold of an unknown protein sequence by fitting the sequence into a structural database and selecting the best-fitting fold. The comparison emphasizes matching of secondary structures, which are most evolutionarily conserved. Therefore, this approach can identify structurally similar proteins even without detectable sequence similarity.

The algorithms can be classified into **two categories**,

pairwise energy based and profile based.

The pairwise energy–based method was originally referred to as *threading* and the profile-based method was originally defined as *fold recognition*.

Pairwise Energy Method

In the pairwise energy based method, a protein sequence is searched for in a structural fold database to find the best matching structural fold using energy-based criteria.

Profile Method

In the profile-based method, a profile is constructed for a group of related protein structures. The structural profile is generated by superimposition of the structures to expose corresponding residues. Statistical information from these aligned residues is then used to construct a profile.

The profile scores contain information for secondary structural types, the degree of solvent exposure, polarity, and hydrophobicity of the amino acids.

To predict the structural fold of an unknown query sequence, the query sequence is first predicted for its secondary structure, solvent accessibility, and polarity. The predicted information is then used for comparison with propensity profiles of known structural folds to find the fold that best represents the predicted profile.

3D-PSSM (www.bmm.icnet.uk/∼3dpssm/) is a web-based program that employs the structural profile method to identify protein folds.

GenThreader (http://bioinf.cs.ucl.ac.uk/psipred/index.html) is a web-based program that uses a hybrid of the profile and pairwise energy methods.

Fugue (www-cryst.bioc.cam.ac.uk/∼fugue/prfsearch.html) is a profile-based fold recognition server. It has precomputed structural profiles compiled from multiple alignments of homologous structures, which take into account local structural environment such as secondary structure, solvent accessibility, and hydrogen bonding status.

AB INITIO PROTEIN STRUCTURAL PREDICTION

- Both homology and fold recognition approaches rely on the availability of template structures in the database to achieve predictions.
- If no correct structures exist in the database, the methods fail.
- However, proteins in nature fold on their own without checking what the structures of their homologs are in databases.
- Obviously, there is
- some information in the sequences that provides instruction for the proteins to "find" their native structures. Early biophysical studies have shown that most proteins fold spontaneously into a stable structure that has near minimum energy. This structural state is

called the *native state*. This folding process appears to be non random; however, its mechanism is poorly understood.

- The limited knowledge of protein folding forms the basis of ab initio prediction.
- The ab initio prediction method attempts to produce all-atom protein models based on sequence information alone without the aid of known protein structures.

The following web program is such an example using this approach:

Rosetta (www.bioinfo.rpi.edu/~bystrc/hmmstr/server.php) is a web server that predicts protein three-dimensional conformations using the ab initio method. This relies on a “mini-threading” method. The method first breaks down the query sequence into many very short segments (three to nine residues) and predicts the secondary structure of the small segments using a hidden Markov model–based program, HMMSTR.

Protein Structure Visualization:

- Once a protein structure has been solved, the structure has to be presented in a three dimensional view on the basis of the solved Cartesian coordinates.
- Before computer visualization software was developed, molecular structures were represented by physical models of metal wires, rods, and spheres.
- With the development of computer hardware and software technology, sophisticated computer graphics programs have been developed for visualizing and manipulating complicated three-dimensional structures.
- The computer graphics help to analyze and compare protein structures to gain insight to functions of the proteins.

The main feature of computer visualization programs is

- interactivity, which allows users to visually manipulate the structural images through a graphical user interface.

At the touch of a mouse button, a user can move, rotate, and zoom an atomic model on a computer screen in real time, or examine any portion of the structure in great detail, as well as draw it in various formsin different colors.

- Further manipulations can include changing the conformation of a structure by protein modeling or matching a ligand to an enzyme active site through docking exercises.

- The visualization program should also be able to produce molecular structures in different styles, which include wire frames, balls and sticks, space-filling spheres, and ribbons.

Wire-frame diagram

is a line drawing representing bonds between atoms.

The wire frame is the simplest form of model representation and is useful for localizing positions of specific residues in a protein structure, or for displaying a skeletal form of a structure when Cα atoms of each residue are connected.

Balls and sticks

are solid spheres and rods, representing atoms and bonds, respectively.

These diagrams can also be used to represent the backbone of a structure.

Space-filling representation

each atom is described using large solid spheres with radii corresponding to the van der Waals radii of the atoms.

Ribbon diagrams use cylinders or spiral ribbons to represent α-helices and broad, flat arrows to represent β-strands. This type of representation is very attractive in that it allows easy identification of secondary structure elements and gives a clear view of the overall topology of the structure.

Some widely used and freely available software programs for molecular graphics:

RasMol (http://rutgers.rcsb.org/pdb/help-graphics.html#rasmol download)

- is a command-line–based viewing program that calculates connectivity of a coordinate file and displays wireframe, cylinder, stick bonds, α-carbon trace, space-filling (CPK) spheres, and ribbons.
- It reads both PDB and mmCIF formats and can display a whole molecule or specific parts of it.
- It is available in multiple platforms: UNIX, Windows, and Mac.

RasTop (www.geneinfinity.org/rastop/)

- is a new version of RasMol forWindows with a more enhanced user interface.

Swiss-PDBViewer (www.expasy.ch/spdbv/)

- is a structure viewer for multiple platforms.
- It is essentially a Swiss-Army knife for structure visualization and modeling because it incorporates so many functions in a small shareware program.
- It is capable of structure visualization, analysis, and homology modeling.
- It allows display of multiple structures at the same time in different styles, by charge distribution, or by surface accessibility.
- It can measure distances, angles, and even mutate residues.
- In addition, it can calculate molecular surface, electrostatic potential, Ramachandran plot, and so on.
- The homology modeling part includes energy minimization and loop modeling.

Molscript (www.avatar.se/molscript/)

- is a UNIX program capable of generating wire-frame, space-filling, or ball-and-stick styles. In particular, secondary structure elements can be drawn with solid spirals and arrows representing α-helices and β-strands, respectively.

Grasp (http://trantor.bioc.columbia.edu/grasp/)

- is a UNIX program that generates solid molecular surface images and uses a gradated coloring scheme to display electrostatic charges on the surface.

WebMol (www.cmpharm.ucsf.edu/cgi-bin/webmol.pl)

- is a web-based program built based on a modified RasMol code and thus shares many similarities with RasMol.
- It runs directly on a browser of any type as an applet and is able to display simple line drawing models of protein structures.

- It also has a feature of interactively displaying Ramachandran plots for structure model evaluation.

Chime (www.mdlchime.com/chime/)

- Is a plug-in for web browsers; it is not a stand alone program and has to be invoked in a web browser.
- The program is also derived from RasMol and allows interactive display of graphics of protein structures inside a web browser.

Cn3D (www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml)

- is a helper application for web browsers to display structures in the MMDB format from the NCBI's structural database.
- It can be used on- or offline as a stand-alone program.
- It is able to render three-dimensional molecular models and display secondary structure cartoons.
- The drawback is that it does not recognize the PDB format.

Review Questions:

Short Answer Questions **(2 Marks)**

1. Define gene prediction.
2. Define ORF?
3. Define RBS?
4. Draw the structure of an amino acid?
5. What are the levels of structural organization of proteins?

6. What are the methods of protein tertiary structure prediction?
7. Define threading?
8. What are the tools used for visualizing a protein structure?
9. Define homology modeling?
10. Define propensity?
11. Name any four hydrophobic amino acids?
12. Name any four hydrophilic amino acids?
13. Name four amino acids that forms α-helix?
14. Name four amino acids that form β-sheets?
15. What are coils?
16. Define quarternary structure of a protein?
17. Write any two differences between prokaryotic and eukaryotic gene?
18. Write the three stop codons?
19. How can you identify a splice junction in a eukaryotic gene?
20. How is transcription terminated in a prokaryotic gene?

Essay Answer Questions: **(6 & 8 Marks)**

1. Describe a prokaryotic gene structure with diagram?
2. Describe a eukaryotic gene structure with diagram?
3. Write notes on secondary structure of a protein?
4. Describe protein secondary structure prediction methods.
5. How is tertiary structure of a protein predicted using homology based methods?
6. Describe threading method of tertiary structure prediction?
7. Describe some protein visualization tools

Further Readings:

Jin Xiong (2006) Essential Bioinformatics, Cambridge University Press.

Applications of bioinformatics – en.wikipedia.org/wiki/**Bioinformatics**

Attwood TK and Parry- Smith DJ (2006) Introduction to Bioinformatics. Pearson Education Ltd.

Question	Opt 1	Opt 2	Opt 3	Opt 4	Answer
Unit I					
The term bioinformatics was coined by	paulien Hogeweg	Ben Hesper	Cris Nolan	paulien Hogeweg and	paulien Hogeweg and Ben Hesper
common activity in bioinformatics includes	food preservation	different DNA and proteir	Blotting	medicine	different DNA and protein to compare
which of the following is an sequence anaylsis method	DNA and protein stru	gene and promoter findin	prediction	classification	gene and promoter finding
major research area of bioinformatics includes	drug design	drug discovery	structure prediction	all	all
the first phage X174 was sequenced in the year	1977	1978	1976	1975	#
which sequencing method is used now a days to sequence all genome	chemical degradation	shot gun	chain termination	nano pore sequencin	shot gun
which method is used for marking the gene and biological features in a DNA sequence	Genome annotation	analysis of Gene expressic	computational evoluior	none of the above	Genome annotation
the first genome annotation software was designed in	1996	1995	1994	1998	#
which was first annatated free livin organism genome	bacteria	fungi	virus	animal	virus
The study of origin and descent of species is called as	Genome annotation	analysis of Gene expressic	computational evoluior	none of the above	computational evoluionary biology
microarray is the method used to	analysis of Gene expr	Genome annotation	structure prediction	computational evolui	analysis of Gene expression
which method is used for analysing the gene expression	MPSS	SAGE	EST	DST	EST
bioimformatics is being used in following field such as	bio-weapon creation	vetinary science	insect resistance	all	all
Bioinformatics consists of --------------- subfields?	three	two	five	six	two
Dr Owen White, designed ------------ software system	protein structure	RNA function	genome annotation	expressed cDNA	Genome annotation
measuring of mRNA levels with multiple techniques includes	MPSS	TSTR	EGAS	TSE	MPSS
bioinformatics is the application of stastistics and computer science in the field of	molecular biology	micro biology	biotechnology	biology	molecular biology
bioinformatics is a union of biology with --------	programmers	informatics	physics	biotechnology	Informatics
Margaret Dayhoff developed the first protein sequence database called	SWISS PROT	PDB	Atlas of protein sequen	Protein sequence dat	Atlas of protein sequence and structure
Step wise method for solving problems in computer science is called	flowchart	sequential design	procedure	algorithm	algorithm
The first published completed gene sequence was of	M 13 phage	T 4 phage	ϕ X174	lambda phage	ϕ X174
The term used to refer something ‘performed on computer or computer simulation	dry lab	web lab	invitro	insilico	insilico
‘Laboratory work using chemicals, drugs etc using water’ is referred as	dry lab	web lab	wet lab	insilico	wet lab
‘Laboratory work using computers and computer generated models generally offline’ is referred as	dry lab	web lab	wet lab	insilico	dry lab
NCBI was established during	1988	1989	1990	1991	#
Application of bioinformatics include	data storage and mar	drug designing	understand relationshi	all of the above	all of the above
The computational methodology that tries to find the best matching between two molecule, a receptor an	molecular matching	molecular docking	molecular fitting	molecule affinity che	molecular docking
Proteomics is the study of	set of proteins	set of proteins in a specifi	entire set of expressed	none of these	entire set of expressed proteins in a cell
The process of finding relative location of genes on a chromosome is called as	gene tracing	genome mapping	genome walking	chromosome walking	genome mapping
A compound that has desirable properties to become a drug is called as	lead	find	fit drug	fit compound	lead
Genome annotation software system eas developed by	Luscombe	Sidney Brenner	Robert Brown	Dr. Owen White	Dr. Owen White
Each record in a database is called as	entry	file	record	ticket	entry
Literature databases include	MEDLINE and PubME	MEDLINE and PDB	PubMED and PDB	MEDLINE and PDS	MEDLINE and PubMED
Which of the following is an E.coli model organism database	EcoGene	EcoBase	EcoSeq	ColGene	EcoGene
Which of the following is a protein sequence database	DDBJ	EMBL	GenBank	PIR	PIR
GenBank, the nucleic acid sequence database is maintained by	Brookhaven laborator	DNA database of Japan (D	European Molecular Bi	National Centre for B	National Centre for Biotechnology Information
Submission to GenBank are made using	BankIt and Sequin	BankIt and BankIn	Sequin and BankIn	Entrez	BankIt and Sequin
STAG is maintained by	Brookhaven laborator	DNA database of Japan (D	European Molecular Bi	National Centre for B	DNA database of Japan (DDBJ)
A comprehensive database for the study of human genetics and molecular biology is	PDB	STAG	OMIM	PSD	OMIM
All the following are protein sequence databases except	PIR	PSD	SWISS PROT	EMBL	EMBL
The information retrieval tool of NCBI GenBank is	Entrez	STAG	SeqIn	text search	Entrez
The first secondary database developed was	PRINTS	PROSITE	PDB	PIR	PROSITE
Which of the following is a sequence alignment tool	BLAST	PRINT	PROSITE	PIR	BLAST
MAtDB is a model organism database for	Mouse	Human	E.coli	Arabidopsis	Arabidopsis
The Modern bioinformatics can be classified into………………..broad categories	one	two	three	four	two
Richard Owen elaborated the distinction of homology and analogy during	1843	1976	1875	1876	#
GenBank Release 3 was made public at	1988	1982	1987	1989	#
National Center for Biotechnology Information (NCBI) was developed in the year	1987	1789	1988	1987	#
BLAST: fast sequence similarity searching was developed in which year	1990	1988	1982	1843	#
First bacterial genome was completely sequenced in the year	1879	1995	1965	1995	#
Pauling's theory of molecular evolution was coined in	1995	1988	1845	1962	#
First plant genome sequenced – Arabidopsis was published in______	1990	1988	2000	1843	#
How many steps are involved in genome project	3 step	5 step	1 step	2 step	2 step
Structural annotation consists of the identification of genomic elements to	biochemical function	biological function	expression	d)gene structure	gene structure
A working draft of the human genome was released in	2000	2005	2002	2004	#
Needleman-Wunsch algorithm was developed in-------	1970	1981	1979	1971	#
Human beings have___----------- pairs of chromosomes	22	23	46	44	#
In Shotgun sequencing, all the DNA from an organism is first fractured into millios of ------ pieces	large	medium	small	micro	small
Automated sequencing can read upto ---- bases at a time	900	500	1500	300	#
short oligonucleotide analysis package software was developed by	Biocon	NCBI	BGI	PGI	BGI
A complete draft of the human genome was released during	2004	2003	2000	2005	#
Human genome project revealed that Humans contains approximately ---- genes	20500	21300	25000	20000	#
Unit II					
The unit of genetic information is the___________ or cistron.	Genom	Codon	Gene	Anticodon	Gene
Chromatin consists of a long double stranded ___________molecules	RNA	DNA	Subunit	tDNA	DNA
The mammalian ribosome contains the number of major nucleoprotein subunits.	1	2	4	5	2
Each transfer RNA molecule contains the number of nucleotides _____	J.B.Sumner	Koshland	Menten	Fisher	J.B.Sumner
Gene is a segment of the DNA molecule containing base pairs about	300	400	500	600	#
The sequences recognized by RNA polymerase are called	Terminator	Promoter	Both of the above	Non of the above	Promoter
Optimum temperature for an enzyme-catalyzed reaction is____________	30°C-40°C	25°C-40°C	35°C-45°C	85°C-90°C	25°C-40°C
DNA is refered as ___________	Transforming factor	Range constants	Transplantation factor	Heterogenous factor	Transforming factor
Messenger RNA has a molecuar weight of__________	15000 to 30000	20000 to 35000	25000 to 40000	30000 to 50000	30000 to 50000
The 60S subunit contains 5s rRNA a 5.8S;rRNA and a___________	18S	30S	28S	40S	28S
DNA is denatured by __________	Acid	Cooling	chillness	buffer	acid
A chemical bond formed between two molecules when the carboxyl group of one molecule reacts with the	Hydrophobic interacti	Hydrophilic interaction	Disulphide bonds	Peptide bonds	Peptide bonds
cGMP is antagonistic to __________	cAMP	CTP	ATP	GTP	cAMP
Michaelis-Menten model describes __________	Enzyme stability	Enzyme specificity	Enzyme kinetics	None of the above	Enzyme kinetics
cGMP is formed from __________by the enzyme adenyl cyclase	ATP	GDP	CTP	CDP	ATP
An example for a semipermeable membrane is ___________	Plasma membrane	mitochondria	nuclear membrane	RNA	Plasma membrane
__________ a procedure to remove waste products and excess fluid from the blood when the kidneys stop	Osmosis	Reverse osmosis	Electrophoresis	Dialysis	Dialysis
____________is a functional group consisting of the formula RN=C=NR.	Gluteraldehyde	Carbodiimides	Carbondioxide	Glycoside	Carbodiimides
The lactam form is the predominant tautomer of	Uracil	Cytosine	Adenine	Xanthine	Uracil
The chemical name 2-amino-6-oxypurine is said to be________	Adenine	Xanthine	Guanine	Hypoxinthine	Guanine
The initiation of DNA synthesis requires priming by a short length of	RNA	DNA	Hydroxyl group	Alkyl group	RNA
which of the following enzymes are used to join fragements of DNA?	DNA ligase	b)DNA polymerase	primase	endonuclease	DNA ligase
Semiconservative replication of DNA was confirmed by the which scientist's experiments ?	Meselson and Stalh's	watson and crick	griffith's experiment	Mendels experiment	Meselson and Stalh's experiment
Unwinding of DNA is done by	helicase	ligase	hexonuclease	toposiomerase	helicase
During the replication of DNA, the sythesis of DNA on lagging strand takes place in segments, these segme	Satellite segments	Double helix segments	kornberg segments	Okazaki segments	Okazaki segments
Which of the following nitrogenous bases is found in DNA but is not found in RNA?	Adenine	Thymine	Guanine	Uracil	Uracil
Nucleotides have a nitrogenous base attached to a sugar at the------	1' Carbon	2' Carbon	3' Carbon	5' carbon	5' Carbon
The DNA of a certain organism has been found to have guanine as 30% of its bases. What percentage of its	30%	60%	20%	40%	#
The Process of DNA replication involves:	Multiple origins of re	Binding of ribosomes to o	Continous synthesis on	Conservative replciat	Multiple origins of replication per chromosme in eukaryotes
RNA is less stable than DNA because	RNAses are ubiquitou	RNAses are extremely he	denatured RNAses reac	d)RNA is chemically s	RNAses are ubiquitous
By convention, a DNA strand is written in an orientation that places the template strand	reading in a 5'→3’ di	reading in a 3'→5' directi	reading in a 2' → 5' dir	reading in a 1' →3' d	reading in a 5'→3’ direction
The structural feature that allows DNA to replicate itself is the	sugar-phosphate bac	complementary pairing o	phosphodiester bondir	twisting of the molec	complementary pairing of the bases
If the base order in one strand of DNA is 5’-CATTAG-3’, what is the order of bases in the complementary D	3’-GTAATC-5’	b. 3’-GUAAUC-5 ’	3’-CATTAG-5’	5’-GTAATC	3’-GTAATC-5’
In metabolically active cells the proportion of mRNAs is _____ and the proportion of rRNAs is ______	55-60%, 25-30%	25-30%; 55-60%	80-90%, 2.5-5%	2.5-5%; 80-90%	80-90%, 2.5-5%
A nucleotide is made up of which of the following chemical components?	A nitrogenous base, a	A nitrogenous base, an an	A nitrogenous base, an	A nitrogenous base, a	A nitrogenous base, a phosphate group, and a pentose sugar
Genetic code is	the sequence of nitrc	is a triplet code	is non-overlapping	anticodon	is a triplet code
A nucleoside is compsoed of	a base + a sugar	a base+ a sugar + phosphe	a base+ a phosphate	lipids+ proteins	a base + a sugar
genetic mutation occurs in	protein	RNA	DNA	Nucleus	DNA
DNA is present in	viruses, prokaryotes	prokaryotes and eukaryot	only in eukaryotes	in some viruses, prok	in some viruses, prokaryotes and eukaryotes
The two strands in a DNA double helix is joined by	Co-valent bond	hydrogen bond	ionic bond	phosphodiester bond	hydrogen bond
The basic repeating units of a DNA molecule is	nuceloside	nucleotide	histones	amino acids	nucleotide
Adjacent nucleotides are joined by	covalent bond	phosphodiester bond	ionic bond	peptide bond	phosphodiester bond
The length of one turn of DNA is	3.4 angstrom	34 angstrom	20 angstrom	3.04 angstrom	34 angstrom
The type of sugar in DNA are	triose	tetrose	c)pentose	hexose	pentose
The width of DNA molecule is	15 angstrom	3.4 angstrom	20 angstrom	25 angstrom	20 angstrom
The length of DNA having 23 base pairs is	78 angstrom	78.4 angstrom	78.2 angstrom	74.8 angstrom	74.8 angstrom
Left handed DNA	A-DNA	B-DNA	c)Z-DNA	C-DNA	Z-DNA
Z-DNA have a	double helical nature	Zig-zag apperarance	uracil base	single stranded natur	Zig-zag apperarance
Chargaff's rule states that in a double stranded DNA molecule	Concentration of dec	Concentration of Deoxyac	Concentration of Deoxy	Concentration of Dec	Concentration of deoxyadenosine (nucleotides equals that of thymidine (T) nucleotides
Choose the incorrect statement out of the following	Double stranded DN/	The B-form is usually four	single turn of B-DNA ab	d)The distance spann	single turn of B-DNA about the axis of the molecule contains six base pairs
When the DNA molecule is twisted in the direction opposite from the clockwise turns of the right-handed (	Z form	A form	Positive supercoils	Negative supercoils	Negative supercoils
Choose the correct statement out of the following?	The common form oc	The coding strand is copie	The two strands of the	The G-C bonds are mı	The G-C bonds are much more resistant to denaturation than A-T rich regions
RNA is a polymer of purine and pyrimidine ribonucleotides linked together by	Hydrogen bonds	hydrophobic interactions	vander wal's forces	3'-5' phosphodiester	3'-5' phosphodiester linkages
Which of the following types of RNA participate in RNA processing?	t-RNA	r-RNA	small nuclear RNA (snR	Small interfering RNA	small nuclear RNA (snRN
The anticodon region is an imprtant structural component of	mRNA	DNA	rRNA	tRNA	tRNA
The small nuclear RNAs are rich in	uracil	cytosine	thymine	adenine	uracil
The Z-DNA helix	has fewer base pairs	has alternating GC sequer	tends to be found at th	is inhibited by methyl	has alternating GC sequences

Microarrays	Are used for analysis	Are made of glass	Are smaller than DNA c	Contain DNA sequen	Contain DNA sequences
Which of the following has the highest percentage of modified bases?	hnRNA	b)t-RNA	sn-RNA	m-RNA	t-RNA
Choose the incorrect statement about an RNA?	has a single stranded	Does not obey chargaff's	c)Does not exhibit wats	instead of uracil cont	instead of uracil contains the ribonucleotides of thymine
Which of the following statements is incorrect about tRNA species?	they have the smalle	All tRNAs have between 1	tRNA molecules contai	They make up 15% o	All tRNAs have between 10-20 nucleotides
Semi conservative replication of DNA was first demonstrated in	E.Coli	S.penumonae	S.typhimuriam	D.melanogaster	E.Coli
Mode of DNA replication in E.Coli is	Conservative and uni	b)Semiconservative and u	Conservative and bidire	semiconservative an	semiconservative and bidirectional
When DNA replication starts	The phosphodiester	the bonds between the ni	the leading strand prod	the hydrogen bonds	the hydrogen bonds between the nucleotides of two strand break
The elongation of the leading strand during DNA synthesis	Progresses away fro	occur in 3'-5' direction	produces okazaki fragm	Depend on the actio	Depend on the action of DNA polymerase
Eukaryotes differ from prokaryote in mechanism of DNA replciation due to-----	Different enzyme for	use of DNA primer rather	unidirectional rather th	discontinous rather t	discontinous rather than semidiscontinous replication
The accepted hypothesis for DNA replication is	conservative theory	b)dispersive theory	semi-conservative theo	evolutionary theory	semi-conservative theory
Unit III					
Biological sequences are generated at ________ rates	exponential	average	low	fast	exponential
Process by which sequences are compared is	docking	alignment	interaction	none of these	alignment
Alignment of two sequences is known as	global alignment	sequence alignment	pairwise alignment	multiple alignment	pairwise alignment
Building block of DNA is	amino acids	glycosides	nucleotides	nucleosides	nucleotides
By comparing sequences through alignment, patterns of ________ can be verified	conservation	differntiation	similarity	declinination	conservation
Variation in the sequence alignment is because of	substitutions	insertion	deletion	all of these	all of these
When significant similarity is seen between two sequence, that means they belong to	different family	same family	close family	none of these	same family
When two sequence have common ancestors, that means sequences are	homologous	heterologous	identical	autologous	homologous
Alignment method suitable for aligning closely related sequences is	sequence alignment	pairwisw alignment	global alignment	local alignment	global alignment
The alignment method suitable for finding out conserved patterns in DNA or protein sequences is	multiple alignment	global alignment	local alignment	pairwise alignment	local alignment
The procedure of aligning many sequence simultaneously is	multiple alignment	global alignment	local alignment	pairwise alignment	multiple alignment
The alignment method that tries to align regions with high level eithout considering alignment of rest of se	sequence alignment	pairwisw alignmwnt	global alignmwnt	local alignment	local alignment
Which is not a sequence alignment tool	Rasmol	BLAST	FASTA	Clustal W	Rasmol
Which of the following is multiple sequence alignment tool	clustal W	chime	pdb	Rasmol	Clustal W
If two sequences shows 30% similarity, they are referred as	Safe zone	twilight zone	midday zone	midnight zone	safe zone
If two sequences are referred as in twilight zone, then it exhibits _________ similarity	20-30%	50-60%	40-50%	70-80%	20-30%
Sequence_____refers to the percentage of matches of the same amino acid residues between two aligned	similarity	identity	complexity	simplicity	identity
When sequences are aligned from beginning to end , it is called as	sequence alignment	pairwisw alignment	global alignment	local alignment	global alignment
Most basic sequence alignment method is	word method	dot plot method	dynamic method	none of these	dot plot method
In dot matrix method two sequences are compared in _______	2D matrix	3D matrix	4D matrix	all of these	2D matrix
Which webserver uses dot matrix	BLAST	FASTA	dottup	clustal W	dottup
Which is a web-based program that uses a variant of the Smith–Waterman algorithm	BLAST	dot plot method	rasmol	LALIGN	LALIGN
Multiple sequence alignment is an natural extension of	local alignment	global alignment	pairwise alignment	all of these	pairwise alignment
The scoring function for multiple sequence alignment is based on the concept of	difference of pairs	multiplication of pairs	division of pairs	sum of pairs	sum of pairs
DCA stands for	divide and conquer al	divide and control alignm	dot and control alignm	dot and conquer alig	divide and conquer alignment
Which is the most common scoring matrix	PAM	BLOSUM	both A and B	none of these	Both A and B
Who developed PAM	Needlemann	wunsch	Dayhoff et al.	smith	Dayhoff et. Al
How is the E-value related to score	lower E value, more s	lower E value, less score	no relation	none of these	lower Evalue, more score
The driving force behind evolution is natural selection in which ____ forms are eliminated	misfit	unfit	similar	different	unfit
Who developed neighbour joining method	Heinkoff and Heinkof	Saitou and Nei	Karlin and Astchul	Dayhoff et. Al	Saitou and Nei
Branching pattern in a phylogenetic tree is called as	tree homology	tree neurology	tree topology	tree heterology	tree topology
The branch path depicting an ancestor–descendant relationship on a tree is called a	topology	taxan	lineage	dichtomy	lineage
When a number of taxa share more then one common ancestors, it is referred as	paraphylectic	topology	dichtomy	lineage	paraphylectic
The phylogeny with multifurcating branches is called	taxan	dichtomy	lineage	polytomy	polytomy
A phylogenetic tree can be	rooted	unrooted	both A and B	none of these	Both A and B
Which tree does not give knowledge of common ancestors	unrooted	rooted	pointed	branched	unrooted
How many ways can define root of a tree	3	2	1	4	2
What is an assumption, by which molecular sequences evolve at constant rates	molecular time	molecular way	molecular approach	molecular clock	molecular clock
Phylogenetic tree in which branch legnth is scaled is known as	phylogram	dendogram	cladogram	caleidogram	phylogram
Cladogram is a tree in which branch legnth is	unscaled	scaled	small	large	unscaled
In phylogram, branch represents evolutionary _______	similarity	divergence	distance	all of these	divergence
How many steps are needed to construct a molecular phylogenetic tree	3	6	8	5	5
What does PHYLIP stands for	PHYLogeny Inference	PHYLogenetic Image Proc	PHYLogeny Inference P	PHYLogenetic Inferen	PHYLogenetic Inference package
Multiple sequence alignment can be used to create	functional model	phylogenetic tree	algorithms	all of these	phylogenetic tree
What infromation does sequence alignment produce	evolutionary relation	functional relationships	structural relationships	all of these	all of these
Multiple state-of-the-art alignment program is	T-Coffee	T-tea	BLAST	FASTA	T-Coffee
How many tree construction methods and programs are available currently	3	5	1	2	2
Clustering type algorithm includes	neighbour joining alg	neighbour distancing algo	both A and B	none of these	neighbour joining algorithm
What is the statistical technique that tests the sampling errors of a phylogenetic tree	bootlegging	bootstrapping	boottrapping	bootnegging	bootstrapping
In nonparametric bootstrapping method, process is repeated	1000-10000 times	10-100 times	100-1000 times	50-100 times	100-1000 times
Unique requiremnt of database searching is	sensitivity	selectivity	speed	time	selectivity
In databse searching, there are ______ fundamental types of algorithm	4	7	10	2	2
Fundamental algorithms used in data searching are	exhaustive and heuris	exhaustive and intrinsic	extrinsic and intrinsic	extrinsic and ehausti	exhaustive and heuristic
Both BLAST and FASTA use a heuristic _____ method for fast pairwise sequence alignment	letter	node	word	sentence	word
BLAST tool was delevloped in	1992	1999	1990	1988	#
First step in a BLAST search is known as	seeding	leading	extending	including	seeding
What uses protein sequences as queries to search against a protein sequence database	BLASTN	BLASTP	BLASTX	TBLASTX	BLASTP
BLASTN is used to query nucleotide sequence from	nucleotide database	protein database	PDB	structure database	nucleotide database
BLASTP is used to query protein sequence from	protein database	nucleotide database	translated database	specific database	protein database
CLUSTAL-W is used to align	Single sequence	multiple sequences	no sequence	two sequences	multiple sequences
Unit IV					
Each record should contain a number of fields that hold the actual data iteam such as	phone number, addre	weight	place	quality	phone number, address, date
Biological databases have a higher level of requirement, known as	knowledge discovery	entry	value	market qurey	knowledge discovery
Depending on the types of data structures, database management systems can be classified into	three	four	five	two	two
Database management systems can be classified into two types such as	Relational databases	Object-oriented database	functional databases	metadatabase	Relational databases
Biological databases can be divided into	5categories	4categories	3categories	2categories	3categories
which one of them is not an biological database	primary databases	secondary databases	specialized databases	derived database	derived database
Example of secondary databases is	SWISS-PROT	RDP	IMGT	REBASE	SWISS-PROT
The most popular retrieval systems for biological database is	Entrez	LIMITS	HISTORY	CLIPBOARD	Entrez
GOBASE - is a specialised database of	Ribosomal Database	restriction enzyme sites	organelle genomes	transcription factor b	organelle genomes
several options which is common to all NCBI databases	Limits	Preview/Index	Send to Clipboard	clipping	Preview/Index
#NAME?	tbl2asn,	Barcode Submission Tool,	BankIt	EST, STS, and GSS	BankIt
data from GenBank can be searched and retrieved by	CoreNucleotide	dbEST	chart	figure	CoreNucleotide
Flat files contain three sections such as	Header	Sequence entry	Features	function	Sequence entry
The minimum length required for submission in GenBank is	60 bp	50 bp	30 bp	20 bp	50 bp
The GenBank database is divided into ------------- divisions	18	17	16	15	#
Iin GenBank MAM refers to ?	other vertebrate seq	plant, fungal, and algal se	other mammalian sequ	patent sequences	other mammalian sequences
The GI system of sequence identifiers runs parallel to the accession.version system, which was implement	Feb-98	Feb-99	Feb-97	Feb-99	#
GenBank flat file AF165912 which refers to ?	gene, promoter, TATA	protein bind, gene, 5'UTR	alternatively spliced m	alternatively spliced	gene, promoter, TATA signal, mRNA, 5'UTR, CDS,3'UTR
Protein sequence database are classified into how many types	five	four	two	three	two
which of the following is not an Primary protein sequence databases	MIPS	BLOCKS	TrEMBL	OWL	BLOCKS
NRDB Primary protein sequence database refers to	Martinsried Institute	Non-Redundant Database	Protein Information Re	Translated EMBL	Non-Redundant Database
which of the following is not a Secondary protein sequence database	SWISS – PROT	Profiles	Pfam	IDENTIFY	SWISS – PROT
which of the following is a Secondary protein sequence database	PROSITE	Profiles	PRINTS	Genbank	PROSITE
how many types of Secondary protein sequence databases are present currently	7	6	5	4	6
PIR was Developed by Margaret Dayhoff during	1970s	1980s	1960s	1990s	1960s
Based on data quality and annotation level, PIR database divided into----- sections	three	two	one	four	four
Protein Information Resource 1 contains------	fully classified and an	unverified entries	artefactual sequences	preliminary entries,	fully classified and annotated entries
Major function of PIR 2 is	Conceptual translatio	Conceptual translations o	Conceptual translation	protein identification	Conceptual translations of genetically engineered
SWISS-PROT established by the Department of Medical Biochemistry at the University of Geneva and EMB	1985	1986	1988	1983	#
which database informs the protein by the name it is known	SWISS – PROT	Profiles	Pfam	PRINTS	SWISS – PROT
how many types of identification lines are there in SWISS-PORT	5	7	8	9	7
ID line which are used in PROSITE entry is	PATTERN	MATRIX	RULE	ruler	MATRIX
ID line which does'nt belongs to PROSITE entry is	PATTERN	MATRIX	IDENTIFY	RULE	IDENTIFY
The entry name consists of from -----------------uppercase alphanumeric characters.	1 to 20	2 to 21	5 to 50	3 to 21	2 to 21
The format of the NR line is	NR /QUALIFIER=data	NR /QUALIFIER=data;	/QUALIFIER=data;	NR =data;	NR /QUALIFIER=data;
which one is the protein domain library sequences database	SBASE	BLOCKS	PRINTS	Pfam	SBASE
different approach to pattern recognition, termed "fingerprinting" is used by which database	SBASE	BLOCKS	PRINTS	Pfam	PRINTS
which datdbase is used to create protein family or domain signatures	BLOCKS	SBASE	Pfam	PRINTS	Pfam
In structural database the structure of biomolecules are obtained by	X-Ray	IR ray	UV ray	β-ray	X-ray
The aim of most protein structure databases is to organize	protein size	protein structures	protein function	protein energy	protein structure
The number of known protein structures are available through	Nucleic Acid Databas	Protein Data Bank	The Cambridge Crystall	none of the above	Protein Data Bank
The Cambridge Crystallographic Data Centre provides database of structures of -----	DNA	RNA	larger protein molecule	smaller protein mole	smaller protein molecules
The Protein Data Bank (PDB) was established in	1972	1973	1974	1971	#
which of the following is an macromolecular 3D structure database	RCSB PDB	MMDB	EBI structure database	database structure	RCSB PDB
EBI structure databases are of how many types	4	3	5	6	5
catalytic sites and residues are identified in enzymes using ------------------structural data base	PDBsum	HSSP	CSA	DSSP	CSA
CATH is a hierarchical classification of protein domain structures, which clusters proteins in ---------------typ	four	three	two	five	four
which database ia used to identify structural and evolutionary relationships between all proteins whose st	SCOP	CATH	PSdB	PDBsum	SCOP
which database provide high quality pictures of biological macromolecules with known three-dimensional	SWISS-MODEL	ModBase	SWISS-3D IMAGE	Bibliographic databas	SWISS-3D IMAGE
The best known bibliographic database is	EMBASE	BIOSIS	CAB International	MEDLINE	MEDLINE
Unit V					

What is the prerequisite for detailed functional annotation of genes and genomes	Gene prediciton	Sequence alignment	BLAST	CLUSTAL-W	Gene prediction
The ultimate goal in gene prediciton is to describe all the genes computationally with nearly ------accuracy	100.00%	75.00%	50.00%	10.00%	#
Which can significantly reduce the amount of experimental verification work required	Gene prediciton	CLUSTAL-W	BLAST	Sequence alignment	Gene prediction
How many categories of gene prediction programs are available	one	two	three	four	two
Which method predicts genes based on the given sequence alone	Ab-initio	homology	alignment	phylogenetics	Ab-initio
Which is the statistical description of coding regions	gene content	homology	structure	alignment	gene content
Which method makes predictions based on significant matches of the query sequence with sequences of k	Ab-initio	homology	alignment	phylogenetics	homology
The size of prokaryotic genome ranges from	0.05 to 1 Mbp	0.5 to 10 Mbp	50 to 100 Mbp	500 to 1000 Mbp	0.5 to 10 Mbp
In prokaryote, the gene density in the genomes is	high	medium	low	very low	high
In prokaryote, gene content in the genome is	25.00%	50.00%	75.00%	100.00%	#
In prokaryotes, majority of genes have a start codon?	ATG	CTG	GTG	TGT	ATG
Prokaryotes includes bacteria and archaea genomes with sizes ranging from	10 to 20Mbp	0.5 to 10Mbp	0.8 to 30Mbp	10 to 15 Mbp	0.5 to 10Mbp
UNIX program from TIGR that uses the IMM algorithm to predict	coding regions	protein suructure	ribosomes binding sites	gene functoin	protein suructure
FGENESB is a web based program that is specifically trained for	human sequences	plant sequences	bacterial sequences	none of the above	bacterial sequences
In prokaryote, the ribosomal binding site is called as	Shine-Delgarno	coding	non-coding	termination	Shine-Delgarno
Many prokaryotic genes are transcribed together as	operon	individual	Co-expressed	differentially express	operon
In prokaryote, end of the operon is characterized by a transcription termination signal called as	p- independent termi	shine-delgarno sequences	codon	operon	p- independent terminator
In eukaryote, the genome size ranges from	25Mbp to 800Gbp	10Mbp to 670Gbp	50Mbp to 100Gbp	100Mbp to 500Gbp	10Mbp to 670Gbp
In eukaryote, coding sequences are called as	introns	exons	terminator	promoter	introns
Splicing is the process of removing	introns	exons	exons and joining intro	terminator	introns
Proteins have major role of	regulatory function	transport	enzymatic	growth	enzymatic
The covelent bond connecting the two amino acid is called as	peptide bond	polypeptide	dipeptide	tripeptide	peptide bond
The protein structure can be organised into	3 level	2 level	5 level	4 level	4 level
Linear amino acid sequences of protein is	primary structure	secondary structure	tertiary structure	quaternary structure	primary structure
coiled loop is a type of	primary structure	secondary structure	tertiary structure	quaternary structure	secondary structure
intrinsic tendency of each residue in a helix is determined by	GOR method	Chou-Fasman algorithm	homology based metho	PSIPRED	Chou-Fasman algorithm
GOR method examines a window of every-------residues	twenty	fifteen	seventeen	ninteen	seventeen
homology based methods were developed in the year	1984	1986	1988	1990	#
expand PHD	profile newton fromh	profile network from heid	postal network from he	peculiar network fro	profile network from heidelberg
there are how many computational approaches present to predict tertiary structure of protein	1	3	5	7	3
Homology modelling is also known as	interactive modeling	competitive modeling	comparative modeling	Non-comparative mo	comparative modeling
Homology modeling consists of how many steps	5	6	7	8	6
As a rule of thumb, a database protein should have ----- sequence similarity	10.00%	20.00%	30.00%	40.00%	#
Which one of these is freely available protein modelling programs	Swiss-Model	BLAST	FASTA	Ab-initio	Swiss-model
Most current side chain prediction programs uses the concept of	totomer	rotamers	both A and B	none of the above	rotamers
The pairwise energy–based method for protein folding was originally referred to as	fold	threading	fold recognition	profile	threading
Which of these is a profile based fold recognition server	GenThreader	ModBase	3Dcrunch	3D-JIGSAW	GenThreader
Fold recognition is also known as	pairwise method	pairwise energy based me	profile based method	proton method	profile based method
Protein structure that has minimum energy level is known as	naive state	novel state	native state	none of the above	native state
Which of these is a protein viewing program	RasMol	RasTop	adobe	andriod	RasMol
Simplest form of protein model representation is	Wire-frame	balls and sticks	ribbon	Space-filling	Wire-frame
In balls and sticks model of protein, balls represents	bonds	molecules	amino acids	atoms	atoms
In ribbon diagrams of protein model, Beta sheets is represented as	cylinders	spiral ribbons	flat arrows	balls	Flat arrows
RasMol is a ------ based viewing program	web based	command line based	visual basic	microsoft	command line based
Rosetta is web-server that predicts protein 3D conformation using -------	GOR method	Chou-Fasman algorithm	Ab-initio method	BLAST	Ab-initio method
Rosetta method first breaks down the query sequence into many very short					
Segments of size	3-9 residues	1-3 residues	5-7 residues	6-9 residues	3-9 residues
procheck is a UNIX program that is able to check general physicochemical paameters such as	chirality	bond angles	bond legnth	width	bond angles
Ab-initio method was developed in year	1960	1970	1980	1990	#
It has been estimated that nearly --- of residues of a protein fold into either α-helices and β-strands.	25.00%	50.00%	75.00%	95.00%	#
Amino acids in a secondary structure is stabilized by	vander wal's force	covalent bond	hydrogen bond	ionic bond	hydrogen bond
Which protein diagram shows localized position of specific residues in protein	ball and sticks	wire frame	space filling	ribbons model	Wire-frame
Protein model that represents backbone of a protein structure is	ball and sticks	wire frame	space filling	ribbons model	balls and sticks
In space-filling representations of protein model each atom is represented as	cylinders	cones	spheres	ribbons	spheres
Which protein models give clear representation of overall topology of the structure	Wire-frame	balls and sticks	space filling	ribbons model	ribbons model
Chime program is derived fom	pubMed	NCBI	PDB	RasMol	RasMol
Which web based servers displays protein secondary structure cartoons	WebMol	Chime	Cn3D	Grasp	Cn3D
Molscript runs on	UNIX	DOS	Windows	android	UNIX