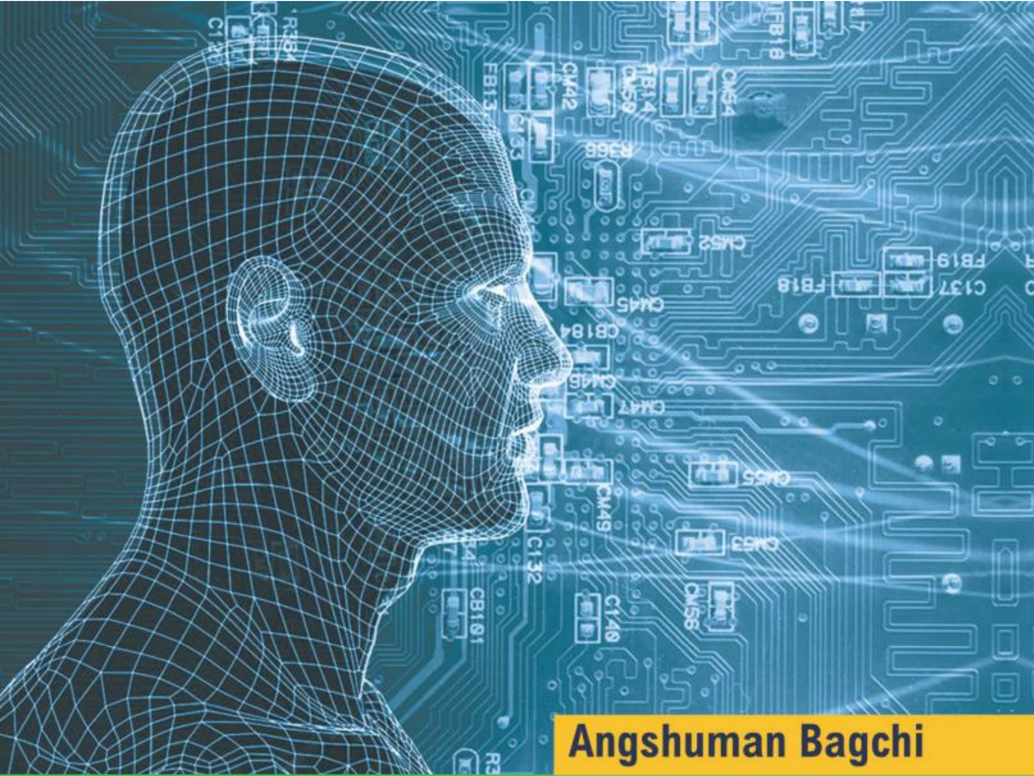


Angshuman Bagchi

Introduction to Bioinformatics

Alpha Science

INTRODUCTION TO BIOINFORMATICS

INTRODUCTION TO BIOINFORMATICS

Angshuman Bagchi

Alpha Science International Ltd.
Oxford, U.K.

Introduction to Bioinformatics
168 pgs.

Angshuman Bagchi
Department of Biochemistry and Biophysics
University of Kalyani
Nadia

ALPHA SCIENCE INTERNATIONAL LTD.

7200 The Quorum, Oxford Business Park North
Garsington Road, Oxford OX4 2JZ, U.K.

www.alphasci.com

ISBN 978-1-78332-377-7

Preface

Bioinformatics is an interdisciplinary subject which uses the knowledge of Biology, Physics, Chemistry, Statistics, Mathematics and Computer Science. It deals with the building and development of Biological Databases and Software tools. It uses the power of computations to analyze and solve biological problems. The prerequisites for Bioinformatics study are good quality computers with high processing speeds. However, the analyses and interpretations of the Biological data require human intervention. Bioinformatics is largely but not exclusively a computer based study. Computers are needed to perform the repetitive tasks that are at the core of many Bioinformatics analyses. Computers are also needed for their abilities to solve problems. The bioinformatics study focuses primarily on the following aspects:

- Database creation for storage and management of large biological data
- Development of algorithms and statistics to analyze the relationships between various types of biological data
- Use of the software tools to come up with new biological hypothesis and to provide biological insights into the biological phenomenon

The use of bioinformatics in analyzing the biological data became popular in the early 1990s. However, previously bioinformatics study was confined mainly to the management of biological sequence information. The popularity of bioinformatics as an independent branch of science came into reckoning after the advancement of sequencing technologies. This led to a rapid outburst of biological sequence data especially nucleotide sequence data which are subsequently being stored in GenBank.

Bioinformatics to a large extent depends on computational power. However, other branches of science like Physics, Statistics, Mathematics, and Chemistry play equally important roles in analyzing the biological data by computational means. Thus, an attempt has been made to come up with an overview of the important aspects of this interdisciplinary science.

This book is meant for the Bachelors and Masters students of Bioinformatics. The second chapter of the book deals with the popular and important biological databases. The third chapter focuses on the biological sequence alignment methods. The fourth chapter presents the molecular modelling techniques and the last chapter presents the Phylogenetic analyses.

The details of some of the important web tools and web servers are also presented. I incorporated some multiple choice questions for competitive examination.

At the end I would like to wish the readers all the very best.

Contents

CHAPTER

1

Introduction

Bioinformatics is a multi-disciplinary science. The study of Bioinformatics is related to solving problems of molecular biology and genetics with the help of computational tools. In doing so, concepts from computer science, mathematics and statistics, physics and chemistry are considered. The basic framework of Bioinformatics lies on the use of computational tools to solve biological problems which are data intensive and often incorporate large-scale biological information. Bioinformatics takes the help of the existing data that are available and tries to extract new information from them. In general the following classes of problems are dealt in Bioinformatics:

- Analyses of biological sequences to extract new information from them
- Annotations of genome sequences
- Identification of phylogenetic lineages
- Modelling of the biological pathways
- Prediction of three-dimensional structures and functions of biological molecules
- Simulations of biological molecules under different physiological conditions
- Making repositories of collected biological information

The Bioinformatics analyzes the biological data in several steps. However, the following steps can be considered as the general ones:

- Collection of biological sequence information
- Building a computational model from the collected sequence information
- Testing the generated model for real life biological data

In this chapter, the basic aspects of Bioinformatics will be discussed. The interrelationships of Bioinformatics with other scientific fields will be considered. The important terminologies in Bioinformatics will be analyzed.

1.1 HISTORY OF BIOINFORMATICS

After the discovery of DNA structure and protein sequencing (1950-70s), protein and nucleotide sequence storing and alignments were being performed manually. A major advancement in the field of molecular biology in the last two decades (1980-2000s) brought about the exponential increase of biological data which need a time and cost effective technology to manage all these. Along with this there was also an increased ratio between performance and cost of computer hardware. Hence, the field of Bioinformatics evolved and took a central role in modern day biological science.

1.1.1 Computational System Biology

The better understanding of complex biological systems requires a computational biology platform that applies the techniques of computer science, applied mathematics and statistics to solve the biological problems. After mid-1990s the Human Genome Project and rapid advances in DNA sequencing technology culminated in explosive growth in the field of computational biology. This necessitates the analyses of biological data to produce meaningful information using algorithms from graph theory, artificial intelligence, soft computing, data mining, and image processing and computer simulation. The advent of databases and their principles led to the concepts of sequence comparisons. Sequential alignments led to three-dimensional model building, validation of models and interaction studies.

Computational biology involves:

- Prediction of three-dimensional structures of proteins and nucleic acids
- Prediction of binding interfaces of biological macro or micro molecules
- Interaction study between biological molecules
- Functional domain/motif prediction
- Computational phylogentics
- Micro RNA reglation and post transcriptional modification
- Mutational analysis

1.2 BIOLOGICAL DATABASES

The rapid progress in gene and protein identification and sequencing led to the collection of information related to genomics, transcriptomics, proteomics, metabolomics and this required those data to be stored which can be accessible and later updated and retrieved easily from all over the world.

1.2.1 Primary Database

A database consisting of data derived directly from experiments such as nucleotide or protein sequences and three-dimensional structures of protein/nucleic acid are known as primary database. There are several types of primary databases

A. Genome Database

1. Sequence Database – NCBI, GenBank, DDBJ, EMBL, TrEMBL
2. Structural Database – Nucleic Acid Database (NDB), RNABase

B. Protein Database

1. Sequence Database – Swiss-Prot
2. Structural Database – Protein Data Bank (PDB)

Brief aspects of some important databases are presented in the following section. The detailed analyses of the databases are given in the respective chapters.

NCBI-The National Center for Biotechnology Information (NCBI): It is the part of the United States National Library of Medicine (NLM), a branch of the National Institutes of Health (NIH). NCBI maintains a series of databases relevant to biomedicine and is an important resource for bioinformatics tools and services. Major databases include GenBank for DNA sequences and PubMed, a bibliographic database for the biomedical literature.

PDB comprises of

- Protein Data Bank of Europe (PDBe)
- Protein Data Bank of Japan (PDBj)
- Research Collaboratory for Structural Bioinformatics in USA (RCSB)

The three dimensional atomic coordinates of biological macromolecules are deposited in PDB, which are obtained by X-Ray crystallography, NMR spectroscopy or Cryo-Electron Microscopy (EM) around the globe. The structure files may be viewed using one of several free and open source computer programs as well as paid software tools like Jmol, Pymol, Rasmol, VMD, UCSF Chimera, Swiss-PDB Viewer and Discovery Studio.

Swiss-Prot: SWISS-PROT is an annotated protein sequence database having an equal partnership between the European Molecular Biology Laboratory (EMBL) and the Swiss Institute of Bioinformatics (SIB). The SWISS-PROT database distinguishes itself from other protein sequence databases by three distinct criteria: (*i*) annotations, (*ii*) minimal redundancy and (*iii*) integration with other databases.

1.2.2 Secondary Database

It is also known as curated database which includes the results of analysis of primary databases and other significant data in the form of conserved sequences, signature sequences, secondary structures, active site residues of proteins etc. In the following section, a brief analysis of some of the important secondary databases are presented.

RefSeq: The Reference Sequence (Refseq) database is an annotated, non-redundant and curated collection of nucleotide sequences and their protein products.

UniProtKB: It is the high quality annotated non-redundant protein sequence database, generated by manual interventions, which brings together experimental results, computed features and scientific conclusions and reviewed section of the UniProt Knowledgebase (UniProtKB).

Prosite: A database of protein domains, families and functional sites as well as associated patterns and profiles to identify them.

Pfam: Protein family database based on alignments and HMMs. Pfam alignments are available from searching a variety of databases with different accession numbers (e.g. all UniProt and NCBI).

InterPro: It provides a classification among protein families, motifs and domains on functional basis to predict functional domain or site.

SCOP: The Structural Classification of Proteins (SCOP) database is largely a manual classification of protein structural domains based on similarities of their structures and amino acid sequences.

CATH: A database of Protein Structure Classification which provides information on the evolutionary relationships of protein domains.

Dali: The Dali Database and server is based on all-against-all 3D structure comparison and FSSP structural classification of protein structures in the Protein Data Bank (PDB).

1.2.3 Composite Database

This type of database is a combination of many other databases. The following examples are some of the most important composite databases.

NRDB: It is a so-called non-redundant composite database of the following sources: PDB, SWISS-PROT, SWISS-PROT update, PIR, GenPept and GenPept update.

OWL: It is a non-redundant composite database of 4 major publicly-available primary sources: SWISS-PROT, PIR, GenBank (translation) and NRL-3D.

1.3 BIOLOGICAL SEQUENCE ANALYSIS

Since the sequencing of bacteriophage ØX-174 in 1977, the nucleotide sequences of thousands of organisms have constantly been decoded and thereby getting stored in different biological databases. The nucleotide sequences of the organisms are being analyzed by various scientific groups to decode the underlying information. This is called genome annotations. The most common annotations involve:

- identifications of protein coding genes
- identifications and analyses of genes that lead to RNA biogenesis
- detection of the presence of regulatory sequences, structural motifs and repetitive sequences in the genes.

The nucleotide information of genes within a species or between different species can be compared to identify the similarity between the gene products, viz., the proteins, RNAs or the nucleotides themselves. The databases contain large amounts of data. The genome information of an organism is also very huge. With such large chunks of data, it is simply impractical to analyze biological sequences manually. Thus, Bioinformatics tools are routinely used to analyze the sequence information. The most widely used tool for genome analyses is BLAST. The method used for the purpose is called pair-wise sequence alignment. However, there are other tools available that provide different sets of results. A common practice in genome analyses is to compare the characteristics of different genomes. For such purposes, the technique that is used is called the multiple sequence alignment.

1.3.1 Pair-wise Sequence Alignment

Dayhoff et al., in 1978 studied the sequence alignments among 34 superfamilies of protein sequences and listed the accepted point mutations i.e., replacement of amino acid by another residue by natural selection. Based on accepted mutation and probabilitities of occurrence of each amino acids they generated a mutation probability matrix known as PAM. PAM1 matrix was generated for closely related protein sequences having atleast 85% sequence identities. Later the PAM1 matrix is multiplied by iteslf for several times to generate several other PAM matrices. One of such derived matrices, the PAM250 matrix, is used for sequence alignment of proteins with 20% amino acid sequence identitities. In other words, the PAM250 matrix is used for sequnce comparison of evolutionarily distantly related proteins. PAM mutation probability matrix is then converted into a scoring matrix called log-odds matrix or relatedness odds matrix. Another log-odds matrix developed by Henikoff and Henikoff (1992, 1996) is called BLOcks SUbstitution Matrix (BLOSUM) using BLOCKS database which consists of more than 500 groups of local alignments of distantly related protein sequences. High value of BLOSUM (BLOSUM90) and low value PAM matrices (PAM30) are used for conserved protein sequences whereas low BLOSUM and high PAM numbers are used for distantly related proteins. BLOSUM62 matrix merged 62% or more identical sequences into one alignment; thus proteins having less than 62% sequnence identities are analyzed by this BLOSUM62 matrix and it is the default scoring matrix used by most BLAST algorithms for searching sequence similarities between sequences. Two types of pair-wise sequence alignments are :

(*i*) **Global alignment technique:** It is derived from the Needleman–Wunsch algorithm (1970), which is based on dynamic programming which attempts to align every residue in every query sequence. It is most useful when the sequences in the query set are similar and of roughly equal length.

(*ii*) **Local alignment technique:** It is more useful for dissimilar sequences that are suspected to contain regions of similarity or similar sequence motifs within their larger sequence contexts. The Smith–Waterman algorithm (1981) is a general local alignment method which is also based on dynamic programming algorithm, for scoring the matrix and it uses trace-back procedure to obtain the optimal alignment.

Dynamic programming: Alignments of amino acid sequences of proteins use a substitution matrix to assign scores to amino acid matches or mismatches, and a gap penalty for matching an amino acid in one sequence to a gap in the other. The nucleotide sequences of DNA and RNA may be used in scoring the matrix. In practice, a simple scoring scheme is generally used to score the sequence alignments; a positive score is used for a match, a less positive or a negative score for mismatch and a negative score for gap penalty.

Though the local alignment is accurate but it is relatively slow when pairwise local alignment is used for a query sequence against an entire database. The computational running space and time then become a significant issue. Two other popular rapid, heuristic version of local alignment algorithms have been developed based on identifying short words or k-tuples. Such processes would involve 1/2 residues for protein and upto 6 bases for DNA searching. The locally aligned regions are called high-scoring segment pairs or HSPs. The expect (E) value is the measure of statistical significance which represents the number of hits one can expect to obtain by chance; that means the lower the E-value is the more significant is the alignment. The E-value decreases exponentially with high score (S) value corresponding to better alignments.

- *Word methods:* The heuristic methods are significantly more efficient than dynamic programming. These methods are especially useful in large-scale database searches and best known for their implementation in the database search tools in a number of web portals, such as EMBL FASTA and NCBI BLAST. NCBI BLAST program is a collection of different tools like blastp and PSI-BLAST.
- *blastp*: This program compares a protein query sequence against a protein sequence database. In this software program the default matrix is BLOSUM62 with a word size 6, gap existence 11 and extension value 1.
- *PSI-BLAST:* This blast program is specilized, advanced and more sensitive position-specific BLAST search which is used for distantly related proteins sharing the same conserved three

dimensional structures but low sequence identities. An initial *blastp* search is used to perform a multiple sequence alignment. By analyzing this alignment it then creates a specialized position-specific scoring matrix which again is used as query iteratively. The orignial query is used as template for generating the PSSM profile.

1.3.2 Multiple Sequence Alignment

The alignments of more than two protein or DNA sequences are very useful to study the homologous group and it provides structure-function and evolution of the protein or gene families. It is more sensitive than pairwise alignment to detect homologs, to find conserved domain or motifs or consensus regions. The most commonly used algorithm for multiple sequence alignment is derived from Feng and Doolittle's (1987, 1990) progressive alignment method. The basic strategy of this method is calculaing pairwise alignment scores between all sequences and then align two closely related sequences followed by addition of sequences progressively to the alignment. The most popular web version of multiple sequence alignment tool is ClustalW program. To generate a guide tree, this program uses mostly distance matrix instead of similarity matrix. The two main features of a guide tree are its branching order or topology and branching length which is proportional to evolutionary distance. The two main methods of tree construction are:

(*a*) unweighted pair group method of arithmatic averages (UPGMA) and

(*b*) neighbour-joining method.

The multiple sequence alignment output arranges the sequences as presented in the guide tree, i.e., two closely related sequences create a pairwise alignment and then sequences are added one by one. The newer web version of multiple sequence alignment is Clustal Omega which uses seeded guide trees and HMM profile-profile techniques to generate alignments.

1.3.3 Molecular Phyllogeny

Previously the classification and phylogenetic studies were done based upon morphological, anatomical, physiological, paleontological characteristics. But after the huge progress in molecular biology the evolutionary relationships among different species have been being studied using their protein or nucleotide sequences. A phylogenetic tree is a graphical representation of relatedness among sequences or ancestral origin of the sequences. The node is the intersection or terminating point of the tree which represents the taxonomic units (taxa or taxons, can be internal node or external node). An operational taxonomic unit (OTU) is the available sequence which is used for tree analysis. Branches define the relatedness of the sequences and branch length sometimes are scaled to represent the number of changes occurring between sequences. The tree generated from scaling is called a phylogram. On the other hand, the trees may be unscaled where the branch length is not proportional to the number of changes. Such trees are called cladogram. A clade or monophyletic group is a group of sequences that share a common ancestor. A rooted tree represents the common ancestor whereas unrooted tree does not define the evolutionary path leading to their common ancestors. A phylogenetic tree generation using molecular sequence data involve multiple sequence alignment using different approaches, viz., distance-based methods or character based methods.

In distance based methods the tree is constructed by calculating the distances between molecular sequences. In the construction of guide tree in multiple sequence alignment, the distance based methods used are unweighted pair group method of arithmatic mean (UPGMA) and neighbor-joining (NJ). UPGMA is simple tree making algorithm where the sequences make cluster based on distance matrix and it is always a rooted tree.

The two main character based methods are maximun parsimony (MP) and maximum likelihood (ML). Maximum parsimony method is applied where the sequences are closely identical thereby building a tree with shortest branch lengths.

Maximum likelihood method on the other hand uses the knowledge of speciation to produce sequence clusters.

1.3.3.1 Neighbour Joining (NJ) Method

A star like tree is generated and then pairwise comparisons are made to identify two closely related sequences, the neighbors, followed by joining them until the topology of the full tree is completed.

1.3.3.2 Maximum Likelihood (ML) Method

It is the most computationally intensive method but the most flexible method when large amounts of evolutionary changes are found in the distantly related sequences. It uses the probability distribution statistical method to generate the highest probable tree. The mutation of each residue calculated is based on substitution model where probability of each mutation is calculated and are then aligned.

1.3.3.3 Tree Evaluation and Bootstrapping

To evalute the tree, the most common method applied is randomizing the test or bootstrapping analysis. Bootstrapping analysis describes the robustness of the topology of a tree that means it checks the consistency of the branching order. After developing the tree, the positions of the aligned sequences are randomly sampled from the multiple sequence alignment with replacements and then are assembled into new randomized data sets, the so-called bootstrapped samples. A large number of bootsrap replicates are generated. The bootsrap tree is compared with the original tree and the bootstrap value denotes the frequencies of observed data to obtain each clade.

Protein Domain Identification and Mutational Analysis: A conserved protein family represents the functional aspect of the amino acid sequences. Multiple alignment of the sequences obtained from the blastp search would find the conserved functional domains. Multiple sequence alignment (MSA) reveals some point mutations within protein domains. Some mutations are synonymous substitutions and semi-conservative mutations in the MSA profile. Point mutations were identified from the particular sequences and then short streches of sequences are again searched using blastp and new functional domains are found. The relative mutational probabilities of different amino acid residues are different according to their genetic codes. Some amino acid residues such as asparagine and serine undergo substitutions very frequently, while other residues (notably tryptophan and cysteine) are very rarely mutated. Dayhoff et al calculated the relative mutabilities of amino acid residues which describe how often each amino acid is likely to change over a short evolutionary period. To calculate the relative mutability they divided the number of times each amino acid was observed to mutate by the overall frequency of occurrence of that amino acid. The less mutable residues probably have important structural and functional roles in proteins such that replacing them resulted in some significant changes. The most common substitutions are Glu for Asp (both acidic), Ser for Ala and Thr (both are hydroxylated), and Ile for Val (both hydrophobic and of similar sizes). The relative mutabilities depend on to the physico-chemical properties of the amino acids. With reference to the genetic code it is also explained in such a way that common amino acid substitutions tend to require only single nucleotide change and least mutable residues are coded by only one or two codons. The low mutability of these amino acids suggest that substitutions are not tolerated by natural selections.

Secondary Structure Prediction: Secondary structure prediction techniques aim to predict the local secondary structures of proteins based on the knowledge of their sequences. The first generation secondary structure prediction methods in 1960-1970s were based on probabilities of a particular amino acid for a particular secondary structure. The method is referred to as Chou-Fasman method. The second generation methods untill early 1990s consider the local environment of the adjacent residues typically 3-51 segments. Modern protein secondary structure prediction methods are based on exploiting evolutionary information contained in multiple sequence alignments such as recognizing the patterns of hydrophobicity, conservation, sequence edge effects etc. In relation to the databases of known protein structures and modern machine learning methods, such as neural networks, multiple linear regressions, k-nearest neighborhood and support vector machines, the accuracy of secondary-structure prediction is raised up to 80% for globular proteins. Further boost up in prediction accuracy is highly limited by varying local conformations under native conditions which are not reflected in crystal structures due to packing constraints and local contacts. In addition, dramatic conformational changes related to the protein's function or environment could complicate the situation a bit more. There are a number of secondary structure prediction servers such as Network Protein Analysis Server (NPS@). At NPS@, a blastp search of the sequence is performed against the SWISS-PROT database. These results are filtered and then aligned by CLUSTAL Omega. The resulting alignment is the input for the following methods: GOR I, GOR III, GOR IV, DSC, DPM, PHD, Predator, SOPM, HNN. GOR (Garnier, Osguthorpe and Robson) method is one of the popular methods utilizing information theory that predict locations of alpha-helix and beta-strand using amino acid sequence. DSC (Discrimination of protein Secondary structure Class) is based on dividing secondary structure prediction into the basic concepts and then use of simple and linear statistical methods to combine the concepts for prediction. DPM method has been called the 'double prediction method' and consists of a first prediction of the secondary structure from a new algorithm which uses parameters of the type described by Chou and Fasman, and the prediction of the class of the proteins from their amino acid composition. PHD program achieved 70% accuracy level through combining neural networks and evolutionary information. PREDATOR method is based on recognition of potentially hydrogen-bonded residues in a single amino acid sequence. Self-optimized prediction method (SOPM) checks against an updated release of the Kabsch and Sander database, 'DATABASE.DSSP'. The first step of the SOPM is to build sub-databases of protein sequences and their known secondary structures drawn from 'DATABASE.DSSP'. The second step is to submit each protein of the sub-database to a secondary structure prediction tool using a predictive algorithm based on sequence similarity. The third step is to iteratively determine the predictive parameters that optimize the prediction quality on the whole sub-database. The HNN (Hierarchical Neural Network) prediction methods are made up of two networks: a sequence-to-structure network and a structure-to-structure network. Kabsch and Sander in 1983, developed the Dictionary of Protein Secondary Structure known as DSSP which is a set of simple and physically motivated criteria for secondary structure prediction, programmed as a pattern-recognition process of hydrogen-bonded and geometrical features extracted from X-Ray crystallographic coordinates of the atoms of biomolecules. Cooperative secondary structure is recognized as repeats of the elementary hydrogen-bonding patterns "turn" and "bridge." Repeating turns are "helices," repeating bridges are "ladders," connected ladders are "sheets." Geometric structure is defined in terms of the concepts of torsion and curvature of differential geometry. Local chain "chirality" is the torsional handedness of four consecutive C-alpha positions and is positive for right-handed helices and negative for ideal twisted beta-sheets. Curved pieces are defined as "bends."

Trans-membrane Helix Prediction: Experimental determination of the 3D structure of transmembrane protein is a challenge as they do not crystalize and are hardly tractable by NMR spectroscopy. Two major classes of membrane proteins are known as:

(*a*) proteins which insert alpha helices into lipid bilayer (intergral) and

(*b*) proteins that form pores by beta strand barrels (porins).

There is not much experimental information available regarding porins like strand barrels. However, knowing the precise location of transmembrane helics, the 3D structure can be predicted by exploring all possible conformations. The basic properties of transmembrane helices (TMHs) found are:

1. Trans-membrane helices are predominantly apolar and between 12 and 35 residues long.
2. Globular regions between membrane helices are typically shorter than 60 residues.
3. Most proteins having trans-membrane helical regions have a specific distribution of the positively charged amino acids Arginine and Lysine coined the "positive-inside-rule". Connecting loop regions at the inside of the membrane have more positive charges than loop regions at the outside.

4. Long globular regions (>60 residues) differ in their composition from those globular regions subject to the “inside-out-rule”.

Most methods for predictions of the structures of trans-membrane helices simply compile the hydrophobicity along the sequence and predict a segment to be a trans-membrane helix if the respective hydrophobicity exceeds some given threshold. Use of evolutionary information also improves the prediction abilities of trans-membrane helices significantly by carefully filtering the results from PSI-BLAST searches. The trans-membrane topology was predicted from the amino acid sequences by averaging the results from six different programs: DAS, HMMTOP, TMHMM, TMPRED, TOPPRED II and GENEIOUSPRO. Dense Alignment Surface (DAS) optimizes the use of hydrophobicity plots. TMHMM is the most advanced and seemingly the most accurate current method to predict membrane helices. It embeds a number of statistical preferences and rules into a Hidden Markov Model to optimize the prediction of the localization of membrane helices and their orientations. Similar concepts are used for HMMTOP. TMPRED algorithm is based on the statistical analysis of TMbase, a database of naturally occurring trans-membrane proteins. The prediction is made using a combination of several weight-matrices for scoring. TOPPRED II averages the GES-scale of hydrophobicity using a trapezoid window.

Three Dimensional Modelling: Now-a-days in structural biology field macromolecular structure development is necessary for the interactions study. So the determination of three dimensional structures of protein or nucleic acids using different experimental and theoretical technique is mandatory for this purpose. Most widely used experimental techniques for the three dimensional structure predictions are X-Ray Crystallography, NMR and cryo-EM. Both NMR and X-Ray methods of structure determination generate co-ordinate data that are usually deposited in the Protein Data Bank (PDB). However, these techniques require huge time, expertise and resource. With the progression of structural biology research, molecular modelling is becoming an important method of choice to obtain the 3-dimensional structures of biological macromolecules such as Proteins, DNA etc relatively more quickly and easily. It helps to bridge the gap between the available sequence and structure information by providing reliable and accurate models.

There are mainly three major methods for prediction of three dimensional structure of protein.

1. Homology modelling or comparative modelling
2. Theading /fold recognition
3. *Ab-initio* structure prediction

The details of these methods are presented in Chapter 4.

Model Refinement/Optimizaion: The models of the biomolecules obtained from the aforementioned techniques may have many errors. Therefore, it is strongly recommended to check the stereo-chemical qualities of the models. Such methods are called model optimizations. Model optimization is applied for the whole protein not the isolated loop regions. Energy minimization is done either by restraining the atom positions and /or only for a few hundred steps. Molecular dynamics simulation removes big errors but inefficient force-fields may introduce atom clashes or small errors.

Validation of Models: Main sources of errors in Homology Modelling are

1. *Poor alignment:* The target template sequence identity is very less. When the target-template sequence identity is 90% and above then model accuracy is comparable with crystal structures. At the 50-90% identity range there are around 1.5Å rms deviations between the structures of the target and the template with considerable larger local errors but below 25% sequence identity the generated model might have large errors.
2. *Errors in template:* If the template itself contains errors then the generated model quality would be very low.

The structure of modeled protein must be validated through validation server: Structure Analysis and Verification Server (SAVES) which contains software programs for checking and validating protein structures during and after model refinement. The following programs are generally used for checking model qualities.

Procheck: The software tool checks the stereo-chemical qualities of protein structures through Ramachandran Plot by analyzing overall and residue by residue geometry. The plots include: Ramachandran plots, both for the protein as a whole and for each type of amino acid; $\chi^1 - \chi^2$ plots (side chain torsion angles) for each amino acid type; main-chain bond lengths and bond angles; secondary structure plot; deviations from planarity of planar side chains and so on. Another important check is the presence of bad contacts which is the distance between any pair of non-bonded atoms being smaller than the sum of their van der Waals radii.

What_check: The tool provides an enormous number of checks and detailed overall summary of the stereo-chemical parameters of the residues in the model. Directional Atomic Contact Analysis implemented in this program is used to analyze non-bonded contacts.

Verify3D: The tool examines the compatibility of an atomic model (3D) with its own amino acid sequence (1D) by assigning a structural class based on its location and environment (alpha, beta, loop, polar, non-polar etc) and comparing the results to good structures.

ERRAT: The tool analyzes the statistics of non-bonded atom-atom interactions with respet to a database of reliable high resolution structures.

Prove: The tool compares atomic volumes against a set of pre-calculated standard values where atoms are treated like hard spheres and calculates a statistical Z-score deviation for the model from highly resolved and refined PDB-deposited structures.

Rampage: The tool provides stereo-chemical quality of protein structure in Ramachandran Plot, where residues having proper ψ and Φ main-chain torsion angles are clustered in allowed regions and the percentage of residues in the disallowed regions are given.

In this chapter a brief overview of the different aspects of Bioinformatics are provided. The details of these topics are presented in the subsequent chapters.

The analyses of biological data are heavily dependent on mathematics, statistics and computer programming. Therefore, the basics of mathematics, statistics and programming languages (Bio-Perl, Bio-Python and Matlab) are discussed in the following sections.

Mathematics and Statistics in Bioinformatics: Mathematics is routinely used in the analyses of Biological data. The rapid growth of genomic data necessitates the use of Mathematics in Biology.

One of the main applications of Mathematics is the Chaos theory. The Chaos theory is a special field of Mathematics which is used to analyze the dynamical behaviors of Biological Systems. From time immemorial, the biologists have been keeping track of populations of different species by analyzing the population models. Most biological models are continuous. However, recently scientists have started implementing chaotic models in certain populations. A study on models of Canadian lynx showed that there was chaotic behavior in the population growth of the species concerned. Chaos can also be found in ecological systems, such as hydrology. Another important biological application of chaos theory is found in cardiotocography.

Another important application of mathematics in solving Biological problems is in the field of Evolutionary Biology. Evolutionary Biology depends on extensive mathematical theorizing. The traditional approach in this area which involves Mathematics and which includes complications from genetics is population genetics. Most population geneticists propose that the appearance of new alleles occurs by mutation; the appearance of new genotypes occurs by genetic recombination; and changes in the existing allele frequencies and genotypes occur at a small number of gene loci. Considering infinitesimal effects at a large number of gene loci together with the assumption of linkage equilibrium or quasi-linkage equilibrium, the quantitative genetics method was devised. Another important avenue of population genetics is molecular phylogenetics. The traditional population genetics model deal with information associated with alleles and genotypes, and is frequently stochastic.

However, many population genetics models assume that population sizes are constant. The variable sizes of populations, in the absence of genetic variation information, are treated in the field of population dynamics. The first principle of population dynamics is known as the Malthusian growth model. Another important aspect of mathematical Bioinformatics is the Lotka–Volterra predator-prey model. There are various models of the spread of infections which have been proposed and analyzed by Mathematical Biology. These theories provide important results that may be applied to health policy decisions. Another important field of Mathematical Bioinformatics is the evolutionary game theory. This theory involves the selection of inherited phenotypes, without genetic complications. The other avenues of Mathematical Biology are as follows:

- Mechanics of biological tissues
- Theoretical enzymology and enzyme kinetics
- Cancer modelling and simulation
- Modelling the movement of interacting cell population
- Mathematical modelling of cell cycle
- Mathematical modelling of intra-cellular dynamics

Application of Mathematics in the analysis of cell cycle: One of the important aspects of Bioinformatics is the analysis of eukaryotic cell cycle. It is one of the most studied topics and a cell cycle mysregulation leads to cancer. The analysis of cell cycle is performed by Mathematics and involves simple calculus. There is a generic eukaryotic cell cycle model representing a particular eukaryotic cell depending on the values of the parameters like concentrations of the proteins in different cellular organelles.

A set of ordinary differential equations are used to determine the properties of the dynamical systems.

A few typical statistical concepts used in Bioinformatics are:

Hidden Markov Model (HMM): Hidden Markov Model (HMM) is a statistical principle in which the biological system under consideration is being modelled by assuming the system to have unobserved (i.e. hidden) states. In simpler Markov Models, the properties of the biological systems are directly visible to the observer. This would help to identify the transition parameters between the system variables. However, in the Hidden Markov Model, the properties of the biological systems are not directly visible; only the output becomes visible. In a simple language, a *Hidden Markov Model (HMM)* is a statistical Model which is used to describe the evolution of observable events. The observable events depend on internal factors. The internal factors are not directly observable. The observed events are called 'symbol' and the invisible factors underlying the observation are called 'state'. An HMM is made up of two stochastic processes, viz., an invisible process of hidden states and a visible process of observable symbols. HMMs have varied applications in Bioinformatics. They are used in biological sequence analyses, gene finding methods, and protein structure prediction methods to name a few.

Machine Learning: Machine learning is a subfield of computer science. It involves the development of algorithms that learn how to make predictions based on available data. The machine learning has a number of emerging applications in the field of bioinformatics. Before the advent of machine learning algorithms, bioinformatics algorithms had to be explicitly programmed manually. However, this makes the solution of the biological problems extremely difficult. Machine learning techniques help the algorithm to consider the automatic feature learning from the dataset. The machine learning algorithms can learn how to combine multiple features of the input data into a more abstract set of features which can further be manipulated to come up with a generalized Model to solve the problem. Machine learning approaches are known to be a very efficient technique to solve biological problems. However, the most important pre-requisite of the process is that the dataset used for the purpose should be large enough. Machine learning has been applied to the following six main subfields of bioinformatics like: Genomics, Proteomics, Microarrays, Systems Biology, Evolutionary Biology, and Text Mining.

Application of Machine Learning in Genomics: Genomics is the study of the genomes which is the complete DNA sequence of organisms.

After the advent of sophisticated genome sequencing techniques, the number of available sequences is growing exponentially. Though the raw data are constantly being deposited in the databases but it is a challenge to annotate the raw data to extract relevant biological information from the raw sequence data. Unfortunately, the biological annotations of the raw sequence data are a daunting task and require a lot of involvements. This necessitates the use of machine learning algorithms. The machine learning methodologies are able to extract features to identify the specific patterns in the biological sequences. This further would help to explain the plausible functional annotations. One of such examples is the prediction of the locations of the protein-coding genes within a given DNA sequence. This is a problem known as computational gene prediction. The prediction of genes is commonly performed through a combination of extrinsic and intrinsic searches. In case of the extrinsic search, the input DNA sequence is allowed to run through a large database of sequences. The database contains a set of known genes for which sufficient information are already available. The machine learning tools extract features from the known gene sets and use those features to build a model. The new nucleotide sequences are fed into the model generated from the available features as mentioned before and are analysed. Machine learning algorithms are also used for the problem of multiple sequence alignment. Multiple sequence alignment involves aligning many DNA or amino acid sequences in order to determine regions of similarity that could indicate a shared evolutionary history. Multiple sequence alignment can also be used to detect and visualize genome rearrangements.

Application of Machine Learning in Proteomics: Proteins are polymers of amino acids. They are the main functional molecules in the cellular systems. Proteins gain their functions from specific protein folding mechanisms. Through the protein folding, the proteins conform into a three-dimensional structure. This protein structure is composed of a number of layers , viz., the primary structure (i.e. the sequence string of amino acids), the secondary structure (alpha helices and beta sheets), the tertiary structure (the three-dimensional structure of the protein), and the quaternary structure (the relative orientations of the different chains in the protein with respect to each other in the three dimension).

Machine learning algorithms mainly aim to predict the secondary structures of proteins from the amino acid sequences of the proteins. After generating the secondary structures of the proteins, the tertiary and quaternary structures of the proteins are subsequently built. This is a very important avenue in structural bioinformatics as solving the true structure of a protein in-vitro is an incredibly expensive and time-intensive process. Such wet-lab experimental methods would require the Systems that can accurately predict the structure of a protein by analyzing the amino acid sequence directly. As in the case of genome analyses, prior to the advent machine learning techniques, the structure prediction work was performed manually. Today, through the use of automatic feature learning, the best machine learning techniques are known to be able to achieve a prediction accuracy of 82-84%. The current state-of-the-art machine learning technique for prediction of protein structure is referred to as DeepCNF (deep convolutional neural fields). This technique relies on the machine learning model of artificial neural networks to achieve an accuracy of approximately 84% when the method is applied to classify the amino acid sequences of a protein into one of three structural classes (helix, sheet, or coil). Machine learning has also been applied to other proteomics problems such as protein side-chain prediction, protein loop modelling, and protein contact map prediction.

Application of Machine Learning in Systems Biology: Systems Biology is the study of the emergent behaviours obtained from complex interactions of simple biological components in a system. Such components are the biological macromolecules like, DNA, RNA, proteins, and metabolites. Machine learning techniques have been used to aid in the modelling of these complex interactions that occur in the biological Systems in various different domains such as genetic networks, signal transduction networks, and metabolic pathways. A specialized machine learning technique, known as the Probabilistic Graphical Models is used for determining the structure-function relationship between different variables. This approach is one of the most commonly used methods for modelling genetic networks. In addition to that, the machine learning techniques have been applied to other Systems Biology problems such as identification of transcription factor binding sites with the help of a technique known as Markov chain optimization. Genetic algorithm is another machine learning technique which is based on the natural process of evolution. Genetic algorithm has constantly been used to model genetic networks and regulatory structures.

Machine learning tools are also used for various other Systems Biology applications which include the task of enzyme function prediction, high throughput microarray data analysis, and analysis of genome-wide association studies for a better understanding of the potential markers of Multiple Sclerosis, protein function prediction, and identification of NCR-sensitivity of genes in yeast.

Application of Machine Learning in text mining: Now-a-days, there is a tremendous increase in the appearance of biological literature in the form of biological publications. Therefore, it is very much a necessity to extract relevant and important information from a huge collection of available biological publications. This task is referred to as knowledge extraction. The extraction of the relevant information from available biological data is a daunting task. Such extraction process leads to collection of biological data to be fed into a machine learning tool to generate new biological knowledge. Machine learning techniques can be used for this knowledge extraction task with the help of techniques such as natural language processing to extract the useful information from human-generated reports in a biological literature database. Another variation of the text mining technique is called the text nailing which has been applied to the search for novel drug targets. This technique requires the examination of information stored in biological databases and journals. Furthermore, it has been observed that annotations of proteins in protein databases often require extraction of additional information from biological literature. Machine learning techniques have been applied to automatic annotation of the function of genes and proteins for the determination of the subcellular localization of a protein, analysis of DNA-expression arrays, large-scale protein interaction analysis, and molecule interaction analysis.

Final Words: Bioinformatics is a multidisciplinary subject. It requires knowledge from Physics, Chemistry, Biology, Mathematics, Statistics and

Computer Science. Day-by-day it is becoming an important field of study. There are various biological problems. The basic aim of the bioinformatics study is to facilitate the experimental biologists. Bioinformatics experiments are comparatively cheaper and reproducible. At the same time, result from bioinformatics analyses, would provide detailed insights into the problem in hand. The different aspects of bioinformatics will be discussed in the subsequent chapters in a greater detail.

CHAPTER

2

Biological Databases

2.1 INTRODUCTION

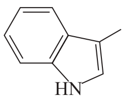
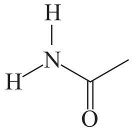
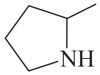
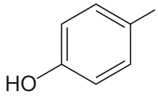
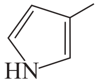
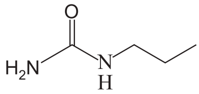
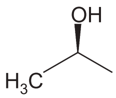
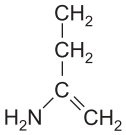
The basic aim of this chapter is to deal with vaious aspetcs of Biological Databases emphasizing on the distinction between datatypes, data organizations and usablity of the data. The chapter focuses on sequence and structural databases. Necessary definitions are also provided. Web addresses of the biological databases are also added.

2.2 SOME IMPORTANT TERMINOLOGIES AND DEFINITIONS

(a) **Amino acids:** Amino acids are biologically important molecules bearing amino (-NH_2) and carboxyl terminal ends (-COOH). The amino acids differ in their side chains. The amino acids are broadly classified as polar charged (acidic, basic), polar uncharged and hydrophobic (aromatic, aliphatic) on the basis of the side chains. Amino acids make peptide linkages among themselves to build the protein chain.

(b) **Representations of amino acids:** Amino acids are represented by single and three letter codes.

Name of the amino acid	Single letter code	Three letter code	Side chain structure	Type
Glycine	G	Gly	H-	Polar neutral
Alanine	A	Ala	CH_3-	Hydrophobic aliphatic
Leucine	L	Leu	$(CH_3)_2$-CH-CH_2-	Hydrophobic aliphatic

Methionine	M	Met	CH_3-S-CH_2-CH_2-	Hydrophobic aliphatic
Phenylalanine	F	Phe	C_6H_5-CH_2-	Hydrophobic aromatic
Tryptophan	W	Trp		Hydrophobic aliphatic
Lysine	K	Lys	NH_2-$(CH_2)_3$-CH_2-	Polar basic
Serine	S	Ser	OH-CH_2-	Polar neutral
Asparagine	N	Asn	H_2N-CO-CH_2-	Polar neutral
Aspartic Acid	D	Asp	HOOC-CH_2-	Polar acidic
Asparagine/ Aspartate	B	Asx		Polar neutral
Proline	P	Pro		Hydrophobic aliphatic
Valine	V	Val	$(CH_3)_2$-CH-	Hydrophobic aliphatic
Isoleucine	I	Ile	CH_3-CH_2-CH-CH_3	Hydrophobic aliphatic
Cysteine	C	Cys	SH-CH_2-	Polar neutral
Tyrosine	Y	Tyr		Hydrophobic aromatic
Histidine	H	His		Hydrophobic aromatic
Arginine	R	Arg		Polar Neutral
Threonine	T	Thr		Polar Neutral
Glutamine	Q	Gln	H_2N-CO-CH_2-CH_2-	Polar Neutral
Glutamic Acid	E	Glu	HOOC-CH_2- CH_2-	Polar acidic
Glutamine/ Glutamate	Z	Glx		Polar Neutral

(c) Databases: Databases are collections of biological data. They store a vast amount data. There are many different types of databases based on the types of information being stored.

2.3 THE PRIMARY SEQUENCE DATABASES

With advent of sequencing techniques, more and more laboratories started working on systematic organizations of available data. The endeavors lead to the inventions of primary sequence databases. The biological databases are built using nucleotide sequences, amino acid sequences. The most important primary databases are:

For Nucleotides

(*a*) EMBL

(*b*) DDBJ

(*c*) NCBI-GenBank

For Proteins

(*a*) PIR

(*b*) SWISS-PROT

(*c*) UniProtKB/TrEMBL

2.4 THE NUCLEOTIDE SEQUENCE DATABASES

The most important DNA sequence databases are EMBL, GenBank and DDBJ. These databases exchange data on a daily basis to ensure a wholesome coverage of the sequence information among all of them.

2.4.1 EMBL-EBI (http://www.ebi.ac.uk/)

EMBL-EBI supplies the nucleotide sequence data from life science experiments. These data are freely available. The database contains sequences submitted directly by scientists individually from genome sequencing groups as well as from the scientific literature and patents. The database collaborates with DDBJ and GenBank and entries are exchanged between these databases on a regular basis.

There is exponential rate of growth of DNA sequence data. Till date the EMBL database contains more than a billion of nucleotide sequence entries.

EMBL can be searched by SRS (Sequence Retrieval System). This is used to link the important DNA and protein sequence databases to analyze the sequence motifs, structure etc. The databases are linked to MEDLINE to search for the available literature references.

Data submission to EMBL

Many journals have made it mandatory for authors to submit the newly found nucleotide sequence information to the EMBL, GenBank or DDBJ database prior to publication in order to ensure its availability to scientists. The data can be submitted via the webtool http://www.ebi.ac.uk/embl/Submission/webin.html.

Database entries produced at the sequencing site can be deposited and updated directly by the submitters using FTP or email. Groups producing large volumes of genome sequence data over an extended period of time are advised to contact the database at ku.ca.ibe@sbusatad.

Sequence manipulations

EMBL-Align is a public data set of both protein and nucleotide multiple sequence alignments. This server can be queried from the EBI Sequence Retrieval Server (SRS) server. It was developed in response to the need for permanent electronic storage and standardized presentation of alignment data from phylogenetic and population analyses. This is a dedicated web-based tool for submission of multiple sequence alignments in all common alignment formats which are available at http://www.ebi.ac.uk/embl/Submission/align_top.html.

Sequence retrieval

The EBI Genomes server at http://www.ebi.ac.uk/genomes/can be used to access the several thousand completely sequenced genomic components. Proteome analysis information on all completely sequenced organisms is available at http://www.ebi.ac.uk/proteome/.

Whole Genome Shotgun (WGS) data

WGS data are available at ftp://ftp.ebi.ac.uk/pub/databases/embl/wgs/. The largest data set is that of *Rattus norvegicus*. While many of the WGS data sets are not annotated, some biological features are present in some of the sets. The WGS data set for *Anopheles gambiae* strain *PEST* is an example with annotation. WGS data sets can now be searched using the FASTA program.

Sequence Retrieval System (SRS)

The EMBL Nucleotide Sequence Database can be accessed via the EBI SRS server at http://srs.ebi.ac.uk/. In SRS, the data are available in the following libraries:

(*i*) EMBL: the database in its entirety by means of a virtual library comprising EMBLRELEASE, EMBLNEW, EMBLTPA and EMBLWGS;

(*ii*) EMBLRELEASE: library containing the latest official release of the EMBL Nucleotide Sequence Database;

(*iii*) EMBLNEW: library containing updated and new entries created since the last official release;

(*iv*) EMBLTPA: library containing TPA entries;

(*v*) EMBLWGS: library containing WGS entries;

(*vi*) EMBLCON: library containing CON entries.

Sequence searching

There is sequence analysis toolbox in EMBL available at http://www.ebi.ac.uk/Tools/. Sequence similarity manipulations can be performed interactively over the WWW as well as by email. Users can search the EMBL Nucleotide Sequence Database as a whole or by individual taxonomic division.

The most commonly used algorithms available are FASTA and WU-BLAST, permitting comparisons between nucleotide query sequences and the nucleotide or protein databases as well as searches of protein query sequences against the nucleotide databases.

The FASTA service for genomes and proteomes (http://www.ebi.ac.uk/fasta33/genomes.html) enables users to search interactively completed genomes and proteomes. The same searches can be performed by email (ku.ca.ibe@atsafpg). User instructions are available by sending an email with the word HELP in the body of the message to ku.ca.ibe@atsafpg. WGS data sets are now available for searching.

Sequence analysis

One of the most important sequence manipulation techniques is the multiple sequence alignment. This is performed by ClustalW. The results from ClustalW can be used in phylogenetic analyses of the species involved. The presence of specific domains and motifs in the amino acid sequences can be identified using InterProScan and others. The EBI also provides interactive sequence analysis resources based on the European Molecular Biology Open Software Suite (EMBOSS) (http://www.emboss.org/).

2.4.2 DDBJ (http://www.ddbj.nig.ac.jp/)

DDBJ stands for the DNA Data Bank of Japan. DDBJ first started in 1986 as a collaborative effort with EMBL and GenBank. The database is maintained by the National Institute of Genetics, Japan. DDBJ accepts nucleotide sequences from the entire world using the world wide web.

DDBJ has several tools for annotations of biological sequence data.

For database search the following tools are available:

getentry: To retrieve sequence data by accession numbers, etc.

ARSA: For all-round Retrieval of Sequence and Annotation

TXSearch: For retrieval of data from unified taxonomy database.

BLAST: For sequnce similarity search

VecScreen: It is a Vector Search to screen contamination in nucleic acid sequences.

DRA Search: This tool is to search metadata by keywords and retrieve data

For Phylogenetic s:

ClustalW: To perform multiple sequence alignment and to draw phylogenetic trees

WABI (Web API for Biology): WABI is the Web API to use DDBJ's data retrieval system without mediating web browsers. Following services provide the Web APIs as well as web browser forms:

WABI BLAST(Instruction in Japanese only)

getentry WebAPI

ARSA WebAPI

For Next Generation Sequencing:

DDBJ Read Annotation Pipeline: This pipeline is used to analyze the High-throughput data obtained from next generation sequencing techniques.

But this tool requires registration.

For Genome Analysis:

MiGAP: This is a mechanical annotation tool for microbial genomes. But this tool requires registration.

GTPS: This tool is for reannotation of bacterial genomes using a new common protocol.

GTOP: This tool is to analyze genome to protein structure and identification of protein function

AOE: Statistics and trends of gene expression data

Gendoo: Functional profiling of gene and disease features for omics analysis

GGGenome: A ultrafast sequence search

GGRNA: A Google-like, ultrafast search engine for genes and transcripts

RefEx: Reference Expression dataset for tissue transcriptome

2.4.3 NCBI-GenBank (https://www.ncbi.nlm.nih.gov/genbank/)

GenBank is the repository of nucleic acid as well as amino acid sequence information maintained by National Institute of Health (NIH), USA. This database contains annotated sequence information of the available sequences. In GenBank the sequence data are updated on a regular basis. A new sequence in GenBank is released approximately in every two months. All the data of the GenBank are available from the ftp site. An important feature of a GenBank entry is that the entry contains release notes. The release notes give the detailed aspect of the current version of GenBank entry. However, the previous release notes for the same GenBank entry also remain available. The GenBank entries can be searched and disseminated freely over the World Wide Web.

GenBank accepts DNA and m-RNA sequences. It also accepts un-annotated sequence information obtained from Next Generation Sequencing (NGS) methods but such sequences are stored in Sequence Read Archive. In no circumstances GenBank accepts the following data types: Non-contiguous nucleotide or amino acid sequences, nucleic acid sequences of primer regions, amino acid sequences of proteins having missing nucleotide sequence submission, a mixture of mRNA and DNA sequences.

Expressed sequence tags and sequences obtained from genomic surveys are accepted in GenBank. GenBank also stores information about the whole genome sequences of organisms.

Each sequence entry in GenBank is given an accession number which serves as a unique identifier of the sequence.

There are various tools available in GenBank to deposit the sequence from a user. The tools are: BankIt, Sequin, tbl2asn, Submission Portal, Barcode Submission Tool.

A GenBank entry is linked to many other databases for cross references. The most important aspect of a GenBank entry is that it is fully annotated and all the annotations are self-explanatory.

2.5 FINAL WORDS

There are different nucleotide sequence databases available. The databases are comprehensive and provide a large chunk of sequence information. However, the basic problem of all these databases is the use of different file formats. Thus, to maintain a uniformity between the different data file formats, a new initiative has been adopted and it is called The International Nucleotide Sequence Database Collaboration (INSDC). It is available at http://www.insdc.org. INSDC is created on the basis of its partnership with various other databases. The basic spirit of the database is to come up with the proper establishment of specific formats and protocols of collecting nucleotide sequence information. INSDC also aims to incorporate the bibliographic information of the sequence and most importantly a uniform accession number for all the entries in the database which would make it easier to retrieve the sequence. INSDC has collaborations with EMBL, DDBJ and NCBI-GenBank.

2.6 THE PROTEIN SEQUENCE DATABASES

Like nucleotide sequence databases, there are many different databases that store information about the amino acid sequences of proteins. The most important sequence databases are PIR, SWISS-PROT and UniProtKB/TrEMBL.

2.6.1 PIR (http://pir.georgetown.edu/)

It is an amino acid sequence database. It is maintained by Georgetown University Medical Center. This database is an integrated repository of protein information. It supports researches in genomics and proteomics. PIR stores the protein information in a database called the Protein Sequence Database (PIR-PSD). PIR-PSD is a sub-database in the domain of PIR and it has amino acid sequences of proteins and the amino acid sequnces are properly annotated. The database generally covers the amino acid sequences of nealy all the taxonomic groups. The amino acid sequences are grouped into superfamilies, families and domains. These annotations provide an in-depth knowledge of the protein product from a gene. Not only the sequence level information about the proteins, the

PIR database is also linked with litterature database to search for suitable bibliographic references for the proteins and their annotations. As sub databases PIR stores some other important information in terms of:

NREF: It is a non-redundant reference database. This database maintains constant cross-talking with other existing databases like Swiss-Prot, PDB etc. The NREF database can be used in searching for sequences using various sequence analyses tools like BLAST.

iProClass: This is an integrated database of protein family, function, structure information, pathways, interactions, genes and gene ontology information and taxonomic results.

iPTMnet: This is a resource that provides information about the various types of post-translational modifications in proteins. This resource covers the Systems Biology aspects of proteins.

***i*ProLINK (*i*ntegrated Protein Literature, INformation and Knowledge)**: This is a resource that links the litterature references of proteins with the sequence information available in PIR. The resource is aimed at providing a wholesum knowledge of the proteins.

PIRSF: This is a tool that is meant to gather information from other resources, viz., UniProtKB in a comprehensive manner. This tool is also used to derive evolutionary relationships from the amino acid sequences present in the UniProtKB. One of the interesting features of the tool is to analyze the protein relationships using the entire protein. This is unique in the sense that other such resources perform the task using only the specific domain information in the protein. The tool helps in proper annotations of biological functions of proteins for which specific domain information is not available.

PRO: This tool is called the protein ontology tool. This tool is specific for analyzing the gene ontological details of the available amino acid sequences of the proteins. The ontological information provide information about the specific gene ontology terms like protein complex clusters, protein othologous isoforms etc. The tool also provides information about the protein product in taxon specific and taxon neutral manner. There are sub ontologies of PRO. They are

ProEvo: This sub-tool of PRO provides the evolutionary information of proteins on the basis of available evolutionary relationships.

ProForm: This sub-tool of PRO provides protein related information from available gene locus.

ProComp: This sub-tool of PRO analyzes the protein complexes and provides the information regarding the various aspects of the protein complexes.

There are other resources available in PRO. They are as follows:

Text Search: This is to identify the details of a specific annotation. This can be done for a single annotation or for a number of annotations commonly referred to as the Batch search.

Sequence analysis: This is done by the tool Peptide Match. The tool is used to match a specific peptide sequence with the existing sequences present in the databases. The commonly used databases for searching purposes are UniProtKB and UniRef100.

There are other tools available for sequence analysis. They are tools for multiple sequence analysis and pairwise sequence analysis. The multiple sequence analysis is commonly performed by ClustalW. For pairwise alignment Search is used.

The theoretical molecular weight of a protein can be computed from the amino acid sequence of a protein. Using this concept a tool was developed and the tool is used in PRO.

The information regarding the same protein are stored in different databases and these proteins bear different identifiers. The mapping of the different protein identifiers can also be done with the tools available in PRO.

2.6.2 SWISS-PROT (http://web.expasy.org/docs/swiss-prot_guideline.html)

One of the vital protein sequence databases is SWISS-PROT. SWISS-PROT is a database which contains the protein sequence information alongwith all the necessary annotations. The database aims to provide high-end sequence information about proteins, which include the functional characterizations, domain information, post-translational modifications, information about protein variants and so on. SWISS-PROT tries to minimize the data redundancy and is linked to many other databases.

The organization of the database and the style of annotations in the SWISS-PROT make it one of the most popular protein sequence databases. The most important part of a SWISS-PROT entree is the CC and FT lines. The CC lines provide the detailed comments regarding the protein. It gives us the ideas about the protein function, its sub-cellular location, the post-translational modifications associated with the protein, its tissue specificity and so on. FT lines represent the structural details of the protein, like the local secondary structure, the ligand binding sites, presence of transmembrane domains. Most importantly, the SWISS-PROT provides the three-dimensional structural details of proteins if available.

2.6.3 TrEMBL (http://web.expasy.org/docs/swiss-prot_guideline.html)

SWISS-PROT is a protein sequence database which provides a high level of annotations. The database stores data pertaining to the functional description of a protein, the presence of domains in the protein, post-translational modification sites in the proteins and the different isoforms of the protein. The database contains a minimal level of redundancy and is linked to other databases. **TrEMBL**, on the other hand, is a computer-annotated supplement of SWISS-PROT. This database contains all the translations of EMBL nucleotide sequences which are not yet integrated in SWISS-PROT database. TrEMBL data are divided into two main sections: SP-TrEMBL and REM-TrEMBL. SP-TrEMBL (Swiss-Prot TrEMBL) contains the data that will be eventually incorporated into Swiss-Prot with Swiss-Prot accession numbers.

SP-TrEMBL data are arranged in subsections as follows:

arc.dat: This is for Archaea.

fun.dat: This is for Fungi

hum.dat: This is for Human

inv.dat: This is for Invertebrates

mam.dat: This is for Other Mammals

mhc.dat: This is for MHC proteins

org.dat: This is for Organelles

phg.dat: This is for Bacteriophages

pln.dat: This is for Plants

pro.dat: This is for Prokaryotes

rod.dat: This is for Rodents

unc.dat: This is for unclassified results

vrl.dat: This is for Viruses

vrt.dat: This is for Other Vertebrates

REM-TrEMBL is also referred to as REMaining TrEMBL database which stores the data that are not ultimately included in Swiss-Prot. Consequently REM-TrEMBL entries are not given any accession numbers.

2.6.4 The Reference Sequence Database (RefSeq) (https://www.ncbi.nlm.nih.gov/RefSeq)

The Reference Sequence (RefSeq) database comprises of data from publicly available nucleotide sequences (DNA, RNA) and their protein products. These data are already curated and annotated. This database is hosted by National Center for Biotechnology Information (NCBI). The unique feature of the database is that it stores only a single record for each biomolecule (DNA, RNA or protein). The data are obtained from a variety of organisms involving viruses to bacteria to eukaryotes. The database contains non-redundant sequence data

The database provides separate and linked records for the genomic DNA, the gene transcripts, and the protein products from the genes. However, RefSeq captures data from a limited number of organisms for which sufficient data are available.

Accession type	Description of the term
AC_	The entry having this accession code signifies complete genomic molecule, usually alternate assembly

NC_	The entry having this accession code signifies complete genomic molecule, usually reference genome assembly. The reference genome is selected as a representative from the set of best genomic sequences available
NG_	The entry having this accession code signifies incomplete genomic molecule
NT_	The entry having this accession code signifies genomic contigs obtained from cloning experimentations. A sequence contig is a continuous stretch of sequences bounded by gap or a run of 10 or more same nucleotides
NM_	The entry having this accession code signifies data pertaining to m-RNA molecules
NR_	The entry having this accession code signifies data pertaining to non-coding RNA molecules
XM_	The entry having this accession code signifies data pertaining to predicted protein coding m-RNA molecules
XR_	The entry having this accession code signifies data pertaining to predicted non-coding RNA molecules
AP_	The entry having this accession code signifies data pertaining to proteins
NP_	The entry having this accession code signifies data pertaining to proteins obtained from genomes represented by AC_ or NC_

2.6.4 Interpro Database (http://www.ebi.ac.uk/interpro/)

InterPro database consists of data required for functional analysis of protein sequences. These are done by classifying the protein sequences into different protein families on the basis of the presence of functional domains and other important sites. For classification purposes, the proteins are classified using predictive models, called signature sequences. The signature sequences were obtained from different other databases. The database has a facility that allows a user to search for the signature sequences in proteins using a tool called InterProScan. The most important aspect of the database is that it collects information from various other resources and combines them in a single comprehensive form. The Interpro entry is classified into mainly four different categories. They are

Homologous Superfamily: This represents a group of proteins that orginate from a common ancestor as reflected by similarity in their three-dimensional structures. However, the members of the superfamily often have very low similarity at the sequence level.

Family: A protein family in Interpro database consists of a group of proteins that orginate from a common ancestor as reflected by their functional similarities. The proteins belonging to the same family have similarities in their amino acid sequences or in their tertiary structure.

Domain: A protein domain in Interpro database is a distinct functional, structural or sequence unit. Such a unit may exist in a variety of biological contexts.

Repeat: A repeat in an Interpro entry is a short stretch of sequence. Such a sequence is found to be repeated within a protein.

Site: A site in an Interpro entry is another short stretch of sequence. It contains one or more conserved residues and would reflect the presence of active sites, binding sites, post-translational modification sites and conserved sites.

The usefulness of Interpro database is manifold:

(*a*) It is comprehensive and non-redundant.

(*b*) It gathers information from various other resources and makes a single searchable form.

(*c*) The database helps users to annotate their queries related to proteins. The Interpro database can be used to analyse the presence of specific domains or motifs in a protein. However, Interpro database can not be used for genome annotations. The database also is not of much help to perform structural alignments of proteins.

2.7 THE PROTEIN STRUCTURE DATABASE

One of the most important biological databases is the protein sequence-structure database, known as the Protein Data Bank (PDB: https://www.rcsb.org/pdb/home/home.do). PDB is a repository that stores the three-dimensional coordinates of the atoms of the amino acid residues in proteins, and protein-ligand complexes. The three-dimensional structures of the proteins are determined by X-ray crystallography, NMR, Electron Microscopy and a combinations of all these of structure determination tools. The database stores information from all forms lives like bacteria, yeast, plants, flies, other animals and humans. The database not only stores static information but contains several tools to process and analyze the data. In general, a PDB entry contains the sequence and structural information together. It also provides the experimental details of the structure determination. The amino acid sequences in a PDB entry are linked to their corresponding structural details. A PDB entry contains the necessary annotations which provide information for the structure and function of the protein. A PDB entry is linked to various other structural databases like CATH and SCOP as well the litterature database Pubmed. The tools available in PDB would help to find the proteins having similar sequences and structures to an existing protein. The pdb data can be freely downloaded for analysis purposes.

2.7. A. CATH DATABASE

The CATH or Class-Architecture-Topology-Homology-S.O.L.I.D database is a freely available online repository for providing information on the evolutionary relationships of protein domains. The protein domains obtained from the protein structures deposited in the PDB are classified within the CATH structural model as:

CLASS: At the Class (C) level, the protein domains are determined according to the contents of their secondary structural elements. The Class level in CATH has the following categories: all alpha, all beta, a mixture of alpha and beta, or little secondary structure.

Architecture: At the Architecture (A) level, the protein domains are classified on the structural arrangements of their secondary structural elements in three-dimensional space.

Topology/Fold: At the Topology/Fold (T) level, the protein domains are classified on the mode of connections between the different secondary structural elements.

Homology: The protein domains are classified to be belonging to the same three-dimensional space if they are related by evolution.

Sequence Family (S35): The protein domains are classified as belonging to the same sequence family if they have >= 35% sequence similarities among themselves.

Orthologous Family (O60): The protein domains are classified as belonging to the same orthologous family if they have >= 60% sequence similarities among themselves.

Like (S95): The protein domains are classified as belonging to having the same domain structure if they have >= 95% sequence similarities among themselves.

Identical (S100): The protein domains are classified as belonging to having exactly the same domain structure if they have >= 100% sequence similarities among themselves.

Domain Counter (D): The protein domains which are unique in nature are classified as belonging to DOMAIN COUNTER.

If there are some sequence data available with no experimental validations of their structural classes, the tool Gene3D , which is a part of CATH database, is used to annotate the sequence data.

2.7. B. SCOP DATABASE

The SCOP database is for the structural classification of proteins. The SCOP database has the following components: Protein types, Evolutionary events, Structural classes and Protein relationships.

Protein Types: In this category the proteins are classified into the following classes: soluble, membrane, fibrous and intrinsically disordered. The protein type class is based on the characteristic sequence and structural features of proteins.

Evolutionary Events: In this category, the various types of structural aberrations in proteins belonging to the same class are considered.

Structural Class: The proteins are classified as per the presence of different types of secondary structural elements in them. The structural classes are all alpha, all beta, alpha beta and alpha/beta.

Family: Proteins are considered to be belonging to the same structural families if they have the same structural domains.

Superfamily: Proteins having the same structural domains but belonging to different structural classes are considered to be in the same superfamily.

Hyperfamily: Proteins belonging to different superfamilies but having the same structural domains are considered to be present in the same Hyperfamily.

The SCOP is a very useful database to analyze the structural features of a protein.

2.7. C. FSSP DATABASE

The FSSP database is a collection of three-dimensional folds of proteins obtained from the Protein Data Bank. The database stores the data based on all-with-all structural comparisons of the proteins present in the Protein Data Bank. The database contains different protein chains belonging to different fold families. The mutual sequence identities between the amino acid sequences of the proteins belonging to each fold fanily are <25% among themselves. In the database, the data are stored as per the three-dimensional coordinates of the atoms of the amino acid residues.

2.8 COMBINATION DATABASES

There are some other databases which provide information about all types of biological systems. Some of those databases are Gene Ontology Database (GO Database), DAVID (the **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery), Pubmed,

2.8.1 Gene Ontology Database (GO Database) (http://www.geneontology.org/page/go-database)

It is a major computational initiative to collect and present the different features of genes and gene products from all species. The main aims behind the development of the database are manifold.

1. To maintain and develop a controlled vocabulary of various features of genes and gene products

2. To extract structural and functional features from the genes and gene products and to present the data in unified manner
3. To provide different tools for easy access of all aspects of the genetic data

The term Ontology is defined as the representation of things that can be detected by means of experimentations. The GO project comes up with such representations of the biological terms associated with gene product. The GO covers mainly the three domains:

- **Cellular component**: It is the regions inside or outside of a cell where the gene product exhibits its function
- **Molecular function**: It is the biological activity of the gene product
- **Biological process**: It is the biological pathway where the gene product exhibits its functionality.

2.8.2. DAVID (the Database for Annotation, Visualization and Integrated Discovery) (https://david.ncifcrf.gov/home.jsp)

DAVID (the **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery) is a repository developed by the Laboratory of Immunopathogenesis and Bioinformatics (LIB). The tools present in the DAVID Bioinformatics Resources help the user to provide functional annotations of large lists of genes derived from genomic studies, e.g. microarray and proteomics studies. The DAVID Bioinformatics Repository comprises of the DAVID Knowledgebase and five other integrated, web-based functional annotation tools:

- DAVID Gene Functional Classification Tool
- DAVID Functional Annotation Tool
- DAVID Gene ID Conversion Tool
- DAVID Gene Name Viewer
- DAVID NIAID Pathogen Genome Browser.

The DAVID repository can be used in the following purposes:

- To identify enriched biological annotations of proteins
- To discover functional aspects of genes and gene products
- To analyze the cluster redundant annotation terms
- To visualize genes involved in metabolic pathways
- To search for functionally related genes
- To analyze a list of interacting proteins
- To explore gene names by batch processing
- To link a gene with diseases
- To identify protein functional domains and motifs
- To study literature references
- To convert gene identifiers from one type to another.

2.8.3. Pubmed (www.ncbi.nlm.nih.gov/pubmed/)

PubMed is a free search engine to access the MEDLINE database of references and abstracts of literatures on life sciences and biomedical topics. The United States National Library of Medicine (NLM) at the National Institutes of Health holds the database as part of the Entrez system for the retrieval of literature data.

MEDLINE database through Pubmed provides access to:

- Very old literature references from the print version of I*ndex Medicus,* back to 1951 and even earlier
- Literature references to some specific journals before they were indexed in Index Medicus and MEDLINE, for instance S*cience,* B*MJ,* and A*nnals of Surgery*
- Very recent literature entries to records for an article before it is indexed with Medical Subject Headings (MeSH) and added to MEDLINE
- In addition to journal references MEDLINE provides access to a collection of books for which full-texts may be available.

A sister database of Pubmed is PubMed Central (PMC). It is a free digital storehouse that contains publicly accessible full-text scholarly articles published within the biomedical and life sciences journal literature. PubMed Central is much more than just a database of documents. Submissions of any literature data into PMC would undergo proper indexing and formatting resulting in a proper user-friendly appearance of the literature.

FINAL WORDS

The biological sequences are not limited to nucleic acids and proteins but there are carbohydrates as well. None-the-less, the advents of new sequences are inreasing day-by-day. The databases are used to store the large numbers of sequence information. The databases are constantly being upgraded and there are cross-talks between the different databases. However, not all the databases have the proper annotations of biological data. The authenticity of the data may therefore be checked before experimentations.

A few other important biological databases are mentioned in the following sections:

Carbohydrate Structure Databases

1. **EuroCarbDB:** A database for both sequences and structures of carbohydrates.
2. **Carbohydrate Structure Database (CSDB):** This database stores sequence and structural information about naturally occurring carbohydrates from bacteria, plants and fungi.

Signal Transduction Pathway Databases

1. **Cancer Cell Map:** This is a database containing a selected set of browsable and searchable human cancer pathways.
2. **Netpath:** This database is a curated resource of signal transduction pathways occurring in humans
3. **NCI-Nature Pathway Interaction Database:** This pathway provides information regarding biomolecular interactions

4. **Reactome:** This database provides the information regarding the biological pathways that include metabolic processes to hormonal signalling.
5. **SignaLink Database:** This database is an integrated repository to analyze signaling pathways, cross-talks, transcription factors, miRNAs and regulatory enzymes.

Protein-Protein and Other Molecular Interactions

1. **BIND:** This is a Biomolecular Interaction Network Database:
2. **BioGRID: I**t is a database for Interaction Datasets
3. **DIP:** This is a Database of Interacting Proteins
4. **IntAct molecular interaction database:** This is a database for analysis of protein–protein, protein–small molecule and protein–nucleic acid interactions.
5. **STRING:** This database is for analyses of protein interaction networks.
6. **MINT:** This database is for analysis of bio-molecular interaction data.
7. **Pfam:** A protein familiy database
8. **PROSITE:** A database which stores protein motifs
9. **CDD:** A database which stores the information about conserved domains in proteins
10. **PRINTS:** A database that contains information about the fingerprint regions in protein sequences. The fingerprint regions are particular sequence motifs in the proteins which confer them with specific characteristics.

SOME IMPORTANT QUESTIONS AND ANSWERS

1. Margaret Dayhoff built the first protein sequence database called

(a) SWISS PROT

(b) PDB

(c) Atlas of protein sequence and structure

(d) Protein sequence databank

2. Each record in a database is known as

(a) entry (b) file

(c) record (d) ticket

3. Literature databases are

(a) MEDLINE and PubMED (b) MEDLINE and PDB

(c) PubMED and PDB (d) MEDLINE and PDS

4. FlyBase is a specialized database containing information about

(a) biodiversity (b) model organism

(c) literature (d) biomolecules

5. Which of the following is known as E. *coli mo*del organism database

(a) EcoGene (b) EcoBase

(c) EcoSeq (d) ColGene

6. Which of the following is a database of protein sequence

(a) DDBJ (b) EMBL

(c) GenBank (d) PIR

7. GenBnak, the nucleic acid sequence database, is housed and maintained by

(a) Brookhaven laboratory

(b) DNA Database of Japan (DDBJ)

(c) European Molecular Biology Laboratory (EMBL)

(d) National Centre for Biotechnology Information (NCBI)

8. Submission to GenBank is done via

(a) BankIt and Sequin

(b) BankIt and BankIn

(c) Sequin and BankIn

(d) Entrez

9. STAG is housed and maintained by

(a) Brookhaven laboratory

(b) DNA Database of Japan (DDBJ)

(c) European Molecular Biology Laboratory (EMBL)

(d) National Centre for Biotechnology Information (NCBI)

10. A database of human genetics and molecular biology is

(a) PDB

(b) STAG

(c) OMIM

(d) PSD

11. Find the odd one out

(a) PIR

(b) PSD

(c) SWISS PROT

(d) EMBL

12. The information retrieval tool of NCBI GenBank is known as

(a) Entrez

(b) STAG

(c) SeqIn

(d) text search

13. The first invented secondary database was

(a) PRINTS

(b) PROSITE

(c) PDB

(d) PIR

14. Find out the sequence alignment tool

(a) BLAST

(b) PRINT

(c) PROSITE

(d) PIR

15. MAtDB contains data for

(a) Mouse (b) Human

(c) *E.coli* (d) Arabidopsis

ANSWERS

1. (c)	**2.** (a)	**3.** (a)	**4.** (b)	**5.** (a)
6. (d)	**7.** (d)	**8.** (a)	**9.** (b)	**10.** (c)
11. (d)	**12.** (a)	**13.** (b)	**14.** (a)	**15.** (d)

CHAPTER

3 Biological Sequence Alignment

3.1 INTRODUCTION

Biological Sequence Alignment is one of the most established and effective protocols in Bioinformatics. The final goal of the sequence alignment techniques is to find similar sequences to an existing unknown sequence to come up with some idea about the possible structure-function activity of the unknown sequence. However, the sequence alignment techniques have manifold applications, viz.,

(a) Gene Finding

(b) Functional Prediction

(c) Assembly of Genome Sequences

3.2 SOME IMPORTANT TERMS:

The phenomenon of Biological Sequence Alignment often encompass the following terms:

(a) Query Sequence: It is the unknown sequence. It is the sequence which is given as an input to find suitable other sequences from the databases.

(b) Subject: It is the sequence present in the database with which the unknown sequence is compared.

(c) Pair-wise Sequence Alignment: It is the alignment of two sequences together.

(d) Multiple Sequence Alignment: It is the alignment between more than two sequences.

(e) Global Sequence Alignment: It is the alignment between sequences encompassing the entire lengths of the sequences. The alignment may contain matches, mismatches and gaps.

Seq1: A T T T T G G G C A C G

Seq2: A T T T T G - - C T C G

In the above sequence alignment, there are 9 exact matches, 1 mismatch and two gaps. Gaps would represent deletion mutations. In the above alignment there are two deletions in Seq2 corresponding to the regions of gaps. It can also be mentioned that in Seq1 there are two insertions. The gaps are therefore referred to as INDEL (Insertion Deletion). There is one mismatch where there is an A in Seq1 and T in the corresponding place in Seq2.

(f) Local Sequence Alignment: It is the alignment of parts of sequences. Local alignment is generally used to identify the similarities in some specific regions of proteins, like the active sites.

(g) Score of the Sequence Alignment: It is the measure of the effectiveness of the sequence alignment. The total score of a sequence alignment is the sum total score obtained from the individual positions and includes matches, mismatches and gaps. The scores are calculated using available substitution matrices.

(h) Substitution Matrix: A substitution matrix is a collection of scores for aligning biological sequences (nucleotides or amino acids) with one another. These scores generally reflect the relative ease with which one nucleotide or amino acid may mutate into or substitute for another nucleotide or amino acid. These scores are used to measure similarity in sequence alignments. Two of the most important substitution matrices are PAM and BLOSUM.

(i) PAM Matrix: PAM or Point Accepted Mutation is one of the first amino acid substitution matrices. This matrix was built by calculating the differences in closely related proteins. PAM1 matrix would therefore reflect that there is 1% change in the sequences of the amino acid residues in proteins belonging to the same family. PAM1 matrix is used as the basis to calculate the other matrices. The higher the number of the PAM matrix, the lower is the sequence similarities between the sequences.

(j) BLOSUM Matrix: BLOSUM or Block Substitution Matrix is obtained by calculating the differences in evolutionarily divergent proteins. The substitution probabilities are computed using the conserved blocks in the multiple sequence alignments of evolutionarily divergent proteins. It is therefore an empirical rule to use a higher numbered BLOSUM matrix for aligning two closely related sequences and a lower number for more divergent sequences.

(k) E-value: This is the Expectation value. It signifies the reliability of a sequence alignment. The lesser the E-value is the higher is the chance of having a significant sequence alignment.

(l) Low-Complexity region: A Low-Complexity region in a sequence alignment represents a series of repeated amino acids or nucleic acids which do not make any specific sense. Such Low-Complexity regions must not be taken into consideration while performing the sequence alignments.

3.3 BIOLOGICAL IMPORTANCE OF SEQUENCE ALIGNMENT

Sequence alignments are very much useful in bioinformatics to identify the sequence similarity and to produce phylogenetic trees. The technique is also used to build homology models of protein structures. However, the biological significance and relevance of sequence alignments are not always clear. Alignments are often assumed to reflect some degree of evolutionary changes between sequences descended from a general common ancestor. However, it is formally possible that convergent evolution could occur to generate apparent similarity between proteins that are evolutionarily unrelated but still are able to perform similar functions having similar structures.

3.4 SEQUENCE ALIGNMENT ALGORITHMS

There are different methods of biological sequence alignment techniques. The techniques include:

Dot-Plot: This is probably the oldest way of comparing biological sequences. A dot plot is a visual representation of the comparison between two sequences. It looks like a rectangular array where each axis of the array represents one of the two sequences to be compared. First, a window length is fixed to compare the sequences. Whenever one window in one sequence resembles another window in the other sequence, a dot or short diagonal is put at the corresponding position of the rectangular array. Thus, a diagonal line is drawn when two sequences share similarities over their entire length from one corner of the dot plot to the diagonally opposite corner. This might be considered as the global alignments between the sequences.

Dynamic Programming Algorithm: In biological sequence alignments, the algorithm proposed by Needleman–Wunsch, is considered to be one of the most important useful applications of Dynamic Programming Algorithm. The algorithm first breaks the entire sequence into small pieces and compares the sequences to come up with a solution which are then joined to build the total alignment. It is considered to be the best option for global alignment. The algorithm can be explained by the following example:

Step1: Two DNA sequences (S1 = GCATGCU and S2 = GATTACA) are considered. A table is made with the DNA sequences as follows:

		G	C	A	T	G	C	U
G								
A								
T								
T								
A								
C								
A								

Step 2: The best possible arrangement between S1 and S2 are calculated as:

S1: GCATG-CU

S2: G-ATTACA

The table is then filled up as follows:

		G	C	A	T	G	C	U
	0	–1	–2	–3	–4	–5	–6	–7
G	–1							
A	–2							
T	–3							
T	–4							
A	–5							
C	–6							
A	–7							

In the next step, the scores to be used for filling up the table are to be calculated either from top, left or top-left (diagonal). The scores calculated from top and left would result in Indels. However, only a score calculated from the diagonal would represent an exact match between the sequences. We need to keep a track of the cell which is used to fill in the rows. As there are no top or top left cells for the second row, -1, -2 etc. are used to fill the row. The same is true for the second column as well. Following this rule, the entire table is filled up as follows:

		G	C	A	T	G	C	U
	0	–1	–2	–3	–4	–5	–6	–7
G	–1	1	0	–1	–2	–3	–4	–5
A	–2	0	0	1	0	–1	–2	–3
T	–3	–1	–1	0	2	1	0	–1
T	–4	–2	–2	–1	1	1	0	–1
A	–5	–3	–3	–1	0	0	0	–1
C	–6	–4	–2	–2	–1	–1	1	0
A	–7	–5	–3	–1	–2	–2	0	0

Smith-Waterman Algorithm: The Smith-Waterman algorithm is basically used for local sequence alignment. Local sequence alignment aims to find local regions of similarities between nucleotide or protein sequences. The algorithm is simple and analogous to the Needleman–Wunsch algorithm. The difference lies in the fact that instead of using negative numbers in the cells, the cells are filled with zero values. The basic motivation behind the Smith-Waterman algorithm is to provide reliable degrees of similarities between local regions of nucleotide or protein sequences which are otherwise distantly related. Local alignment techniques are used in FASTA and BLAST.

In database search techniques, such as BLAST, statistical methods are used to evaluate the likelihood of a particular alignment between regions of sequences arising by chance given the size and composition of the database being used for searching. In particular, the probability of finding a given alignment by chance increases if the database consists only of sequences from the same organism as used in the query sequence. Repetitive sequences in the database or query can also distort both the search results and the assessment of statistical significance scores. BLAST automatically filters such repetitive sequences in the query to avoid apparent hits that are statistical artifacts. There are several versions of BLAST alignments.

- **blastp:** This tool uses the amino acid sequence of a protein as the query sequence to find its similar proteins in a protein sequence database.
- **blastn:** This tool uses the nucleotide sequence as the query sequence to find its similar nucleotides in a nucleotide sequence database.
- **blastx:** This tool uses a nucleotide sequence as the query sequence which is translated in all reading frames to generate a protein which is then used as the query sequence to search for similar sequences against a protein sequence database.
- **tblastn:** This tool uses a protein sequence as the query sequence which is translated in all reading frames to generate a nucleotide

sequence which is then used as the query sequence to search for similar sequences against a nucleotide sequence database.

- **tblastx:** This tool uses the nucleotide query sequence with six frame translation against the six-frame translations of a nucleotide sequence database.
- **BLAST2:** This is a tool that allows the user to upload his or her own sequence of interest to be compared with a sequence database or with another sequence.

3.5 STRUCTURAL ALIGNMENT

Structural alignments are the techniques to determine the similarities between the structures of proteins or RNAs. These methods use information about the secondary and tertiary structure of the protein or RNA molecule to identify the aligned regions in the molecules. Such methods consider not only the structures but also the sequences of the concerned molecules. The structural alignment methods typically come up with a local sequence alignment between the biomolecules under considerations. The local sequence alignment is produced as a part of the structural alignment results. The methods are generally used for comparison between two or more structures of biomolecules and their corresponding sequences. The only limitation of the work is that the alignment methodologies are dependent on the availability of structural information such that they can only be used for the analysis of sequences of biomolecules for which structural information are already available. The structural information is generally obtained from X-ray crystallography or NMR spectroscopy. It is, however, an important fact that both the protein and RNA structures remain highly conserved during evolution. Therefore, alignments involving structural information can be considered to be more reliable in case of sequences which are obtained from very distantly related organisms. Such sequences are known to be diverged so extensively that sequence comparison cannot reliably detect their similarities.

The most important aspect of structure based sequence alignments is that such alignments can be used as the "gold standard" for the determination of protein structures by homology modelling techniques for which sequence similarity between the target and template is the most vital information. In case homology-based protein structure prediction methodologies, the structurally similar regions in proteins should align properly in their sequence parts. Such structurally conserved regions, if found to be conserved in sequences as well, would provide additional reliability of the structure prediction method. However, the structural similarity scores cannot be used in homology based protein structure prediction methods directly as the target sequence does not possess any structure.

The most important methods for structural alignments are: DALI, SSAP and Combinatorial Extensions.

DALI (http://ekhidna.biocenter.helsinki.fi/dali_lite/start)

DALI search method is one of the most important structural alignment tools. This method generates a distance matrix by considering the amino acid sequences extracted from the structures of the proteins provided as inputs. This method generates the structural alignment between the proteins and the method of identification of structural similarities depends on the contact similarity patterns generated from the amino acid sequences of the proteins. The entire protein sequences are divided into several hexa-peptides and a similarity analysis is done between these hexa-peptides. Then, the DALI program creates pair-wise sequence alignments between the proteins and the process is repeated to find neighboring sequences from the protein data bank. The tool then constructs a FSSP which is a fold classification method using structural alignments between neighboring protein structures.

SSAP (https://doi.org/10.1016/S0076-6879(96)66038-8)

SSAP is the Sequential Structure Alignment Program. It is based on dynamic programming algorithm for alignments of structures. This method is based on the analysis of atom-to-atom vectors in structure space as comparison points. It has been extended since its original description to include multiple as well as pairwise alignments, and has been used in the construction of the CATH (Class, Architecture, Topology, Homology) hierarchical database classification of protein folds.The CATH database can be accessed at CATH Protein Structure Classification.

FINAL WORDS

There are many methods to determine the biological sequence alignments. All the methods are fairly accurate. The approximations used in the algorithms are also quite good. The new methods that are being developed are extensions of the previous methods. Not only the sequences of DNA and proteins are analyzed but the sequences of RNA, such as expressed sequence tags and full-length mRNAs, are also aligned to sequenced genomes in order to find the locations of genes. This sequencing also provides information about alternative splicing and RNA editing. The biological sequence alignment techniques are also important parts of genome assembly. In genome assembly techniques, the sequences are aligned to the whole genomes to find overlapping portions. These would provide information about contigs. The contigs are long stretches of sequences. The analysis of nucleic acid sequences would also provide information about the Single Nucleotide Polymorphism (SNP).

The techniques for analyses of the biological sequences have found their applications in non-biological fields as well. The most important of such applications is in natural language processing and in social sciences. In these fields, the Needleman-Wunsch algorithm is usually used to identify the optimal matching. The word selection methods in natural-language processing were dependent on the process of multiple sequence alignment techniques from bioinformatics to generate linguistic versions of computer-generated mathematical proofs. On the other hand, in the field of historical and comparative linguistics, sequence alignment tools have been used to partially automate the comparison method by which linguists traditionally reconstruct languages. Interestingly, business and marketing professional also use multiple sequence alignment techniques to analyze a series of purchases over time.

Some important software tools

BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi): All purpose sequence alignment tool.

CUDASW++ (http://cudasw.sourceforge.net/homepage.htm): Smith-Waterman sequence alignment tool.

DIAMOND (http://ab.inf.uni-tuebingen.de/software/diamond/welcome.html): Multi-purpose sequence alignment tool.

FASTA (https://fasta.bioch.virginia.edu/fasta_www2/): A tool for sequence comparison.

HMMER (http://hmmer.org/): A sequence database search tool.

LAMBDA (http://www.seqan.de/apps/lambda/): A tool for local sequence alignment.

USEARCH (http://drive5.com/usearch/): A tool for ultra-fast sequence alignment.

PSISEARCH (https://www.ebi.ac.uk/Tools/sss/psisearch/): It is used for searching of distant homologues.

ScalaBLAST (https://omics.pnl.gov/software/ScalaBLAST.php): A multipurpose sequence alignment tool.

Sequilab (http://www.sequilab.org/): A multipurpose software suit.

SSEARCH (https://www.ebi.ac.uk/services): A tool for Smith-Waterman sequence alignment. It is slower than FASTA but gives a better result.

SWAPHI(http://swaphi.sourceforge.net/homepage.htm#latest): It is used for searching on Intel Xeon PHI co-processor.

AlignMe(http://www.bioinfo.mpg.de/AlignMe): This is for the alignment of membrane proteins.

SWIPE (http://dna.uio.no/swipe/): It is used for Smith-Waterman sequence searching.

BioConductor (http://www.bioconductor.org/): This is a software suite for analyses of high-throughput sequence data.

Bio-Perl (http://bioperl.org/): Perl programming language for Bioinformatics.

DNASTAR(https://www.dnastar.com/t-sub-solutions-molecular-biology-sequence-alignment.aspx): A multipurpose software suite for sequence analyses.

Genome Compiler (http://www.genomecompiler.com/): A tool for multipurpose sequence analyses.

G-PAS (http://gpualign.cs.put.poznan.pl/project-gpu-pairAlign.html): A tool for global sequence alignment.

JALIGNER (http://jaligner.sourceforge.net/): A software tool for dynamic programming.

ROBETTA (https://web.archive.org/web/20150819163428/http://www.robetta.org/): A protein- sequence to structure alignment tool

LALIGN (https://embnet.vital-it.ch/software/LALIGN_form.html): A tool for local similarity search.

NW-align (https://zhanglab.ccmb.med.umich.edu/NW-align/): A tool for Needleman-Wunsch dynamic programming.

Matcher (https://galaxy.pasteur.fr/?form=matcher): A tool for local alignment.

Stretcher (https://galaxy.pasteur.fr/?form=matcher): A sequence alignment tool.

MCALIGN (http://www.homepages.ed.ac.uk/pkeightl//mcalign/mcinstructions.html): A tool for alignment of non-coding DNA sequences.

PATH (http://bioinfo.lifl.fr/path/path.php): A tool for searching of distant protein homologues.

SEQALIGN (http://www-hto.usc.edu/software/seqaln/seqaln-query.html): A tool for sequence alignment.

SWIFT (https://bibiserv.cebitec.uni-bielefeld.de/swift/): A tool for local alignment.

UGENE (http://ugene.net/): A repository of sequence alignment tools.

YASS (http://bioinfo.lifl.fr/yass/yass.php): A software tool to search for genomic similarity.

ALE (http://www.red-bean.com/ale/): Software tool for multiple sequence alignment.

BALI-PHY (http://bali-phy.org/): This tool is for multiple sequence alignment. It has a downloadable version.

CHAOS-DIALIGN (http://dialign.gobics.de/chaos-dialign-submission) : A tool for both pair-wise and multiple sequence alignment.

CLAUTALW2 (https://www.ebi.ac.uk/Tools/msa/clustalw2/): A tool for multiple sequence alignment

CODONALIGNER (www.codoncode.com/aligner): A tool for sequence assembly and alignment.

COMPASS (prodata.swmed.edu/compass/compass_advanced.php): A tool for multiple alignments of protein sequences with a statistical comparison of the alignments.

DECIPHER (http://decipher.cee.wisc.edu/): R programming language platform for sequence analysis.

DIALIGN (http://dialign-tx.gobics.de/): A tool for segment based multiple sequence alignment.

DNADyanamo (http://www.bluetractorsoftware.co.uk/): A tool for analysis of DNA sequences.

DNASTAR (https://www.dnastar.com/t-sub-solutions-molecular-biology-sequence-alignment.aspx): A tool for analysis of DNA sequences.

MULTALIN (http://multalin.toulouse.inra.fr/multalin/): A tool for analysis of biological sequences.

LAGAN (http://genome.lbl.gov/vista/lagan/submit.shtml): A tool for analysis of biological sequences.

MUSCLE (http://www.drive5.com/muscle/): A tool for sequence alignment.

PMFastR (http://genome.ucf.edu/PMFastR/): A tool for analysis of RNA sequences ePROBALIGN (http://probalign.njit.edu/probalign/login): A tool for analysis of RNA and Protein sequences.

PROBCONS (http://probcons.stanford.edu/index.html): A probabilistic approach to multiple sequence alignment of proteins.

PROMALS3D (http://prodata.swmed.edu/promals3d/promals3d.php): A tool for multiple sequence alignment of protein sequences and structures.

REVTRANS (http://www.cbs.dtu.dk/services/RevTrans/): A tool that takes the DNA sequence and converts it to amino acid sequence of proteins and then performs the sequence alignment.

T-COFFEE (http://www.tcoffee.org/): A tool for multiple sequence alignment.

GENOME-VISTA (http://genome.lbl.gov/cgi-bin/GenomeVista): A tool for whole genome analysis.

FLAK (http://www.flakbio.com/): A tool for comparison of whole genomes.

PLASTRNA (http://plastrna.njit.edu/rnagenome/index): A tool for the analysis of non-coding RNA.

SOME IMPORTANT QUESTIONS AND ANSWERS

1. The method of aligning two sequences by searching for similar sequence patterns that is in the same order in the sequences is called

(a) sequence alignment

(b) pair wise alignment

(c) multiple sequence alignment

(d) all of these

2. A sequence alignment method that is suitable for aligning closely related sequences is called

(a) multiple sequence alignment

(b) pair wise alignment

(c) global alignment

(d) local alignment

3. A sequence alignment method that tries to align the entire sequence is called

(a) multiple sequence alignment

(b) pair wise alignment

(c) global alignment

(d) local alignment

4. A sequence alignment method suitable for finding out conserved patterns in DNA or protein sequences is known as

(a) multiple sequence alignment

(b) pair wise alignment

(c) global alignment

(d) local alignment

5. A sequence alignment method of aligning many sequences simultaneously is known as

(a) multiple sequence alignment

(b) pair wise alignment

(c) global alignment

(d) local alignment

6. Sequence alignment methods help

(a) to identify evolutionary relationships

(b) to identify the functions of newly synthesized genes

(c) to propose new members of gene families

(d) all of these

7. A sequence alignment method that tries to align regions with high level of matches without considering the alignment of rest of the sequences is referred to as

(a) multiple sequence alignment

(b) pair wise alignment

(c) global alignment

(d) local alignment

8. Find the odd one out

(a) Rasmol

(b) BLAST

(c) FASTA

(d) Clustal W

9. Which of the following servers is a sequence alignment tool provided by NCBI

(a) Chime

(b) BLAST

(c) FASTA

(d) Clustal W

10. Which of the following is a multiple sequence alignment tool?

(a) Clustal W

(b) Chime

(c) Dismol

(d) PDB

ANSWERS

1. (b)	**2.** (c)	**3.** (c)	**4.** (d)	**5.** (a)
6. (d)	**7.** (d)	**8.** (a)	**9.** (b)	**10.** (a)

CHAPTER

4

Molecular Modelling

4.1 INTRODUCTION

Molecular modelling techniques represent the combination of all theoretical methods and computational techniques used to model or mimic the behaviour of molecules. The techniques are used in various different fields like computational chemistry, drug design, computational biology and materials science for studying molecular systems ranging from small chemical systems to large biological molecules and material assemblies. The common feature of molecular modelling techniques is the atom level description of the molecular systems. This may include treating atoms as the smallest individual unit (the molecular mechanics approach), or explicitly modelling electrons of each atom (the quantum chemistry approach).

The molecular modelling techniques include the following:

(a) Homology Modelling

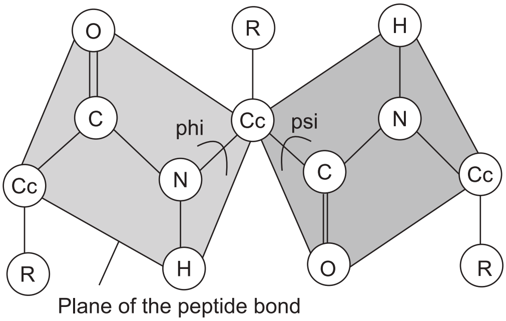
(b) Threading

(c) Ab Initio Modelling

(d) Molecular Docking

Before beginning a few words on protein structure are worth mentioning.

Proteins are polymers made up of amino acids. The amino acids form the protein by losing a water molecule and are linked by peptide linkages (Figure 1).



The R groups represent the side chains

Figure 1: Peptide linkage. The central Ca atom joins the two planes together. The peptide linkage is the linkage between C=O and N-H groups. The peptide linkage is planar and generally the O atom of the C=O group and the H atom of the N-H group remain trans to each other. The dihedral angle between the N-Ca is called the φ angle and that between the Ca- C' is called the ψ angle. The peptide linkages can assume only some specific values of these angles and those values are called the allowed values. The distributions of the values of φ and ψ angles are obtained from Ramachandran plot.

By convention, a chain under 40 amino acid residues is often identified as a peptide, rather than a protein. To be able to perform their biological functions, proteins fold into one or more specific spatial conformations. The underlying forces for the formation of the spatial conformations of proteins are the various types of non-covalent interactions such as hydrogen bonding, ionic interactions, Van der Waals forces, and hydrophobic packing. The understanding of protein functions at the molecular level requires the identification of their three-dimensional structures.

The protein structure is divided mainly into four categories:

(*i*) **The primary structure:** It is the linear sequence of the amino acid residues in the protein.

(*ii*) **The secondary structure:** It is the highly regular local sub-structure in a protein. The major secondary structures present in a protein are the alpha helix, beta strand and turn. These secondary structures have definite patterns of hydrogen bonds between the main-chain peptide groups. They have a regular geometry and are known to possess specific values of the dihedral angles ψ and φ on the Ramachandran plot. The amino acids Ala, Glu, Leu, Met are mostly observed in the helical regions of the amino acids. On the other hand Pro, Gly, Tyr, Ser are not so common in the helical regions. The hydrogen bonding pattern in alpha helix is n to n + 4, where the carbonyl oxygen of the peptide bond of the nth amino acid forms hydrogen bonding with the –NH portion of the peptide bond of the n + 4 th amino acid. There are two other types of helical structures in proteins. They are called the 3_{10}-helix and π-helix. In the 3_{10}-helix the hydrogen bonding pattern is n to n + 3. It is tighter than alpha helix. On the other hand, the π-helix has n to n + 5 hydrogen bonding pattern. It is looser than alpha helix. The secondary structures sometimes form super secondary structures. The different beta strands combine together to form a sheet like structure called beta sheet. In the beta sheets the different beta strands remain hydrogen bonded. The beta sheets can be of two different types: parallel and anti-parallel. In anti-parallel beta sheets there are strands for which one sides remain buried whereas the other sides remain exposed. Anti-parallel beta sheets are more commonly observed and in them the hydrophobic and hydrophilic amino acids alternate. In the parallel beta sheets the strands remain mostly buried. The terminal amino acid residues in parallel beta strands are hydrophilic whereas the central amino acids are hydrophobic.

(*iii*) **The super-secondary structure:** A super-secondary structure is defined as a compact three-dimensional protein structure of several adjacent elements of secondary structure. These are also referred to as motifs. The protein motifs also possess loops and other unstructured regions.

(*iv*) **The tertiary structure:** Tertiary structure of a protein is the three-dimensional structure of a monomeric or multimeric protein molecule. The secondary structural elements, viz., alpha helix and beta strands, are folded into compact shapes of tertiary structure.

The folding of the secondary structural elements is facilitated by non-specific hydrophobic interactions as well as salt bridges, disulfide linkages and hydrogen bonds.

(*v*) **The quaternary structure:** The quaternary structure is the three-dimensional structure of a multimeric protein. It represents the relative orientations of the different protein chains with respect to each other. The quaternary structure is also stabilized by hydrophobic interactions as well as salt bridges, disulfide linkages and hydrogen bonds.

(*vi*) **Protein domains:** A domain or a structural domain in a protein molecule is an element of the protein's overall structure that is self-stabilizing. The domains have the capacity of being folded independently of the rest of the protein chain. Domains take part in the biological function of the protein they belong to.

(*vii*) **Protein fold:** It is the general protein architecture. A protein fold may be defined by the way the secondary structural elements in the protein are arranged relative to each other in space.

Proteins perform most if not all of the biological processes. Thus, the basic aim of the molecular modelling techniques is to analyze the structure function relationships of proteins. In most of the cases proteins interact among themselves to form protein-protein complexes or proteins interact with other different ligands (such as, macro-molecules like DNA, RNA, carbohydrates, lipids; small organic or inorganic molecules or ions). Molecular modelling techniques are used to build the protein-protein or protein-ligand complexes. The bindings of proteins may be transient or permanent. The interfaces of the interacting proteins have several different characteristics. The identification of the interface residues may shed light in many important aspects like drug development, analyses of molecular pathways, generation of protein mimetics as well as understanding of disease mechanisms and development of new docking methodologies to build structural models of protein complexes. In the next few pages, the basics of protein interactions are presented.

Amino acid conformations in proteins

There are 20 different amino acids. They are classified based on the physico-chemical characteristics of their side chains. The various physicochemical properties of side chains are:

Size, shape, hydrophobicity, charge, hydrogen bonding

The general properties of amino acids:

Glycine (Gly) (G): The amino acid G has the following characteristics:

- It is the smallest amino acid
- It has no side chain
- It is highly flexible and is commonly observed in turns
- It is generally present as the outlier in Ramachandran plot
- It is ambivalent. It can be located inside or outside of the protein structure; buried or solvent exposed

Alanine (Ala) (A): The amino acid A has the following characteristics:

- It is small
- It is ambivalent. It can be located inside or outside of the protein structure; buried or solvent exposed

Valine (Val) (V) : The amino acid V has the following characteristics:

- It has a large side chain
- It is hydrophobic
- It remains usually buried in the protein cor58e

Leucine (Leu) (L): The amino acid L has the following characteristics:

- It has a large side chain with an additional –CH_2 group attached to the side chain as compared to V. It is larger than V.
- It is hydrophobic
- It remains usually buried in the protein core

Isoleucine (Ile) (I): The amino acid I has the following characteristics:

- It has a branched side chain
- It is hydrophobic
- It remains usually buried in the protein core
- It is important for protein stability and ligand binding

Phenylalanine (Phe) (F): The amino acid F has the following characteristics:

- It has a nonpolar, hydrophobic side chain
- It is an aromatic amino acid containing an aromatic ring as the side chain
- It remains usually buried in the protein core
- The π electrons of the aromatic ring of the side chain often stack with other aromatic residues
- The π stacking increases protein stability

Methionine (Met) (M): The amino acid M has the following characteristics:

- It contains sulfur-containing, non-polar flexible side chain
- It remains usually buried in the protein core
- The side chain is able to form S-mediated hydrogen bonds

Proline (Pro) (P): The amino acid P has the following characteristics

- It is a cyclic, chain imino acid
- The side chain is unable to form traditional hydrogen bonding interactions like some other amino acids
- The amino acid is able to cause a bend of α-helical structure of a protein
- The amino acid is very rigid and is generally present as *cis*-Pro
- This amino acid is important for protein folding
- It is generally present at the end of α-helix, turn, loops

Serine (Ser) (S): The amino acid S has the following characteristics:

- The amino acid contains a -OH group in the side chain
- The side chain of the amino acid remains uncharged at physiological pH

Threonine (Thr) (T): The amino acid T has the following characteristics:

- The amino acid has a small, hydrophilic side chain
- The amino acid takes part in post-translational modifications
- This amino acid is found to be present in many enzymes
- This amino acid is also involved in metal ion binding

Tyrosine (Tyr) (Y): The amino acid Y has the following characteristics:

- This amino acid has an aromatic side chain with a hydrophilic –OH group attached to the aromatic ring
- This amino acid helps in identification of proteins by UV-Vis spectroscopy
- It may remain buried in the protein core or exposed to the solvents
- This amino acid is found in be present in many enzymes
- The amino acid takes part in post-translational modifications

Tryptophan (Trp) (W): The amino acid W has the following characteristics:

- It has the largest side chain
- The side chain of this amino acid is aromatic
- This amino acid helps in identification of proteins by UV-Vis spectroscopy
- The nitrogen atom present in the side of the amino acid can form hydrogen bonds
- This amino acid is found to be present in many enzymes
- The nitrogen atom of the side chain often remains exposed to surface

Cysteine (Cys) (C): The amino acid C has the following characteristics:

- It contains sulfur-containing, non-polar flexible side chain
- The side chain of the amino acid is small
- The side of the amino acid can form disulfide linkages
- This amino acid is found to be present in many enzymes
- This amino acid is also involved in metal ion binding

AsparagiNe (Asn) (N): The amino acid N has the following characteristics:

- The amino acid has uncharged polar side chain
- It may remain buried in the protein core or exposed to the solvents
- The amino acid takes part in post-translational modifications

Glutamine (Gln) (Q): The amino acid Q has the following characteristics:

- The amino acid has uncharged polar side chain
- The side chain of this amino acid is longer than that of N
- It may remain buried in the protein core or exposed to the solvents

Histidine (His) (H): The amino acid H has the following characteristics:

- The amino acid has aromatic side chain
- The side chain of the amino acid remains positively charged when both the nitrogen atoms are protonated
- This amino acid is found to be present in many enzymes
- This amino acid is also involved in metal ion binding

Aspartate (Asp) (D): The amino acid D has the following characteristics:

- The amino acid has negatively charged side chain

- This amino acid is found to be present in many enzymes
- This amino acid is also involved in metal ion binding

Glutamate (Glu) (E): The amino acid E has the following characteristics

- The amino acid has negatively charged side chain
- This amino acid is found to be present in many enzymes
- This amino acid is also involved in metal ion binding

Lysine (Lys) (K): The amino acid K has the following characteristics:

- The amino acid has positively charged side chain
- This amino acid is found to be present in many enzymes
- The amino acid has long flexible side chain

Arginine (Arg) (R): The amino acid R has the following characteristics:

- The amino acid has positively charged side chain
- This amino acid is found to be present in many enzymes
- The amino acid has long flexible side chain

The amino acid side chains can have different conformations. They are called rotamers. A database is present where the three dimensional coordinates of the atoms of the side chains of the amino acids in possible different conformations are present. Such a library is called rotamer library. In general the staggered conformation of the side chain is the most favoured conformation.

Basic definitions:

- **Monomer:** A single unit of an assembly.
- **Polymer:** An assembly of single monomeric units i.e., Polymer = $(\text{Monomer})_n$

If n = 2, the polymer is called dimer

If n = 3, the polymer is called trimer etc.

- **Protomer:** The monomeric constituent units of a protein having two or more monomeric protein chains.

- **Homomer:** A protein with the same monomeric constituents.
- **Heteromer:** A protein with different monomeric constituents.
- **Obligate and non-obligate protein-protein complexes:** A protein-protein complex where the individual partners (protomers) are not stable by themselves, such as, the Arc repressor dimer. If the individual protomers can exist on their own, the complex is then referred to as non-obligate. As for an example, the antibody-antigen complexes are non-obligate ones.
- **Transient and permanent complexes:** If the associations between the protomers in a protein-protein complex is weak and are in a dynamic equilibrium in solution where it is broken and formed continuously the complex is called transient complex, such as, the non-obligate homo-dimer of sperm lysine.

On the other hand, if the associations between the protomers require molecular switch to break, the complex is called permanent complex. The hetero-trimeric G protein forms a permanent complex in presence of GDP.

It is to be noted that PPIs cannot be distinctly classified as transient and permanent rather a continuum exists between them. The stabilities of all these complexes depend upon physiological conditions and cellular environments.

- **Accessible surface area (ASA):** It is the fraction of the total van der Waal's surface of an atom that can come in contact with other atoms, specifically water. In case of proteins the accessible surface area is calculated for each atom of each amino acid residue.
- **Interface patch:** The interface is the contact area of the proteins. It involves those residues of the proteins, which have the ASA of their side chains decreased by >1 $Å^2$ on complex formation.
- **Relative accessible surface area (RSA):** It is defined as the ratio of ASA to the maximal accessibility of each amino acid.
- **Gap volume and Gap index:** The gap volume is defined as the volume enclosed between any two protein molecules delimiting the boundary by defining a maximum allowed distance from both

the interfaces. The Gap index is calculated as *Gap index = gap volume / interface ASA*

- **Surface patch:** Surface residues are the amino acid residues of a protein with a relative accessible surface area of > 5%. A surface patch is the central surface accessible residue and *n* nearest surface accessible neighbors, where *n* is the size of the patch in terms of the number of residues. A mean relative ASA for each patch is calculated as

Patch ASA (Å^2) = sum of the RSAs of the amino acid residues in the patch / number of amino acid residues in the patch

- **Solvation potential:** This is the measure of the propensity of an amino acid to get solvated. It is used to quantify the tendency of a patch to be exposed to solvent or buried in the interface of a protein-protein complex.
- **Residue interface propensities (RIP):** It represents the tendency of the amino acid residues of a protein to be on the interface of the protein-protein complex. The patch interface propensity (PIP) is calculated as

PIP = sum of the natural logarithms of the RIPs of the amino acid residues in the patch / number of amino acid residues in the patch

- **Hydrophobicity:** The surface patch hydrophobicities are defined as

  Patch Hydrophobicity = Hydrophobicity value of the amino acid residues in a patch / number of amino acid residues in the patch

- **Planarity:** This quantity of the surface patch is evaluated by calculating the root mean square (rms) deviation of all the atoms present on the surface patch from the least square plane that passes through the atoms.
- **Protrusion index (PI):** This quantity gives an idea of how much a residue sticks out from the surface of a protein. The patch PI is calculated as

Patch PI = sum of the PIs of the amino acid residues in the patch / number of amino acid residues in the patch

A few aspects of protein-ligand interactions

Ligand is a substance that binds to a biomolecule, mainly proteins, to serve some specific purposes. In protein-ligand binding, the ligand is usually a signaling molecule that binds to a specific site on a target protein. In case of protein-ligand binding the binding occurs by intermolecular forces, such as ionic bonds, hydrogen bonds and van der Waals forces. However, association between the ligand and its receptor protein is usually reversible. It is well established that an actual irreversible covalent bonding between a ligand and its target molecule is rare in biological systems.

A ligand binding to its receptor protein alters its chemical conformation i.e., the three-dimensional shape of the receptor protein. It is known that the conformational state of a receptor protein determines its functional state. In biological systems, the ligands can be of different types. They include substrates, inhibitors, activators, and neurotransmitters. The tendency or strength of ligand binding to its receptor protein is called affinity. The binding affinity of ligands is determined not only by *direct* interactions, but also by solvent effects that are known to play a dominant *indirect* role in driving non-covalent receptor-ligand bindings in solutions.

The interactions of most ligands with the binding sites of their corresponding receptors can be characterized in terms of binding affinity. In general, high-affinity ligand binding results from greater intermolecular forces between the ligand and its receptors. On the other hand, low-affinity ligand binding involves less intermolecular force between the ligand and its receptor. Another important aspect of high affinity binding is that high-affinity binding involves a longer residence time for the ligand at its receptor binding site whereas the low-affinity binding leads to comparatively lesser binding time. High-affinity binding of ligands to their receptors is often physiologically important and in these cases some of the binding energy can be used to cause a conformational change in the receptor site. This results in some altered behavior of an associated ion channel or enzyme.

A ligand is called the agonist for that particular receptor if the ligand can bind to the receptor, alter the function of the receptor, and trigger a physiological response. The agonist for a specific receptor can be characterized on the basis of how much physiological response can be triggered and in terms of the concentration of the agonist that is required to produce the physiological response. For agonists binding to high-affinity ligand binding sites in receptors, a relatively low concentration of a ligand is adequate to maximally occupy a ligand-binding site and trigger a physiological response. On the other hand, low-affinity binding implies that a relatively high concentration of a ligand is required before the binding site in the receptor is maximally occupied and the maximum physiological response to the ligand is achieved.

On a similar note, an antagonistic ligand is the one that after binding to its receptor fails to stimulate the receptor and brings about a change in the receptor.

The ligands can be of the type selective and non-selective. Selective ligands have their tendencies to bind to very limited types of receptor proteins. On the other hand, non-selective ligands bind to several types of receptor proteins. This has got a huge pharmaceutical importance. Drugs that are non-selective tend to have more adverse effects, because they bind to several other receptors in addition to the one generating the desired effect. This leads to side effects.

There is another class of ligands referred to as bivalent ligands. The bivalent ligands consist of two connected molecules as ligands, and are used in scientific research to detect receptor dimers and to investigate their properties. Bivalent ligands are usually large molecules and they tend not to be 'drug-like', limiting their applicability in clinical settings.

Molecular modelling techniques utilize all the basic definitions of protein structures and come up with plausible models of single proteins, protein-protein and protein-ligand complexes. In the next few pages the basic principles of molecular modelling techniques will be described. As mentioned earlier molecular modelling techniques depend on molecular mechanics and quantum chemistry principles.

A brief overview of molecular mechanics principles

Molecular mechanics is one of the basic aspects of molecular modelling techniques. The underlying principle of molecular mechanics is the use of classical mechanics or Newtonian mechanics to describe the physical basis behind the models. In molecular models the atoms are considered to be a combination of the nucleus and electrons collectively and they are depicted as point charges with an associated mass. The atoms are considered to be joined by chemical bonds. The bonds are like springs and they follow Hook's law. There are non-bonded interactions in the molecules which are known as van der Waals forces. The van der Waals forces are described by Lennard-Jones potential. On the other hand, the electrostatic interactions are computed based on Coulomb's law. Atoms are assigned coordinates in Cartesian space or internal coordinates. The atoms can also be assigned velocities in dynamical simulations. The atomic velocities are related to the temperature of the system which is a macroscopic quantity. The collective mathematical expression of all the interaction terms is known as a potential function and is related to the system internal energy (U) which is a thermodynamic quantity equal to the sum of potential and kinetic energies. Methods to minimize the potential energy are known as energy minimization techniques (e.g., steepest descent and conjugate gradient). On the other hand, the methods that model the behaviour of the system with propagation of time are known as molecular dynamics. The energy terms considered are presented as:

$$E = E_{\text{bonds}} + E_{\text{angle}} + E_{\text{dihedral}} + E_{\text{non-bonded}}$$

$$E_{\text{non-bonded}} = E_{\text{electrostatic}} + E_{\text{van der Waals}}$$

This function, referred to as a potential function, computes the molecular potential energy as a sum of energy terms that describe the deviation of bond lengths, bond angles and torsion angles away from equilibrium values, plus terms for non-bonded pairs of atoms expressed by van der Waals and electrostatic interactions. The set of parameters consisting of equilibrium bond lengths, bond angles, partial charge values, force constants and van der Waals parameters are collectively known as the force field. However, different implementations of molecular mechanics use different mathematical expressions and different parameters for the potential function. The common force fields in use today have been developed by using high level quantum calculations and/or fitting to existing experimental data. The technique known as energy minimization is used to find positions of zero gradients for all atoms. In other words, a local energy minimum is generally found. Lower energy states are more stable and are commonly investigated because of their role in chemical and biological processes. A molecular dynamics simulation, on the other hand, computes the behaviour of a system as a function of time. It involves solving Newton's laws of motion, principally the second law, F=m.a.

Where, F = Exerted force, m = Mass of the body, a = acceleration

Integration of Newton's laws of motion, using different integration algorithms, leads to atomic trajectories in space and time. The force on an atom is defined as the negative gradient of the potential energy function. The energy minimization technique is useful to obtain a static picture of the molecular system to compare the states of similar systems. On the other hand, molecular dynamics provides information about the dynamic processes with the intrinsic inclusion of temperature effects. The molecules can be modeled either in vacuum or in the presence of a solvent such as water. The modeled structures may be simulated in vacuum or in presence of a solvent such as water. The former type of simulation is called the *gas-phase* simulations, while the latter which include the presence of solvent molecules is referred to as *explicit solvent* simulations. However, the inclusion of solvent molecule is very much time consuming. Thus, the effect of solvent may be estimated using an empirical mathematical expression. Such simulations are known as *implicit solvation* simulations. The details of molecular mechanics force-fields are presented in the following section.

Molecular Mechanics Force Field

Force-fields link a structure's atomic coordinates to a potential energy. Force fields for molecular systems can be interpreted in terms of a relatively simple four component picture of the intra- and inter-molecular forces within the system. The molecular system's potential energy (E) in a given conformation is a sum of individual energy terms.

$$E= E_{\text{covalent}} + E_{\text{noncovalent}}$$

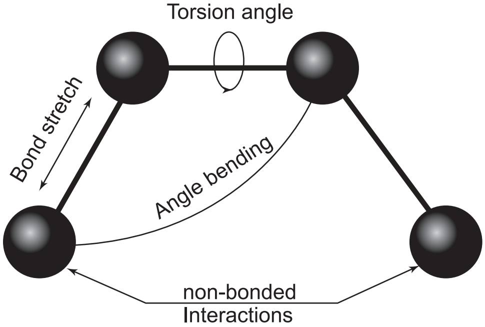
where, the components of the covalent and non covalent contributions are given by the following summations:

$$E_{\text{covalent}} = E_{\text{bond}} + E_{\text{angle}} + E_{\text{dihedral}}$$

$$E_{\text{noncovalent}} = E_{\text{van der Waals}} + E_{\text{electrostatic}}$$

One functional form of force field is

$$V(r) = \sum_{\text{bonds}} k_b(b-b_0)^2 + \sum_{\text{angles}} k_\theta(\theta-\theta_0)^2$$
$$+ \sum_{\text{dihedrals}} k_\phi(1+\cos(n\phi-\phi_0)) \sum_{\text{impropers}} k_\psi(\psi-\psi_0)^2$$
$$+ \sum_{\substack{\text{non-bonded}\\ \text{pairs }(i,j)}} 4\varepsilon j\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)\right] + \sum_{\substack{\text{non-bonded}\\ \text{pairs }(i,j)}} \frac{q_i q_j}{\varepsilon_D r_{ij}}$$



V(r) denotes the potential energy, which is a function of the positions (r) of N particles usually atoms. The first term denotes the interaction between pairs of bonded atoms, a harmonic potential that gives the increase in energy as the bond length b deviates from its reference value b_0. The second term is a summation over all valence angles in the molecule and the third and fouth terms are torsional potential that model how the energy changes as bond rotates. The non-bonded term is usually modelled using a Coulomb potential term for electrostatic and a Lennard-Jones potential for van der Waals interactions.This is calculated between all pairs of atoms (i and j) that are in different molecules or in the same molecule but separated by at least three bonds, i.e.; have a 1 n relationship where $n \geq 4$.

Common molecular nodelling force fields and software tools:

One of the widely used packages is AMBER (Assisted Model Building with energy refinement). AMBER was developed with an aim to simulate the biological macro molecule especially for protein and nucleic acid. Later on the force field was modified for the simulation of carbohydrate. The (Chemistry at Harvard Macromolecular Mechanics) CHARMM, is a very important and highly capable force field and it is also a software package used for molecular mechanics and dynamics simulations of biological macromolecules such as proteins, nucleic acids and carbohydrates. GROningen MOlecular Simulation (GROMOS) is another versatile force field that comes as part of the GROMOS software suite. This tool is a versatile package to perform molecular mechanics and dynamics simulations. It is primarily designed for biological molecules like proteins, lipids and nucleic acids that have a lot of complicated bonded interactions. The package includes a fully automated topology builder for proteins, even the multimeric structures of proteins.

CVFF (Consistent Valence Force Field) is also used broadly for small molecules and macromolecules.

CFF (Consistent Force-Field) is a family of force-fields adapted to a broad variety of organic compounds, and it includes force fields for polymers, metals, etc.

Optimized Potential for Liquid Simulations (OPLS) force-field is optimized to fit experimental properties of liquids, such as density and heat of vaporization, in addition to fitting gas-phase torsional profiles.

Periodic Boundary Conditions

The proper treatment of boundaries and boundary effects is crucial to simulation methods as it enables 'macroscopic' properties to be calculated from simulations using relatively small numbers of particles. Periodic boundary conditions (PBC) are particularly useful for simulating a part of a bulk system with no surfaces present, unless the behaviour near the walls (surface) are of interest. The bulk is assumed to be composed of the primary / unit cell surrounded by its exact replicas and thus forms an infinite lattice in 3D space. The image cells not only have the same size and shape as the primary one but also contain particles that are images of the particles in the primary cell. Moreover the cells are space filling, i.e., separated by open boundaries so particles can freely enter or leave any cell in coordination so that total number of particles remain conserved. The number of image cells needed depends on the range of intermolecular forces. When the forces are sufficiently short ranged (e.g. in truncated Lennard-Jones model), only image cells that adjoin the primary cell are needed (minimum image convention). For squares in two dimensions, there are eight adjacent image cells whereas for cubes in three dimensions, the number of adjacent images is twenty-six. In principle, any cell shape can be used, provided it fills all of space by translation operations of the central box in three dimensions. Five shapes satisfy this condition: the cube (and its close relation, the parallelepiped), the hexagonal prism, the truncated octahedron, the rhombic dodecahedron and the ‘elongated’ dodecahedron. The truncated octahedron and the rhombic dodecahedron provide periodic cells that are approximately spherical, thus reduces the number of solvent molecules needed to be added to solvate the solute and saves CPU time as well. But it is often sensible to choose a periodic cell that reflects the underlying geometry of the system like the hexagonal prism is appropriate for solvated DNA. For some simulations, it is inappropriate to use standard PBC in all directions. PBC may also cause difficulties when simulation is performed in homogeneous systems or systems that are not at equilibrium. For example, when studying the adsorption of molecules onto a surface, it is clearly inappropriate to use the usual PBC for motion perpendicular to the surface, although it is applied to motion parallel to the surface. PBCs are not always used in simulation in such systems like liquid droplets or van der Waals clusters as these systems inherently contain a boundary. Moreover, to avoid interference from non-bonded interactions, according to minimum image convention, the periodic images must be separated by a distance equal to or larger than a cut-off distance. In practice, a solute-box distance equal to the cut-off (generally 1 nm) is used. For long-range electrostatic interactions, lattice sum methods such as Ewald Sum, Particle Mesh Ewald (PME) and Particle-Particle Particle-Mesh (P3M) algorithms are used.

4.2 Energy Minimisation

Energy minimisation is widely used in molecular modelling and is an integral part of some techniques like conformational search procedures or preparation of a system prior to other type of calculations like Monte Carlo or molecular dynamics simulation in order to relieve any unfavourable interactions in the initial configuration.

Potential energy of all except very simplest system is a complicated, multidimensional function of the coordinates. For a system with N atoms the energy is a function of 3N −6 internal or 3N Cartesian coordinates. It is therefore impossible to visualize the entire energy surface except for some simple cases where the energy is a function of just one or two coordinates. In biomolecular modelling, minimum points on the energy surface would correspond to stable states of the system. There might be a very large number of minima on the energy surface. The minimum with the very lowest energy is known as the global energy minimum. Other minima are called local energy minima. To achieve the geometry of a conformation with its minimum points of energy surface, minimisation algorithm is applied. The highest point on the pathway between two minima, the saddle point, and the transition structures are of special interest. All these minima and saddle points are stationary points on the energy surface, where the first derivative of the energy function is zero and second derivative is positive with respect to all the coordinates. Molecular mechanics minimisations are always performed by Cartesian coordinates, where the energy is a function of 3N variables.

Most of the minimisation methods gradually change the coordinates of the atoms as they move the system closer and closer to the minimum on the potential energy surface of the molecule. In all these methods, a numerical test is applied to the new geometry to decide if a minimum is reached. For example, the slope may be tested to see whether it is zero within some numerical tolerance or not. If the criterion is not met, then the formula is applied again to make another change in the geometry, until the minimum is reached. In practice, it is rarely possible to identify the exact location of minima and/or saddle point. Minimisations are usually converged by tracking,

(a) energy difference between successive steps, or

(b) root mean square deviation (RMSD) between successive configurations, or

(c) calculating root mean square gradients of the energy with respect to the coordinates.

Derivative minimization algorithms: In this method, it is obvious to calculate the derivatives of the energy wih respect to the variables, i.e., the cartesian or internal coordinates. It is more useful as it provides information about the shape of the energy surface and if used properly it could significantly enhance the efficiency with which the minima are located. The direction of the first derivative of the energy, the gradient, indicates where the minimum lies, and the magnitude of the gradient indicates the steepness of the local slope. The energy of the system can be lowered by moving each atom in response to the force the acting on it. The second derivatives indicate the curvature of the function, information that can be used to predict where the function will change direction, i.e. pass through a minimum or some other stationary point.

First-order derivative methods use the first derivatives, i.e., the gradients, whereas second-order methods use both first and second derivatives. The Newton-Raphson method is the simplest second-order method.

First–order minimization methods: Two most frquently used first-order minimisation algorithms are steepest descent and conjugate gradients methods.

Steepest Descent (SD): Steepest descent method moves in the way in which the geometry is first minimized in the direction parallel to the net force, i.e., opposite to the direction in which the gradient is the largest or, steepest at the initial point, hence its name is steepest. Once a minimum in the first direction is reached, a second minimization is carried out starting from that point and moving in the steepest remaining direction. This process continues until a minimum has been reached in all directions within a sufficient tolerance. The direction of the gradient is determined by the largest inter-atomic forces. So SD is a good method for relieving the highest-energy features in an initial configuration. The method is generally robust even when the starting point is far from a minimum, where the harmonic approximation to the energy surface is often a poor assumption. However, it is known for slow progress near minima. In SD method both the gradients and the direction of successive steps are orthogonal and show oscillatory behaviour.

Conjugate gradient method (CG): Nonlinear conjugate gradient methods form another popular type of energy minimization scheme for large-scale problems where memory and computational performance are important. In CG method, the first portion of the search takes place in the direction parallel to the net force, just as in SD. However, to avoid oscillatory behavior of SD as it moves toward the narrow minima, the CG method uses the information from the previous direction in the next search. This allows the method to move rapidly to the minimum in fewer steps. In CG, the gradients at each point are orthogonal but the directions are conjugate and hence the name. The line search method is usually used, although an arbitrary step method is also possible. CG method is usually attributed as the best choice for general use, even though time per iteration is longer compared to steepest descents for complex systems.

The ideal minimization algorithm

There are many factors that must be taken into account when choosing the most appropriate minimisation algorithm (or combination of algorithms) for a given problem. The ideal minimisation algorithm is the one that provides the answer as quickly as possible, using the least amount of memory. Availability of analytical derivatives and the robustness of the method are also required to be taken into account. No single minimisation method has yet proved to be the best for all molecular modelling problems. In particular, a method that works well with quantum mechanics (for fewer atoms) may not be the most suitable for use with molecular mechanics due to size of the system under consideration and differential computational requirement for calculating energy in both methods. For bio-molecular systems, SD and CG methods are mostly used. Generally, SD performs better than CG when the starting structure is far from the minimum. However, CG takes over once the initial strain has been removed. In case of small molecules Newton-Raphson method might be used after initial minimisation by a more robust method.

Long range Electrostatic Calculations

Long-range interactions (that decay slower than r^{-n}, where n is the dimensionality of the system), form a complex part of molecular simulations, as their range is often greater than half the box length. The charge-charge interaction, which decays as r^{-1}, is particularly problematic like in case of charged species, molten salts and during calculation of such properties like as dielectric constant. Moreover, an increase in the cut-off distance as well in the box size within which all calculations are required to be performed will increase computational costs unnecessarily. A variety of methods are developed to deal with long-range electrostatics: Ewald summation and related methods, reaction-field and image charge method, cell multipole methods during ionization, and minimization, equilibration and production phases of simulation.

Ewald summation method: Ewald summation, named after Paul Peter Ewald, was developed to study the energetics of ionic crystals, where a particle interacts with all the other particles in the simulation box and with all of their images in an infinite array of periodic cells. It has been extensively used in simulations involving highly charged systems (such as ionic melts, solid state materials) and where electrostatic effects are important (lipid bilayers, proteins and DNA). The Ewald summation method is computaionally quite expensive to implement. A popular approach is the particle-mesh method of Hockney and Eastwood, which uses the nearest 27 points in three dimensions. From this grided charge density it is possible to calculate (through Fast Fourier Transform algorithm) the potential due to the Gaussian distributions at the grid points, which by interpolation gives rise to the desired potential at each of the particles. A number of variants of Ewald summation was later developed and all of which use Fast Fourier Transform algorithm but differ in other aspects of implementation. These are particle-mesh Ewald method and particle-particle-particle-mesh approach.

An early application of the particle-mesh Ewald (PME) method was the MD simulation of a crystal of the bovine pancreatic trypsin inhibitor. Ewald summation were in close agreement with those derived from the crystallographic temperature factors unlike the non-Ewald mehod. The highly charged nature of DNA simulation using PME are often much more stable and trajectories remain closer to the experimental structures. PME algorithm is almost linear as it scales as N log(N), which is substantially faster and more accurate than the ordinary Ewald summation on medium to large systems. The method is based on interpolation of the reciprocal space Ewald sums and evaluation of the resulting convolutions using Fast Fourier Transform algorithms. Timings and accuracies are presented for three large crystalline ionic systems. However on very small systems, the simple Ewald method gets a better preference to avoid the overhead in setting up grids and transforms.

Molecular dynamics with continuous potential

The first simulation using continuous potential was performed by Rahman, with Argon in 1964 and molecular liquid (water) in 1971. Under the influence of continuous potential the motions of all the particles are coupled together, giving rise to a many-body problem that cannot be solved analytically. However, for biological molecules, mainly constrained dynamics is performed.

Constrained dynamics

The conformational behavior of a flexible molecule is usually a complex superposition of different motions. The high frequency motions (e.g. bond vibrations) are usually of less interest than the lower frequency modes, which often correspond to major conformational changes. Unfortunately, the time step of a molecular dynamics simulation is dictated by the highest frequency motion present in the system. It would therefore be of considerable benefit to be able to increase the time step without prejudicing the accuracy of the simulation. Constraint dynamics enables individual internal coordinates or combinations of specified coordinates to be constrained, or ‘fixed’ during the simulation without affecting the other internal degrees of freedom. A constraint is a requirement that the system is forced to satisfy, i.e. bonds or angles are forced to adopt specific values throughout a simulation.

Shake: The most widely used algorithm for performing constrained dynamics for large molecules is SHAKE. SHAKE is an iterative method; it resets bonds and angles to prescribed values by moving the bonded particles parallel to the old bond directions. Sequentially SHAKE changes a set of unconstrained coordinates, all bonds, to a set of coordinates that fulfill a list of distance constraints, using a set of references. As the bonds are coupled, this procedure has to be repeated until the desired accuracy is reached. The number of iterations can be decreased using over-relaxation. SHAKE is simple and numerically stable because it resets all constraints within a prescribed tolerance. SHAKE exhibits the problem with large displacements – it fails to provide any solution as the coupled bonds are handled sequentially. Thus, correcting one bond may tilt a coupled bond so far that the method does not converge. Due to its iterative nature it is difficult to parallelize.

Lincs: LINear Constraint Solver (LINCS) is an algorithm that resets bonds to their correct lengths after an unconstrained update. The method is non-iterative, as it always uses two steps. In the first step, the projections of the new bonds on the old bonds are set to zero. In the second step, a correction is applied for the lengthening of the bonds due to rotation. The algorithm is inherently stable, as the constraints themselves are reset instead of derivatives of the constraints. In this way, the algorithm eliminates drift. The constraint problem reduces to a linear matrix equation as the second derivatives of the constraint equations are set to zero. However, in a finite discretization scheme corrections are necessary to achieve accuracy and stability. LINCS does this by implementing an efficient solver for the matrix equation, a velocity correction that prevents rotational lengthening and a length correction that improves accuracy and stability. Although the derivation of the algorithm is based on matrices, no matrix-matrix multiplications are needed and only the nonzero matrix elements have to be stored, making the method useful for very large molecules. LINCS has broader convergence conditions; the sparse constrain coupling matrix is inverted using a power series, which converges rapidly because the constraining influence is relatively local; therefore, parallelisation of the algorithm is straightforward. The latter is of major importance for simulations of large molecules on parallel computers. Moreover, the LINCS algorithm is three to four times faster than the SHAKE algorithm, at the same accuracy. Thus use of LINCS enables time step to be guided by physical conditions only, independent of constrained algorithm.

Overviews of biological modelling techniques:

(a) Homology Modelling

Homology modelling, also known as comparative modelling of protein, is the method to build an atomic-resolution model of a protein from its amino acid sequence on the basis of an experimental three-dimensional structure of a related homologous protein. The protein for which the model is being built is called the target protein. On the other hand, the other homologous protein on which the model of the target protein is built is called the template. Homology modelling technique relies on the identification of one or more known protein structures which are likely to resemble the structure of the target sequence. The amino acid sequence of the target protein is used as a query sequence to search structural databases in order to find the structures of suitable homologous proteins having sufficient sequence identities with the amino acid sequence of the target protein. The underlying principle of the method is based on the assumptions that protein structures are more conserved than protein sequences. However, the typical cut-off value for sequence identity to reliably predict the structures of proteins from the sequence alignments of homologous proteins is 20%. Evolutionarily related proteins have somewhat similar sequences and naturally occurring homologous proteins have similar protein structures. It has been observed that three-dimensional protein structure is evolutionarily more conserved than would be expected on the basis of sequence conservation alone. The sequence alignment between the target and template protein sequences and template structure are then used to produce a structural model of the target. Because protein structures are more conserved than protein sequences, detectable levels of sequence similarity usually imply significant structural similarity between the proteins.

The quality of the homology model is dependent on the accuracy of the sequence alignment and the quality of the template structure. However, the homology modelling techniques can be complicated by the presence of alignment gaps (commonly called INDELS) that indicate a structural region present in the target but not in the template. Sometimes the template structures also have missing amino acid residues arising out of poor resolution in the experimental procedure (usually X-ray crystallography) used to solve the structure. The quality of a homology model declines with decreasing sequence identity. A typical good quality homology model built from the alignment of a template and a target protein sequences sharing 70% sequence identity should have ~1–2 Å root mean square deviation (RMSD) between the matched C^{α} atoms. However, with decreasing sequence identities (below 25% level) the RMSD is only 2–4 Å . However, the errors are significantly higher in the loop regions, where the amino acid sequences of the target and template proteins may be completely different.

Regions of the model that are constructed without a template are generally much less accurate than the rest of the model; the technique is called loop modelling. Errors in side chain packing and position also increase with decreasing sequence identities between the target and the template. Variations in these packing configurations have been suggested as a major reason for poor model quality at low identity. Taken together, these various atomic-position errors are significant and limit the use of homology models for purposes that require atomic level resolution of structural data, such as drug design and predictions of protein–protein interaction. It is also known that even the quaternary structure of a protein may be difficult to predict from homology models of its subunit(s). Nevertheless, homology models can be useful in reaching qualitative conclusions about the molecular biochemistry of the query sequence.

This technique is especially employed in formulating hypotheses about why certain residues are conserved, which may in turn lead to wet-lab experiments to test those hypotheses. For example, the spatial arrangement of conserved residues may suggest whether a particular residue is conserved to stabilize the folding, to participate in binding of some small molecule, or to foster association with another protein or nucleic acid.

Homology modelling can produce high-quality structural models when the amino acid sequences of the target and template are closely related. This has inspired the formation of a structural genomics consortium dedicated to the production of corresponding representative experimental structures for all classes of protein folds. The method of homology modelling is based on the observation that the tertiary structure of a protein is better conserved than the primary structure i.e., the amino acid sequence. Thus, even proteins that have diverged appreciably in sequence but still share detectable sequence similarities will also share common structural properties; particularly the overall fold of the protein will be similar. Because it is difficult and time-consuming to obtain experimental structures from methods such as X-ray crystallography and protein NMR for every protein of interest, homology modelling can provide useful structural insights for generating hypotheses about a protein's function and directing further experimental work. In other words, two proteins may share a similar fold even if their evolutionary relationship is so distant that it cannot be discerned reliably.

Steps in model building

The homology modelling technique can be broken down into four sequential steps:

1. template selection
2. target-template alignment
3. model construction
4. model assessment.

The first two steps as mentioned above are often essentially performed together. This is because the most common methods of identifying templates rely on the production of sequence alignments. The key to a good model quality is to have a proper sequence alignment.

The step of model construction can be done in several different ways. Given a template and a sequence alignment between the target and the template, the information contained therein is used to generate a three-dimensional structural model of the target, represented as a set of Cartesian coordinates for each atom in the protein. Three major classes of model generation methods have been proposed:

(a) **Fragment assembly:** This is the original method of homology modelling. This method relied on the assembly of a complete model from conserved structural fragments identified in closely related solved structures. As for an example, a modelling study of serine proteases in mammals identified a sharp distinction between "core" structural regions conserved in all experimental structures in the class, and variable regions typically located in the loops where the majority of the sequence differences were localized. Thus, the target proteins could be modeled by first constructing the conserved core and then substituting variable regions from other proteins in the set of solved structures. However, the up-to-date implementations of this method differ mainly in the way they deal with regions that are not conserved or that lack a template. The variable regions, i.e, the loop regions are often constructed with the help of fragment libraries.

(b) **Segment matching:** This method divides the entire target protein into short segments of amino acid residues and each segment is matched with different structural templates. The selection of the template is done with sequence similarity, comparison of alpha carbon coordinates and steric clashes between the side chain atoms.

(c) **Satisfaction of spatial restraints:** The most commonly used technique of homology modelling is by the satisfaction of spatial restraints. This method of homology modelling takes its inspiration from calculations required to construct a three-dimensional structure from data generated by NMR spectroscopy. In this method, one or more target-template alignments are used to construct a set of geometrical criteria that are then converted to probability density functions (pdfs) for each restraint. A more recent expansion of the method applies the spatial-restraint model to electron density maps derived from cryo-electron microscopy studies, which provide low-resolution information that is not usually itself sufficient to generate atomic-resolution of structural models.

The assessment of the homology models is an important task. This is generally done in two different ways; the statistical potentials and physics based energy calculations. Both these methods produce energy values of the modeled proteins.

Statistical potentials are empirical methods based on observed residue-residue contact frequencies among proteins of known structures in the Protein Data Bank (PDB)—the structural database for macromolecules. They assign a probability or energy score to each possible pair wise interactions between amino acids and combine these pair wise interaction scores into a single score for the entire homology model.

Physics-based energy calculations generally aim to capture the inter-atomic interactions that are physically responsible for protein stability in solution, especially the van der Waals and electrostatic interactions. These calculations are performed using a molecular mechanics force field. Such physics based method is based on the energy landscape hypothesis of protein folding, which predicts that a protein's native state is also its energy minimum.

The accuracies of the structures obtained by homology modelling are highly dependent on the sequence identity between target and template. Above 50% sequence identity, the homology models tend to be reliable, with only minor errors in side chain packing and rotameric state. The overall RMSD between the modeled and the experimental structure are found to be around 1Å. This error is comparable to the typical resolution of a structure solved by Nuclear Magnetic Resonance (NMR) techniques.

When the sequence identity is in the range of 30–50%, errors can be more severe and are often located in loops. Below 30% sequence identity, serious errors occur which sometimes result in the basic fold being mis-predicted. This low-identity region is often referred to as the "twilight zone" within which homology modelling is extremely difficult. In such cases homology modelling methods are possibly less suited than fold recognition.

The other less popular molecular modelling techniques are Threading and Ab Initio methods.

(b) Threading

Protein threading, also known as fold recognition, is a method of modelling protein structures. The method is used to model those proteins which have the same fold as proteins of known structures, but do not have homologous proteins with known structure. The threading method differs from the homology modelling method of structure prediction as the threading method is used for proteins which do not have their homologous protein structures deposited in the PDB, whereas homology modelling is used for those proteins which have homologous protein structures present in PDB. Threading works by using statistical knowledge of the relationship between the structures deposited in the PDB and the sequence of the target protein, the one to be modeled. The prediction is made by "threading" (i.e. placing, aligning) each amino acid in the target sequence to a position in the template structure, and progressively evaluating how well the target fits the template. The best-fit template is then selected and then the structural model of the sequence is built based on the alignment with the chosen template. Protein threading is based on two basic observations: that the total number of different folds in nature is fairly small (approximately 1300); and that nearly 90% of the new structures submitted to the PDB in the past three years have similar structural folds to the ones already in the PDB.

Comparison with homology modelling

Homology modelling and protein threading are both template-based methods to predict the structures of proteins from their amino acid sequences. There is practically no rigorous boundary between them in terms of prediction techniques. However, homology modelling is for those protein targets which have homologous proteins with known structure. On the other hand, protein threading is for those protein targets with only fold-level homology found. In a nutshell, homology modelling is for "easier" targets and protein threading is for comparatively "harder" targets.

(c) Ab initio method

In ab initio methods, an initial effort is made to analyze the secondary structures (alpha helix, beta sheet, beta turn, etc.) from primary structure of a protein. This is done by utilization of physicochemical parameters and neural network algorithms. From that point, algorithms predict the tertiary folding in a protein. One drawback to this strategy is that it is not yet capable of incorporating the locations and orientations of side chains in amino acid.

The next important molecular modelling technique is the molecular docking. The molecular docking is one of the most widely used techniques. This technique is required to generate models of protein and protein ligand complexes.

(d) Molecular Docking

In the field of molecular modelling, docking is the method which predicts the preferred orientation of one molecule bound to a second molecule in a complex. Knowledge of the preferred orientation in turn may therefore be used to predict the strength of association or binding affinity between the two molecules using a scoring function. The associations between biologically relevant molecules such as proteins, nucleic acids, carbohydrates, and lipids play very vital roles in signal transduction. Furthermore, the relative orientation of the two interacting partners (protein with another protein or a ligand) may affect the type of signal produced. Therefore, docking is useful for predicting both the strength and type of signal produced. Docking is frequently used to predict the binding orientation of small molecule drug candidates to their protein targets in order to predict the affinity and activity of the small molecule. Hence docking plays an important role in the rational drug design endeavors. Given the biological and pharmaceutical significance of molecular docking, considerable efforts have been directed towards improving the methods to predict the structure of the complex obtained by docking.

As the subject is still very young, researches are going on to come up with many other plausible ways of molecular techniques to build the structures of proteins and protein ligand complexes. In the present scenario, the present aim of the chapter is to give a fair idea about the basics of the molecular modelling principles as utilized in the field of protein-protein and protein ligand complexes. There are very few articles available that deal with the basics of such techniques. In the present chapter emphasis is given on the elucidation of the existing knowledge about the molecular modelling techniques. The chapter is intended for both the computer scientists and molecular biologists. The basics of molecular biology are discussed. The protein structure is presented and these are linked to the theory of computations. The chapter may be used as a first hand guide for the researchers interested in the field of Molecular Modelling and Drug Design.

Prediction of protein-protein and protein-ligand interactions

The protein-protein and protein-ligand interaction prediction methodologies can be broadly classified into two categories, viz., the experimental determination techniques and the computational. In most of the cases the two types of methods are combined together to complement each other.

4.3 EXPERIMENTAL METHODOLOGIES

The prediction of protein-protein and protein-ligand interactions requires the determination of the quaternary structures of the proteins. This needs the background knowledge of the subunit composition of the system under investigation. The subunit composition of the proteins may be determined by applying chemical methods like introducing chemical cross-links between the polypeptide chains. This may also be done by comparing the molecular masses of the native protein and its constituent chains. The molecular masses of the subunits are obtained using denaturing gel electrophoresis.

The most accurate and important method of prediction of structures of protein-protein and protein-ligand complexes is X-ray crystallography. However, there are several other experimental techniques for the purpose, viz., NMR spectroscopy, fluorescence resonance energy transfer, yeast two-hybrid, affinity purification/mass spectrometry, protein chips to name a few.

X-ray crystallography

X-ray crystallography is a method to determine the arrangement of atoms within a crystal. In this method, a beam of X-rays are allowed to strike a crystal and then it scatters into many different directions. The electron density map of the molecule can be generated from the angles and intensities of these scattered X-ray beams to build a three-dimensional picture of the distribution of electrons within the crystal. This leads to the determinations of the mean positions of the atoms in the crystal, as well as, their chemical bonds with which the atoms are held together and the disorder and various other structural properties. A fully grown crystal is mounted on the apparatus called goniometer. After that the crystal is rotated gradually and X-rays are passed through it, which produce a diffraction pattern of spots regularly spaced in two dimensions. These are known as reflections. A three-dimensional electron density model of the whole crystal is then created from the images of the crystal which are taken by rotating it at different orientations with the help of Fourier transforms, as well as with previous chemical data for the crystallized sample. Small crystals or deformities in crystal packing lead to erroneous results. X-ray crystallography is related to several other methods for determining atomic structures. Similar diffraction patterns can be produced by scattering electrons or neutrons, which are likewise interpreted as a Fourier transform.

The structure of Hexamethylenetetramine was solved in 1923, and this happened to be the structure of the first organic molecule to be solved by X-ray crystallography. After that, structures of a number of important bioorganic molecules like, porphyrin, corrin and chlorophyll were solved.

X-ray crystallography of biological molecules took off with Dorothy Crowfoot Hodgkin, who solved the structures of cholesterol (1937), vitamin B12 (1945) and penicillin (1954). In 1969, she succeeded in solving the structure of insulin.

X-crystallography has a widespread use in the elucidation of protein structure function relationship, mutational analysis and drug design. The challenge lies in the prediction of structures of membrane proteins, like ion channels and receptors as it is difficult to find appropriate system for them to crystallize as these proteins are integral parts of the cell membranes and it is hard to get the protein part out from the membrane component.

Nuclear Magnetic Resonance (NMR) Techniques

Protein nuclear magnetic resonance spectroscopy (usually abbreviated as protein NMR) is a field of structural biology in which NMR spectroscopy is used to obtain information about the structure and dynamics of protein-protein and protein-ligand complexes. Structure determination by NMR spectroscopy usually consists of following phases, each using a separate set of highly specialized techniques. The sample is prepared; resonances are assigned; restraints are generated and a structure is calculated and validated.

FÖRSTER Resonance Energy Transfer / Fluorescence Resonance Energy Transfer (FRET)

It is a mechanism describing energy transfer between two molecules both of which should be sensitive to light. This method is used to study protein dynamics, protein-protein and protein-ligand interactions. For FRET analysis of protein interactions, the cyan fluorescent protein (CFP)-yellow fluorescent protein (YFP) pair, which is the colour variants of the green fluorescent protein (GFP), is currently the most useful protein pair that is being employed in biology. The interactions between the proteins are determined by the amount of energy that is being transferred between the proteins, thereby creating a large emission peak of YFP obtained by the overlaps of the individual fluorescent emission peaks of CFP and YPF as the two proteins are near to each other.

Yeast Two-Hybrid System

One of the important methods to analyze the physical interactions between proteins or proteins with ligand is the yeast two-hybrid screening. The basic principle behind the process comes from the fact that the close proximity and modularity of the activating and the binding domains of most of the eukaryotic transcription factors lead to the interactions between themselves albeit indirectly. This system often utilizes a genetically engineered strain of yeast which does not possess the biosynthetic machinery required for the biosynthesis of amino acids or nucleic acids. Yeast cells do not survive on the media lacking these nutrients. In order to detect the interactions between the proteins, the transcription factor is divided into two domains called the binding (BD) and the activator domain (AD). Genetically engineered plasmids are made to produce a protein product with the DNA binding domain (BD) attached onto a protein. Another such plasmid codes for a protein product having the activation domain (AD) tagged to another protein. The protein fused to the BD may be referred to as the bait protein and is typically a known protein that is used to identify the new binding partners. The protein fused to the AD may be referred to as the prey protein and can either be a single known protein or a collection of known or unknown proteins. The transcription of the reporter gene(s) occurs if and only if the AD and the BD of the transcription factors are connected bringing the AD close to the transcription start site of the reporter gene, which justifies the presence of physical interactions between the bait and the prey proteins. Thus, a fruitful interaction between the proteins fused together determines the phenotypic change of the cell.

Affinity Purification

It is a technique to study protein-protein and protein-ligand interactions. It involves creating a fusion protein with a designed piece, the tag, on the end. The protein of interest with the tag first binds to beads coated with Immunoglobulin G. Then the tag is broken apart by an enzyme. Finally a different part of the tag binds reversibly to beads of a different type. After the protein of interest has been washed through two affinity columns, it can be examined for binding partners.

Protein Chips / Protein Microarray

This method is sometimes referred to as a protein binding microarray. It provides a multiplex approach to identify protein-protein and protein-ligand interactions. This method is mainly applied to identify transcription factor protein-activation, or to identify the targets of biologically active small molecules. On a piece of glass, different protein molecules or DNA binding sequences of proteins are affixed orderly at different locations forming a microscopic array. A commonly used microarray is obtained by affixing antibodies which bind antigen molecules from cell lysate solutions. These antibodies can easily be spotted with appropriate dyes.

The aforementioned experimental tools are routinely used in laboratories to detect protein-protein and protein-ligand interactions. However, these methods are not full proof. On top of that, these experimental tools are labor-intensive and time-consuming. The methods are expensive and often they include false positives. These factors necessitate the use of other methods in order to verify the results. All these led to the development of computational methods which are capable of prediction of protein-protein and protein-ligand interactions with sufficient accuracies.

Computational Predictions of Protein-*Protein and Protein-Ligand Interactions*

Computational methodologies to predict protein-protein and protein-ligand interactions can broadly be classified as numerical value-based and probabilistic. Both of them involve training over a dataset containing protein structural and sequence information.

Numerical methods use a function of the form $F = f(p_i, p_j \in n_i, x)$

Where, p_i = input data for the amino acid residue i under consideration.

p_j = the corresponding properties of the spatially neighboring residues and $j \in n_i$

x = the collection of coefficients to be determined by training

The value of F determines the characteristics of residue under consideration. It can either be I for interface or be N for non-interface.

If the value of F is above a certain threshold i it is considered to be in I state otherwise in N state.

The value-based methods are classified as

- **Linear regression:** This method is used to predict the values of the unknown variables from a set of known variables. It also tests the relationship among the variables. In case of predictions of protein-protein and protein-ligand interactions solvent accessibilities of the amino acid residues of the proteins are taken as inputs. The different amino acid residues have different solvent accessibility values depending on whether they are exposed or buried. Based on a collection of such values of known proteins linear regression methods may be used to predict the nature of amino acids in unknown proteins.
- **Scoring function:** This is a general knowledge based approach. Scoring functions are based on empirical energy functions made up of contributions from various experimentally validated data. In this approach, information are generated from know protein-protein and protein-ligand complexes. The approach takes into account various types of physico-chemical parameters of the protein complexes, like solvation potential, solvent accessibilities, interaction free energies and entropies for example. These data are used to generate the scoring functions for the individual atoms of the amino acids constituting the protein complexes. The functions can then be used in case of unknown proteins to predict its mode of binding.
- **Support Vector Machines (SVMs)**: This is a supervised learning method for classification, function approximation, signal processing and regression analysis. In this method, the input data are divided into two different sets, viz., the interface (I) and non-interface (N) sets. The SVM will create a separating hyper-plane which maximizes the margin between the two different types of the data. The basic principle of SVM is first to train it with a set of known data to create a classifier. The classifier is then used to predict whether an amino acid residue is in I or N state by giving

it a probability score. SVM training is done by using a training file consisting of feature vectors generated using various information about the protein-protein and protein-ligand complexes. The typical information used for the purpose are hydrophobicities, accessible surface areas, electrical charges, sequence similarity and sequence conservation scores of amino acids etc. These information are combined into feature vectors and are used as input to train the SVM.

- **Neural network:** This is an interconnected group artificial neurons (In biological perspective, neurons are connectors that transmit information via chemical signaling between cells), which are basically an adaptive system. There are hidden layers whose output is fed into a final output node. Protein information like the solvent accessibilities, free energy of interactions etc. are fed into the hidden layers to train them. Next the trained model can be used to produce the output from a set of unknown data.
- **Random Forests:** This is a combination of tree predictors. Each of the tree is dependent on the respective values of a random vector that is sampled independently following the same distribution for all the trees in the whole forest generated by the method. The more the number of trees the less is the error. Due to the law of Large Numbers there are no over-fitting problems.

The other class of computational prediction methodology is the use of probabilistic methods. The probabilistic methods are employed to find the conditional probability p(s| $x_1,.,x_k$), where s is either I or N for the range of input data $x_1,.,x_k$ for a residue under consideration. Interface is predicted if the value of $p(s|\, x_1,.,x_k)$ becomes greater than a threshold value.

- **Naïve Bayesian**: This is a supervised machine learning method. It takes input data, which are assumed to be independent. Naïve Bayes is a well-known machine learning tool. This is a classifier that assumes no dependencies between the variables to predict the class of the object under study.

- **Bayesian network**: In this case, the input data are dependent on each other. So a joint probability is calculated. For two dependent input data, x_1 and x_2, their joint probability $P(S|x_1,x_2)$ is calculated as $P(x_1, x_2|S)$.

  $P(S)$ = fraction of state S in the training dataset. In case of protein-protein interactions S represents whether an amino acid residue of the protein under study is in the interface or not.

- **Hidden Markov Model (HMM):** Hidden Markov Models (HMMs) belong to the class of directed graphical models. HMMs define a factored probability distribution of p(x,y). The mathematical formulation of the model is expressed as

  $$p(x,y) = \prod p(x_i|y_i)p(y_i|y_{i-1})$$

  This model is often referred to as a generative model. The term $p(x_i|y_i)$ is considered to be the probability that the observed result x_i is generated from the input feature y_i. The second term, $p(y_i \mid y_{i-1})$ is the first-order Markov assumption term. This term represents that the probability of a label variable y_i is not related to the other label variables $y_{i\text{-}1}$ which are used in the study.

  In case of prediction of protein-protein and protein-ligand interactions this method takes into account a multiple sequence alignment (MSA) of known proteins. The amino acid residues which are found to be conserved in the MSA are used to construct a sequence profile which is used to predict the nature of the amino acids from the protein of interest.

4.4 SOLUTIONS AND RECOMMENDATIONS

One of the major advantages of the use of molecular modelling techniques in solving structural problems in biology is the speed of calculations. A user can run several modelling jobs in parallel and is therefore able to perform with much higher speed. There are several experimental methods (like Yeast-two-hybrid, X-Ray Crystallography and Nuclear Magnetic Resonance spectroscopy) available for the analysis of protein interactions.

Side by side a number of computational algorithms have been developed to predict amino acid residues in protein-protein interfaces with accuracies typically in the range of 70%. There is no denying that the experimental approaches are more accurate and would give more comprehensive and reliable results; but they are time consuming and labour intensive besides being expensive. As alternatives to the experimental approaches, computational methods have been developed. They are comparatively less accurate but often give an overall idea of the whole process. There are various computational approaches with somewhat varying degrees of accuracies. The most important aspect is that computational approaches are cost effective, and requires less time. A word of caution is that none of the methods are cent percent accurate. To properly predict a PPI a number of information is needed. The best way to perform an experiment is first to use the computational algorithms to find a PPI and then test that via experimental means.

This is a list of docking servers and software tools. This is given for an easy reference of the computational tools available to study protein interactions.

http://protein3d.ncifcrf.gov/~keskino

http://dockground.bioinformatics.ku.edu/

http://www.ces.clemson.edu/compbio/protcom/

http://mips.gsf.de/proj/ppi/

http://mips.gsf.de/proj/yeast/CYGD/interaction/

http://dip.doe-mbi.ucla.edu/dip/Main.cgi

http://www.thebiogrid.org/

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

http://www.hprd.org/

http://wilab.inha.ac.kr/hpid/

http://www.ihop-net.org/UniPub/iHOP

http://insilico.csie.ntu.edu.tw:9999/point/

http://point.bioinformatics.tw/

http://www.compbio.dundee.ac.uk/www-pips

http://www.jcvi.org/mpidb/about.php

http://www.molecularconnections.com/home/en/home/products/NetPro

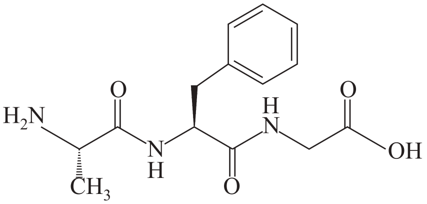
http://www.proteinlounge.com/inter_home.asp

http://itolab.cb.k.u-tokyo.ac.jp/Y2H/

http://mips.gsf.de/genre/proj/mpact/index.html

SOME IMPORTANT QUESTIONS AND ANSWERS

1. What is Molecular Modelling?
2. What is it good for?
3. What are the goals of using force fields?
4. How are the parameters in the force fields calculated?
5. What are topological properties of molecules?
6. What are structural parameters of molecules?
7. What are other properties of molecules?
8. What are structural properties of molecules?
9. What are electron densities?
10. What is electrostatic potential?
11. What is electrostatic potential map?
12. What is the correct name for the following peptide.



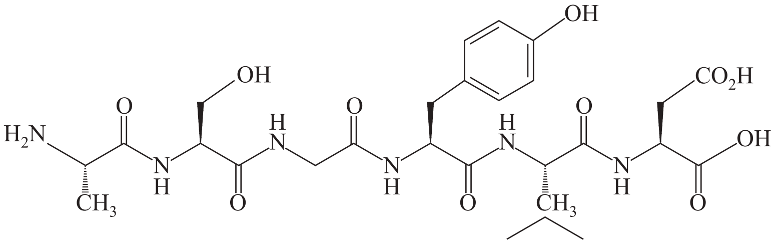
(a) L-alanyl-L-phenylalanyl-glycine

(b) glycyl-L-phenylalanyl-L-alanine

(c) L-phenylalanyl-L-alanyl-glycine

(d) L-alanyl-glycyl-L-phenylalanine

13. Identify the single letter code from the following single letter to represent the structure below



(a) DVYGSA

(b) ASGYVD

(c) EVFGSA

(d) ASGFVE

14. Identify the true statement about a peptide bond (RCONHR')?

(a) It is non planar.

(b) It is capable of forming a hydrogen bond.

(c) The cis configuration is favoured over the trans configuration.

(d) Single bond rotation is permitted between nitrogen and the carbonyl group.

15. Identify the incorrect statement about protein secondary structure?

(a) The alpha helix, beta pleated sheet and beta turns are examples of protein secondary structure.

(b) The ability of peptide bonds to form intramolecular hydrogen bonds is important to secondary structure.

(c) The steric influence of amino acid residues is important to secondary structure.

(d) The hydrophilic/hydrophobic character of amino acid residues is important to secondary structure.

16. Which of the following terms would refer to the arrangement of different protein subunits in a multiprotein complex?

(a) primary structure

(b) secondary structure

(c) tertiary structure

(d) quaternary structure

17. Which of the following terms would refer to the order in which amino acids are linked together in a protein?

(a) primary structure

(b) secondary structure

(c) tertiary structure

(d) quaternary structure

18. Which of the following terms would refer to the overall three-dimensional shape of a protein?

(a) primary structure

(b) secondary structure

(c) tertiary structure

(d) quaternary structure

19. Which of the following terms would refer to ordered region in a protein?

(a) primary structure

(b) secondary structure

(c) tertiary structure

(d) quaternary structure

20. What is the strongest form of intermolecular bonding that could be formed involving the residue of the amino acid serine.

(a) ionic bond

(b) hydrogen bond

(c) van der Waals interactions

(d) none of the above

ANSWERS

1. Molecular Modelling is the science that deals with the description of the atomic and molecular interactions. These interactions govern microscopic and macroscopic behaviors of physical systems

2. The utility of molecular modelling is to make connections between the microscopic and the macroscopic world provided by the theory of statistical mechanics

3. The basic aim of the force fields is to define the empirical potential energy functions V(r) to model the molecular interactions. These potential energy functions need to be differentiable for the computations of the forces acting on each atom: $F= -\Delta V(r)$

4. At first the theoretical analytical functional forms of the interactions are derived. For that purpose, the entire system is divided into a number of atom types which differ by their atomic number and chemical environment. Thus, the carbons in C=O or C-C are not of the same type as they are present in different chemical environments. The bonding parameters, like the bond enthalpy and bond free energy values, are then determined so as to reproduce the interactions between the various atom types by fitting procedures. The experimental enthalpy values of the covalent bonds are used in CHARMM force field. On the other hand, GROMOS and AMBER force fields are based on experimental free energies.

5. The topological properties are the descriptions of the covalent connectivity of the molecules to be modeled.

6. It is the initial or the starting conformation of a molecule, obtained from an X-ray structure, NMR data or a theoretical model.

7. These are the different types of forces acting on the molecule based on the force-field parameters.

8. These are the thermodynamic ensembles corresponding to the experimental conditions.

9. Electron densities require no prior knowledge about bonding which is quite different from the traditional methods. Therefore, they can be used to elucidate bonding. Electron densities also provide information about the transition states in absence of experimental evidences.

10. The electrostatic potential is the interaction energy of a point positive charge (known as an electrophile) with the electrons of a molecule. Negative electrostatic potentials would indicate areas that are prone to electrophilic attack.

11. The electrostatic potential map of a molecule provides the information about the distribution of charges in the molecule. It also gives information about the delocalization of electric charges.

12. (a)

13. (b)

14. (b)

15. (d)

16. (d)

17. (a)

18. (c)

19. (b)

20. (b)

CHAPTER

5 Phylogenetic Analysis

5.1 INTRODUCTION

Phylogenetic analysis deals with the study of evolutionary history of organisms. It also analyzes the relationships between various species. The Phylogenetic relationships are drawn by using common heritable elements like DNA sequences, r-RNA sequences, amino acid sequences of proteins. The results of these analyses are presented in the form of a tree known as the Phylogenetic tree. On the other hand, taxonomy which is generally used synonymously with Phylogenetic study is the branch of science that deals with the nomenclature, classification and identification of organisms. The purpose of Phylogenetic study is manifold as follows:

1. To study the evolutionary history: The main aim of a Phylogenetic study is to reconstruct the evolutionary history of the organism. The species diversities are also studied using this technique.

2. Another important application of the study is to determine the closest relative of the organism of interest.

3. With the help of Phylogenetic analyses, the functions of new genes may be identified or at least predicted.

4. Similarly the origin of the evolved gene of interest may be determined.

5. Phylogenetic studies based on biological sequence information could provide valuable information with accurate descriptions of

patterns of relatedness between the species before the advent of modern day molecular sequencing methods.

6. Phylogenetic studies are also useful in forensic sciences. Phylogenetic analyses are used to assess the DNA samples submitted in court cases to solve the problems of paternity and other crimes.

7. Phylogenetic analyses are routinely used to analyze the mode of pathogen outbreak. The Phylogenetic analyses are capable of elucidating the characters of new pathogenic species, their associations with their hosts and further their modes of transmissions.

8. Another important aspect of Phylogenetic studies is that the study provides ideas to conservation biologists to identify the nature of a species. The study also provides tentative ideas regarding the probabilities of a species to get extinct.

5.2 IMPORTANCE OF MOLECULAR SEQUENCE DATA

The Phylogenetic studies are almost entirely based on molecular sequence information. The mostly used molecular sequence data are the nucleotide sequences of DNA, r-RNA and the amino acid sequences of proteins. The molecular sequence data are used for Phylogenetic studies for the following reasons:

1. DNA is mainly the genetic material. It contains the genetic information and it is inherited from the ancestors.

2. Another important molecular sequence data that are mostly used in Phylogenetic analyses are the r-RNA. r-RNA is present in all cells. Thus, r-RNA encoding genes are sequenced and the r-RNA sequence is analyzed to identify taxonomic group of the organism. From the use of r-RNA sequence data the related species are also identified and consequently the rates of species divergence can be estimated.

3. The amino acid sequence of a protein can also be used as the starting material for Phylogenetic studies. The amino acid

sequences of proteins are the fundamental building blocks responsible for the structure and function of the protein.

4. The most important aspect of the molecular sequences is that the molecular sequences are highly specific and rich in information.

A word of caution regarding the use of molecular sequence information for Phylogenetic analyses is that DNA sequences can only be used if there are high sequence identities between them. Otherwise, the DNA sequences are to be converted to amino acid sequences for the analysis purpose.

However, a situation might arise where the molecular sequence data are not available. For example, when analyzing the ancient fossil samples, the molecular sequence data might not be available. In such situations, morphological features may be used for the analysis purposes. It is redundant to mention that such procedures are less accurate as the same morphological feature might arise from a large number of independent evolutionary events. It may further be noted that biochemical properties can also be used in Phylogenetic studies.

However, an empirical guiding principle for Phylogenetic analyses is as follows:

(a) In order to construct a Phylogenetic relationship between closely related species a rapidly evolving molecule, like mitochondrial DNA, is used.

(b) On the other hand, for finding the Phylogenetic relationship between diverse species, a highly conserved molecule like ribosomal RNA (r-RNA) is to be used.

Phylogenetic studies can be analyzed in three major ways. They are

- Phenetics: In this type of Phylogenetic analyses, the different species are grouped on the basis of their phenotypic resemblances. In such studies, all characteristics are taken into considerations.
- Cladistics: In this type of Phylogenetic analyses, the different species are grouped only with those species that share the similar characteristics.

- Evolutionary systematics: In this type of Phylogenetic study, both the phonetic and cladistics principles are used.

Among the aforementioned methods cladistics is considered to be the best method for Phylogenetic analyses as it incorporates the latest evolutionary theory.

5.3 SOME IMPORTANT TERMINOLOGIES AND DEFINITIONS

The Phylogenetic studies often involve the following terms:

A. Phylogenetic tree: It is the representation of evolutionary relationships between organisms. It is basically a tree like diagram where the branches of the tree would represent the speciation events. A Phylogenetic tree can be of different types.

(a) The most commonly used one is the Dendrogram. A Dendrogram is a broad term which encompasses all the different types of Phylogenetic trees.

(b) A Phylogram is a branching diagram representing Phylogenetic relationship. A Phylogram provides an estimate of a Phylogenetic relationship where the lengths of the tree are proportional to the extent of inferred evolutionary change.

(c) A Cladogram is a Phylogenetic diagram which is used to show the relations which is used to reveal the relationship between the ancestors and the descendants. In a Cladogram the lines would represent the different organisms coming from a common ancestor.

(d) A Chronogram is another type of Phylogenetic tree that is constructed on the basis of evolutionary time scale.

(e) A Spindle Diagram is a special type of Phylogenetic relationship which would represent the evolutionary history and the distribution of various taxonomic classes.

(f) Rooted Phylogenetic tree: A rooted Phylogenetic tree is a directed tree with a unique node which would represent the most common ancestor of all the species used in the study.

(g) Unrooted Phylogenetic tree: An unrooted Phylogenetic tree would depict the relatedness between the different species under study without the knowledge of a common ancestor.

(h) Homologs: Studies of protein and gene evolution are known to involve the comparison of *homologous species.* The sequences that have common origins but may or may not have common activity are considered to be belonging to homologous species. However, homologous species are considered to have a common ancestor.

(i) Orthologs: Orthologous species are homologs produced by speciation. These species are known to contain genes derived from a common ancestor that diverged due to divergence of the organisms they are associated with. However, the orthologous genes *would tend to have similar functions.*

(j) Paralogs: Paralogous species are homologs produced by gene duplications. These species are known to contain genes produced by gene duplications. However, the paralogous genes would tend to have different functions.

(k) Xenologs: Xenologs are homologous species produced by horizontal gene transfer between two organisms. The G+C content are found to be vastly different between the organisms. However, the xenologous genes tend to have similar functions.

(l) Last Common Ancestor (LCA): The species from which the new species are evolved.

(m) Mitochondrial DNA (mt-DNA): The DNA that is available in the mitochondria. Mt-DNA has a higher mutation rate than the nuclear DNA and therefore evolves more rapidly than the nuclear DNA.

(n) Ribosomal RNA (r-RNA): The ribosomal RNA has a specific secondary structure. It is a slowly evolving molecule.

(o) Distance matrix method: This is a basic method which involves the selection of two of the most closely related species from a bunch of species used for Phylogenetic study. These two species are clustered together and the operation is repeated until only one cluster is left.

(p) Unweighted Pair Group Method using Arithmetic Mean (UPGMA): In this method, the distance between the different species is calculated as arithmetic mean as all the species are given equal statistical weights.

(q) Weighted Pair Group Method using Arithmetic Mean (WPGMA): In this method, the distance between the different species is calculated on the basis of the population size of the cluster to which the species belongs. The cluster with a higher population is given a larger statistical weightage.

(r) Neighbour Joining (NJ): This is clustering method. This method is based on the principle of UPGMA technique. In this method, the closely related species as obtained from multiple sequence alignment are connected by internal nodes and they are referred to as Neighbours. The process is repeated until all the species under consideration are joined together to form a Phylogenetic tree.

(s) Maximum parsimony method: This method is based on the principle of finding the mutations that are needed to generate a sequence from the other. This is generally obtained from the analyses of the multiple sequence alignment results. The grouping of the species in the Phylogenetic tree is done by finding the species having sequences with minimum number of mutations.

(t) Maximum likelihood method: This method is also based on the analyses of the results of multiple sequence alignment. However, the method allows user intervention in selecting the closely related species which might not be reflected in the multiple sequence alignment.

5.4 SOME STATISTICAL ASPECTS IN PHYLOGENETICS

(i) **Akaike information criterion (AIC):** It is an estimator obtained from information theory. It measures the distance between the true model and the estimated model. The value of the AIC is derived from a maximum likelihood estimate that is penalized for the number of estimable parameters in the model. The selection of the model by AIC is based on a comparison between nested and non-nested models. The AIC framework also helps in the assessment of the uncertain parameters in model selection.

(ii) **Bayes factors:** It is a Bayesian analog to the likelihood ratio test. In this case, the fitness of the model is derived from integration over all possible parameter values that are required for fitness evaluation. The most important aspect of Bayes factors is that the model selection process does not require an a priori model selection criterion.

(iii) **Bayesian information criterion (BIC):** This parameter is similar to the AIC as it compares transformed likelihoods. However, it differs by penalizing the sample size and the number of estimable parameters. The BIC is an approximation of the natural logarithm of the Bayes factor. This makes this parameter as computationally more reliable than Bayes factors for Phylogenetic purposes. However, BIC generally selects less complex models than those selected by Bayes factors.

(iv) **Branch length estimations:** It represents the number of estimated changes in each lineage used in the Phylogenetic analyses. It is usually calculated as the number of substitutions per site. The extent of the measurements of the branch lengths depends on the actual number of substitutions in the sequence. This is generally reflected as the adequacy of the model.

(v) **Heterotachy:** It is the property of change in evolutionary rate at a given sequence position through time. This is very vital as any uncorrected data may lead to mismatching in Phylogenetic study. This is a source of generation of systematic error.

(vi) Jukes-Cantor model: It is the simplest substitution model that is used for DNA. There are different assumptions. The model assumes equal base frequencies and equal mutation rates in the DNA sequences. The only parameter of this model is the overall substitution rate. This variable turns out to be a constant when the mean-rate of base substitutions is normalized to 1.

(vii) Likelihood ratio test (LRT): It is a method to compare the maximum likelihood estimates of two nested models obtained using the same one data set. The significance of the models is assessed by an arbitrarily attained significance value (usually $p = 0.05$). It is to be noted that the test requires that one of the nested models being compared is the true model obtained from the data.

(viii) Model averaging: It is an information theoretic technique and it makes formal inferences based on a set of adequate models. This parameter is generally used when there are more than one models obtained for the same dataset with reasonably same accuracies. Model averaging techniques judge the relative importance of the individual models by calculating the relative weights. This technique utilizes the incorporations of uncertainties in model selection and in the estimation of model parameters.

(ix) ModelTest: It is a widely used computer programing technique that quickly fits one of the candidate models from a set of built models to a nucleotide data set using hierarchical likelihood ratio tests. This technique builds a neighbor-joining Jukes-Cantor tree. Then the maximized model likelihoods are estimated from the tree using the program PAUP. ModelTest also comes up with an AIC score for each model.

(x) Nested models: These are the statistical models which include all special cases of the most general model in the candidate set.

(xi) Phylogeny estimation: There are various approaches for inferring evolutionary relationships among living organisms (i.e. a phylogeny). All the methods for molecular data depend upon an implicit or explicit mathematical model which describes the evolution of aligned nucleotide or amino acid sequence characters. Thus, a Phylogenetic tree is referred to as an estimation of the true phylogeny of a group of organisms.

(xii) Systematic error: In Phylogenetic study, there are errors which arise due to the use of an inappropriate model of character evolution. The error is accentuated by the addition of sufficient data points. This process makes it 'systemic'. In Phylogenetics, less number of parameters might lead to under-fitting problems. On the other hand, the presence of too many parameters in a dataset would lead to over-fitting problem. All these might lead to create systematic error.

5.5 DIFFERENT ASPECTS OF PHYLOGENETIC STUDIES

The important features of Phylogenetic analyses are the visualizations of the species relationships. The visualizations are done by different types of Phylogenetic trees. The Phylogenetic trees have some important features as follows:

1. **The topology of the tree:** The topology of a Phylogenetic tree refers to the relatedness among the various species used for the construction of the tree. Phylogenetic trees having the same patterns of species distributions would reflect the similar kinds of biological relatedness among the species.

2. **Branches of the tree and branch lengths:** The branch lengths of the trees would represent the transmission of the genetic data among the species. The branch lengths would reflect the genetic changes or the mutations that are responsible for the speciation. A longer branching would indicate more variations among the species. The differences in genetic data among the species are measured by calculating the number of changes in the nucleotide or amino acid sequences.

3. **Nodes in Phylogenetic trees:** The nodes in a Phylogenetic tree are the end points in a Phylogenetic tree. The nodes would typically

represent a species in the Phylogenetic tree. The nodes can be of three different types:

(a) Tip: The sequences used to construct the Phylogenetic trees are present at the single terminal branches. They are called tips or external nodes.

(b) Internal nodes: The nodes where the different branches meet in the Phylogenetic trees are called internal nodes. The internal nodes would represent the ancestral sequence.

(c) Root: The root is the most recent common ancestor of all the sequences used for the construction of the Phylogenetic tree. It is the oldest part of the Phylogenetic tree. The root node in a Phylogenetic tree would determine the direction of flow of the genetic information. However, it is always difficult to identify the common ancestor of all the species and consequently, the addition of a root node in the Phylogenetic tree is not an easy task. The following figure would depict the rooted and un-rooted trees.

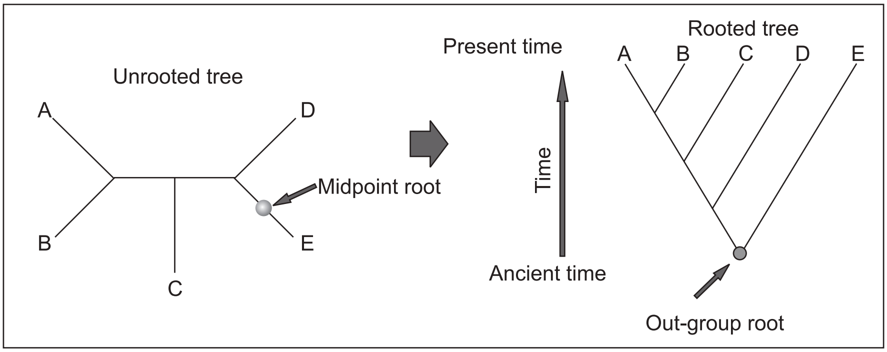


Figure: A, B, C, D and E are the tip nodes.

It is therefore very important to put a root node in a Phylogenetic tree as this is the most critical aspect of a Phylogenetic analysis. The presence of the root node would be the most critical aspect of determining the direction of flow of genetic information.

5.6 POSITING OF ROOT NODE IN A PHYLOGENETIC TREE

As mentioned earlier that the presence of a root node is the most critical aspect of a Phylogenetic analysis as this root node would guide the direction of flow of genetic data among the species. However, the positioning of the root node in a Phylogenetic tree is based on the following two assumptions:

1. Addition of the root node as and out-group: In this assumption a specific molecular information (sequence of DNA, r-RNA or amino acid sequence of a protein) is chosen in such a way that the sequence belongs to the same biological family of the sequences of interest to be used for the Phylogenetic analysis but it is more distantly related to them. The aforementioned specific molecular information is then considered as the out-group. The positioning of the root then simply depends on the sequence similarity measures between the sequences of interest and the out-group sequence.
2. Addition of the root to the midpoint of the Phylogenetic tree: In this assumption, it is considered that all the sequences are being evolved at the same time. The root node in such cases should be placed at the midpoint of the two longest branches.

5.7 STEPS FOR CONSTRUCTION OF PHYLOGENETIC TREES

There is no short-cut ways to construct a Phylogenetic tree. However, while building a Phylogenetic tree the following points are to be considered:

1. **To identify a sequence of interest:** The first step in Phylogenetics study is the presence of a sequence of interest for which the Phylogenetic relationship is to be studied. In principle anything starting from a single gene to the whole genome may be used for Phylogenetic analysis. However, the choice of sequence of interest is guided by the availability of the sequence. For example to study the microbial evolution pattern, the small subunit of r-RNA is generally used. However, the main drawback of using the r-RNA sequence is that the rates of evolution of the r-RNA sequences vary between different species. Another problem associated with the use of the r-RNA sequence is that the rate of evolution of r-RNA sequence is very slow, which makes it a difficult choice to study the Phylogenetic relationships of rapidly evolving genes.

2. **To gather molecular sequence information:** The other related sequences are to be gathered from existing databases.

3. **Alignment of the sequences:** The quality of a Phylogenetic tree depends on a proper alignment of the biological sequence under consideration. A faulty sequence alignment would produce an erroneous tree. The details of sequence alignment methods can be obtained from chapter 3.

4. **Determining the tree building method (substitution model):** The most commonly used methods for tree building are based on distance matrices. The method starts with creating n numbers of single clusters containing one species from the list of species under consideration based on the sequence alignments. This would create a $n \times n$ matrix. Then differences between the clusters are determined using a distance function. The clusters with the minimum distances between them are grouped together. This process is repeated until one cluster which is the most distant one from the rest of the clusters is left. There are different varieties of the distance matrix methods available. One of them is the Unweighted Pair Group Method Using Arithmetic Mean (UPGMA) method. In this method the distances between the clusters are determined as a simple average as each candidate is weighted equally. The UPGMA method assumes that the evolution occurs at the same rate on all branches of the tree and the distances between the nodes of the tree are additive. In the Weighted Pair Group Method Using Arithmetic Mean (WPGMA) method the clusters are weighted according to their sizes. The method of Neighbour Joining (NJ) is somewhat similar to UPGMA but it does not use the principle of addition. This method chooses the species on the basis of multiple sequence alignment results. The species with the maximum numbers of matching residues are considered to be belonging to the same cluster. Another method called Maximum Parsimony Method uses the information about the minimum numbers of mutations required to convert one sequence to another. This is generally achieved from the

multiple sequence alignment results. The Maximum Likelihood Method uses a user defined expectation model for determining the relatedness between the sequences.

5. **Building of Phylogenetic trees:** The Phylogenetic trees are constructed by using the information from the substitution model. The popular software tools for such analyses are PAUP and PHYLIP. There are other packages also like MEGA and MacClade which are also used these days.

6. **Evaluations of the trees:** There are no specific ways to verify the reliability of a Phylogenetic tree. If the different methods of tree construction would come up with similar results then it would indicate a good tree quality. Another way to analyze the reliability of a Phylogenetic tree is bootstrapping. In this method, the data are randomly sampled from any position within a multiple sequence alignment. A good Phylogenetic tree would have a high percentage of bootstrapping score.

7. **Determining the errors associated with the tree building method:** All the methods have their own short-comings. The results of Phylogenetic analyses are to be interpreted considering the inherent approximations associated with the methods employed.

8. **To come up with new biological ideas:** The ultimate goal of the Phylogenetic analysis is to identify the positioning of the sequence of interest in the realm of biological environment. A proper Phylogenetic analysis is capable of providing insights into the identity of the sequence of interest.

5.8 THE RELIABILITY OF A PHYLOGENETIC ANALYSIS

The most important aspect of Phylogenetics is the reliability issue. This is very critical for all types of computational analyses. In case of phylogenetics, there is no existing information regarding the ancestor for the sequence of interest. It is also a fact that some sequences are more informative than others. However, the most commonly used method of providing the confidence about a Phylogenetic analysis is bootstrapping. In the following figure the bootstrap values are given in a Phylogenetic tree.

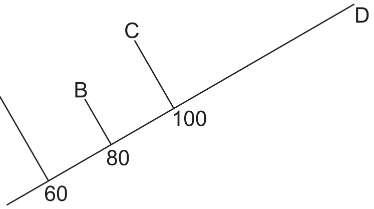


Figure: A Phylogenetic tree with bootstrap values of 60, 80 and 100

From the above figure it is clear that the Phylogenetic analyses have been performed 100 times with A, B, C and D. Out of 100 such replicates, 80% of times B appeared at the same position as presented in the tree while the same for A and C are 60% and 100% respectively. Thus, it is conclusive that C and D are distant neighbours and C is related to B.

All said and done, all the Phylogenetic analyses methods have their own shortcomings. Among them the major ones are:

(a) Limitations in sequence alignment methodologies

(b) Not being able to differentiate between the differences in rates of evolutions in the different parts of the same sequence

(c) Not being able to differentiate between the differences in rates of evolutions of the same sequence in the different species.

Different labs are constantly trying to sort out these problems. A few recent papers justifying such endeavors are presented in following section:

1. **RAxML version 8:** a tool for Phylogenetic analysis and post-analysis of large phylogenies, Alexandros Stamatakis, Bioinformatics, Volume 30, Issue 9, 1 May 2014, Pages 1312–1313, https://doi.org/10.1093/bioinformatics/btu033

2. **Verdant:** automated annotation, alignment and Phylogenetic analysis of whole chloroplast genomes, Michael R. McKain, Ryan H. Hartsock, Molly M. Wohl, Elizabeth A. Kellogg, Bioinformatics, Volume 33, Issue 1, 1 January 2017, Pages

130–132, https://doi.org/10.1093/bioinformatics/btw583

3. **ClockstaR:** choosing the number of relaxed-clock models in molecular Phylogenetic analysis, Sebastián Duchêne, Martyna Molak, Simon Y. W. Ho, Bioinformatics, Volume 30, Issue 7, 1 April 2014, Pages 1017–1019, https://doi.org/10.1093/bioinformatics/btt665

4. **MulRF:** a software package for Phylogenetic analysis using multi-copy gene trees,

   Ruchi Chaudhary, David Fernández-Baca, John Gordon Burleigh, Bioinformatics, Volume 31, Issue 3, 1 February 2015, Pages 432–433, https://doi.org/10.1093/bioinformatics/btu648

5. Unrealistic Phylogenetic trees may improve Phylogenetic footprinting, Martin Nettling, Hendrik Treutler, Jesus Cerquides, Ivo Grosse, Bioinformatics, Volume 33, Issue 11, 1 June 2017, Pages 1639–1646, https://doi.org/10.1093/bioinformatics/btx033

6. **markophylo:** Markov chain analysis on Phylogenetic trees, Utkarsh J. Dang, G. Brian Golding, Bioinformatics, Volume 32, Issue 1, 1 January 2016, Pages 130–132, https://doi.org/10.1093/bioinformatics/btv541

5.9 DETERMINATION OF EVOLUTIONARY RATE AND TIME

It is a well established fact that the genetic change in a lineage is measured as the product of evolutionary rate and time of divergence.

Genetic change: The genetic change is expressed as the number of mutations or substitutions per site.

Evolutionary rate: The evolutionary rate is the amount of genetic change per unit time. The time scale is measured in year.

Time of divergence: The time of divergence is the total time required for the entire speciation.

The mathematical form of calculation of the genetic change is

Genetic change = Evolutionary rate X Time of divergence

However, the appropriate calculation of the genetic change would require the knowledge of the root node.

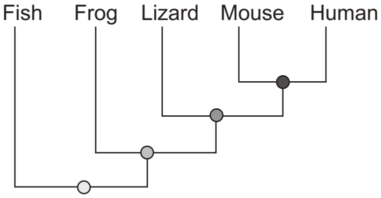


Figure: Evolution of vertebrates with their ancestors represented by the circles

From the above figure it can be concluded that human is closely related to mouse as both of them have a common ancestor expressed by the black circle. Similarly frogs share a recent ancestor with fish represented by another circle.

FINAL WORDS

Phylogenetic analyses would involve the use of biological sequences. The most commonly used sequences are amino acid sequences and DNA sequences. However, the choice of the biological sequence depends on the problem. For the evolutionary study of primates mitochondrial DNA is preferred. To study the relatedness between divergent species, Ribosomal RNA sequence is generally used. However, the success behind a good Phylogenetic analysis depends on the accuracy of the sequence alignment methods. The problem of sequence alignment can be solved by developing a robust alignment method. It is also to be noted that there is no guarantee that a Phylogenetic tree would represent a true evolutionary pathway. However, in order to solve the problem different methods of tree building may be employed. If the same result is obtained by all the methods, the Phylogenetic analyses may be considered to represent a true evolutionary picture. In general, bootstrapping is commonly used to randomly sample the available data. In an ideal case, the Phylogenetic trees obtained from bootstrapping analyses should match the original tree. However, in reality a bootstrapping result of 70% for any given branch in Phylogenetic tree is considered to provide nearly 95% confidence level about the correctness of the branch.

Some computational resources for Phylogenetic analyses

Ensemble (http://www.ensembl.org/index.html): This is a repository of genomic information

Ensembl compara (http://www.ensembl.org/info/genome/compara/index.html): This contains Phylogenetic information which can be freely downloaded.

ClustalW2Phylogeny(http://www.ebi.ac.uk/Tools/phylogeny/simple_phylogeny/): This tool can be used to construct Phylogenetic trees.

EMBOSS (https://www.ebi.ac.uk/Tools/sfc/emboss_seqret/): Another tool for Phylogenetic analysis.

Phyllip (http://evolution.genetics.washington.edu/phylip.html): A free software package for Phylogenetic analyses.

Dendroscope (http://ab.inf.uni-tuebingen.de/software/dendroscope/): It is a freely available tool for Phylogenetic study.

HyPhy (http://www.hyphy.org/): It is an online tool for testing the hypothesis of Phylogenetics.

MEGA (http://www.megasoftware.net/): It is a comprehensive tool for Phylogenetic analyses.

PAUP (http://www.paup.sc.fsu.edu/): It is a comprehensive tool for Phylogenetic analyses.

Splitstree (http://www.splitstree.org/): It is a tool for building un-rooted tree.

Treefinder (http://www.treefinder.de/): It is a comprehensive tool for Phylogenetic analyses.

Ugene (http://ugene.net/): It is a bioinformatics software

T-REX (www.trex.uqam.ca.): A webserver for Phylogenetic study

SOME IMPORTANT QUESTIONS AND ANSWERS

1. Overall phenotypic similarity may not always reflect evolutionary relationships—

(a) due to convergent evolution

(b) because of variation in rates of evolutionary change of different kinds of characters

(c) due to homoplasy

(d) all of the above.

2. Cladistics

(a) is based on overall similarity of phenotypes

(b) requires distinguishing similarity due to inheritance from a common ancestor from other reasons for similarity

(c) is not affected by homoplasy

(d) none of the above.

3. The principle of parsimony in Phylogenetics

(a) helps evolutionary biologists distinguish among competing Phylogenetic hypotheses

(b) does not require that the polarity of traits be determined

(c) is a way to avoid having to use outgroups in a Phylogenetic analysis

(d) cannot be applied to molecular traits.

4. The concept of Phylogenetic species

(a) depends on whether individuals from different populations can successfully breed

(b) is indistinguishable from the biological species concept

(c) does not apply to allopatric populations

(d) is based on evolutionary independence among populations.

5. Parsimony would suggest that parental care in birds, crocodiles, and some dinosaurs

(a) evolved independently, multiple times by convergent evolution

(b) evolved once in an ancestor common to all three groups

(c) is a homoplastic trait

(d) is not a homologous trait.

6. The reappearance of lost traits, especially if they are complex

(a) can be identified with Phylogenetic analyses

(b) never happens

(c) is not an example of a reversal

(d) does not affect interpretation of evolutionary relationships.

7. The term molecular clock pertaining to evolutionary biology and Phylogenetics

(a) refers to a group of proteins that induce endogenous circadian rhythms in animals

(b) is an undisputed assumption that all biological molecules evolve at a constant rate

(c) may help provide a way to estimate the absolute timing of historical events in evolution

(d) applies only to organisms that reproduce sexually.

8. A taxonomic group containing a common ancestor, which leaves out a descendant group is known as

(a) paraphyletic

(b) monophyletic

(c) polyphyletic

(d) a good cladistic group.

9. The forelimbs of birds and rhinoceros

(a) are homologous and symplesiomorphic

(b) are not homologous but are symplesiomorphic

(c) are homologous and synapomorphic

(d) are not homologous but are synapomorphic.

10. In order to determine polarity in different states of a character

(a) there must be a fossil record of the groups in question

(b) genetic sequence data must be available

(c) an appropriate name for the taxonomic group must be selected

(d) an outgroup must be identified.

11. A paraphyletic group

(a) includes an ancestor and all of its descendants

(b) an ancestor and some of its descendants

(c) descendants of more than one common ancestor

(d) all of the above.

12. Sieve tubes and sieve elements

(a) are homoplastic because they have different function

(b) are homologous because they have similar function

(c) are homoplastic because their common ancestor was single-celled

(d) are structures involved in transport within animals.

13. The phylogeny of dinosaurs leading to birds

(a) demonstrates that the first function of feathers was flight

(b) demonstrates that feathers and wings evolved simultaneously

(c) suggests that complex characters evolve rapidly, in one step

(d) reveals many transitional forms between modern birds and their ancestors.

14. The Phylogenetic analysis of HIV suggests

(a) a single origin of HIV from primates

(b) multiple origins of HIV from several different primate species

(c) multiple origins of HIV from a single primate species

(d) that SIV originated from HIV.

ANSWERS

1. (d) **2.** (b) **3.** (a) **4.** (d) **5.** (b) **6.** (a)

7. (c) **8.** (a) **9.** (a) **10.** (d) **11.** (b) **12.** (c)

13. (d) **14.** (b)

Few More Important Questions and Answers

1. The term Bioinformatics was invented by

(a) Paulien Hogeweg, 1979.

(b) Dr Margaret Oakley Dayhoff, 1976.

(c) Robert Ledley, 1978. D. David W Mount, 1977.

2. Find the odd one out

(a) GenBank.

(b) DDBJ.

(c) EMBL.

(d) TREMBL.

3. Find the odd one out

(a) SWISSPROT.

(b) EMBL.

(c) DDBJ.

(d) NCBI.

4. Find the secondary database

(a) DDBJ.

(b) PROSITE.

(c) NRDB.

(d) OWL.

5. Find a composite database.

(a) PROSITE.

(b) DDBJ.

(c) NRDB.

(d) EMBL.

6. Find a primary protein structure database.

(a) PDB.

(b) PubChem.

(c) ChemBank.

(d) SCOP.

7. Find a secondary protein structure database

(a) PubChem.

(b) PDB.

(c) ChemBank.

(d) SCOP.

8. FASTA format starts with which symbol?

(a) /.

(b) *.

(c) >.

(d) #.

9. A complementary DNA database is

(a) SwissProt.

(b) GenBank.

(c) UniSTS.

(d) NRDB.

10. Find a bibliographic database.

(a) PubMed.

(b) Entrez.

(c) PIR.

(d) EBI.

11. A life science search engine is

(a) PubMed.

(b) Entrez.

(c) Mozilla.

(d) EBI.

12. PubMed is maintained by

(a) NCBI.

(b) EMBL.

(c) DDBJ.

(d) SWISSPROT.

15. Entrez is maintained by

(a) SWISSPROT.

(b) EMBL.

(c) DDBJ.

(d) NCBI.

16. Find a similarity search tool.

(a) BLAST.

(b) CLUSTALW.

(c) CLUSTALX.

(d) RASMOL.

17. BLAST contains ------- programs.

(a) Four.

(b) Five.

(c) Six.

(d) Seven.

18. BLAST accepts ----- sequence format.

(a) GenBank.

(b) EMBL.

(c) FASTA.

(d) PIR.

19. Which tool compares protein sequence against protein databases?

(a) blastp.

(b) blastn.

(c) blastx.

(d) tblastx.

20. Which tool compares nucleotide sequence against DNA databases?

(a) blastn.

(b) blastp.

(c) tblastx.

(d) tblastn.

21. Which tool compares translated nucleotide query sequence against protein databases?

(a) blastp.

(b) tblastn.

(c) blastx

(d) tblastx.

22. Which tool compares protein sequence against translated nucleotide databases?

(a) blastp.

(b) tblastx.

(c) blastn.

(d) tblastn.

23. Which tool compares translated nucleotide query sequence against translated nucleotide databases?

(a) blastp.

(b) blastn.

(c) tblastx.

(d) tblastn.

24. Which Evalue provides the more significant the hit?

(a) lower. (b) higher.

(c) average. (d)superior.

25. SwissProt is housed and maintained by.

(a) NCBI. (b) NBRF.

(c) SIB. (d) DDBJ.

26. The full form of ExPASy is

(a) Expert Protein Analysis Server.

(b) Exponential Protein Analysis Server.

(c) Expert Protein Analysis System.

(d) Exponential Protein Analysis System.

27. Which web resource is maintained by NCBI for the retrieval of gene based information?

(a) LocusLink. (b) PDB.

(c) MSD. (d) PRF.

28. EST is

(a) Expressed Sequence Tag.

(b) Expressed Site Tag.

(c) Expressed Structure Tag.

(d) Expressed Symbol Tag.

29. SNP is

(a) Small Nucleic Polymorphism.

(b) Single Nucleic Polymorphism.

(c) Single Nucleotide Polymorphism.

(d) Small Nucleotide Polymorphism.

30. What is Evalue?

(a) The chance that a random sequence could achieve a better score than the query.

(b) The chance that a homologous sequence could achieve a similar score to the query.

(c) The chance that a random sequence could achieve a worse score than the query.

(d) The chance that a homologous sequence could achieve a better score than the query.

31. PROSITE is

(a) A database of protein structures.

(b) A database of protein sequences.

(c) A database of protein motifs.

(d) Options a and b.

37. Fingerprint is

(a) A protein family discriminator built from a set of regular expressions.

(b) A protein family discriminator built from a set of conserved motifs.

(c) A cluster of protein sequences gathered from a BLAST search.

(d) A cluster of protein sequences gathered from a FASTA search.

38. What are the conserved regions in multiple sequence alignments?

(a) Reflect areas of structural importance.

(b) Reflect areas of functional importance.

(c) Reflect areas of both functional and structural importance.

(d) Reflect areas likely to be of functional and/or structural importance.

39. How is a motif identified?

(a) COPIA. (b) Patternhunter.

(c) PROSPECT. D. BLAST.

43. A PATTERN is

(a) block. (b) profile.

(c) regular expression. (d) fuzzy regular expression.

44. InterPro is.

(a) integrated protein family.

(b) integrated protein sequence.

(c) integrated protein structure.

(d) integrated protein interaction.

45. More than two sequences are aligned by

(a) Multiple sequence alignment

(b) Pairwise sequence alignment

(c) Global alignment

(d) Local alignment

46. The family of related genes in an organism is called

(a) orthologs. (b) zoologs.

(c) paralogs. (d) xenologs.

47. PAM and BLOSUM are

(a) Distance Matrix (b) Sequence search method

(c) Both (d) None.

48. Which characteristics of a protein family are defined by the existence of a multiple sequence alignment of a group of homologous sequences?

(a) Domains. (b) DNA.

(c) Proteins. (d) RNA.

49. ClustalW accepts inputs in the ---- format.

(a) genbank.

(b) embl.

(c) pdb.

(d) fasta.

50. Comparison of two sequences is called

(a) Global alignment

(b) Local alignment

(c) Pairwise sequence alignment

(d) Multiple sequence alignment

51. Comparison of more than two sequences is called

(a) Global alignment

(b) Local alignment

(c) Pairwise sequence alignment

(d) Multiple sequence alignment

52. Highly similar sequences are aligned by

(a) Global alignment

(b) Local alignment

(c) Pairwise sequence alignment

(d) Multiple sequence alignment

53. An optimum alignment _____ number of matches and ______ the number of gaps.

(a) minimizes, maximizes.

(b) maximizes, minimizes.

(c) degrades, upgrades.

(d) upgrades, degrades.

54. Multiple sequence alignment method is known as.

(a) global. (b) local.

(c) progressive. (d) nonprogressive.

55. A Phylogenetic tree showing the branching representing evolutionary distance is

(a) Phylogram.

(b) Cladogram.

(c) A guide tree.

(d) Cardiogram.

56. Pfam database is obtained from .

(a) SMART.

(b) PRINTS.

(c) PROSITE. D. PRODOM.

57. High quality in Pfam is given by

(a) PfamA.

(b) PfamB.

(c) PfamC. D. PfamD.

58. Low quality of Pfam is given by

(a) PfamD.

(b) PfamB.

(c) PfamA. D. PfamC.

59. Full form of CDD is

(a) Conserved Domain Database.

(b) Conserved Dictionary Database.

(c) Conserved Domain Dictionary.

(d) Conserved Dictionary Database.

60. Which BLAST program is used to search conserved domains?

(a) BLASTN.

(b) BLASTP.

(c) SNPBLAST.

(d) PSIBLAST.

61. Which is RPS BLAST?

(a) PHIBLAST.

(b) PSIBLAST.

(c) BLASTN.

(d) TBLASTX.

62. The PRINTS database is linked to

(a) SwissProt/TrEMBL.

(b) SwissProt/EMBL.

(c) PIR/TrEMBL.

(d) PIR/EMBL.

63. Dotmatrix is

(a) as the coordinates of a two dimensional graph.

(b) are represented in the form of trees.

(c) as the coordinates of a 3D graph.

(d) not represented as graph.

64. PAM is

(a) Parallel Align Mutation.

(b) Point Altered Mutation.

(c) Point Accepted Mutation.

(d) Point Arranged Mutation.

65. Local alignment is done by

(a) Needleman and Wunsch. (b) PAM.

(c) SmithWaterman. (d) All the above.

66. Global alignment is done by

(a) Needleman and Wunsch. (b) SmithWaterman.

(c) BLAST. (d) PAM .

67. PAM was developed by

(a) Needleman and Wunsch, 1976.

(b) SmithWaterman, 1978.

(c) Dayhoff et al., 1978.

(d) Henikoff 1992.

68. BLOSUM is

(a) Blocks of Amino Acid Substitution Mutation.

(b) Basic Amino Acid Substitution Mutation.

(c) Blocks of Amino Acid Substitution Matrix.

(d) Basic Amino Acid Substitution Matrix.

69. PAM matrices are obtained from .

(a) Needleman and Wunsch.

(b) SmithWaterman.

(c) Dayhoff model.

(d) Markov model.

70. A mismatch refers to

(a) gap.

(b) deletion.

(c) insertion.

(d) substitution mutation.

71. Gene duplication leads to

(a) orthologs.

(b) paralogs.

(c) xenologs.

(d) zoologs.

79. Speciation leads to

(a) zoologs.

(b) paralogs.

(c) xenologs.

(d) orthologs.

80. The paired dots in the sequence alignments would determine

(a) conserved substitutions.

(b) semiconserved substitutions.

(c) gaps.

(d) identity.

81. A single dot in the sequence alignments would determine

(a) identity.

(b) semiconserved substitutions.

(c) conserved substitutions.

(d) gaps.

82. BLAST2 is used for comparison of .

(a) two. (b) three.

(c) four. (d) five.

83. Two main ways to construct guide tree in progressive alignment is

(a) UPGMA and Neighbor joining method.

(b) Maximum Parsimony.

(c) Maximum Likelihood.

(d) all the above.

84. Genetic recombination is shown by

(a) Progressive.

(b) Dynamic Programming.

(c) Genetic Algorithm.

(d) Hidden Markov Model.

85. Related protein sequences are used for

(a) PAM1.

(b) PAM45.

(c) PAM60.

(d) PAM250.

86. Distantly related proteins are used in

(a) PAM1.

(b) PAM250.

(c) PAM60.

(d) PAM45.

87. A Twilight Zone is

(a) Where alignments appear plausible and are statistically significant.

(b) Where alignments may appear plausible to the eye, but are no longer statistically significant.

(c) Where alignments neither appear plausible nor statistically significant.

(d) Where alignments share 30% identity.

88. Protein domains are presented by

(a) class. (b) architecture.

(c) taxonomy. (d) homologs.

89. CATH and SCOP are meant for

(a) the structural family to which a protein belongs.

(b) the generic family to which a protein belongs.

(c) homologous proteins.

(d) analogous proteins.

90. The coordinates of the amino acid residues in proteins are present in

(a) CATH. (b) SCOP.

(c) PDBsum. (d) PDB.

91. Homology modelling is

(a) Due to low sequence similarity between proteins of unknown and known structure, the structure is predicted from first principles.

(b) Due to high sequence similarity between proteins of unknown and known structure, the same function is assumed for both.

(c) Due to high sequence similarity between proteins of unknown and known structure, the structure of the latter is used as a template to model the former.

(d) A protein of unknown structure is compared against a library of fold templates to find the best match.

92. Threading is

(a) Due to low sequence similarity between proteins of unknown and known structure, the structure is predicted from first principles.

(b) Due to high sequence similarity between proteins of unknown and known structure, the same function is assumed for both.

(c) Due to high sequence similarity between proteins of unknown and known structure, the structure of the latter is used as a template to model the former.

(d) A protein of unknown structure is compared against a library of fold templates to find the best match.

93. PDBID is expressed as

(a) SMILES. (b) ROSDAL.

(c) WLN. (d) ALPHANUMERIC.

94. PDB does not accept the structures obtained by

(a) Xray crystallography. (b) NMR.

(c) Mass spectrometry. (d) Comparative modelling.

95. Fold recognition is

(a) Comparative modelling. (b) Threading.

(c) Abinitio. (d) Homology modelling.

96. SCOP is

(a) Similar Classification of Proteins.

(b) Structural Classification of Proteins.

(c) Similar Characterization of Proteins.

(d) Similar Classification of Proteins.

97. What is present in SCOP?

(a) primary structure of protein.

(b) secondary structure of protein.

(c) protein fold.

(d) mutated protein sequences.

98. Template based protein modelling is done in

(a) comparative modelling. (b) surface modelling.

(c) threading. (d) abinitio prediction.

99. Gaps are allowed in

(a) Maximum parsimony.

(b) Maximum likelihood.

(c) Neighbor Joining.

(d) Unweighted pair group method with arithmetic mean.

100. Distance based method is

(a) Unweighted pair group method with arithmetic mean.

(b) JukesCantor.

(c) Neighbuor Joining

(d) Jukes-Cantor

ANSWERS

1. (a)	**2.** (d)	**3.** (a)	**4.** (b)	**5.** (c)
6. (a)	**7.** (d)	**8.** (c)	**9.** (c)	**10.** (a)
11. (b)	**12.** (a)	**15.** (d)	**16.** (a)	**17.** (b)
18. (c)	**19.** (a)	**20.** (a)	**21.** (c)	**22.** (d)
23. (c)	**24.** (a)	**25.** (c)	**26.** (c)	**27.** (a)
28. (a)	**29.** (c)	**30.** (b)	**31.** (c)	**37.** (b)
38. (b)	**39.** (a)	**43.** (c)	**44.** (a)	**45.** (a)
46. (c)	**47.** (a)	**48.** (a)	**49.** (d)	**50.** (c)
51. (d)	**52.** (a)	**53.** (b)	**54.** (c)	**55.** (a)
56. (d)	**57.** (a)	**58.** (b)	**59.** (a)	**60.** (d)
61. (b)	**62.** (a)	**63.** (a)	**64.** (c)	**65.** (c)
66. (a)	**67.** (c)	**68.** (c)	**69.** (c)	**70.** (d)
71. (b)	**79.** (d)	**80.** (a)	**81.** (b)	**82.** (a)
83. (a)	**84.** (c)	**85.** (a)	**86.** (b)	**87.** (c)
88. (a)	**89.** (a)	**90.** (d)	**91.** (c)	**92.** (d)
93. (d)	**94.** (d)	**95.** (b)	**96.** (b)	**97.** (c)
98. (a)	**99.** (a)	**100.** (a)		

Further Reading

Allen, M. P., & Tildesley, D. J. (1989) *Computer Simulation of Liquids*, 1989, Oxford University Press, ISBN 0-19-855645-4.

Alberts, J., Lewis, R., & Roberts, W. (2008) *Molecular Biology of the Cell.*

Anfinsen, C. (1972). The formation and stabilization of protein structure. *Biochem. J.* **128** (4): 737–749.

Attwood, T.K, Parrysmith D-J (2001). Introduction to Bioinformatics. 1st edition. Published by Benjamin Cummings. ISBN-10: 0582327881, ISBN-13: 978-0582327887.

Baker, D., Sali, A. (2001). Protein structure prediction and structural genomics. *Science* **294** (5540): 93–96.

Baron, R., Setny, P., & Andrew M.J. (2010). Water in Cavity-Ligand Recognition". *Journal of the American Chemical Society* **132** (34): 12091–12097.

Baxevanis A. D., Francis Ouellette, B.F (2004). Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. Wiley-Blackwell; 3rd Revised edition. ISBN-10: 0471478784, ISBN-13: 978-0471478782.

Branden, C., & Tooze, J. (1999) *Introduction to Protein Structure*, Second Edition. Garland Publishing Inc., 19 Union Square West, NY 10003.

Bradford et al., (2006). Insights into protein–protein interfaces using a Bayesian network prediction method. *J. Mol. Biol.* 362 (2).

Bradford, J.R. & Westhead, D.R. (2005). Improved prediction of protein–protein binding sites using a support vector machines approach. *Bioinformatics.* 21(8).

Bradford, J.R. et al., (2006). Insights into Protein–Protein Interfaces using a Bayesian Network Prediction Method. *J.Mol.Biol.* 362(2).

Bogan, A.A. &Thorn, K.S. (1998) Anatomy of hot spots in protein interfaces. *J.Mol.Biol.* 280(1).

Bogan, A.A. &Thorn, K.S. (2001). ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics* 17(3).

Bowie, J.U., Lüthy, R., & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253 (5016): 164–170.

Brown, S.M. (2000) Bioinformatics: A Biologist's Guide to Biocomputing and the Internet. Eaton Pub Co. ISBN-10: 188129918X, ISBN-13: 978-1881299189.

Buehler, L.K, Rashidi, H.H (2005). Bioinformatics Basics: Applications in Biological Science and Medicine. CRC Press. ISBN 9780849312830.

Burgoyne, N.J. & Jackson, R.M. (2006) Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces. *Bioinformatics* 22(11).

Cavanagh, J., Fairbrother, W.J., Palmer, A.G. III, Skelton, N.J., Rance, M. (2006) *Protein NMR Spectroscopy: Principles and Practice,* Second Edition, Academic Press.

Chiang, Y.S., Gelfand, T.I., Kister, A.E., & Gelfand, I.M. (2007). New classification of supersecondary structures of sandwich-like proteins uncovers strict patterns of strand assemblage. *Proteins.* 68 (4): 915–921.

Chothia, C., & Lesk, A.M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J* 5 (4): 823–6.

Chung, S.Y., & Subbiah, S. (1996) A structural explanation for the twilight zone of protein sequence homology. *Structure* 4: 1123–27.

Clavarie J-M, Notredame C. (2006) Bioinformatics For Dummies. Second Edition. Wiley. ISBN: 978-0-470-08985-9.

Creighton T.E. (1992) *Proteins: Structures and Molecular Properties*, Second Edition. W. H. Freeman, NY.

Deniz, A. et al., (2000) Single-molecule protein folding: Diffusion fluorescence resonance energy transfer studies of the denaturation of chymotrypsin inhibitor 2. *Proc. Natl Acad. Sci. USA*, 97(10).

Dill, Ken A. et al. (2007). The protein folding problem: when will it be solved? Current Opinion in Structural Biology, 17:342–346.

Drenth, J. (2007) *Principles of Protein X-ray Crystallography*, Second Edition. Springer.

Ezkurdia *et al.* (2009) Progress and challenges in predicting protein-protein interaction sites. *Briefings in Bioinforamtics* 10(3).

Fariselli, P. et. al., (2002). Prediction of protein–protein interaction sites in heterocomplexes with neural networks. *Eur. J. Biochem* 269 (5).

Fiser, A, & Sali, A. (2003). "ModLoop: automated modelling of loops in protein structures". *Bioinformatics* 19 (18): 2500–1.

Frenkel, D., & Smit, B.(1996) *Understanding Molecular Simulation: From Algorithms to Applications* ISBN 0-12-267370-0.

Friedrich et al., (2006). Modelling interaction sites in protein domains with interaction profile hidden Markov models. *Bioinformatics.* 22(23).

Fresht, A. (1999) *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*, First Edition. W. H. Freeman, NY.

Gibas, C., Jambeck P. (2001) Developing Bioinformatics Computer Skills An Introduction to Software Tools for Biological Applications. O'Reilly Media.

Govindarajan, S., Recabarren, R., & Goldstein, R.A. (1999). Estimating the total number of protein folds. *Proteins.* 35 (4): 408–414.

Jones, D.T., Taylor, W.R., & Thornton, J.M. (1992). A new approach to protein fold recognition. *Nature* **358** (6381): 86–89.

Jones, S. & Thornton, J.M. (1997). Analysis of protein-protein interaction sites using surface patches. *J.Mol.Biol.* 272(1).

Jones, S. & Thornton, J.M. (1997). Thornton, Prediction of protein-protein interaction sites using patch analysis. *J.Mol.Biol.* 272(1).

Kaczanowski, S., & Zielenkiewicz, P. (2010). Why similar protein sequences encode similar three-dimensional structures?. *Theoretical Chemistry Accounts* **125**: 643–650.

Kitchen, D.B., Decornez, H., Furr, J.R., & Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews. Drug discovery***3** (11): 935–949.

Koike, A., & Takagi, T. (2004). Prediction of protein–protein interaction sites using support vector machines. *Protein Eng. Des. Sel.* 17(2).

Kyte, J (2006). *Structure in Protein Chemistry*, Second Edition. Garland Publishing Inc., 19 Union Square West, NY 10003

Leach, A. R. (2001). *Molecular Modelling: Principles and Applications*, ISBN 0-582-38210-6.

Lengauer, T., & Rarey, M. (1996). Computational methods for biomolecular docking. *Curr. Opin. Struct. Biol.* **6** (3): 402–406.

Lesk, A. (2008). Introduction to Bioinformatics. OUP Oxford; 3 edition, **ISBN-10:** 0199208042.

ISBN-13: 978-0199208043.

Levitt, M. (1992). Accurate modelling of protein conformation by automatic segment matching. *J Mol Biol* **226** (2): 507–33.

Mount, D.M. (2004). *Bioinformatics: Sequence and Genome Analysis* 2nd ed. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY.

Nelson, D. L., & Cox, M. L. (2008) Principles of Biochemistry 5th Edition. M. W. H. Freeman; June 15, 2008. ISBN 1429224169.

Ofran, Y. & Rost, B (2003). Analysing six types of protein–protein interfaces. *J.Mol.Biol.* 325(2).

Orego, C, Jones, D. , Thornton, J. (2002) Bioinformatics: Genes, Proteins and Computers. Taylor & Francis; 1 edition. ISBN-10: 1859960545, ISBN-13: 978-1859960547.

Ramachandran K.I Deepa, G. & Krishnan Namboori. P.K. (2008) *Computational Chemistry and Molecular Modelling Principles and Applications*. ISBN 978-3-540-77302-3 Springer-Verlag GmbH.

Rapaport,D. C. (2004) *The Art of Molecular Dynamics Simulation*, 2004, ISBN 0-521-82568-7.

Sali, A, Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234** (3): 779–815.

Voet, D., & Voet, J.G. (2010) *Biochemistry*, Second Edition. John Wiley & Sons, INC, NY.

Xiong, J. (2006). Essential Bioinformatics. Cambridge University Press. ISBN-10: 0521600820, ISBN-13: 978-0521600828.

Zhang Y (2008). Progress and challenges in protein structure prediction. *Curr Opin Struct Biol* **18** (3): 342–348.

Zemla, A. (2003). LGA – A Method for Finding 3-D Similarities in Protein Structures. *Nucleic Acids Research* 31 (13): 3370–3374.

Zvelebil, M, Baum, Jeremy O. (2007). Understanding Bioinformatics. Garland science publisher. ISBN: 9780815340249.

Index