SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF BIO AND CHEMICAL ENGINEERING
DEPARTMENT OF BIOTECHNOLOGY

Unit 1 – Bioinformatics – SBB3201

I. INTRODUCTION TO BIOINFORMATICS

Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biologicaldata. As an interdisciplinary field of science, bioinformatics combines computer science, statistics, mathematics, and engineering to analyze and interpret biological data. Bioinformatics has been used for in silico analyses of biological queries using mathematical and statistical techniques. Bioinformatics derives knowledge from computer analysis of biological data. These can consist of the information stored in the genetic code, but also experimental results from various sources, patient statistics, and scientific literature. Research in bioinformatics includes method development for storage, retrieval, and analysis of the data. Bioinformatics is a rapidly developing branch of biology and is highly interdisciplinary, using techniques and concepts from informatics, statistics, mathematics, chemistry, biochemistry, physics, and linguistics. It has many practical applications in different areas of biology and medicine.

Bioinformatics: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.

Computational Biology: The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.

"Classical" bioinformatics: "The mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information."

The National Center for Biotechnology Information (NCBI 2001) defines bioinformatics as: "Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline. There are three important sub-disciplines within bioinformatics: the development of new algorithms and statistics with which to assess relationships among members of large data sets; the analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures; and the development and implementation of tools that enable efficient access and management of different types of information

Even though the three terms: bioinformatics, computational biology and bioinformation infrastructure are often times used interchangeably, broadly, the three may be defined as follows:

1. bioinformatics refers to database-like activities, involving persistent sets of data that are maintained in a consistent state over essentially indefinite periods of time;

2. computational biology encompasses the use of algorithmic tools to facilitate biological analyses; while

3. bioinformation infrastructure comprises the entire collective of information management systems, analysis tools and communication networks supporting biology. Thus, the latter may be viewed as a computational scaffold of the former two.

There are three important sub-disciplines within bioinformatics:

- the development of new algorithms and statistics with which to assess relationships among members of large data sets;
- the analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures;
- and the development and implementation of tools that enable efficient access and management of different types of information

Bioinformatics definition - other sources

- Bioinformatics or computational biology is the use of mathematical and informational techniques, including statistics, to solve biological problems, usually by creating or using computer programs, mathematical models or both. One of the main areas of bioinformatics is the data mining and analysis of the data gathered by the various genome projects. Other areas are sequence alignment, protein structure prediction, systems biology, protein-protein interactions and virtual evolution. (source: www.answers.com)
- Bioinformatics is the science of developing computer databases and algorithms for the purpose of speeding up and enhancing biological research. (source: www.whatis.com)
- "Biologists using computers, or the other way around. Bioinformatics is more of a

tool than a discipline.(source: An Understandable Definition of Bioinformatics , The O'Reilly Bioinformatics Technology Conference, 2003) (4)

- The application of computer technology to the management of biological information. Specifically, it is the science of developing computer databases and algorithms to facilitate and expedite biological research.(source: Webopedia)
- Bioinformatics: a combination of Computer Science, Information Technology and Genetics to determine and analyze genetic information. (Definition from BitsJournal.com)
- Bioinformatics is the application of computer technology to the management and analysis of biological data. The result is that computers are being used to gather, store, analyse and merge biological data.(EBI - 2can resource)
- Bioinformatics is concerned with the creation and development of advanced information and computational technologies to solve problems in biology.
- Bioinformatics uses techniques from informatics, statistics, molecular biology and high-performance computing to obtain information about genomic or protein sequence data.

Bioinformaticist versus a Bioinformatician

A bioinformaticist is an expert who not only knows how to use bioinformatics tools, but also knows how to write interfaces for effective use of the tools.

A bioinformatician , on the other hand, is a trained individual who only knows to use bioinformatics tools without a deeper understanding.

Aims of Bioinformatics

In general, the aims of bioinformatics are three-fold.

1. The first aim of bioinformatics is to store the biological data organized in form of a database. This allows the researchers an easy access to existing information and submit new entries. These data must be annoted to give a suitable meaning or to assign its functional characteristics. The databases must also be able to correlate between different hierarchies of information. For example: GenBank for nucleotide and protein sequence information, Protein Data Bank for 3D macromolecular structures, etc.

2. The second aim is to develop tools and resources that aid in the analysis of data. For example: BLAST to find out similar nucleotide/amino-acid sequences, ClustalW to align two or more nucleotide/amino-acid sequences, Primer3 to design primers probes for PCR techniques, etc.

3. The third and the most important aim of bioinformatics is to exploit these computational tools to analyze the biological data interpret the results in a biologically meaningful manner.

Goals

The goals of bioinformatics thus is to provide scientists with a means to explain

1. Normal biological processes
2. Malfunctions in these processes which lead to diseases
3. Approaches to improving drug discovery

To study how normal cellular activities are altered in different disease states, the biological data must be combined to form a comprehensive picture of these activities. Therefore, the field of bioinformatics has evolved such that the most pressing task now involves the analysis and interpretation of various types of data. This includes nucleotide and amino acid sequences, protein domains, and protein structures. The actual process of analyzing and interpreting data is referred to as computational biology.

Important sub-disciplines within bioinformatics and computational biology include:

- Development and implementation of computer programs that enable efficient access to, use and management of, various types of information
- Development of new algorithms (mathematical formulas) and statistical measures that assess relationships among members of large data sets. For example, there are methods to locate a gene within a sequence, to predict protein structure and/or function, and to cluster protein sequences into families of related sequences.

The primary goal of bioinformatics is to increase the understanding of biological processes. What sets it apart from other approaches, however, is its focus on developing and

applying computationally intensive techniques to achieve this goal. Examples include: pattern recognition, data mining, machine learning algorithms, and visualization. Major research efforts in the field include sequence alignment, gene finding, genome assembly, drug design, drug discovery, protein structure alignment, protein structure prediction, prediction of gene expression and protein–protein interactions, genome-wide association studies, the modeling of evolution and cell division/mitosis.

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

Tools: Used in three areas

- Molecular Sequence Analysis
- Molecular Structural Analysis
- Molecular Functional Analysis

Over the past few decades, rapid developments in genomic and other molecular research technologies and developments in information technologies have combined to produce a tremendous amount of information related to molecular biology. Bioinformatics is the name given to these mathematical and computing approaches used to glean understanding of biological processes.

Common activities in bioinformatics include mapping and analyzing DNA and protein sequences, aligning DNA and protein sequences to compare them, and creating and viewing 3-D models of protein structures.

Bioinformatics encompasses the use of tools and techniques from three separate disciplines; molecular biology (the source of the data to be analyzed), computer science (supplies the hardware for running analysis and the networks to communicate the results), and the data analysis algorithms which strictly define bioinformatics. For this reason, the editors have decided to incorporate events from these areas into a brief history of the field.

A SHORT HISTORY OF BIOINFORMATICS

- 1933 A new technique, electrophoresis, is introduced by Tiselius for separating proteins in solution.

- 1951 Pauling and Corey propose the structure for the alpha-helix and beta-sheet (Proc. Natl. Acad. Sci. USA, 27: 205-211, 1951; Proc. Natl. Acad. Sci. USA, 37: 729- 740, 1951).
- 1953 Watson and Crick propose the double helix model for DNA based on x-ray data obtained by Franklin and Wilkins (Nature, 171: 737-738, 1953).
- 1954 Perutz's group develop heavy atom methods to solve the phase problem in protein crystallography.
- 1955 The sequence of the first protein to be analyzed, bovine insulin, is announced by F. Sanger.
- 1969 The ARPANET is created by linking computers at Stanford and UCLA.
- 1970 The details of the Needleman-Wunsch algorithm for sequence comparison are published.
- 1972 The first recombinant DNA molecule is created by Paul Berg and his group.
- 1973 The Brookhaven Protein Data Bank is announced (Acta. Cryst. B, 1973, 29: 1746).
- Robert Metcalfe receives his Ph.D. from Harvard University. His thesis describes Ethernet.
- 1974 Vint Cerf and Robert Kahn develop the concept of connecting networks of computers into an "internet" and develop the Transmission Control Protocol (TCP).
- 1975 Microsoft Corporation is founded by Bill Gates and Paul Allen.
- Two-dimensional electrophoresis, where separation of proteins on SDS polyacrylamide gel is combined with separation according to isoelectric points, is announced by P. H. O'Farrell (J. Biol. Chem., 250: 4007-4021, 1975).
- E. M. Southern published the experimental details for the Southern Blot technique of specific sequences of DNA (J. Mol. Biol., 98: 503-517, 1975).
- 1977 The full description of the Brookhaven PDB (http://www.pdb.bnl.gov) is published (Bernstein, F.C.; Koetzle, T.F.; Williams, G.J.B.; Meyer, E.F.; Brice, M.D.; Rodgers, J.R.; Kennard, O.; Shimanouchi, T.; Tasumi, M.J.; J. Mol. Biol., 1977, 112:, 535).
- Allan Maxam and Walter Gilbert (Harvard) and Frederick Sanger (U.K. Medical Research Council), report methods for sequencing DNA.
- 1980 The first complete gene sequence for an organism (FX174) is published. The gene consists of 5,386 base pairs which code nine proteins.

- Wuthrich et. al. publish paper detailing the use of multi-dimensional NMR for protein structure determination (Kumar, A.; Ernst, R.R.; Wuthrich, K.; Biochem. Biophys. Res. Comm., 1980, 95:, 1).
- IntelliGenetics, Inc. founded in California. Their primary product is the IntelliGenetics Suite of programs for DNA and protein sequence analysis.
- 1981 The Smith-Waterman algorithm for sequence alignment is published.
- IBM introduces its Personal Computer to the market.
- 1982 Genetics Computer Group (GCG) created as a part of the University of Wisconsin of Wisconsin Biotechnology Center. The company's primary product is The Wisconsin Suite of molecular biology tools.
- 1983 The Compact Disk (CD) is launched.
- 1984 Jon Postel's Domain Name System (DNS) is placed on-line.
- The Macintosh is announced by Apple Computer.
- 1985 The FASTP algorithm is published.
- The PCR reaction is described by Kary Mullis and co-workers.
- 1986 The term "Genomics" appeared for the first time to describe the scientific discipline of mapping, sequencing, and analyzing genes. The term was coined by Thomas Roderick as a name for the new journal.
- Amoco Technology Corporation acquires IntelliGenetics.
- NSFnet debuts.
- The SWISS-PROT database is created by the Department of Medical Biochemistry of the University of Geneva and the European Molecular Biology Laboratory (EMBL).
- 1987 The use of yeast artifical chromosomes (YAC) is described (David T. Burke, et. al., Science, 236: 806-812).
- The physical map of E. coli is published (Y. Kohara, et. al., Cell 51: 319-337).
- 1988 The National Center for Biotechnology Information (NCBI) is established at the National Cancer Institute.
- The Human Genome Initiative is started (Commission on Life Sciences, National Research Council. Mapping and Sequencing the Human Genome, National Academy Press: Washington, D.C.), 1988.
- The FASTA algorithm for sequence comparison is published by Pearson and Lupman.
- A new program, an Internet computer virus designed by a student, infects 6,000 military computers in the US.

- 1989 The Genetics Computer Group (GCG) becomes a private company.
- Oxford Molecular Group, Ltd. (OMG) founded in Oxford, UK by Anthony Marchington, David Ricketts, James Hiddleston, Anthony Rees, and W. Graham Richards. Primary products: Anaconda, Asp, Cameleon and others (molecular modeling, drug design, protein design).
- 1990 The BLAST program (Altschul, et. al.) is implemented.
- Molecular Applications Group is founded in California by Michael Levitt and Chris Lee. Their primary products are Look and SegMod which are used for molecular modeling and protein design.
- InforMax is founded in Bethesda, MD. The company's products address sequence analysis, database and data management, searching, publication graphics, clone construction, mapping and primer design.
- 1991 The research institute in Geneva (CERN) announces the creation of the protocols which make-up the World Wide Web.
- The creation and use of expressed sequence tags (ESTs) is described (J. Craig Venter, et. al., Science, 252: 1651-1656).
- Incyte Pharmaceuticals, a genomics company headquartered in Palo Alto California, is formed.
- Myriad Genetics, Inc. is founded in Utah. The company's goal is to lead in the discovery of major common human disease genes and their related pathways. The Company has discovered and sequenced, with its academic collaborators, the following major genes: BRCA1, BRCA2, CHD1, MMAC1, MMSC1, MMSC2, CtIP, p16, p19, and MTS2.
- 1992 Human Genome Systems, Gaithersburg Maryland, is formed by William Haseltine.
- The Institute for Genomic Research (TIGR) is established by Craig Venter.
- Genome Therapeutics announces its incorporation.
- Mel Simon and coworkers announce the use of BACs for cloning.
- 1993 CuraGen Corporation is formed in New Haven, CT.
- Affymetrix begins independent operations in Santa Clara, California
- 1994
- Netscape Comminications Corporation founded and releases Navigator, the commercial version of NCSA's Mozilla.
- Gene Logic is formed in Maryland.

- The PRINTS database of protein motifs is published by Attwood and Beck.
- Oxford Molecular Group acquires IntelliGenetics.
- 1995 The Haemophilus influenzea genome (1.8 Mb) is sequenced.
- The Mycoplasma genitalium genome is sequenced.
- 1996 Oxford Molecular Group acquires the MacVector product from Eastman Kodak.
- The genome for Saccharomyces cerevisiae (baker's yeast, 12.1 Mb) is sequenced.
- The Prosite database is reported by Bairoch, et.al.
- Affymetrix produces the first commercial DNA chips.
- 1997 The genome for E. coli (4.7 Mbp) is published.
- Oxford Molecular Group acquires the Genetics Computer Group.
- LION bioscience AG founded as an integrated genomics company with strong focus on bioinformatics. The company is built from IP out of the European Molecular Biology Laboratory (EMBL), the European Bioinformatics Institute (EBI), the German Cancer Research Center (DKFZ), and the University of Heidelberg.
- Paradigm Genetics Inc., a company focussed on the application of genomic technologies to enhance worldwide food and fiber production, is founded in Research Triangle Park, NC.
- deCode genetics publishes a paper that described the location of the FET1 gene, which is responsible for familial essential tremor, on chromosome 13 (Nature Genetics).
- 1998 The genomes for Caenorhabditis elegans and baker's yeast are published.
- The Swiss Institute of Bioinformatics is established as a non-profit foundation.
- Craig Venter forms Celera in Rockville, Maryland.
- PE Informatics was formed as a Center of Excellence within PE Biosystems. This center brings together and leverages the complementary expertise of PE Nelson and Molecular Informatics, to further complement the genetic instrumentation expertise of Applied Biosystems.
- Inpharmatica, a new Genomics and Bioinformatics company, is established by University College London, the Wolfson Institute for Biomedical Research, five leading scientists from major British academic centers and Unibio Limited.
- GeneFormatics, a company dedicated to the analysis and prediction of protein structure and function, is formed in San Diego.

- Molecular Simulations Inc. is acquired by Pharmacopeia
- 1999 deCode genetics maps the gene linked to pre-eclampsia as a locus on chromosome 2p13.
- 2000 The genome for Pseudomonas aeruginosa (6.3 Mbp) is published.
- The A. thaliana genome (100 Mb) is secquenced.
- The D. melanogaster genome (180Mb) is sequenced.
- Pharmacopeia acquires Oxford Molecular Group.
- 2001 The human genome (3,000 Mbp) is published.
- 2002 Chang Gung Genomic Research Center established.
- -Bioinformatics Center, -Proteomics Center, -Microarray Center

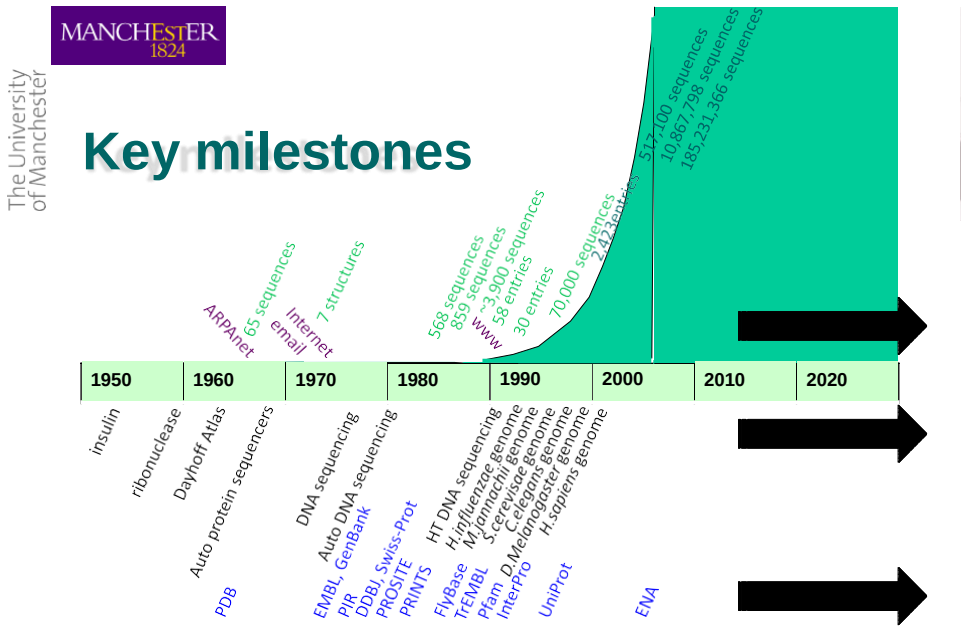


Figure 1

Applications

Bioinformatics joins mathematics, statistics, and computer science and information technology to solve complex biological problems. These problems are usually at the molecular level which cannot be solved by other means. This interesting field of science has many applications and research areas where it can be applied.

All the applications of bioinformatics are carried out in the user level. Here is the biologist including the students at various level can use certain applications and use the output in their research or in study. Various bioinformatics application can be categorized under following groups:

- Sequence Analysis
- Function Analysis
- Structure Analysis

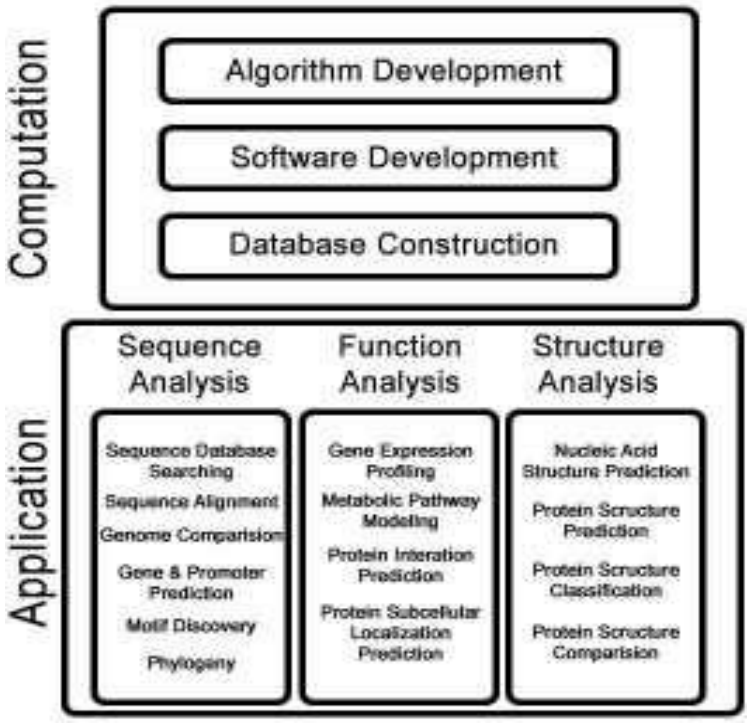


Figure 2

Sequence Analysis: All the applications that analyzes various types of sequence information and can compare between similar types of information is grouped under Sequence Analysis.

Function Analysis: These applications analyze the function engraved within the sequences and helps predict the functional interaction between various proteins or genes. Also expressional analysis of various genes is a prime topic for research these days.

Structure Analysis: When it comes to the realm of RNA and Proteins, its structure plays a vital role in the interaction with any other thing. This gave birth to a whole new branch

termed Structural Bioinformatics with is devoted to predict the structure and possible roles of these structures of Proteins or RNA

Sequence Analysis:

The application of sequence analysis determines those genes which encode regulatory sequences or peptides by using the information of sequencing. For sequence analysis, there are many powerful tools and computers which perform the duty of analyzing the genome of various organisms. These computers and tools also see the DNA mutations in an organism and also detect and identify those sequences which are related. Shotgun sequence techniques are also used for sequence analysis of numerous fragments of DNA. Special software is used to see the overlapping of fragments and their assembly.

Prediction of Protein Structure:-

It is easy to determine the primary structure of proteins in the form of amino acids which are present on the DNA molecule but it is difficult to determine the secondary, tertiary or quaternary structures of proteins. For this purpose either the method of crystallography is used or tools of bioinformatics can also be used to determine the complex protein structures.

Genome Annotation:-

In genome annotation, genomes are marked to know the regulatory sequences and protein coding. It is a very important part of the human genome project as it determines the regulatory sequences.

Comparative Genomics:-

Comparative genomics is the branch of bioinformatics which determines the genomic structure and function relation between different biological species. For this purpose, intergenomic maps are constructed which enable the scientists to trace the processes of evolution that occur in genomes of different species. These maps contain the information about the point mutations as well as the information about the duplication of large chromosomal segments.

Health and Drug discovery:

The tools of bioinformatics are also helpful in drug discovery, diagnosis and disease management. Complete sequencing of human genes has enabled the scientists to make medicines and drugs which can target more than 500 genes. Different computational tools and drug targets has made the drug delivery easy and specific because now only those cells can be targeted which are diseased or mutated. It is also easy to know the molecular basis of a disease.

Application of Bioinformatics in various Fields

Molecular medicine

The human genome will have profound effects on the fields of biomedical research and clinical medicine. Every disease has a genetic component. This may be inherited (as is the case with an estimated 3000-4000 hereditary disease including Cystic Fibrosis and Huntingtons disease) or a result of the body's response to an environmental stress which causes alterations in the genome (eg. cancers, heart disease, diabetes.). The completion of the human genome means that we can search for the genes directly associated with different diseases and begin to understand the molecular basis of these diseases more clearly. This new knowledge of the molecular mechanisms of disease will enable better treatments, cures and even preventative tests to be developed.

Personalised medicine

Clinical medicine will become more personalised with the development of the field of pharmacogenomics. This is the study of how an individual's genetic inheritence affects the body's response to drugs. At present, some drugs fail to make it to the market because a small percentage of the clinical patient population show adverse affects to a drug due to sequence variants in their DNA. As a result, potentially life saving drugs never make it to the marketplace. Today, doctors have to use trial and error to find the best drug to treat a particular patient as those with the same clinical symptoms can show a wide range of responses to the same treatment. In the future, doctors will be able to analyse a patient's genetic profile and prescribe the best available drug therapy and dosage from the beginning.

Preventative medicine

With the specific details of the genetic mechanisms of diseases being unravelled, the development of diagnostic tests to measure a persons susceptibility to different diseases may become a distinct reality. Preventative actions such as change of lifestyle or having treatment at the earliest possible stages when they are more likely to be successful, could result in huge advances in our struggle to conquer disease.

Gene therapy

In the not too distant future, the potential for using genes themselves to treat disease may become a reality. Gene therapy is the approach used to treat, cure or even prevent disease by changing the expression of a persons genes. Currently, this field is in its infantile stage with clinical trials for many different types of cancer and other diseases ongoing.

Drug development

At present all drugs on the market target only about 500 proteins. With an improved understanding of disease mechanisms and using computational tools to identify and validate new drug targets, more specific medicines that act on the cause, not merely the symptoms, of the disease can be developed. These highly specific drugs promise to have fewer side effects than many of today's medicines.

Microbial genome applications

Microorganisms are ubiquitous, that is they are found everywhere. They have been found surviving and thriving in extremes of heat, cold, radiation, salt, acidity and pressure. They are present in the environment, our bodies, the air, food and water. Traditionally, use has been made of a variety of microbial properties in the baking, brewing and food industries. The arrival of the complete genome sequences and their potential to provide a greater insight into the microbial world and its capacities could have broad and far reaching implications for environment, health, energy and industrial applications. For these reasons, in 1994, the US Department of Energy (DOE) initiated the MGP (Microbial Genome Project) to sequence genomes of bacteria useful in energy production, environmental cleanup, industrial processing and toxic waste reduction. By studying the genetic material of these organisms, scientists can begin to understand these microbes at a very fundamental level and isolate the genes that give them their unique abilities to survive under extreme conditions.

Waste cleanup

Deinococcus radiodurans is known as the world's toughest bacteria and it is the most radiation resistant organism known. Scientists are interested in this organism because of its potential usefulness in cleaning up waste sites that contain radiation and toxic chemicals.

Climate change Studies

Increasing levels of carbon dioxide emission, mainly through the expanding use of fossil fuels for energy, are thought to contribute to global climate change. Recently, the DOE (Department of Energy, USA) launched a program to decrease atmospheric carbon dioxide levels. One method of doing so is to study the genomes of microbes that use carbon dioxide as their sole carbon source.

Alternative energy sources

Scientists are studying the genome of the microbe Chlorobium tepidum which has an unusual capacity for generating energy from light

Biotechnology

The archaeon Archaeoglobus fulgidus and the bacterium Thermotoga maritima have potential for practical applications in industry and government-funded environmental remediation. These microorganisms thrive in water temperatures above the boiling point and therefore may provide the DOE, the Department of Defence, and private companies with heat-stable enzymes suitable for use in industrial processes Other industrially useful microbes include, Corynebacterium glutamicum which is of high industrial interest as a research object because it is used by the chemical industry for the biotechnological production of the amino acid lysine. The substance is employed as a source of protein in animal nutrition. Lysine is one of the essential amino acids in animal nutrition. Biotechnologically produced lysine is added to feed concentrates as a source of protein, and is an alternative to soybeans or meat and bonemeal. Xanthomonas campestris pv. is grown commercially to produce the exopolysaccharide xanthan gum, which is used as a viscosifying and stabilising agent in many industries. Lactococcus lactis is one of the most important micro-organisms involved in the dairy industry, it is a non-pathogenic rod-shaped bacterium that is critical for

manufacturing dairy products like buttermilk, yogurt and cheese. This bacterium, Lactococcus lactis ssp., is also used to prepare pickled vegetables, beer, wine, some breads and sausages and other fermented foods. Researchers anticipate that understanding the physiology and genetic make- up of this bacterium will prove invaluable for food manufacturers as well as the pharmaceutical industry, which is exploring the capacity of L. lactis to serve as a vehicle for delivering drugs.

Antibiotic resistance

Scientists have been examining the genome of Enterococcus faecalis-a leading cause of bacterial infection among hospital patients. They have discovered a virulence region made up of a number of antibiotic-resistant genes that may contribute to the bacterium's transformation from harmless gut bacteria to a menacing invader. The discovery of the region, known as a pathogenicity island, could provide useful markers for detecting pathogenic strains and help to establish controls to prevent the spread of infection in wards.

Forensic analysis of microbes

Scientists used their genomic tools to help distinguish between the strain of Bacillus anthryacis that was used in the summer of 2001 terrorist attack in Florida with that of closely related anthrax strains.

The reality of bioweapon creation

Scientists have recently built the virus poliomyelitis using entirely artificial means. They did this using genomic data available on the Internet and materials from a mail-order chemical supply. The research was financed by the US Department of Defence as part of a biowarfare response program to prove to the world the reality of bioweapons. The researchers also hope their work will discourage officials from ever relaxing programs of immunisation. This project has been met with very mixed feeelings

Evolutionary studies

The sequencing of genomes from all three domains of life, eukaryota, bacteria and archaea means that evolutionary studies can be performed in a quest to determine the tree of life and the last universal common ancestor.

Crop improvement

Comparative genetics of the plant genomes has shown that the organisation of their genes has remained more conserved over evolutionary time than was previously believed. These findings suggest that information obtained from the model crop systems can be used to suggest improvements to other food crops. At present the complete genomes of Arabidopsis thaliana (water cress) and Oryza sativa (rice) are available.

Insect resistance

Genes from Bacillus thuringiensis that can control a number of serious pests have been successfully transferred to cotton, maize and potatoes. This new ability of the plants to resist insect attack means that the amount of insecticides being used can be reduced and hence the nutritional quality of the crops is increased.

Improve nutritional quality

Scientists have recently succeeded in transferring genes into rice to increase levels of Vitamin A, iron and other micronutrients. This work could have a profound impact in reducing occurrences of blindness and anaemia caused by deficiencies in Vitamin A and iron respectively. Scientists have inserted a gene from yeast into the tomato, and the result is a plant whose fruit stays longer on the vine and has an extended shelf life.

Development of Drought resistance varieties

Progress has been made in developing cereal varieties that have a greater tolerance for soil alkalinity, free aluminium and iron toxicities. These varieties will allow agriculture to succeed in poorer soil areas, thus adding more land to the global production base. Research is also in progress to produce crop varieties capable of tolerating reduced water conditions.

Veterinary Science

Sequencing projects of many farm animals including cows, pigs and sheep are now well under way in the hope that a better understanding of the biology of these organisms will have huge impacts for improving the production and health of livestock and ultimately have benefits for human nutrition.

Comparative Studies

Analysing and comparing the genetic material of different species is an important method for studying the functions of genes, the mechanisms of inherited diseases and species evolution. Bioinformatics tools can be used to make comparisons between the numbers, locations and biochemical functions of genes in different organisms.
Organisms that are suitable for use in experimental research are termed model organisms. They have a number of properties that make them ideal for research purposes including short life spans, rapid reproduction, being easy to handle, inexpensive and they can be manipulated at the genetic level.

An example of a human model organism is the mouse. Mouse and human are very closely related (>98%) and for the most part we see a one to one correspondence between genes in the two species. Manipulation of the mouse at the molecular level and genome comparisons between the two species can and is revealing detailed information on the functions of human genes, the evolutionary relationship between the two species and the molecular mechanisms of many human diseases.

Table 1

Data source	Data size	Bioinformatics topics
Raw DNA sequence	11.5 million sequences (12.5 billion bases)	Separating coding and non-coding regions Identification of introns and exons Gene product prediction Forensic analysis
Protein sequence	400,000 sequences (~300 amino acids each)	Sequence comparison algorithms Multiple sequence alignments algorithms Identification of conserved sequence motifs
Macromolecular structure	15,000 structures (~1,000 atomic coordinates each)	Secondary, tertiary structure prediction 3D structural alignment algorithms Protein geometry measurements Surface and volume shape calculations Intermolecular interactions Molecular simulations (force-field calculations, molecular movements, docking predictions)
Genomes	300 complete genomes (1.6 million – 3 billion bases each)	Characterisation of repeats Structural assignments to genes Phylogenetic analysis Genomic-scale censuses (characterisation of protein content, metabolic pathways) Linkage analysis relating specific genes to diseases
Gene expression	largest: ~20 time point measurements for ~6,000 genes in yeast	Correlating expression patterns Mapping expression data to sequence, structural and biochemical data
Other data		
Literature	11 million citations	Digital libraries for automated bibliographical searches Knowledge databases of data from literature
Metabolic pathways		Pathway simulations

Definitions of Fields Related to Bioinformatics

Bioinformatics has various applications in research in medicine, biotechnology, agriculture etc.

Following research fields has integral component of Bioinformatics

1. **Computational Biology:** The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.
2. **Genomics:** Genomics is any attempt to analyze or compare the entire genetic complement of a species or species (plural). It is, of course possible to compare genomes by comparing more-or-less representative subsets of genes within genomes.
3. **Proteomics:** Proteomics is the study of proteins - their location, structure and function. It is the identification, characterization and quantification of all proteins involved in a particular pathway, organelle, cell, tissue, organ or organism that can be studied in concert to provide accurate and comprehensive data about that system. Proteomics is the study of the function of all expressed proteins. The study of the proteome, called proteomics, now evokes not only all the proteins in any given cell, but also the set of all protein isoforms and modifications, the interactions between them, the structural description of proteins and their higher-order complexes, and for that matter almost everything 'post-genomic'."
4. **Pharmacogenomics:** Pharmacogenomics is the application of genomic approaches and technologies to the identification of drug targets. In Short, pharmacogenomics is using genetic information to predict whether a drug will help make a patient well or sick. It Studies how genes influence the response of humans to drugs, from the population to the molecular level.
5. **Pharmacogenetics:** Pharmacogenetics is the study of how the actions of and reactions to drugs vary with the patient's genes. All individuals respond differently to drug treatments; some positively, others with little obvious change in their conditions and yet others with side effects or allergic reactions. Much of this variation is known to have a genetic basis. Pharmacogenetics is a subset of pharmacogenomics which uses genomic/bioinformatic methods to identify genomic correlates, for example SNPs (Single Nucleotide Polymorphisms), characteristic of particular patient response profiles and use those markers to

inform the administration and development of therapies. Strikingly such approaches have been used to "resurrect" drugs thought previously to be ineffective, but subsequently found to work with in subset of patients or in optimizing the doses of chemotherapy for particular patients.

6. **Cheminformatics:**

   Chemical informatics: 'Computer-assisted storage, retrieval and analysis of chemical information, from data to chemical knowledge.' This definition is distinct from Chemoinformatics which focus on drug design. ***chemometrics:*** The application of statistics to the analysis of chemical data (from organic, analytical or medicinal chemistry) and design of chemical experiments and simulations. ***computational chemistry:*** A discipline using mathematical methods for the calculation of molecular properties or for the simulation of molecular behavior. It also includes, e.g., synthesis planning, database searching, combinatorial library manipulation

7. **Structural genomics or structural bioinformatics** refers to the analysis of macromolecular structure particularly proteins, using computational tools and theoretical frameworks. One of the goals of structural genomics is the extension of idea of genomics, to obtain accurate three-dimensional structural models for all known protein families, protein domains or protein folds Structural alignment is a tool of structural genomics.
8. **Comparative genomics:** The study of human genetics by comparisons with model organisms such as mice, the fruit fly, and the bacterium E. coli.
9. **Biophysics:** The British Biophysical Society defines biophysics as: "an interdisciplinary field which applies techniques from the physical sciences to understanding biological structure and function".
10. **Biomedical informatics / Medical informatics:** "Biomedical Informatics is an emerging discipline that has been defined as the study, invention, and implementation of structures and algorithms to improve communication, understanding and management of medical information."
11. **Mathematical Biology:** Mathematical biology also tackles biological problems, but the methods it uses to tackle them need not be numerical and need not be implemented in software or hardware. It includes things of theoretical interest which are not necessarily algorithmic, not necessarily molecular in nature, and are

not necessarily useful in analyzing collected data.

12. **Computational chemistry:** Computational chemistry is the branch of theoretical chemistry whose major goals are to create efficient computer programs that calculate the properties of molecules (such as total energy, dipole moment, vibrational frequencies) and to apply these programs to concrete chemical objects. It is also sometimes used to cover the areas of overlap between computer science and chemistry.
13. **Functional genomics:** Functional genomics is a field of molecular biology that is attempting to make use of the vast wealth of data produced by genome sequencing projects to describe genome function. Functional genomics uses high-throughput techniques like DNA microarrays, proteomics, metabolomics and mutation analysis to describe the function and interactions of genes.
14. **Pharmacoinformatics:** Pharmacoinformatics concentrates on the aspects of bioinformatics dealing with drug discovery
15. **In silico ADME-Tox Prediction:** Drug discovery is a complex and risky treasure hunt to find the most efficacious molecule which do not have toxic effects but at the same time have desired pharmacokinetic profile. The hunt starts when the researchers look for the binding affinity of the molecule to its target. Huge amount of research requires to be done to come out with a molecule which has the reliable binding profile. Once the molecules have been identified, as per the traditional methodologies, the molecule is further subjected to optimization with the aim of improving efficacy. The molecules which show better binding is then evaluated for its toxicity and pharmacokinetic profiles. It is at this stage that most of the candidates fail in the race to become a successful drug.
16. **Agroinformatics / Agricultural informatics:** Agroinformatics concentrates on the aspects of bioinformatics dealing with plant genomes.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF BIO AND CHEMICAL ENGINEERING
DEPARTMENT OF BIOTECHNOLOGY

Unit 2 – Bioinformatics – SBB3201

II. DATABASES

Biological databases

Biological databases are libraries of life sciences information, collected from scientific experiments, published literature, high-throughput experiment technology, and computational analysis. They contain information from research areas including genomics, proteomics, metabolomics, microarray gene expression, and phylogenetics. Information contained in biological databases includes gene function, structure, localization (both cellular and chromosomal), clinical effects of mutations as well as similarities of biological sequences and structures.

Why databases?

- Means to handle and share large volumes of biological data
- Support large-scale analysis efforts
- Make data access easy and updated
- Link knowledge obtained from various fields of biology and medicine

Features

- Most of the databases have a web-interface to search for data
- Common mode to search is by Keywords
- User can choose to view the data or save to your computer
- Cross-references help to navigate from one database to another easily

Biological databases can be broadly classified into sequence and structure databases. Nucleic acid and protein sequences are stored in sequence databases and structure database only store proteins. These databases are important tools in assisting scientists to analyze and explain a host of biological phenomena from the structure of biomolecules and their interaction, to the whole metabolism of organisms and to understanding the evolution of species. This knowledge helps facilitate the fight against diseases, assists in the development of medications, predicting certain genetic diseases and in discovering basic relationships among

species in the history of life.

A biological database is a collection of data that is organized so that its contents can easily be accessed, managed, and updated. There are two main functions of biological databases:

- **Make biological data available to scientists.**
    - As much as possible of a particular type of information should be available in one single place (book, site, and database). Published data may be difficult to find or access and collecting it from the literature is very time-consuming. And not all data is actually published explicitly in an article (genome sequences!).
- **To make biological data available in computer-readable form.**
    - Since analysis of biological data almost always involves computers, having the data in computer-readable form (rather than printed on paper) is a necessary first step.

Data Domains

- Types of data generated by molecular biology research:
  – Nucleotide sequences (DNA and mRNA)
  – Protein sequences
  – 3-D protein structures
  – Complete genomes and maps

Sequence Databases

Nucleic acid sequence databases

EMBL • GenBank • DDBJ

Main protein sequence databases

Swiss Prot

also TREMBL, GenPept

Often integrated with other databases

Structure databases

- NDB, wwPDB, BMRB, CSD, EMDB

Biological databases can be broadly classified into sequence and structure databases. Sequence databases are applicable to both nucleic acid sequences and protein sequences, whereas structure database is applicable to only Proteins. The first database was created within a short period after the Insulin protein sequence was made available in 1956. Incidentally, Insulin is the first protein to be sequenced. The sequence of Insulin consisted of just 51 residues (analogous to alphabets in a sentence) which characterize the sequence. Around mid nineteen sixties, the first nucleic acid sequence of Yeast tRNA with 77 bases (individual units of nucleic acids) was found out. During this period, three dimensional structures of proteins were studied and the well known Protein Data Bank was developed as the first protein structure database with only 10 entries in 1972. This has now grown in to a large database with over 10,000 entries. While the initial databases of protein sequences were maintained at the individual laboratories, the development of a consolidated formal database known as SWISS-PROT protein sequence database was initiated in 1986 which now has about 70,000 protein sequences from more than 5000 model organisms, a small fraction of all known organisms. These huge varieties of divergent data resources are now available for study and research by both academic institutions and industries. These are made available as public domain information in the larger interest of research community through Internet (www.ncbi.nlm.nih.gov) and CDROMs (on request from www.rcsb.org). These databases are constantly updated with additional entries.

Databases in general can be classified in to **primary**, **secondary** and **composite** databases. A **primary** database contains information of the sequence or structure alone. Examples of these include Swiss-Prot & PIR for protein sequences, GenBank & DDBJ for Genome sequences and the Protein Databank for protein structures.

A **secondary** database contains derived information from the primary database. A secondary sequence database contains information like the conserved sequence, signature sequence and active site residues of the protein families arrived by multiple sequence alignment of a set of related proteins. A secondary structure database contains entries of the PDB in an organized way. These contain entries that are classified according to their structure like all alpha proteins, all beta proteins, etc. These also contain information on conserved secondary structure motifs of a particular protein. Some of the secondary database created and hosted by various researchers at their individual laboratories includes SCOP, developed at Cambridge University; CATH developed at University College of London, PROSITE of Swiss Institute of Bioinformatics, eMOTIF at Stanford.

Composite database amalgamates a variety of different primary database sources, which obviates the need to search multiple resources. Different composite database use different primary database and different criteria in their search algorithm. Various options for search have also been incorporated in the composite database. The National Center for Biotechnology Information (NCBI) which hosts these nucleotide and protein databases in their large high available redundant array of computer servers, provides free access to the various persons involved in research. This also has link to OMIM (Online Mendelian Inheritance in Man) which contains information about the proteins involved in genetic diseases.

Primary databases

I. Primary database

1. It is also known as archival database
2. Databases consisting of data derived experimentally such as nucleotide sequences and three dimensional structures are known as primary databases.
3. Experimental results are directly submitted into database by researchers across the globe
4. Example: Gen bank, DDBJ, SWISS-PROT

Contain sequence data such as nucleic acid or protein

Example of primary databases include :

Protein Databases

- SWISS-PROT
- TREMBL
- PIR

Nucleic Acid Databases

- EMBL
- Genbank
- DDBJ

II. Secondary database

1. It is also known as curated database
2. Databases consisting of data derived from the analysis of primary data such as sequences, secondary structures etc
3. It contains results of analysis of primary databases and significant data in the form of conserved sequences, signature sequences, active site residues of proteins etc.

Secondary databases

Or sometimes known as pattern databases

Contain results from the analysis of the sequences in the primary databases

Example of secondary databases include : PROSITE, Pfam, BLOCKS, PRINTS

Composite databases

Combine different sources of primary databases.

Make querying and searching efficient and without the need to go to each of the primary databases.

Example of composite databases include : NRDB – Non-Redundant DataBase, OWL

Genbank

GenBank, the National Institutes of Health (NIH) genetic sequence database, is an annotated collection of all publicly available nucleotide and protein sequences. The records within GenBank represent, in most cases, single, contiguous stretches of DNA or RNA with annotations. GenBank files are grouped into divisions; some of these divisions are phylogenetically based, whereas others are based on the technical approach that was used to generate the sequence information. Presently, all records in GenBank are generated from direct submissions to the DNA sequence databases from the original authors, who volunteer their records to make the data publicly available or do so as part of the publication process. GenBank, which is built by the National Center for Biotechnology Information (NCBI), is part of the International Nucleotide Sequence Database Collaboration, along with its two partners, the DNA Data Bank of Japan (DDBJ, Mishima, Japan) and the European Molecular Biology Laboratory (EMBL) nucleotide database from the European Bioinformatics Institute (EBI, Hinxton, UK). All three centers provide separate points of data submission, yet all three centers exchange this information daily, making the same database (albeit in slightly different format and with different information systems) available to the community at-large.

Only original sequences can be submitted to GenBank. Direct submissions are made to GenBank using BankIt, which is a Web-based form, or the stand-alone submission program, Sequin. Upon receipt of a sequence submission, the GenBank staff examines the originality of the data and assigns an accession number to the sequence and performs quality assurance checks. The submissions are then released to the public database, where the entries are retrievable by Entrez or downloadable by FTP. Bulk submissions of Expressed Sequence Tag (EST), Sequence-tagged site (STS), Genome Survey Sequence (GSS), and High- Throughput Genome Sequence (HTGS) data are most often submitted by large-scale sequencing centers. The GenBank direct submissions group also processes complete microbial genome sequences.

THE GENBANK FLATFILE: A DISSECTION The GenBank flatfile (GBFF) is the elementary unit of information in the GenBank database. It is one of the most commonly used formats in the representation of biological sequences. At the time of this writing, it is the format of exchange from GenBank to the DDBJ and EMBL databases and vice versa. The DDBJ flat file format and the GBFF format are now nearly identical to the

GenBank format. Subtle differences exist in the formatting of the definition line and the use of the gene feature. EMBL uses line-type prefixes, which indicate the type of information present in each line of the record.

The GBFF can be separated into three parts: the header, which contains the information (descriptors) that apply to the whole record; the features, which are the annotations on the record; and the nucleotide sequence itself. All major nucleotide database flat files end with

// on the last line of the record. The header is the most database-specific part of the record. The various databases are not obliged to carry the same information in this segment, and minor variations exist, but some effort is made to ensure that the same information is carried from one to the other.

The first line of all GBFFs is the Locus line:

Locus name

The locus name was originally designed to help group entries with similar sequences: the first three characters usually designated the organism; the fourth and fifth characters were used to show other group designations, such as gene product; for segmented entries, the last character was one of a series of sequential integers.

Sequence length

Number of nucleotide base pairs (or amino acid residues) in the sequence record.

Molecule Type

The type of molecule that was sequenced Genbank division

The GenBank division to which a record belongs is indicated with a three letter abbreviation. In this example, GenBank division is PRI.

The GenBank database is divided into 18 divisions:

1. PRI - primate sequences
2. ROD - rodent sequences
3. MAM - other mammalian sequences

4. VRT - other vertebrate sequences
5. INV - invertebrate sequences
6. PLN - plant, fungal, and algal sequences
7. BCT - bacterial sequences
8. VRL - viral sequences
9. PHG - bacteriophage sequences
10. SYN - synthetic sequences
11. UNA - unannotated sequences
12. EST - EST sequences (expressed sequence tags)
13. PAT - patent sequences
14. STS - STS sequences (sequence tagged sites)
15. GSS - GSS sequences (genome survey sequences)
16. HTG - HTG sequences (high-throughput genomic sequences)
17. HTC - unfinished high-throughput cDNA sequencing
18. ENV - environmental sampling sequences

Modification date

The date in the LOCUS field is the **date of last modification**. The sample record shown here was last modified on

Definition

Brief description of sequence; includes information such as source organism, gene name/protein name, or some description of the sequence's function

Accession

The unique identifier for a sequence record. An accession number applies to the complete record and is usually a combination of a letter(s) and numbers, such as a single letter followed by five digits (e.g., U12345) or two letters followed by six digits (e.g., AF123456). Accession numbers do not change, even if information in the record is changed at the author's request.

Version

If there is any change to the sequence data (even a single base), the version number will be increased, e.g., U12345.1 → U12345.2, but the accession portion will remain stable.

GI

"GenInfo Identifier" sequence identification number, in this case, for the nucleotide sequence. If a sequence changes in any way, a new GI number will be assigned. GI sequence identifiers run parallel to the new **accession.version** system of sequence identifiers

Keywords

Word or phrase describing the sequence. If no keywords are included in the entry, the field contains only a period.

Source

Free-format information including an abbreviated form of the organism name, sometimes followed by a molecule type.

Features

Information about genes and gene products, as well as regions of biological significance reported in the sequence. These can include regions of the sequence that code for proteins and RNA molecules, as well as a number of other features.

The **location of each feature** is provided as well, and can be a single base, a contiguous span of bases, a joining of sequence spans, and other representations. If a feature is located on the complementary strand, the word "complement" will appear before the base span

Source: Mandatory feature in each record that summarizes the length of the sequence, scientific name of the source organism, and Taxon ID number. Can also include other information such as map location, strain, clone, tissue type, etc., if provided by submitter.

Taxon: A stable unique identification number for the taxon of the source organism. A taxonomy ID number is assigned to each taxon

CDS:

Coding sequence; region of nucleotides that corresponds with the sequence of amino acids in a protein (location includes start and stop codons). The CDS feature includes an amino acid translation <1…206 Base span of the biological feature indicated to the left, in this case, a CDS feature Gene

A region of biological interest identified as a gene and for which a name has been assigned. The base span for the gene feature is dependent on the furthest 5' and 3' features.

Origin

The ORIGIN may be left blank, may appear as "Unreported," or may give a local pointer to the sequence start, usually involving an experimentally determined restriction cleavage site or the genetic locus (if available). This information is present only in older records.

The sequence data begin on the line immediately below ORIGIN.

DNA Data Bank of Japan

The DNA Data Bank of Japan (DDBJ) is a biological database that collects DNA sequences. It is located at the National Institute of Genetics (NIG) in the Shizuoka prefecture of Japan. It is also a member of the International Nucleotide Sequence Database Collaboration or INSDC. It exchanges its data with European Molecular Biology Laboratory at the European Bioinformatics Institute and with GenBank at the National Center for Biotechnology Information on a daily basis. Thus these three databanks contain the same data at any given time.

DDBJ began data bank activities in 1986 at NIG and remains the only nucleotide sequence data bank in Asia. Although DDBJ mainly receives its data from Japanese researchers, it can accept data from contributors from any other country.

DDBJ is primarily funded by the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT). DDBJ has an international advisory committee which consists of nine members, 3 members each from Europe, US, and Japan. This committee advises DDBJ about its maintenance, management and future plans once a year. Apart from this DDBJ also has an international collaborative committee which advises on various technical issues related to international collaboration and consists of working-level participants.

The format of DDBJ is similar to that of Genbank.

EMBL

The European Molecular Biology Laboratory (EMBL) is a molecular biology research institution supported by 21 member states, three prospect and two associate member states. EMBL was created in 1974 and is an intergovernmental organisation funded by public research money from its member states. Research at EMBL is conducted by approximately 85 independent groups covering the spectrum of molecular biology.

The Laboratory operates from five sites: the main laboratory in Heidelberg, and outstations inHinxton (the European Bioinformatics Institute (EBI), in England), Grenoble (France), Hamburg (Germany), and Monterotondo (near Rome). EMBL groups and laboratories perform basic research in molecular biology and molecular medicine as well as training for scientists, students and visitors. The organization aids in the development of services, new instruments and methods, and technology in its member states. Each of the different EMBL sites have a specific research field. The EMBL-EBI is a hub for bioinformatics research and services, developing and maintaining a large number of scientific databases, which are free of charge. At Grenoble and Hamburg, research is focused on structural biology. EMBL's dedicated Mouse Biology Unit is located in Monterotondo. Many scientific breakthroughs have been made at EMBL, most notably the first systematic genetic analysis of embryonic development in the fruit fly by Christiane Nüsslein-Volhard and Eric Wieschaus, for which they were awarded the Nobel Prize in Physiology or Medicine in 1995

EMBL format

A sequence file in EMBL format can contain several sequences. One sequence entry starts with an identifier line ("ID"), followed by further annotation lines. The start of the sequence is marked by a line starting with "SQ" and the end of the sequence is marked by two slashes ("//").

UniProt

UniProt is a comprehensive, high-quality and freely accessible database of protein sequence and functional information, many entries being derived from genome sequencing projects. It contains a large amount of information about the biological function of proteins derived from the research literature. Universal Protein resource, a central repository of protein data created by combining the Swiss-Prot, TrEMBL and PIR-PSD databases.

The UniProt consortium comprises the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB), and the Protein Information Resource (PIR). EBI, located at the Welcome Trust Genome Campus in Hinxton, UK, hosts a large resource of bioinformatics databases and services. SIB, located in Geneva, Switzerland, maintains the ExPASy(Expert Protein Analysis System) servers that are a central resource for proteomics tools and databases. PIR, hosted by the National Biomedical Research Foundation (NBRF) at the Georgetown University Medical Center in Washington, DC, USA, is heir to the oldest protein sequence database, Margaret Dayhoff's Atlas of Protein Sequence and Structure, first published in 1965.[2] In 2002, EBI, SIB, and PIR joined forces as the UniProt consortium

SWISSPROT

SWISS-PROT is an annotated protein sequence database, which was created at the Department of Medical Biochemistry of the University of Geneva and has been a collaborative effort of the Department and the European Molecular Biology Laboratory (EMBL), since 1987. SWISS-PROT is now an equal partnership between the EMBL and the Swiss Institute of Bioinformatics (SIB). The EMBL activities are carried out by its Hinxton Outstation, the European Bioinformatics Institute (EBI). The SWISS-PROT protein sequence database consists of sequence

entries. Sequence entries are composed of different line types, each with their own format. For standardisation purposes the format of SWISS-PROT (see http://www.expasy. ch/txt/userman.txt) follows as closely as possible that of the EMBL Nucleotide Sequence Database.

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 (Cross references).

Annotation

In SWISS-PROT two classes of data can be distinguished: the core data and the annotation. For each sequence entry the core data consists of the sequence data; the citation information (bibliographical references) and the taxonomic data (description of the biological source of the protein), while the annotation consists of the description of the following items:

- Function(s) of the protein
- Post-translational modification(s). For example carbohydrates, phosphorylation, acetylation, GPI-anchor, etc.
- Domains and sites. For example calcium binding regions, ATP-binding sites, zinc fingers, homeoboxes, SH2 and SH3 domains, etc.
- Secondary structure. For example alpha helix, beta sheet, etc.
- Quaternary structure. For example homodimer, heterotrimer, etc.
- Similarities to other proteins
- Disease(s) associated with deficiencie(s) in the protein
- Sequence conflicts, variants, etc.

Minimal redundancy

Many sequence databases contain, for a given protein sequence, separate entries which correspond to different literature reports. In SWISS-PROT we try as much as possible to merge all these data so as to minimise the redundancy of the database

Integration with other databases

It is important to provide the users of biomolecular databases with a degree of integration between the three types of sequence-related databases (nucleic acid sequences, protein sequences and protein tertiary structures) as well as with specialised data collections. Cross-references are provided in the form of pointers to information related to SWISS-PROT entries and found in data collections other than SWISS-PROT. For example the sample sequence mentioned above contains, among others, DR (Databank Reference) lines that point to EMBL, PDB, OMIM, Pfam and PROSITE.

TREMBL: A COMPUTER ANNOTATED SUPPLEMENT TO SWISS-PROT

Maintaining the high quality of sequence and annotation in SWISS-PROT requires careful sequence analysis and detailed annotation of every entry. This is the rate-limiting step in the production of SWISS-PROT. On one hand we do not wish to relax the high editorial standards of SWISS-PROT and it is clear that there is a limit to how much we can accelerate the annotation procedures. On the other hand, it is also vital that we make new sequences available as quickly as possible. To address this concern, we introduced in 1996 TrEMBL (Translation of EMBL nucleotide sequence database). TrEMBL consists of computer-annotated entries derived from the translation of all coding sequences (CDSs) in the EMBL database, except for CDSs already included in SWISS-PROT.

We have split TREMBL into two main sections, SP-TREMBL and REM-TREMBL. SP- TREMBL (SWISS-PROT TREMBL) contains entries (~55 000) which should be incorporated into SWISS-PROT. SWISS-PROT accession numbers have been assigned to these entries. SP- TREMBL is partially redundant against SWISS-PROT, since ~30 000 of these SP-TREMBL entries aie only additional sequence reports of proteins already in SWISS-PROT. REM-TREMBL (REMaining TREMBL) contains those entries (~15 000) that we do not wish to include in SWISS- PROT. This section is organized into four subsections. Most REM-TREMBL entries are immunoglobulins and T-cell receptors. We have stopped entering immunoglobulins and T-cell receptors into SWISS-PROT,

because we want to keep only germ line gene-derived translations of these proteins in SWISS-PROT and not all known somatic recombinant variations of these proteins. Another category of data which will not be included in SWISS-PROT is synthetic sequences. A third subsection consists of fragments with less than seven amino acids. The last subsection consists of CDS translations where we have strong evidence to believe that these CDS are not coding for real proteins.

The creation of TREMBL as a supplement to SWISS-PROT was not only for the purpose of producing a more complete and up to date protein sequence collection. Also to achieve a deeper integration of the EMBL nucleotide sequence database with SWISS-PROT + TREMBL.

Structure of a sequence entry

The entries in the SWISS-PROT data bank are structured so as to be usable by human readers as well as by computer programs. The explanations, descriptions, classifications and other comments are in ordinary English. Wherever possible, symbols familiar to biochemists, protein chemists and molecular biologists are used. Each sequence entry is composed of lines. Different types of lines, each with their own format, are used to record the various data which make up the entry.

Each line begins with a two-character line code, which indicates the type of data contained in the line. The current line types and line codes and the order in which they appear in an entry, are shown below:

ID - Identification.
AC - Accession number(s). DT - Date.
DE - Description. GN - Gene name(s).
OS - Organism species. OG - Organelle.
OC - Organism classification. RN - Reference number.
RP - Reference position. RC - Reference comments.
RX - Reference cross-references. RA - Reference authors.
RL - Reference location. CC - Comments or notes.
DR - Database cross-references. KW Keywords.
FT - Feature table data. SQ - Sequence header.

(blanks) sequence data. // - Termination line.

Protein Information Resource

The Protein Information Resource (PIR), located at Georgetown University Medical Center (GUMC), is an integrated public bioinformatics resource to support genomic and proteomic research, and scientific studies. PIR was established in 1984 by the National Biomedical Research Foundation (NBRF) as a resource to assist researchers and costumers in the identification and interpretation of protein sequence information. Prior to that, the NBRF compiled the first comprehensive collection of macromolecular sequences in the Atlas of Protein Sequence and Structure, published from 1964-1974 under the editorship of Margaret Dayhoff.

Dr. Dayhoff and her research group pioneered in the development of computer methods for the comparison of protein sequences, for the detection of distantly related sequences and duplications within sequences, and for the inference of evolutionary histories from alignments of protein sequences.

The Protein Information Resource (PIR) produces the largest, most comprehensive, annotated protein sequence database in the public domain, the PIR-International Protein Sequence Database, in collaboration with the Munich Information Center for Protein Sequences (MIPS) and the Japan International Protein Sequence Database (JIPID).

PIR, MIPS and JIPID constitute the PIR-International consortium that maintains the PIR- International Protein Sequence Database (PSD), the largest publicly distributed and freely available protein sequence database. The database has the following distinguishing features.

- It is a comprehensive, annotated, and non-redundant protein sequence database, containing over 142 000 sequences as of September 1999. Included are sequences from the completely sequenced genomes of 16 prokaryotes, six archaebacteria, 17 viruses and phages, >100 eukaryote organelles and Saccharomyces cerevisiae.

• The collection is well organized with >99% of entries classified by protein family and >57% classified by protein superfamily.

• PSD annotation includes concurrent cross-references to other sequence, structure, genomic and citation databases, including the public nucleic acid sequence databases ENTREZ, MEDLINE, PDB, GDB, OMIM, FlyBase, MIPS/Yeast, SGD/Yeast, MIPS/Arabidopsis and TIGR. Where these databases are publicly and freely accessible and provide suitable WWW access, the cross-references presented on the PIR WWW site are hot-linked so that searchers can consult the most current data.

• The PIR is the only sequence database to provide context cross-references between its own database entries. These cross-references assist searchers in exploring relationships such as subunit associations in molecular complexes, enzyme–substrate interactions, activation and regulation cascades, as well as in browsing entries with shared features and annotations.

• Interim updates are made publicly available on a weekly basis, and full releases have been published quarterly since 1984.

It is split into 4 distinct section (PIR1-PIR4).
PIR1: contains fully classified and annotated entries.
PIR2: includes preliminary entries not been thoroughly reviewed, contain redundancy. PIR3: contains unverified entries.
PIR4: fall into one of four categories.

- Conceptual translations of art factual sequences.
- Conceptual translations of sequences that are not transcribed or translated.
- Protein sequences or conceptual translations that are extensively genetically engineered
- Sequences that are not genetically encoded and not produced on ribosomes.

Protein data bank

The Protein Data Bank (PDB) is a crystallographic database for the three-dimensional structural data of large biological molecules, such as proteins and nucleic acids. The data, typically obtained by X-ray crystallography, NMR spectroscopy, or, increasingly, cryo-electron microscopy, and submitted by biologists and biochemists from around the world, are freely accessible on the Internet via the websites of its member organisations (PDBe, PDBj, and RCSB). The PDB is overseen by an organization called the Worldwide Protein Data Bank, wwPDB.

The PDB is a key resource in areas of structural biology, such as structural genomics. Most major scientific journals, and some funding agencies, now require scientists to submit their structure data to the PDB. Many other databases use protein structures deposited in the PDB. For example, SCOP and CATH classify protein structures, while PDBsum provides a graphic overview of PDB entries using information from other sources, such as Gene ontology

The Protein Data Bank (PDB) at Brookhaven National Laboratory (BNL), is a database containing experimentally determined three-dimensional structures of proteins, nucleic acids and other biological macromolecules. The archives contain atomic coordinates, citations, primary and secondary structure information, crystallographic structure experimental data, as well as hyperlinks to many other scientific databases.

Protein Data Bank (PDB) format is a standard for files containing atomic coordinates. Structures deposited in the Protein Data Bank at the Research Collaboratory for Structural Bioinformatics (RCSB) are written in this standardized format. The complete PDB file specification provides for a wealth of information, including authors, literature references, and the identification of substructures such as disulfide bonds, helices, sheets, and active sites.

Protein Data Bank format consists of lines of information in a text file. Each line of information in the file is called a record. A file generally contains several

different types of records, which are arranged in a specific order to describe a structure.

Table PDB Record Types

Selected Protein Data Bank Record Types	
Record Type	
ATOM	atomic coordinate record containing the x,y,z orthogonal Angstrom coordinates for atoms in standard residues (amino acids and nucleic acids).
HETATM	atomic coordinate record containing the x,y,z orthogonal Angstrom coordinates for atoms in nonstandard residues. Nonstandard residues include inhibitors, cofactors, ions, and solvent. The only functional difference from ATOM records is that HETATM residues are by default not connected to other residues. Note that water residues should be in HETATM records.
TER	indicates the end of a chain of residues. For example, a hemoglobin molecule consists of four subunit chains which are not connected. TER indicates the end of a chain and prevents the display of a connection to the next chain.
SSBOND	defines disulfide bond linkages between cysteine residues.
HELIX	indicates the location and type (right-handed alpha, *etc.*) of helices. One record per helix.
SHEET	indicates the location, sense (anti-parallel, *etc.*) and registration with respect to the previous strand in the sheet (if any) of each strand in the model. One record per strand.

The Protein Data Bank (pdb) file format is a textual file format describing the three- dimensional structures of molecules held in the Protein Data Bank. The pdb format accordingly provides for description and annotation of protein and nucleic acid structures including atomic coordinates, observed sidechain rotamers, secondary structure assignments, as well as atomic connectivity. Structures are often deposited with other molecules such as water, ions, nucleic acids, ligands and so on, which can be described in the pdb format as well. The Protein Data Bank also keeps data on biological macromolecules in the newer mmCIF file format.

A typical PDB file describing a protein consists of hundreds to thousands of lines like the following (taken from a file describing the structure of a synthetic collagen-like peptide):

HEADER, TITLE and AUTHOR records provide information about the researchers who defined the structure; numerous other types of records are available to provide other types of information.

REMARK records can contain free-form annotation, but they also accommodate standardized information; for records describe how to compute the coordinates of the experimentally observed multimer from those of the explicitly specified ones of a single repeating unit.

SEQRES records give the sequences of the three peptide chains (named A, B and C), which are very short in this example but usually span multiple lines.

ATOM records describe the coordinates of the atoms that are part of the protein. For example, the first ATOM line above describes the alpha-N atom of the first residue of peptide chain A, which is a proline residue; the first three floating point numbers are its x, y and z coordinates and are in units of Ångströms. The next three columns are the occupancy, temperature factor, and the element name, respectively.

HETATM records describe coordinates of hetero-atoms which are not part of the protein molecule.

Secondary Databases

- A biological database is a large, organized body of persistent data, usually associated with computerized software designed to update, query, and retrieve components of the data stored within the system.
- The chief objective of the development of a database is to organize data in a set of structured records to enable easy retrieval of information.
- Based on their contents, biological databases can be either primary

database or secondary databases.

- Among the two, secondary databases have become a biologist's reference library over the past decade or so, providing a wealth of information on just any research or research product that has been investigated by the research community.
- Sequence annotation information in the primary database is often minimal.
- To turn the raw sequence information into more sophisticated biological knowledge, much post-processing of the sequence information is needed.
- This begs the need for secondary databases, which contain computationally processed sequence information derived from the primary databases.
- Thus, secondary databases comprise data derived from the results of analyzing primary data.
- Secondary databases often draw upon information from numerous sources, including other databases (primary and secondary), controlled vocabularies and the scientific literature.
- They are highly curated, often using a complex combination of computational algorithms and manual analysis and interpretation to derive new knowledge from the public record of science.
- The amount of computational processing work, however, varies greatly among the secondary databases; some are simple archives of translated sequence data from identified open reading frames in DNA, whereas others provide additional annotation and information related to higher levels of information regarding structure and functions.

Importance of secondary databases

- Secondary databases contain information derived from primary sequence data which are in the form of regular expressions (patterns), Fingerprints, profiles blocks or Hidden Markov Models.
- The type of information stored in each of the secondary databases is different. But in secondary databases, homologous sequences may be gathered together in multiple alignments.

- In multiple alignments, there are conserved regions that show little or no variation between the constituent sequences. These conserved regions are called motifs.
- Motifs reflect some vital biological role and are crucial to the structure of the function of the protein. This is the importance of the secondary database.
- So by concentrating on motifs, we can find out the common conserved regions in the sequences and study the functional and evolutionary details or organisms.

Some of the common secondary databases include:

Prosite

- It was the first secondary database developed.
- Protein families usually contain some most conserved motifs which can be encoded to find out various biological functions.
- So by using such a database tool, we can easily find out the family of proteins when a new sequence is searched. This is the importance of PROSITE.
- Within PROSITE motifs are encoded as a regular expression (called patterns).
- Entries are deposited in PROSITE in two distant files. The first file gives the pattern and lists all matches of pattern, whereas the second one gives the details of family, description of the biological role, etc.
- The process used to derive patterns involves the construction of multiple alignment and manual inspection.
- So PROSITE contains documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them.
- A set of databases collects together patterns found in protein sequences rather than the complete sequences. PROSITE is one such pattern database.
- The protein motif and pattern are encoded as “regular expressions”.
- The information corresponding to each entry in PROSITE is of the two forms – the patterns and the related descriptive text.

Prints

- Most protein families are characterized by several conserved motifs.
- All of these motifs can be an aid in constructing the `signatures" of different families. This principle is highlighted in constructing PRINT database.
- Within PRINTS motifs are encoded as unweighted local alignments. So small initial multiple alignments are taken to identify conserved motifs.
- Then these regions are searched in the database to find out similarities.
- Results are analyzed to find out the sequences which matched all the motifs within the fingerprint.
- PROSITE and PRINTS are the only manually annotated secondary databases. The print is a diagnostic collection of protein fingerprints.
- In the PRINTS database, the protein sequence patterns are stored as "fingerprints". Afingerprint is a set of motifs or patterns rather than a single one.
- The information contained in the PRINT entry may be divided into three sections. In addition to entry name, accession number and number of motifs, the first section contains cross-links to other databases that have more information about the characterized family.
- The second section provides a table showing how many of the motifs that make up the fingerprint occurs in the how many of the sequences in that family.
- The last section of the entry contains the actual fingerprints that are stored as multiple aligned sets of sequences, the alignment is made without gaps. There is, therefore, one set of aligned sequences for each motif.

Blocks

- The limitations of the above two databases led to the formation of Block database.
- In this database, the motifs (here called Blocks) are created automatically by highlighting and detecting the most conserved regions of each family of proteins.
- Block databases are fully automated.
- Keyword and sequence searching are the two important features of this type of database.

- Blocks are ungapped Multiple Sequence Alignment representing conserved protein regions.

Pfam

- Pfam contains the profiles used using Hidden Markov models.
- HMMs build the model of the pattern as a series of the match, substitute, insert or delete states, with scores assigned for alignment to go from one state to another.
- Each family or pattern defined in the Pfam consists of the four elements. The first is the annotation, which has the information on the source to make the entry, the method used and some numbers that serve as figures of merit.
- The second is the seed alignment that is used to bootstrap the rest of the sequences into the multiple alignments and then the family.
- The third is the HMM profile.
- The fourth element is the complete alignment of all the sequences identified in that family.

3.7.1 SCOP

The SCOP (Structural Classification of Proteins) database maintained at the MRC Laboratory of Molecular Biology and Centre for Protein Engineering describes structural and evolutionary relationships between proteins of known structure (Murzin *et al.*, 1995). Because current automatic structure comparison tools cannot reliably identify all such relationships, SCOP has been constructed using a combination of manual inspection and automated methods. The task is complicated by the fact that protein structures show such variety, ranging from small, single domains to vast multi-domain assemblies. In some cases (e.g., some modular proteins), it may be meaningful to discuss a protein structure at the same time both at the multi-domain level and at the level of its individual domains.

SCOP classification

Proteins are classified in a hierarchical fashion to reflect their structural and evolutionary relatedness. Within the hierarchy there are many levels, but principally these describe the family, superfamily and fold. The boundaries between these levels may be subjective, but the higher levels generally reflect the clearest structural similarities.

- ***Family***: Proteins are clustered into families with clear evolutionary relationships if they have sequence identities ≥30%. But this is not an absolute measure—in some cases (e.g., the globins), it is possible to infer common descent from similar structures and functions in the absence of significant sequence identity (some members of the globin family share only 15% identity).

- ***Superfamily:*** Proteins are placed in superfamilies when, in spite of low sequence id... their structural and functional characteristics suggest a common evolutionary origi...
- ***Fold:*** Proteins are classed as having a common fold if they have the same ... secondary structures in the same arrangement and with the same topology, whether or ... they have a common evolutionary origin. In these cases, the structural similarities ... have arisen as a result of physical principles that favour particular packing arrange... and fold topologies.

SCOP is accessible for keyword interrogation via the MRC Laboratory Web server.

3.7.2 CATH

The CATH (Class, Architecture, Topology, Homology) database is a hierarchical do... classification of protein structures maintained at UCL (Orengo *et al.*, 1997). The reso... is largely derived using automatic methods, but manual inspection is necessary wh... automatic methods fail. Different categories within the classification are identified by m... of both unique numbers (by analogy with the **enzyme classification** or **E.C. system** ... enzymes) and descriptive names. Such a numbering scheme allows efficient computati... manipulation of the data. There are five levels within the hierarchy:

- ***Class*** is derived from gross secondary structure content and packing. Four classes of do... are recognised: (i) mainly-α, (ii) mainly-β, (iii) α−β, which includes both alternating ... and α+β structures, and (iv) those with low secondary structure content.
- ***Architecture*** describes the gross arrangement of secondary structures, ignoring ... connectivities; it is currently assigned manually using simple descriptions of ... secondary structure arrangements (e.g., barrel, roll, sandwich, etc.).
- ***Topology*** gives a description that encompasses both the overall shape and the connecti... of secondary structures. This is achieved by means of structure comparison algorith... that use empirically derived parameters to cluster the domains. Structures in which at l... 60% of the larger protein matches the smaller are assigned to the same topology leve...
- ***Homology*** groups domains that share ≥35% sequence identity and are thought to sh... a common ancestor, i.e. are homologous. Similarities are first identified by sequ... comparison and subsequently by means of a structure comparison algorithm.
- ***Sequence*** provides the final level within the hierarchy, whereby structures wi... homology groups are further clustered on the basis of sequence identity. At this le... domains have sequence identities >35% (with at least 60% of the larger dom... equivalent to the smaller), indicating highly similar structures and functions.

CATH is accessible for keyword interrogation via UCL's Biomolecular Structure ... Modelling Unit Web server.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF BIO AND CHEMICAL ENGINEERING
DEPARTMENT OF BIOTECHNOLOGY

Unit 3 – Bioinformatics – SBB3201

III. SEQUENCE ANALYSIS

Pairwise alignment

Pairwise sequence alignment methods are used to find the best-matching piecewise (local) or global alignments of two query sequences. Pairwise alignments can only be used between two sequences at a time, but they are efficient to calculate and are often used for methods that do not require extreme precision (such as searching a database for sequences with high similarity to a query). The three primary methods of producing pairwise alignments are dot- matrix methods, dynamic programming, and word methods.

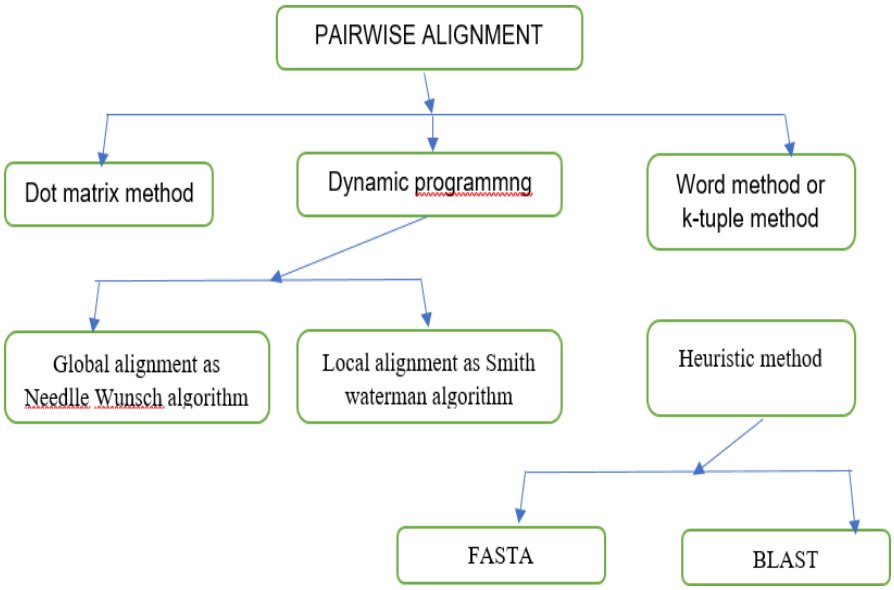


Figure 1

Dot plots

Dot plots are probably the oldest way of comparing sequences (Maizel and Lenk). A dot plot is a visual representation of the similarities between two sequences. Each axis of a rectangular array represents one of the two sequences to be compared. A window length is fixed, together with a criterion when two sequence windows are deemed to be similar. Whenever one window in one sequence resembles another a window in the other sequence, a dot or short diagonal is drawn at the corresponding position of the array. Thus, when two sequences share similarity over their entire length a diagonal line will extend from one corner of the dot plot to the diagonally opposite corner. If two sequences only share patches of similarity this will be revealed by diagonal stretches.

Dynamic programming

The technique of dynamic programming can be applied to produce global alignments via the Needleman-Wunsch algorithm, and local alignments via the Smith-Waterman algorithm. In typical usage, protein alignments 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. DNA and RNA alignments may use a scoring matrix, but in practice often simply assign a positive match score, a negative mismatch score, and a negative gap penalty. (In standard dynamic programming, the score of each amino acid position is independent of the identity of its neighbors, and therefore base stacking effects are not taken into account. However, it is possible to account for such effects by modifying the algorithm.) A common extension to standard linear gap costs, is the usage of two different gap penalties for opening a gap and for extending a gap. Typically the former is much larger than the latter, e.g. -10 for gap open and -2 for gap extension. Thus, the number of gaps in an alignment is usually reduced and residues and gaps are kept together, which typically makes more biological sense. The Gotoh algorithm implements affine gap costs by using three matrices. Needleman - Wunsch Algorithm

The Needleman–Wunsch algorithm is an algorithm used in bioinformatics to align protein or nucleotide sequences. It was one of the first applications of dynamic programming to compare biological sequences. The algorithm was developed by Saul B. Needleman and Christian D. Wunsch and published in 1970. The algorithm essentially divides a large problem (e.g. the full sequence) into a series of smaller problems and uses the solutions to the smaller problems to reconstruct a solution to the larger problem. It is also sometimes referred to as the optimal matching algorithm and the global alignment technique. The Needleman–Wunsch algorithm is still widely used for optimal global alignment, particularly when the quality of the global alignment is of the utmost importance.

The Smith–Waterman algorithm performs local sequence alignment; that is, for determining similar regions between two strings or nucleotide or protein sequences. Instead of looking at the total sequence, the Smith–Waterman algorithm compares segments of all possible lengths and optimizes the similarity measure.

The algorithm was first proposed by Temple F. Smith and Michael S. Waterman in 1981. Like the Needleman–Wunsch algorithm, of which it is a variation, Smith–Waterman is a dynamic programming algorithm.

As such, it has the desirable property that it is guaranteed to find the optimal local alignment with respect to the scoring system being used (which includes the substitution matrix and the gap-scoring scheme). The main difference to the Needleman–Wunsch algorithm is that negative scoring matrix cells are set to zero, which renders the (thus positively scoring) local alignments visible. Backtracking starts at the highest scoring matrix cell and proceeds until a cell with score zero is encountered, yielding the highest scoring local alignment. One does not actually implement the algorithm as described because improved alternatives are now available that have better scaling and are more accurate.

Word methods or k-tuple methods or heuristic methods

BLAST - Basic Local Alignment Search Tool

The BLAST algorithm was developed by Altschul, Gish, Miller, Myers and Lipman in 1990. The motivation for the development of BLAST was the need to increase the speed of FASTA by finding fewer and better hot spots during the algorithm. The idea was to integrate the substitution matrix in the first stage of finding the hot spots. The BLAST algorithm was developed for protein alignments in comparison to FASTA, which was developed for DNA sequences.

Different types of BLAST

blastn compares your query nucleotide sequence with database nucleotide sequences

blastp compares your query protein sequence with database of protein sequences that were derived form cDNA of interest blastx first translates your query sequence into amino acids in six reading frames (three forward and three back) then compares the protein sequences with protein databases

tblastn compares your query protein sequence with the database after translating each nucleotide sequence into protein using all six reading frames (This algorithm takes a long time, but is more likely to find distantly related sequences than the blastn, blastx, and blastp.)

tblastx translates both the query nucleotide sequence and the database sequences in all six reading frames and then compares the protein sequences (This, like tblastn, is very time consuming, but finds more results).

Algorithm

1. Remove low-complexity region or sequence repeats in the query sequence.
2. Make a k-letter word list of the query sequence.
3. List the possible matching words.
4. Organize the remaining high-scoring words into an efficient search tree.
5. Repeat step 3 to 4 for each k-letter word in the query sequence.
6. Scan the database sequences for exact matches with the remaining high-scoring words.
7. Extend the exact matches to high-scoring segment pair (HSP).
8. List all of the HSPs in the database whose score is high enough to be considered.
9. Evaluate the significance of the HSP score.
10. Make two or more HSP regions into a longer alignment.
11. Show the gapped Smith-Waterman local alignments of the query and each of the matched database sequences.
12. Report every match whose expect score is lower than a threshold parameter E.

In Bioinformatics, **BLAST** for **B**asic **L**ocal **A**lignment **S**earch **T**ool is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of proteins or the nucleotides of DNA sequences. A BLAST search enables a researcher to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold.

Different types of BLASTs are available according to the query sequences. For example, following the discovery of a previously unknown gene in the mouse, a scientist will typically perform a BLAST search of the human genome to see if humans carry a similar gene; BLAST will identify sequences in the human genome that resemble the mouse gene based on similarity of sequence. The BLAST algorithm and program were designed by Stephen Altschul, Warren Gish, Webb Miller, Eugene Myers, and David J. Lipman at the National Institutes of Health and was published in the Journal of Molecular Biology in 1990 and cited over 50,000 times.

Background

BLAST is one of the most widely used bioinformatics programs for sequence searching. It addresses a fundamental problem in bioinformatics research. The heuristic algorithm it uses is much faster than other approaches, such as calculating an optimal alignment. This emphasis on speed is vital to making the algorithm practical on the huge genome databases currently available, although subsequent algorithms can be even faster.

Before BLAST, FASTA was developed by David J. Lipman and William R. Pearson in 1985.

Before fast algorithms such as BLAST and FASTA were developed, doing database searches for protein or nucleic sequences was very time consuming because a full alignment procedure (e.g., the Smith–Waterman algorithm) was used.

While BLAST is faster than any Smith-Waterman implementation for most cases, it cannot "guarantee the optimal alignments of the query and database sequences" as Smith-Waterman algorithm does. The optimality of Smith-Waterman "ensured the best performance on accuracy and the most precise results" at the expense of time and computer power.
BLAST is more time-efficient than FASTA by searching only for the more significant patterns in the sequences, yet with comparative sensitivity. This could be further realized by understanding the algorithm of BLAST introduced below.

Examples of other questions that researchers use BLAST to answer are:

- Which bacterial species have a protein that is related in lineage to a certain protein with known amino-acid sequence
- What other genes encode proteins that exhibit structures or motifs such as ones that have just been determined

BLAST is also often used as part of other algorithms that require approximate sequence matching.

The BLAST algorithm and the computer program that implements it were developed by Stephen Altschul, Warren Gish, and David Lipman at the U.S. National Center for Biotechnology Information (NCBI), Webb Miller at the Pennsylvania State University, and Gene Myers at the University of Arizona. It is available on the web on the NCBI website. Alternative implementations include AB-BLAST (formerly known as WU-BLAST), FSA-BLAST (last updated in 2006), and ScalaBLAST.

Input

Input sequences (in FASTA or Genbank format) and weight matrix.

Output

BLAST output can be delivered in a variety of formats. These formats include HTML, plain text, and XML formatting. For NCBI's web-page, the default format for output is HTML. When performing a BLAST on NCBI, the results are given in a graphical format showing the hits found, a table showing sequence identifiers for the hits with scoring related data, as well as alignments for the sequence of interest and the hits received with corresponding BLAST scores for these. The easiest to read and most informative of these is probably the table.

If one is attempting to search for a proprietary sequence or simply one that is unavailable in databases available to the general public through sources such as NCBI, there is a BLAST program available for download to any computer, at no cost. This can be found at BLAST+ executables. There are also commercial programs available for purchase. Databases can be found from the NCBI site, as well as from Index of BLAST databases (FTP).

Process

Using a heuristic method, BLAST finds similar sequences, by locating short matches between the two sequences. This process of finding similar sequences is called seeding. It is after this first match that BLAST begins to make local alignments. While attempting to find similarity in sequences, sets of common letters, known as words, are very important. For example, suppose that the sequence contains the following stretch of letters, GLKFA. If a BLAST was being conducted under normal conditions, the word size would be 3 letters. In this case, using the given stretch of letters, the searched words would be GLK, LKF, KFA. The heuristic algorithm of BLAST locates all common three-letter words between the sequence of interest and the hit sequence or sequences from the database. This result will then be used to build an alignment. After making words for the sequence of interest, the rest of the words are also assembled. These words must satisfy a requirement of having a score of at least the threshold T, when compared by using a scoring matrix. One commonly used scoring matrix for BLAST searches is BLOSUM62, although the optimal scoring matrix depends on sequence similarity. Once both words and neighborhood words are assembled and compiled, they are compared to the sequences in the database in order to find matches. The threshold score T determines whether or not a particular word will be included in the alignment. Once seeding has been conducted, the alignment which is only 3 residues long, is extended in both directions by the algorithm used by BLAST. Each extension impacts the score of the alignment by either increasing or decreasing it. If this score is higher than a pre-determined T, the alignment will be included in the results given by BLAST. However, if this score is lower than this pre-determined T, the alignment will cease to extend, preventing the areas of poor alignment from being included in the BLAST results. Note that increasing the T score limits the amount of space available to search, decreasing the number of neighborhood words, while at the same time speeding up the process of BLAST.

Program

The BLAST program can either be downloaded and run as a command-line utility "blastall" or accessed for free over the web. The BLAST web server, hosted by the NCBI, allows anyone with a web browser to perform similarity searches against constantly updated databases of proteins and DNA that include most of the newly sequenced organisms.

The BLAST program is based on an open-source format, giving everyone access to it and enabling them to have the ability to change the program code. This has led to the creation of several BLAST "spin-offs".

There are now a handful of different BLAST programs available, which can be used depending on what one is attempting to do and what they are working with. These different programs vary in query sequence input, the database being searched, and what is being compared. These programs and their details are listed below:

BLAST is actually a family of programs (all included in the blastall executable). These include

Nucleotide-nucleotide BLAST (blastn)

This program, given a DNA query, returns the most similar DNA sequences from the DNA database that the user specifies.

Protein-protein BLAST (blastp)

This program, given a protein query, returns the most similar protein sequences from the protein database that the user specifies.

Position-Specific Iterative BLAST (PSI-BLAST) (blastpgp)

The program is used to find distant relatives of a protein. First, a list of all closely related proteins is created. These proteins are combined into a general "profile" sequence, which summarises significant features present in these sequences. A query against the protein database is then run using this profile, and a larger group of proteins is found. This larger group is used to construct another profile, and the process is repeated.

By including related proteins in the search, PSI-BLAST is much more sensitive in picking up distant evolutionary relationships than a standard protein-protein BLAST.

Nucleotide 6-frame translation-protein (blastx)

This program compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database.

Nucleotide 6-frame translation-nucleotide 6-frame translation (tblastx)

This program is the slowest of the BLAST family. It translates the query nucleotide sequence in all six possible frames and compares it against the six-frame translations of a nucleotide sequence database. The purpose of tblastx is to find very distant relationships between nucleotide sequences.

Protein-nucleotide 6-frame translation (tblastn)

This program compares a protein query against the all six reading frames of a nucleotide sequence database.

Large numbers of query sequences (megablast)

When comparing large numbers of input sequences via the command-line BLAST, "megablast" is much faster than running BLAST multiple times. It concatenates many input sequences together to form a large sequence before searching the BLAST database, then post-analyzes the search results to glean individual alignments and statistical values.

Of these programs, BLASTn and BLASTp are the most commonly used because they use direct comparisons, and do not require translations. However, since protein sequences are better conserved evolutionarily than nucleotide sequences, tBLASTn, tBLASTx, and BLASTx, produce more reliable and accurate results when dealing with coding DNA. They also enable one to be able to directly see the function of the protein sequence, since by translating the sequence of interest before searching often gives you annotated protein hits.

FASTA

FASTA is a DNA and protein sequence alignment software package first described (as FASTP) by David J. Lipman and William R. Pearson in 1985. Its legacy is the FASTA format which is now ubiquitous in bioinformatics.

The original FASTP program was designed for protein sequence similarity searching. FASTA added the ability to do DNA:DNA searches, translated protein:DNA searches, and also provided a more sophisticated shuffling program for evaluating statistical significance. There are several programs in this package that allow the alignment of protein sequences and DNA sequences.

FASTA is pronounced "fast A", and stands for "FAST-All", because it works with any alphabet, an extension of "FAST-P" (protein) and "FAST-N" (nucleotide) alignment.

The current FASTA package contains programs for protein:protein, DNA:DNA, protein:translated DNA (with frameshifts), and ordered or unordered peptide searches. Recent versions of the FASTA package include special translated search algorithms that correctly handle frameshift errors (which six-frame-translated searches do not handle very well) when comparing nucleotide to protein sequence data.

In addition to rapid heuristic search methods, the FASTA package provides SEARCH, an implementation of the optimal Smith-Waterman algorithm. A major focus of the package is the calculation of accurate similarity statistics, so that biologists can judge whether an alignment is likely to have occurred by chance, or whether it can be used to infer homology. The FASTA package is available from fasta.bioch.virginia.edu. The web-interface to submit sequences for running a search of the European Bioinformatics Institute (EBI)'s online databases is also available using the FASTA programs.

The FASTA file format used as input for this software is now largely used by other sequence database search tools (such as BLAST) and sequence alignment programs (Clustal, T-Coffee, etc.).

FASTA takes a given nucleotide or amino acid sequence and searches a corresponding sequence database by using local sequence alignment to find matches of similar database sequences.

The FASTA program follows a largely heuristic method which contributes to the high speed of its execution. It initially observes the pattern of word hits, word-to-word matches of a given length, and marks potential matches before performing a more time-consuming optimized search using a Smith-Waterman type of algorithm.

The size taken for a word, given by the parameter kmer, controls the sensitivity and speed of the program. Increasing the kmer value decreases number of background hits that are found. From the word hits that are returned the program looks for segments that contain a cluster of nearby hits. It then investigates these segments for a possible match.

There are some differences between fastn and fastp relating to the type of sequences used but both use four steps and calculate three scores to describe and format the sequence similarity results. These are:

Identify regions of highest density in each sequence comparison. Taking a kmer to equal 1 or 2.

In this step all or a group of the identities between two sequences are found using a look up table. The kmer value determines how many consecutive identities are required for a match to be declared. Thus the lesser the kmer value: the more sensitive the search. kmer=2 is frequently taken by users for protein sequences and kmer=4 or 6 for nucleotide sequences.

Short oligonucleotides are usually run with kmer= 1. The program then finds all similar **local regions**, represented as diagonals of a certain length in a dot plot, between the two sequences by counting kmer matches and penalizing for intervening mismatches. This way, **local regions** of highest density matches in a diagonal are isolated from background hits. For protein sequences BLOSUM50 values are used for scoring kmer matches. This ensures that groups of identities with high similarity scores contribute more to the local diagonal score than to identities with low similarity scores. Nucleotide sequences use the identity matrix for the same purpose. The best 10 local regions selected from all the diagonals put together are then saved.

Rescan the regions taken using the scoring matrices. trimming the ends of the region to include only those contributing to the highest score.

Rescan the 10 regions taken. This time use the relevant scoring matrix while rescoring to allow runs of identities shorter than the kmer value. Also while rescoring conservative replacements that contribute to the similarity score are taken. Though protein sequences use the BLOSUM50 matrix, scoring matrices based on the minimum number of base changes required for a specific replacement, on identities alone, or on an alternative measure of similarity such as PAM, can also be used with the program. For each of the diagonal regions rescanned this way, a subregion with the maximum score is identified. The initial scores found in step1 are used to rank the library sequences. The highest score is referred to as init1 score.

In an alignment if several initial regions with scores greater than a CUTOFF value are found, check whether the trimmed initial regions can be joined to form an approximate alignment with gaps. Calculate a similarity score that is the sum of the joined regions penalising for each gap 20 points. This initial similarity score (initn) is used to rank the library sequences. The score of the single best initial region found in step 2 is reported (init1). Here the program calculates an optimal alignment of initial regions as a combination of compatible regions with maximal score. This optimal alignment of initial regions can be rapidly calculated using a dynamic programming algorithm. The resulting score initn is used to rank the library sequences. This joining process increases sensitivity but decreases selectivity. A carefully calculated cut-off value is thus used to control where this step is implemented, a value that is approximately one standard deviation above the average score expected from unrelated sequences in the library. A 200-residue query sequence with kmer 2 uses a value 28.

This step uses a banded Smith-Waterman algorithm to create an optimised score (opt) for each alignment of query sequence to a database (library) sequence. It takes a band of 32 residues centered on the init1 region of step2 for calculating the optimal alignment. After all sequences are searched the program plots the initial scores of each database sequence in a histogram, and calculates the statistical significance of the "opt" score. For protein sequences, the final alignment is produced using a full Smith-Waterman alignment. For DNA sequences, a banded alignment is provided.

FASTA cannot remove low complexity regions before aligning the sequences as it is possible with BLAST. This might be problematic as when the query sequence contains such regions, e.g. mini- or microsatellites repeating the same short sequence frequent times, this increases the score of not familiar sequences in the database which only match in this repeats, which occur quite frequently. Therefore the program PRSS is added in the FASTA distribution package. PRSS shuffles the matching sequences in the database either on the one-letter level or it shuffles short segments which length the user can determine. The shuffled sequences are now aligned again and if the score is still higher than expected this is caused by the low complexity regions being mixed up still mapping to the query. By the amount of the score the shuffled sequences still attain PRSS now can predict the significance of the score of the original sequences. The higher the score of the shuffled sequences the less significant the matches found between original database and query sequence.

The FASTA programs find regions of local or global similarity between Protein or DNA sequences, either by searching Protein or DNA databases, or by identifying local duplications within a sequence. Other programs provide information on the statistical significance of an alignment. Like BLAST, FASTA can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

Multiple Sequence Alignment (MSA)

A multiple sequence alignment (MSA) is a sequence alignment of three or more biological sequences, generally protein, DNA, or RNA. In many cases, the input set of query sequences are assumed to have an evolutionary relationship by which they share a lineage and are descended from a common ancestor. From the resulting MSA, sequence homology can be inferred and phylogenetic analysis can be conducted to assess the sequences' shared evolutionary origins. Visual depictions of the alignment as in the image at right illustrate mutation events such as point mutations (single amino acid or nucleotide changes) that appear as differing characters in a single alignment column, and insertion or deletion mutations (indels or gaps) that appear as hyphens in one or more of the sequences in the alignment. Multiple sequence alignment is often used to assess sequence conservation of protein domains, tertiary and secondary structures, and even individual amino acids or nucleotides.

Multiple sequence alignment also refers to the process of aligning such a sequence set. Because three or more sequences of biologically relevant length can be difficult and are

almost always time-consuming to align by hand, computational algorithms are used to produce and analyze the alignments. MSAs require more sophisticated methodologies than pairwise alignment because they are more computationally complex. Most multiple sequence alignment programs use heuristic methods rather than global optimization because identifying the optimal alignment between more than a few sequences of moderate length is prohibitively computationally expensive.

1 Dynamic programming - T-Coffee

2 Progressive alignment construction- hill-climbing algorithm, ClustalW

3 Iterative methods - MUSCLE (multiple sequence alignment by log-expectation)

4 Hidden Markov models- HMMER

6 Genetic algorithms and simulated annealing

7 Phylogeny methods

8. Motif finding

ClustalW

Clustal is a series of widely used computer programs used in Bioinformatics for multiple sequence alignment. ClustalW software algorithm is used for global alignments.

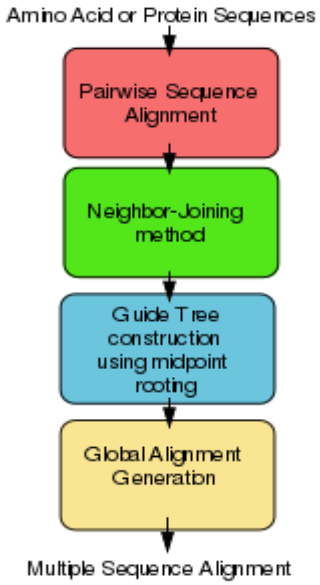


Fig. 2

ClustalW like the other Clustal tools is used for aligning multiple nucleotide or protein sequences in an efficient manner. It uses progressive alignment methods, which align the most similar sequences first and work their way down to the least similar sequences until a global alignment is created. ClustalW is a matrix-based algorithm, whereas tools like T-Coffee and Dialign are consistency-based. ClustalW has a fairly efficient algorithm that competes well against other software. This program requires three or more sequences in order to calculate a global alignment, for pairwise sequence alignment (2 sequences) use tools similar to EMBOSS, LALIGN.

Algorithm

ClustalW uses progressive alignment methods. In these, the sequences with the best alignment score are aligned first, then progressively more distant groups of sequences are aligned. This heuristic approach is necessary due to the time and memory demand of finding the global optimal solution. The first step to the algorithm is computing a rough distance matrix between each pair of sequences, also known as pairwise sequence alignment. The next step is a neighbor-joining method that uses midpoint rooting to create an overall guide

tree. The guide tree is then used as a rough template to generate a global alignment.

Accuracy and Results

The algorithm ClustalW uses provides a close-to-optimal result almost every time. However, it does exceptionally well when the data set contains sequences with varied degrees of divergence. This is because in a data set like this, the guide tree becomes less sensitive to noise. ClustalW was one of the first algorithms to combine pairwise alignment and global alignment in an attempt to be speed efficient, and it worked, but there is a loss in accuracy that other software doesn't have due to this.

ClustalW, when compared to other MSA algorithms, performed as one of the quickest while still maintaining a level of accuracy. There is still much to be improved compared to its consistency-based competitors like T-Coffee. The accuracy for ClustalW when tested against MAFFT, T-Coffee, Clustal Omega, and other MSA implementations had the lowest accuracy for full-length sequences. It had the least RAM memory demanding algorithm out of all the ones tested in the study. While ClustalW recorded the lowest level of accuracy among its competitors, it still maintained what some would deem acceptable. There have been updates and improvements to the algorithm that are present in ClustalW2 that work to increase accuracy while still maintaining its greatly valued speed.

PHYLOGENETIC ANALYSIS

How to construct a Phylogenetic tree?

- A phylogenetic tree is a visual representation of the relationship between different organisms, showing the path through evolutionary time from a common ancestor to different descendants.
- Similarities and divergence among related biological sequences revealed by sequence alignment often have to be rationalized and visualized in the context of phylogenetic trees. Thus, molecular phylogenetics is a fundamental aspect of bioinformatics.
- Molecular phylogenetics is the branch of phylogeny that analyzes genetic, hereditary molecular differences, predominately in DNA sequences, to gain information on an organism’s evolutionary relationships.
- The similarity of biological functions and molecular mechanisms in living organisms strongly suggests that species descended from a common ancestor. Molecular phylogenetics uses the structure and function of molecules and how they change over time to infer these evolutionary relationships.

- From these analyses, it is possible to determine the processes by which diversity among species has been achieved. The result of a molecular phylogenetic analysis is expressed in a phylogenetic tree.

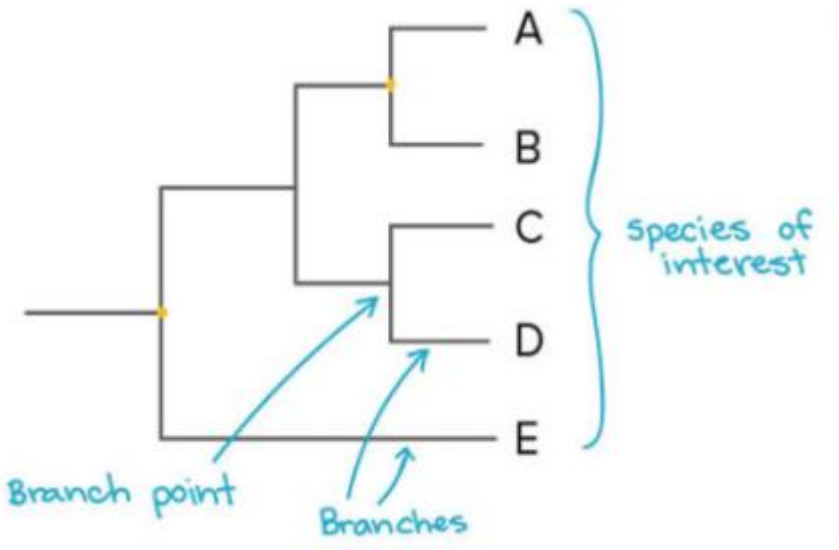


Figure 3

Phylogenetic Analysis and the Role of Bioinformatics

Molecular data that are in the form of DNA or protein sequences can also provide very useful evolutionary perspectives of existing organisms because, as organisms evolve, the genetic materials accumulate mutations over time causing phenotypic changes. Because genes are the medium for recording the accumulated mutations, they can serve as molecular fossils. Through comparative analysis of the molecular fossils from a number of related organisms, the evolutionary history of the genes and even the organisms can be revealed.

However, phylogeny inference are notoriously difficult endeavours because the number of solutions increases explosively with the number of taxa and the tremendous number of new questions in evolutionary biology that could be investigated through the use of larger taxon samplings.

But with the development and use of computational and an array of bioinformatics tools, the ability to analyze large data sets in practical computing times, and yielding an optimal or near-optimal solutions with high probability are being possible. In response to this trend, much of the current research in phyloinformatics (i.e., computational phylogenetics) concentrates on the development of more efficient heuristic approaches.

Steps in Phylogenetic Analysis

The basic steps in any phylogenetic analysis include:

1. **Assemble and align a dataset**

- The first step is to identify a protein or DNA sequence of interest and assemble a dataset consisting of other related sequences.
- DNA sequences of interest can be retrieved using NCBI BLAST or similar search tools.
- Once sequences are selected and retrieved, multiple sequence alignment is created.
- This involves arranging a set of sequences in a matrix to identify regions of homology.
- There are many websites and software programs, such as ClustalW, MSA, MAFFT, and T-Coffee, designed to perform multiple sequence on a given set of molecular data.

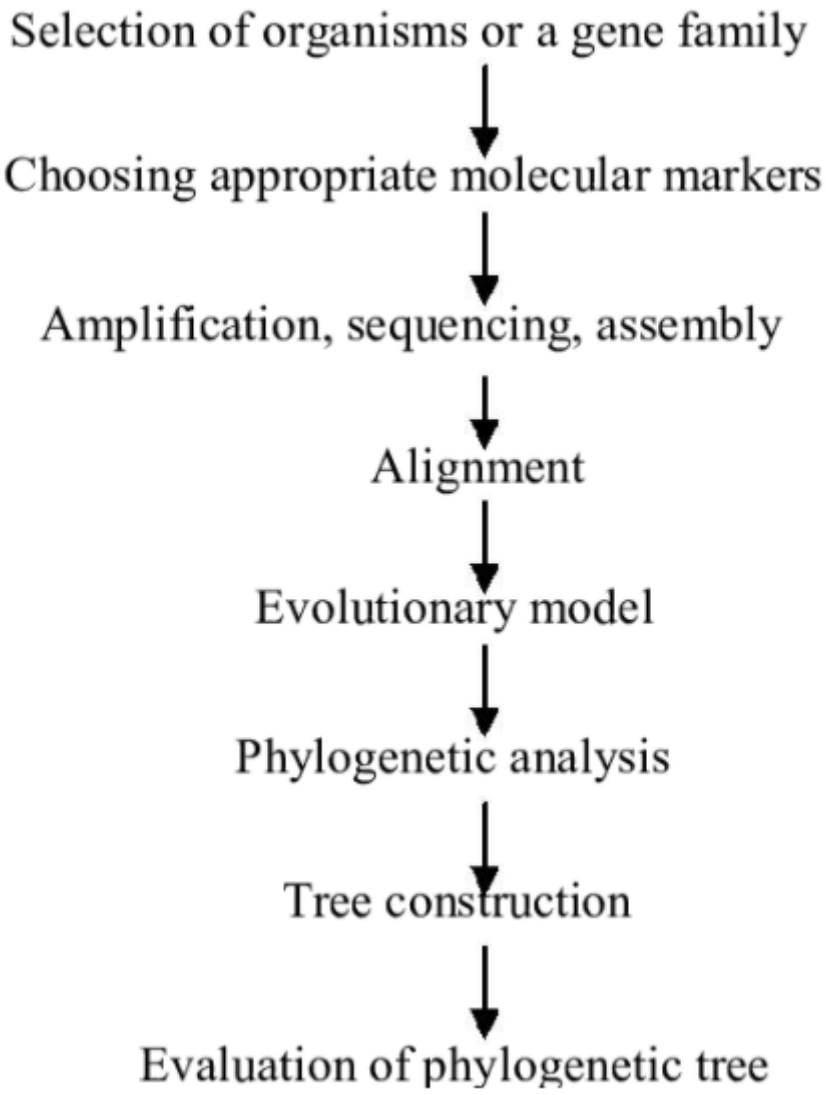


Figure 4

2. **Build (estimate) phylogenetic trees** from sequences using computational methods and stochastic models

- To build phylogenetic trees, statistical methods are applied to determine the tree topology and calculate the branch lengths that best describe the phylogenetic relationships of the aligned sequences in a dataset.
- The most common computational methods applied include distance-matrix methods, and discrete data methods, such as maximum parsimony and maximum likelihood.
- There are several software packages, such as Paup, PAML, PHYLIP, that apply these most popular methods.

3. **Statistically test and assess the estimated trees.**

- Tree estimating algorithms generate one or more optimal trees.
- This set of possible trees is subjected to a series of statistical tests to evaluate whether one tree is better than another – and if the proposed phylogeny is reasonable.
- Common methods for assessing trees include the Bootstrap and Jackknife Resampling methods, and analytical methods, such as parsimony, distance, and likelihood.

Bioinformatics Tools for Phylogenetic Analysis

- There are several bioinformatics tools and databases that can be used for phylogenetic analysis.
- These include **PANTHER, P-Pod, PFam, TreeFam, and the PhyloFacts structural phylogenomic encyclopedia**.
- Each of these databases uses different algorithms and draws on different sources for sequence information, and therefore the trees estimated by PANTHER, for example, may differ significantly from those generated by P-Pod or PFam.
- As with all bioinformatics tools of this type, it is important to test different methods, compare the results, then determine which database works best (according to consensus results) for studies involving different types of datasets.

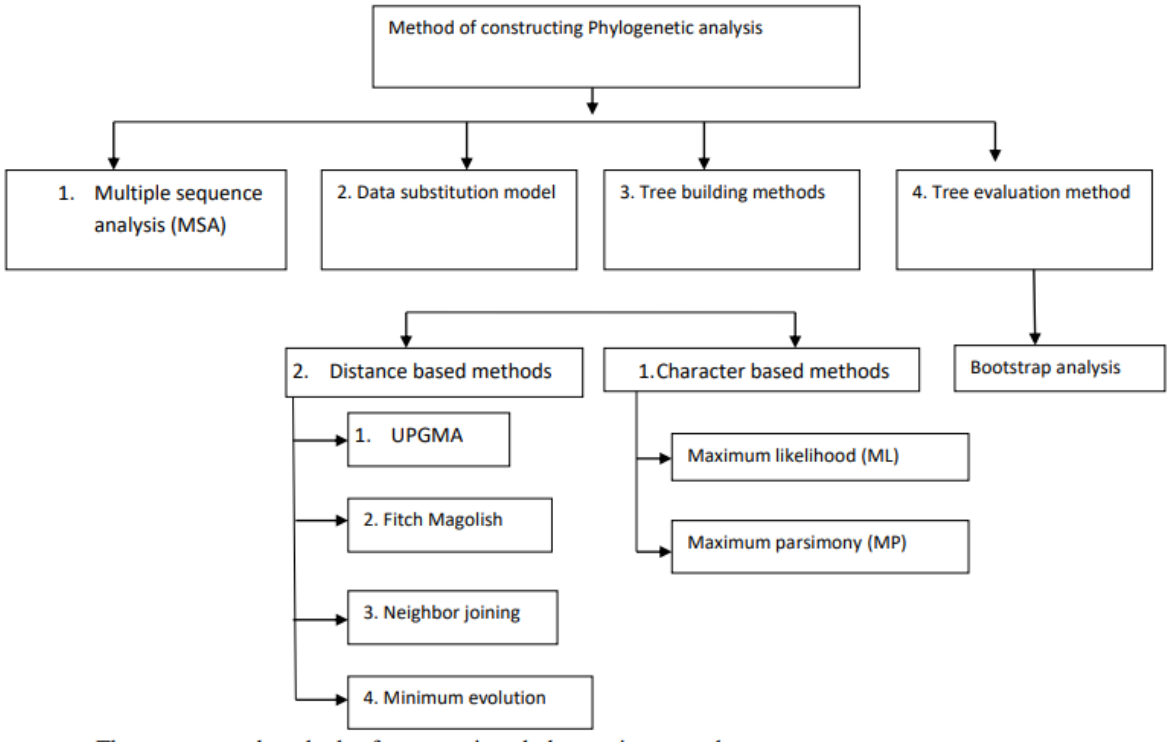


Figure 5

There are several methods of constructing phylogenetic trees - the most common are:

- distance methods
- character based methods

All these methods can only provide estimates of what a phylogenetic tree might look like for a given set of data. Most good methods also provide an indication of how much variation there is in these estimates. Distance methods: Preferred for work with immunological data, frequency data, or data with some impreciseness in its methods. Very rapid, and easily permits statistical tests e.g. bootstrapping. Derives some measure of similarity or difference between the input sequences.

UPGMA Cluster algorithm. Links least different pairs of seqs, sequentially (so that when one pair is formed, they become a single entity). (Invalid) assumptions made: 1. Rate of change equal among all sequences. 2. Branch lengths correlate with the expected phenotypic distance between sequences, which corresponds to a proportional measure of time. o NJ Corrects several assumptions made in the UPGMA method. Yields an unrooted tree. o Fitch and Margoliash Does not try to find pairs of least different sequences, but tries to find trees that

fulfil an optimum criterion. Yields an unrooted tree. Character based methods Popular for reconstructing ancestral relationships. o Maximum parsimony: Evaluates all possible trees. Infers the number of evolutionary events implied by a particular topology. The most likely tree is then one that requires the minimum number of evolutionary changes needed to explain the observed data. Problems: Most parsimonious tree may not be unique; difficult to make valid statistical statements if there are many steps in a tree; branches with particularly rapid rates of change tend to attract one another, especially when the sequence lengths are small. o Maximum likelihood: Very slow. Preferred when homoplasies (convergences of a particular character at a site) are expected to be concentrated in a few sites only, whose identities are known in advance. The method works by estimating, for all nucleotide positions in a sequence, what the probability of having a particular nucleotide at a particular site is, based on whether or not its ancestors had it (and the transition/transversion ratio). These probabilities are summed over the whole sequence, for both branches of a bifurcating tree. The product of the two probabilities gives you the likelihood of the tree up to this point. With more sequences, the estimation is done recursively at every branch point. Since each site evolves independently, the likelihood of the phylogeny can be estimated at every site. This process can only be done in a reasonable amount of time with four sequences. If there are more than four sequences, basic trees can be made for sets of four sequences, and then extra sequences added to the tree and the process of finding the maximum likelihood re-estimated. The order in which the sequences are added and the initial sequences chosen to start the process critically influences the resulting tree. To prevent any bias, the whole process is done multiple times with random choices for the order of the sequences. A majority rule consensus tree is then chosen as the final tree. To create a phylogenetic tree, you must first have an alignment. This can be created using ClustalW. ClustalW can also create a tree file for you (if you choose 'nj', 'phylip', or 'dist' from the "Tree type" pull-down menu.) However, you have more control over the tree if you simply choose to create an alignment in ClustalW (do not choose a tree type in this case, because then the alignment itself will not be presented). Copy the alignment (including the title, so that the PHYLIP programs recognise the alignment format as ClustalW), and paste it into the text-entry box provided for alignments in one of the following programs in the PHYLIP suite of programs. PHYLIP will convert the format of your alignment to Phylip format automatically. However, occasionally, especially in cases where the alignment is very large, this automatic conversion may cause errors. You can also convert the alignment yourself using SQUIZZ.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF BIO AND CHEMICAL ENGINEERING
DEPARTMENT OF BIOTECHNOLOGY

Unit 4 – Bioinformatics – SBB3201

IV. Protein Analysis

The spatial arrangement of atoms in a protein is called a conformation. The term conformation refers to a structural state that can, without breaking any covalent bonds, interconvert with other structural states. A change in conformation could occur, for example, by rotation about single bonds. Of the innumerable conformations that are theoretically possible in a protein containing hundreds of single bonds, one generally predominates. This is usually the conformation that is thermodynamically the most stable, having the lowest Gibbs' free energy (G). Proteins in their functional conformation are called native proteins.

There Are Four Levels of Architecture in Proteins

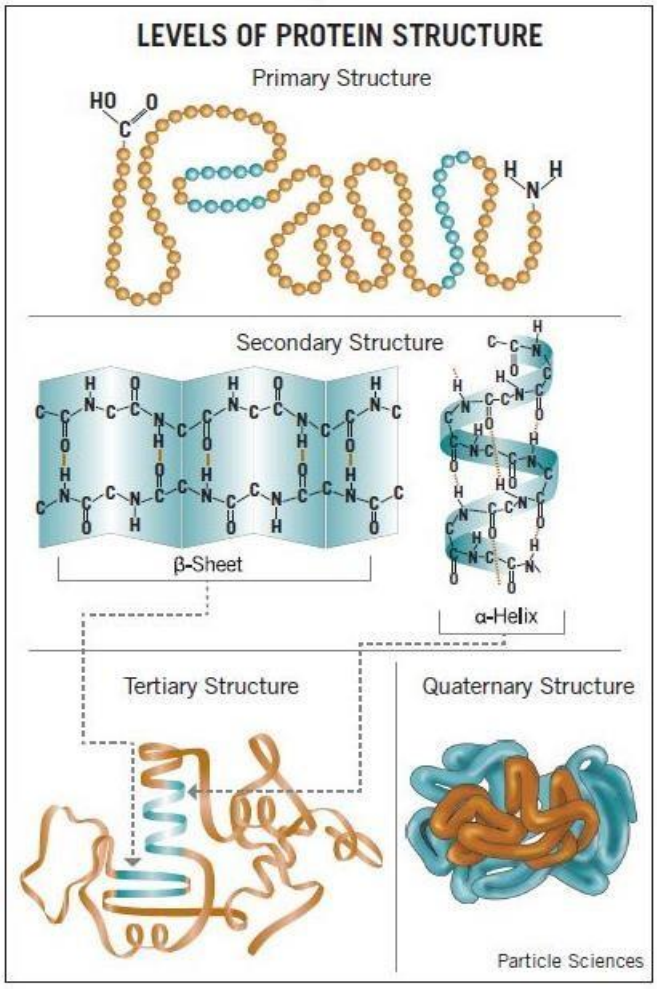


Figure 1 Levels of structure in proteins

Conceptually, protein structure can be considered at four levels (Fig. 1). **Primary structure** includes all the covalent bonds between amino acids and is normally defined by the sequence of peptide-bonded amino acids and locations of disulfide bonds. The relative spatial arrangement of the linked amino acids is unspecified. Polypeptide chains are not free to take up any three-dimensional structure at random. Steric constraints and many weak interactions stipulate that some arrangements will be more stable than others.

Secondary structure refers to regular, recurring arrangements in space of adjacent amino acid residues in a polypeptide chain. There are a few common types of secondary structure, the most prominent being the a helix and the β conformation.

Tertiary structure refers to the spatial relationship among all amino acids in a polypeptide; it is the complete three-dimensional structure of the polypeptide. The boundary between secondary and tertiary structure is not always clear. Several different types of secondary structure are often found within the three-dimensional structure of a large protein. Proteins with several polypeptide chains have one more level of structure: **quaternary structure**, which refers to the spatial relationship of the polypeptides, or subunits, within the protein.

Protein Secondary Structure

Several types of secondary structure are particularly stable and occur widely in proteins. The most prominent are the α helix and β conformations. Using fundamental chemical principles and a few experimental observations, Linus Pauling and Robert Corey predicted the existence of these secondary structures in 1951, several years before the first complete protein structure was elucidated.

In considering secondary structure, it is useful to classify proteins into two major groups: fibrous proteins, having polypeptide chains arranged in long strands or sheets, and globular proteins, with polypeptide chains folded into a spherical or globular shape. Fibrous proteins play important structural roles in the anatomy and physiology of vertebrates, providing external protection, support, shape, and form. They may constitute one-half or more of the total body protein in larger animals. Most enzymes and peptide hormones are globular proteins. Globular proteins tend to be structurally complex, often containing several types of secondary structure; fibrous proteins usually consist largely of a single type of secondary structure. Because of this structural simplicity, certain

fibrous proteins played a key role in the development of the modern understanding of protein structure and provide particularly clear examples of the relationship between structure and function; they are considered in some detail after the general discussion of secondary structure.

The Peptide Bond is Rigid and Planar

In the peptide bond, the π-electrons from the carbonyl are delocalized between the oxygen and the nitrogen. This means that the peptide bond has ~40% double bond character. This partial double bond character is evident in the shortened bond length of the C–N bond. The length of a normal C–N single bond is 1.45 Å and a C=N double bond is 1.25 Å, while the peptide C–N bond length is 1.33 Å.

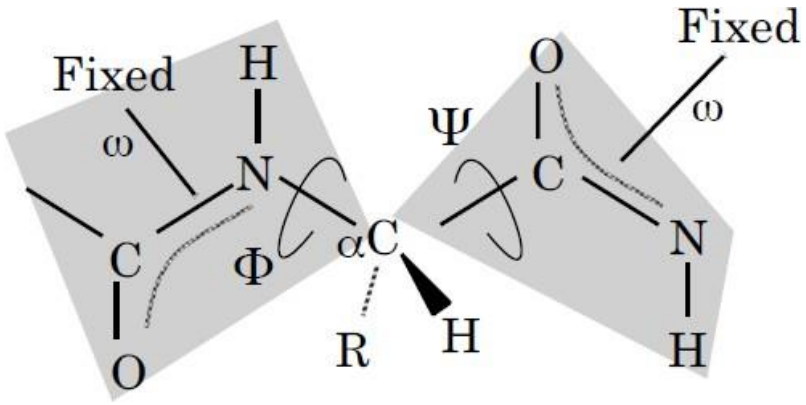


Figure 2 Peptide Bond – Dihedral angles

Because of its partial double bond character, rotation around the N–C bond is severely restricted. The peptide bond allows rotation about the bonds from the α- carbon, but not the amide C–N bond. Only the Φ and Ψ torsion angles can vary reasonably freely. In addition, the six atoms in the peptide bond (the two α-carbons, the amide O, and the amide N and H) are coplanar. Finally, the peptide bond has a dipole, with the O having a partial negative charge, and the N amide having a partial positive charge.

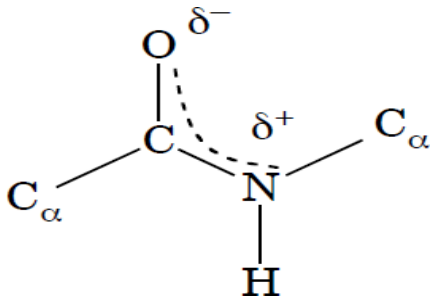


Figure 3 Peptide Bond – Dipole

This allows the peptide bond to participate in electrostatic interactions, and contributes to the hydrogen bond strength between the backbone carbonyl and the Namide proton.

Allowed values for Φ and Ψ are graphically revealed when Ψ is plotted versus Φ in a **Ramachandran plot**, introduced by G. N. Ramachandran.

The Ramachandran Plot

In a polypeptide the main chain N-Calpha and Calpha-C bonds relatively are free to rotate. These rotations are represented by the torsion angles phi and psi, respectively.

G N Ramachandran used computer models of small polypeptides to systematically vary phi and psi with the objective of finding stable conformations. For each conformation, the structure was examined for close contacts between atoms. Atoms were treated as hard spheres with dimensions corresponding to their van der Waals radii. Therefore, phi and psi angles which cause spheres to collide correspond to sterically disallowed conformations of the polypeptide backbone.

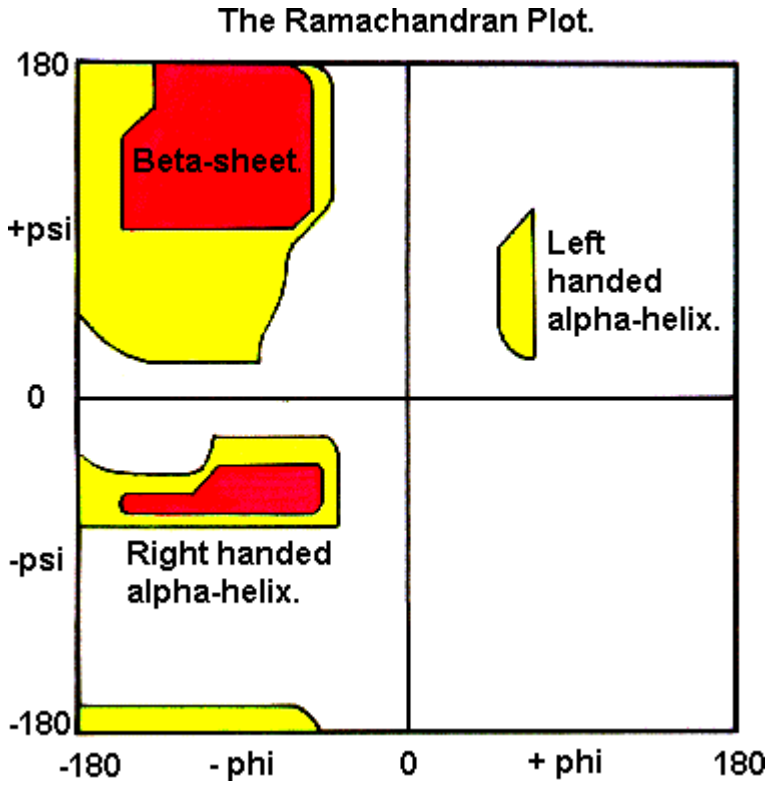


Figure 4 Ramachandran Plot

In the diagram above the white areas correspond to conformations where atoms in the polypeptide come closer than the sum of their van der Waals radii. These regions are sterically disallowed for all amino acids except glycine which is unique in that it lacks a side chain. The red regions correspond to conformations where there are no steric clashes, ie these are the allowed regions namely the alpha-helical and beta-sheet conformations. The yellow areas show the allowed regions if slightly shorter van der Waals radi are used in the calculation, ie the atoms are allowed to come a little closer together. This brings out an additional region which corresponds to the left-handed alpha-helix.

L-amino acids cannot form extended regions of left-handed helix but occasionally individual residues adopt this conformation. These residues are usually glycine but can also be asparagine or aspartate where the side chain forms a hydrogen bond with the main chain and therefore stabilizes this otherwise unfavorable conformation. The 3(10) helix occurs close to the upper right of the alpha-helical region and is on the edge of allowed region indicating lower stability.

Disallowed regions generally involve steric hindrance between the side chain C-beta methylene group and main chain atoms. Glycine has no side chain and therefore can adopt phi and psi angles in all four quadrants of the Ramachandran plot. Hence it frequently occurs in turn regions of proteins where any other residue would be sterically hindered.

Secondary structure

The term secondary structure refers to the local conformation of some part of a polypeptide. The discussion of secondary structure most usefully focuses on common regular folding patterns of the polypeptide backbone. A few types of secondary structure are particularly stable and occur widely in proteins. The most prominent are the α-helix and β-sheet. Using fundamental chemical principles and a few experimental observations, Pauling and Corey predicted the existence of these secondary structures in 1951, several years before the first complete protein structure was elucidated.

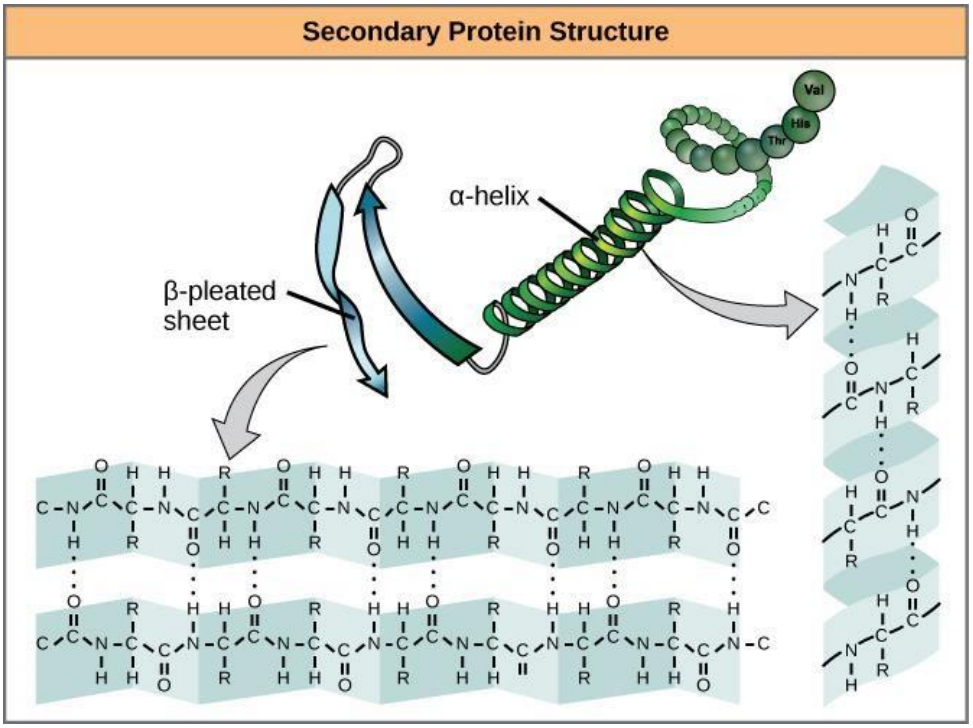


Figure 5 Secondary protein structures

Alpha helix (**α-helix**)

The **alpha helix** (**α-helix**) is a common secondary structure of proteins and is a right hand- coiled or spiral conformation (helix) in which every backbone N-H group donates a hydrogen bond to the backbone C=O group of the amino acid four residues earlier ($i+4 \rightarrow i$ hydrogen bonding). This secondary structure is also sometimes called a classic **Pauling–Corey–Branson alpha helix** (see below). The name **3.613-helix** is also used for this type of helix, denoting the number of residues per helical turn, and 13 atoms being involved in the ring formed by the hydrogen bond. Among types of local structure in proteins, the α- helix is the most regular and the most predictable from sequence, as well as the most prevalent.

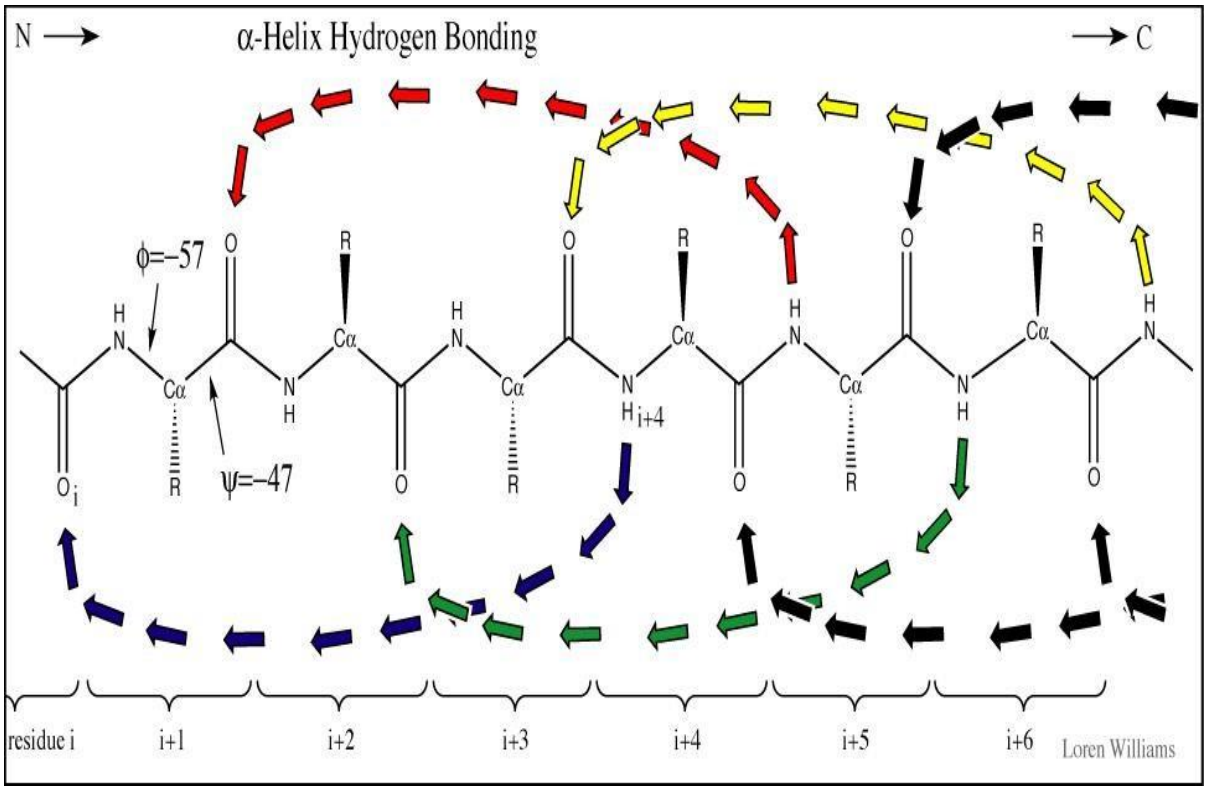


Figure 6 α Helix H-Bonding

PROPERTIES

The amino acids in an α helix are arranged in a right-handed helical structure where each amino acid residue corresponds to a 100° turn in the helix (i.e., the helix has 3.6 residues per turn), and a translation of 1.5 Å (0.15 nm) along the helical axis.

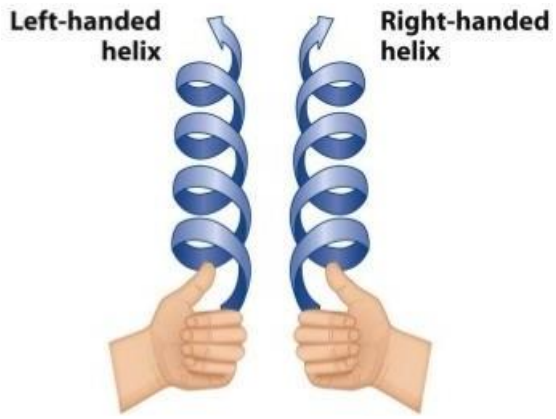


Figure 7 α Helix – Left handed & Right handed

Short pieces of left-handed helix sometimes occur with a large content of achiral glycine amino acids, but are unfavorable for the other normal, biological L-amino acids.

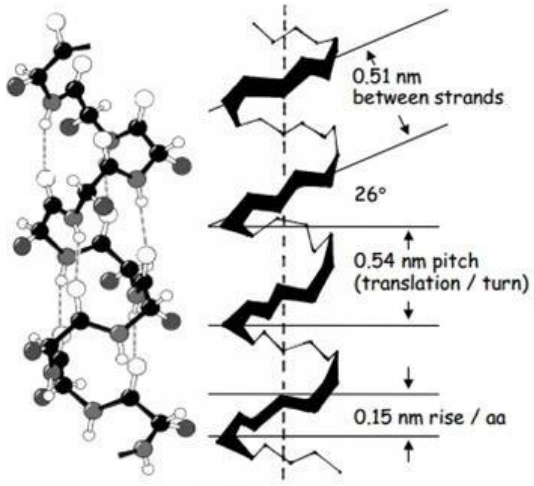


Figure 8 α Helix Pitch

The pitch of the alpha-helix (the vertical distance between consecutive turns of the helix) is

5.4 Å (0.54 nm), which is the product of 1.5 and 3.6. What is most important is that the N- H group of an amino acid forms a hydrogen bond with the C=O group of the amino acid *four* residues earlier; this repeated $i + 4 \rightarrow i$ hydrogen bonding is the

most prominent characteristic of an α-helix.

Similar structures include the 3_{10} helix ($i+3 \rightarrow i$ hydrogen bonding) and the π- helix (i+5→i hydrogen bonding). The α helix can be described as a 3.6_{13} helix, since the i + 4 spacing adds 3 more atoms to the H-bonded loop compared to the tighter 3_{10} helix, and on average, 3.6 amino acids are involved in one ring of α helix. The subscripts refer to the number of atoms (including the hydrogen) in the closed loop formed by the hydrogen bond.

Residues in α-helices typically adopt backbone (φ, ψ) dihedral angles around (-60°, -45°), as shown in the image at right. In more general terms, they adopt dihedral angles such that the ψ dihedral angle of one residue and the φ dihedral angle of the *next* residue sum to roughly - 105°. As a consequence, α-helical dihedral angles, in general, fall on a diagonal stripe on the Ramachandran diagram (of slope -1), ranging from (-90°, -15°) to (-35°, -70°). For comparison, the sum of the dihedral angles for a 3_{10} helix is roughly -75°, whereas that for the π-helix is roughly -130°.

Table 1- Structural features of the three major forms of protein helices

Geometry attribute	**α-helix**	**310 helix**	**π-helix**
Residues per turn	3.6	3.0	4.4
Translation per residue	1.5 Å (0.15 nm)	2.0 Å (0.20 nm)	1.1 Å (0.11 nm)
Radius of helix	2.3 Å (0.23 nm)	1.9 Å (0.19 nm)	2.8 Å (0.28 nm)
Pitch	5.4 Å (0.54 nm)	6.0 Å (0.60 nm)	4.8 Å (0.48 nm)

Pauling and Corey predicted a second type of repetitive structure, the β conformation -an **extended** state for which angles phi = -135^{o} and psi = $+135^{o}$; the polypeptide chain **alternates** in direction, resulting in a zig-zag structure for the peptide chain. Note the shaded circle around R; the extended strand arrangement also allows the **maximum space and freedom of movement for a side chain**. The repeat between identically oriented R-groups is 7.0 Å, with 3.5 Å per amino acid, matching the fiber diffraction data for beta-keratins.

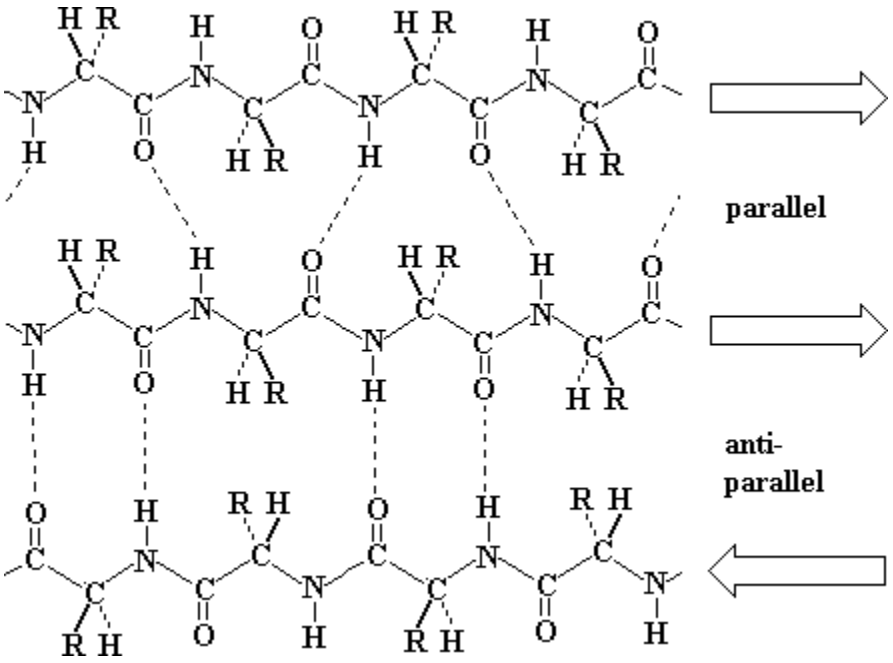


Figure 9 Parallel and Antiparallel beta sheets

Pauling's extended state model matched the spacing of fibroin exactly (3.5 and 7.0 Å). In the extended state, H-bonding NH and CO groups point out at 90^{o} to the strand. If extended strands are lined up side by side, H-bonds bridge from strand to strand. Identical or opposed strand alignments make up parallel or antiparallel beta sheets (named for beta keratin). Antiparallel beta-sheet is significantly more stable due to the well aligned H-bonds.

Table 2 – Dihedral angles in beta sheets

Angle	Antiparallel	Parallel
Φ	–139°	–119°
Ψ	135°	113°

The amino acids have side chains which **disrupt secondary structure**, and are known as **secondary structure breakers**:

side chain H is too small to protect backbone H-bond: **Gly**

side chain linked to alpha N, **has no N-H** to H-bond; **Pro**

rigid structure due to ring restricts to phi = -60^{o};

H-bonding side chains compete directly with **Asp, Asn, Ser**

backbone H-bonds

Clusters of breakers give rise to regions known as **loops or turns** which mark the boundaries of regular secondary structure, and serve to link up secondary structure segments.

Protein Tertiary Structure

Tertiary structure refers to the three-dimensional arrangement of all atoms in a protein. Tertiary structure is formed by the folding in three dimensions of the secondary structure elements of a protein. While the α helical secondary structure is held together by interactions between the carbonyl and amide groups within the backbone, tertiary structure is held together by interactions between R-groups of residues brought together by folding. Disulfide bonds are also counted under the category of tertiary structure interactions. Proteins that are compact are known as globular proteins.

Examination of protein structures resolved by X-ray diffraction and NMR has revealed a variety of folding patterns common to many different proteins. However, even within these folds, distinct substructures or structural **motifs**, i.e. distinctive arrangements of elements of secondary structure, have been described. The term **supersecondary structure** has been coined to describe this level of organisation, which is intermediate between secondary and tertiary.

Motifs or folds, are particularly stable arrangements of several elements of the secondary structure. Supersecondary structures are usually produced by packing side chains from adjacent secondary structural elements close to each other.

Rules for secondary structure

• Hydrophobic side groups must be buried inside the folds, therefore, layers must be created (β−α−β; α− α).

• α-helix and β-sheet, if occur together, are found in different structural layers.

• Adjacent polypeptide segments are stacked together.

• Connections between secondary structures do not form knots.

• The β-sheet is the most stable.

Motif

• Secondary structure composition, e.g. all α, all β, segregated α+β, mixed α/β

• Motif = small, specific combinations of secondary structure elements, e.g. β-α-β loop

Helix super secondary structures

Helix-Turn-Helix Motif

Also called the alpha-alpha type (αα-type). The motif is comprised of two antiparallel helices connected by a turn. The helix-turn-helix is a functional motif and is usually identified in proteins that bind to DNA minor and major grooves, and Calcium-binding proteins.

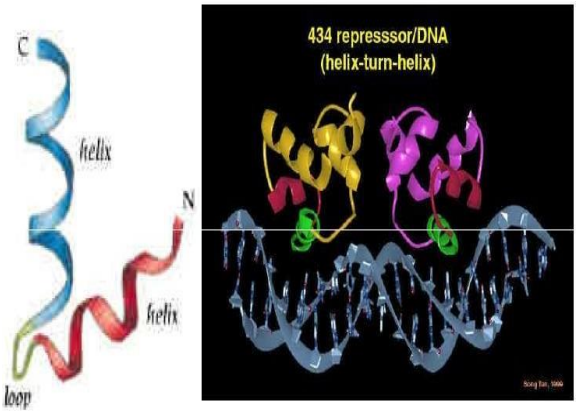


Figure 10 DNA binding Helix-turn-Helix motif

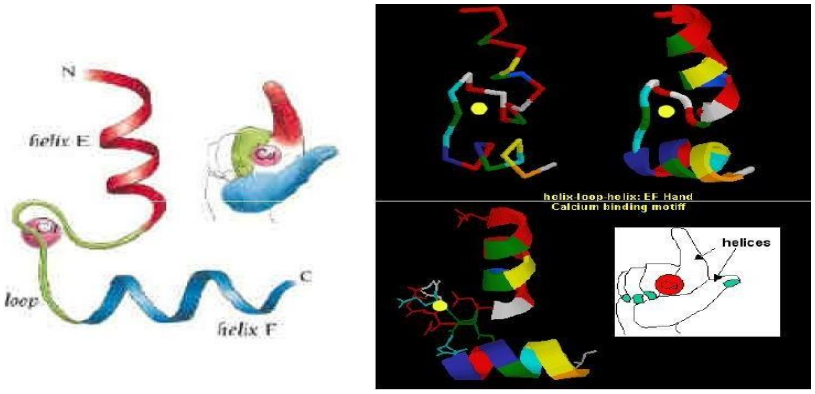


Figure 11 Calcium binding (EF Hand- Calcium binding) motif

Helix-hairpin-helix: Involved in DNA binding

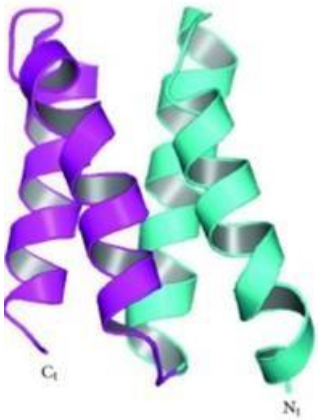


Figure 12 Helix-hairpin-helix

Alpha-alpha corner

Short loop regions connecting helices which are roughly perpendicular to one another

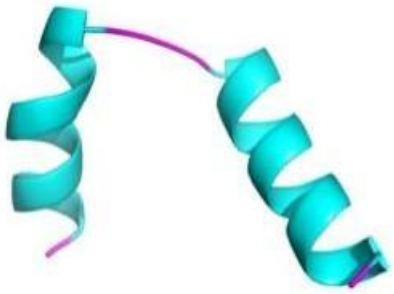


Figure 13 Alpha-alpha corner

Sheet super secondary structures

All beta tertiary structural domains can occur in proteins with one domain (eg. concanavalin A, superoxide dismutase), and occurs at least once in proteins with two domains (eg. chymotrypsin), or three domains (eg. OmpF).

The beta strands making up these domains are all essentially antiparallel and form structures to achieve stable packing arrangements within the protein.

There are presently (as of version 1.39) about 70 subclasses listed in SCOP for this

domain, and some examples of these are outlined below.

Beta barrels

This is the most abundant beta-domain structure and as the name suggests the domain forms a 'barrel-like' structure. The beta barrels are not geometrically perfect and can be rather distorted.

There are three main types:

1. Up-and-down barrels
2. Greek key barrels
3. Jelly roll (Swiss roll) barrels

Up-and-down beta-sheets or beta-barrels

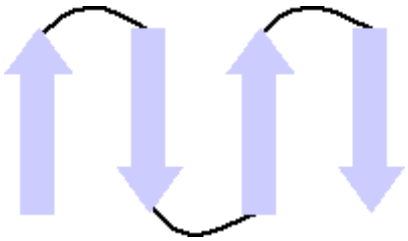


Figure 14 Beta barrels

The simple topology of an up-and-down barrel (named because the beta strands follow each other in sequence in an up-and-down fashion). Usually, the loops joining the beta strands do not crossover the 'ends' of the barrel.

Greek key barrels

These are barrels formed from two, or more, Greek Key motifs. It is a stable structure. The Greek key barrel consists of four anti-parallel Beta strands where one strand changes the topology direction. Hydrogen bonding occurs between strands 1:4, and strands 2:3. Strand 2 then folds over to form the structural motif.

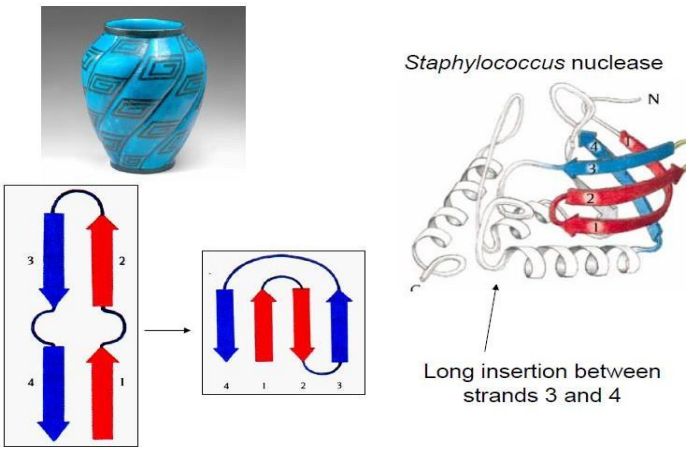


Figure 14 Greek key barrels

Jelly roll barrels

These barrels are formed from a 'Greek Key-like' structure called a jelly roll. Supposedly named because the polypeptide chain is wrapped around a barrel core like a jelly roll (swiss roll). It is a stable structure. This structure is found in coat proteins of spherical viruses, plant lectin concanavalin A, and hemagglutinin protein from influenza virus.

The essential features of a jelly roll barrel are that:

- it is like an inverted 'U' (which is often seen twisted and distorted in proteins)
- it is usually divided into two beta sheets which are packed against each other
- most jelly roll barrels have eight strands although any even number greater than 8 can form a jelly roll barrel
- it folds such that hydrogen bonds exist between strands 1 and 8; 2 and 7; 3 and 6; and 4 and 5

Figure 15 Jelly roll barrels

Beta sandwich

A beta sandwich is essentially a 'flattened' beta barrel with the two sheets packing closely together (like a sandwich!). The first and last strands of the sandwich do not hydrogen bond to each other to complete a 'barrel' structure.

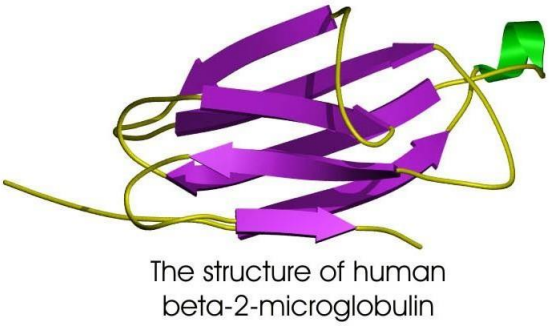


Figure 16 Beta sandwich in beta 2 microglobulin

Aligned or Orthogonal beta strands

Beta strands in barrels or sandwich structures can be orientated in two general ways:

- where the strands in two sheets are almost aligned, and in the same orientation, to each other and form an 'aligned beta' structure (eg. gamma crystallin)

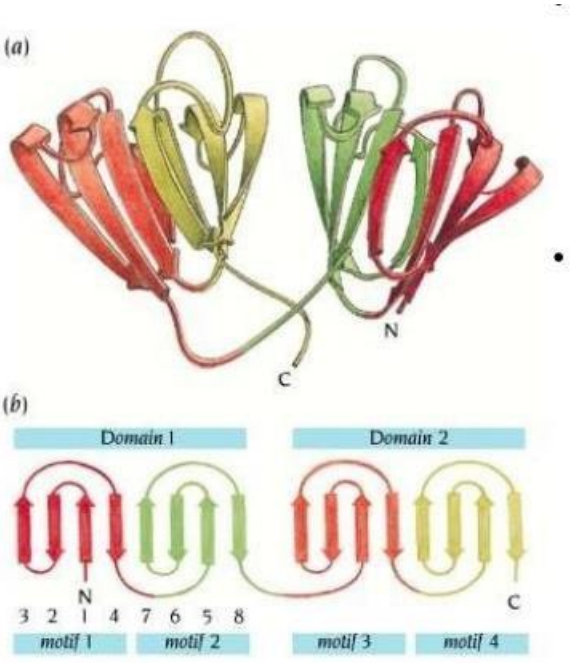


Figure 17 Aligned beta strands

- where the strands, in at least two sheets, are roughly perpendicular to each other and form an 'orthogonal beta' structure.

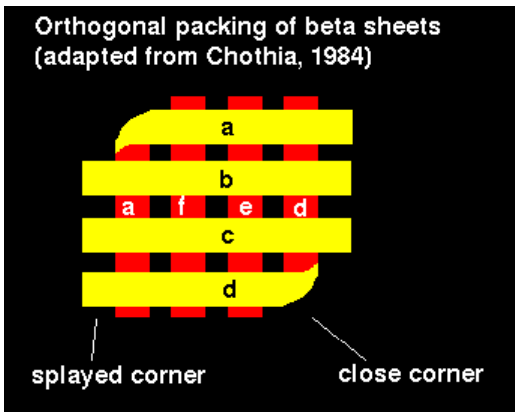


Figure 18 Orthogonal beta sheets

Beta-hairpin: two antiparallel beta strands connected by a “hairpin” bend, i.e. beta-turn 2 x antiparallel beta-strands + beta-turn = beta hairpin

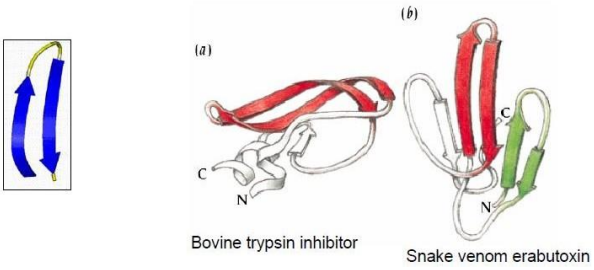


Figure 19 Beta hairpin

Beta-beta corner

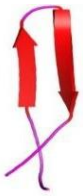


Figure 20 Beta-beta corner

- Two antiparallel beta strands which form a beta hairpin can change direction abruptly. The angle of the change of direction is about 90 degrees and so the structure is known as a 'beta corner'
- The abrupt angle change is achieved by one strand having a glycine residue (so there is no steric hindrance from a side chain) and the other strand having a beta bulge (where the hydrogen bond is broken).
- no known function

α/β Topologies

Beta-Helix-Beta Motif

An important and widespread supersecondary structural motif in proteins is known as the β- α-β motif (Beta-Alpha-Beta motif). The motif consists of two parallel Beta strands that is connected via an alpha helix (with two turns). The motif is found in most proteins that contain parallel beta strands, and the axis of the Helix and the Strands are roughly parallel to each other with all three elements forming a hydrophobic core due to shielding. The β-α-β motif may be structurally or functionally involved. The Loop that

connects the C-terminal of first Beta strand and N-terminal of Helix is frequently involved in ligand binding functions, and the motif itself is frequently found in ion channels.

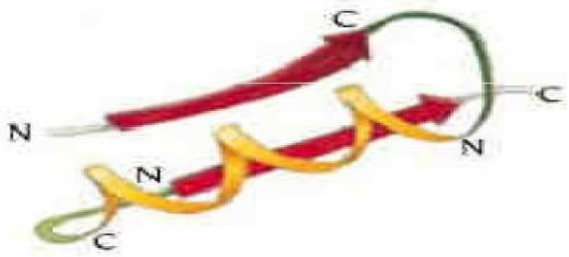


Figure 21 Beta-Helix-Beta Motif

The β - α - β - α - β subunit, often present in nucleotide-binding proteins, is named the **Rossman Fold**, after Michael Rossman

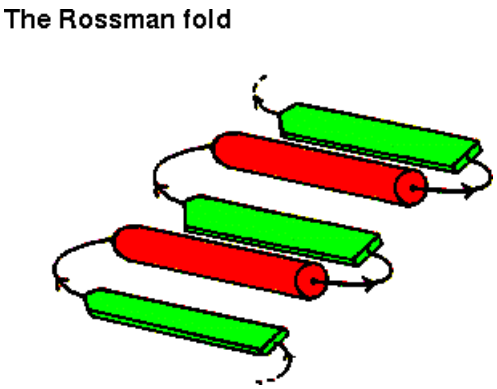


Figure 22 Rossman fold

α/β horse shoe

17-stranded parallel b sheet curved into an open horseshoe shape, with 16 a-helices packed against the outer surface. It doesn't form a barrel although it looks as though it should. The strands are only very slightly slanted, being nearly parallel to the central `axis'.

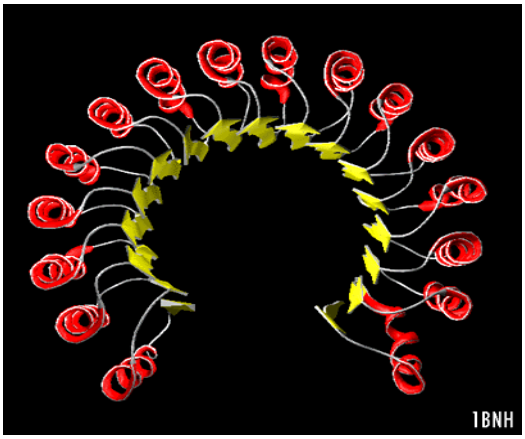


Figure 23 α/β horse shoe

α/β barrels

Consider a sequence of eight α/β motifs:

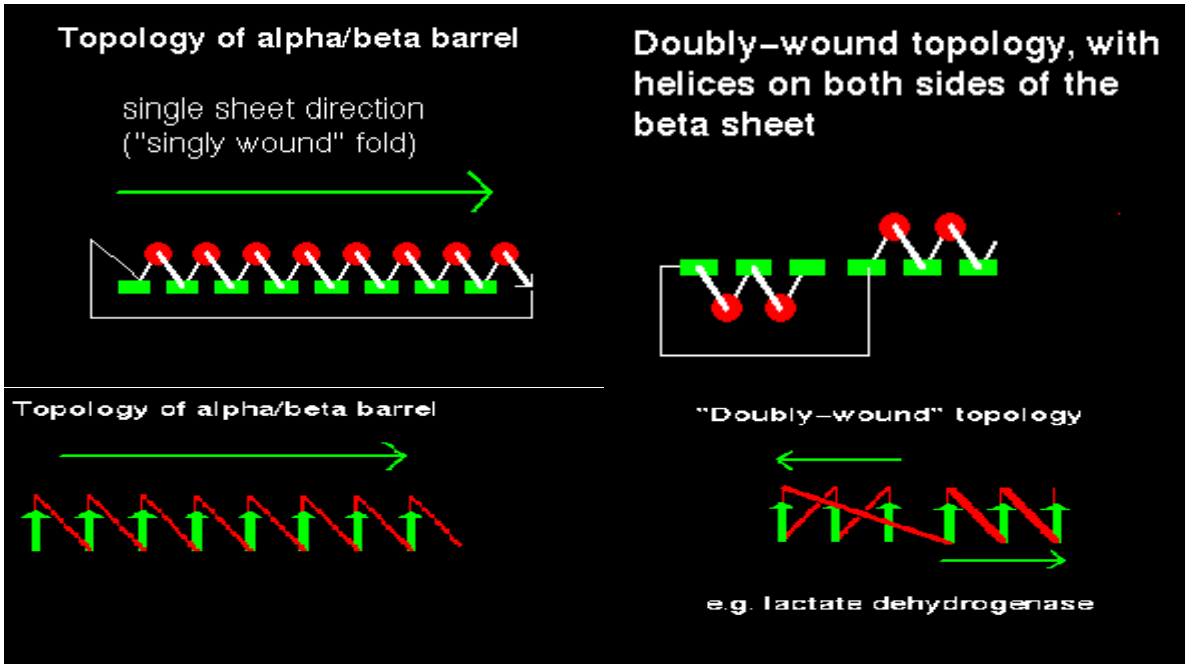


Figure 24: Topology of α/β barrel

If the first strand hydrogen bonds to the last, then the structure closes on itself forming a barrel-like structure. This is shown in the picture of triose phosphate isomerase.

Note that the "staves" of the barrel are slanted, due to the twist of the b sheet. Also notice that there are effectively four layers to this structure. The direction of the sheet does not change (it is anticlockwise in the diagram). Such a structure may therefore be described as **singly wound**.

In a structure which is open rather than closed like the barrel, helices would be situated on only one side of the b sheet if the sheet direction did not reverse. Therefore open a/b structures must be **doubly wound** to cover both sides of the sheet.

The chain starts in the middle of the sheet and travels outwards, then returns to the centre via a loop and travels outwards to the opposite edge:

Doubly-wound topologies where the sheet begins at the edge and works inwards are rarely observed.

Alpha+Beta Topologies

This is where we collect together all those folds which include significant alpha and beta secondary structural elements, but for which those elements are `**mixed**', in the sense that they do NOT exhibit the wound alpha-beta topology. This class of folds is therefore referred to as **α+ β**

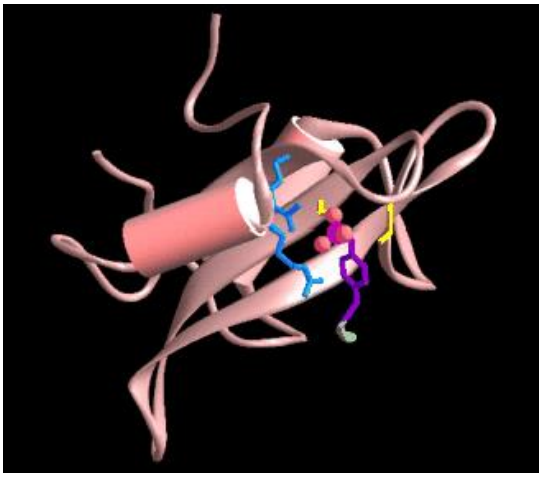


Figure 25 Alpha+Beta Topology

Domains

Domains are stable, independently folded, globular units, often consisting of combinations of motifs vary from 25 to 300 amino acids, average length – 100.

large globular proteins may consist of several domains linked by stretches of polypeptide. Separate domain may have distinct functions (eg G3P dehydrogenase). In many cases binding site formed by cleft between 2 domains frequently correspond to exon in gene

- Some examples of domains:

1. involving α-helix 4-helix bundle globin fold

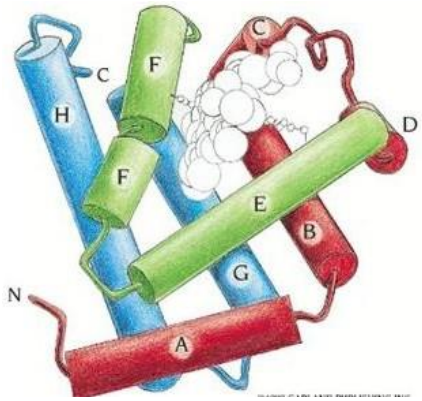


Figure 27 α-helix 4-helix bundle globin fold

The globin fold is found in its namesake globin protein families: hemoglobins and myoglobins, as well as in phycocyanins. Because myoglobin was the first protein whose structure was solved, the globin fold was thus the first protein fold discovered.

2. parallel β-sheets

hydrophobic residues on both sides, therefore must be buried.

□ □ barrel: 8 β strands each flanked by an antiparallel α-helix eg triose phosphate isomerase.)

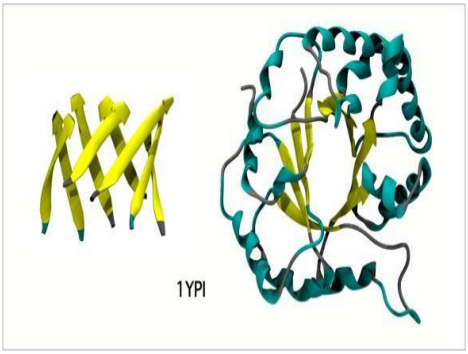


Figure 28 Parallel beta sheets

3. antiparallel β -sheet

Hydrophobic residues on one side, one side can be exposed to environment, minimum structure 2 layers

Sheets arranged in a barrel shape. More common than parallel β - barrels eg. immunoglobulin

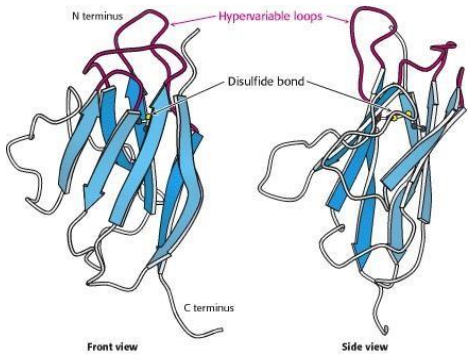


Figure 29 Antiparallel beta sheets

The **immunoglobulin domain** is a type of protein domain that consists of a 2-layer sandwich of 7-9 antiparallel β-strands arranged in two β-sheets with a Greek key topology, consisting of about 80 amino acids.

The backbone switches repeatedly between the two β-sheets. Typically, the pattern is (N- terminal β-hairpin in sheet 1)-(β-hairpin in sheet 2)-(β-strand in sheet 1)-(C-terminal β- hairpin in sheet 2). The cross-overs between sheets form an "X", so that the N- and C-

terminal hairpins are facing each other.

Members of the immunoglobulin superfamily are found in hundreds of proteins of different functions. Examples include antibodies, the giant muscle kinase titin, andreceptor tyrosine kinases. Immunoglobulin-like domains may be involved in protein–protein and protein– ligand interactions.

Example of Tertiary Structure: Myoglobin and Hemoglobin

Myoglobin and hemoglobin are hemeproteins whose physiological importance is principally related to their ability to bind molecular oxygen.

Myoglobin

Single polypeptide chain (153 amino acids).

No disulfide bonds 8 right handed alpha helices form a hydrophobic pocket which contains heme molecule protective sheath for a heme group

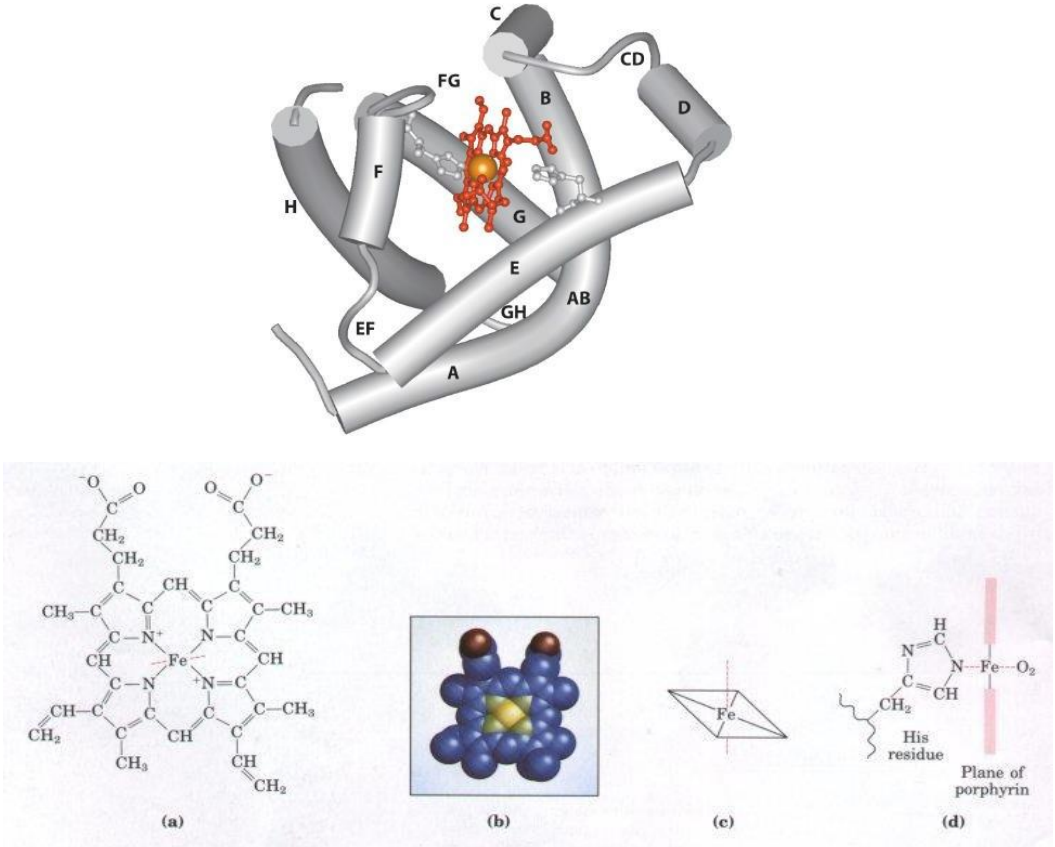


Figure 30 Myoglobin structure

Myoglobin is a monomeric heme protein found mainly in muscle tissue where it serves as an intracellular storage site for oxygen During periods of oxygen deprivation oxymyoglobin releases its bound oxygen which is then used for metabolic purposes The tertiary structure of myoglobin is that of a typical water soluble globular protein Its secondary structure is unusual in that it contains a very high proportion (75%) of α-helical secondary structure A myoglobin polypeptide is comprised of 8 separate right handed a-helices, designated A through H, that are connected by short non helical regions Amino acid R-groups packed into the interior of the molecule are predominantly hydrophobic in character while those exposed on the surface of the molecule are generally hydrophilic, thus making the molecule relatively water soluble.

Each myoglobin molecule contains one heme prosthetic group inserted into a hydrophobic cleft in the protein Each heme residue contains one central coordinately bound iron atom that is normally in the Fe 2+ , or ferrous, oxidation state The oxygen carried by hemeproteins is bound directly to the ferrous iron atom of the heme prosthetic group. The heme group is located in a crevice Except for one edge, non polar side chains surround the heme Fe 2+ is octahedrally coordinated Fe 2+ covalently bonded to the imidazole group of histidine 93 (F8) O 2 held on the other side by histidine 64 (E7)

Hydrophobic interactions between the tetrapyrrole ring and hydrophobic amino acid R groups on the interior of the cleft in the protein strongly stabilize the heme protein conjugate. In addition a nitrogen atom from a histidine R group located above the plane of the heme ring is coordinated with the iron atom further stabilizing the interaction between the heme and the protein. In oxymyoglobin the remaining bonding site on the iron atom (the 6th coordinate position) is occupied by the oxygen, whose binding is stabilized by a second histidine residue Carbon monoxide also binds coordinately to heme iron atoms in a manner similar to that of oxygen, but the binding of carbon monoxide to heme is much stronger than that of oxygen. The preferential binding of carbon monoxide to heme iron is largely responsible for the asphyxiation that results from carbon monoxide poisoning.

Hemoglobin

Oxygen transporter Four polypeptide chains Tetramer Each chain has a heme group Hence four O 2 can bind to each Hb Two alpha (141 amino acids) and two beta (146 amino acids) chains

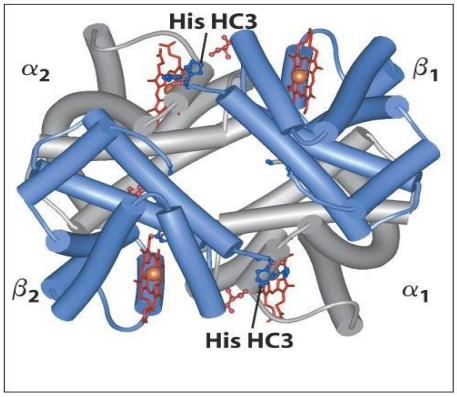


Figure 31 Hemoglobin structure

Hemoglobin is an [α(2):β(2)] tetrameric hemeprotein found in erythrocytes where it is responsible for binding oxygen in the lung and transporting the bound oxygen throughout the body where it is used in aerobic metabolic pathways Each subunit of a hemoglobin tetramer has a heme prosthetic group identical to that described for myoglobin. Although the secondary and tertiary structure of various hemoglobin subunits are similar, reflecting extensive homology in amino acid composition, the variations in amino acid composition that do exist impart marked differences in hemoglobin's oxygen carrying properties In addition, the quaternary structure of hemoglobin leads to physiologically important allosteric interactions between the subunits, a property lacking in monomeric myoglobin which is otherwise very similar to the α-subunit of hemoglobin

Quaternary structure

3-dimensional relationship of the different polypeptide chains (subunits) in a multimeric protein, the way the subunits fit together and their symmetry relationships.

• Only in proteins with more than one polypeptide chain; proteins with only one chain have no quaternary structure.

• Each polypeptide chain in a multichain protein = a subunit • 2-subunit protein = a dimer, 3 subunits = trimeric protein, 4 = tetrameric • homo(dimer or trimer etc.): identical subunits • hetero(dimer or trimer etc.): more than one kind of subunit (chains with different amino acid sequences) • different subunits designated with Greek letters – e.g., subunits of a heterodimeric protein = the "α subunit" and the "β subunit".

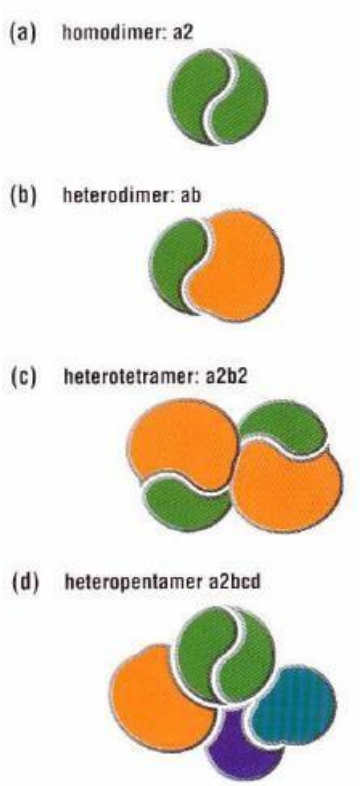


Figure 32 Protein subunits

– NOTE: This use of the Greek letters to differentiate different polypeptide chains in a multimeric protein has nothing to do with the names for the secondary structures α helix and β conformation.

• Some protein structures have very complex quaternary arrangements; e.g., mitochondrial ATP synthase, viral capsids

Symmetry in quaternary structures

• Simplest kind of symmetry = rotational symmetry

• Individual subunits can be superimposed on other identical subunits (brought into coincidence) by rotation about one or more rotational axes.

• If the required rotation = 180° (360°/2), protein has a 2-fold axis of symmetry (e.g., Cro repressor protein above).

• If the rotation = 120° (360°/3), e.g., for a homotrimer, the protein has a 3-fold symmetry axis. Rotational symmetry in proteins: Cyclic symmetry: all subunits are related by rotation about a single n-fold rotation axis (C2 symmetry has a 2-fold axis, 2 identical subunits; C3 symmetry has a 3-fold axis, 3 identical subunits, etc.)

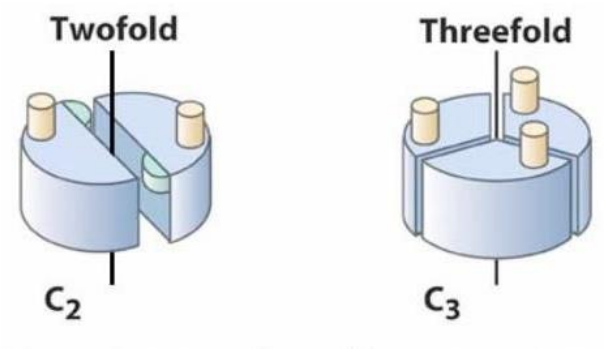


Figure 33 Two common folds of symmetry

Forces that stabilize Protein Structure

Proteins are formed of amino acids linked together by the following types of bonds

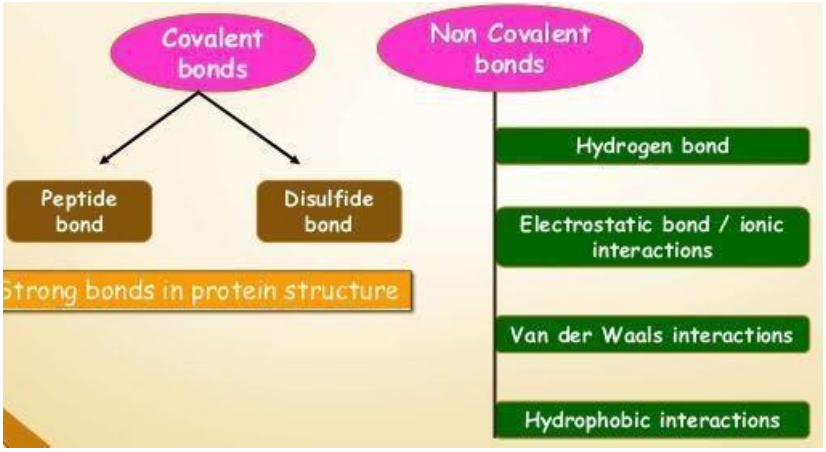


Figure 34 Forces stabilizing protein structure

Covalent Bonds - Disulfide Bridges

Covalent bonds are the strongest chemical bonds contributing to protein structure. Covalent bonds arise when two atoms share electrons.

In addition to the covalent bonds that connect the atoms of a single amino acid and the covalent peptide bond that links amino acids in a protein chain, covalent bonds between

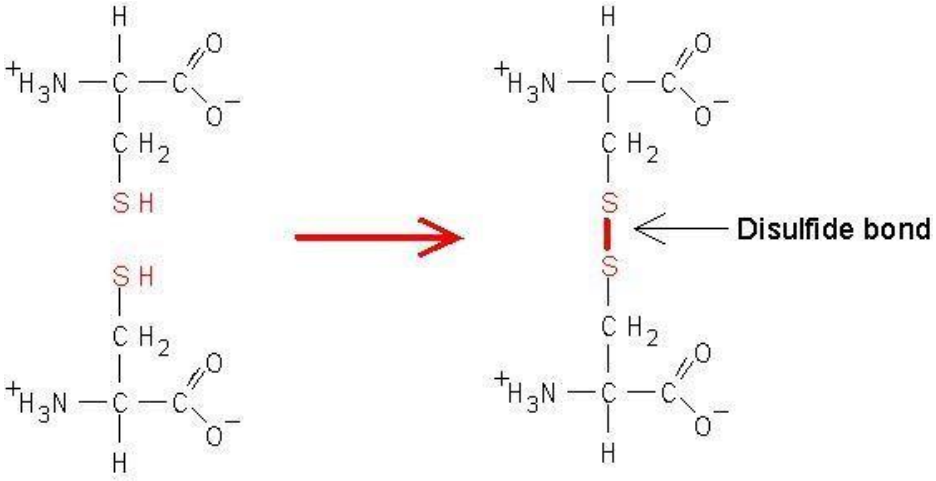


Figure 35 Disulphide bond

Non-covalent bonds Electrostatic Interactions

Ionic Bonds - Salt Bridges

Ionic bonds are formed as amino acids bearing opposite electrical charges are juxtaposed in the hydrophobic core of proteins. Ionic bonding in the interior is rare because most charged amino acids lie on the protein surface. Although rare, ionic bonds can be important to protein structure because they are potent electrostatic attractions that can approach the strength of covalent bonds. An ionic bond-salt bridge between a negatively charged O on the sidechain of glutamic acid lies 2.8 Å from the positively charged N on the amino terminus (lysine) is shown here.

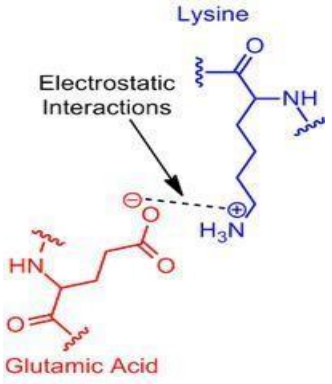


Figure 36 Electrostatic bond

Hydrogen Bonds

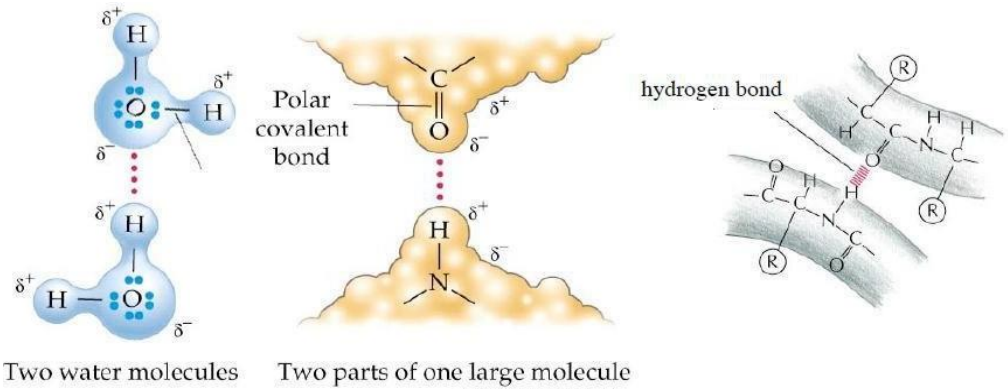


Figure 37 Hydrogen bond

Hydrogen bonds are a particularly strong form of dipole-dipole interaction. Because atoms of different elements differ in their tendencies to hold onto electrons -- that is, because they have different electronegativities -- all bonds between unlike atoms are polarized, with more electron density residing on the more electronegative atom of the bonded pair. Separation of partial charges creates a dipole, which you can think of as a mini-magnet with a positive and a negative end. In any system, dipoles will tend to align so that the positive end of one dipole and the negative end of another dipole are in close proximity. This alignment is favorable.

Hydrogen bonds are dipole-dipole interactions that form between heteroatoms in which one heteroatom (e.g. nitrogen) contains a bond to hydrogen and the other(e.g. oxygen) contains an available lone pair of electrons. You can think of the hydrogen in a hydrogen bond as being shared between the two heteroatoms, which is highly favorable. Hydrogen bonds have an ideal X-H-X angle of 180°, and the shorter they are, the stronger they are. Hydrogen bonds play an important role in the formation of secondary structure. Alpha helices are hydrogen bonded internally along the backbone whereas beta strands are hydrogen bonded to other beta strands. Side chains can also participate in hydrogen bonding interactions. You should be able to list the side chains that can participate in hydrogen bonds now that you know the structures of the side chains. Because hydrogen bonds are directional, meaning the participating dipoles must be aligned properly for a hydrogen bond to form (another w ay of saying it is that the hydrogen bonding angle must be larger than about 135°, with an optimum of 180°), and because unfavorable alignment of participating dipoles is repulsive, hydrogen bonds between side chains play key roles in determining the unique structures that different proteins form.

Hydrophobic Bonds

Hydrophobic bonds are a major force driving proper protein folding. Burying the nonpolar surfaces in the interior of a protein creates a situation where the water molecules can hydrogen bond with each other without becoming excessively ordered. Thus, the energy of the system goes down.

Therefore, an important factor governing the folding of any protein is the distribution of its polar and nonpolar amino acids. The nonpolar (hydrophobic) side chains in a protein such as those belonging to phenylalanine, leucine, isoleucine, valine, methionine and tryptophan tend to cluster in the interior of the molecule (just as hydrophobic oil droplets coalesce in water to form one large droplet). In contrast, polar side chains such as those belonging to arginine, glutamine, glutamate, lysine, etc. tend to arrange themselves near the outside of the molecule, where they can form hydrogen bonds with water and with other polar molecules. There are some polar amino acids in protein interiors, however, and these are very important in defining the precise shape adopted by the protein because the pairing of opposite poles is even more significant than it is in water.

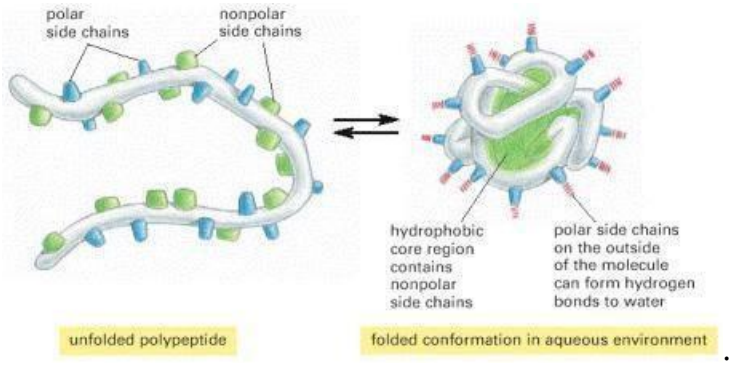


Figure 38 Hydrophobic bonds

Van der Waals Forces

The Van der Waals force is a transient, weak electrical attraction of one atom for another. Van der Waals attractions exist because every atom has an electron cloud that can fluctuate, yielding a temporary electric dipole. The transient dipole in one atom can induce a complementary dipole in another atom, provided the two atoms are quite close. These short- lived, complementary dipoles provide a weak electrostatic attraction, the Van der Waals force. Of course, if the two electron clouds of adjacent atoms are too close, repulsive forces come into play because of the negatively-charged electrons. The appropriate distance required for Van der Waals attractions differs from atom to atom, based on the size of each electron cloud, and is

referred to as the Van der Waals radius. The dots around atoms in this and other displays represent Van der Waals radii.

Van der Waals attractions, although transient and weak, can provide an important component of protein structure because of their sheer number. Most atoms of a protein are packed sufficiently close to others to be involved in transient Van der Waals attractions.

Van der Waals forces can play important roles in protein-protein recognition when complementary shapes are involved. This is the case in antibody-antigen recognition, where a "lock and key" fit of the two molecules yields extensive Van der Waals attractions.

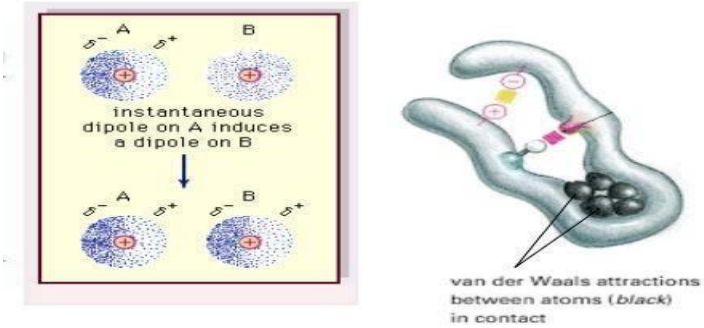


Figure 39 Van der Waals forces

PROTEIN STRUCTURE PREDICTION

Proteins are one of the major biological macromolecules performing a variety functions such as enzymatic catalysis, transport, regulation of metabolism, nerve conduction, immune response etc. The three-dimensional structure of a protein is responsible for its function.

Sequence-Structure Gap and the Need for Structure Prediction

With the advent of recombinant DNA technology it has become possible to determine the amino acid sequences of proteins quite rapidly. However, determining the three dimensional structure of proteins is a time consuming task and hence there exists a vast gap between the number of proteins of known amino acid sequence and that of known structures. This is called as the sequence-structure gap. As the knowledge of the 3-D structure of a protein is very essential to understand its function, it is imperative to develop techniques to predict the structure of a protein from its amino acid sequence.

Basis for Structure Prediction:

The classic experiments carried out by C.B. Afinson in the 60's on the enzyme ribonuclease led to the conclusion that the information to specify the 3-D structure of a protein resides in its amino acid sequence. Within the cell a newly synthesized protein chain spontaneously folds into the compact globular structure to perform its function. Thus nature has an algorithm to fold proteins to their native structures. Efforts have been directed for the past four decades to discover nature's algorithm and computational methods have been developed to predict the structure of proteins from their sequences.

Approaches to Structure Prediction

Prediction of protein structures can be classified into two major categories viz.

(i) Prediction of secondary structure and

(ii) Prediction of tertiary (3-D) structure.

Prediction of secondary structure of proteins attempts to locate segments of the polypeptide

chain adopting the α-helical or β-strand structure. Regions that are devoid of these regular secondary structural elements are considered to adopt coil conformation.

In tertiary structure prediction, one attempts to predict the three-dimensional structure of a protein or the native structure. While so far this has remained an elusive goal, different methods have been developed to press forward to the attainment of this goal.

Secondary structure prediction

What?

- Given a protein sequence (primary structure)

GHWIATRGQLIREAYEDYRHFSSECPFIP

(C=Coils H=Alpha Helix E=Beta Strands)

CEEEEECHHHHHHHHHHHHCCCHHCCCCCC

- 1 st step in prediction of protein structure.
- Technique concerned with determination of secondary structure of given polypeptide by locating the Coils Alpha Helix Beta Strands in polypeptide

Why?

- secondary structure —tertiary structure prediction
- Protein function prediction
- Protein classification
- Predicting structural change
- detection and alignment of remote homology between proteins
- on detecting transmembrane regions, solvent-accessible residues, and other important features of molecules
- Detection of hydrophobic region and hydrophilic region

Prediction methods

Chou-Fasman method

- Based on the propensities of different amino acids to adopt different secondary structures

- Predictions are made using a rules-based approach to identify groups of amino acids with shared secondary structure propensities

Garnier, Osguthorpe, Robson (GOR) method

- Statistical method of secondary structure prediction based on information theory & Bayesian probability

Multiple Sequence Alignment (MSA) methods

- Performs secondary structure prediction on a multiple sequence alignment as opposed to a single protein sequence

Neural network-based methods

- Example: **P**rofile network from **H**ei**d**elberg (PHD)

Chou-Fasman method:

1. Alpha Helix Prediction:

A. Nucleate a helix by scanning for groups of 6 residues with at least 4 helix formers (Hα and hα) and no more than 1 helix breaker (Bα and bα).

- Two Iα residues count as one helix former for nucleating a helix

B. Propagate predicted helix in both directions until reach a four residue window with average propensity (Pα) < 1.0

C. The average propensity (Pα) for a predicted helix must be Pα > 1.03 and Pα > Pβ

2. Beta Strand Prediction:

A. Nucleate a β-strand by scanning for groups of 5 residues with at least 3 strand formers (Hβ and hβ) and no more than 1 strand breaker (B$ and b$).

B. Propagate predicted β-strand in both directions until reach a four residue window with average propensity (Pβ) < 1.0

C. The average propensity (Pβ) for a predicted β-strand must be Pβ > 1.05 and Pβ > Pα

3. Resolving conflicting predictions:

(regions with both α-helix and β-strand assignment)

- If average Pα > average Pβ, then the region is alpha helix
- If average Pβ > average Pα, then the region is beta strand

Chou-Fasman algorithm:

• Later versions of the algorithm included predictions for turns

• The original algorithm contained additional rules about the location of certain residues (e.g., proline) in α-helices and β-strands

• More recent versions of the algorithm have used sequential tetrapeptide average propensities to predict secondary structure

• The propensity values have also been variously recalculated with larger protein data sets (original data sets based on 15 and 29 proteins)

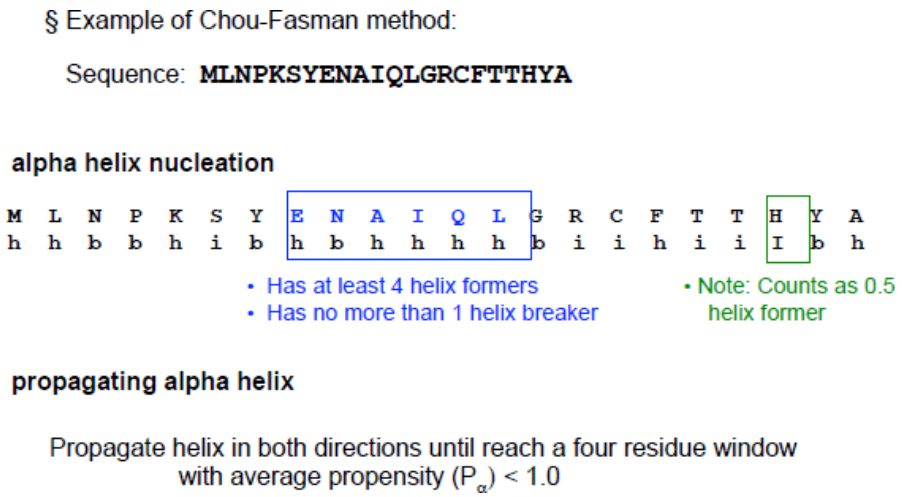


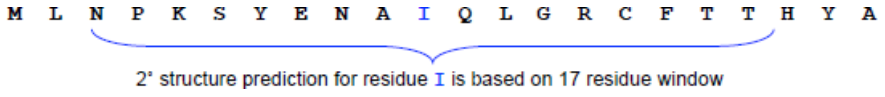
Figure 40

GOR (Garnier,Osguthorpe,Robson) Method

Key difference: Chou-Fasman uses individual amino acid propensities, while GOR incorporates information about neighboring amino acids to make prediction

A 20 x 17 matrix of directional information values for each secondary structure class was calculated from a database of known structures

These matrices are used to predict the secondary structure of the central (9th) residue in a 17 residue window:



The secondary structure class with highest information score over 17 residue window is selected as the prediction for the central residue of the window (e.g., I is predicted to be α-helix)

Multiple sequence alignment method

A multiple sequence alignment arranges protein sequences into a rectangular array with the goal that residues in a given column are homologous (derived from a common ancestor), superposable (in a 3D structural alignment - α helix / β sheet) or play a common functional role (catalytic sites, nuclear localisation signal, protein-protein interaction sites,...). Uses BLAST to identify homologous protein sequence fragments in a protein structure database (PDB).

```
VTISCTGSSSNIGAG-NHVKWYQQLPG
VTISCTGTS-NIGS-ITVNWYQQLPG-
LRLS-CSVSGFIFSS-YAMYWVRQAPG
-LS-LTCTVSGTSFDDYYSTWVRQPPG
PEVTCVVVDVSHEDPQVKFN-WYVDG-
A--TLVCTISDFYPGAVTVA-WKADS-
AALGCTVKDYFPEPVTVSWN--SG---
VSLTCTVKGFYPSD--IAVEWESNG--
```

Goal: try to have a maximum of identical/similar residues in a given column of the alignment

```
VTISCTGSSSNIGAG-NHVKWYQQLPG
VTISCTGTSSNIGS--ITVNWYQQLPG
LRLSCSVSGFIFSS--YAMYWVRQAPG
LSLTCTVSGTSFDD--YYSTWVRQPPG
PEVTCVVVDVSHEDPQVKFNWYVDG--
ATLVCTISDFYPGA--VTVAWKADS--
AALGCTVKDYFPEP--VTVSWNSG---
VSLTCTVKGFYPSD--IAVEWESNG--
```

Table 1

Main Criteria for building a multiple sequence alignment

Criterion	*Meaning*
Structure similarity	**Amino acids that play the same role in each structure are in the same column.** Structure superposition programs are the only ones that use this criterion.
Evolutionary similarity	**Amino acids or nucleotides related to the same amino acid (or nucleotide) in the common ancestor of all the sequences are put in the same column.** No automatic program explicitly uses this criterion, but they all try to deliver an alignment that respects it.
Functional similarity	**Amino acids or nucleotides with the same function are in the same column.** No automatic program explicitly uses this criterion, but if the information is available, you can force some programs to respect it or you can edit your alignment manually.
Sequence similarity	**Amino acids in the same column are those that yield an alignment with maximum similarity.** Most programs use sequence similarity because it is the easiest criterion. When the sequences are closely related, structure, evolutionary and functional similarities are equivalent to sequence similarity.

What are the applications of multiple sequence alignment

§ Protein structure and function prediction

Figure 41

§ Phylogenetic inference

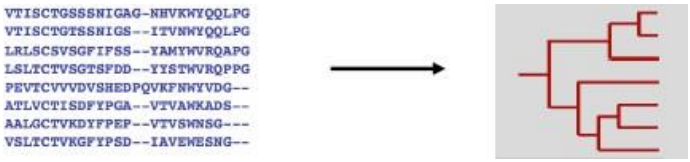


Figure 42

§ Detecting similarities between sequences (closely or distantly related) and conserved regions / motifs in sequences.

§ Detection of structural patterns (hydrophobicity/hydrophilicity, gaps etc), thus assisting improved prediction of secondary and tertiary structures and loops and variable regions.

§ Predict features of aligned sequences like conserved positions which may have structural or functional importance.

§ Computing consensus sequence.

§ Making patterns or profiles that can be further used to predict new sequences falling in a given family.

§ Deriving profiles or Hidden Markov Models that can be used to remove distant sequences (outliers) from protein families.

§ Inferring evolutionary trees / linkage.

How is a multiple sequence alignment used?

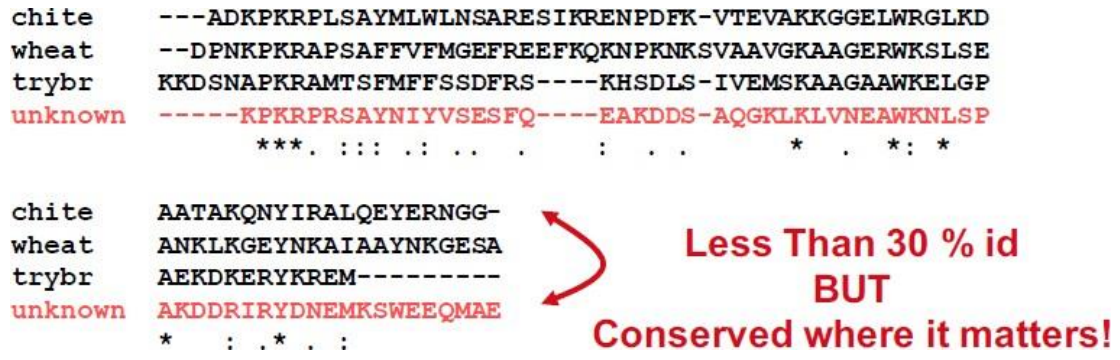


Figure 43

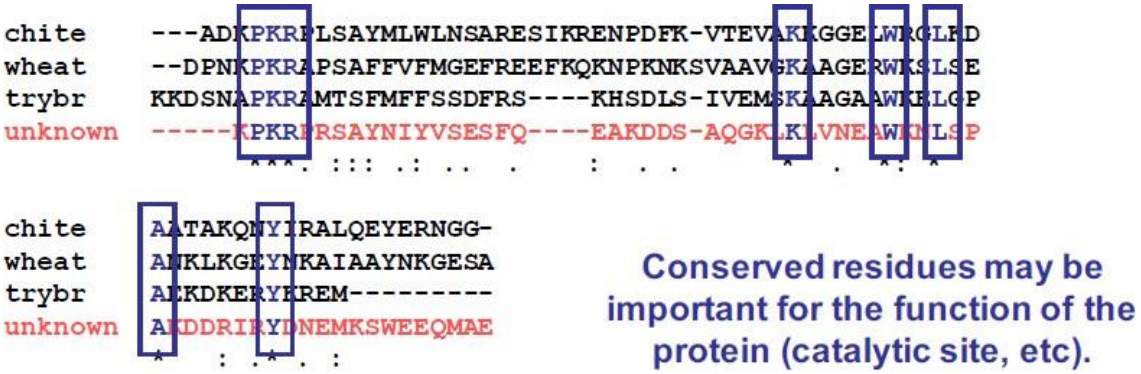


Figure 44

Neural network secondary structure prediction methods

Artificial neural networks (ANN), with both statistical (linear regression and discriminant analysis) and artificial intelligence roots, are information processing units that that are modeled after the brain and its 100 billion neurons. In a neuron, the distal and proximal dendrites receive signals and communicate to the cell body, which in turn communicates with other neurons via its axon and its terminals.

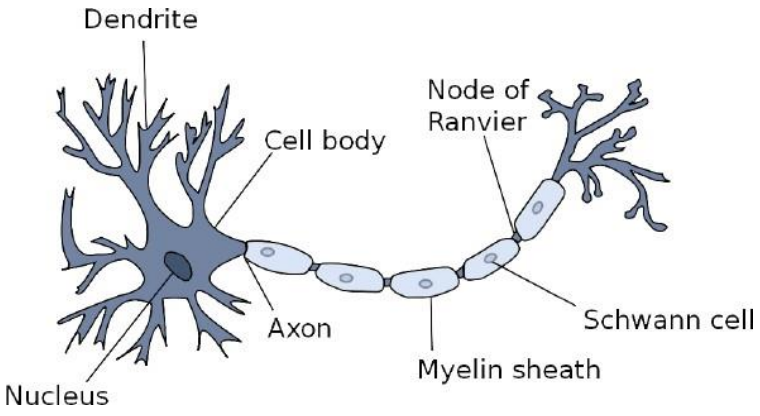


Figure 45

Similarly, an ANN receives inputs (dendrites) that are processed with influence by weights to become outputs (axon).

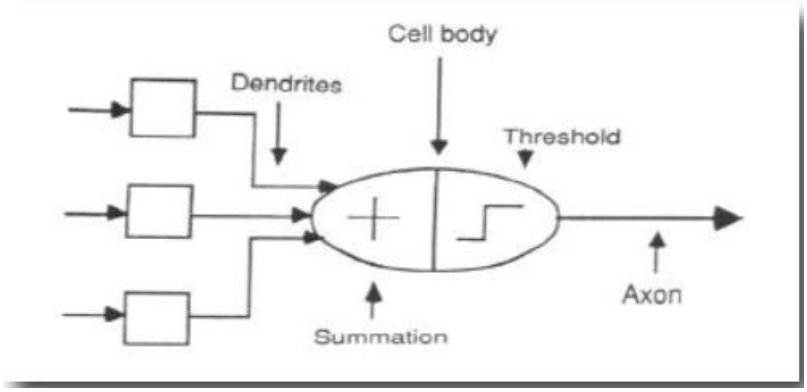


Figure 47

The neurons or nodes interconnect with informational flows (unidirectional or bidirectional) at various weights or strengths. The simplest architecture is the perceptron, which consists of 2 layers (input and output layers) that are separated by a linear discrimination function (10). In a multi-layer perceptron (MLP) model, there are three layers: the input nodes, the hidden nodes layer, and the output nodes.

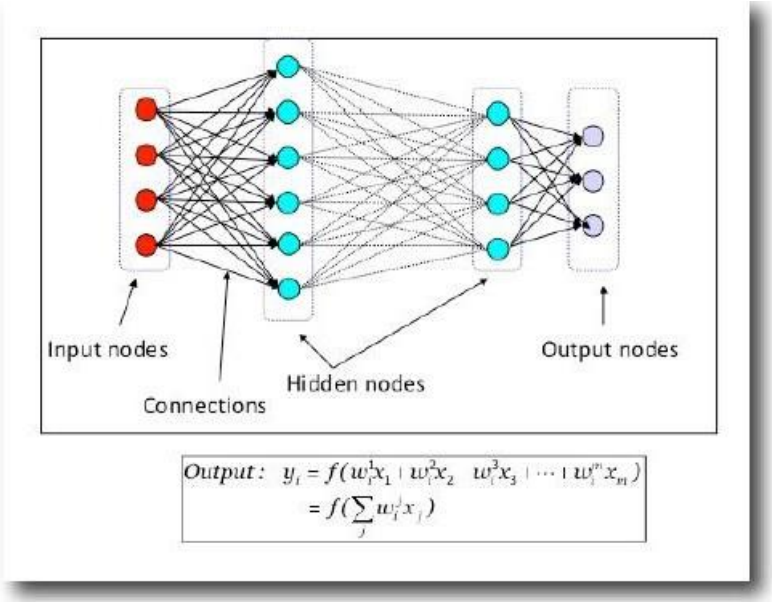


Figure 48

Learning/ Training

In a feed-forward neural network architecture, a unit will receive input from several nodes or neurons belonging to another layer. These highly interconnected neurons therefore form an infrastructure (similar to the biological central nervous system) that is capable of learning by successfully perform pattern recognition and classification tasks. Training of the ANN is a process in which learning occurs from representative data and the knowledge is applied to the new situation.

This training or learning process occurs by arranging the algorithms so that the weights of the ANN are adjusted to lead to the final desired output. The learning in neural networks can be supervised (such as the multilayer perceptron that trained with sets of input data) or unsupervised (such as the Kohonen self-organizing maps which learn by finding patterns). Neural networks can also perform both regression and classification.

The ANN learning process consists of both a forward and a backward propagation process. The forward propagation process involves presenting data into the ANN whereas the important backward propagation algorithm determines the values of the weights for the nodes during a training phase. This latter process is accomplished by directing the errors for input values backwards so that corrections for the weights can be made to minimize the error of actual and desired output data. A recurrent neural network is a series of feed-forward neural networks sharing the same weights and is good for time series data. ANN can therefore extract patterns or detect trends from complicated and imprecise data sets.

Application of ANN to bioinformatics needs the following strategy:

Extraction of features from molecular sequences to serve as training/prediction data; preprocessing that consists of feature selection and encoding into vectors of real numbers; neural network for training or prediction; post processing that consists of output encoding from the neural network; and finally the myriad of applications (such as sequence analysis, gene expression data analysis, or protein structure prediction).

In secondary structure prediction, neural network methods are trained using sequences with known secondary structure, and then asked to predict the secondary structure of proteins of unknown structure

§ Example: **P**rofile network from **H**ei**d**elberg (PHD) uses multiple sequence alignment with neural network methods to predict secondary structure.

Protein structure prediction is one of the most important goals pursued by bioinformatics and theoretical chemistry; it is highly important in medicine (for example, in drug design) andbiotechnology (for example, in the design of novel enzymes). Every two years, the performance of current methods is assessed in the CASP experiment (Critical Assessment of Techniques for Protein Structure Prediction). A continuous evaluation of protein structure prediction web servers is performed by the community project CAMEO3D.

Accuracy of Secondary Structure Prediction

§ Prediction accuracy

- Accuracy is usually measured by Q3 (or Qindex) value
- For a single conformation state, i:

$$Q_i = \frac{\text{number of residues correctly predicted in state i}}{\text{number of residues observed in state i}} * 100\%$$

• where i is either helix, strand, or coil. For all three states:

$$Q_3 = \frac{\text{number of residues correctly predicted}}{\text{number of all residues}} * 100\%$$

§ Accuracy of prediction methods

- A random prediction has a Q3 value of ~ 33-38%
- Chou-Fasman method typically has a Q3 ~ 56-60%
- GOR method (depending upon version) has a Q3 ~ 60-65%
- MSA, neural network methods have Q3 ~70%

PROTEIN TERTIARY STRUCTURES: PREDICTION FROM AMINO ACID SEQUENCES

The biological function of a protein is often intimately dependent upon its tertiary structure. X-ray crystallography and nuclear magnetic resonance are the two most mature experimental methods used to provide detailed information about protein structures. However, to date the majority of the proteins still do not have experimentally determined structures available. As at December 2000, there were about 14 000 structures available in the protein data bank (PDB, http://www.pdb.org), and there are about 10 106 000 sequence records sequences in GenBank (http://www.ncbi.nlm.nih.gov/Genbank). Thus theoretical methods are very important tools to help biologists obtain protein structure information. The goal of theoretical research is not only to predict the structures of proteins but also to understand how protein molecules fold into the native structures. The current methods for protein structure prediction can be roughly divided into three major categories: comparative modelling; threading; and ab initio prediction. For a given target protein with unknown structure, the general procedure for predicting its structure is described below:

Comparative modelling

It is based on two major observations:

1. The structure of a protein is uniquely determined by its amino acid sequence. Knowing the sequence should, at least in theory, suffice to obtain the structure.

2. During evolution, the structure is more stable and changes much slower than the associated sequence, so that similar sequences adopt practically identical structures, and distantly related sequences still fold into similar structures. This relationship was first identified by Chothia and Lesk (1986) and later quantified by Sander and Schneider (1991).

Thanks to the exponential growth of the Protein Data Bank (PDB), Rost (1999) could recently derive a precise limit for this rule, shown in Figure below. As long as the length of two sequences and the percentage of identical residues fall in the region marked as “safe,” the two sequences are practically guaranteed to adopt a similar structure

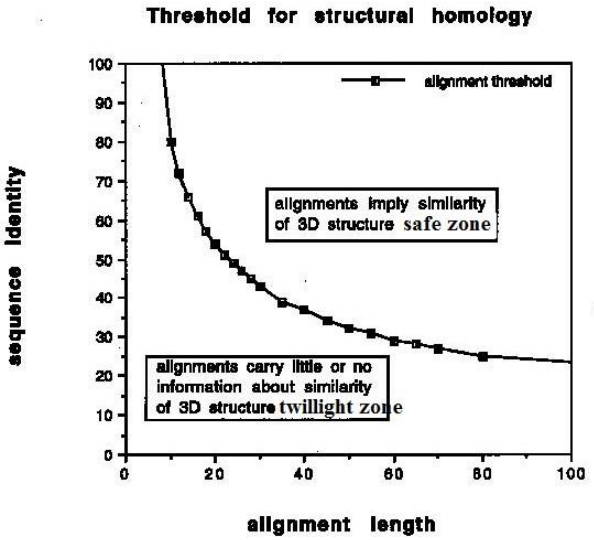


Figure 49

For a sequence of 100 residues, for example, a sequence identity of 40% is sufficient for structure prediction. When the sequence identity falls in the safe homology modeling zone, we can assume that the 3D-structure of both sequences is the same.

The known structure is called the template, the unknown structure is called the target. Homology modeling of the target structure can be done in 7 steps:

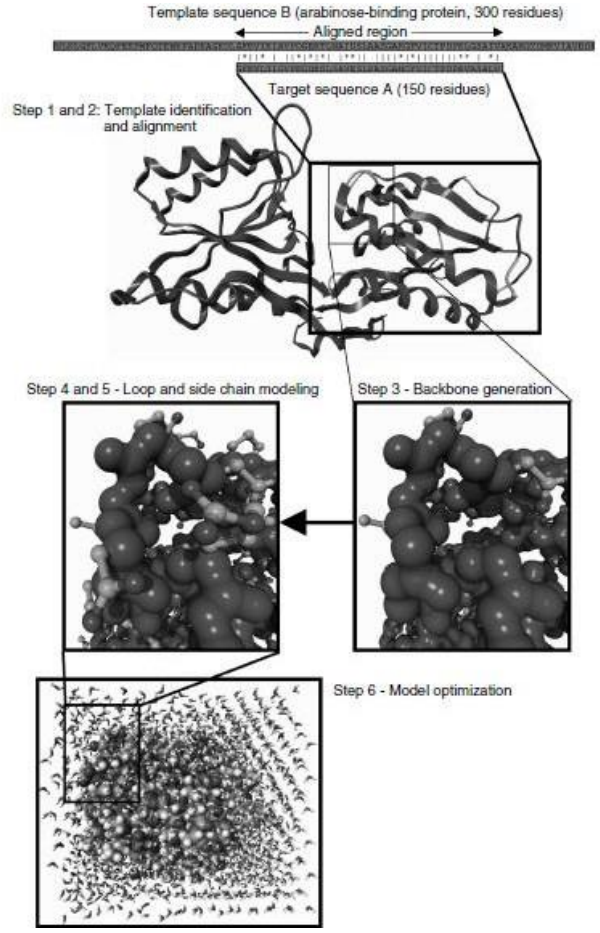


Figure 50

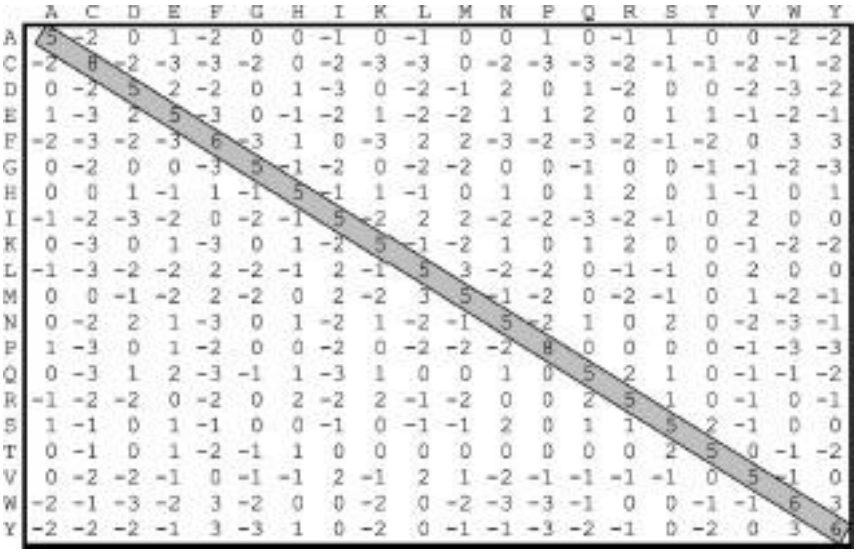
1: Template recognition and initial alignment

In the safe homology modeling zone, the percentage identity between the sequence of interest and a possible template is high enough to be detected with simple sequence alignment programs such as BLAST or FASTA. To identify these hits, the program compares the query sequence to all the sequences of known structures in the PDB using mainly two matrices:

1. A residue exchange matrix (A). The elements of this 20 ∗ 20 matrix define the likelihood

that any two of the 20 amino acids ought to be aligned. It is clearly seen that the values along the diagonal (representing conserved residues) are highest, but one can also observe that exchanges between residue types with similar physicochemical properties (for example F →

Y) get a better score than exchanges between residue types that widely differ in their properties.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	5	-2	0	1	-2	0	0	-1	0	-1	0	0	1	0	-1	1	0	0	-2	-2
C	-2	8	-2	-3	-3	-2	0	-2	-3	-3	0	-2	-3	-3	-2	-1	-1	-2	-1	-2
D	0	-2	5	2	-2	0	1	-3	0	-2	-1	2	0	1	-2	0	0	-2	-3	-2
E	1	-3	2	5	-3	0	-1	-2	1	-2	-2	1	1	2	0	1	1	-1	-2	-1
F	-2	-3	-2	-3	6	-3	1	0	-3	2	2	-3	-2	-3	-2	-1	-2	0	3	3
G	0	-2	0	0	-3	5	-1	-2	0	-2	-2	0	0	-1	0	0	-1	-1	-2	-3
H	0	0	1	-1	1	-1	5	-1	1	-1	0	1	0	1	2	0	1	-1	0	1
I	-1	-2	-3	-2	0	-2	-1	5	-2	2	2	-2	-2	-3	-2	-1	0	2	0	0
K	0	-3	0	1	-3	0	1	-2	5	-1	-2	1	0	1	2	0	0	-1	-2	-2
L	-1	-3	-2	-2	2	-2	-1	2	-1	5	3	-2	-2	0	-1	-1	0	2	0	0
M	0	0	-1	-2	2	-2	0	2	-2	3	5	-1	-2	0	-2	-1	0	1	-2	-1
N	0	-2	2	1	-3	0	1	-2	1	-2	-1	5	-2	1	0	2	0	-2	-3	-1
P	1	-3	0	1	-2	0	0	-2	0	-2	-2	-2	8	0	0	0	0	-1	-3	-3
Q	0	-3	1	2	-3	-1	1	-3	1	0	0	1	0	5	2	1	0	-1	-1	-2
R	-1	-2	-2	0	-2	0	2	-2	2	-1	-2	0	0	2	5	1	0	-1	0	-1
S	1	-1	0	1	-1	0	0	-1	0	-1	-1	2	0	1	1	5	2	-1	0	0
T	0	-1	0	1	-2	-1	1	0	0	0	0	0	0	0	0	2	5	0	-1	-2
V	0	-2	-2	-1	0	-1	-1	2	-1	2	1	-2	-1	-1	-1	-1	0	5	-1	0
W	-2	-1	-3	-2	3	-2	0	0	-2	0	-1	-3	-3	-1	0	0	-1	-1	6	3
Y	-2	-2	-2	-1	3	-3	1	0	-2	0	-1	-1	-3	-2	-1	0	-2	0	3	6

Figure 51

A* A typical residue exchange or scoring matrix used by alignment algorithms. Because the score for aligning residues A and B is normally the same as for B and A, this matrix is symmetric.

2. An alignment matrix (B). The axes of this matrix correspond to the two sequences to align, and the matrix elements are simply the values from the residue exchange matrix for a given pair of residues. During the alignment process, one tries to find the best path through this matrix, starting from a point near the top left, and going down to the bottom right. To make sure that no residue is used twice, one must always take at least one step to the right and one step down. A typical alignment path is shown in Figure B. At first sight, the dashed path in the bottom right corner would have led to a higher score. However, it requires the opening of an additional gap in sequence A (Gly of sequence B is skipped). By comparing thousands of sequences and sequence families, it became clear that the opening of gaps is about as unlikely as at least a couple of nonidentical residues in a row. The jump roughly in the middle of the matrix, however, is justified, because after the jump we earn lots of points (5,6,5), which would have been (1,0,0) without the jump. The alignment algorithm therefore subtracts an “opening penalty” for every new gap and a much smaller “gap extension penalty” for every residue that is skipped in the alignment. The gap extension penalty is smaller simply because one gap of three residues is much more likely than three gaps of one residue each. In practice, one just feeds the query sequence to one of the countless BLAST servers on the web, selects a search of the PDB, and obtains a list of hits—the modeling templates and

corresponding alignments.

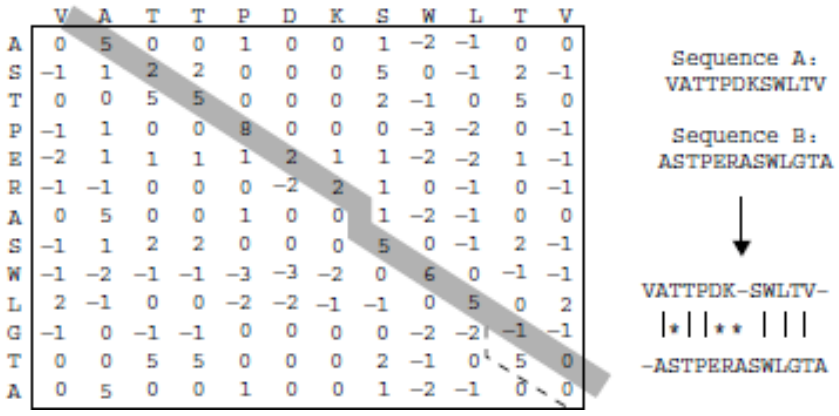


Figure 52

B: The alignment matrix for the sequences VATTPDKSWLTV and ASTPERASWLGTA, using the scores from Figure A. The optimum path corresponding to the alignment on the right side is shown in gray. Residues with similar properties are marked with a star (*). The dashed line marks an alternative alignment that scores more points but requires opening a second gap

2: Alignment correction

Having identified one or more possible modeling templates using the fast methods described above, it is time to consider more sophisticated methods to arrive at a better alignment. Sometimes it may be difficult to align two sequences in a region where the percentage sequence identity is very low.

One can then use other sequences from homologous proteins to find a solution. A pathological example is shown in C:

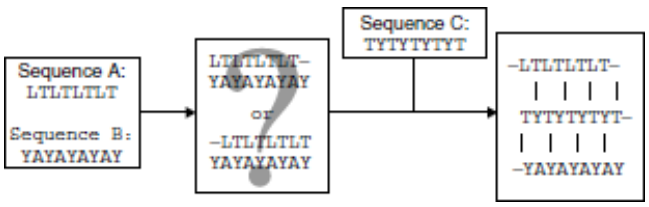


Figure 53

C: A pathological alignment problem. Sequences A and B are impossible to align, unless one considers a third sequence C from a homologous protein.

Suppose you want to align the sequence LTLTLTLT with YAYAYAYAY. There are two equally poor possibilities, and only a third sequence, TYTYTYTYT, that aligns easily to both of them can solve the issue.

The example above introduced a very powerful concept called "multiple sequence alignment." Many programs are available to align a number of related sequences, for example CLUSTALW, and the resulting alignment contains a lot of additional information.

Think about an Ala → Glu mutation. Relying on the matrix in Figure A, this exchange always gets a score of 1. In the 3D structure of the protein, it is however very unlikely to see such an Ala → Glu exchange in the hydrophobic core, but on the surface this mutation is perfectly normal. The multiple sequence alignment implicitly contains information about this structural context. If at a certain position only exchanges between hydrophobic residues are observed, it is highly likely that this residue is buried. To consider this knowledge during the alignment, one uses the multiple sequence alignment to derive position specific scoring matrices, also called profiles. When building a homology model, we are in the fortunate situation of having an almost perfect profile—the known structure of the template. We simply know that a certain alanine sits in the protein core and must therefore not be aligned with a glutamate. Multiple sequence alignments are nevertheless useful in homology modeling, for example, to place deletions (missing residues in the model) or insertions (additional residues in the model) only in areas where the sequences are strongly divergent.

A typical example for correcting an alignment with the help of the template is shown in Figures D and E. Although a simple sequence alignment gives the highest score for the wrong answer (alignment 1 in Fig. D), a simple look at the structure of the template reveals that alignment 2 is correct, because it leads to a small gap, compared to a huge hole associated with alignment 1.

		1	2	3	4	5	6	7	8	9	10	11	12	13
Template		PHE	ASP	ILE	CYS	ARG	LEU	PRO	GLY	SER	ALA	GLU	ALA	VAL
Model(bad)	1	PHE	ASN	VAL	CYS	ARG	ALA	PRO	---	---	---	GLU	ALA	ILE
Model(good)	2	PHE	ASN	VAL	CYS	ARG	---	---	---	ALA	PRO	GLU	ALA	ILE

Figure 54

D: Example of a sequence alignment where a three-residue deletion must be modeled. While the first alignment appears better when considering just the sequences (a matching proline at position 7), a look at the structure of the template leads to a different conclusion (Figure E)

Figure 55

E: Correcting an alignment based on the structure of the modeling template (Cα-trace shown in black). While the alignment with the highest score (dark gray) leads to a gap of 7.5 A between residues 7 and 11, the second option (white) creates only a tiny hole of˚ 1.3 A between residues 5 and 9. This can easily be accommodated by small backbone shifts. (The normal Cα−Cα distance of 3.8 A has been subtracted).

3: Backbone generation

When the alignment is correct, the backbone of the target can be created. The coordinates of the template-backbone are copied to the target. When the residues are identical, the side-chain coordinates are also copied. Because a PDB-file can always contain some errors, it can be useful to make use of multiple templates.

4: Loop modeling

Often the alignment will contain gaps as a result of deletions and insertions. When the target

sequence contains a gap, one can simply delete the corresponding residues in the template. This creates a hole in the model, this has already been discussed in step 2. When there is an insertion in the target, shown in Figure B, the template will contain a gap and there are no backbone coordinates known for these residues in the model. The backbone from the template has to be cut to insert these residues. Such large changes cannot be modeled in secondary structure elements and therefore have to be placed in loops and strands. Surface loops are, however, flexible and difficult to predict. One way to handle loops is to take some residues before and after the insertion as "anchor" residues and search the PDB for loops with the same anchor-residues. The best loop is simply copied in the model. This is shown in Figure G. The two residues which are colored green in Figure G are used as anchor, the best loop with the inserted resisdues was found in the database and placed in the model.

Figure 56

F: Target sequence (green) with insertion (grey box) results in a gap in the template

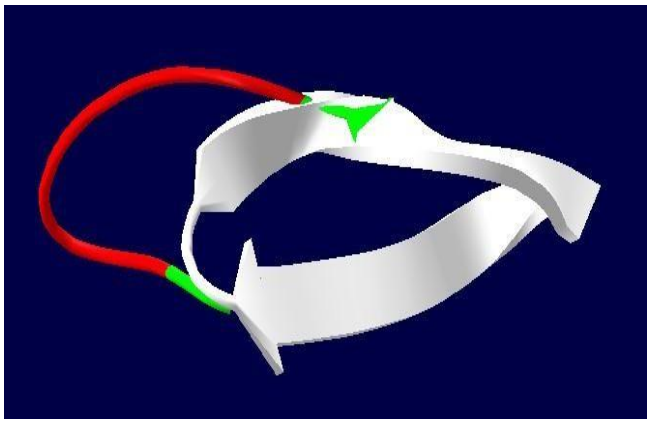


Figure 57

F: The red loop is modeled with the green residues as anchor residues. The insertion of 2 residues results in a longer loop

5: Side-chain modelling

Now it is time to add side-chains to the backbone of the model. Conserved residues were already copied completely. The torsion angle between C-alpha and C-beta of the other residues can also be copied to the model because these rotamers tend to be conserved in similar proteins. It is also possible to predict the rotamer because many backbone configurations strongly prefer a specific rotamer. As shown in Figure G, the backbone of this tyrosine strongly prefers two rotamers and the real side-chain fits in one of them. There are libraries based upon the backbone of the residues flanking the residue of interest. By using these libraries the best rotamer can be predicted. This last method is used by Yasara.

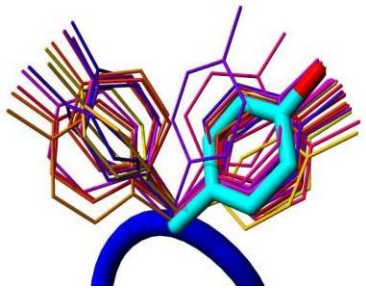


Figure 58

G: Prefered rotamers of this tyrosin (colored sticks) the real side-chain (cyan) fits in one of them.

6: Model optimization

The model has to be optimized because many structural artifacts can be introduced while the model protein is being built

- ❑ Substitution of large side chains for small ones
- ❑ Strained peptide bonds between segments taken from difference reference proteins
- ❑ Non optimum conformation of loops

Energy Minimisation is used to produce a chemically and conformationally reasonable model protein structure

Two mainly used optimisation algorithms are

- ➢ Steepest Descent
- ➢ Conjugate Gradients

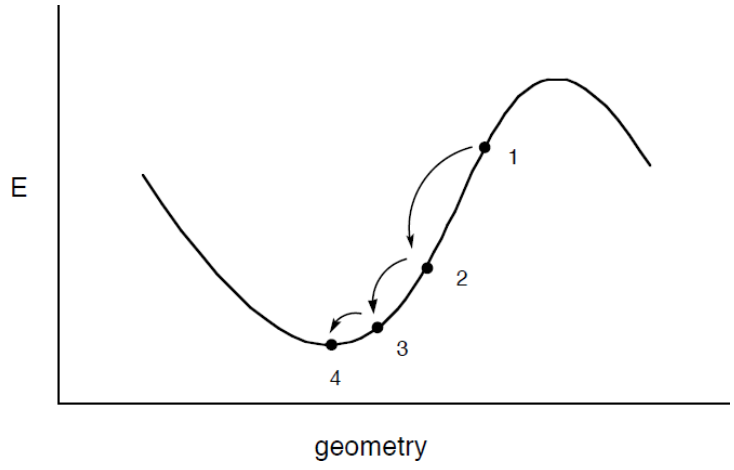


Figure 59

The process of energy minimization changes the geometry of the molecule in a step-wise fashion until a minimum is reached.

Molecular Dynamics is used to explore the conformational space a molecule could visit, Molecular dynamics (MD) is a computer simulation method for studying the physical movements of atoms and molecules

7: Model validation

The models we obtain may contain errors. These errors mainly depend upon two values.

1. The percentage identity between the template and the target.
   If the value is > 90% then accuracy can be compared to crystallography, except for a few individual side chains. If its value ranges between 50-90 % r.m.s.d. error can be as large as 1.5 Å, with considerably more errors. If the value is <25% the alignment turns out to be difficult for homology modeling, often leading to quite larger errors.

2. The number of errors in the template.
   Errors in a model become less of a problem if they can be localized. Therefore, an essential step in the homology modeling process is the verification of the model. The errors can be estimated by calculating the model's energy based on a force field. This method checks to see if the bond lengths and angles are in a normal range. However, this method cannot judge if the model is correctly folded. The 3D distribution functions can also easily identify misfolded proteins and are good indicators of local model building problems.

Modeller

Modeller is a program for comparative protein structure modelling by satisfaction of spatial restraints. It can be described as "Modeling by satisfaction of restraints" uses a set of restraints derived from an alignment and the model is obtained by minimization of these restraints. These restraints can be from related protein structures or NMR experiments. User gives an alignment of sequences to be modelled with known structures. Modeller calculates a model with all non hydrogen atoms. It also performs comparison of protein structures or sequences, clustering of proteins, searching of sequence databases.

THREADING

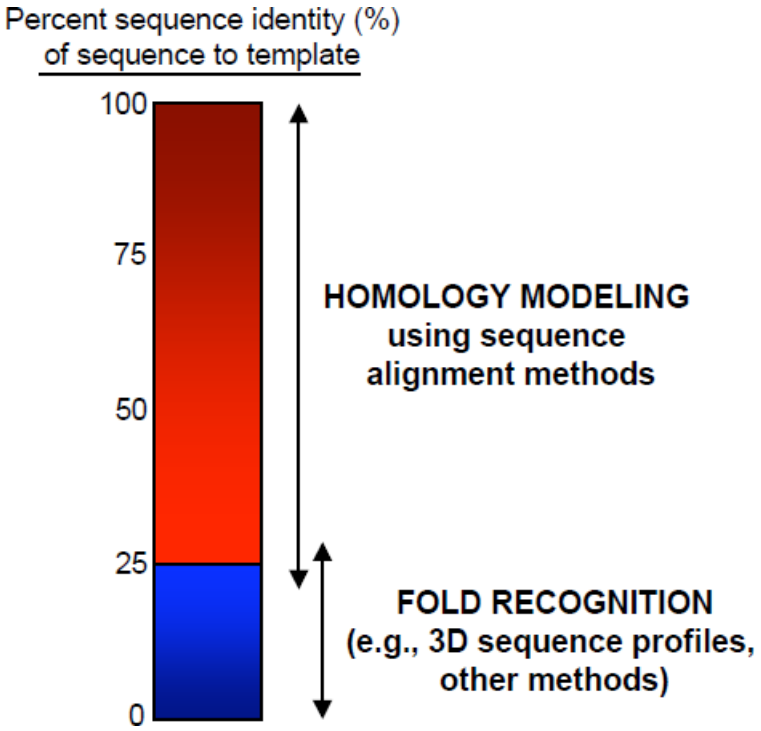


Figure 60

Threading or Fold recognition is a method to identify proteins that have similar 3D structure (fold), but limited or non existent sequence homology. The threading and sequence-structure alignment approachs are based on the observation that many protein structures in the PDB are very similar. For example, there are many 4-helical bundles, TIM barrels, globins, etc. in the set of solved structures.

As a result of this, many scientists have conjectured there are only a limited number of " unique" protein folds in nature. Estimates vary considerably, but some predict that are fewer than 1000 different protein folds. Thus, one approach to the protein structure prediction problem is to try to determine the structure of a new sequence by finding its best fit" to some fold in a library of structures.

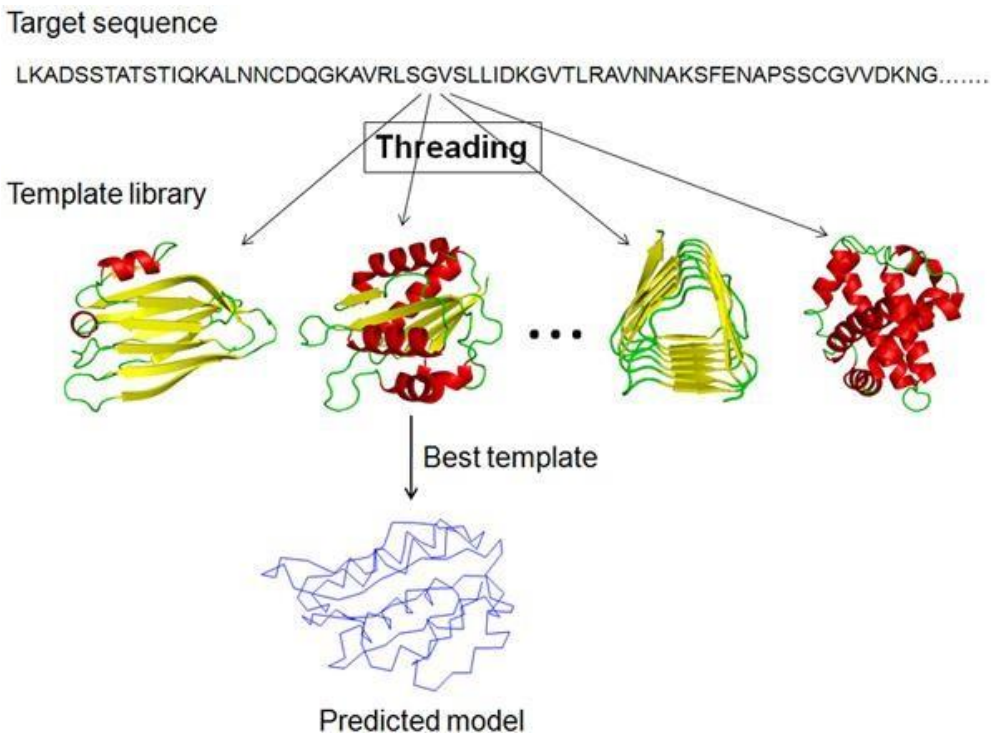


Figure 61

Given a new sequence and a library of known folds, the goal is to _figure out which of the folds (if any) is a good fit to the sequence.

Fold recognition methods include:

• 3D profiles (and protein threading)

- Align sequence to structure

• Profile-based alignment methods that integrate sequence and structural (2D or 3D) information

- e.g., 3D-PSSM or PHYRE software

As a subproblem to fold recognition, we must solve the sequence-structure alignment problem.

Namely, given a solved structure T for a sequence t_1 t_2 ….. t_n = t and a new sequences s_1 s_2 …. s_m = s, we need to find the best match" between s and T. This actually consists of two subproblems:

- Evaluating (scoring) a given alignment of s with a structure T.

- Efficiently searching over possible alignments.

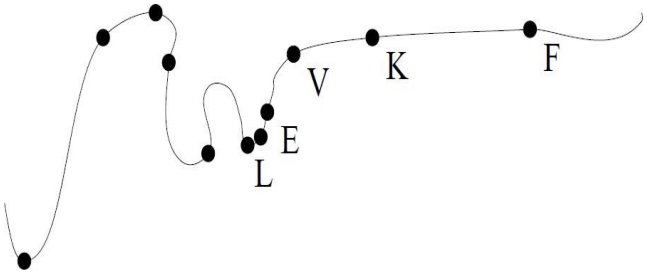


Figure 62

Example: New sequence s=LEVKF, and its best alignment to a particular structure.

There are at least three approaches to the sequence-structure alignment problem.

1. The first method is to just use protein sequence alignment. That is, find the best sequence alignment between the new sequence s and the sequence t with structure T. This is then used to infer the structural alignment: if s_i aligns with t_j , s_i's position in the 3D structure is the same as t_j 's. Scoring in this case is based on amino-acid similarity matrices (e.g., you could use the PAM-250 matrix), and the search algorithm is dynamic programming (O(nm) time). This is a non- physical method; that is, it does not use structural information. The major limitation of this method is that similar structures have lots of sequence variability, and thus sequence alignment may not be very helpful. Hidden Markov model techniques have the same problem.

2. The second method we will describe, the 3D profile method, actually uses structural information. The idea here is that instead of aligning a sequence to a sequence, we align a sequence to a string of descriptors that describe the 3D environment of the target structure. That is, for each residue position in the structure, we determine:

_ how buried it is (buried, partly buried or exposed)

_ the fraction of surrounding environment that is polar (polar or apolar)

_ the local secondary structure (α-helix, β-sheet or other)

3. Our third method for sequence-structure alignments uses contact potentials. Most "threading" methods today fall into this category.
Typically, these methods model interactions in a protein structure as a sum over pairwise interactions.

A general paradigm of protein threading consists of the following four steps:

1. Construct a library of core fold templates
2. A scoring (or objective) function is used to evaluate the placement of a sequence in a core template
3. Search for optimal alignments between the sequence and each core fold template
4. Select the core fold template that best aligns (fits) with the protein sequence
    - The 3D model is derived from the optimal alignment (or 'threading') of the sequence to the best scoring structural template

The construction of a structure template database

Select protein structures from the protein structure databases as structural templates. This generally involves selecting protein structures from databases such as PDB, FSSP, SCOP, or CATH, after removing protein structures with high sequence similarities.

Threading alignment

Align the target sequence with each of the structure templates by optimizing the designed scoring function. This step is one of the major tasks of all threading-based structure prediction programs that take into account the pairwise contact potential; otherwise, a dynamic programming algorithm can fulfill it.

Threading prediction

Select the threading alignment that is statistically most probable as the threading prediction. Then construct a structure model for the target by placing the backbone atoms of the target sequence at their aligned backbone positions of the selected structural template.

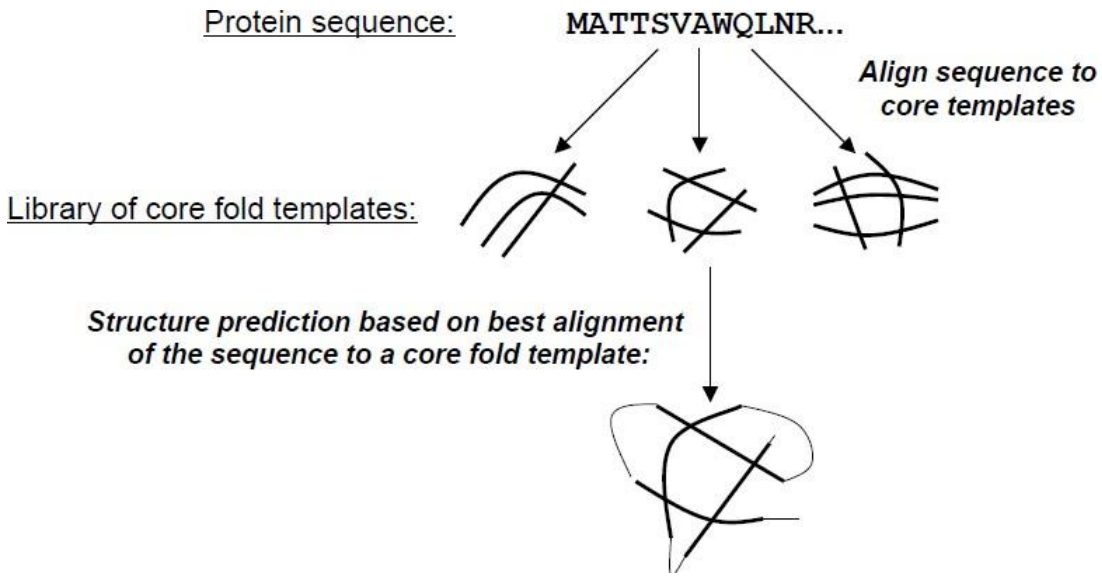


Figure 64

AB INITIO PREDICTION METHOD

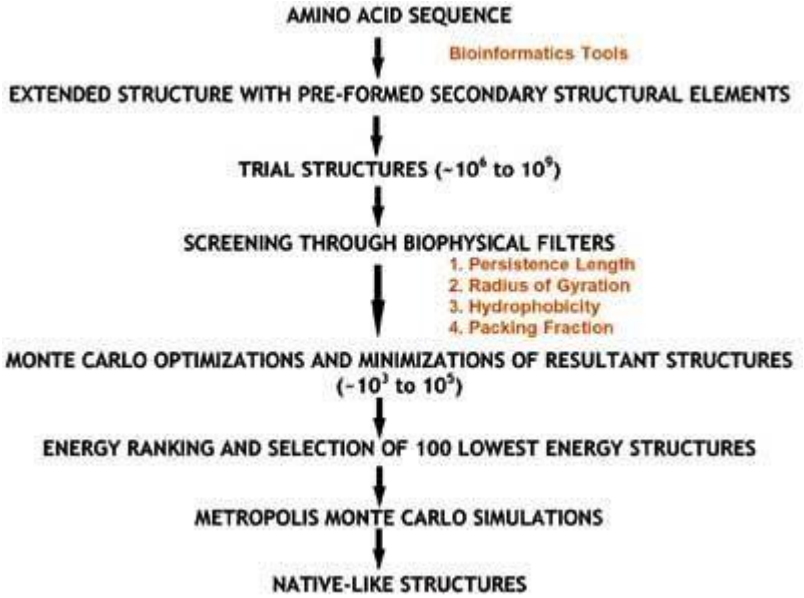


Figure 65

Ab initio, or de novo approaches predict a protein structure and folding mechanism from knowledge only of its amino acid sequence. Often the term ab initio is interpreted as applied to an algorithm based entirely on physico-chemical interactions. On the other hand, the most successful ab initio methods utilize information from the sequence and structural databases in some form. Basic idea of an ab initio algorithm: search for the native state which is presumably in the minimum energy conformation. Usually an ab initio algorithm consists of multiple steps with different levels of approximated modeling of protein structure.

For a consideration of side chains in ab initio predictions, a so-called united residue approximation (UNRES) is frequently used:

- Side chains are represented by spheres ("side-chain centroids", SC). Each centroid represents all the atoms belonging to a real side chain. A van der Waals radius is introduced for every residue type.

- A polypeptide chain is represented by a sequence of Cα atoms with attached SCs and peptide group centers (p) centered between two consecutive Cα atoms.

- The distance between successive Cα atoms is assigned a value of 3.8 Å (a virtual-bond length, characteristic of a planar trans peptide group CO-NH).

- It is assumed that Cα - Cα - Cα virtual bond angles have a fixed value of 90° (close to what is observed in crystal structures). - The united side chains have fixed geometry, with parameters being taken from crystal data.

The only variables in this model of protein conformation are virtual-bond torsional angles γ.

The energy function for the simplified chain can be represented as the sum of the hydrophobic, hydrophilic and electrostatic interactions between side chains and peptide groups (potential functions dependent on the nature of interactions, distances and dimensions of side chains). The parameters in the expressions for contact energies are estimated empirically from crystal structures and all-atom calculations.

An example of the algorithm for structure prediction using UNRES:

1. Low-energy conformations in UNRES approximation are searched using Monte Carlo energy minimization. A cluster analysis is then applied to divide the set of low-energy conformations whose lowest-energy representatives are hereafter referred to as structures. Structures having energies within a chosen cut-off value above the lowest energy structure are saved for further stages of the calculation.

2. These virtual-bond united-residue structures are converted to an all-atom backbone (preserving distances between α-carbons).

3. Generation of the backbone is completed by carrying out simulations in a "hybrid" representation of the polypeptide chain, i.e. with an all-atom backbone and united side chains (still subject to the constraints following the UNRES simulations, so that some or even all the distances of the virtual-bond chain are substantially preserved). The simulations are performed by a Monte Carlo algorithm.

4. Full (all-atom) side chains are introduced with accompanying minimization of steric overlaps, allowing both the backbone and side chains to move. Then Monte

Carlo simulations explore conformational space in the neighborhood of each of the low-energy structures.

Monte Carlo algorithms start from some (random) conformation and proceed with (quasi)randomly introduced changes, such as rotations around a randomly selected bond. If the change improves energy value, it is accepted. If not, it may be accepted with a probability dependent on energy increase. The procedure is repeated with a number of iterations, leading to lower energy conformations. A function defining higher energy acceptance probability is usually constructed 25 with a parameter that leads to lower probabilities in the course of simulation ("cooling down" the simulation) in order to achieve convergence and stop the algorithm.

Combinations of approaches

Many of the modern packages for protein structure predictions attempt to combine various approaches, algorithms and features. One of the most successful examples is Rosetta - ab initio prediction using database statistics.

Rosetta is based on a picture of protein folding in which local sequence fragments (3-9 residues) rapidly alternate between different possible local structures. The distribution of conformations sampled by an isolated chain segment is approximated by the distribution adopted by that sequence segment and related sequence segments in the protein structure database. Thus the algorithm combines both ab initio and fold recognition approaches.

Folding occurs when the conformations and relative orientations of the local segments combine to form low energy global structures. Local conformation are sampled from the database of structures and scored using Bayesian logic:

P(structure | sequence) = P(structure) x P(sequence | structure) / P(sequence).

For comparisons of different structures for a given sequence, P(sequence) is constant. P(structure) may be approximated by some general expression favouring more compact structures. P(sequence | structure) is derived from the known structures in the database by assumptions somewhat similar to those used in fold recognition, for instance by estimating probabilities for pairs of amino acids to be at particular distance and computing the probability of sequence as the product over

all pairs).

Non-local interactions are optimized by a Monte Carlo search through the set of conformations that can be built from the ensemble of local structure fragments.

In the standard Rosetta protocol, an approximated protein representation is used: backbone atoms are explicitly included, but side chains are represented by centroids (so-called low- resolution refinement of protein structure). The low-resolution step can be followed by high- resolution refinement, with all-atom protein representation. Similar stepwise refinement protocols can be used to improve predictions yielded by other methods, for instance, in loops (variable regions) of homology-modeling structures.

In recent CASP experiments (Critical Assessment of Structure Prediction), the Rosetta approach turned out to be one of the most successful prediction methods in the novel fold category. Obviously, none of prediction approaches is ideal. Therefore it is reasonable to try to combine the best features of many different procedures or to derive a consensus, meta- prediction. For instance, the 3D-Jury system generated meta predictions using models produced by a set of servers. The algorithm scored various models according to their similarities to each other.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF BIO AND CHEMICAL ENGINEERING

DEPARTMENT OF BIOTECHNOLOGY

Unit 5 – Bioinformatics – SBB3201

V. EXPLORING BIOLOGICAL INFORMATION

GENE PREDICTION

Gene prediction by computational methods for finding the location of protein coding regions is one of the essential issues in bioinformatics.

Gene prediction basically means locating genes along a genome. Also called gene finding, it refers to the process of identifying the regions of genomic DNA that encode genes.

This includes protein coding genes, RNA genes and other functional elements such as the regulatory genes.

Importance of Gene Prediction

Helps to annotate large, contiguous sequences

Aids in the identification of fundamental and essential elements of genome such as functional genes, intron, exon, splicing sites, regulatory sites, gene encoding known proteins, motifs, EST, ACR, etc.

Distinguish between coding and non-coding regions of a genome

Predict complete exon – intron structures of protein coding regions

Describe individual genes in terms of their function

It has vast application in structural genomics, functional genomics, metabolomics, transcriptomics, proteomics, genome studies and other genetic related studies including genetics disorders detection, treatment and prevention.

Bioinformatics and the Prediction of Genes

With databases of human and model organism DNA sequences increasing quickly with time, it has become almost impossible to carry out the conventional painstaking experimentation on living cells and organisms to predict genes.

Formerly, statistical analysis of the rates of homologous recombination of several different genes could determine their order on a certain chromosome, and information from many such experiments could be combined to create a genetic map specifying the rough location of known genes relative to each other.

However, today, the frontiers of bioinformatics research are making it increasingly possible to predict the function of such a deluge of genes based on its sequence alone.

Methods of Gene Prediction

Two classes of methods are generally adopted:

A. Similarity based searches

It is a method based on sequence similarity searches.

It is a conceptually simple approach that is based on finding similarity in gene sequences between ESTs (expressed sequence tags), proteins, or other genomes to the input genome.

This approach is based on the assumption that functional regions (exons) are more conserved evolutionarily than nonfunctional regions (intergenic or intronic regions).

Once there is similarity between a certain genomic region and an EST, DNA, or protein, the similarity information can be used to infer gene structure or function of that region.

Local alignment and global alignment are two methods based on similarity searches. The most common local alignment tool is the BLAST family of programs, which detects sequence similarity to known genes, proteins, or ESTs.

Two more types of software, PROCRUSTES and GeneWise, use global alignment of a homologous protein to translated ORFs in a genomic sequence for gene prediction.

A new heuristic method based on pairwise genome comparison has been implemented in the software called CSTfinder.

B. Ab- initio prediction

It is a method based on gene structure and signal-based searches.

It uses gene structure as a template to detect genes

Ab initio gene predictions rely on two types of sequence information: signal sensors and content sensors.

Signal sensors refer to short sequence motifs, such as splice sites, branch points, polypyrimidine tracts, start codons and stop codons.

On the other hand content sensors refer to the patterns of codon usage that are unique to a species, and allow coding sequences to be distinguished from the surrounding non-coding sequences by statistical detection algorithms. Exon detection must rely on the content sensors.

The search by this method thus relies on the major feature present in the genes.

Many algorithms are applied for modeling gene structure, such as Dynamic Programming, linear discriminant analysis, Linguist methods, Hidden Markov Model and Neural Network.

Based on these models, a great number of ab initio gene prediction programs have been developed. Some of the frequently used ones are GeneID, FGENESH, GeneParser, GlimmerM, GENSCAN etc.

World Wide Web

What is the Internet? What is the World Wide Web? How are they related?

The Internet is an international network (a collection of connected, in this case, computers) – networked for the purpose of communication of information. The Internet offers many software services for this purpose, including:

- World Wide Web
- E-mail
- Instant messaging, chat
- Telnet (a service that lets a user login to a remote computer that the user has login privileges for)
- FTP (File Transfer Protocol) – a service that lets one use the Internet to copy files from one computer to another

The Web was originally designed for the purpose of displaying "public domain" data to anyone who could view it. Although this is probably the most popular use of the Web today, other uses of the Web include:

- Research, using tools such as "search engines" to find desired information.
- A variety of databases are available on the Web (this is another "research" tool). One example of such a database: a library's holdings.
- Shopping – most sizable commercial organizations have Web sites with forms you can fill out to specify goods or services you wish to purchase. Typically, you must include your credit card information in this form. Typically, your credit card information is safe – the system is typically automated so no human can see (and steal) your credit card number.
- We can generalize the above: Web forms can be filled out and submitted to apply for admission to a university, to give a donation to a charity, to apply for a job, to become a member of an organization, do banking chores, pay bills, etc.

- Listen to music or radio-like broadcasts, view videos or tv-like broadcasts.
- Some use the Web to access their e-mail or bulletin board services such as Blackboard.
- Most "browsers" today are somewhat like operating systems, in that they can enable a variety of application programs. For example, a Word, Excel, PowerPoint document can be placed on the Web and viewed in its "native" application.

Some terminology you should know:

- **Browser:** A program used to view Web documents. Popular browsers include Microsoft Internet Explorer (IE), Netscape, Opera; an old text-only browser called *Lynx* is still around on some systems; etc. The browsers of Internet Service Providers (ISPs) like AOL, Adelphia, Juno, etc., are generally one of the above, with the ISP's logo displayed. Most browsers work alike, today. There may be minor (for example, what IE calls "Favorites," Netscape calls "Bookmarks").
- **A Web document is called a "page."** A collection of related pages is a "site." A Web site typically has a "home page" designed to be the first, introductory, page a user of the site views.
- A Web page typically has an "address" or URL (Universal Reference Locator). You can view a desired page by using any of several methods to inform your browser of the URL whose page you wish to view. The home page of a site typically has a URL of the form
  http://www.DomainName.suffix

  where the "DomainName" typically tells you something about the identity of the "host" or "owner" of the site, and the "suffix" typically tells either the type of organization of the owner or its country. Some common suffixes include:

  - ✓ edu – An educational institution, usually a college or university.
  - ✓ com – A commercial site – a company
  - ✓ gov – a government site
  - ✓ org – an organization that's non-profit
  - ✓ net – an alternative to "com" for network service providers

  Also, the Internet originally was almost entirely centered in the US. As it spread to other countries, it became common for sites outside the US to use a suffix that's a 2-letter country abbreviation: "ca" (without quotation marks) for Canada; "it" for Italy; "mx" for Mexico; etc.

A page that isn't a home page will typically have an address that starts with its site's home page address, and has appended further text to describe the page. For example, the Niagara University home page is at http://www.niagara.edu/ and the Niagara University Academics page is at http://www.niagara.edu/academic.htm.

Navigating:

- One way to reach a desired page is to enter its URL in the "Address" textbox.
- You can click on a *link* (usually underlined text, or a graphic may also serve as a link; notice that the mouse cursor changes its symbol, typically to a hand, when hovering over a Web link) to get to the page addressed by the link.
- The Back button may be used to retrace your steps, revisiting pages recently visited.
- You can click the Forward button to retrace your steps through pages recently Backed out of.
- Notice the drop-down button at the right side of the Address textbox. This reveals a menu of URLs recently visited by users of the browser on the current computer. You may click one of these URLs to revisit its page.
- Favorites (what Netscape calls "Bookmarks") are URLs saved for the purpose of making revisits easy. If you click a Favorite, you can easily revisit the corresponding page.

How do we find information on the Web? Caution: Don't believe everything you see on the Web. Many Web sites have content made up of hate literature, political propaganda, unfounded opinions, and other content of dubious reliability. Therefore, you should try to use good judgment about the sites you use for research.

Strategies for finding information on the Web include:

- Often, you can make an intelligent guess at the URL of a desired site. For example, you might guess the UB Web site is http://www.ub.edu (turned out to be the University of Barcelona) or http://www.buffalo.edu (was correct); similarly, if you're interested in the IRS Web site, you might try http://www.irs.gov – and it works. Similarly, you might try, for Enron, we might try http://www.enron.com – and this redirected us to the page http://www.enron.com/corp/.
- "Search engines" are Web services provided on a number of Web sites, allowing you to enter a keyword or phrase describing the topic you want information for. You may then

click a button to activate the search. A list of links typically appears, and you may explore these links to find (you hope) the information you want. Note: if you use a phrase of multiple words, and don't place that phrase in quotation marks, you may get links by virtue of matching all the words separately – *e.g.*, "Diane" and "Pilarski" separately appeared in a document that matched the phrase "Diane Pilarski" without quotation marks; but the same link did not appear when we searched for "Diane Pilarski" with quotation marks. Also, you may find if the phrase you enter is someone's name, that many people have the same name.

- Another strategy: Some Web sites (including some that offer search engines) have "Web directories" or "indices" – classifications of Web pages. A good example: The Yahoo! site at http://www.yahoo.com has such a Web directory. You can work your way through the directory, often, to find desired information.

WEB BROWSER

A web browser (commonly referred to as a browser) is a software application for accessing information on the World Wide Web. Each individual web page, image, and video is identified by a distinct Uniform Resource Locator (URL), enabling browsers to retrieve these resources from a web server and display them on a user's device.

A web browser is not the same thing as a search engine, though the two are often confused. For a user, a search engine is just a website, such as google.com, that stores searchable data about other websites. But to connect to a website's server and display its web pages, a user must have a web browser installed on their device.

As of March 2019, more than 4.3 billion people use a browser, which is about 55% of the world's population.

The most popular browsers are Chrome, Firefox, Safari, Internet Explorer, and Edge.

History

The first web browser, called WorldWideWeb, was created in 1990 by Sir Tim Berners-Lee. He then recruited Nicola Pellow to write the Line Mode Browser, which displayed web pages on dumb terminals; it was released in 1991.

Nicola Pellow and Tim Berners-Lee in their office at CERN.

Marc Andreessen, lead developer of Mosaic and Navigator

1993 was a landmark year with the release of Mosaic, credited as "the world's first popular browser". Its innovative graphical interface made the World Wide Web system easy to use and thus more accessible to the average person. This, in turn, sparked the Internet boom of the 1990s when the Web grew at a very rapid rate. Marc Andreessen, the leader of the Mosaic team, soon started his own company, Netscape, which released the Mosaic-influenced Netscape Navigator in 1994. Navigator quickly became the most popular browser.

Microsoft debuted Internet Explorer in 1995, leading to a browser war with Netscape. Microsoft was able to gain a dominant position for two reasons: it bundled Internet Explorer with its popular Microsoft Windows operating system and did so as freeware with no restrictions on usage. Eventually the market share of Internet Explorer peaked at over 95% in 2002.

WorldWideWeb was the first web browser.

In 1998, desperate to remain competitive, Netscape launched what would become the Mozilla Foundation to create a new browser using the open source software model. This work evolved into Firefox, first released by Mozilla in 2004. Firefox reached a 28% market share in 2011.

Apple released its Safari browser in 2003. It remains the dominant browser on Apple platforms, though it never became a factor elsewhere.

The last major entrant to the browser market was Google. Its Chrome browser, which debuted in 2008, has been a huge success. It steadily took market share from Internet Explorer and became the most popular browser in 2012. Chrome has remained dominant ever since.

In terms of technology, browsers have greatly expanded their HTML, CSS, JavaScript, and multimedia capabilities since the 1990s. One reason has been to enable more sophisticated websites, such as web applications. Another factor is the significant increase of broadband connectivity, which enables people to access data-intensive web content, such as YouTube streaming, that was not possible during the era of dial-up modems.

Function

The purpose of a web browser is to fetch information resources from the Web and display them on a user's device.

This process begins when the user inputs a URL, such as https://en.wikipedia.org/, into the browser. Virtually all URLs on the Web start with either http: or https: which means the browser will retrieve them with the Hypertext Transfer Protocol. In the case of https:, the communication between the browser and the web server is encrypted for the purposes of security and privacy. Another URL prefix is file: which is used to display local files already stored on the user's device.

Once a web page has been retrieved, the browser's rendering engine displays it on the user's device. This includes image and video formats supported by the browser.

Web pages usually contain hyperlinks to other pages and resources. Each link contains a URL, and when it is clicked, the browser navigates to the new resource. Thus the process of bringing content to the user begins again.

Settings

Web browsers can typically be configured with a built-in menu. Depending on the browser, the menu may be named Settings, Options, or Preferences.

The menu has different types of settings. For example, users can change their home page and default search engine. They also can change default web page colors and fonts. Various network connectivity and privacy settings are also usually available.

Privacy

During the course of browsing, cookies received from various websites are stored by the browser. Some of them contain login credentials or site preferences. However, others are used for tracking user behavior over long periods of time, so browsers typically provide settings for removing cookies when exiting the browser.Finer-grained management of cookies requires a browser extension.

Features

The most popular browsers have a number of features in common. They allow users to set bookmarks and browse in a private mode. They also can be customized with extensions, and some of them provide a sync service.

Most browsers have these user interface features:

Allow the user to open multiple pages at the same time, either in different browser windows or in different tabs of the same window.

Back and forward buttons to go back to the previous page visited or forward to the next one.

A refresh or reload button to reload the current page.

A stop button to cancel loading the page. (In some browsers, the stop button is merged with the reload button.)

A home button to return to the user's home page.

An address bar to input the URL of a page and display it.

A search bar to input terms into a search engine. (In some browsers, the search bar is merged with the address bar.)

There are also niche browsers with distinct features. One example is text-only browsers that can benefit people with slow Internet connections or those with visual impairments.

Security

Web browsers are popular targets for hackers, who exploit security holes to steal information, destroy files, and other malicious activity. Browser vendors regularly patch these security holes, so users are strongly encouraged to keep their browser software updated. Other protection measures are antivirus software and avoiding known-malicious websites.

INTERNET

The **Internet** is a global system of interconnected computer networks that use the standard Internet protocol suite (TCP/ IP) to serve billions of users worldwide. It is a *network of networks* that consists of millions of private, public, academic, business, and government networks, of local to global scope, that are linked by a broad array of electronic, wireless and optical networking technologies. The Internet carries a vast range of information resources and services, such as the inter- linked hypertext documents of the World Wide Web (WWW) and the infrastructure to support electronic mail.

Uses of Internet

Internet has been the most useful technology of the modern times which helps us not only in our daily lives, but also our personal and professional lives developments. The internet helps us achieve this in several different ways.

For the students and educational purposes the internet is widely used to gather information so as to do the research or add to the knowledge of various subjects. Even the business professionals and the professionals like doctors, access the internet to filter the necessary information for their use. The internet is therefore the largest encyclopedia for everyone, in all age categories. The internet has served to be more useful in maintaining contacts with friends

and relatives who live abroad permanently.

Advantages of Internet:

E-mail: Email is now an essential communication tools in business. With e-mail you can send and receive instant electronic messages, which works like writing letters. Your messages are delivered instantly to people anywhere in the world, unlike traditional mail that takes a lot of time. Email is free, fast and very cheap when compared to telephone, fax and postal services.

24 hours a day - 7 days a week: Internet is available, 24x7 days for usage.

Information: Information is probably the biggest advantage internet is offering. There is a huge amount of information available on the internet for just about every subject, ranging from government law and services, trade fairs and conferences, market information, new ideas and technical support. You can almost find any type of data on almost any kind of subject that you are looking for by using search engines like google, yahoo, msn, etc.

Online Chat: You can access many 'chat rooms' on the web that can be used to meet new people, make new friends, as well as to stay in touch with old friends. You can chat in MSN and yahoo websites.

Services: Many services are provided on the internet like net banking, job searching, purchasing tickets, hotel reservations, guidance services on array of topics engulfing every aspect of life.

Communities: Communities of all types have sprung up on the internet. Its a great way to meet up with people of similar interest and discuss common issues.

E-commerce: Along with getting information on the Internet, you can also shop online. There are many online stores and sites that can be used to look for products as well as buy them using your credit card. You do not need to leave your house and can do all your shopping from the convenience of your home. It has got a real amazing and wide range of products from household needs, electronics to entertainment.

Entertainment: Internet provides facility to access wide range of Audio/ Video songs, plays films. Many of which can be downloaded. One such popular website is YouTube.

Software Downloads: You can freely download innumerable, softwares like utilities, games, music, videos, movies, etc from the Internet.

Limitations of Internet

Theft of Personal information: Electronic messagessent over the Internet can be easily snooped and tracked, revealing who is talking to whom and what they are talking about. If you use the Internet, your personal information such as your name, address, credit card, bank details and other information can be accessed by unauthorized persons. If you use a credit card or internet banking for online shopping, then your details can also be 'stolen'.

Negative effects on family communication: It is generally observed that due to more time spent on Internet, there is a decrease in communication and feeling of togetherness among the family members.

Internet addiction: There is some controversy over whether it is possible to actually be addicted to the Internet or not. Some researchers, claim that it is simply people trying to escape their problems in an online world.

Children using the Internet has become a big concern. Most parents do not realize the dangers involved when their children log onto the Internet. When children talk to others online, they do not realize they could actually be talking to a harmful person. Moreover, pornography is also a very serious issue concerning the Internet, especially when it comes to young children. There are thousands of pornographic sites on the Internet that can be easily found and can be a detriment to letting children use the Internet.

Virus threat: Today, not only are humans getting viruses, but computers are also. Computers are mainly getting these viruses from the Internet. Virus is is a program which disrupts the normal functioning of your computer systems. Computers attached to internet are more prone to virus attacks and they can end up into crashing your whole hard disk.

Spamming: It is often viewed as the act of sending unsolicited email. This multiple or vast emailing is often compared to mass junk mailings. It needlessly obstruct the entire system. Most spam is commercial advertising, often for dubious products, get-rich-quick schemes, or quasi-legal services. Spam costs the sender very little to send — most of the costs are paid for by the recipient or the carriers rather than by the sender

SERVICES OF INTERNET - E-mail, FTP, Telnet

Email, discussion groups, long-distance computing, and file transfers are some of the important services provided by the Internet. Email is the fastest means of communication. With email one can also send software and certain forms of compressed digital image as an attachment. News groups or discussion groups facilitate Internet user to join for various kinds of debate, discussion and news sharing. Long-distance computing was an original inspiration for development of ARPANET and does still provide a very useful service on Internet. Programmers can maintain accounts on distant, powerful computers and execute programs. File transfer service allows Internet users to access remote machines and retrieve programs, data or text.

E-Mail (Electronic Mail)

E-mail or Electronic mail is a paperless method of sending messages, notes or letters from one person to another or even many people at the same time via Internet. E-mail is very fast compared to the normal post. E-mail messages usually take only few seconds to arrive at their destination. One can send messages anytime of the day or night, and, it will get delivered immediately. You need not to wait for the post office to open and you don't have to get worried about holidays. It works 24 hours a day and seven days a week. What's more, the copy of the message you have sent will be available whenever you want to look at it even in the middle of the night. You have the privilege of sending something extra such as a file, graphics, images etc. along with your e-mail. The biggest advantage of using e- mail is that it is cheap, especially when sending messages to other states or countries and at the same time it can be delivered to a number of people around the world.

It allows you to compose note, get the address of the recipient and send it. Once the mail is received and read, it can be forwarded or replied. One can even store it for later use, or delete. In e-mail even the sender can request for delivery receipt and read receipt from the recipient.

Features of E-mail:

- One-to-one or one-to-many communications
- Instant communications

- Physical presence of recipient is not required
- Most inexpensive mail services, 24-hours a day and seven days a week
- Encourages informal communications

Components of an E-mail Address

As in the case of normal mail system, e-mail is also based upon the concept of a recipient address. The email address provides all of the information required to get a message to the recipient from any where in the world. Consider the e-mail ID.

john@hotmail.com

In the above example john is the username of the person who will be sending/ receiving the email. Hotmail is the mail server where the username john has been registered and com is the type of organization on the internet which is hosting the mail server.

FTP (File Transfer Protocol)

File Transfer Protocol, is an Internet utility software used to uploaded and download files. It gives access to directories or folders on remote computers and allows software, data and text files to be transferred between different kinds of computers. FTP works on the basis of same principle as that of Client/ Server. FTP "Client" is a program running on your computer that enables you to communicate with remote computers. The FTP client takes FTP command and sends these as requests for information from the remote computer known as FTP servers. To access remote FTP server it is required, but not necessary to have an account in the FTP server. When the FTP client gets connected, FTP server asks for the identification in terms of User Login name and password of the FTP client. If one does not have an account in the remote FTP server, still he can connect to the server using anonymous login.

Using anonymous login anyone can login in to a FTP server and can access public archives; anywhere in the world, without having an account. One can easily Login to the FTP site with the username anonymous and e-mail address as password.

Objectives of FTP:

- Provide flexibility and promote sharing of computer programs, files and data
- Transfer data reliably and more efficiently over network
- Encourage implicit or indirect use of remote computers using Internet
- Shield a user from variations in storage systems among hosts.

The basic steps in an FTP session

Start up your FTP client, by typing ftp on your system's command line/ 'C>' prompt (or, if you are in a Windows, double-click on the FTP icon).

Give the FTP client an address to connect. This is the FTP server address to which the FTP client will get connected

Identify yourself to the FTP remote site by giving the Login Name

Give the remote site a password

Remote site will verify the Login Name/ Password to allow the FTP client to access its files

Look directory for files in FTP server

Change Directories if required

Set the transfer mode (optional);

Get the file(s) you want, and

Quit.

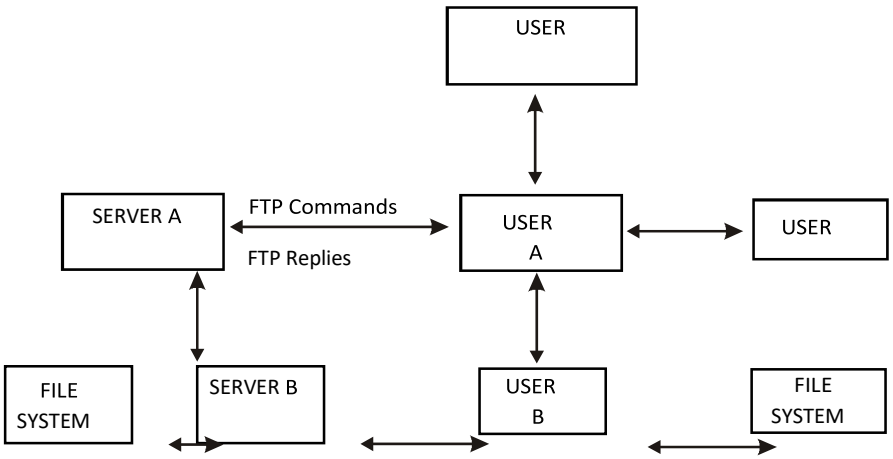


Figure 1

Telnet (Remote Computing)

Telnet or remote computing is telecommunication utility software, which uses available telecommunication facility and allows you to become a user on a remote computer. Once you gain access to remote computer, you can use it for the intended purpose. The TELNET works in a very step by step procedure. The commands typed on the client computer are sent to the local Internet Service Provider (ISP), and then from the ISP to the remote computer that you have gained access. Most of the ISP provides facility to TELENET into your own account from another city and check your e-mail while you are travelling or away on business.

The following steps are required for a TELNET session

- Start up the TELNET program
- Give the TELNET program an address to connect (some really nifty TELNET packages allow you to combine steps 1 and 2 into one simple step)
- Make a note of what the "escape character" is
- Log in to the remote computer,
- Set the "terminal emulation"
- Play around on the remote computer, and
- Quit.

EMBnet

The European Molecular Biology network (EMBnet) is an international scientific network and interest group that aims to enhance bioinformatics services by bringing together bioinformatics expertises and capacities. On 2011 EMBnet has 37 nodes spread over 32 countries. The nodes include bioinformatics related university departments, research institutes and national service providers.

Operations

The main task of most EMBnet nodes is to provide their national scientific community with access to bioinformatics databanks, specialised software and sufficient computing resources and expertise. EMBnet is also working in the fields of bioinformatics training and software development. Examples of software created by EMBnet members are: EMBOSS, wEMBOSS, UTOPIA.

EMBnet represents a wide user group and works closely together with the database producers such as EMBL's European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (Swiss-Prot), the Munich Information Center for Protein Sequences (MIPS), in order to provide a uniform coverage of services throughout Europe. EMBnet is registered in the Netherlands as a public foundation (Stichting).

Since its creation in 1988, EMBnet has evolved from an informal network of individuals in charge of maintaining biological databases into the only worldwide organization bringing bioinformatics professionals to work together to serve the expanding fields of genetics and molecular biology. Although composed predominantly of academic nodes, EMBnet gains an important added dimension from its industrial members. The success of EMBnet is attracting increasing numbers of organizations outside Europe to join.

EMBnet has a tried-and-tested infrastructure to organise training courses, give technical help and help its members effectively interact and respond to the rapidly changing needs of biological research in a way no single institute is able to do.

In 2005 the organization created additional types of node to allow more than one member per country. The new category denomination is "associated node".

Coordination and organization

EMBnet is governed by the Annual General Meetings (AGM), and is coordinated by an Executive Board (EB) that oversees the activities of three project committees:

Education and Training committee (E&T). Educational support includes a series of courses organised in the member countries and languages, the committee works as well on the continued development of on-line accessible education materials.

Publicity and Public Relations committee (P&PR). This committee is responsible for promoting any type of EMBnet activities, for the advertisement of products and services provided by the EMBnet community, as well as for proposing and developing new strategies aiming to enhance EMBnet's visibility, and to take care of public relationships with EMBnet communities and related networks/societies.

Technical Manager committee (TM). The TM PC provides assistance and practical help to the participating nodes and their users.

THE NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION (NCBI)

The National Center for Biotechnology Information (NCBI) is part of the United States National Library of Medicine (NLM), a branch of the National Institutes of Health (NIH). The NCBI is located in Bethesda, Maryland and was founded in 1988 through legislation sponsored by Senator Claude Pepper.

The NCBI houses a series of databases relevant to biotechnology and 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. Other databases include the NCBI Epigenomics database. All these databases are available online through the Entrez search engine. NCBI was directed by David Lipman, one of the original authors of the BLAST sequence alignment program and a widely respected figure in bioinformatics. He also led an intramural research program, including groups led by Stephen Altschul (another BLAST co-author), David Landsman, Eugene Koonin, John Wilbur, Teresa Przytycka, and Zhiyong Lu. David Lipman stood down from his post in May 2017.

GenBank

NCBI has had responsibility for making available the GenBank DNA sequence database since 1992.GenBank coordinates with individual laboratories and other sequence databases such as those of the European Molecular Biology Laboratory (EMBL) and the DNA Data Bank of Japan (DDBJ).

Since 1992, NCBI has grown to provide other databases in addition to GenBank. NCBI provides Gene, Online Mendelian Inheritance in Man, the Molecular Modeling Database (3D protein structures), dbSNP (a database of single-nucleotide polymorphisms), the Reference Sequence Collection, a map of the human genome, and a taxonomy browser, and coordinates with the National Cancer Institute to provide the Cancer Genome Anatomy Project. The NCBI assigns a unique identifier (taxonomy ID number) to each species of organism.

The NCBI has software tools that are available by WWW browsing or by FTP. For example, BLAST is a sequence similarity searching program. BLAST can do sequence comparisons against the GenBank DNA database in less than 15 seconds.

NCBI Bookshelf

The "NCBI Bookshelf is a collection of freely accessible, downloadable, on-line versions of selected biomedical books. The Bookshelf covers a wide range of topics including molecular biology, biochemistry, cell biology, genetics, microbiology, disease states from a molecular and cellular point of view, research methods, and virology. Some of the books are online versions of previously published books, while others, such as Coffee Break, are written and edited by NCBI staff. The Bookshelf is a complement to the Entrez PubMed repository of peer-reviewed publication abstracts in that Bookshelf contents provide established perspectives on evolving areas of study and a context in which many disparate individual pieces of reported research can be organized.

Basic Local Alignment Search Tool (BLAST)

BLAST is an algorithm used for calculating sequence similarity between biological sequences such as nucleotide sequences of DNA and amino acid sequences of proteins. BLAST is a powerful tool for finding sequences similar to the query sequence within the same organism or in different organisms. It searches the query sequence on NCBI databases and servers and post the results back to the person's browser in chosen format.

Input sequences to the BLAST are mostly in FASTA or Genbank format while output could be delivered in variety of formats such as HTML, XML formatting and plain text. HTML is the default output format for NCBI's web-page. Results for NCBI-BLAST are presented in graphical format with all the hits found, a table with sequence identifiers for the hits having scoring related data, along with the alignments for the sequence of interest and the hits received with analogous BLAST scores for these

Entrez

The Entrez Global Query Cross-Database Search System is used at NCBI for all the major databases such as Nucleotide and Protein Sequences, Protein Structures, PubMed, Taxonomy, Complete Genomes, OMIM, and several others. Entrez is both indexing and retrieval system having data from various sources for biomedical research. NCBI distributed the first version of Entrez in 1991, composed of nucleotide sequences from PDB and GenBank, protein sequences from SWISS-PROT, translated GenBank, PIR, PRF , PDB and associated abstracts and citations from PubMed. Entrez is specially designed to integrate the data from several different sources, databases and formats into a uniform information model and retrieval system which can efficiently retrieve that relevant references, sequences and structures.