Semester II

18BCP204 **BIOINFORMATICS** **4H-4C**

Instruction hours / week: L: 4 T:0 P:0 **Marks:** Internal: **40** External: **60** Total: **100**
End Semester Exam: 3 Hours

Course objectives

To make the students

- To make students understand the essential features of the interdisciplinary field of science for better understanding the biological data.
- To create students opportunity to interact with algorithms, tools and data in current scenario.
- To make the students look at a biological problem from a computational point of view.
- To find out the methods for analyzing the expression, structure and function of proteins, and understanding the relationships between species.

Course outcomes (CO's)

1. The student will choose biological data, submission and retrieval from databases.
2. The students will be able to experiment pair wise and multiple sequence alignment and will analyze the secondary and tertiary structures of protein sequences.
3. The student will understand the data structure (databases) used in bioinformatics and interpret the information (especially: find genes; determine their functions), understand and be aware of current research and problems relating to this area.

UNIT I: Concepts of Bioinformatics

Definition, concepts of Bioinformatics: Objectives, History of Bioinformatics, Milestones, Genome sequencing projects, Human Genome Project- Science, applications and ELSI.
Introduction to Biological databases: Types of databases, sequence databases-nucleic acid sequence databases, GenBank, protein sequence database, Swiss-Prot, PIR, motif database-PROSITE, structural databases, bibliographic databases and organism specific databases-GMOD- Searching and retrieval of data-Entrez and SRS.

UNIT II: Sequence Alignment

Introduction to sequence Alignment: Pairwise and multiple sequence alignment, substitution matrices, Dynamic programming algorithms-Needleman and Wunsch and Smith-Waterman, Similarity searching programs, BLAST, FASTA, Multiple sequence alignment – CLUSTAL, Introduction and application to phylogenetic trees, basic terminologies, Phylogenetic analysis-PHYLIP theory of phylogeny, tree building methods.

UNIT III: Protein prediction strategies and programs

Protein Secondary Structure Prediction, three dimensional structure prediction-Comparative modeling, threading, Concepts of Molecular modeling, Model refinement, evaluation of the model, protein folding and visualization of molecules – Visualization tools-RasMol, Deep View.

UNIT IV: Gene Identification and Prediction
Genome sequencing, Genome database-SWISS-2D PAGE database, Gene Mark, Gene Scan, Pattern Recognition, Global gene expression studies-DNA Micro array.

UNIT V: Applications of Bioinformatics
Applications of Bioinformatics-Molecular medicine, biotechnology, agricultural, Computer Aided Drug Designing-structure and ligand based drug designing, ADME profiles, QSAR. receptors, docking, Introduction to molecular dynamics simulation.

SUGGESTED READINGS

1. Lesk, A.M., (2014). Introduction to Bioinformatics, 4th edition. Oxford University Press, Oxford.

2. Attwood, K., and Parry-Smith, J., (2003). Introduction to Bioinformatics, Pearson Education, Singapore.

3. Baxevanis, A.D., and Quellette, B.F.F., (2001). Practical Guide to the Analysis of Genes and Proteins, John Wiley & Sons, New York.

4. Mount, D.W., 2013. Bioinformatics: Sequence and Genome Analysis. 2nd edition, Cold Spring Harbour Laboratory Press, New York.

5. Ignacimuthu, S., (2013). Basic Bioinformatics, 2nd edition Alpha Science Intl Ltd Chennai.

6. Rastogi, S.C., Mendiratta, N and Rastogi, P., (2004). Bioinformatics – Concepts, Skills, Applications. CBS Publishers & Distributiors, New Delhi.

7. Rastogi S.C and Mendiratta, N., (2006). Bionformatics Methods and applications Genomics, Proteomics and Drug Discovery 2nd Edition, Parag Rastogi Publication, India.

8. Sundararajan, S., and Balaji, R., (2003). Introduction to Bioinformatics, Himalaya Publishing House, Mumbai.

KARPAGAM ACADEMY OF HIGHER EDUCATION

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

Coimbatore – 641 021.

LECTURE PLAN
DEPARTMENT OF BIOCHEMISTRY

STAFF NAME: Dr. S. RAJAMANIKANDAN
SUBJECT NAME: BIOLOGICAL DATABASES AND ANALYSIS
SUB.CODE:18BCP212
SEMESTER: II CLASS: I M.Sc (BC)

S.NO	NAME OF THE EXPERIMENT	SUPPORT MATERIALS
1.	Biological databanks, sequence databases, structure databases, specialized databases	T1:117-139
2.	Data base file formats	T1:117-139
3.	Data retrieval tools and methods (PUBMED, ENTREZ, SRS)	T1:117-139
4.	Sequence similarity searching (NCBI-BLAST, FASTA)	T1:134-144; 145-146
5.	Protein sequence analysis (ExPASY proteomics tools)	W1
6.	Multiple sequence alignment (Clustal-W)	W2
7.	Gene structure and function prediction (using ORF finder, Genscan, Genemark)	W3-W5
8.	Molecular phylogeny (PHYLIP)	W6
9.	Sequence analysis using EMBOSS	W7
10.	Protein structure visualization-RASMOL (menu function and command line entries), Deep view	T2 (201-211, 195-196)

REFERENCES	
T1	Arthur M. Lesk, (2005). Introduction to Bioinformatics, 2nd edition, Published by Oxford University Press, New Delhi-110001.
T2	Mani K., Vijayaraj N, (2002). Bioinformatics for Beginers, Kalaikathir Achchagam, Coimbatore.
W2	https://web.expasy.org/protparam/
W3	https://www.ebi.ac.uk/Tools/msa/clustalw2/
W4	https://www.ncbi.nlm.nih.gov/orffinder/
W5	http://genes.mit.edu/GENSCAN.html
W6	http://exon.gatech.edu/GeneMark/
W7	http://evolution.genetics.washington.edu/phylip.html
W8	https://www.ebi.ac.uk/Tools/emboss/

UNIT-I

SYLLABUS

Definition, concepts of Bioinformatics: Objective, History of Bioinformatics, Milestones, Genome sequencing projects, Human Genome Projects-Science, applications and ELSI.
Introduction to Biological databases: Types of databases, sequence databases-nucleic acid sequence databases, Genbank, protein sequence database, Swiss-Prot, PIR, motif databases-PROSITE, structural databases, bibliographic databases and organism specific databases-GMOD-searching and retrieval of data-Entrez and SRS.

INTRODUCTION

Bioinformatics is a new discipline that addresses the need to manage and interpret the data that in the past decade was massively generated by genomic research. This discipline represents the convergence of genomics, biotechnology and information technology, and encompasses analysis and interpretation of data, modeling of biological phenomena, and development of algorithms and statistics. Bioinformatics is by nature a cross-disciplinary field that began in the 1960s with the efforts of Margaret O. Dayhoff, Walter M. Fitch, Russell F. Doolittle and others and has matured into a fully developed discipline. However, bioinformatics is wide-encompassing and is therefore difficult to define. For many, including myself, it is still a nebulous term that encompasses molecular evolution, biological modeling, biophysics, and systems biology. For others, it is plainly computational science applied to a biological system. Bioinformatics is also a thriving field that is currently in the forefront of science and technology. Our society is investing heavily in the acquisition, transfer and exploitation of data and bioinformatics is at the center stage of activities that focus on the living world. It is currently a hot commodity, and students in bioinformatics will benefit from employment demand in government, the private sector, and academia.

With the advent of computers, humans have become ‘data gatherers’, measuring every aspect of our life with inferences derived from these activities. In this new culture, everything can and will become data (from internet traffic and consumer taste to the mapping of galaxies or human behavior). Everything can be measured (in pixels, Hertz, nucleotide bases, etc), turned into collections of numbers that can be stored (generally in bytes of information), archived in databases, disseminated (through cable or wireless conduits), and analyzed. We are expecting giant pay-offs from our data: proactive control of our world (from earthquakes and disease to finance and social stability), and clear understanding of chemical, biological and cosmological processes. Ultimately, we expect a better life. Unfortunately, data brings clutter and noise and its interpretation cannot keep pace with its accumulation. One problem with data is its multi-dimensionality and how to uncover underlying signal (patterns) in the most parsimonious way (generally using nonlinear approaches.

Another problem relates to what we do with the data. Scientific discovery is driven by falsifiability and imagination and not by purely logical processes that turn observations into understanding. Data will not generate knowledge if we use inductive principles.

The gathering, archival, dissemination, modeling, and analysis of biological data falls within a relatively young field of scientific inquiry, currently known as 'bioinformatics', 'Bioinformatics was spurred by wide accessibility of computers with increased compute power and by the advent of *genomic*s. Genomics made it possible to acquire nucleic acid sequence and structural information from a wide range of genomes at an unprecedented pace and made this information accessible to further analysis and experimentation. For example, sequences were matched to those coding for globular proteins of known structure (defined by crystallography) and were used in high-throughput combinatorial approaches (such as DNA microarrays) to study patterns of gene expression. Inferences from sequences and biochemical data were used to construct metabolic networks. These activities have generated terabytes of data that are now being analyzed with computer, statistical, and machine learning techniques. The sheer number of sequences and information derived from these endeavors has given the false impression that imagination and hypothesis do not play a role in acquisition of biological knowledge. However, bioinformatics becomes only a science when fueled by hypothesis-driven research and within the context of the complex and ever changing living world.

The science that relates to bioinformatics has many components. It usually relates to biological molecules and therefore requires knowledge in the fields of biochemistry, molecular biology, molecular evolution, thermodynamics, biophysics, molecular engineering, and statistical mechanics, to name a few. It requires the use of computer science, mathematical, and statistical principles. Bioinformatics is in the cross roads of experimental and theoretical science. Bioinformatics is not only about modeling or data 'mining', it is about understanding the molecular world that fuels life from evolutionary and mechanistic perspectives. It is truly inter-disciplinary and is changing. Much like biotechnology and genomics, bioinformatics is moving from applied to basic science, from developing tools to developing hypotheses.

Definition

- Bioinformatics is the recording, annotation, storage, analysis, and searching/retrieval of nucleic acid sequence (genes and RNAs), protein sequence and structural information. This includes databases of the sequences and structural information as well methods to access, search, visualize and retrieve the information.
- Sequence data can be used to make predictions of the functions of newly identified genes.
- Estimate evolutionary distance in phylogeny reconstruction, determine the active sites of enzymes, construct novel mutations and characterize alleles of genetic diseases to name just a few uses.
- Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline.
- The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned.

There are three important sub-disciplines within bioinformatics involving computational biology:

- 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.

History of Bioinformatics

- The history of biology in general, B.C. and before the discovery of genetic inheritance by G. Mendel in 1865, is extremely sketch and inaccurate. This was the start of Bioinformatics history.
- G. Mendel is known as the "Father of Genetics". He did experiment on the cross-fertilization of different colors of the same species.
- Mendel illustrated that the inheritance of traits could be more easily explained if it was controlled by factors passed down from generation to generation.
- The understanding of genetics has advanced remarkably in the last thirty years. In 1972, Paul berg made the first recombinant DNA molecule using ligase.
- In that same year, Stanley Cohen, Annie Chang and Herbert Boyer produced the first recombinant DNA organism.
- In 1973, two important things happened in the field of genomics.
- The advancement of computing in 1960-70s resulted in the basic methodology of bioinformatics. 1990s when the INTERNET arrived when the full fledged bioinformatics field was born.

Chronological History of Bioinformatics

- 1953 - Watson & Crick proposed the double helix model for DNA based x-ray data obtained by Franklin & Wilkins.

- 1954 - Perutz's group develops 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 Standford 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 DataBank is announced. Robert Metcalfe receives his Ph.D from Harvard University. His thesis describes Ethernet.
- 1974 - Vint Cerf and Robert Khan 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'Farrel.
- 1988 - The National Centre 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 Lipmann. 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, UK by Anthony Marchigton, David Ricketts, James Hiddleston, Anthony Rees, and W. Graham Richards. Primary products: Anaconds, 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. 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.

Major events in Computational Methods and Computational Biology

- 1993 - CuraGen Corporation is formed in New Haven, CT. Affymetrix begins independent operations in Santa Clara, California.
- 1994 - Netscape Communications 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.
- 1995 - *The Haemophilus* influenza genome (1.8) is sequenced. The *Mycoplasma genitalium* genome is sequenced.
- 1996 - 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 focused 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 *Caenorhabitis 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. Gene Formatics, a company dedicated to the analysis and predication 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 sequenced. The *D. melanogaster* genome (180 Mb) is sequenced. Pharmacopeia acquires Oxford Molecular Group.
- 2001 - The human genome (3,000 Mbp) is published.

Genome sequencing

- 1990s did advances in sequencing technology make it feasible to sequence the entire genome of anything more complex than a bacterium.

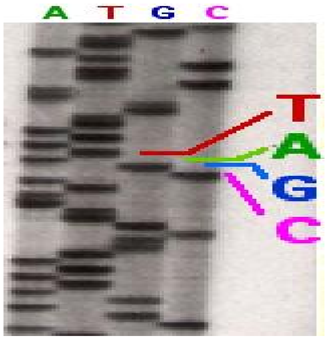
- DNA sequencing includes several methods and technologies that are used for determining the order of the nucleotide base adenine, guanine, cytosine, and thymine in a molecule of DNA.
- Knowledge of DNA sequences has become indispensable for basic biological research, other research branches utilizing DNA sequencing, and in numerous applied fields such as diagnostic, biotechnology, forensic biology and biological systematic.
- The advent of DNA sequencing has significantly accelerated biological research and discovery.
- The rapid speed of sequencing attained with modern DNA sequencing technology has been instrumental in the sequencing of the human genome, in the Human Genome Project.
- Related projects, often by scientific collaboration across continents, have generated the complete DNA sequences of many animal, plant, and microbial genomes.
- The first DNA sequences were obtained in the early 1970s by academic researchers using laborious methods based on two-dimensional chromatography. Following the development of dye-based sequencing methods with automated analysis,
- DNA sequencing has become easier and orders of magnitude faster.

Maxam–Gilbert sequencing

- In 1976–1977, Allan Maxam and Walter Gilbert developed a DNA sequencing method based on chemical modification of DNA and subsequent cleavage at specific bases.
- Although Maxam and Gilbert published their chemical sequencing method two years after the ground-breaking paper of Sanger and Coulson on plus-minus sequencing.
- Maxam–Gilbert sequencing rapidly became more popular, since purified DNA could be used directly, while the initial Sanger method required that each read start be cloned for production of single-stranded DNA.

- However, with the improvement of the chain-termination method, Maxam-Gilbert sequencing has fallen out of favor due to its technical complexity prohibiting its use in standard molecular biology kits, extensive use of hazardous chemicals, and difficulties with scale-up.
- The method requires radioactive labeling at one 5' end of the DNA and purification of the DNA fragment to be sequenced.
- Chemical treatment generates breaks at a small proportion of one or two of the four nucleotide bases in each of four reactions (G, A+G, C, C+T).
- For example, the purines (A+G) are depurinated using formic acid, the guanines (and to some extent the adenines) are methylated by dimethyl sulfate, and the pyrimidines (C+T) are methylated using hydrazine.
- The addition of salt (sodium chloride) to the hydrazine reaction inhibits the methylation of thymine for the C-only reaction.
- The modified DNAs are then cleaved by hot piperidine at the position of the modified base. The concentration of the modifying chemicals is controlled to introduce on average one modification per DNA molecule. Thus a series of labeled fragments is generated, from the radio labeled end to the first "cut" site in each molecule.
- The fragments in the four reactions are electrophoresed side by side in denaturing acrylamide gels for size separation.
- To visualize the fragments, the gel is exposed to X-ray film for autoradiography, yielding a series of dark bands each corresponding to a radio labeled DNA fragment, from which the sequence may be inferred.
- Chemical sequencing", this method led to the Methylation Interference Assay used to map DNA-binding sites for DNA-binding proteins.

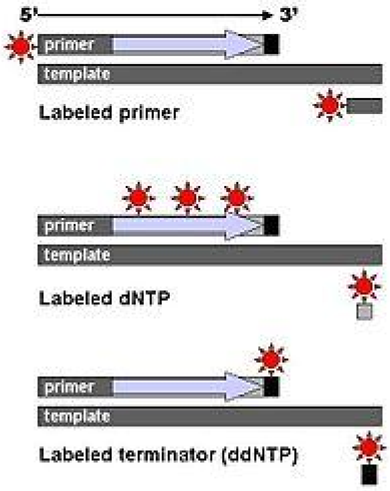
Chain-termination methods



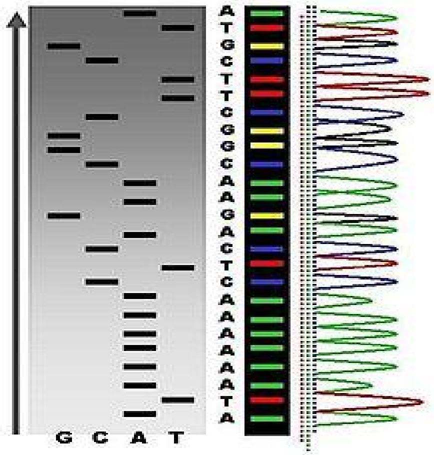
- Because the chain-terminator method is more efficient and uses fewer toxic chemicals and lower amounts of radioactivity than the method of Maxam and Gilbert, it rapidly became the method of choice.
- The key principle of the Sanger method was the use of dideoxynucleotide triphosphates (ddNTPs) as DNA chain terminators.
- The classical chain-termination method requires a single-stranded DNA template, a DNA primer, a DNA polymerase, normal deoxynucleotide triphosphates (dNTPs), and modified nucleotides (dideoxyNTPs) that terminate DNA strand elongation.
- These ddNTPs will also be radioactively or fluorescently labelled for detection in automated sequencing machines.
- The DNA sample is divided into four separate sequencing reactions, containing all four of the standard deoxynucleotides (dATP, dGTP, dCTP and dTTP) and the DNA polymerase.
- To each reaction is added only one of the four dideoxynucleotides (ddATP, ddGTP, ddCTP, or ddTTP) which are the chain-terminating nucleotides, lacking a 3'-OH group required for the formation of a phosphodiester bond between two nucleotides, thus terminating DNA strand extension and resulting in DNA fragments of varying length.
- The newly synthesized and labeled DNA fragments are heat denatured, and separated by size by gel electrophoresis on a denaturing polyacrylamide-urea gel with each of the four reactions run in one of four individual lanes (lanes A, T, G, C);

the DNA bands are then visualized by autoradiography or UV light, and the DNA sequence can be directly read off the X-ray film or gel image.

- In the image on the right, X-ray film was exposed to the gel, and the dark bands correspond to DNA fragments of different lengths.
- A dark band in a lane indicates a DNA fragment that is the result of chain termination after incorporation of a dideoxynucleotide (ddATP, ddGTP, ddCTP, or ddTTP). The relative positions of the different bands among the four lanes are then used to read (from bottom to top) the DNA sequence.



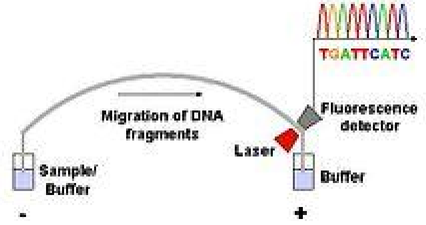
- DNA fragments are labeled with a radioactive or fluorescent tag on the primer, in the new DNA strand with a labeled dNTP, or with a labeled ddNTP.
- Technical variations of chain-termination sequencing include tagging with nucleotides containing radioactive phosphorus for radio labeling, or using a primer labeled at the 5' end with a fluorescent dye.
- Dye-primer sequencing facilitates reading in an optical system for faster and more economical analysis and automation.
- The later development by Leroy Hood and coworkers–of fluorescently labeled ddNTPs and primers set the stage for automated, high-throughput DNA sequencing.



Sequence ladder by radioactive sequencing compared to fluorescent peaks

- Chain-termination methods have greatly simplified DNA sequencing.
- For example, chain-termination-based kits are commercially available that contain the reagents needed for sequencing, pre-aliquot and ready to use.
- Limitations include non-specific binding of the primer to the DNA, affecting accurate read-out of the DNA sequence, and DNA secondary structures affecting the fidelity of the sequence.

Dye-terminator sequencing



Capillary electrophoresis

- Dye-terminator sequencing utilizes labeling of the chain terminator ddNTPs, which permits sequencing in a single reaction, rather than four reactions as in the labeled-primer method.
- In dye-terminator sequencing, each of the four dideoxynucleotide chain terminators is labeled with fluorescent dyes, each of which emits light at different wavelengths.

- Owing to its greater expediency and speed, dye-terminator sequencing is now the mainstay in automated sequencing.
- Its limitations include dye effects due to differences in the incorporation of the dye-labeled chain terminators into the DNA fragment, resulting in unequal peak heights and shapes in the electronic DNA sequence trace chromatogram after capillary electrophoresis
- This problem has been addressed with the use of modified DNA polymerase enzyme systems and dyes that minimize incorporation variability, as well as methods for eliminating "dye blobs".
- The dye-terminator sequencing method, along with automated high-throughput DNA sequence analyzers, is now being used for the vast majority of sequencing projects.
- Very few genes in a cell are actually active in protein production at any given time.
- Different sections of DNA may be dormant or active over the life of a cell, their expression triggered by other genes and changes in the cell's internal environment.
- Genes interact, why they express at certain times and not others, and how the mechanisms of gene suppression and activation work are all topics of intense interest in microbiology.
- Chromosomes, which range in size from 50 million to 250 million bases, must first be broken into much shorter pieces (*sub cloning step*).
- Each short piece is used as a template to generate a set of fragments that differ in length from each other by a single base that will be identified in a later step (*template preparation and sequencing reaction steps*).
- The fragments in a set are separated by gel electrophoresis (*separation step*).
- New fluorescent dyes allow separation of all four fragments in a single lane on the gel.

- The final base at the end of each fragment is identified (*base-calling step*). This process recreates the original sequence of As, Ts, Cs, and Gs for each short piece generated in the first step.
- Automated sequencers analyze the resulting electropherograms, and the output is a four-color chromatogram showing peaks that represent each of the four DNA bases.
- After the bases are "read," computers are used to assemble the short sequences into long continuous stretches that are analyzed for errors, gene-coding regions, and other characteristics.
- Finished sequences are submitted to major public sequence databases, such as GenBank. Human Genome Project sequence data are thus freely available to anyone around the world
- Full genome sequencing (FGS), also known as whole genome sequencing (WGS), complete genome sequencing, or entire genome sequencing, is a laboratory process that determines the complete DNA sequence of an organism's genome at a single time.
- This entails sequencing all of an organism's chromosomal DNA as well as DNA contained in the mitochondria and, for plants, in the chloroplast. Almost any biological sample even a very small amount of DNA or ancient DNA-can provide the genetic material necessary for full genome sequencing.
- Such samples may include saliva, epithelial cells, bone marrow, hair (as long as the hair contains a hair follicle), seeds, plant leaves, or anything else that has DNA-containing cells.
- Because the sequence data that is produced can be quite large (for example, there are approximately six billion base pairs in each human diploid genome), genomic data is stored electronically and requires a large amount of computing power and storage capacity.
- Full genome sequencing would have been nearly impossible before the advent of the microprocessor, computers, and the Information Age.

- Unlike full genome sequencing, DNA profiling only determines the likelihood that genetic material came from a particular individual or group; it does not contain additional information on genetic relationships, origin or susceptibility to specific diseases.
- Also unlike full genome sequencing, SNP genotyping covers less than 0.1% of the genome.
- Almost all truly complete genomes are of microbes; the term "full genome" is thus sometimes used loosely to mean "greater than 95%". The remainder of this article focuses on nearly complete human genomes.
- In general, knowing the complete DNA sequence of an individual's genome does not, on its own, provide useful clinical information, but this may change over time as a large number of scientific studies continue to be published detailing clear associations between specific genetic variants and disease.
- The first nearly complete human genomes sequenced were J. Craig Venter's James Watson's a Han Chinese a Yoruban from Nigeria a female leukemia patient (and Seong-Jin Kim
- There are currently over 60 nearly complete human genomes publicly available.
- Sequencing of nearly an entire human genome was first accomplished in 2000 partly through the use of shotgun sequencing technology.
- The Institute for Genomic Research (TIGR) to sequence the entire genome of the bacterium *Haemophilus influenzae* in 1995, and then by Celera Genomics to sequence the entire fruit fly genome in 2000.

The Mechanics of Sequencing

- The goal of genome sequencing projects is to record all of the genetic information contained in a given organism - that is, create a sequential list of the base pairs comprising the DNA of a particular plant or animal.

- Since chromosomes consist of long, unbroken strands of DNA, a very convenient way to sequence a genome would be to unravel each chromosome and read off the base pairs like punch tape.
- Unfortunately, there is no machine available that can read a single strand of DNA Instead, scientists have to use a cruder, shotgun technique that first chops the DNA into short pieces and then tries to reassemble the original sequence based on how the short fragments overlap.
- Various alignment algorithms to look for overlaps between short DNA fragments.

Finding the Genes

- Once a reliable DNA sequence has been established, there still remains the task of finding the actual genes (coding regions) embedded within the DNA strand.
- Large proportion of the DNA in a genome is non-coding.
- Finding the coding regions is an important step in genome analysis, but it is not the end of the road.
- Emergent behaviour is very hard to simulate, because there is no way to infer the simple rules from the complex behaviour Still, computers give us a way to try many different rule sets and test hypotheses about gene interaction.

Unexplored Territory

- Only a very few organisms have had their genome fully sequenced
- Many technical and computational challenges remain before sequencing becomes an automatic process
- Some species are still very difficult for us to sequence, and much remains to be learned about the role and origin of that entire non-coding DNA.

Human Genome Projects

Goals:

- Identify all the approximate 30,000 genes in human DNA
- Determine the sequences of the 3 billion base pairs that make up human DNA

- Sequence the genomes of other model organisms including *Escherichia coli*, yeast (*Saccharomyces cerevisiae*), the fruit fly *Drosophila melanogaster*, the worm *Caenorhabditis elegans* and the laboratory mouse,
- Store this information in databases
- Improve tools for data analysis
- Transfer related technologies to the private sector and
- Address the ethical, legal and social issues (ELSI) that may arise from the project.

Milestones:

- 1990: project initiated as joint effort of US Department of Energy and the National Institutes of Health
- June 2000: Completion of a working draft of the entire human genome
- February 2001: Analyses of the working draft are published
- April 2003: HGP sequencing is completed and project is declared finished two years ahead of schedule.

What does the draft human genome sequence tell us?

By the numbers

- The human genome contains 3 billion chemical nucleotide bases (A, C, T, and G).
- The average gene consists of 3000 bases, but sizes vary greatly, with the largest known human gene being dystrophin at 2.4 million bases.
- The total number of genes is estimated at around 20,000- 25,000 much lower than previous estimates of approximately 100,000.
- Almost all (99.9%) nucleotide bases are exactly the same in all people.
- The functions are unknown for over 50% of discovered genes.

How it's arranged

- The human genome's gene-dense "urban centers" are predominantly composed of the DNA building blocks G and C.

- In contrast, the gene-poor "deserts" are rich in the DNA building blocks A and T. The GC and AT rich regions usually can be seen through a microscope as light and dark bands on chromosomes.
- Genes appear to be concentrated in random areas along the genome, with vast expanses of non-coding DNA between.
- Stretches of up 30,000 C and G bases repeating over and over often occur adjacent to gene rich areas, forming a barrier between the genes and the "junk DNA". These CpG islands are believed to help regulated gene activity.
- Chromosome 1 has the most genes (2968), and the Y chromosome has the fewest (231).

Future Challenges:

What we still don't know

- Gene number, exact locations, and functions
- Gene regulation
- DNA sequence organization
- Chromosomal structure and organization
- Non-coding DNA types, amount, distribution, information content and functions
- Coordination of gene expression, protein synthesis, and post-translational events
- Interaction of proteins in complex molecular machines
- Predicted vs experimentally determined gene function
- Evolutionary conservation among organisms
- Protein conservation (Structure and function)
- Proteomes (total protein content and function) in organisms
- Correlation of SNPs (single-base DNA variations among individuals) with health and disease
- Disease susceptibility prediction based on gene sequence variation
- Genes involved in complex traits and multigene diseases

- Complex systems biology including microbial consortia useful for environmental restoration
- Developmental genetics, genomics

Anticipated Benefits of Genome Research

- Molecular medicine
- Improve diagnosis of disease
- Detect genetic predispositions to disease
- Create drugs based on molecular information
- Use gene therapy and control systems as drugs
- Design “custom drugs” based on individual genetic profiles

Microbial genomics

- Rapidly detect and treat pathogens (disease causing microbes) in clinical practice
- Develop new energy sources (biofuels)
- Monitor environments to detect pollutants
- Protect citizenry from biological and chemical warface
- Clean up toxic waste safely and efficiently

Risk assessment

- Evaluate the health risks faced by individuals who may be exposed to radiation (including low levels in industrial areas) and to cancer causing chemicals and toxins.
- Bioarchaeology, Anthropology, Evolution and Human migration
- Study evolution through germline mutations in lineages
- Study migration of different population groups based on maternal inheritance
- Study mutations on the Y chromosome to trace lineage and migration of males
- Compare breakpoints in the evolution of mutations with ages of populations and historical events.

DNA identification

- Identify potential suspects whose DNA may match evidence left at crime scenes

- Exonerate persons wrongly accused of crimes
- Identify crime and catastrophe victims
- Establish paternity and other family relationships
- Identify endangered and protected species as an aid to wildlife officials
- Detect bacteria and other organisms that may pollute air, water, soil and food
- Match organ donors with recipients in transplant programs
- Determine pedigree for seed or livestock breeds
- Authenticate consumables such as caviar and wine

Agriculture, Livestock, Breeding, and Bioprocessing

- Grow disease, insect, and drought resistant crops
- Breed healthier, more productive, diseases resistant farm animals
- Grow more nutritious produce
- Develop biopesticides
- Incorporate edible vaccines incorporated into food products
- Develop new environmental cleanup uses for plants like tobacco
- Cellulosis biomass research for bioenergy

Anticipated benefits

- Improved diagnosis of disease
- Earlier detection of genetic predispositions to disease
- Rational drug design
- Gene therapy and control systems for drugs
- Personalized, custom drugs

ELSI (Ethical, Legal, and Social Issues)

Privacy and confidentiality of genetic information

- Fairness in the use of genetic information by insurers, employers, courts, schools, adoption agencies, and the military, among others.
- Psychological impact, stigmatization and discrimination due to an individual's genetic differences.

- Reproductive issues including adequate and informed consent and use of genetic information in reproductive decision making.
- Clinical issues including the education of doctors and other health service providers, people identified with genetic conditions, and the general public about capabilities, limitations, and social risks; and implementation of standards and quality control measures.
- Uncertainties associated with gene tests for susceptibilities and complex conditions (e.g., heart disease, diabetes, and alzheimer's disease).
- Fairness in access to advanced genomic technologies.
- Conceptual and philosophical implications regarding human responsibility, free will vs genetic determinism and concepts of health and disease.
- Health and environmental issues concerning genetically modified (GM) foods and microbes.
- Commercialization of products including property rights (patents, copyrights, and trade secrets) and accessibility of data and materials.

INTRODUCTION TO BIOLOGICAL DATABASES

Data

Data is unprocessed facts and figures without any added interpretation or analysis

Information

Information is data that has been interpreted so that it has meaning for the user.

Database

Is a usually large collection of data organized especially for rapid search and retrieval.

There are many different types of database but for routine sequence analysis, the following are initially the most important

- Primary database
- Secondary database
- Composite database

Primary Database

- Primary databases are produced with experimentally derived data such as nucleotide sequence, protein sequence or macromolecular structure.
- Experimental results are submitted directly into the database by researchers, and the data are essentially archival in nature.
- Once given a database accession number, the data in primary databases are never changed: they form part of the scientific record.

Secondary Database

- 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).
- They are highly curated, often using a complex combination of computational algorism and manual analysis and interpretation to derive new knowledge from the public record of science.

	Primary Database	**Secondary Database**
Synonyms	Archival Database	Curated Database; Knowledgebase
Source of Data	Direct submission of experimentally-derived data from researchers	Results of analysis, literature research and interpretation, often of data in primary database
Examples	ENA, GenBank and DDBJ (Nucleotide sequence) Array Express Archieve and GEO (functional genomics data) Protein Data Bank (PDB coordinates of three dimensional macromolecular	InterPro (protein families, motifs and domains) UniProt Knowledgebase (sequence and functional information on proteins) Ensemble (variation, function, regulation and more layered onto whole genome sequences)

	structure	

Composite databases

- Collection of various primary database sequences
- Renders sequence searching highly efficient as it searches multiple resources
- Example: NRDB (non redundant database), OWL, MIPSX, SWISSPROT, TrEMBL

Nucleic acid Sequence databases

GenBank

- GenBank ® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences
- GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.
- A GenBank release occurs every two months and is available from the ftp site.
- The release notes for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank.

Access to GenBank

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

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

CoreNucleotide (the main collection)

- The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB.

dbEST (Expressed Sequence Tags)

- The EST database is a collection of short single-read transcript sequences from GenBank. These sequences provide a resource to evaluate gene expression, find potential variation, and annotate genes.

dbGSS (Genome Survey Sequences)

- The GSS database is a collection of unannotated short single-read primarily genomic sequences from GenBank including random survey sequences clone-end sequences and exon-trapped sequences.

GenBank Data Usage

- The GenBank database is designed to provide and encourage access within the scientific community to the most up to date and comprehensive DNA sequence information.
- Therefore, NCBI places no restrictions on the use or distribution of the GenBank data. However, some submitters may claim patent, copyright, or other intellectual property rights in all or a portion of the data they have submitted.
- NCBI is not in a position to assess the validity of such claims, and therefore cannot provide comment or unrestricted permission concerning the use, copying, or distribution of the information contained in GenBank.

Confidentiality

- Some authors are concerned that the appearance of their data in GenBank prior to publication will compromise their work.
- GenBank will, upon request, withhold release of new submissions for a specified period of time.
- A date must be specified; we cannot hold a sequence indefinitely pending publication.
- However, if a paper citing the sequence or accession number is published prior to the specified date, the sequence will be released upon publication.
- In order to prevent the delay in the appearance of published sequence data, we urge authors to inform us of the appearance of the published data.
- As soon as it is available, please send the full publication data--all authors, title, journal, volume, pages and date--to the following address: update@ncbi.nlm.nih.gov

Submission to GenBank

There are several options for submitting data to GenBank:

BankIt, a WWW-based submission tool with wizards to guide the submission process

tbl2asn, a command-line program, automates the creation of sequence records for submission to GenBank using many of the same functions as Sequin. It is used primarily for submission of complete genomes and large batches of sequences and is available by FTP for use on MAC, PC and Unix platforms.

Submission Portal , a unified system for multiple submission types. Currently only ribosomal RNA (rRNA), rRNA-ITS or Influenza sequences can be submitted with the GenBank component of this tool.

Sequin, NCBI's stand-alone submission tool with wizards to guide the submission process is available by FTP for use on for MAC, PC, and UNIX platforms.

EMBL

- European molecular Biology Laboratory
- Nucleic acid database from EBI (European Bioinformatics Institute)
- Produced in collaboration with DDBJ and GenBank
- Search engine – SRS (sequence Retrieval System)
- Keeping with the tremendous growth in field of computational biology, a need was felt to establish an independent and parallel research institute that would act not just as a mirror housing the GenBank nucleotide resources of NCBI, but would also develop matching databases and analysis tools. The European Molecular Biology Laboratory (EMBL) was thus established in 1974 and is now supported with funding from 20 members states of the European Union, Israel and Australia. EMBL currently operates five research institutes in different countries with main institute at Heidelberg, Germany.

The Five institutes of EMBL with their core research activities are

- EMBL Heidelberg (Germany)
- EMBL Grenoble (France)- Structural Biology

- EMBL-European Bioinformatics Institute (Hinxton, UK)- Bioinformatics
- EMBL Hamburg (Germany)-Structural Biology
- EMBL Monterotondo (Italy)-Mouse Biology

The broad goals of EMBL are

- Basic research in Molecular biology
- Training manpower i.e. students, scientist and visitors
- Develop new tools, technologies and methods
- Offer service to the research community
- Transfer technology to industry for commercialization

The following are the broad categories of databases at EBI-EMBL

- Biological ontologies
- Literature
- Functional Genomics or microarray
- Nucleotides
- Pathways and networks
- Protein
- Proteomics
- Small molecules
- Structure

DDBJ

- DNA databank of Japan
- Started in 1986 in collaboration with GenBank
- Produced and maintained at NIG (National Institute of Genetics)
- DDBJ was established in the year 1986 at the National Institute of Genetics (NIG), Japan with support from the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT). Later on for its efficient functioning, Center for Information Biology (CIB) was established at NIG in 1995. In 2004, NIG was made a member of Research Organization of Information and Systems.

- The functioning and maintenance of DDBJ is monitored by an international advisory committee consisting of 9 members from Japan, Europe and USA. The committee reviews the functioning of DDBJ and reports the progress of DDBJ in database issue of Nucleic acid Research Journal every year. Since its inception there has been a tremendous increase in the number of sequence submitted to DDBJ.

Roles of DDBJ

As a member of INSDC, primary objective of DDBJ is to collect sequence data from researchers all over the world and to issue a unique accession number for each entry. The data collected from the submitters is made publically available and anyone can access the data through data retrieval tools available at DDBJ. Everyday data submitted at either DDBJ or EMBL or NCBI is exchanged, therefore at any given time these three databases contain same data. Following are the steps along with snapshots showing data retrieval from DDBJ using getenetry.

- Open the homepage of DDBJ
- Click on the search/Analysis link on the menu bar
- Click on getentry link
- Type in the accession number in the search box and click on search
- Desired sequence will be retrieved.

Software development

DDBJ team continuously focuses on developing new software which can be used for data analysis. For example, WINA (A window Analysis Program for the number of synonymous and nonsynonymous nucleotide substitutions) has been developed by DDBJ. It is tool which helps in visualizing the difference in accumulation of both synonymous and nonsynonymous nucleotide substitutions.

Training courses

DDBJ also focuses on providing teaching assistance on bioinformatics. It conducts Bioinformatics training course which teaches analysis of data.

SwissProt

- Annotated sequence database established in 1986
- Consists of sequence entries of different lie formats
- Similar format to EMBL

SwissProt is an annotated protein sequence database which was formulated and managed by Amos Bairoch in 1986. It was established collaboratively by the Department of Medical Biochemistry at the University of Geneva and European Molecular Biology Laboratory (EMBL). Later it shifted to European Bioinformatics Institute (EBI) in 1994 and finally in April 1998, it became a part of Swiss Institute of Bioinformatics (SIB). In 1996, TrEMBL was added as an automatically annotated supplement to Swiss-Prot database. Since 2002, it is maintained by the UniProt consortium and information about a protein sequence can be accessed via the Uniprot website. The universal protein resource is the most widespread protein sequence catalog comprising of EBI, SIB and PIR.

There are four main features of Swiss-Prot

High quality annotation

It is achieved through manually creating the protein sequence entries. It is processed through 6 stages.

Sequence curation: In this step, identical sequences are extracted through blast search and then the sequence form the related gene and same organism are incorporated into a single entry. It makes sure that the sequence is complete, correct and ready for further curation steps.

Sequence analysis: It is performed by using various sequence analysis tools. Computer predictions are manually reviewed and important results are selected for integration.

Literature curation: In this step, important publications related to the sequence are retrieved from literature databases. The whole text of each article is scanned manually and relevant information if gathered and supplemented to the entry.

Family based curation: Putative homolog's are determined by reciprocal Blast searches and phylogenetic resources which are further evaluated, curated, annotated and propagated across homologous proteins to ensure data consistency.

Evidence attribution: All information incorporated to the sequence entry during manual annotation is linked to the original source so that users can trace back the origin of data and evaluate its.

Quality assurance, integration and update: each completely annotated entry undergoes quality assurance before integration into Swiss-Prot and is updated as new data become available.

Minimum redundancy: during manual annotation, all entries belonging to identical gene and form similar organism are merged into a single entry containing complete information. This results in minimal redundancy.

Integration with other databases: Swiss-Prot is presently cross-referenced to move than 50 specialized documentation files. Documentation file section provides an updated descriptive list of all document files.

PIR

- Protein Information Resource
- A division of National Biomedical Research Foundation (NBRF) in U.S
- One can search for entries or do sequence similarity search at PIR site.

In year 1984, National Biomedical Research Foundation (NBRF) developed PIR for identification and interpretation of information on protein sequences. This database was actually derived from Atlas of Protein Sequence and Structure, which was developed by Margaret O Dayhoff in the year 1964. Four years later in 1988, PIR along with NBRF, Munich Information Centre for Protein Sequence (MIPS) and the Japan International **Protein Information Database, developed an organization referred as PIR – international with four main aims.**

- To create an organized, non redundant, comprehensive protein database to study structural, functional and evolutionary relationships.
- To generate information on biological origin of protein sequences
- To make database easily accessible in public domain

- To enable cross reference with other databases for presenting structural information of biomolecules.

TrEMBL

TrEMBL stands for automatic Translations of European Molecular Biology Laboratory nucleotide sequences. It is a protein sequence database consisting of unreviewed computer annotated translations of new DNA sequence in the nucleotide sequence databases. Swiss-Prot only includes entries validated by expert curators.

This database was created in 1996 as a computer-annotated supplementary database to Swiss-Prot. With the invent of high throughput sequencing techniques, there is an immense flow of new sequence data from the genome projects and Swiss-Prot is falling behind to provide quick database annotation. To address this problem, a Swiss-Prot buffer called TrEMBL was created. It allows very rapid access to sequence data from the genome projects, without having to compromise the quality of Swiss-Prot.

TrEMBL sequences are produced at the EBI from GenBank entries and annotated mostly computationally using sequence homology as a main principle. It also contains protein sequences selected from the literature and protein entries submitted directly by the researchers. TrEMBL unreviewed entries are kept separated from the Swiss-Prot manually annotated entries so as to maintain the high quality data of later.

The Key features of TEMBL are:

Automatic annotation: It is performed by transferring data from well-labeled entries of Swiss-Prot to unannotated entries in TrEMBL. This process raises the standard of annotation in TrEMBL next to the level of Swiss-Prot, thus improving the quality of data.

Redundancy removal: Full length sequence belonging to same organism and showing 100% identify are fused into a single entry to curtail redundancy.

Evidence attribution: Since TrEMBL contains data from a variety of sources, evidence attribution helps in identifying the source of individual data items. It allows automatic update of data if the underlying data source changes.

It has been dissected into two parts: SP-TrEMBL and REM-TrEMBL

SP-TrEMBL (Swiss-Prot TrEMBL) is a collection of sequences that will be finally upgraded to Swiss-Prot after their manual annotation is finished.

REM-TrEMBL (Remaining TrEMBL) stores those sequences that will never be incorporated in Swiss-Prot. E.g immunoglobulins and T-Cell receptors, fragments of fewer than 8 amino acids, synthetic sequences, patented sequences and coding sequences that do not code real proteins.

Structural Databases

NDB

- A repository of three dimensional structural information about nucleic acids
- The NDB is supported by funds from the national science foundation and the department of energy
- The NDB follows the dictionaries and formats used by the worldwide protein data bank
- Search the NDB by ID
- Enter an NDB ID or PDB ID
- Atlas, Deposit Data, Download Data, Search, Education, Standards, Tools, Links
- The NDB Atlas provides summary information and images for each structure in the database. The Atlas is first divided by experimental type and then by structure type.

Features include

- Image of the asymmetric and biological units, and crystal packing pictures for nucleic acid structures from X-Ray crystallographic experiments
- Image of the average and ensemble structure form NMR experiments
- Links to coordinate files, experimental data files
- Tables of derived data, including torsion angles and hydrogen bonding classifications
- Special features for RNA structures, including images of secondary and tertiary structure

- Each page in Atlas is generated directly from NDB using an XML translator which formats the data contained in each NDB files.
- Pictures present on the Atlas pages were generated by different software
- Blocview
- RNAview
- MaxIT
- In the nucleotide block models, adenine is red, thymine is blue, cytosine is yellow, guanine is green, and uracil is cyan. In the atom stick models, carbon is black, oxygen is red, nitrogen is blue, and phosphates are orange.
- The Atlas is first divided by how the structures were determined: by X-ray crystallographic or NMR experiments. Gallery index pages, which include images for each structure on the page, and plain text index pages are offered.
- The NDB processes data for the crystal structures of nucleic acids. Structures can be deposited to the NDB and PDB at the same time using ADIT (the AutoDep Input Tool).
- ADIT accepts coordinates in PDB or mmCIF format and structure factor files. All other information is entered into ADIT by the author.
- Coordinate files can be downloaded from download option
- Basic detail of DNA and RNA is given in education option
- X-plor parmeters and geometries are given in standard option

Tools and Software's

RNA viewer- RNA 2-dimensional structure using the RNAview program

Base pair Viewer- RNA base pairs using the BPView program

DNA binding prediction for protein structures are HTHQuery and predictdnahth these predicts whether given three dimensional protein structure contains a DNA-binding Helix-turn-Helix (HTH) structural motif.

QPROF (Query of PROtein Features) A web utility for secondary similarity search of protein three dimensional structure.

RNAView Program Quickly generate display of RNA/DNA secondary structures with tertiary interactions.

RNAMLview Program Display and/or edit RNAView 2-dimensional diagrams

3DNA A software package for the analysis, rebuilding and visualization of three dimensional nucleic acid structures.

Freehelix98 described in "DNA bending: the prevalence of kinkiness and the virtues of normality" Richard E. Dickerson Nucleic Acid Research

PDB

Protein Data Bank

- Structural data from the PDB can be freely accessed at
- It is very large global repository for processing and distribution of 3D macromolecular structure data such as protein, nucleic acids
- Depositors to PDB have derived the structures using variety of tools and techniques like X-ray crystal structure determination, NMR, cryoelectron microscopy and theoretical modeling.
- The database provides access at no charge on internet to structural data as well as methods to visualize the structure and to download structural information.
- It is a primary data and databases derived from PDB are called secondary databases like SCOP and CATH.
- The PDB is overseen by an organization called world wide protein data bank (wwPDB): a consortium whose partners comprise: the research collaborator for structural bioinformatics, the macromolecular structure database at the European bioinformatics, the protein data bank Japan at Osaka university and more recently the BioMagResBank at the University of Wisconsin-Madison.
- In 1971 Walter Hamilton of BNL (Brookhaven National Laboratory) agreed to setup the data bank at Brookhavan and then he died in 1983.
- Then Tom Koeztle took over direction of PDB

- Then in 1998 PDB was transferred to RCSB (Research Collaboratory Structural Bioinformatics)
- Then in 2003, with formation of wwPDB, the PDB became an international organization
- Most structures are determined by X-ray diffraction and about 15% of structure by NMR and few by Cryo-electron microscopy. In the past, number of structures in PDB has grown nearly exponentially.
- The file format initially used by PDB was called PDB file format.
- Around 1996, mmCIF (Macro Molecular Crystallographic Information File) started to phased in.
- Then in 2005 XML version of above format called PDBML was described.
- The structure file can be downloaded in any of these three formats.
- Each structure published in PDB receives a four character alpha numeric identifier, its PDB ID
- The Structure files may be viewed using one of the several open source computer programs. Some other free but not open source programs include VMD, MDLChime, Swiss PDB Viewer, started Biochem and Sirius.
- PDB Wiki is a website for community annotation of PDB structures.

Motif Database

Prosite

It was initiated and is maintained by Amos Bairoch and colleagues, now at the Swiss Institute of Bioinformatics. It is based on the proteins sequences in SWISS-PROT. It aims at describing characteristic patterns for some domain families using regular expressions, and contains about 1400 patterns, rules and profile/matrices. It is being maintained, but it is fair to say that it has been superceded in practical terms by other search methods and databases, such as Pfam (mentioned before, and discussed later).

PROSITE makes a distinction between patterns and rules, which are both described by regular expressions:

- A pattern is intended to capture the characteristic fingerprint of a protein domain family.
- A rule, on the other hand, is intended to highlight features in a protein sequence that does not necessarily have anything to do with a specific protein family. For example, potential glycosylation sites and phosphorylation sites can be found in many protein sequences, and have little to do with the family of a protein.

Patterns and rules are described using the same notation. Unfortunately, the PROSITE notation for sequence patterns is different from the UNIX-type regular expressions**.** However, the concepts are the same, and it is not so difficult to translate a PROSITE pattern into a UNIX-type regular expression.

As an example, let us use the PROSITE pattern CBD_FUNGAL (accession code PS00562). The preceding link shows a nicer view of the entry. Below is the original text entry as it is given in the downloadable PROSITE data file.

```
ID   CBD_FUNGAL; PATTERN.
AC   PS00562;
DT   DEC-1991 (CREATED); NOV-1997 (DATA UPDATE); JUL-1998 (INFO UPDATE).
DE   Cellulose-binding domain, fungal type.
PA   C-G-G-x(4,7)-G-x(3)-C-x(5)-C-x(3,5)-[NHG]-x-[FYWM]-x(2)-Q-C.
NR   /RELEASE=38,80000;
NR   /TOTAL=21(18); /POSITIVE=21(18); /UNKNOWN=0(0); /FALSE_POS=0(0);
NR   /FALSE_NEG=1; /PARTIAL=0;
CC   /TAXO-RANGE=??E??; /MAX-REPEAT=4;
CC   /SITE=1,disulfide; /SITE=7,disulfide; /SITE=9,disulfide;
CC   /SITE=16,disulfide;
DR   Q00023, CEL1_AGABI, T; Q12714, GUN1_TRILO, T; P07981, GUN1_TRIRE, T;
DR   P07982, GUN2_TRIRE, T; P43317, GUN5_TRIRE, T; P46236, GUNB_FUSOX, T;
DR   P46239, GUNF_FUSOX, T; P45699, GUNK_FUSOX, T; P15828, GUX1_HUMGR, T;
DR   Q06886, GUX1_PENJA, T; P13860, GUX1_PHACH, T; P00725, GUX1_TRIRE, T;
DR   P19355, GUX1_TRIVI, T; Q92400, GUX2_AGABI, T; P07987, GUX2_TRIRE, T;
```

DR P49075, GUX3_AGABI, T; P46238, GUXC_FUSOX, T; P50272, PSBP_PORPU, T;

DR O59843, GUX1_ASPAC, N;

DO PDOC00486;

//

The central line is the PA line, which contains the pattern. Let us go through this pattern step by step.

PA C-G-G-x (4,7)-G-x(3)-C-x(5)-C-x(3,5)-[NHG]-x-[FYWM]-x(2)-Q-C.

Let us go through the elements in the pattern to see what they mean:

- Each non-x letter defines one particular type of amino-acid residue in that position in the pattern. Here, we must have a tripeptide Cys-Gly-Gly in the beginning of the matching segment of a protein chain. The dash characters '-' add no information to the pattern, and are added to make the pattern slightly easier to read.
- The notation x(4,7) means that at least 4 and at most 7 residues of any type may occur at this position. This corresponds to the notation.{4,7} in a UNIX-type regular expression.
- The notation [NHG] means the same thing as in a UNIX-type regular expression: in this position any of the residues within the brackets may be chosen. One and only one such residue must be at this position.
- The notation x(2) means that exactly two residues of any type may occur at this position. This corresponds to the notation **..** or .{2,2}in a UNIX-type regular expression.
- The notation {GP} (not shown in this example) means that all residues except Gly and Pro are allowed in this position.

The lines marked DR are the protein sequence entries in SWISS-PROT that match (character T) or do not match (character N) the regular expression. In this case, the protein GUX1_ASPAC does not match the PROSITE rule, although it should; it is a false negative.

PUBMED

PubMed is a free resource developed and maintained by the National Center for Biotechnology Information (NCBI), a division of the U.S. National Library of Medicine (NLM), at the National Institutes of Health (NIH).

PubMed comprises over 22 million citations and abstracts for biomedical literature indexed in NLM's MEDLINE database, as well as from other life science journals and online books. PubMed citations and abstracts include the fields of biomedicine and health, and cover portions of the life sciences, behavioral sciences, chemical sciences, and bioengineering. PubMed also provides access to additional relevant websites and links to other NCBI resources, including its various molecular biology databases.

PubMed uses NCBI's Entrez search and retrieval system. PubMed does not include the full text of the journal article; however, the abstract display of PubMed citations may provide links to the full text from other sources, such as directly from a publisher's website or PubMed Central (PMC).

Data Source

MEDLINE

- The primary component of PubMed is MEDLINE, NLM's premier bibliographic database, which contains over 19 million references to journal articles in life sciences, with a concentration on biomedicine.
- The majority of journals selected for MEDLINE are based on the recommendation of the Literature Selection Technical Review Committee (LSTRC), an NIH-chartered advisory committee of external experts analogous to the committees that review NIH grant applications. Some additional journals and newsletters are selected based on NLM-initiated reviews in areas that are special priorities for NLM or other NIH components (e.g., history of medicine, health services research, AIDS, toxicology and environmental health, molecular biology, and complementary medicine). These reviews generally also involve consultation with an array of NIH and outside experts or, in some cases, external organizations with which NLM has special collaborative arrangements.

Non-MEDLINE

In addition to MEDLINE citations, PubMed also contains:

- In-process citations, which provide a record for an article before it is indexed with NLM Medical Subject Headings (MeSH) and added to MEDLINE or converted to out-of-scope status.
- Citations that precede the date that a journal was selected for MEDLINE indexing.
- Some OLDMEDLINE citations that have not yet been updated with current vocabulary and converted to MEDLINE status.
- Citations to articles that are out-of-scope (e.g., covering plate tectonics or astrophysics) from certain MEDLINE journals, primarily general science and general chemistry journals, for which the life sciences articles are indexed with MeSH for MEDLINE.
- Citations to some additional life science journals that submit full-text articles to PubMed Central and receive a qualitative review by NLM.

Journal Selection Criteria

Journals that are included in MEDLINE are subject to a selection process. The Fact Sheet on *Journal Selection for Index Medicus®/MEDLINE®* describes the journal selection policy, criteria, and procedures for data submission.

History

PubMed was first released in January 1996 as an experimental database under the Entrez retrieval system with full access to MEDLINE. The word "experimental" was dropped from the website in April 1997, and on June 26, 1997, free MEDLINE access via PubMed was announced at a Capitol Hill press conference. Use of PubMed has grown exponentially since its introduction: PubMed searches numbered approximately 2 million for the month of June 1997, while current usage typically exceeds 3.5 million searches per day.

PubMed was significantly redesigned in 2000 to integrate new features such as LinkOut, Limits, History, and Clipboard. PubMed began linking to PubMed Central full-text articles and the Bookshelf's initial book, *Molecular Biology of the Cell*.

The Entrez Programming Utilities, E-Utilities, and the Cubby (My NCBI subsequently replaced the Cubby) also were released.

In 2002, the PubMed database programming was completely redesigned to work directly from XML files, and two new NCBI databases, Journals (now the NLM Catalog) and MeSH, were created to provide additional search capabilities for PubMed.

GMOD (Generic model/many/my organism database)

There are three main GMOD components that are fundamentally about databases, and several more that help you manage databases or that use (or can use) databases to accomplish their purpose.

GMOD's database related components are:

Chado

Chado is the modular database schema of GMOD. Chado is about organizing your data in a database so that you can manage it and can connect other GMOD components to it (either directly or via data exports). When someone speaks of the GMOD Schema they are speaking about Chado.

BioMart

BioMart is a data warehouse package tailored for biological data. It takes existing databases (for example, the FlyBase Chado database), transforms them into a data warehouse and then provides a web interface for supporting arbitrary queries against the data.

InterMine

InterMine also integrates multiple data sources into a single data warehouse. It has a core data model based on the sequence ontology and supports several biological data formats. It is easy to extend the data model and integrate your own data, Java and Perl APIs and an XML format to help import custom data. A web application allows creation of custom queries, includes template queries (web forms to run 'canned' queries) and can upload and operate on lists of data. Many aspects of the web app can be configured and branded.

Entrez and SRS

Entries in sequence databases can be thought of consisting in two main parts: the sequence itself and the information about the sequence such as any unique identifiers assigned to it, what organism it is from, who deposited it in the database and where it is referred to by entries in other databases. The information about the data is called metadata. Searching the metadata of entries in sequence and sequence-related databases can yield large amounts of information as well as sequence sets that can be used for further analysis. Two of the major interfaces allowing you to search the metadata of sequences (and carry out other analysis tasks as well) are the Sequence Retrieval Service (SRS) at the EBI and Entrez at the NCBI. In the associated lecture and practical, we will focus on using SRS. We will look at Entrez in the context of blast searching of sequence databases later in the course.

Possible questions

1. Define Bioinformatics? Explain the objectives, milestones and goals of Bioinformatics
2. What are biological databases? Discuss in detail
3. Write short notes on application of bioinformatics in
    a) Molecular medicine
    b) Microbial genomics
    c) Agriculture
    d) Biotechnology
4. Discuss in detail about human genome project add a note on its current developments.
5. Explain bibliographic databases and its uses.
6. Explain in details about the historical background of Bioinformatics
7. Write a short notes on Ethical, Legal and social issues in human genome project.
8. Define biological databases? Distinguish primary and secondary databases.
9. Discuss in brief about PROSITE.
10. Explain in brief about nucleic acid sequence databases.
11. Explain in detail about organism specific databases.

KAPRAGAM ACADEM YOF HIGHER EDUCATION

DEPARTMENT OF BIOCHEMISTRY

I- M.Sc Biochemistry

18BCP204-Bioinformatics

S. No	Unit	Questions	Option I	Option II	Option III	Option IV	Answer
1	1	The vector used for cloning the human genome fragment	YAC	Plasmid	Plasmid vector	Cosmid vector	**YAC**
2	1	______nucleotides make up the human genome	1,00,000	2 Billion	3 Billion	5000	**3 Billion**
3	1	Gene rich "Urban Centres" consisting predominantly nucleotides	A and T	G and C	A and C	T and G	**G and C**
4	1	GC regions can be seen through microscope on chromosomes as	White bands	Dark bands	Light bands	Both b and c	**Dark bands**
5	1	Gene Poor " Desert Regions " are predominated with nucleotides	G and C	A and T	G and A	C and T	**A and T**
6	1	Largest number of genes are seen in	Chromosome 1	Chromosome Y	Both a and b	Chromosome x	**Chromosome 1**
7	1	Repeated sequences that do not code for proteins is	RNA	tRNA	mRNa	Junk DNA	**Junk DNA**
8	1	AT region can be seen through the microscope on the chromosome as	Dark bands	Light bands	both a and b	very light band	**Dark bands**
9	1	Least number of genes are present in	Chromosome 1	Y chromosome	both a and b	Chromosome 2	**Y chromosome**
10	1	The single base difference is referred to as	SNP	PCR	NMR	PAM	**SNP**
11	1	Largest human genome consist of bases	6 million	1 lakh	2.4 million	3.4 million	**3.4 million**
12	1	NMR is published in the year	1985	1980	1972	1992	**1980**
13	1	The human genome consists of base	3 billion	30,000	1 lakh	2 lakh	**30,000**
14	1	____% of genes have unknown functions in human genome	25%	1%	50%	100%	**50%**
15	1	HGP is completed on	2001	2000	2003	2004	**2003**
16	1	The sources of X-ray crystallography are	X-tubes	rotating anode generator	synchronators	all the above	**all the above**
17	1	The human genome project was completed in the year of	2003	2005	2004	2002	**2003**
18	1	The first sequenced protein was	Insulin	Melanine	Heamoglobin	Keratin	**Insulin**
19	1	Genbank seated at	NIH	EMBL	DDBJ	PDB	**NIH**
20	1	Insulin consist of residues	52	51	53	54	**53**
21	1	Protein was first sequenced in	1954	1955	1956	1958	**1958**
22	1	The genome for *Saccharomyces cerevisiae*	1.1 Mb	12 Mb	12.1 Mb	10 Mb	**12.1 Mb**
23	1	The sequence of the first protein to be analysed, bovine insulin is announed by	1956	1955	1978	1986	**1955**
24	1	The ARPANET is created by linking computers at Standford and UCLA.	1969	1979	1986	1970	**1969**
25	1	The first recombinant DNA molecule is created by	Crick	Watson	Berg	Paul Berg	**Paul Berg**
26	1	The FASTA algorithm for sequence comparison is published by	Pauling	Peter	Pearson and Lipman	Berg	**Pearson and Lipman**
27	1	Microsoft Corporation is founded by Bill Gates and Paul Allen in	1975	1985	2005	2006	**1975**
28	1	DDBJ began in the year	1960	1975	1986	2003	**1986**
29	1	Prosite is	Primary Database	Secondary Database	Tertiary Database	All the above	**Secondary Database**
30	1	The NCBI is established at the national cancer institute in	1986	1988	1974	1994	**1988**
31	1	The links for NCBI is	www.ncbi.co.in	www.ncbi.nih.gov	www.ncbi.com	www.ncbi.nih.gov	www.ncbi.nih.gov
32	1	The links for EMBL is	www.embl.com	www.ebi.uk	www.ebi-ac.uk/embl	www.embl.ac.uk	www.ebi-ac.uk/embl
33	1	The links for DDBJ is	WWW.ddbj.nig.ac.jp	www.ddbj.com	www.ddbj.ac.in	www.ddbj.org	www.ddbj.nig.ac.jp
34	1	The link for SWISS PROT is	www.swissprot.com	www.expansy.org.ncbi	www.swiss.ac.in	www.expansy.com	**www.expansy.org.ncbi**
35	1	Swiss prot is a ______sequence database	Nucleotide	Protein	both a and b	None of the above	**Protein**
36	1	_______ is a principal DNA sequence database	NBRF	Gen bank	STS	None of the above	**Gen bank**
37	1	Translation of all coding sequences in EMBL is	Swissprot	DDBJ	TrEMBL	all of the above	**TrEMBL**
38	1	Print database is otherwise known as	Nucleotide database	Pattern Database	protein database	Structural database	**Pattern Database**
39	1	Trembl is used to convert	Protein sequence to nucleotide sequence	Translate nucleotide sequence into amino acids	Translate nucleotide to protein	both a and b	**Translate nucleotide sequence into amino acids**
40	1	Which electrophoresis is used in proteome databases	1D gel	2D gel	3D gel	SDS	**2D gel**
41	1	Protein sequence database was developed at	NBRF	NCBI	EMBL	NCBS	**NBRF**
42	1	The number of chromosomes in *Drosophila Melanogaste*	13 pairs	3 pairs	4 pairs	6 pairs	**13 pairs**
43	1	The *Drosophila melanogaster* genome is sequenced using ________ approach.	Shot-gun	High resolution and physical mapping	Parellel	Vertical	**Shot-gun**
44	1	*Saceharomyces cerevisiae* is also known as	Tape worm	Baker's yeast	House mouse	BAC	**Baker's yeast**
45	1	*Arabidopsis thaliana* is a member of the family.	Brassicaceae	Nematoda	Drosophilidae	Homo sapiens	**Brassicaceae**
46	1	The number of chromosomes of *Arabidopsis thaliana* is	5	10	15	20	**5**
47	1	*Arabidopsis thaliana* contains _______ bases	160 millions	150 millions	140 million	all the above	**150 millions**
48	1	Protein sequence database was developed at	NBRF	NCBI	EMBL	NCBS	**NBRF**
49	1	Expansion of PIR	Protein Information Research	Protein Information Resource	Protein Integral research	Protein Information Results	**Protein Information Resource**
50	1	DB is the major respiratory of ________ structures	DNA	RNA	Protein	All of the above	**All of the above**
51	1	The first secondary database have been developed was	PROSITE	PDB	PIR	GENBANK	**PDB**
52	1	With in PROSITE, Motifs are encoded as	Regular expression	Patterns	motif	fold	**Patterns**
53	1	SWISS-prot is a sequence _______ database	nucleotide	protein	OMIM	OMIA	**protein**
54	1	The nucleic acid sequence database are collections of	datas	queries	entites	Indices	**datas**
55	1	Fields are used to create for relational databases	Entities	Queries	Indices	codons	**Queries**
56	1	The first nucleic acid sequence was announced in the year	1954	1964	1956	1966	**1964**
57	1	The first secondary database have been developed was	PROSITE	PDB	PIR GENBANK	Motif	**PROSITE**
58	1	With in PROSITE,Motifs are encoded as	Regular expression	genes	Patterns	All of the above	**Patterns**
59	1	Entries are deposited in	GENBANK.	PROSITE.	PDB.	SWISS-PROT	**PDB.**
60	1	Comments are list of	Accesion of the numbers	Swiss Prot Identification	Codes of the true matches	Any possible matches which are often fragments	**Codes of the true matches**
61	1	The documentation files concludes with appropriate	Geographic references	Bibliographic References,	Structural References,	Statistical Refer	**Bibliographic References**
62	1	α–helix is disrupted by certain amino acid like	proline	arginine	histidine	lysine	**proline**
63	1	α–helix is stabilized by	hydrogen bonds	disulphide bonds	salt bridges	electrostatic bonds	**hydrogen bonds**
64	1	Gene that have arisen from a common ancestors is called	homologous	orthologous	pralogous	xenologous	**homologous**
66	1	______ is a short conserved pattern of amino acids.	motif	contigs	oligonucleotides	blocks	**motif**
66	1	______ are collections of overlapping sequence that are obtained in a sequencing project	motif	contigs	oligonucleotides	patterns	**contigs**
67	1	Protein sequence database was developed at	NBRF	NCBI	EMBL	NCBS	**NBRF**
68	1	Expansion of PIR	Protein Information Research	Protein Information Resourses	Protein Integral research	Protein Information Results	**Protein Information Resourses**
69	1	DB is the major respiratory of ______ structures	DNA	RNA	Protein	All of the above	**Protein**
70	1	The documentation files concludes with appropriate	Geographic references	Bibliographic References	Structural References	Statistical References	**Bibliographic References**
71	1	Motif finding, also known as profile analysis, constructs global MSAs that attempt to align short conserved among the sequences in the query set	DALI	MSA	sequence motifs	Fold	**sequence motifs**
72	1	Population of identical cells or molecules, derived from a single ancestor	cluster	sequence	clone	vector	**clone**
73	1	An organism's basic complement of DNA is called	Proteome	genome	gene	protein	**genome**
74	1	______contains experimentally determined 3D protein structures.	PIR	PDB	SWISSPROT	MMDB	**MMDB**
75	1	Local alignment was performed by	Pearson and Lipman	Smith and Waterman	Waterman	Crick	**Smith and Waterman**
76	1	A protein sequence database translated nucleotide sequences	TrEMBL	TrEXPASY	TrPDB	TrPIR	**TrEMBL**
77	1	The presence of more than one identical item represents	Alignment	conserved region	hypothetical region	redundancy	**redundancy**
78	1	In the sequence database, each sequence is an	Files	Entry	Query	All the above	**All the above**

UNIT-II

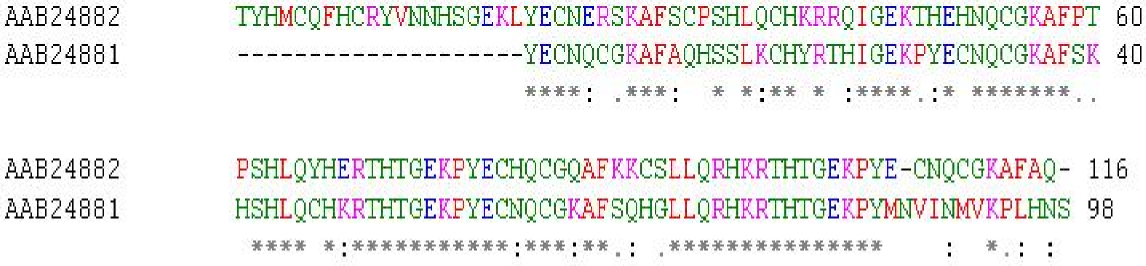
SYLLABUS

Introduction to sequence Alignment: Pairwise and multiple sequence alignment, substitution matrices, dynamic programming algorithms-Needleman and Wunsch and Smith Waterman, similarity searching programs, BLAST, FASTA, Multiple sequence alignment-CLUSTAL, Introduction and application of phylogenetic trees, basic terminologies, Phylogenetic analysis-PHYLIP theory of phylogeny, tree building methods.

INTRODUCTION TO SEQUENCE ALIGNMENT

Sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences.

Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix. Gaps are inserted between the residues so that identical or similar characters are aligned in successive columns.



```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                    ****: .***:  * *:** * :****.:* *******..

AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ- 116
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS 98
                 **** *:***********:***:**.: .***************    :  *.: :
```

A sequence alignment, produced by ClustalW, of two human zinc finger proteins, identified on the left by GenBank accession number.

Single letters: amino acids.

Red: small, hydrophobic, aromatic, not Y.

Blue: acidic. Magenta: basic.

Green: hydroxyl, amine, amide, basic.

Gray: others. "*": identical. ":": conserved substitutions (same colour group). ".": semi-conserved substitution (similar shapes).

Definition of sequence alignment

- Sequence alignment is the procedure of comparing two (pair-wise alignment) or more multiple sequences by searching for a series of individual characters or patterns that are in the same order in the sequences.
- There are two types of alignment: local and global. In global alignment, an attempt is made to align the entire sequence. If two sequences have approximately the same length and are quite similar, they are suitable for the global alignment.
- Local alignment concentrates on finding stretches of sequences with high level of matches**.**

L G P S S K Q T G K G S - S R I W D N

Global alignment

L N - I T K S A G K G A I M R L G D A

- - - - - - - T G K G - - - - - - - -

Local alignment

- - - - - - A G K G - - - - - - - -

Interpretation of sequence alignment

- Sequence alignment is useful for discovering structural, functional and evolutionary information.
- Sequences that are very much alike may have similar secondary and 3D structure, similar function and likely a common ancestral sequence. It is extremely unlikely that such sequences obtained similarity by chance. For DNA molecules with *n* nucleotides such probability is very low *P = 4-n*. For proteins the probability even much lower *P = 20 –n*, where *n* is a number of amino acid residues

- Large scale genome studies revealed existence of horizontal transfer of genes and other sequences between species, which may cause similarity between some sequences in very distant species

Methods of Sequence Alignment

- Dot Matrix analysis
- Dynamic Programming (DP) algorithm
- Word (or) K-tuple methods

Dot matrix analysis

- Comparing for two sequence
- One sequence (A) top of the matrix
- Other one sequence (B) down left side
- Any region of similarity is revealed by a diagonal row of dots.
- Five clear diagonals
- Diagonals are obtained by aligning genomic and cDNA.
- Five diagonals represent the five exons of the gene which was confirmed from the annotated entry of the gene.

Sequence alignment program is to align the two sequences

- To produce highest score a scoring matrix is used to add points to the score for each match and subtract them for each mismatch.
- Matrixes are used for nucleic acid alignment to involve fairly simple match/mismatch scoring schemes.

Parameters used for sequence alignment

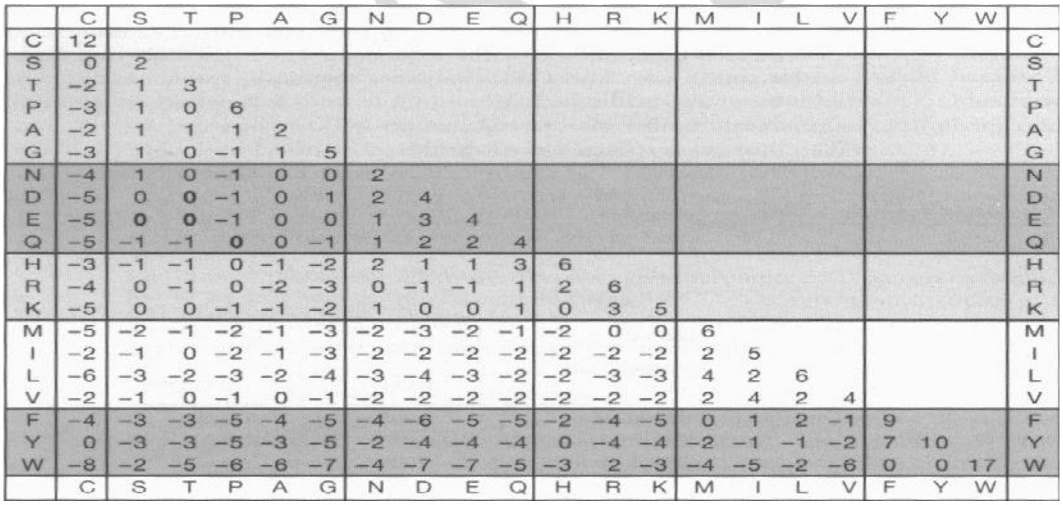
1. scoring matrix
2. Substitution matrices
3. Gap penalty

Scoring matrices

- It is critical to have reasonable scoring schemes accepted by the scientific community for DNA and proteins and for different types of alignments

- The wealth of information accumulated in the gene/protein banks was utilised with dynamic programming procedure to create such matrices for scoring matches and separately penalties for gaps introduction and extensions
- Matrices for DNA are rather similar as there are only two options purine & pyrimidine and match & mismatch
- Proteins are much more complex and the number of option is significant
- PAM and other matrices are represented in log odds scores, which is the ratio of chance of amino acid substitution due to essential biological reason to the chance of random substitution
- There are many different PAMs, which are representing different evolutionary scenarios.
- PAM 250 represents a level of 250% of changes expected in 2500 MY
- PAM is more suitable for studying quite distant proteins, BLOSUM is for more conserved proteins of domains

Scoring matrices: PAM (Percent Accepted Mutation)

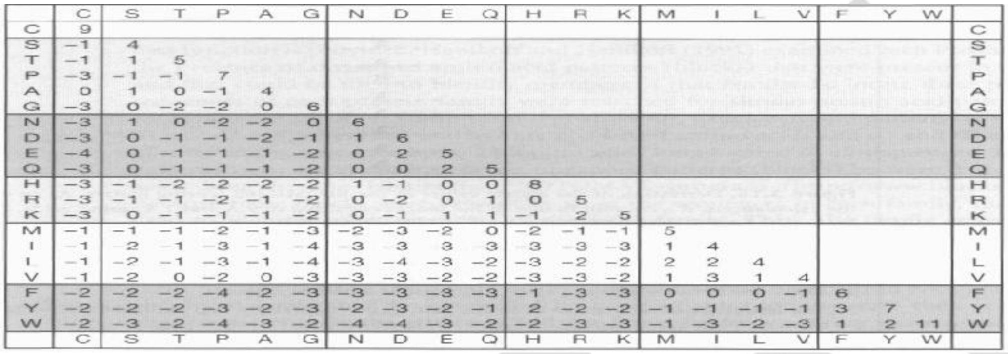


	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	12																				C
S	0	2																			S
T	-2	1	3																		T
P	-3	1	0	6																	P
A	-2	1	1	1	2																A
G	-3	1	0	-1	1	5															G
N	-4	1	0	-1	0	0	2														N
D	-5	0	0	-1	0	1	2	4													D
E	-5	0	0	-1	0	0	1	3	4												E
Q	-5	-1	-1	0	0	-1	1	2	2	4											Q
H	-3	-1	-1	0	-1	-2	2	1	1	3	6										H
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6									R
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5								K
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6							M
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5						I
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6					L
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4				V
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9			F
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10		Y
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

- Amino acids are grouped according to to the chemistry of the side group: (C) sulfhydryl, (STPAG)-small hydrophilic, (NDEQ) acid, acid amide and hydrophilic, (HRK) basic, (MILV) small hydrophobic, and (FYW) aromatic.

- Log odds values: +10 means that ancestor probability is greater, 0 means that the probabilities are equal, -4 means that the change is random. Thus the probability of alignment YY/YY is 10+10=20, whereas YY/TP is –3-5=-8, a rare and unexpected between homologous sequences.

Scoring matrices: BLOSUM62 (BLOcks amino acid SUbstitution Matrices)



	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	−1	4																			S
T	−1	1	5																		T
P	−3	−1	−1	7																	P
A	0	1	0	−1	4																A
G	−3	0	−2	−2	0	6															G
N	−3	1	0	−2	−2	0	6														N
D	−3	0	−1	−1	−2	−1	1	6													D
E	−4	0	−1	−1	−1	−2	0	2	5												E
Q	−3	0	−1	−1	−1	−2	0	0	2	5											Q
H	−3	−1	−2	−2	−2	−2	1	−1	0	0	8										H
R	−3	−1	−1	−2	−1	−2	0	−2	0	1	0	5									R
K	−3	0	−1	−1	−1	−2	0	−1	1	1	−1	2	5								K
M	−1	−1	−1	−2	−1	−3	−2	−3	−2	0	−2	−1	−1	5							M
I	−1	−2	−1	−3	−1	−4	−3	−3	−3	−3	−3	−3	−3	1	4						I
L	−1	−2	−1	−3	−1	−4	−3	−4	−3	−2	−3	−2	−2	2	2	4					L
V	−1	−2	0	−2	0	−3	−3	−3	−2	−2	−3	−3	−2	1	3	1	4				V
F	−2	−2	−2	−4	−2	−3	−3	−3	−3	−3	−1	−3	−3	0	0	0	−1	6			F
Y	−2	−2	−2	−3	−2	−3	−2	−3	−2	−1	2	−2	−2	−1	−1	−1	−1	3	7		Y
W	−2	−3	−2	−4	−3	−2	−4	−4	−3	−2	−2	−3	−3	−1	−3	−2	−3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Ideology of BLOSUM is similar but it is calculated from a very different and much larger set of proteins, which are much more similar and create blocks of proteins with a similar pattern.

Differences between PAM and BLOSUM

1. PAM matrices are based on an explicit evolutionary model (i.e. replacements are counted on the branches of a phylogenetic tree), whereas the BLOSUM matrices are based on an implicit model of evolution.
2. The PAM matrices are based on mutations observed throughout a global alignment, this includes both highly conserved and highly mutable regions. The BLOSUM matrices are based only on highly conserved regions in series of alignments forbidden to contain gaps.
3. The method used to count the replacements is different: unlike the PAM matrix, the BLOSUM procedure uses groups of sequences within which not all mutations are counted the same.
4. Higher numbers in the PAM matrix naming scheme denote larger evolutionary distance, while larger numbers in the BLOSUM matrix naming scheme denote higher sequence similarity and therefore smaller evolutionary distance. Example: PAM150

is used for more distant sequences than PAM100; BLOSUM62 is used for closer sequences than Blosum50.

Substitution matrices

- 210 score possibilities for any possible protein pair.
- 20X20 matrix, where the diagonal gives 100% match between the amino acids.
- Main diagonal are of identical 20 amino acid scores and on each side of diagonal 190 scoring that are similar obtain 210 scoring terms for 20 amino acid combinations.
- Pair of amino acid is termed as log – odds values and these have been scaled and rounded to the nearest integer for computational efficiency known as score matrix or substitution matrix.
- The late Margaret Dayhoff pioneer in protein data basing and comparison
- Dayhoff, MDM (Mutation Data Matrix] (or) PAM (Point (or) Percent Accepted Mutation).
- PAM one such major amino acid scoring or substitution matrix
- BLOSUM series of matrices were created by Steve Henikoff and colleagues.
- Matrices are used BLOSUM 80, 62, 40 and 30.
- PAM matrices used are PAM 120, 160, 250 and 350 matrices.

80 – 100 %	Sequence	identity	BLOSUM80
60 – 80 %	Sequence	identity	BLOSUM62
30 – 60 %	Sequence	identity	BLOSUM45
0 – 30 %	Sequence	identity	BLOSUM30
80 – 100 %	Sequence	identity	PAM20
60 – 80 %	Sequence	identity	PAM60
40 – 60 %	Sequence	identity	PAM120
0 – 40 %	Sequence	identity	PAM350

Gap penalty

- Gap is any maximal consecutive run of spaces in a single string of a given alignment.

- Gap helps to create alignments that better conform to underlying biological models and more closely fit patterns that one excepts to final in meaningful alignment.
- No. of continuous gaps and not only the number of spaces when calculating an alignment mark.

Example

X = a t t c - - g a-t g g a c c
Y = a - - c g t g a t t - - -c c

- Tour gaps containing a total of eight spaces
- 7 matches, no mismatch.
- No. of gaps in the alignment will be denoted as # gaps

Pairwise Sequence Alignment:

- The two sequences are homologous, i.e. they have evolved from a common ancestor.
- Differences between them are due to only two kinds of events, namely insertion
- deletions (*indels*) and substitutions (change of single elements of the sequence -
- amino acids if the sequence is a protein and nucleic acid, if the sequence is DNA).

Two types pairwise sequence alignment

- Needlemen-Wunsch Alogorithm (or) Global Alignment
- Smith-Waterman (or) Local Alignment

Needleman–Wunsch algorithm

- The **Needleman–Wunsch algorithm** performs a global alignment on two sequences (called *A* and *B* here).
- It is commonly used in bioinformatics to align protein or nucleotide sequences.
- The algorithm was published in 1970 by Saul B. Needleman and Christian D. Wunsch.
- The Needleman–Wunsch algorithm is an example of dynamic programming, and was the first application of dynamic programming to biological sequence comparison.

A modern presentation

Scores for aligned characters are specified by a similarity matrix. Here, $S(a,b)$ is the similarity of characters a and b. It uses a linear gap penalty, here called d.

For example, if the similarity matrix were then the alignment.

	A	**G**	**C**	**T**
A	10	-1	-3	-4
G	-1	7	-5	-3
C	-3	-5	9	0
T	-4	-3	0	8

then the alignment:

AGACTAGTTAC
CGA–––GACGT

with a gap penalty of -5, would have the following score:

To find the alignment with the highest score, a two-dimensional array (or matrix) F is allocated. The entry in row i and column j is denoted here by F_{ij}. There is one column for each character in sequence A, and one row for each character in sequence B. Thus, if we are aligning sequences of sizes n and m, the amount of memory used is in $O(nm)$. (Hirschberg's algorithm can compute an optimal alignment in $\Theta(\min\{n,m\})$ space, roughly doubling the running time.

Dotplots

- The most intuitive representation of the comparison between two sequences uses dot-plots.
- One sequence is represented on each axis and significant matching regions are distributed along diagonals in the matrix.

Exercise: Making a dotplot

unix % **dottup**

DNA sequence dot plot

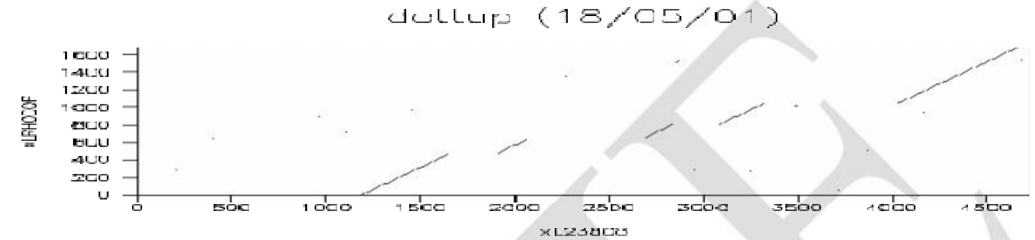
Input sequence: **embl:xl23808**

Second sequence: **embl:xlrhodop**

Word size [4]: **10**

Graph type [x11]:

A window will pop up on your screen that should look something like this:



- The diagonal lines represent areas where the two sequences align well. You can see that there are five clear diagonals.
- Aligning genomic and cDNA - these five diagonals represent the five exons of the gene! If you look at the original EMBL entry for the genomic sequence using SRS, you will see that the annotated entry says that there are five exons in this gene. So our results are in agreement.
- The settings we have used for this example are those that give the best results. dottup looks for exact matches between sequences.
- As we expect the exon regions from the genomic sequence to exactly match the cDNA sequence we can use longer word lengths as we should still get exact matches.
- This gives a very clean plot. If you were to match the cDNA sequence against that of a related sequence, e.g. the rhodopsin from mouse (embl: m55171) then you wouldn't expect long exact matches so should use a shorter word length.

Smith–Waterman algorithm

- The Smith–Waterman algorithm is a well-known algorithm for performing local sequence alignment; that is, for determining similar regions between two 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

Algorithm Explanation

A matrix *H* is built as follows:

if $a_i = b_j$ $w(a_i,b_j) = w(\text{match})$ or if $a_i! = b_j$ $w(a_i,b_j) = w(\text{mismatch})$

Where:

a,b = Strings over the Alphabet Σ

m = length(a)

n = length(b)

$H(i,j)$ - is the maximum Similarity-Score between a suffix of a[1...i] and a suffix of b[1...j]

, '-' is the gap-scoring scheme

Example

Sequence 1 = ACACACTA

Sequence 2 = AGCACACA

w(gap) = 0

w(match) = +2

$w(a,-) = w(-,b) = w(mismatch) = -1$

To obtain the optimum local alignment, we start with the highest value in the matrix (i,j). Then, we go backwards to one of positions (i-1,j), (i,j-1), and (i-1,j-1) depending on the direction of movement used to construct the matrix. We keep the process until we reach a matrix cell with zero value, or the value in position (0,0).

In the example, the highest value corresponds to the cell in position (8,8). The walk back corresponds to (8,8), (7,7), (7,6), (6,5), (5,4), (4,3), (3,2), (2,1), (1,1), and (0,0),

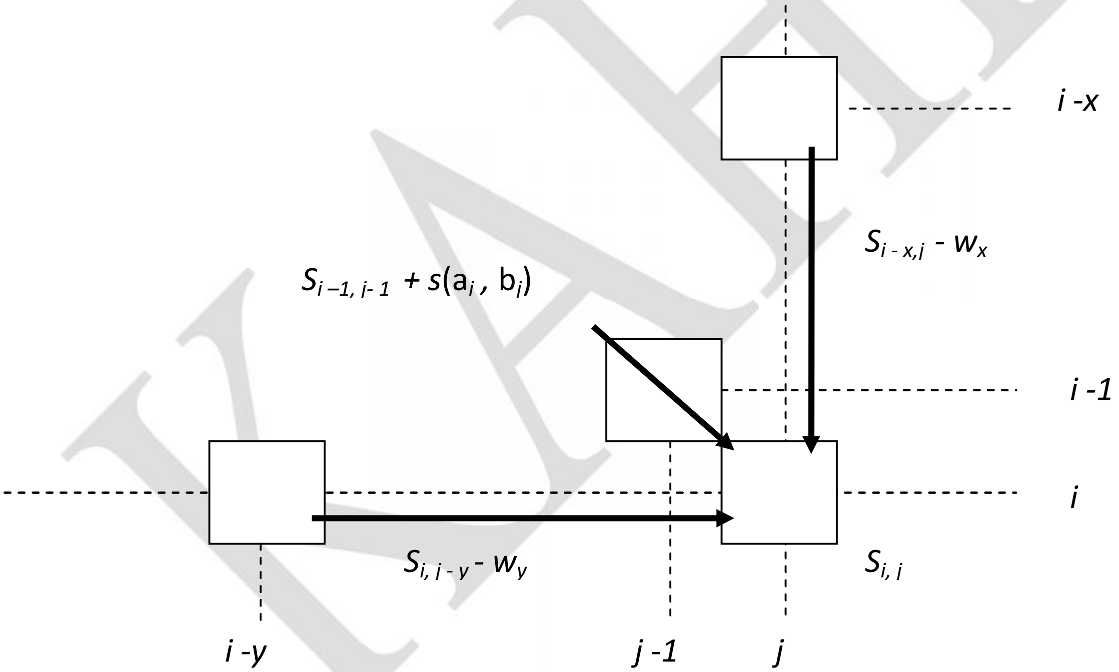
Once we've finished, we reconstruct the alignment as follows: Starting with the last value, we reach (i,j) using the previously-calculated path. A diagonal jump implies there is an alignment (either a match or a mismatch). A top-down jump implies there is a deletion. A left-right jump implies there is an insertion.

Description of the dynamic programming algorithm

- Consider building this alignment in steps, starting from the initial match (V/V) and then sequentially adding a new pair until the alignment is complete, at each stage choosing a pair from all the possible matches that provides the highest score for the alignment up to that point.

- If the full alignment has the highest possible (or optimal) score, then the old alignment from which it was derived (A) by addition of the aligned Y/Y pair must also have been optimal up to that point in the alignment.
- In this manner, the alignment can be traced back to the first aligned pair that was also an optimal alignment.
- The example, which we have considered, illustrates 3 choices: 1. Match the next character(s) in the following position(s); 2. Match the next character(s) to a gap in the upper sequence; 3. Add a gap in the lower sequence.

Formal description of dynamic programming algorithm



- This diagram indicates the moves that are possible to reach a certain position (*i,j*) starting from the previous row and column at position (*i -1, j-1*) or from any position in the same row or column
- Diagonal move with no gap penalties or move from any other position from column *j* or row *i*, with a gap penalty that depends on the size of the gap

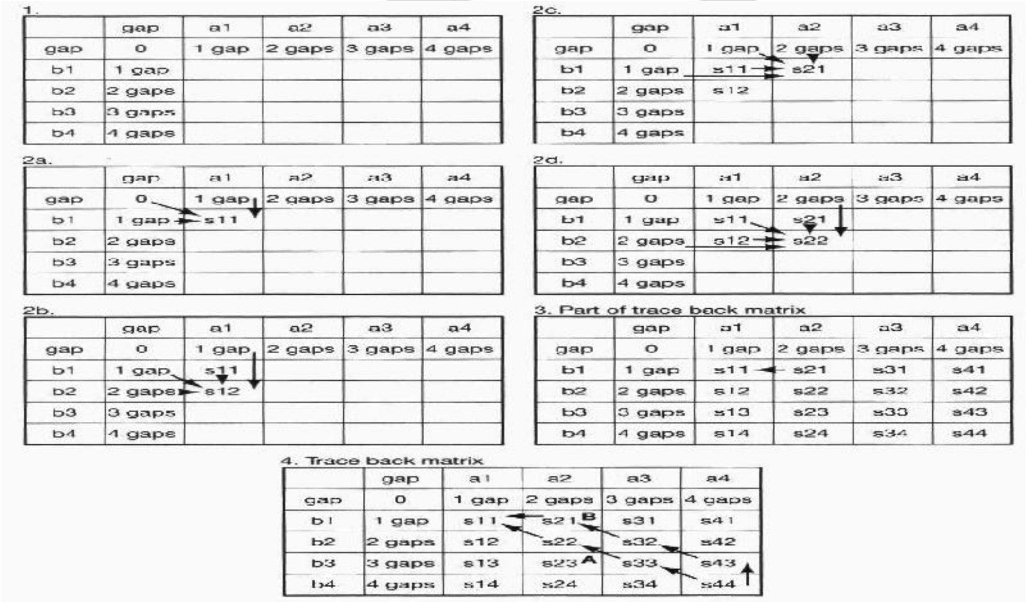
Formal description of dynamic programming algorithm

For two sequences **a** = a1, a2,..a*i* and **b** = b1, b2, ..b*j*, where *Sij* = *S* (a1,…a*i*, b1,…b*j*) then

$$S_{ij} = \max \{ S_{i-1,\, j-1} + s(a_i b_j),$$
$$\max (S_{i-x,\, j} - w_x),$$
$$x \geq 1$$

where S*ij* is the score at position at *i* in sequence **a** and *j* in sequence **b**, s(a*i*b*j*) is score for aligning the character at positions *i* and *j*, *wx* is the penalty for a gap of length *x* in sequence **a**, and *wx* is the penalty for a gap of length *y* in sequence **b**.

Note that S*ij* is a type of running best score as the algorithm moves through every position in the matrix



1.

	gap	a1	a2	a3	a4
gap	0	1 gap	2 gaps	3 gaps	4 gaps
b1	1 gap				
b2	2 gaps				
b3	3 gaps				
b4	4 gaps				

2a.

	gap	a1	a2	a3	a4
gap	0	1 gap	2 gaps	3 gaps	4 gaps
b1	1 gap	s11			
b2	2 gaps				
b3	3 gaps				
b4	4 gaps				

2b.

	gap	a1	a2	a3	a4
gap	0	1 gap	2 gaps	3 gaps	4 gaps
b1	1 gap	s11			
b2	2 gaps	s12			
b3	3 gaps				
b4	4 gaps				

2c.

	gap	a1	a2	a3	a4
gap	0	1 gap	2 gaps	3 gaps	4 gaps
b1	1 gap	s11	s21		
b2	2 gaps	s12			
b3	3 gaps				
b4	4 gaps				

2d.

	gap	a1	a2	a3	a4
gap	0	1 gap	2 gaps	3 gaps	4 gaps
b1	1 gap	s11	s21		
b2	2 gaps	s12	s22		
b3	3 gaps				
b4	4 gaps				

3. Part of trace back matrix

	gap	a1	a2	a3	a4
gap	0	1 gap	2 gaps	3 gaps	4 gaps
b1	1 gap	s11	s21	s31	s41
b2	2 gaps	s12	s22	s32	s42
b3	3 gaps	s13	s23	s33	s43
b4	4 gaps	s14	s24	s34	s44

4. Trace back matrix

	gap	a1	a2	a3	a4
gap	0	1 gap	2 gaps	3 gaps	4 gaps
b1	1 gap	s11	s21 B	s31	s41
b2	2 gaps	s12	s22	s32	s42
b3	3 gaps	s13	s23 A	s33	s43
b4	4 gaps	s14	s24	s34	s44

Alignment A: a1 a2 a3 a4
b1 b2 b3 b4

Alignment B: a1 a2 a3 a4 -

b1 - b2 b3 b4

The highest scoring matrix position is located (in this case s44) and then traced back as far as possible, generating the path shown.

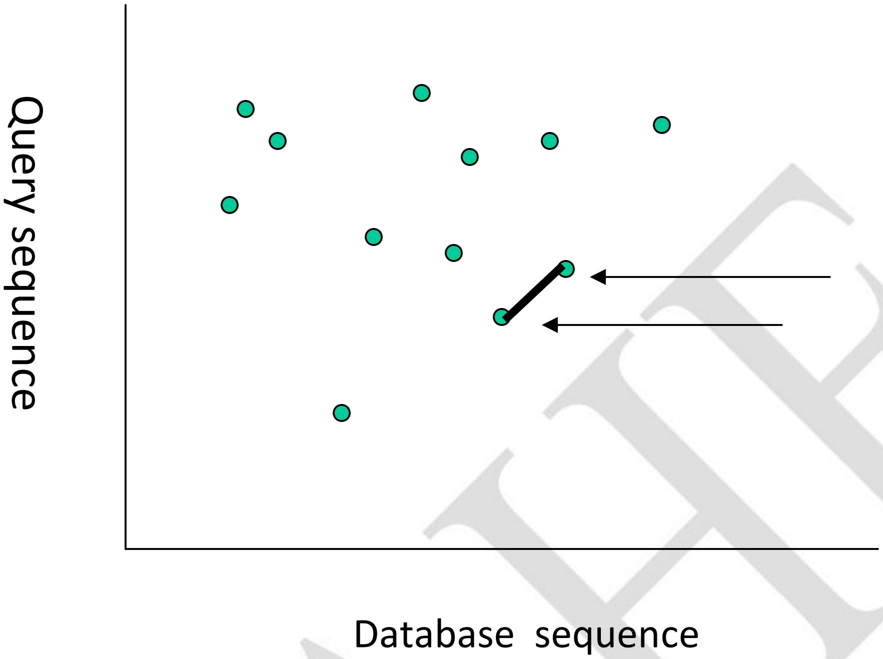
BLAST

- BLAST (**B**asic **L**ocal **A**lignment **S**earch **T**ool) comes under the category of homology and similarity tools.
- It is a set of search programs designed for the Windows platform and is used to perform fast similarity searches regardless of whether the query is for protein or DNA.
- Comparison of nucleotide sequences in a database can be performed. Also a protein database can be searched to find a match against the queried protein sequence.
- NCBI has also introduced the new queuing system to BLAST (Q BLAST) that allows users to retrieve results at their convenience and format their results multiple times with different formatting options.

BLAST procedure

- The steps used by the BLAST algorithm:
- The seq is optionally filtered to remove low-complexity regions (AGAGAG…)
- A list of words of certain length is made
- Using substitution scores matrixes (like PAM or BLOSUM62) the query seq. words are evaluated for matches with any DB seq. and these scores (log) are added
- A cutoff score (T) is selected to reduce number of matches to the most significant ones.
- The above procedure is repeated for each word in the query seq.
- The remaining high-scoring words are organised into efficient search tree and rapidly compared to the DB seq.
- If a good match is found then an alignment is extended from the match area in both directions as far as the score continue to grow. In the latest version of BLAST more time-efficient method is used

- The essence of this method is finding a diagonal connecting ungapped alignments and extending them



- The next step is to determine those high scoring pairs (HSP) of seq., which have score greater than a cutoff score (*S*). S is determined empirically by examining a range of scores found by comparing random seq. and by choosing a value that is significantly greater.
- Then BLAST determines statistical significance of each HSP score. The probability *p* of observing a score *S* equal to or greater than *x* is given by the equation: p (S ñ x) = 1 – exp(-e-λ(x-u)), where u = [log (Km'n')]/λ and K and λ are parameters that are calculated by BLAST for amino acid or nucleotide substitution scoring matrix, n' is effective length of the query seq. and m' is effective length of the database seq.
- On the next step a statistical assessments is made in the case if two or more HSP regions are found and certain matching pairs are put in descending order in the output file as far as their similarity/ score is concerned.

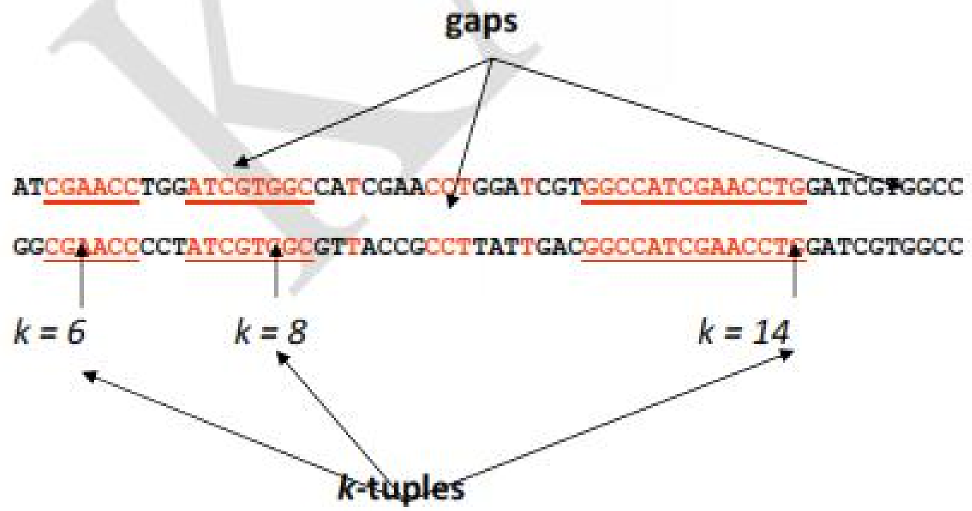
Depending on the type of sequences to compare, there are different programs:

- **blastp** compares an amino acid query sequence against a protein sequence database

- **blastn** compares a nucleotide query sequence against a nucleotide sequence database
- **blastx** compares a nucleotide query sequence translated in all reading frames against a protein sequence database
- **tblastn** compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames
- **tblastx** compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database

FASTA

- FASTA is a program for rapid alignment of pairs of protein and DNA sequences.
- Comparison of all nucleotides or amino acids is not an option, even for powerful
- computers, FASTA instead searches for matching sequence patterns ("words")called *k*-tuples. These patterns comprise *k* consecutive matches in the compared sequences.
- Using *k*-tuples FASTA builds a local alignment.
- Finally FASTA scores this alignment and output a list of sequences similar to a query in the descending order.


gaps
ATCGAACCTGGATCGTGGCCATCGAACCTGGATCGTGGCCATCGAACCTGGATCGTGGCC
GGCGAACCCCTATCGTGGCGTTACCGCCTTATTGACGGCCATCGAACCTGGATCGTGGCC
k = 6
k = 8
k = 14
k-tuples

FASTA performs the following statistical tasks: 1. The average score for DB seq. of the same length is determined, 2. The average score is plotted against the log of average seq. length in each length range, 3. The points are then fitted to a straight line by linear regression, 4. A z score, the number of standard deviations from fitted line, is calculated for each score, 5. Low scoring seq. are removed. 6. A statistical comparison with Z distribution follows, which allows to calculate E ( ) value. If E ( ) = 0, and z score is high two sequences are identical, when E is higher then a threshold level, no clear similarity is observed**.**

Methods used by FASTA to locate sequence similarity:

A. Rapid location of 10 best matching regions in each pair. For DNA seq. k = 4-6, for protein k = 1-2. The highest-density matches identified.

B. The highest-density regions are evaluated using special scoring matrixes (next lecture) and the best initial regions (INIT1) are found (*-the best).

C. Longer regions of identity of score INITN are generated by joining INIT with scores higher than a certain threshold, which include positive scores for similarity and negative for gaps. Optimization procedure follows.

Typical output of FASTA similarity search

Query – Motif2; #282 – is a fragment from a DB

```
>>#282          (18 aa)initn:   48  init1:   48  opt:   71 z-score: 191.0 E(): 6.9e-06
Smith-Waterman score: 71;   61.111% identity in 18 aa overlap
10      20
Motif2 VKTYGFAATSVEEAKEVAEERGK
X:.::.::X.:.:.. :
#282      GFVATSAEEAEEIAKKLG
10
```

Multiple Sequence Alignment

- Often applied to proteins
- Proteins that are similar in sequence are often similar in structure and function.
- Sequence changes more rapidly in evolution than does structure and function

Overview of Methods

- Dynamic programming – too computationally expensive to do a complete search; uses heuristics
- Progressive – starts with pair-wise alignment of most similar sequences; adds to that
- Iterative – make an initial alignment of groups of sequences, adds to these (e.g. genetic algorithms)
- Locally conserved patterns
- Statistical and probabilistic methods

Dynamic Programming

- Computational complexity – even worse than for pair-wise alignment because we're finding all the paths through an n-dimensional hyperspace (We can picture this in 2 or 3 dimensions.)
- Can align about 7 relatively short (200-300) protein sequences in a reasonable amount of time; not much beyond that.
- Let's picture this in 3 dimensions (pp. 146-157 in book). It generalizes to n.
- Consider the pair-wise alignments of each pair of sequences.
- Create a phylogenetic tree from these scores.
- Consider a multiple sequence alignment built from the phylogenetic tree.
- These alignments circumscribe a space in which to search for a good (but not necessarily optimal) alignment of all n sequences.
- Create a phylogenetic tree based on pair-wise alignments (Pairs of sequences that have the best scores are paired first in the tree.)
- Do a "first-cut" msa by incrementally doing pair-wise alignments in the order of "alikeness" of sequences as indicated by the tree. Most alike sequences aligned first.
- Use the pair-wise alignments and the "first-cut" msa to circumscribe a space within which to do a full msa that searches through this solution space.

- The score for a given alignment of all the sequences is the sum of the scores for each pair, where each of the pair-wise scores is multiplied by a weight ϵ indicating how far the pair-wise score differs from the first-cut msa alignment score.
- Does not guarantee an optimal alignment of all the sequences in the group.
- Does get an optimal alignment within the space chosen.

Phylogenetic Tree

- Dynamic programming uses a phylogenetic tree to build a "first-cut" msa
- The tree shows how protein could have evolved from shared origins over evolutionary time.
- See page 143 in *Bioinformatics* by Mount.
- Chapter 6 goes into detail on this.

Progressive Methods

- Similar to dynamic programming method in that it uses the first step (i.e., it creates a phylogenetic tree, aligns the most-alike pair, and incrementally adds sequences to the alignment in order of "alikeness" as indicated by the tree.)
- Differs from dynamic programming method for MSA in that it doesn't refine the "first-cut" MSA by doing a full search through the reduced search space. (This is the computationally expensive part of DP MSA in that, even though we've cut down the search space, it's still big when we have many sequences to align.)
- Generally proceeds as follows:
  - Choose a starting pair of sequences and align them
  - Align each next sequence to those already aligned, one at a time
- Heuristic method – doesn't guarantee an optimal alignment

ClustalW

- Based on phylogenetic analysis
- A phylogenetic tree is created using a pairwise distance matrix and nearest-neighbor algorithm

- The most closely-related pairs of sequences are aligned using dynamic programming
- Each of the alignments is analyzed and a profile of it is created
- Alignment profiles are aligned progressively for a total alignment
- W in ClustalW refers to a weighting of scores depending on how far a sequence is from the root on the phylogenetic tree.

Problems with Progressive Method

- Highly sensitive to the choice of initial pair to align. If they aren't very similar, it throws everything off.
- It's not trivial to come up with a suitable scoring matrix or gap penaties.

Iterative Methods for Multiple Sequence Alignment

- Get an alignment.
- Refine it.
- Repeat until one msa doesn't change significantly from the next.
- An example is genetic algorithm approach

Genetic Algorithms

- A general problem solving method modeled on evolutionary change.
- Create a set of candidate solutions to your problem, and cause these solutions to evolve and become more and more fit over repeated generations.
- Use survival of the fittest, mutation, and crossover to guide evolution.

Evolutionary Change in Genetic Algorithms

- survival of the fittest – the best solutions survive and reproduce to the next generation
- mutation – some solutions mutate in random ways (but they must always remain viable solutions)
- crossover – solutions "exchange parts"

Laying Out the Problem

- What would a candidate solution look like in a multiple sequence alignment program? (an MSA of ~20 proteins)
- How many candidate solutions should there be? (~100)

Evolving to a Next Generation

- Which candidate solutions should survive to the next generation?
  - First, take the top half based on best sum of pairs scores
  - Then randomly select second half, giving more chance to an MSA's being selected in proportion to how good its score is.

Phylogenetic tree

- A **phylogenetic tree** or **evolutionary tree** is a branching diagram or "tree" showing the inferred evolutionary relationships among various biological species or other entities based upon similarities and differences in their physical and/or genetic characteristics.
- The taxa joined together in the tree are implied to have descended from a common ancestor.
- In a **rooted** phylogenetic tree, each node with descendants represents the inferred most recent common ancestor of the descendants and the edge lengths in some trees may be interpreted as time estimates.
- Each node is called a taxonomic unit.
- Internal nodes are generally called hypothetical taxonomic units (HTUs) as they cannot be directly observed.
- Trees are useful in fields of biology such as systematics and comparative phylogenetics.

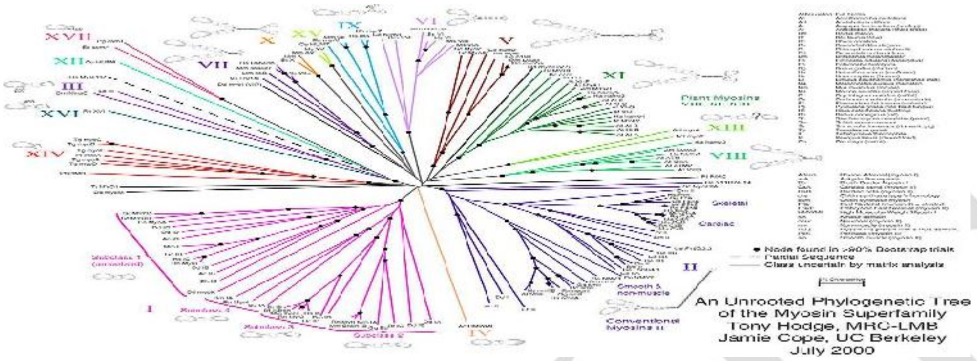
Types

A rooted phylogenetic tree

A rooted tree is used to make inferences about the most common ancestor of the leaves or branches of the tree. Most commonly the root is referred to as an "outgroup"

Unrooted tree:

An unrooted tree is used to make an illustration about the leaves or branches, but not make assumption regarding a common ancestor.



Total rooted trees and total unrooted trees, where *n* represents the number of leaf nodes. Among labeled bifurcating trees, the number of unrooted trees with *n* leaves is equal to the number of rooted trees with $n - 1$ leaves.

dendrogram is a broad term for the diagrammatic representation of a phylogenetic tree.

A **cladogram** is a tree formed using cladistic methods. This type of tree only represents a branching pattern, i.e., its branch lengths do not represent time.

A **phylogram** is a phylogenetic tree that explicitly represents number of character changes through its branch lengths.

A **chronogram** is a phylogenetic tree that explicitly represents evolutionary time through its branch lengths.

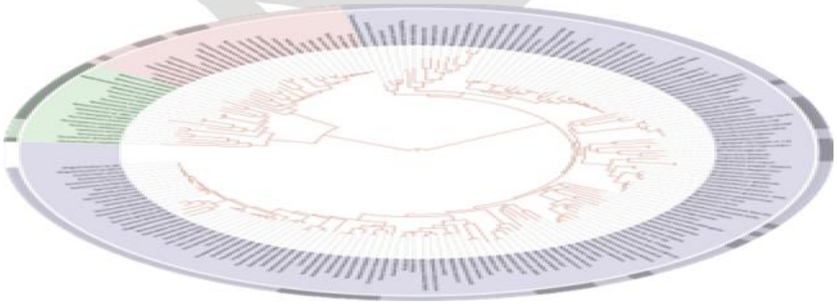
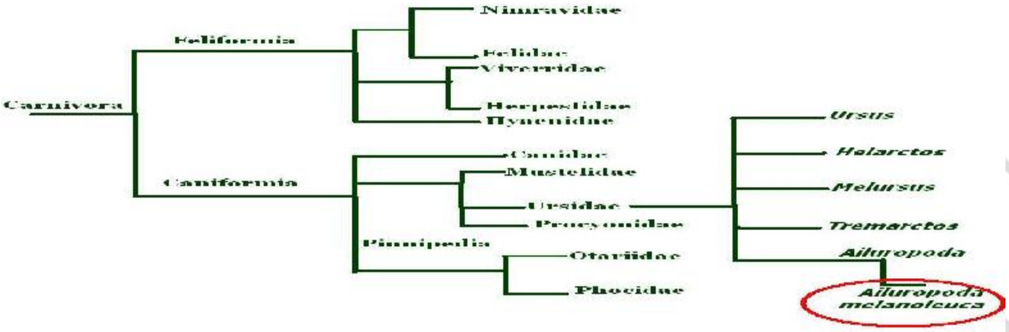
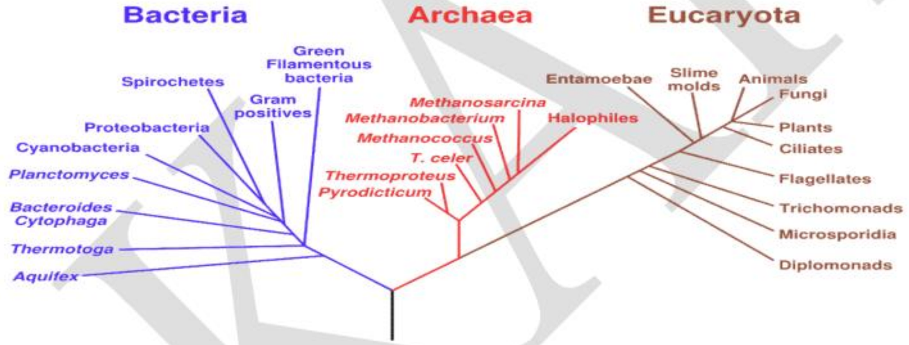


Fig. 2: A highly resolved, automatically generated Tree Of Life, based on completely sequenced genomes.

The agglomerative hierarchical clustering algorithms available in this program module build a cluster hierarchy that is commonly displayed as a tree diagram called a dendrogram. They begin with each object in a separate cluster. At each step, the two clusters that are most similar are joined into a single new cluster. Once fused, objects are never separated. The eight methods that are available represent eight methods of defining the similarity between clusters.



A phylogenetic tree, showing how Eukaryota and Archaea are more closely related to each other than to Bacteria, based on Cavalier-Smith's theory of bacterial evolution.



Tree-building methods can be assessed on the basis of several criteria:

- efficiency
- power
- consistency
- robustness
- falsifiability
- Tree-building techniques have also gained the attention of mathematicians. Trees can also be built using T-theory.

Limitations

- It is important to remember that trees do have limitations. For example, trees are meant to provide insight into a research question and not intended to represent an entire species history.
- Several factors, like gene transfers, may affect the output placed into a tree.
- All knowledge of limitations related to DNA degradation over time must be considered, especially in the case of evolutionary trees aimed at ancient or extinct organisms.

Construction

Phylogenetic trees composed with a nontrivial number of input sequences are constructed using computational phylogenetics methods. Distance-matrix methods such as neighbor-joining or UPGMA, which calculate genetic distance from multiple sequence alignments, are simplest to implement, but do not invoke an evolutionary model. Many sequence alignment methods such as ClustalW also create trees by using the simpler algorithms (i.e. those based on distance) of tree construction. Maximum parsimony is another simple method of estimating phylogenetic trees, but implies an implicit model of evolution (i.e. parsimony). More advanced methods use the optimality criterion of maximum likelihood, often within a Bayesian Framework, and apply an explicit model of evolution to phylogenetic tree estimation.[4] Identifying the optimal tree using many of these techniques is NP-hard,[4] so heuristic search and optimization methods are used in combination with tree-scoring functions to identify a reasonably good tree that fits the data.

Neighbor Joining or UPGMA

UPGMA and Neighbor Joining use a clustering procedure that is commonly found in data mining techniques. The method is simple and intuitive which makes it appealing. The method works by clustering nodes at each stage and then forming a new node on a tree. This process continues from the bottom of the tree and in each step a new node is added, and the tree grows upward. The length of the branch at each step is determined by the difference in heights of the nodes at each end of the branch. UPGMA has built in assumptions that the tree is additive and that all nodes are equally distance from the root. Since a "molecular clock" hypothesis assumption poses biological issues, UPGMA is not used much today, but gave way to a very common approach now termed "Neighbor Joining" Neighbor Joining (NJ) works like UPGMA in that it creates a new distance matrix at each step, and creates the tree based on the matrices. The difference is that NJ does not construct clusters but directly calculates distances to internal nodes. The first step in the NJ algorithm is to create a matrix with the Hamming distance between each node or taxa. The minimal distance is then used to calculate the distance from the two nodes to the node that directly links them. From there, a new matrix is calculated and the new node is substituted for the original nodes that are now joined. The advantage here is that there is not an assumption about the distances between nodes since it is directly calculated.

Steps for building a tree

1. Construct distance matrix
2. Cluster the two shortest distance OTUs into an internal nodes
3. Recalculate the distance matrix
4. Repeat the process until all OTUs are grouped in a single cluster

Maximum Parsimony

Maximum Parsiomony (MP) is probably the most widely and accepted method of tree construction to date The method is different from the previously discussed distance based methods since it uses a character based algorithm. The method works by searching through possible tree structures and assigning a cost to each tree. Parsimony is based on the assumption that the mostly likely tree is the one that requires the fewest number of changes to explain the data in the alignment. The premise that taxa or nodes sharing a common characteristic do so because the inherited that characteristics from a common ancestor.

Conflicts with this major assumption are explained under the term homoplasy. There are three main ways to reserve conflicts: reversal (revert back to original state), convergence (unrelated taxa evolved the same characteristic completely independently) and parallelism (different taxa may have similar mechanisms that cause a characteristic to develop in a certain manner). The tree with the lowest tree score or length, as defined by the number of changes summed along the branches, becomes the most parsimonious tree and is taken as the tree that best represents the evolutionary pattern. Maximum Parsimony is also different from the other methods in that it does not find branch lengths but rather the total overall length in terms of the number of changes. Often MP, finds two or more trees that it deems equal and does not provide a definite answer in how to distinguish which tree represents the actual evolutionary tree. In most cases a strict (majority rule) consensus is used to solve this dilemma Traditional parsimony uses recursion to search for the minimal number of changes within the trees. This done by starting at the leaf of a tree and working up towards the root, which is known as post-order traversal Another version of parsimony, weighted parsimony, adds a cost factor to the algorithm and weights certain scenarios accordingly. An artifact called long-branch attraction sometimes occurs in parsimony and should be handled. The branch length indicates the number of substitutions between two taxa or nodes. Parsimony assumes that all taxa evolve at the same rate and contribute that same amount of information. Long-branch is the phenomenon in which rapidly evolving taxa are placed together on a tree because they have many mutations. Anytime two long branches are present, they may be attracted to one another.

Steps for building a tree

- Start with multiple alignment
- Construct all possible topologies and base on evolutionary changes to score each of these topologies
- Choose a tree with the fewest evolutionary changes as the final tree

Maximum Likelihood

Proposed in 1981 by Felsenstein, Maximum likelihood (ML) is among the most computationally intensive approach but is also the most flexible ML optimizes the likelihood of observing the data given a tree topology and a model of nucleotide evolution Maximum Likelihood finds the tree that explains the observed data with the greatest probability under a specific model of evolution. ML is different from the other methods in that it is based on probability.

One of the big advantages to ML is the ability to make statistical comparisons between topologies and data sets. ML can return several equally likely trees – pro and con depending on the study Maximum Likelihood makes assumptions that the model used is accurate and if the model does not accurately reflect the underlying data set, the method is inconsistent. ML is designed to be robust, but breaching is assumptions can cause problems. A disadvantage of ML is the extensive computation as well as new evidence that suggest there can be multiple maximum likelihood points for a given phylogenetic tree.

PHYLIP

PHYLIP comes with an extensive set of documentation files. These include the main documentation file (this one), which you should read fairly completely. In addition there are files for groups of programs, including ones for the molecular sequence programs, the distance matrix programs, the gene frequency and continuous characters programs, the discrete characters programs, and the tree drawing programs.

Clique

Finds the largest clique of mutually compatible characters, and the phylogeny which they recommend, for discrete character data with two states. The largest clique (or all cliques within a given size range of the largest one) are found by a very fast branch and bound search method. The method does not allow for missing data. For such cases the T (Threshold) option of Pars or Mix may be a useful alternative. Compatibility methods are particular useful when some characters are of poor quality and the rest of good quality, but when it is not known in advance which ones are which.

Consense

Computes consensus trees by the majority-rule consensus tree method, which also allows one to easily find the strict consensus tree. Is not able to compute the Adams consensus tree. Trees are input in a tree file in standard nested-parenthesis notation, which is produced by many of the tree estimation programs in the package. This program can be used as the final step in doing bootstrap analyses for many of the methods in the package.

Contml

Estimates phylogenies from gene frequency data by maximum likelihood under a model in which all divergence is due to genetic drift in the absence of new mutations. Does not assume a molecular clock. An alternative method of analyzing this data is to compute Nei's genetic distance and use one of the distance matrix programs. This program can also do maximum likelihood analysis of continuous characters that evolve by a Brownian Motion model, but it assumes that the characters evolve at equal rates and in an uncorrelated fashion, so that it does not take into account the usual correlations of characters.

Contrast

Reads a tree from a tree file, and a data set with continuous characters data, and produces the independent contrasts for those characters, for use in any multivariate statistics package. Will also produce covariances, regressions and correlations between characters for those contrasts. Can also correct for within-species sampling variation when individual phenotypes are available within a population.

Dnacomp

Estimates phylogenies from nucleic acid sequence data using the compatibility criterion, which searches for the largest number of sites which could have all states (nucleotides)uniquely evolved on the same tree. Compatibility is particularly appropriate when sites vary greatly in their rates of evolution, but we do not know in advance which are the less reliable ones.

Dnadist

Computes four different distances between species from nucleic acid sequences. The distances can then be used in the distance matrix programs. The distances are the JukesCantor formula, one based on Kimura's 2- parameter method, the F84 model used in Dnaml, and the LogDet distance. The distances can also be corrected for gamma distributed and gamma-plus-invariant-sites-distributed rates of change in different sites. Rates of evolution can vary among sites in a prespecified way, and also according to a Hidden Markov model.

Dnainvar

For nucleic acid sequence data on four species, computes Lake's and Cavender's phylogenetic invariants, which test alternative tree topologies. The program also tabulates the frequencies of occurrence of the different nucleotide patterns. Lake's invariants are the method which he calls "evolutionary parsimony".

Dnaml

Estimates phylogenies from nucleotide sequences by maximum likelihood. The model employed allows for unequal expected frequencies of the four nucleotides, for unequal rates of transitions and transversions, and for different (prespecified) rates of change in different categories of sites, and also use of a Hidden Markov model of rates, with the program inferring which sites have which rates. This also allows gamma-distribution and gamma-plus-invariant sites distributions of rates across sites.

Dnamlk

Same as Dnaml but assumes a molecular clock. The use of the two programs together permits a likelihood ratio test of the molecular clock hypothesis to be made.

Dnamove

Interactive construction of phylogenies from nucleic acid sequences, with their evaluation by parsimony and compatibility and the display of reconstructed ancestral bases. This can be used to find parsimony or compatibility estimates by hand.

Dnapars

Estimates phylogenies by the parsimony method using nucleic acid sequences. Allows use the full IUB ambiguity codes, and estimates ancestral nucleotide states. Gaps treated as a fifth nucleotide state. It can also do transversion parsimony. Can cope with multifurcations, reconstruct ancestral states, use 0/1 character weights, and infer branch lengths.

Dnapenny

Finds all most parsimonious phylogenies for nucleic acid sequences by branch-and-bound search. This may not be practical (depending on the data) for more than 10-11 species or so.

Dollop

Estimates phylogenies by the Dollo or polymorphism parsimony criteria for discrete character data with two states (0 and 1). Also reconstructs ancestral states and allows weighting of characters. Dollo parsimony is particularly appropriate for restriction sites data; with ancestor states specified as unknown it may be appropriate for restriction fragments data.

Dolmove

Interactive construction of phylogenies from discrete character data with two states (0 and 1) using the Dollo or polymorphism parsimony criteria. Evaluates parsimony and compatibility criteria for those phylogenies and displays reconstructed states throughout the tree. This can be used to find parsimony or compatibility estimates by hand.

Dolpenny

Finds all most parsimonious phylogenies for discrete-character data with two states, for the Dollo or polymorphism parsimony criteria using the branch-and-bound method of exact search. May be impractical (depending on the data) for more than 10-11 species.

Drawgram

Plots rooted phylogenies, cladograms, circular trees and phenograms in a wide variety of user-controllable formats. The program is interactive. It has an interface in the Java language which gives it a closely similar menu on all three major operating systems. Final output can be to a file formatted for one of the drawing programs, for a ray-tracing or VRML browser, or one at can be sent to a laser printer (such as Postscript or PCL compatible printers), on graphics screens or terminals, on pen plotters or on dot matrix printers capable of graphics. Many of these formats are historic so we no longer have hardware to test them. If you find a problem please report it.

Drawtree

Similar to Drawgram but plots unrooted phylogenies. It also has a Java interface for previews.

Factor

Takes discrete multistate data with character state trees and produces the corresponding data set with two states (0 and 1). Written by Christopher Meacham. This program was formerly used to accomodate multistate characters in Mix, but this is less necessary now that Pars is available.

Fitch

Estimates phylogenies from distance matrix data under the "additive tree model" according to which the distances are expected to equal the sums of branch lengths between the species. Uses the Fitch-Margoliash criterion and some related least squares criteria, or the Minimum Evolution distance matrix method. Does not assume an evolutionary clock. This program will be useful with distances computed from molecular sequences, restriction sites or fragments distances, with DNA hybridization measurements, and with genetic distances computed from gene frequencies.

Gendist

Computes one of three different genetic distance formulas from gene frequency data. The formulas are Nei's genetic distance, the Cavalli-Sforza chord measure, and the genetic distance of Reynolds et. al. The former is appropriate for data in which new mutations occur in an infinite isoalleles neutral mutation model, the latter two for a model without mutation and with pure genetic drift. The distances are written to a file in a format appropriate for input to the distance matrix programs.

Kitsch

Estimates phylogenies from distance matrix data under the "ultrametric" model which is the same as the additive tree model except that an evolutionary clock is assumed. The Fitch-Margoliash criterion and other least squares criteria, or the Minimum Evolution criterion are possible. This program will be useful with distances computed from molecular sequences, restriction sites or fragments distances, with distances from DNA hybridization measurements, and with genetic distances computed from gene frequencies.

Mix

Estimates phylogenies by some parsimony methods for discrete character data with two states (0 and 1). Allows use of the Wagner parsimony method, the Camin-Sokal parsimony method, or arbitrary mixtures of these. Also reconstructs ancestral states and allows weighting of characters (does not infer branch lengths).

Move

Interactive construction of phylogenies from discrete character data with two states (0 and 1). Evaluates parsimony and compatibility criteria for those phylogenies and displays reconstructed states throughout the tree. This can be used to find parsimony or compatibility estimates by hand.

Neighbor

An implementation by Mary Kuhner and John Yamato of Saitou and Nei's "Neighbor Joining Method," and of the UPGMA (Average Linkage clustering) method. Neighbor Joining is a distance matrix method producing an unrooted tree without the assumption of a clock. UPGMA does assume a clock. The branch lengths are not optimized by the least squares criterion but the methods are very fast and thus can handle much larger data sets.

Pars

Multistate discrete-characters parsimony method. Up to 8 states (as well as "?") are allowed. Cannot do Camin-Sokal or Dollo Parsimony. Can cope with multifurcations, reconstruct ancestral states, use character weights, and infer branch lengths.

Penny

Finds all most parsimonious phylogenies for discrete-character data with two states, for the Wagner, Camin-Sokal, and mixed parsimony criteria using the branch-and-bound method of exact search. May be impractical (depending on the data) for more than 10-11 species.

Proml

Estimates phylogenies from protein amino acid sequences by maximum likelihood. The PAM, JTT, or PMB models can be employed, and also use of a Hidden Markov model of rates, with the program inferring which sites have which rates. This also allows gamma distribution and gamma-plus-invariant sites distributions of rates across sites. It also allows different rates of change at known sites.

Promlk

Same as Proml but assumes a molecular clock. The use of the two programs together permits a likelihood ratio test of the molecular clock hypothesis to be made.

Protdist

Computes a distance measure for protein sequences, using maximum likelihood estimates based on the Dayhoff PAM matrix, the JTT matrix model, the PBM model, Kimura's 1983 approximation to these, or a model based on the genetic code plus a constraint on changing to a different category of amino acid. The distances can also be corrected for gamma-distributed and gamma-plus-invariant-sites-distributed rates of change in different sites. Rates of evolution can vary among sites in a prespecified way, and also according to a Hidden Markov model. The program can also make a table of percentage similarity among sequences. The distances can be used in the distance matrix programs.

Protpars

Estimates phylogenies from protein sequences (input using the standard one-letter code for amino acids) using the parsimony method, in a variant which counts only those nucleotide changes that change the amino acid, on the assumption that silent changes are more easily accomplished. percentage similarity among sequences.

Restdist

Distances calculated from restriction sites data or restriction fragments data. The restriction sites option is the one to use to also make distances for RAPDs or AFLPs.

Restml

Estimation of phylogenies by maximum likelihood using restriction sites data (not restriction fragments but presence/absence of individual sites). It employs the JukesCantor symmetrical model of nucleotide change, which does not allow for differences of rate between transitions and transversions. This program is very slow.

Retree

Reads in a tree (with branch lengths if necessary) and allows you to reroot the tree, to flip branches, to change species names and branch lengths, and then write the result out. Can be used to convert between rooted and unrooted trees, and to write the tree into a preliminary version of a new XML tree file format which is under development and which is described in the Retree documentation web page.

Seqboot

Reads in a data set, and produces multiple data sets from it by bootstrap resampling. Since most programs in the current version of the package allow processing of multiple data sets, this can be used together with the consensus tree program Consense to do bootstrap (or delete-half-jackknife) analyses with most of the methods in this package. This program also allows the Archie/Faith technique of permutation of species within characters. It can also rewrite a data set to convert it from between the PHYLIP Interleaved and Sequential forms, and into a preliminary version of a new XML sequence alignment format which is under development and which is described in the Seqboot documentation web page.

Threshml

Reads a tree from a tree file, and a data set with discrete 0/1 characters. Using the threshold model of quantitative genetics, the program runs a Markov Chain Monte Carlo (MCMC) sampler to sample the underlying continuous characters (the liabilities) that cause the discrete characters. The covariances of the liabilities are estimated, as well as the transformation from the liabilities to underlying independently evolving characters.

Treedist

Computes the Branch Score distance between trees, which allows for differences in tree topology and which also makes use of branch lengths. Also computes another distance by Robinson and Foulds that uses branch lengths, and the Symmetric Difference distance between trees, which allows for differences in tree topology but does not use branch lengths.

Possible questions

1. What is sequence alignment? Explain in detail.
2. Explain the steps in constructing a phylogenetic tree.
3. Differentiate the tblastn and tblastx
4. Illustrate about the PHYLIP package
5. Write about the following
    a. BLAST
    b. FASTA

c. ClustalW

d. RasMol

6. Write in detail about different types of BLAST and their significance

7. What is phylogentics tree? Explain about their types and terminologies.

8. Differentiate local and global alignment.

9. Mention the significance and methodology involved in multiple alignment.

10. Write short notes on tree building methods.

11. Define multiple sequence alignment? Write short notes on ClustalW.

KAPRAGAM ACADEM YOF HIGHER EDUCATION

DEPARTMENT OF BIOCHEMISTRY

I- M.Sc Biochemistry

18BCP204-Bioinformatics

S. No	Unit	Questions	Option I	Option II	Option III	Option IV	Answer
1	2	Expansion of BLAST is	Basic arrangement site tool	Basic legal alignment search tool	Basic local alignment search tool	Basic level arrangement of search tool	**Basic local alignment search tool**
2	2	Protein sequenced with protein sequence	Blastp	Blastn	Blastx	Tblastn	**Blastp**
3	2	Translated nucleic acid sequenced with a protein	Blastn	Blastx	TBlastn	TBlastx	**TBlastn**
4	2	Searches for all matching words of length K	FASTA	Blast	FASTAB	Blast2	**FASTA**
5	2	PSI blast is	Protein specific implementation	Protein specific iterative	Paired specific iterative	Paired specific implementation	**Protein specific iterative**
6	2	Algorithm is one of the most widely used multiple sequence alignment programs available free for all programs	GCG	FASTA	Phylip	Clustal W	**Clustal W**
8	2	A sequence starting with start codon and a stop codon on either side of a strech of nucleotide sequence is called	open reading frame	exon	patterns	motif	**open reading frame**
9	2	Protein allows one to input protein sequences and compares the sequences against	Protein	nucleotide	a & b	EST	**Protein**
11	2	Blast was found out by	Thompson et al	Alstchul et al	Both a and b	Crick	**Alstchul et al**
12	2	Due to computational time & costs cannot be used for MSA of a set of sequence	Pairwise alignment	Dynamic programming	Phylogenetic tree	dendograms	**Phylogenetic tree**
13	2	The most popular methods conducting multiple sequence alignment are summarized in the	Ven diagram	Tree diagram	Dot plot	Root	**Root**
14	2	Like pairwise alignment, MSA can also be done	Globally	Locally	Both a & b	phylogenetic	**Both a & b**
15	2	Expand MSA	Multiple sequence alignment	Multiple alignment	Both a & b	multiple sequence	**Multiple sequence alignment**
16	2	Gene recognition and analysis internet link is the expansion of	GRAIL	AAT	FGENEH	MZEF	**MZEF**
17	2	The Needle man Wunsch algorithm relies upon searching:	Molecular nature	local similarity	conserved regions	global similarity	**global similarity**
18	2	a program that is designed to create optimum multiple alignment of sequences	PIR	Clustal X	FASTA	Phylip	**FASTA**
19	2	The phenetic method is applied for comparing organisms at genetic level and is suitable for smaller evolutionary divergence. This method is also called	Parsimony	Maximum likelihood	cladistic	distance	**distance**
20	2	function by breaking down a large problem into a series of smaller problems.	Dynamic programming	MSA	Heuristic method	Pair wise alignment	**Pair wise alignment**
21	2	was optimized local alignment and substitution matrix for its sensitivity.	Fasta	Blast	PAM	BLOSUM	**BLOSUM**
22	2	The triplet of contiguous bases on tRNA that binds to the codon sequences of nucleotides of mRNA is	Codon	termination codon	initiation codon	Anticodon	**Anticodon**
23	2	Program produces graphical version of tree files.	PAUP	PAUPDISPLAY	PAUPSEARCH	PARSIMONY	**PARSIMONY**
24	2	A conserved element of a protein sequence that usually correlates with a particular function is	Database	domain	motif	gaps	**gaps**
25	2	protein sequenced with protein sequence..	Blast p	Blast n	Blast x	T Blast n	**T Blast n**
26	2	Translated nucleic acid sequenced with a protein	Blast n	Blast x	T blast n	T blast x	**T blast x**
27	2	Pairwise algorithm alignment performs well at	<50%	<70%	>50%	40%	**40%**
28	2	Which one of the following is used for aligning and searching nucleotide sequences	BLAST	FASTA	GCG	both a and b	**both a and b**
29	2	Which matrix is used to compare sequences with by distant relationship.	PAM-1	PAM-250	PAN-350	PAM-1000	**PAM-250**
30	2	The most popular tools for similarity search are	BLAST	FASTA	both a & b	clustal-w	**FASTA**
31	2	Which tool used in hidden markov models	GRAIL	Benie	polyphred	BLAST	**GRAIL**
32	2	The two strands in DNA are held together between individual bases by	ionic interactions	hydrogen bonds	salt bridges	hydrophobic interactions	**hydrogen bonds**
33	2	program can be used for finding consensus trees.	PAUP	PIR	PAP	NBRF	**PAUP**
34	2	a powerful tool for scanning databases to find sequences that are similar to query sequence. Blast	FASTA	ClustalW	None of the above	PIR	**FASTA**
35	2	the position of comparing two or more sequences	Alignment	Scoring	MSA	distance	**Alignment**
36	2	a method for data analysis.	algorithm	procedure	preprocessing	processing	**algorithm**
37	2	refers to the position in a sequence alignment	distance score	distance matrix	sequence position	matrices	**distance score**
38	2	provides the graphical methods for comparing two sequences	dot matrix analysis	dynamic programming	dotplot	matrix	**dot matrix analysis**
39	2	an alignment that includes all of the sequences in the alignment	global alignment	local alignment	MSA	CSA	**global alignment**
40	2	an alignment that aligns region of sequences by the highest density of matches	global alignment	local alignment	MSA	CSA	**local alignment**
41	2	a graphical representation of observed changes in MSA.	maximum parsimony tree	parsimony	clustal w	dot matrix	**maximum parsimony tree**
42	2	evoloves to the problem of finding the optimal alignment between two sequences.	dynamic programming	dot matrix	MSA	Parsimony	**dynamic programming**
43	2	a blank position in the alignment between the two sequences.	gap	deletion	insertion	subsitution	**gap**
44	2	a term used to represent either insertion or deletion in sequence alignment.	gap	indel	index	field	**indel**
45	2	the ratio of the likelihoods of two event or out comes.	odd score	even score	alignment score	gaps	**odd score**
46	2	Protein allows one to input protein sequences and compares the sequences against	Other protein sequences	Nucleotide blast	Mega blast	Protein blast	**Protein blast**
47	2	The. is a general local alignment method also based on dynamic programming	N-W algorithm	S-W Algorithm	both a and b	Clustal W	**S-W Algorithm**
48	2	Protein sequenced with protein sequence.	Blast p	Blast n	Blast x	T blast n	**Blast p**
49	2	PSI blast is	Protein specific implementation	Protein specific iterative	Paired specific iterative	Paired specific implementation	**Protein specific iterative**
50	2	algorithm is one of the most widely used multiple sequence alignment programs available freely for all platforms	GCG	FASTA	Phylip	Clustal W	**Clustal W**
51	2	Due to computational and time costs cannot be used for MSA of a set of sequence	Pairwise alignment	Dynamic programming	Phylogenitic tree	Phylip	**Pairwise alignment**
52	2	Like pairwise alignment, MSA can also be done	Globally	Locally	Both a and b	aligment score	**Both a and b**
53	2	Algorithm is used for local alignment.	Smith waterman	Needleman wunch	Lipman	Heuristic	**Smith waterman**
54	2	Searches for all matching words of length K.	Fasta	Blast	Fasta	Blast2	**Fasta**
55	2	are used to find the best-matching piecewise (local) or global alignments of two query sequence	phylogenetic trees	pairwise alignment	MSA	global similarity	**pairwise alignment**
56	2	The dot plots of very closely related sequences will appear as a single line along the matrix's main	arrow	diagonal	points	vertical	**diagonal**
57	2	an extension of pairwise alignment to incorporate more than two sequences at a time. phylogenetic trees	pairwise alignment	MSA	global	diagonal	**pairwise alignment**
58	2	Alignments are also used to aid in establishing evolutionary relationships by constructing phylogenetic trees	pairwise alignment	MSA	similarity	diagonal	**MSA**
59	2	The is a fragment-based method for constructing structural alignments based on contact similarity patterns between successive hexapeptides in the query sequences	DALI	msa	SSAP	dot polot	**DALI**
60	2	a dynamic programming-based method of structural alignment that uses atom-to-atom vectors in structure space as comparison points.	pair wise alignment	msa	SSAP	SSP	**SSAP**
61	2	The scientific discipline concerned with naming organisms is called	taxonomy	cladistics	binomial nomenclature	systematics	**taxonomy**
62	2	A phylogenetic tree that is "rooted" is one	that extends back to the origin of life on Earth	at whose base is located the common ancestor of all taxa depicted on that tree	that illustrates the rampant gene swapping that occurred early in life's history	with very few branch points	**at whose base is located the common ancestor of all taxa depicted on that tree**

63	2	The best classification system is that which most closely	unites organisms that possess similar morphologies.	conforms to traditional, Linnaean taxonomic practices	reflects evolutionary history	reflects the basic separation of prokaryotes from eukaryotes	**reflects evolutionary history**
64	2	A phylogenetic tree constructed using sequence differences in mitochondrial DNA would be most valid for discerning the evolutionary relatedness of	archaeans and bacteria	fungi and animals	Hawaiian silverswords	mosses and ferns	**Hawaiian silverswords**
65	2	The reason that paralogous genes can diverge from each other within the same gene pool, whereas orthologous genes diverge only after gene pools are isolated from each other, is that	having multiple copies of genes is essential for the occurrence of sympatric speciation in the wild	paralogous genes can occur only in diploid species; thus, they are absent from most prokaryotes	polyploidy is a necessary precondition for the occurrence of sympatric speciation in the wild	having an extra copy of a gene permits modifications to the copy without loss of the original gene product	**having an extra copy of a gene permits modifications to the copy without loss of the original gene product**
66	2	The Neighbour-Joining method is	Closely related method	Distance clustering method	Single sequence method	Maximum likelihood method	**Distance clustering method**
67	2	Phylogenetic system of classification is based on	Morphological features	Chemical relationship	Evolutionary relationship	Floral characters	**Evolutionary relationship**
68	2	Similarity between two short fragments results from the	Evolutionary convergence	Evolutionary Divergence	Common ancestor	All of the above	**Evolutionary Divergence**
69	2	PILEUP is the Multiple sequence alignment program which is a part of the	Genetics Computer Group	Computer Genetics Group	Group of Computer Genetics	None of the above	**Genetics Computer Group**
70	2	The two main features of any phylogenetic tree are the	clades and the nodes	topology and the branch lengths	clades and the root	alignment and the bootstrap	**topology and the branch lengths**
71	2	Reconstruction of Phylogenetic tree will be carried out from	Protein sequence	Nucleic acid sequence	Both a and b	None of the above	**Both a and b**
72	2	Phylogeniesis of species can be reconstructed only by comparing	Orthologous genes	Paralogous genes	Both a and b	None of the above	**Orthologous genes**
73	2	____ is a broad term for the diagrammatic representation of a phylogenetic tree	chronogram	Phylogram	cladogram	Dendrogram	**Dendrogram**
74	2	A node in a phylogeny represents a	Common ancestor	Different ancestor	Both a and b	None of the above	**Common ancestor**
75	2	The groups showing similarities due to single ancestors are called________________ groups	monophyletic	diphyletic	triphyletic	ployphyletic	**monophyletic**
76	2	The study of kinds and diversity of organisms and the evolutionary relationships aMong them is called	systematics	genetics	kinetics	mechanics	**systematics**
77	2	The group with unknown evolutionary relationship is	Monophyletic	Polyphyletic	Paraphyletic	None	**Paraphyletic**
78	2	Which of the followings systematics have traditional approach?	Evolutionary	Phyletic	Numerical	All	**Evolutionary**
79	2	Why do scientists apply the concept of maximum parsimony?	to decipher accurate phylogenies	to eliminate analogous traits	to identify mutations in DNA codes	to locate homoplasies	**to decipher accurate phylogenies**
80	2	On a phylogenetic tree, which term refers to lineages that diverged from the same place?	sister taxa	basal taxa	rooted taxa	dichotomous taxa	**sister taxa**
81	2	What do scientists use to apply cladistics?	homologous traits	homoplasies	analogous traits	monophyletic groups	**homologous traits**
82	2	What does the trunk of the classic phylogenetic tree represent?	single common ancestor	pool of ancestral organisms	new species	old species	**single common ancestor**
83	2	To apply parsimony to constructing a phylogenetic tree,	choose the tree that assumes all evolutionary changes are equally probable	choose the tree in which the branch points are based on as many shared derived characters as possible	choose the tree that represents the fewest evolutionary changes, either in DNA sequences or morphology	choose the tree with the fewest branch points	**choose the tree that represents the fewest evolutionary changes, either in DNA sequences or morphology**
84	2	Theoretically, molecular clocks are to molecular phylogenies as radiometric dating is to phylogenies that are based on the	fossil record	geographic distribution of extant species	morphological similarities among extant species	amino acid sequences of homologous polypeptides	**fossil record**
85	2	Concerning growth in genome size over evolutionary time, which of these does *not* belong with the others?	orthologous genes	gene duplications	paralogous genes	gene families	**orthologous genes**
86	2	Cladograms (a type of phylogenetic tree) constructed from evidence from molecular systematics are based on similarities in	morphology	biochemical pathways	habitat and lifestyle choices	mutations to homologous genes	**mutations to homologous genes**
87	2	Phylogenetic hypotheses (such as those represented by phylogenetic trees) are strongest when	they are based on amino acid sequences from homologous proteins, as long as the genes that code for such proteins contain no introns	each clade is defined by a single derived character	they are supported by more than one kind of evidence, such as when fossil evidence corroborates molecular evidence	they are based on a single DNA sequence that seems to be a shared derived sequence	**they are supported by more than one kind of evidence, such as when fossil evidence corroborates molecular evidence**
88	2	he term that is most appropriately associated with *clade* is	paraphyletic	polyphyletic	monophyletic	diphyletic	**monophyletic**
89	2	A taxon, all of whose members have the same common ancestor, is	paraphyletic	polyphyletic	monophyletic	diphyletic	monophyletic
90	2	Shared derived characters are most likely to be found in taxa that are	paraphyletic	polyphyletic	monophyletic	diphyletic	monophyletic
91	2	The importance of computers and of computer software to modern cladistics is most closely linked to advances in	light microscopy	fossil discovery techniques	Linnaean classification.	molecular genetics	**molecular genetics**
92	2	The family that consists of related genes within an organism is called	orthologs	paralogs	zoologs	xenologs	**paralogs**
93	2	The family that consists of related genes in another organism is called	orthologs	paralogs	zoologs	xenologs	**orthologs**
94	2	Which branching diagram is assumed to be an estimate of a phylogeny when branching lengths are proportional to the amount of inferred evolutionary change?	Phylogram	Cladogram	A guide tree	Cardiogram	**Phylogram**
95	2	Gene duplication results in	orthologs	paralogs	zoologs	xenologs	**paralogs**
96	2	Two principal ways to construct guide tree in progressive alignment is	UPGMA and Neighbor joining method	Maximum Parsimony	Maximum Likelihood	all the above	**UPGMA and Neighbor joining method**
97	2	Which of these methods is a distance-based method in tree construction?	Unweighted pair group method with arithmetic mean	Jukes-Cantor	Minimum evolution	Maximum parsimony	**Unweighted pair group method with arithmetic mean**
98	2	Which one of the following is not a character-based method in tree construction?	Maximum parsimony	Minimum likelihood	Minimum evolution method	Neighbor joining	**Neighbor joining**
99	2	A tree representation of a family showing the relationships between members and pattern of inheritance of a given trait is known as	pedigree	physical Map	genetic map	population studies	**pedigree**
100	2	The study of evolutionary relationships is	Phylogenics	Molecular Evolution	Cladogenesis	Cladistics	**Phylogenics**
101	2	A bifurcating branch point in the phylogenetic tree is known as	node	clade	branch	taxon	**node**
102	2	Expand UPGMA	Unweighted Pair Group Method with Arithmetic Mean	Unweighted Pair Group Method with All Mean	Upregulated Gene Method with Arithmetic Mean	Unregulated Genome Method with All Mean	**Unweighted Pair Group Method with Arithmetic Mean**
103	2	One of the most common errors in making and analyzing phylogenetic tree is	using a bad multiple sequence alignment as input	trying to infer the evolutionary relationship of genes or proteins in the tree	trying to infer the age at which genes or proteins diverged from each other	assuming that clades are monophyletic	**using a bad multiple sequence alignment as input**
104	2	Which one of the following tool can be used to generate neighbor joining trees with or without bootstrap values?	ClustalX	BLAST	Swiss-PDB viewer	ChemSketch	**ClustalX**
105	2	Molecular phylogeny can be performed with ______ sequences	only DNA	only RNA	only protein	all the above	**all the above**
106	2	A phylogenetic tree that explicitly represents number of character changes through its branch lengths is	dendogram	cladogram	phylogram	chronogram	**phylogram**
107	2	Which of the following is the character based method?	UPGMA	Maximum Parsimony and Maximum Likelihood	Maximum Likelihood and Neighbor-Joining	Neighbor-Joining	**Maximum Parsimony and Maximum Likelihood**
108	2	Which if the following is not an algorithm for generating phylogenetic trees from molecular data?	Neighbor-joining	Parsimony	Maximum likelihood	Jukes & Cantor	**Jukes & Cantor**

109	2	__________ is a way to judge the reliability of the branches in a tree	bootstrapping	clade	branch tree	chronogram	**bootstrapping**
110	2	OTU stands for	operational taxonomic unit	Outgroups	Only translation units	Outlying units	**operational taxonomic unit**
111	2	A taxon ________	is a species	is a formal grouping at any given level	is a clade	of one type of organism at one level is comparable to another type of organism at the same level	**is a formal grouping at any given level**
112	2	What does a branch point in a phylogenetic tree represent?	A branch point represents a point at which two evolutionary lineages split from a common ancestor	A branch point represents a gene duplication event	A branch point represents a split between two phyla	A branch point represents a place where one species branches off from another	**A branch point represents a point at which two evolutionary lineages split from a common ancestor**
113	2	Which of the following methods to establish phylogenetic relationships among organisms has been developed most recently?	comparing physiology	comparing behavioral patterns	comparing the amino acid sequences of proteins and nucleotide sequences of nucleic acids	comparing morphology	**comparing the amino acid sequences of proteins and nucleotide sequences of nucleic acids**
114	2	Which statement below is true about an outgroup?	The outgroup should be from a lineage known to have diverged before the lineage that includes the ingroup	The outgroup would be found at one of the highest branches of a phylogenetic tree	Outgroup comparison is based on the assumption that homologies present in both the outgroup and ingroup must be derived characters	The outgroup and ingroup display a mixture of shared and derived characters	**The outgroup should be from a lineage known to have diverged before the lineage that includes the ingroup**
115	2	Unlike a regular phylogenetic tree, phylogenetic trees with branch lengths proportional to time can be used to ________	tie polyphyletic clades to a common ancestor	reflect the rate of evolutionary change	hypothesize the relative relatedness between different taxa	represent the chronological time that has passed since two groups diverged from a common ancestor	**represent the chronological time that has passed since two groups diverged from a common ancestor**
116	2	Which statement below is true of parsimonious trees?	Trees can be constructed that are the most parsimonious or the most likely, but not both at the same time	The most parsimonious tree requires the fewest evolutionary events to have occurred in the form of shared derived characters	Given the rules of how morphological traits change over time, a tree can be found that reflects the most likely sequence of evolutionary events	The most parsimonious tree requires the fewest evolutionary events to have occurred in the form of shared ancestral characters	**The most parsimonious tree requires the fewest evolutionary events to have occurred in the form of shared derived characters**
117	2	Paralogous genes	result from gene duplication	are passed from generation to generation in a straight line	cannot diverge in the same gene pool	can only diverge after speciation has taken place	**result from gene duplication**
118	2	What is the evolutionary significance of paralogous genes?	They give the absolute time that two species diverged	They give the absolute time that the gene duplication occurred	None of the listed responses is correct	They increase the size of the genome and provide more opportunity for the evolution of novel characteristics	**They increase the size of the genome and provide more opportunity for the evolution of novel characteristics**

UNIT-III

SYLLABUS

Protein prediction strategies and programs: Protein Secondary Structure Prediction, three dimensional structure prediction-Comparative modeling, threading, Concepts of molecular modeling, Model refinement, evaluation of the model, protein folding and visualization of molecules-Visualization tools- RasMol, Deep Veiw.

Proteins

- Protein: from the Greek word PROTEUO which means "to be first (in rank or influence)"
- Why are proteins important to us:

  Proteins make up about 15% of the mass of the average person

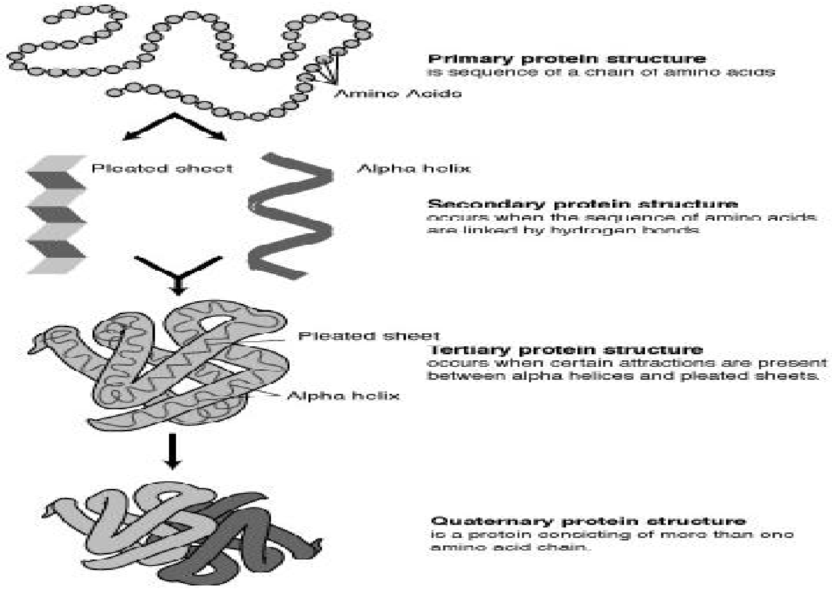
  Enzyme – acts as a biological catalyst

  Storage and transport – Haemoglobin

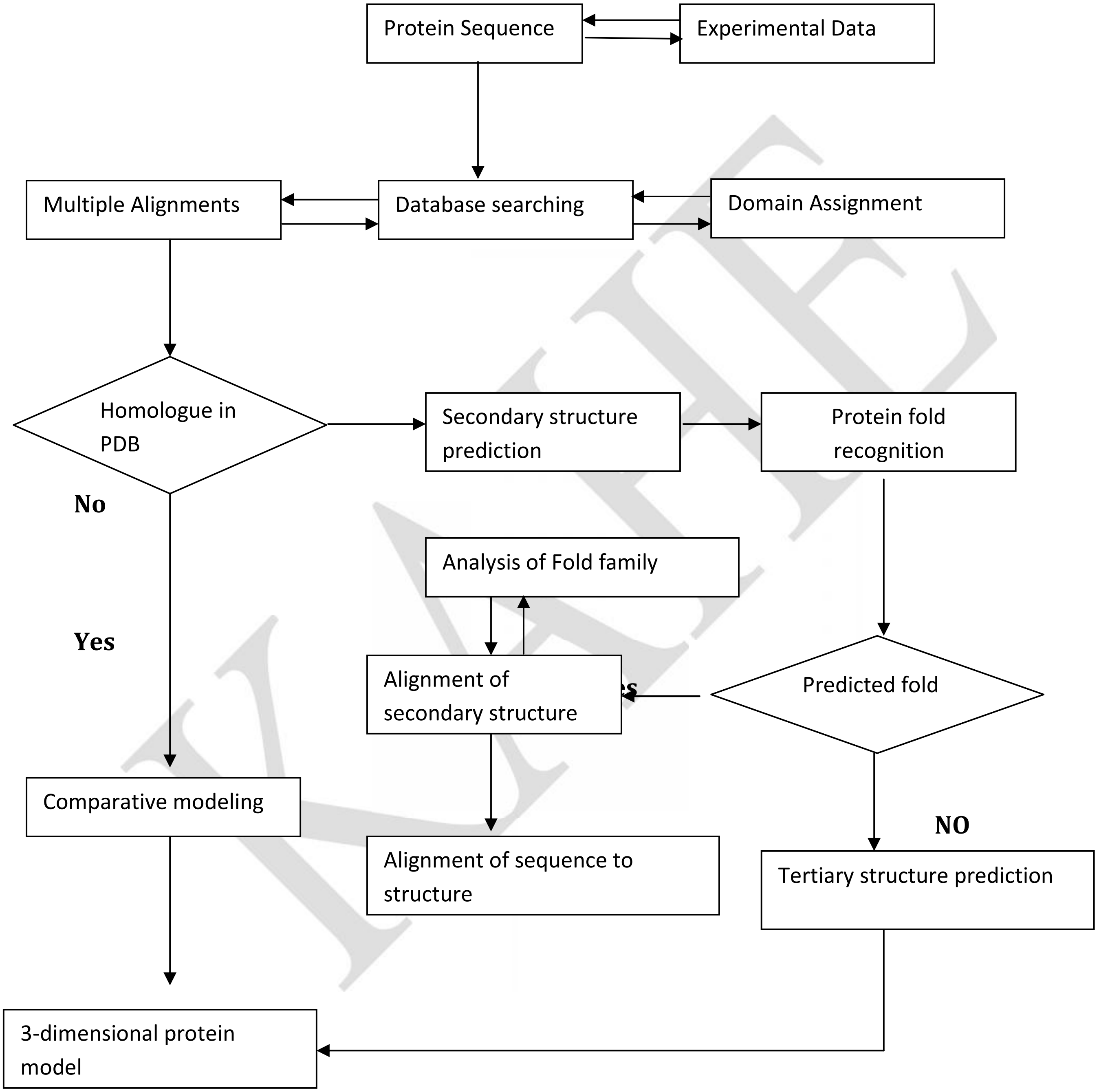
  Antibodies

  Hormones – Insulin

Four levels of protein structure


Primary protein structure
is sequence of a chain of amino acids
Amino Acids
Pleated sheet
Alpha helix
Secondary protein structure
occurs when the sequence of amino acids
are linked by hydrogen bonds
Pleated sheet
Tertiary protein structure
occurs when certain attractions are present
between alpha helices and pleated sheets.
Alpha helix
Quaternary protein structure
is a protein consisting of more than one
amino acid chain.

Flow chart showing steps involved in prediction protein structure



Protein Secondary Structure

- Secondary structure is the term protein chemist give to the arrangement of the peptide backbone in space. It is produced by hydrogen bondings between aminoacids
- The assignment of the SS categories to the experimentally determined three-dimensional (3D) structure of proteins is a non-trivial process and is typically performed by widely used DSSP program
- PROTEIN SECONDARY STRUCTURE consists of : protein sequence and its hydrogen bonding patterns called SS categories
- Databases for protein sequences are expanding rapidly due to the genome sequencing projects and the gap between the number of determined protein structures (PSS – protein secondary structures) and the number of known protein sequences in public
- Protein data banks (PDB) is growing bigger.
- PSSP (Protein Secondary Structure Prediction) research is trying to breach this gap.

Early methods for Secondary Structure Prediction

Chou and Fasman

- Start by computing amino acids propensities to belong to a given type of secondary structure:

$$\frac{P(i/Helix)}{P(i)} \qquad \frac{P(i/Beta)}{P(i)} \qquad \frac{P(i/Turn)}{P(i)}$$

Propensities > 1 mean that the residue type I is likely to be found in the
Corresponding secondary structure type.

Predicting Alpha helices

õ find nucleation site: 4 out of 6 contiguous residues with P(a)>1

õ extension: extend helix in both directions until a set of 4 contiguous residues has an average P(a) < 1 (breaker)

- if average P(a) over whole region is >1, it is predicted to be helical

- a Spiral shape.

Predicting Beta strands

õ find nucleation site: 3 out of 5 contiguous residues with P(b)>1

õ extension: extend strand in both directions until a set of 4 contiguous

õ residues has an average P(b) < 1 (breaker)

õ if average P(b) over whole region is >1, it is predicted to be a strand

Predicting turns

o for each tetrapeptide starting at residue I, compute:

o Pturn (average propensity over all 4 residues)

o F = f(i)*f(i+1)*f(i+2)*f(i+3)

o if Pturn > Pa and Pturn > Pb and Pturn > 1 and F>0.000075

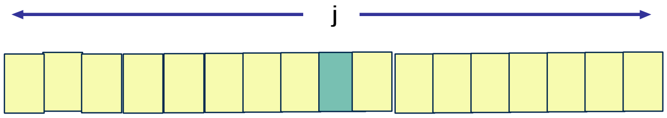
o tetrapeptide is considered a turn.

Random Coils

- prediction of secondary structure of protein though difficult but is important for mainly two reasons.
- Functional properties of proteins depend upon their 3D structure.
- Due to relationships between the ways amino acid sequence arrangement and their corresponding structure.
- No definite rule or an algorithm which characterizes this relationships.

The GOR method

Position-dependent propensities for helix, sheet or turn is calculated for each amino acid. For each position j in the sequence, eight residues on either side are considered.



A helix propensity table contains information about propensity for residues at 17 positions when the conformation of residue j is helical. The helix propensity tables have 20 x 17 entries.
Build similar tables for strands and turns.

GOR simplification:
The predicted state of AAj is calculated as the sum of the position-dependent propensities of all residues around AAj.

GOR can be used at : http://abs.cit.nih.gov/gor/ (current version is GOR IV)

Accuracy

- Both Chou and Fasman and GOR have been assessed and their accuracy is estimated to be Q3=60-65%.
- (initially, higher scores were reported, but the experiments set to measure Q3 were flawed, as the test cases included proteins used to derive the propensities!)

Protein tertiary structure prediction Methods

The biological role of a protein is determined by its function, which is in turn largely determined by its structure. Thus there is enormous benefit in knowing the three dimensional structures of all the proteins. Although more and more structures are determined experimentally at an accelerated rate, it is simply not possible to determine all the protein structures from experiments. As more and more protein sequences are determined, there is pressing need for predicting protein structures computationally. Decades of intense research in this area brought about huge progress in our ability to predict protein structures from sequences only.

The protein structure prediction methods can be broadly divided into three categories:

- Homology modeling,
- Threading or fold recognition, and

- Ab Initio

Essentially, the classification reflects the degree to which different methods utilize the information content available from the known structure database.

Comparative homology modeling

So far protein prediction methods based on homology have been the most successful. Homology modeling is based on the notion that new proteins evolve gradually from existing ones by amino acid substitution, addition, and/or deletion and that the 3D structures and functions are often strongly conserved during this process. Many proteins thus share similar functions and structures and there are usually strong sequence similarities among the structurally similar proteins. Strong sequence similarity often indicates strong structure similarity, although the opposite is not necessarily true. Homology modeling tries to identify structures similar to the target protein through sequence comparison. The quality of homology modeling depends on whether these exists one or more protein structures in the protein structure databases that show significant sequence similarity to the target sequence.

There are usually four steps in homology based protein structure prediction methods:

- Identify one or more suitable structural templates from the known protein structure databases;
- Align the target sequence to the structural template;
- Build the backbone from the alignment, including the loop region and any region that is significantly different from the template; and
- Place the side-chains.

The first two steps, identification of structural templates and alignment of the target sequence onto the parent structures, are usually related. Sequence comparison methods determine sequence similarity by aligning the sequences optimally. The aligned residuals of the structure templates are used to construct the structural model in the second step. The quality of the sequence comparison thus not only determines whether a suitable structural template can be found but also the quality of the alignment between the target sequence and the parent structure, which in turn determines the accuracy of the structural model. Of critical importance is the ability for the sequence comparison to detect remote homologues and to correctly align the target sequence to and parent structure. In the following I discuss the various sequence comparison methods in relation to homology modeling and their range of applicability, accuracy and shortcomings. For comparative modeling, local sequence comparison methods are usually used since the sequence similarity is most likely over segments of the two sequences. The local sequence comparison can either be pair wise or profile based. Pair wise comparisons, such as the widely used BLAST in the early days, can detect sequence similarities better than 30%. A number of tools have also been developed to detect weak homology relationships. Methods like profile and HMM use a statistical profile of a protein family. To further increase the chance of detecting remote homologues, PSI-BLAST and SAM-T98 build the profile or HMM by searching the database iteratively until no new hits are found. Methods such as PSI-BLAST encode the information about a whole protein family for the target sequence in a model to increase the chance of detecting remote homologies. To further increase the detection sensitivity, the sequences in the structure database can also be encoded in profiles. This forms the basis of the profile-profile based comparison methods. With low sequence identities ((<20%), profile-profile methods clearly outperform the other two kinds of methods: profile-profile methods identified more than 90% of homologous pairs, determined from structure-structure similarity comparison, with sequence identity better than 10% and an impressive 38% even for cases with sequence identities between 5% and 9%.

The structure models are constructed from the residuals of the structure template that are aligned to the target sequence in the sequence comparison. The quality of this alignment thus is critical for the accuracy achievable. The aligned residues from sequence comparison are generally different from that from structure-structure comparison though, especially when the sequence identity is low. To assess the ability of the sequence comparison methods to align the sequences correctly, it is instructive to compare the sequence-sequence alignment to the structure-structure alignment of the same pair of proteins. To determine how well the different similarity search methods can detect remote homologies and assess their ability in correctly aligning the sequences, compared various sequence alignment methods to the CE structure alignment of the SCOP protein structures. For sequence identities less than 30%, profile-based comparison methods, such as PSI-BLAST and profile-profile comparison, are all obviously better than the pair wise BLAST method. For example, at 10-15% sequence identity, BLAST aligns only 20% correctly while PSIBLAST and profile-profile comparison can correctly align 40% and 48% respectively. This also indicates that there is still large room for improvement in correctly aligning the target sequence to the target structure. One indication of the accuracy of comparative modeling is the sequence identity between the target and the template. It is believed that if two protein sequences have 50% or higher sequence identity, then the RMSD of the alignable potion between the two structures will normally be less than 1. In the so-called “twilight zone”, with sequence identity between 20%~30%, 95% of the sequences with this level of identity have different structures though. When a structure template can indeed be found within the known protein structure databases in such cases, the backbone RMSD can be expected to be no better than 2. Structurally similar proteins can have low sequence identities in the 8~10% range and can still be identified with sensitive profile-profile based comparison, but the RMSD can be as large as 3~6. The error largely comes from the misalignment from sequence comparison. At such low sequence identity, comparison method that can detect the remote homology as well as align the sequences close to the optimal from structure alignment will be desirable.

Threading or fold recognition:

For evolutionally remotely related proteins, even if the sequence similarity is difficult to detect with sequence comparison methods, there could still be identifiable structural similarity. Structure alignments have been shown to be able to identify homologous protein pairs with sequence similarities less than 10%. When sequence comparison based methods are no longer sensitive enough to recognize the correct fold for the target sequence, fold recognition or threading can still be used to assign the correct fold to the target sequence. Threading or fold recognition is the method by which a library of unique or representative structures is searched for structure analogs to the target sequence, and is based on the theory that there may be only a limited number of distinct protein folds. For example, in an early paper, Chothia postulated that the number of unique protein folds would be on the order of only about 1000 unique protein folds. In another estimation, the number of distinct domains and folds were placed around 7000. Even though the number of new structures solved has been increasing at an accelerated rate (close to 3000 structures solved in 2002), the proportion of new folds, as determined by the CE algorithm (http://cl.sdsc.edu/ce.html), to the total number of new structures solved in a given year decreased from an average of ca. 30% in the 80's steadily down to only ca. 8% in year 2001 (http://www.rcsb.org/pdb/holdings.html). It is reasonable to expect that as more and more protein structures are determined experimentally, we will be able to find close structure analogues in the databases of known structures for almost any protein sequence in the near future.

Threading or fold recognition involves similar steps as comparative modeling. The difference is in the fold identification step. First of all, a structure library needs to be defined. The library can include whole chains, domains, or even conserved protein cores. Once the library is defined, the target sequence will be fitted to each library entry and a energy function is used to evaluate the fit between the target sequence and the library entries to determine the best possible templates. Depending on the algorithms to align the target sequence with the folds and the energy functions to determine the best fits, the threading methods can roughly be divided into four classes.

- The earliest threading methods used the environment of each residue in the structure as the energy function and dynamical programming to evaluate the fit and the alignment.
- Instead of using overly simplified residual environment as the energy function, statistically derived pair wise interaction potentials between residue pairs or atom pairs can be used to evaluate the best possible fits between the target sequence and library folds. In this method, for efficient optimal alignment between the target sequence and the folds, the potential for residual is obtained by summing over all

the pair wise potentials involving i, and then “double dynamical programming” method can be used.

- The third kind of methods does not use any explicit energy function at all. Instead, secondary structures and accessibility of each residue are predicted first and the target sequence and library folds are encoded into strings for the purpose of sequence-structure alignment.
- Finally, sequence similarity and threading can be combined for fold recognition. For large-scale genome wise protein structure prediction, sequence similarity can be first used for the initial alignments and the alignments can be evaluated by threading methods.

The threading methods are limited by the high computational cost since each entry in the whole library of thousands of possible folds needs to be aligned in all possible ways to select the fold(s). Another major bottleneck is the energy function used for the evaluation of the alignment. As these functions are drastically simplified for efficient evaluation, it is not reasonable to expect to be able to find the correct folds in all cases with a single form of energy function. Nevertheless, with the current functions, it is possible to reduce the thousands of possible folds to only a few. Similar to the comparative modeling case, for sequence similarities at protein family level, threading can produce alignments that are accurate to 1 to 3 _, or in the case with low sequence similarity at the super-family level, alignment at the range of 3 to 6_ can still be expected. As more protein structures are determined and sequence comparison methods improve, more and more target sequences fold assignment can be achieved by comparative modeling though. Worth mentioning is the threading program PROSPECT, which performed best in its category in the CASP4 competition. What is unique to PROSPECT is that it is designed to find the globally optimal sequence-structure alignment for the given form of energy function. The divide-and-conquer algorithm is used to speed up the calculation by explicitly avoiding the conformation search space that is shown not to contain the optimal alignment. In several cases that have sequence identity as low as 17%, perfect sequence-structure alignment is still achieved for the alignable potions between the target and template structures. Even in cases that no fold templates exist for the target sequence, important features of the structure are still recognized through threading the target sequence to the structures.

Protein folding

A protein is a polymer of fixed length, composition and structure, made up by a combination of the 20 naturally occurring amino acids. With 20 amino acids it is possible to generate 20 different chains of 200 amino acids each. Only a small fraction is actually used by living organisms. Out of all the possible sequences of amino acids, proteins have been selected by evolution over 10^6-10^7 of years to perform a specific biological task.

The role of protein folding

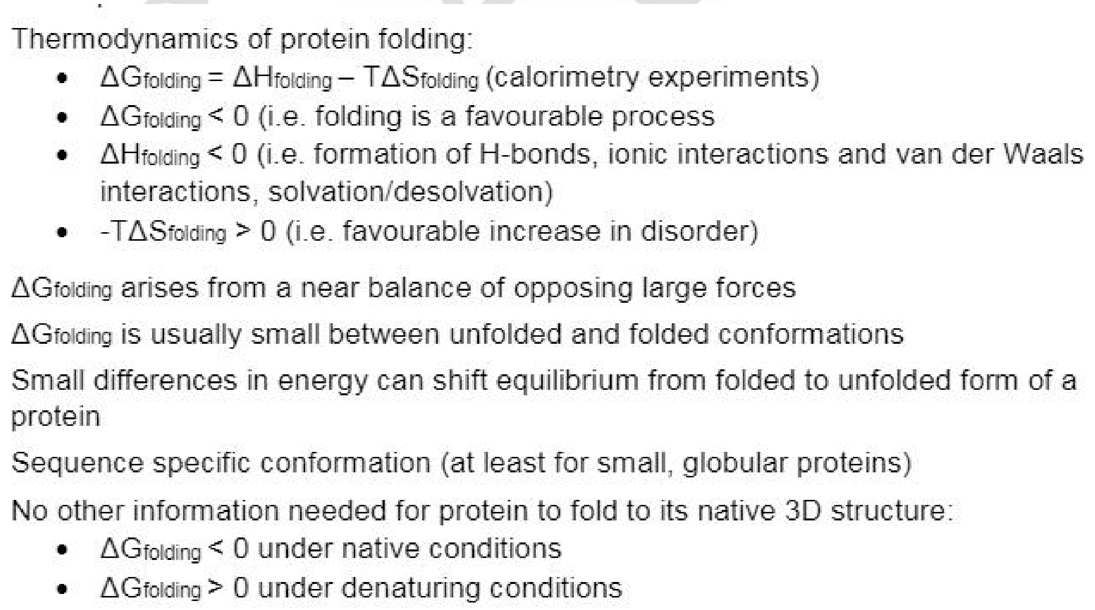
3D structure of each protein arises from folding onto a specific unique conformation with a particular function.

A protein made in the ribosome as a polypeptide chain must evolve to reach its more stable conformation- this can take between 1 to several minutes.

From sequence to 3D structure

3D structure of each protein arises from folding into a specific unique conformation with a particular function.

Thermodynamics of protein folding:



Thermodynamics of protein folding:

- $\Delta G_{folding} = \Delta H_{folding} - T\Delta S_{folding}$ (calorimetry experiments)
- $\Delta G_{folding} < 0$ (i.e. folding is a favourable process
- $\Delta H_{folding} < 0$ (i.e. formation of H-bonds, ionic interactions and van der Waals interactions, solvation/desolvation)
- $-T\Delta S_{folding} > 0$ (i.e. favourable increase in disorder)

$\Delta G_{folding}$ arises from a near balance of opposing large forces

$\Delta G_{folding}$ is usually small between unfolded and folded conformations

Small differences in energy can shift equilibrium from folded to unfolded form of a protein

Sequence specific conformation (at least for small, globular proteins)

No other information needed for protein to fold to its native 3D structure:

- $\Delta G_{folding} < 0$ under native conditions
- $\Delta G_{folding} > 0$ under denaturing conditions

Anfinsen's Experiments

B-mercaptoethanol is reducing agent that breaks disulphide bridges in proteins

Urea disrupts non-covalent interactions

Denaturation leads to complete inactivation of RNase

Dialysis was used to remove urea and air to re-oxidise protein

RNase recovers all activity

Correct tertiary structure of RNase backbone was recovered

Right SH groups must have been adjacent to each other prior to re-oxidation upon correct refolding of backbone because disulphide bridges formed spontaneously with the correct combination of Cys amino acids.

Anfinsen's Dogma

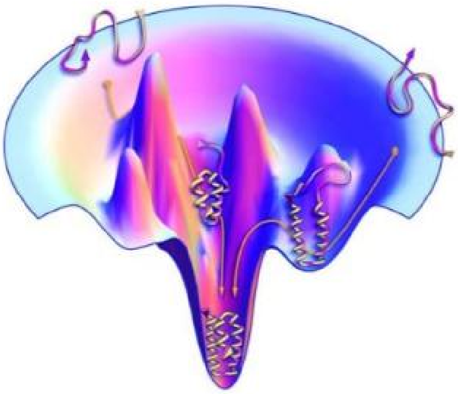
Native structure is determined only by the amino acid sequence of a protein, at least four globular proteins:

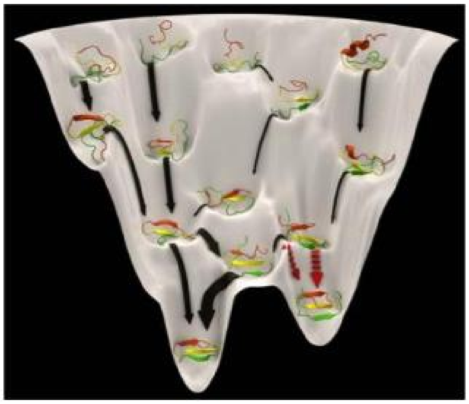
Uniqueness: native structure is the thermodynamically most stable (favoured) and thus unique

Stability: Small changes in surrounding environment do not affect free energy minimum configuration

Kinetic accessibility: path in the conformational free energy surface must be smooth.

Free energy landscape of protein folding





The Levinthal Paradox

If we take a protein 150 amino acids long and assume that if only has two main chain torsional degrees of freedom per amino acid.

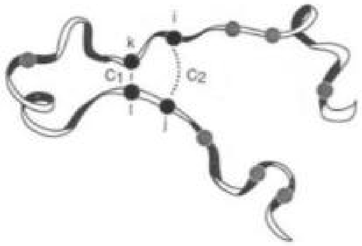
Proteins do not randomly search all possible conformations until they reach the most stable structure.

Cooperatively in protein folding

A protein will reach an optimal conformation without actually undertaking a global conformational search.

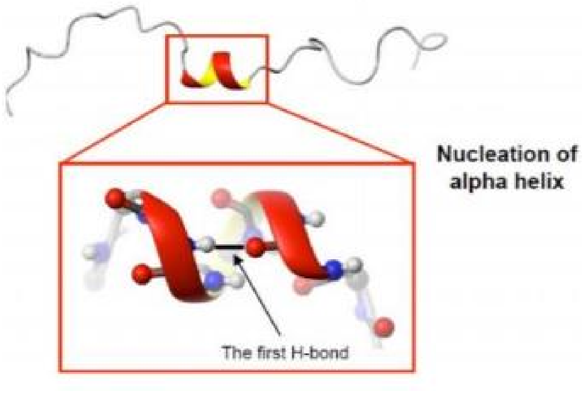
Depends on physic-chemical conditions: pH, temperature, ionic strength, redox potential

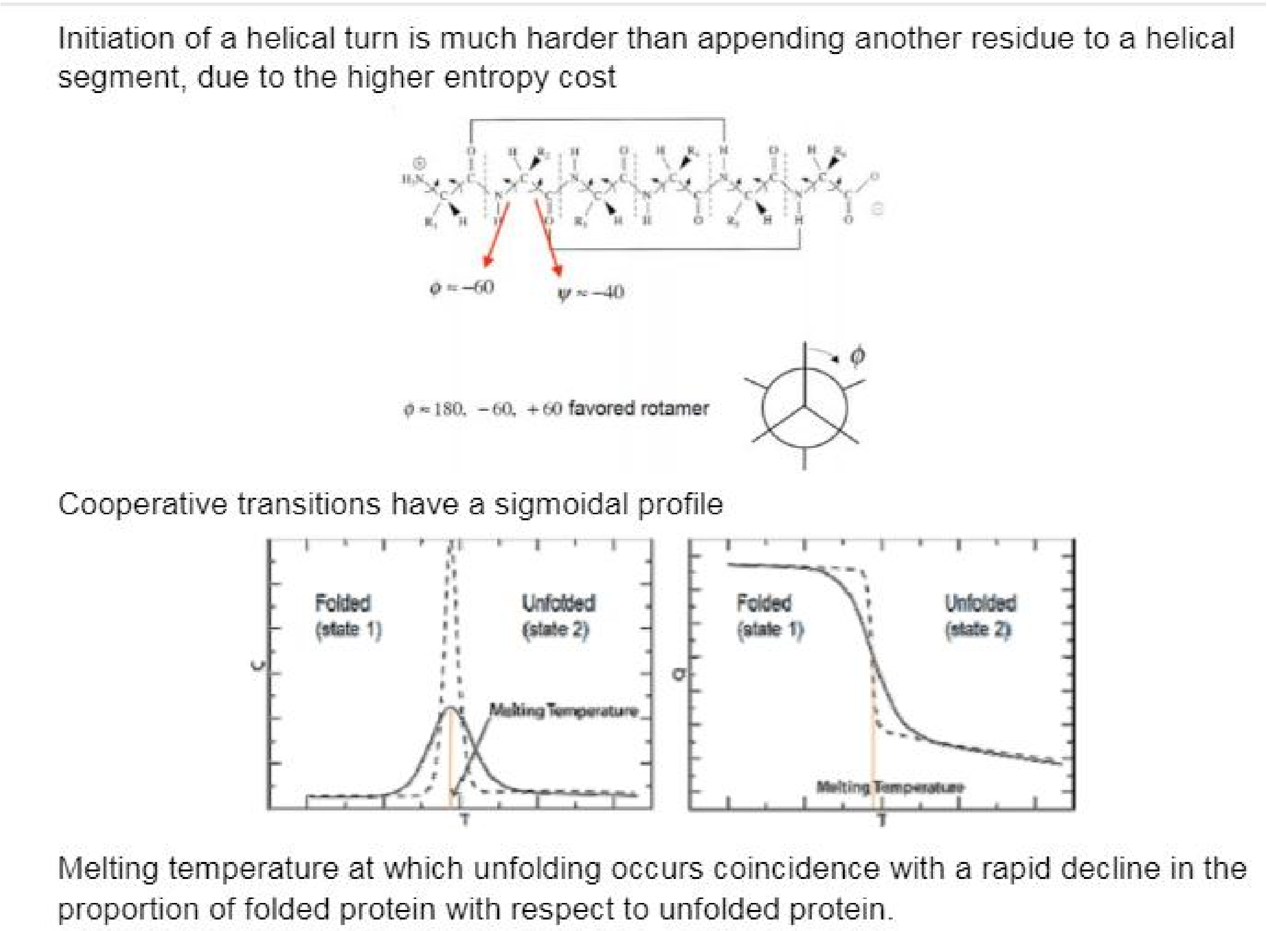
Cooperativity is essential, probability of forming contact C2 is much higher if C1 is formed that in the absence of C1.



The Coil-Helix Transition

The Paradigm for cooperativity in protein folding





Initiation of a helical turn is much harder than appending another residue to a helical segment, due to the higher entropy cost

Cooperative transitions have a sigmoidal profile

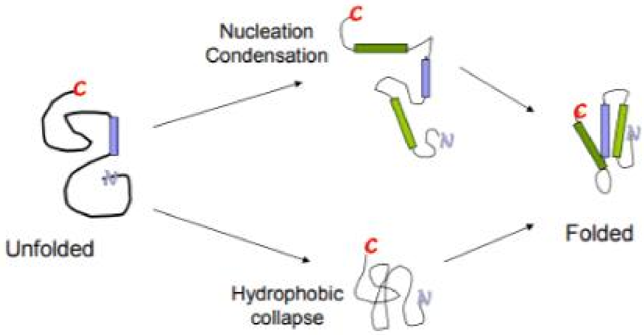
Melting temperature at which unfolding occurs coincidence with a rapid decline in the proportion of folded protein with respect to unfolded protein.

Mechanisms of Protein folding

Two over all mechanisms have been proposed:

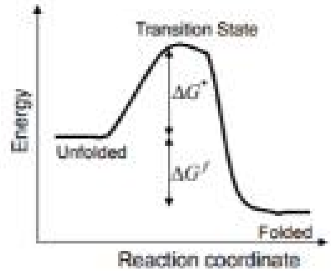
Nucleation-condensation: Some secondary structure motifs are formed and act as template for the formation of tertiary structure

Hydrophobic collapse: Hydrophobic interactions produce a compact structure (molten globule) that subsequently folds into its final state.



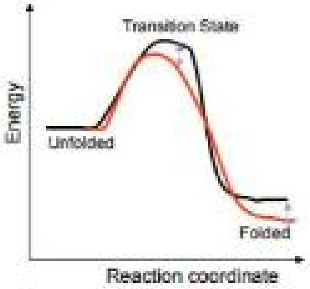
Characterisation of Folding States

All residues in a protein are mutated to Ala one by one and the stability towards denaturation and folding rate of each mutant with respect to the wild type sequence can be measured. The relative free energies of the folded, unfolded and transition state are calculated.



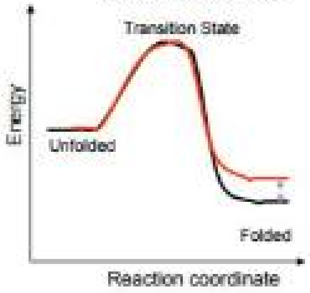
Φ values for each amino acid *i* are calculated as

$$\Phi_i = \frac{\Delta G^{*}_{wt} - \Delta G^{*}_{i}}{\Delta G^{f}_{wt} - \Delta G^{f}_{i}}$$



$$\Phi_i = \frac{\Delta G^{*}_{wt} - \Delta G^{*}_{i}}{\Delta G^{f}_{wt} - \Delta G^{f}_{i}}$$

A Φ value close to 1 indicates that the mutation has the same effect on the folded and the transition state. This can be interpreted as the residue being in a folded conformation at the transition state.



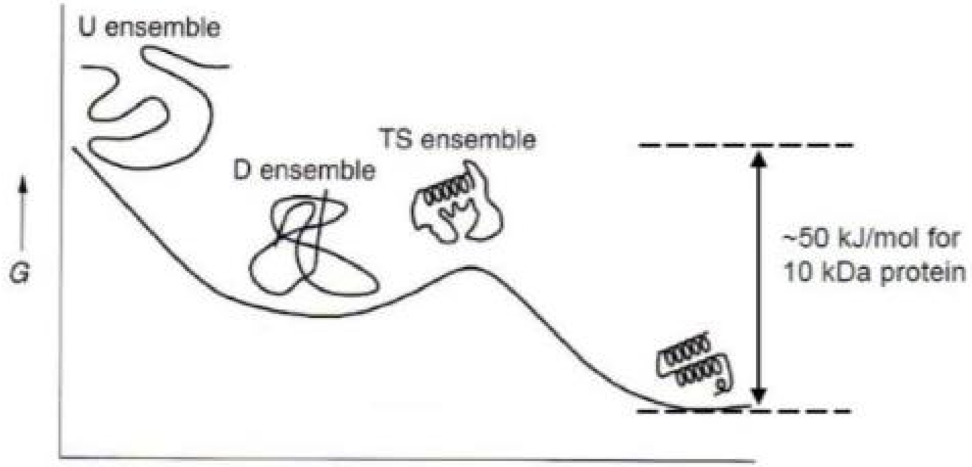
A Φ value close to 0 indicates that the mutation has no effect on the transition state. This can be interpreted as the residue being in an unfolded conformation at the transition state.

φ values of small protein indicate that α-helices are often formed before the transition state while the formation of β-sheets is rate-limiting.

Protein with high α-helix content fold much faster than proteins with large β-sheet content, although there are some exceptions These finding suggest a nucleation-condensation mechanism, at least for small proteins.

More on Thermodynamics of Folding

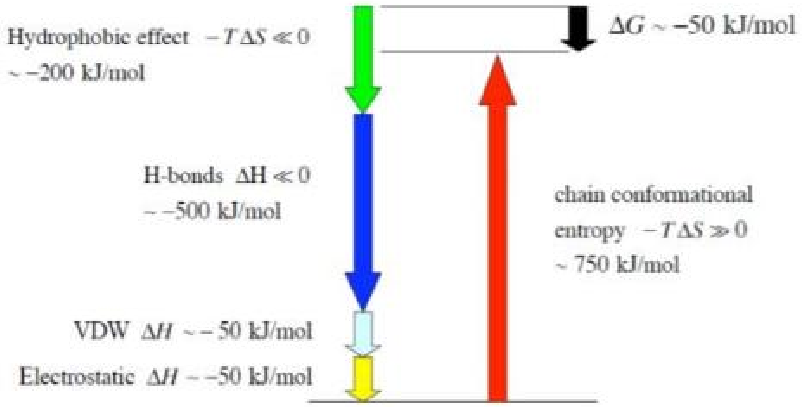
Simplifies free energy profile of protein folding:



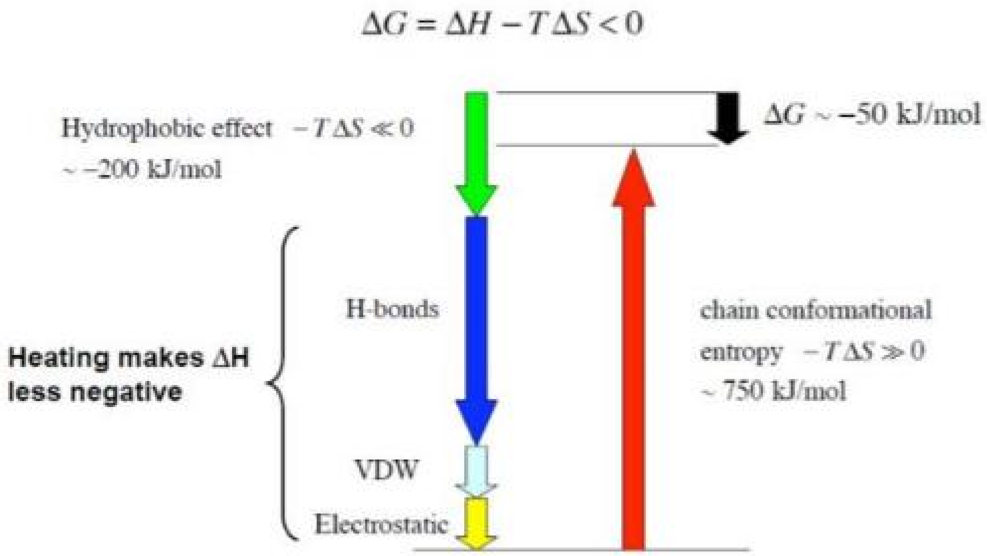
Dissecting the Free Energy of Folding

$$\text{Unfolded} \overset{\Delta G}{\rightleftharpoons} \text{Folded}$$

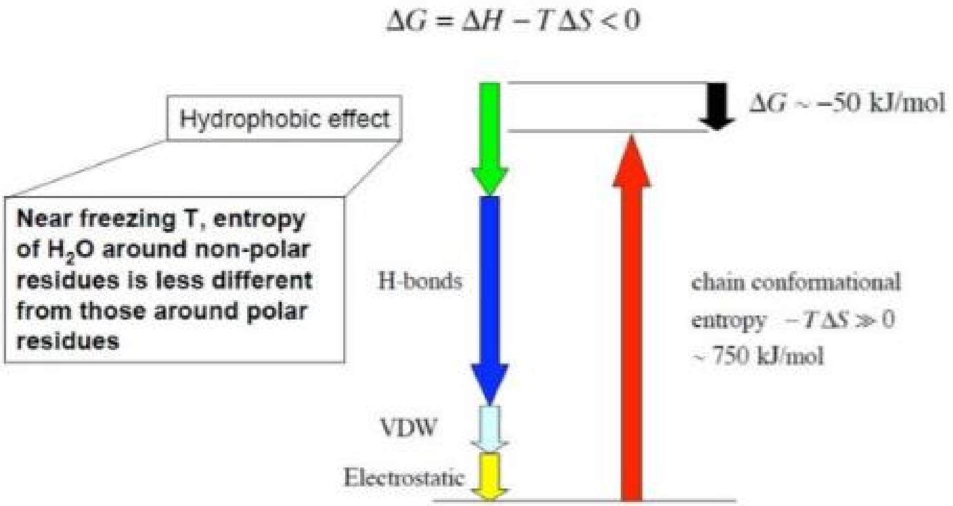
$$\Delta G = \Delta H - T\Delta S < 0, \ \ \Delta G =\sim -50 \text{ kJ/mol}$$


Hydrophobic effect $-T\Delta S \ll 0$
~ -200 kJ/mol
H-bonds $\Delta H \ll 0$
~ -500 kJ/mol
VDW $\Delta H \sim -50$ kJ/mol
Electrostatic $\Delta H \sim -50$ kJ/mol
$\Delta G \sim -50$ kJ/mol
chain conformational
entropy $-T\Delta S \gg 0$
~ 750 kJ/mol

Heat Denaturation Breaks Interactions


$\Delta G = \Delta H - T\Delta S < 0$
Hydrophobic effect $-T\Delta S \ll 0$
~ −200 kJ/mol
ΔG ~ −50 kJ/mol
Heating makes ΔH less negative
H-bonds
VDW
Electrostatic
chain conformational entropy $-T\Delta S \gg 0$
~ 750 kJ/mol

Cold Denaturation Reduces Hydrophobic Effect


$\Delta G = \Delta H - T\Delta S < 0$
Hydrophobic effect
Near freezing T, entropy of H_2O around non-polar residues is less different from those around polar residues
ΔG ~ −50 kJ/mol
H-bonds
VDW
Electrostatic
chain conformational entropy $-T\Delta S \gg 0$
~ 750 kJ/mol

Chaperone-Assisted Protein Folding

The two most important type of chaperones are Hsp60 and Hsp70:

- Hsp60 is found in bacteria, and provide a folding chamber
- Hsp70 is found in all living organisms, and mainly block aggregation

Visualization of Molecules

- Molecules are used two-dimensional (2D) structure and 3D structure.
- Mostly the molecules with interacting three dimensions.
- No. of tools are available for eg. Rotate, flip and otherwise manipulate virtual molecular models of chemicals and macromolecules.
- Small molecules in 3-D download and install on your computer.
- Scientific websites are JMol, Java-based viewer for rendering molecules in 3-D.
- Chime is the program used most for viewing small molecules from websites.
- Macromolecules in 3D download and install.

CHIME

- Chime is a free downloadable
- Its chemical structure visualization Plug-in windows and Macintosh.
- It allows view chemical structure from within popular web browsers Java applets and Java applications**.**
- Chime already be installed for their graphics to work properly.
- Rasmol, Chime shows molecules within a webpage.

Cn3D

- Cn3D is a visualization tool for macromolecules.
- To view 3-dimensional structure from NCBI's Entrez →it's a retrieval service.
- Cn3D is able to correlate structure and sequence information.
    - **Example**: find the residues in a crystal structure that correspond to known
    - disease mutations.
    - = powerful annotations and editing features.
- = right click on the molecule to see the viewing options.

RASMOL

- It's free program

- Developed by Roger A Sayle (1993) University of Edinburgh's, Biocomputing Research unit and the Biomolecular structure Department at Glaxo Research and Development Green Ford, UK.
- Rasmol derived from **Raster** (the array of pixels on a computer screen) molecules.
- Molecular graphics program for visualization of proteins, nucleic acids and small molecules.
- Powerful program aimed at display, teaching and generation of publication quality image.
- Rasmol reads in molecular co-ordinate files in formats like Brookhaven Protein Databanks (PDB).
- Different parts of the molecules displayed and colored independently rest of the molecule or show in different representations simultaneously.
- Molecule may be shown Wire frame, cylinder(deriding), stick bonds, alpha-carbon trace, space filling (CPK) spheres, macromolecular ribbons (either smooth shaded solid ribbons or parallel strands) hydrogen bonding and dot surface.
- Molecule displayed may be rotated, translated, zoomed, z-clipped (slabbed) interactively using either the mouse, the scroll bars, command line or an attached dials box.
- Model can be rotated about the x, y and z axes interactively so that all parts of the molecule can be studied.
- Smaller molecule or layer and in addition it is possible to expand the viewing window up to the full size of the screen.
- Larger picture more elaborate the model, the longer it takes the computer to calculate the appearance of the drawing.

Color schemes are available

CPK

Carbon atoms	→ Pale grey	Oxygen	→ red
Nitrogen	→ blue	Sulphur	→ yellow

Group colouring

- Chain colored with color of the rainbow.
- Blue – N-terminus
- Red – C-terminus
- Useful for following the fold from one end of the chain to the other.
- Shapely and amino colours
- Backbone – Pale gray side chain atoms are all given a colour depends upon the size and the polarity of the side chain.
- Oxygen containing side chain (acids, amides and the hydroxy-amino acids Ser and Thr.
- Various shade of red and basic side chain Blue (Arg, Lus, His)
- Hydrophobic amino acids are mostly Grey; Ile is dark green and Val a Pale Magenta.
- Sulphur containing amino acids (Cys, Met) have muddy yellow colors.
- Trp is yellow and Grey, White

Residue colour list

A (ala) – Pale green
C (cys) – Sandy yellow
D (Asp) – dark magenta
E (glu) – red
F (phe) – grey
G (gly) – White
H (his) – slate blue
I (ile) – dark green
K (lys) – royal blue
W (trp) – yellow

L (leu) - grey
M (met) – pale brown
N (asn) - salmon
P (pro) - grey
Q (gln) – flash pink
R (arg) – navy blue
S (ser) - tomato
T (thr) – orange red
V (val) – pale magenta
Y (tyr) – clay grey

- Rasmol prepared by list of commands from a script file
- Rasmol works well for both small molecules and for large ones such as proteins, DNA, RNA.

Protein Explorer a Rasmol derivatives:

- Protein explorer (PE) enables to explore the 3D structure of any macromolecule.
- Proteins, DNA, RAN, carbohydrates and complexes such as between transcriptional regulatory explorers.
- It is not compatible with Internet Explorer
- Firebox free and is recommended for protein explorer.

Biomodel – 3:

- Developed by Angel Herraez, Lecturer in Biochemistry and molecular Biology at the University of Alcala de Henares (Spain)
- Version V3
- Use J_{MOL}, Java applet to show manipulates the molecular models.

3D – Chemical libraries

- Use chime plug-in

3-D Virtual Chemistry Library

- Molecular database has about 150 molecules divided into six main groups.
- Simple molecule
- Polymers
- Senses
- Medical
- Horrible molecule and
- Interesting molecules
- I addition to structure it also has physical data, history and reactivity of the molecules.

3D Macromolecular structures

Using Cn3D

Entrez molecular modeling Databases

- It contains 3-D macromolecular structure
- Including proteins and polynucleotide

- MMDB contain over 40,000 structures and linked to the rest of the NCBI database
- Including sequences bibliographic citations, taxonomic classifications and sequence and structure neighbors.

Possible Questions

1. Write an account on protein structure prediction with suitable examples.

2. How will you visualize the molecules using RasMol ttol.

3. Explain the steps involved in homology modelling.

4. Discuss in detail about the threading method.

5. Define comparative modelling? Explain the steps involved in modelling the three dimensional structure of the protein.

6. Write an account on protein folding.

7. How will you construct a 3D protein model and validate it? Explain in a stepwise manner.

8. How will you visualize the molecules using Deep View tool

9. Write an account on protein structure prediction with suitable examples.

10. Discuss in some of commands used to visualize the molecules in the visualization software's.

KAPRAGAM ACADEM YOF HIGHER EDUCATION

DEPARTMENT OF BIOCHEMISTRY

I- M.Sc Biochemistry

18BCP204-Bioinformatics

S.NO	Unit 3	Questions	Option I	Option II	Option III	Option IV	Answers
1	3	The secondary prediction method is	nearest neighbour method	hidden morkov model	neural network	all the above	**neural network**
2	3	can be used for homology protein three dimensional structure	MEME	MODELLER	PDGCON	PROSITE	**MODELLER**
3	3	The secondary structure prediction was discovered in1951 by	linux paluing	Corey	both a and b	Michael zhang's	**Corey**
4	3	Genome represents to	entire genetic material	nucleus	gene	Protein	**entire genetic material**
5	3	an interuening region in sequence	intron	exon	EST	all of the above	**exon**
6	3	Protein structure can be measured by	bond angles	torsion angles	Bond length	All the above	**All the above**
7	3	A function of position of two atoms in proteins is	Bond length	Bond angle	Torsion angles	transitional angle	**Bond length**
8	3	A function of position of three atoms in proteins is	Bond lengts	Bond angle	Torsion angles	transitional angle	**Bond angle**
9	3	A function of position of four atoms in proteins is	Bond length	Bond angle	Torsion angles	transitional angle	**Torsion angles**
10	3	Example for α–helical protein is	Keratin	Myoglobin	Collagen	Hemoglobin	**Keratin**
11	3	One of the below given amino acids is known as imino acid	proline	glycine	lysine	Leucine	**proline**
12	3	In X- crystallography, the diffraction pattern is converted in to electron density maps by	mathematical Fourier transform	fingerprinting	ESR	resonance	**mathematical Fourier transform**
13	3	The helical rotation of the DNA double helix is known as	axial rise	helix sense	helix pitch	rotation per residue	**helix sense**
14	3	The first significant macromolecular sequence database was created by	Pearson	M.Dayhoff	Thompson *et al*	Alstch *et al*	**M.Dayhoff**
15	3	The solution obtained after extracting the absorbed substances in chromatography is termed as	Solute	Filterate	Elute	solvent	**Elute**
16	3	a molecular graphics program intended for the visualisation of proteins.	Mol mol	rasmol	both a and b	PDB	**both a and b**
17	3	are macromolecules compared of both RNA and several polypeptides.	Chromosomes	ribosomes	autosomes	genes	**ribosomes**
18	3	Protein sequence determines	genetic variation	genetic disorders	protein structure	domain	**protein structure**
19	3	Sequence within a single species that arose by gene duplication is called	homogenous	paralogous	homologous	orthologus	**paralogous**
20	3	A series of codons which can be translated into protein	anti codon	termionation codon	initiation codon	ORF	**ORF**
21	3	Scattered X-rays cause positive and negative interference, generating an ordered pattern of signals called	reflections	interference	scattering	diffraction	**reflections**
22	3	The gene expression implies	gene function	protein	gene regulation	genetic material	**gene function**
23	3	In the past direct protein sequencing was carried out the process	Sanger method	Edman degradation	spectroscopy	both a & b	**spectroscopy**
24	3	Protein structure can be determined using spectroscopy	IR	UV	NMR	HPLC	**NMR**
25	3	The structure data from databank can be downloaded and fed into the	molecularvisualization tool to visualize the of the molecules	one dimensional structure	two dimensional structure	three dimensional structure	**three dimensional structure**
26	3	Three subfields of Genomics are and	Structural, functional and comparative	clustering, cladistic and distance	maximum likelihood,parsimony and upgma	Phylogenetic	**Structural, functional and comparative**
27	3	Functional genomics is the study of the structure,expression patterns, interactions, and regulations of an encoded by genome.	RNAs and Proteins	DNAs and proteins	genes and proteins	Trna	**RNAs and Proteins**
28	3	Proteomics is the cataloging and analysis of to determine when a protein is expressed.	DNA	Gene	Proteins	Aminoacids	**Proteins**
29	3	The term proteomics indicates proteins expressed by a	DNA	Gene	Genome	Aminoacids	**Genome**
30	3	Proteomics can be divided into and	Expression proteomics and cell-map proteomics	Structural and functional proteomics	Both a and b	conserved regions	**Expression proteomics and cell-map proteomics**
31	3	The tool that calculates the isoelectric point and molecular weight of an input sequence.	Mw/PI	MI/Pw	PI/Mw	Ip/Wm	**PI/Mw**
32	3	Domain helps to attain ___________	Stability of nucleotides	Stability of protein	Stability of gene	Stability of chromosomes	**Stability of protein**
33	3	The link for PDB	www.rcsb.org/pdb	www.pdb.com	www.pdb.org	www.pdb.ac.in	**www.rcsb.org/pdb**
34	3	In PDB protein code repersents in	numeric	alpha	alpha numeric	float	**alpha numeric**
35	3	__ Visualization tools	RASMOL	Deepviewer	RASMOL and Deepviewer	PDB	**RASMOL and Deepviewer**
36	3	The α helix has ___ amino acids per turn with an H bond formed between every fourth residue	3.6	3.4	3.2	3.1	**3.6**
37	3	The activity of a gene is called_______	Gene function	gene expression	gene regulation	gene function	**gene expression**
38	3	β sheets are formed by H bonds between an average of __ consecutive amino acids	5--6	6--7	5--10	7--9	**5--10**
39	3	Expand FSSP	Families of structurally similar proteins	Families similar proteins	Function of structurally similar proteins	function similar protein	**Families of structurally similar proteins**
40	3	A region of secondary structure that is not a α helix, a β sheet	sheets	coils	Secondary structure	turns	**coils**
41	3	A type of architecture that also has a conserved loop structure	blocks	coils	folds	loop	**folds**
42	3	The Chou-Fasman method is roughly __accurate in predicting secondary structures.	70%	80%	90%	50-60%	**50-60%**
43	3	_____ is a widely used software for remote homology detection based on pairwise comparison of hidden Markov models.	Hpredict	predict H	Ipredict	HHpred	**HHpred**
44	3	PDB stands for	Protein Databank	Pattern Databank	Protein Database	Pattern Database	**Protein Databank**
45	3	Which one of the following is gene prediction tool?	Gene scan	Expasy	NCBI	Uniprot	**Gene scan**
46	3	Which one of the following is gene prediction tool?	Swiss Prot	Procheck	Gen Mark	Trimmer	**Gen Mark**
47	3	Comparitive modeling is also called as	Threading	Ab intio	Homology modeling	None of the above	**Homology modeling**
48	3	How many levels of protein structure are	Three	Four	Two	One	**Four**
49	3	Fold recognition is also called as	Threading	Ab intio	Homology modeling	None of the above	**Threading**
50	3	Biologically active protein structure is	Primary Level	Secondary Level	Both a and b	Tertiary Level	**Tertiary Level**
51	3	Linear sequence of amino acids is	Primary Level	Secondary Level	Both a and b	Tertiary Level	**Primary Level**
52	3	Hydrogen bonds stabilises the	Primary Level	Secondary Level	Both a and b	Tertiary Level	**Secondary Level**
53	3	Dimerization of protein monomers is	Primary Level	Secondary Level	Tertiary Level	Quarternary Level	**Quarternary Level**
54	3	Disulphide bonds playing a major role in stablizing the protein structure at	Primary Level	Secondary Level	Tertiary Level	Both a and b	**Tertiary Level**
55	3	Which of the following amino acid is exceptional in Ramachandran plot?	Glycine	Methionine	Lysine	Alanine	**Glycine**
56	3	Which are of the following amino acids is Ramachandran plot exception?	Methionine	Lysine	Alanine	Proline	**Proline**
57	3	Which of the following technique is used in genomics?	SAGE	Protein Purification	Homology Modeling	Virtual screening	**SAGE**
58	3	Which one of the following technique is used in differential gene expression?	SAGE	Protein Purification	Homology Modeling	Virtual screening	**SAGE**
59	3	Which one of the following is not a gene prediction tool?	FGENESH	GeneMark	PDB	GENEID	**PDB**

UNIT-IV

SYLLABUS

Gene Identification and Prediction: Genome sequencing, Genome database: SWISS-2D PAGE, Gene Mark, Gene Scan, Pattern Recognition, Global Gene expression studies-DNA Micro array.

Genome sequencing

Genome database: SWISS-2D PAGE

Annotated two-dimensional polyacrylamide gel electrophoresis (2-D PAGE) and SDS-PAGE.

Established in 1993 and maintained collaboratively by the Central Clinical Chemistry Laboratory of the Geneva University Hospital and the Swiss Institute of Bioinformatics (SIB).

Each SWISS-2DPAGE entry contains textual data on one protein, including

- Mapping procedures.
- Physiological and pathological information.
- Experimental data (isoelectric point, molecular weight, amino acid composition, peptide masses).
- Bibliographical references.
- images showing the experimentally determined location of the protein, as well as a theoretical region computed from the sequence protein, indicating where the protein might be found in the gel.
- Cross-references to Medline and other federated 2-DE databases and molecular databases.

Gene Mark

Gene Mark developed in 1993 was the first gene finding method recognized as an efficient and accurate tool for genome projects. Gene Mark was used for annotation of the first completely sequenced bacteria, *Haemophilus influenzae*, and the first completely sequenced archaea, *Methanococcus jannaschii*. The Gene Mark algorithm uses species specific inhomogeneous Markov chain models of protein-coding DNA sequence as well as homogeneous Markov chain models of non- coding DNA. Parameters of the models are estimated from training sets of sequences of known type. The major step of the algorithm computes a posteriory probability of a sequence fragment to carry on a genetic code in one of six possible frames (including three frames in complementary DNA strand) or to be "non-coding".

Gene Scan

In bioinformatics GENSCAN is a program to identify complete gene structures in genomic DNA. It is a GHMM-based program that can be used to predict the location of genes and their exon-intron boundaries in genomic sequences from a variety of organisms. The GENSCAN Web server can be found at MIT. GENSCAN was developed by Christopher Burge in the research group of Samuel Karlin, Department of Mathematics, Stanford University. It is a general probabilistic model of the gene structure of human genomic sequences which incorporates descriptions of the basic transcriptional, translational and splicing signals, as well as length distributions and compositional features of exons, introns and intergenic regions. Distinct sets of model parameters are derived to account for the many substantial differences in gene density and structure observed in distinct C + G compositional regions of the human genome. In addition, new models of the donor and acceptor splice signals are described which capture potentially important dependencies between signal positions. The model is applied to the problem of gene identification in a computer program, GENSCAN, which identifies complete exon/intron structures of genes in genomic DNA. Novel features of the program include the capacity to predict multiple genes in a sequence, to deal with partial as well as complete genes, and to predict consistent sets of genes occurring on either or both DNA strands. GENSCAN is shown to have substantially higher accuracy than existing methods when tested on standardized sets of human and vertebrate genes, with 75 to 80% of exons identified exactly. The program is also capable of indicating fairly accurately the reliability of each predicted exon. Consistently high levels of accuracy are observed for sequences of differing C + G content and for distinct groups of vertebrates.

Both GeneMark and GeneMark.hmm can be used via the Gene Mark website for the analysis of prokaryotic DNA, with 175 pre-computed species-specific statistical models available. Analysis of DNA from any prokaryotic species is supported by (i) a special version of GeneMark.hmm using a heuristic model calculated from the nucleotide frequencies of an input sequence at least 400 nt long and (ii) a self-training program, GeneMarkS, which can be used for longer sequences on the order of 1 Mb in length. Thus, the DNA of any prokaryote can be analysed, via either a pre-computed species-specific model or a model created on the fly.

As many of the programs at the Gene Mark website share similar interfaces, we use here the prokaryotic GeneMark.hmm program as an exemplar and discuss program-specific differences below, where appropriate.

The GeneMark.hmm web interface accepts as input a single DNA sequence as an uploaded file or as text pasted into a textbox. If a FASTA description line begins the sequence, all text on the line following the 'greater than' symbol (>) is used as the title. In the remainder of the submission, digits and white space characters are ignored and letters other than T, C, A and G (assumed to appear rarely) are converted to N. The interface requires selection of the species name. Selection of a model for the RBS (in the form of a position-specific weight matrix and a spacer length distribution) is optional. In certain cases, such as the crenarchaeote *Pyrobaculum aerophilum*, the RBS model is replaced by a promoter model, which is the dominant regulatory motif located upstream to gene starts in this species. The interface also includes the option of using other types of genetic codes such as the Mycoplasma genetic code.

GeneMark.hmm reports all predicted genes in a format that includes the strand the gene resides on, its boundaries, length in nucleotides and gene class. Class indicates which of the two Markov chain models used in GeneMark.hmm, Typical or Atypical gene model, provided the higher likelihood for the gene sequence. Genes of the Typical class exhibit codon usage patterns specific to the majority of genes in the given species, while Atypical class genes may not follow such patterns and frequently contain significant numbers of laterally transferred genes. The nucleotide sequences of predicted genes and translated protein sequences are available as an output to facilitate further analysis, such as BLAST searching. An option to generate Gene Mark predictions in parallel with the GeneMark.hmm analysis provides important additional information. In this case, Gene Mark is set up to use models derived from the same training data as models for the current run of GeneMark.hmm.

It is worth noting that the GeneMark.hmm and GeneMark algorithms are complementary to each other in the same way as the Viterbi algorithm and the posterior decoding algorithm are. Therefore, though the two algorithms are distinct, they are supposed to generate predictions largely corroborating and validating each other. Differences frequently indicate sequence errors and deviations in gene organization, very short genes, gene fragments, gene overlaps, etc.

Graphical output of the analysis is available in PDF or PostScript format. The graphical output clearly depicts the advantage of using multiple Markov chain models representing different classes of genes. Here, the coding potential graph obtained using the Typical gene model, derived by GeneMarkS, is denoted by a solid black line, and the coding potential graph obtained using the Atypical gene model (derived by a heuristic approach) is denoted by a dotted line. The GeneMark graph also includes indications of frameshift positions (also listed in the text report), which are often sequencing errors but in rare cases are natural and biologically very interesting.

For the GeneMark program, there are several specific options. The window size and step size parameters (96 nt and 12 nt, respectively, by default) define the size of the sliding window and how far this window is moved along the sequence in one step. The threshold parameter determines the minimal average coding potential for an open reading frame (ORF) to be predicted as a gene. There are several options which allow fine-tuning of the Gene Mark graphical output. In addition, there are options supporting the analysis of eukaryotic DNA sequences by Gene Mark including the ability to provide lists of putative splice sites and protein translations of predicted exons. As might be expected, Gene Mark (the posterior decoding algorithm) does not produce high enough resolution for the precise prediction of exon–intron borders. Thus, GeneMark.hmm (the generalized Viterbi algorithm) in its eukaryotic version is the major tool for the identification of exon–intron structures in eukaryotic DNA sequences.

The output of the Gene Mark program consists of a list of ORFs predicted as genes, i.e. those with average coding potential above the selected threshold. Although each predicted gene can have more than one potential start, additional data is provided to help the researcher annotate one of the alternatives as the 'true' one. The start probability (abbreviated 'Start Prob') is derived from the sequences in the windows immediately upstream and downstream of each potential start. RBS information is provided in the form of a probability score along with the position and sequence of the potential RBS (abbreviated 'RBS Prob', 'RBS Site' and 'RBS Seq'). In addition to the list of predicted genes, Gene Mark provides a list of 'regions of interest', spans of significant length between in-frame stop codons where spikes of coding potential are wide enough and may warrant further analysis even if no genes are predicted therein based on automatic comparison with the threshold.

Analysis of prokaryotic DNA sequences for which there is no pre-computed species-specific model can be carried out using a program version which heuristically derives a model for any input sequence >400 nt. This approach has also proven useful for the analysis of inhomogeneous genomes, particularly regions too divergent from the bulk of the genome, such as pathogenicity islands.

If models (including RBS models) have to be computed *de novo* for an anonymous DNA sequence with length of the order of 1 Mb or longer, the GeneMarkS program can be used. This program needs significantly more computational resources; thus, its output is provided via email. A modified version of GeneMarkS tuned for the analysis of viruses of eukaryotic hosts creates a model for the Kozak consensus sequence instead of a two-component RBS model.

The eukaryotic version of GeneMark.hmm is currently available for the analysis of 11 eukaryotic genomes: *Homo sapiens*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Chlamydomonas reinhardtii*, *Drosophila melanogaster*, *Gallus gallus*, *Hordeum vulgare*, *Mus musculus*, *Oryza sativa*, *Triticum aestivum* and *Zea mays*. From the prediction accuracy tables given at the website (http://opal.biology.gatech.edu/GeneMark/plant_accuracy.html), it follows that the latest versions of GeneMark.hmm produce remarkably accurate gene predictions for plant genomes such as rice and *Arabidopsis*. This fact has not escaped the attention of plant genome sequencing consortiums, which have used the program intensively. The analysis of cDNA and EST sequences from eukaryotes, which typically contain no introns, is facilitated by a special version of Gene Mark called GeneMark.SPL. Interestingly, eukaryotic genomes with rare introns present difficulty in terms of collecting enough statistics for the intron and internal exon related models, the important components of a full-fledged eukaryotic gene finder. For this reason, a special interface is available for low eukaryotes such as *Saccharomyces cerevisiae*. Currently, this interface employs versions of prokaryotic Gene Mark and GeneMark.hmm augmented with Kozak start site models instead of the prokaryotic RBS model.

The eukaryotic species-specific models are represented by several variants built for distinct G + C% ranges covering the whole scale of G + C inhomogeneity observed in a particular genome. GeneMark.hmm automatically selects the model variant which fits the G + C% of the input sequence. Note that, in the eukaryotic case, the RepeatMasker program (www.repeatmasker.org), which is frequently used for pre-processing, can introduce a significant number of 'N' characters. These characters do not influence the selection of the Markov chain model used in prediction.

In the graphical output of the eukaryotic version of GeneMark.hmm, the thick horizontal bars (which represent whole genes in the prokaryotic case) indicate predicted exons. Vertical ticks on these bars show the starts and ends of predicted initial and terminal exons, respectively.

For the analysis of virus and phage DNA, the heuristic (for short genomes) and GeneMarkS (for long genomes) options, mentioned above, are recommended. In addition, a database called VIOLIN containing pre-computed reannotations of >1000 virus genomes is available.

Future directions for Gene Mark web software development include detection of several genomic elements currently not predicted by either Gene Mark or GeneMark.hmm, such as rRNA and tRNA genes (which can be mis-predicted as protein-coding genes in low G+C% species) and improving the detection of gene 5′ ends. Currently, the server supports the analysis of sequences masked by tRNAscan or similar programs. The Gene Mark programs will not find genes in these masked areas (sequences of 'N' characters); thus, the predictions will be compatible with this extrinsic information. The detection of exact gene starts remains a challenging problem in gene finding, as many genes have relatively weak patterns indicating sites of translation and transcription initiation. This problem is made especially difficult by the lack of available data sets containing verified gene start locations to be used for training and evaluation. Refinements in the RBS and Kozak models and the potential inclusion of hidden states representing upstream promoter sequences are currently being explored to address this issue.

Patter Recognition

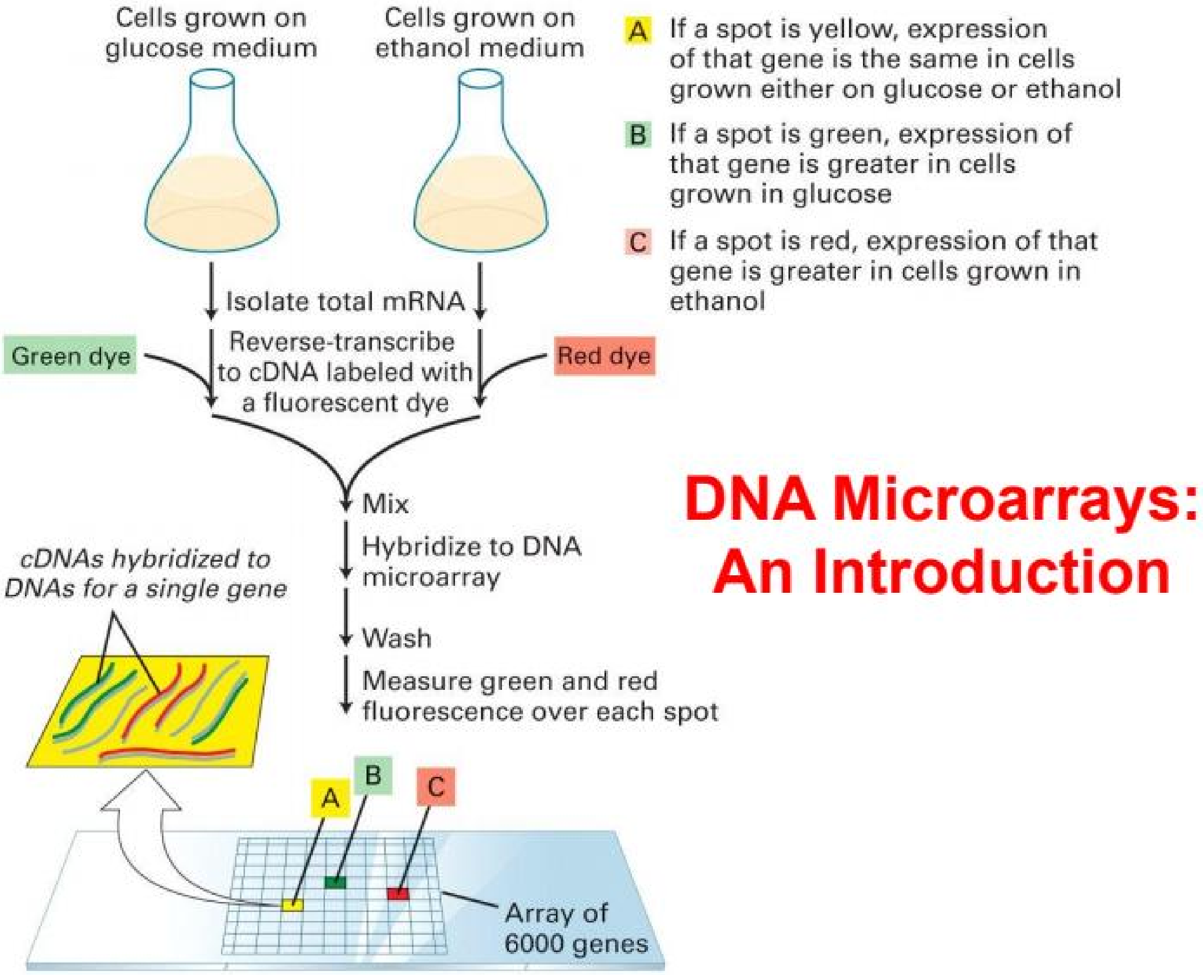
Every accumulation of data in its raw form holds obscure patterns. Pattern recognition deals with the science of transforming and classifying entities on the basis of these patterns. It is a vast field as it deals with data from diverse sources. Data can be of single dimensional nature as in case of stock exchanges and sound, two-dimensional as in case of images, and even multidimensional. It has many applications, for example, in medical science, it provides origins for computer-aided diagnosis (CAD) which supports medical practitioners in interpretations and finding of diseases. It has other typical applications: automatic speech recognition; recognition of text in various categories; and automatic recognition of human faces. Moreover, the genetic and protein structure in living organisms form intrinsic patterns. Data collected from the decomposition of these proteins help to identify them and hence to classify the protein. The ultimate objective is to make machines ideally as intelligent as humans in recognizing such patterns which help to form automated systems for conduction of routine matters.

Bioinformatics deals with development of algorithms and software for understanding the biological data. For analyzing and interpretation of the biological data, bioinformatics uses mathematics, statistics, computer, and engineering. There exists a lot of work in molecular biology using various approaches of bioinformatics like image processing and machine learning.

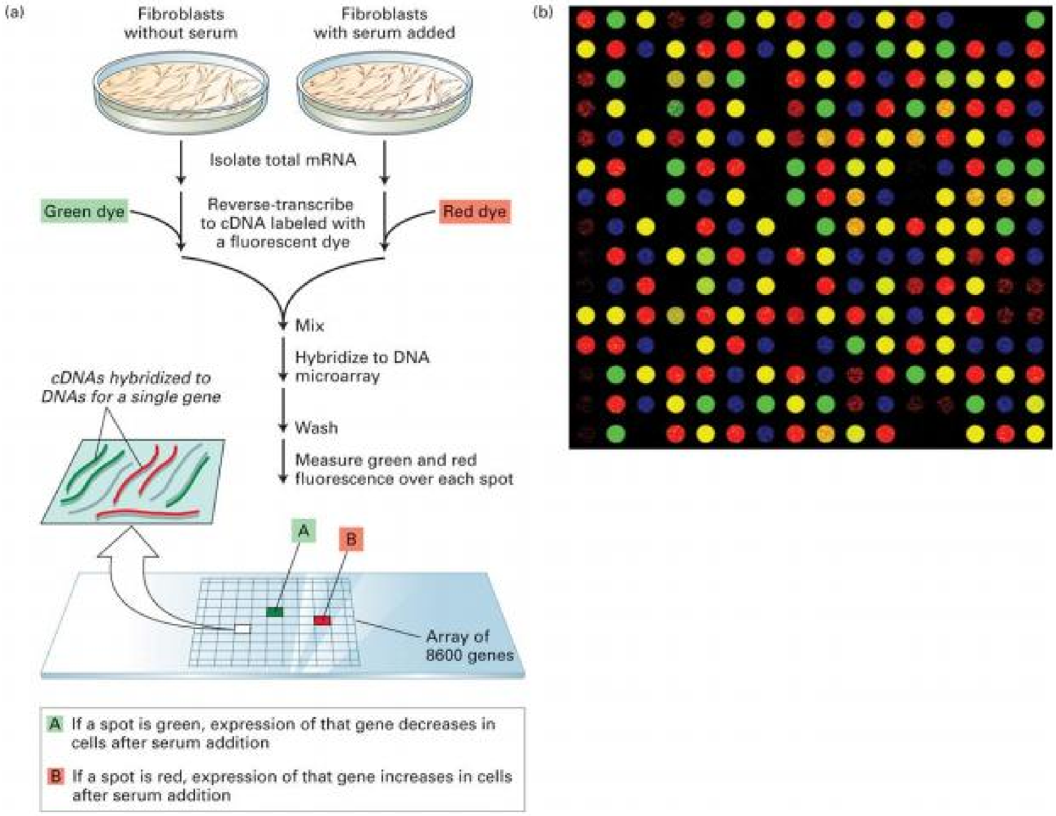
Bioinformatics not just deals with application of pattern recognition for protein classification but it also incorporates use of computational intelligence in protein sequencing, gene expression, comparative genomics, mutation, disease genetics, and molecular interactive networks.

High-throughput measurement technologies, such as cDNA and oligonucleotide microarrays, are changing the practice of biology and medicine. Microarrays provide simultaneous expression (RNA abundance) measurements for thousands of genes and thereby facilitate analysis of the complex multivariate relations among genes. This new capability is being used to promote two major goals of functional genomics: (1) to use gene expression to classify disease on a molecular level; and (2) to discover genes that determine specific cellular phenotypes (diseases) and model their activity in a way that provides quantitative discrimination between normal and abnormal behavior. These goals correspond to diagnosing the presence or type of disease and to developing therapies based on the disruption or mitigation of aberrant gene function contributing to the pathology of a disease. Developing diagnostic tools at the RNA level involves designing expression-based classifiers to discriminate differences in cell state, suchas one type of cancer or another. Engineering therapeutic tools involves synthesizing nonlinear dynamical networks to model gene regulation and deriving intervention strategies to modify network behavior. The classification methods of pattern recognition are clearly associated withdiagnosis, but they also apply to therapy because prediction methods are used to identify gene–gene and gene–phenotype relations in network modeling. In discrete models, prediction of a targetgene value is given via a function of some predictor-gene values. This function is a multinomial classifier.

DNA Microarray


Cells grown on glucose medium
Cells grown on ethanol medium
A If a spot is yellow, expression of that gene is the same in cells grown either on glucose or ethanol
B If a spot is green, expression of that gene is greater in cells grown in glucose
C If a spot is red, expression of that gene is greater in cells grown in ethanol
Isolate total mRNA
Green dye
Reverse-transcribe to cDNA labeled with a fluorescent dye
Red dye
Mix
DNA Microarrays: An Introduction
Hybridize to DNA microarray
cDNAs hybridized to DNAs for a single gene
Wash
Measure green and red fluorescence over each spot
A
B
C
Array of 6000 genes

DNA microarray analysis can reveal differences in gene expression in fibroblasts under different experimental conditions


(a)
Fibroblasts without serum
Fibroblasts with serum added
Isolate total mRNA
Green dye
Reverse-transcribe to cDNA labeled with a fluorescent dye
Red dye
Mix
Hybridize to DNA microarray
Wash
Measure green and red fluorescence over each spot
cDNAs hybridized to DNAs for a single gene
A
B
Array of 8600 genes
A If a spot is green, expression of that gene decreases in cells after serum addition
B If a spot is red, expression of that gene increases in cells after serum addition
(b)

Possible Question

1. Write in detail about DNA Microarray
2. Write an account on gene prediction
3. Application of DNA microarray
4. Enumerate the importance of pattern recognition
5. Write a note on SAGE analysis
6. Discuss the gene prediction tools
7. Compate Genscan vs Gene mark
8. Write note on oligomers and its importance in micro array.
9. Write in detail about differential gene expression analysis.
10. Explain the importance of gene prediction methods in whole genome sequencing
11. Write short notes on Gene Mark.

KAPRAGAM ACADEM YOF HIGHER EDUCATION

DEPARTMENT OF BIOCHEMISTRY
I- M.Sc Biochemistry
18BCP204-Bioinformatics

S. No	Unit	Questions	Option I	Option II	Option III	Option IV	Answer
1	4	After transcription, mRNA goes through processing in eukaryotic cells. Why do prokaryotic cells not use mRNA processing	Prokaryotes do not possess hydrolytic enzymes against which processing protects	In prokaryotes, operons are used to regulate mRNA	mRNA processing only evolved in eukaryotes	Prokaryotes have no nucleus so gene expression occurs all together	**Prokaryotes have no nucleus so gene expression occurs all together**
2	4	During chromosomal replication, DNA is built in the 5' - 3' direction. Why does this occur	Building from 5'- 3' conserves energy	The replication fork runs in this direction	Okazaki fragments prevent building in the opposite direction	Nucleotides are added to the -OH end (3' end) of the sugar backbone	**Nucleotides are added to the -OH end (3' end) of the sugar backbone**
3	4	Alternative splicing is a process that enables the number of proteins produced by an organism to be vastly greater than its number of genes. How is this possible?	Codons can code for more than one amino acid	Recombinant technology is able to translate different proteins from the same gene	Depending on what sections are treated as introns and exons, different proteins can be made from the same gene	The anticodon of tRNA has a wobble effect that allows a variety of translations per gene	**Depending on what sections are treated as introns and exons, different proteins can be made from the same gene**
4	4	Which of the following is not an example of a point mutation	Silent mutation	Nonsense mutation	Missense mutation	Frameshift mutation	**Frameshift mutation**
5	4	codons are triplet nucleotide sequences that play an essential role in translation. Where are codons located	tRNA	mRNA	DNA template strand	rRNA	**mRNA**
6	4	Prokaryotic genes are often clustered into operons. Which of the following is not part of an operon	Genes	Operator	Exon	Promoter	**Exon**
7	4	DNA polymerase and RNA polymerase have similar functions. Which statement is an incorrect description of these enzymes	Both enzymes require a primer	DNA polymerase is used in replication while RNA polymerase is used in transcription	Both enzymes add nucleotides to the 3' end of a chain	The binding of RNA polymerase is preceded by the binding of transcription factors	**Both enzymes require a primer**
8	4	frameshit mutations are the results of what occurrence	Insertions or deletions that are not multiple of three	A mutation that changes an amino acid codon to a stop codon	A mutation that changes one amino acid to another		**Insertions or deletions that are not multiple of three**
9	4	DNA microarrays are used for	DNA variation screening	Gene expression profiling	Microarray comparative genomic hybridization	All of the above	**All of the above**
10	4	The DNA microarray technology that indicates which genes are transcribed in called	DNA variation screening	Gene expression profiling	Microarray comparative genomic hybridization	Antisense	**Gene expression profiling**
11	4	The DNA microarray technology that tracks deletions and amplifications of specific DNA sequences is called	DNA variation screening	Gene expression profiling	Microarray comparative genomic hybridization	Antisense	**Microarray comparative genomic hybridization**

UNIT-V

SYLLABUS

Applications of Bioinformatics: Molecular medicine, Biotechnology, Agricultural, Computer Aided Drug Design (structure and ligand based drug designing), Lead Molecular, Properties, ADME Profiles, QSAR, Receptor Docking, Introduction to molecular dynamics simulation.

Applications of Bioinformatics

- Bioinformatics is the use of IT in biotechnology for the data storage, data warehousing and analyzing the DNA sequences.
- In Bioinformatics knowledge of many branches are required like biology, mathematics, computer science, laws of physics & chemistry, and of course sound knowledge of IT to analyze biotech data.
- Bioinformatics is not limited to the computing data, but in reality it can be used to solve many biological problems and find out how living things works.

Bioinformatics is being used in following fields:

- Molecular medicine
- Personalised medicine
- Preventative medicine
- Gene therapy
- Drug development
- Microbial genome applications
- Waste cleanup
- Climate change Studies
- Alternative energy sources
- Biotechnology

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 (e.g. 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 inheritance 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 makes 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 unraveled, 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 person's 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 glutamic*um 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 bone meal.
- *Xanthomonas campestris* pv. is grown commercially to produce the exopolysaccharide xanthan gum, which is used as a viscosifying and stabilizing 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 bread 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 a 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 Defense 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 feelings

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 aluminum 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.

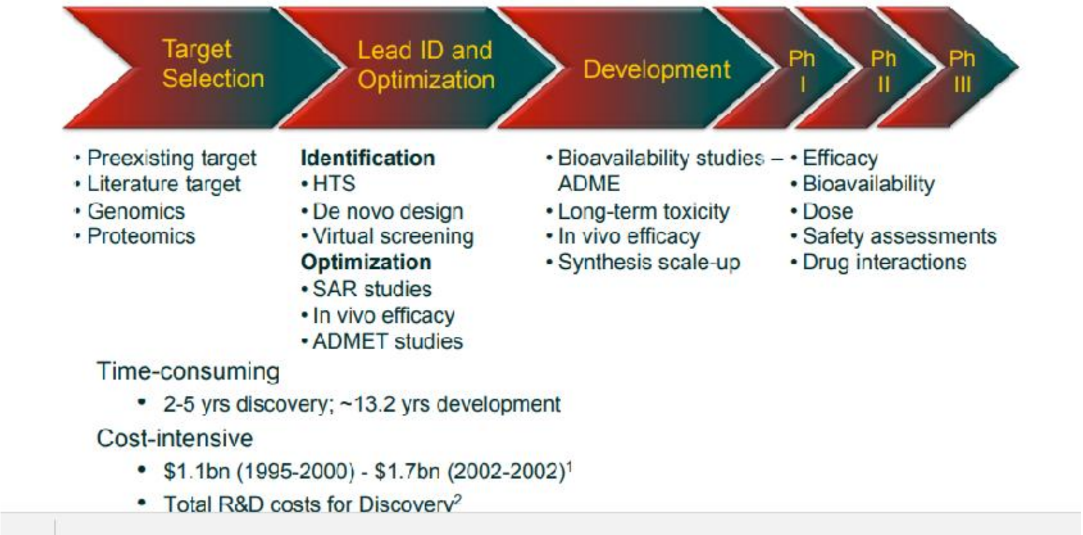
Computer aided drug design

Computer-aided drug design, often called structure based drug design involves using the biochemical information of ligand-receptor interaction in order to postulate ligand refinements. For example, if we know the binding site the steric complementarity of the ligand could be improved to increase the affinity for its receptor. Indeed, using the crystal structure of the complex we can target regions of the ligand that fit poorly within the active site and postulate chemical modifications that lower the energetic potential by making more negative van der Waals terms, thus improving complementarity with the receptor. In a similar fashion, functional groups on the ligand can be changed in order to augment electrostatic complementarity with the receptor. When a target is selected for the design of new lead compounds three different situations can be faced regarding the amount of information of the system that is available:

1) The structure of the receptor is well known and the bioactive conformation of the ligand is not known,
2) Only the bioactive conformation of the ligand is known and

3) The target structure and the bioactive conformation of the ligand are unknown

The best possible starting point is an X-ray crystal structure of the target site. If the molecular model of the binding site is precise enough, one can apply docking algorithms that simulate the binding of drugs to the respective receptor site, like Autodock.24 In the first step the program creates a negative image of the target site through the use of several atom probes that determine affinity potentials for each atom type in the substrate molecule at different points in a grid, place the putative ligands into the site and finally they evaluate the quality of the fit. The program will try a set of different conformers of the ligand in order to obtain the best disposition of the atoms of the molecule for maximizing the scoring function that quantifies ligand receptor interaction. A different strategy for obtaining new lead compounds through rational drug design is the de novo design of ligands with the use of a builder program, like Ligbuilder. 25 This program also determines the shape and the electrostatic properties of the binding site cavity through the use of several atom probes and then it combines from a library of chemical fragments those that better fill the cavity based on steric and electrostatic complementarity.



- Preexisting target
- Literature target
- Genomics
- Proteomics

Identification
- HTS
- De novo design
- Virtual screening

Optimization
- SAR studies
- In vivo efficacy
- ADMET studies

- Bioavailability studies – ADME
- Long-term toxicity
- In vivo efficacy
- Synthesis scale-up

- Efficacy
- Bioavailability
- Dose
- Safety assessments
- Drug interactions

Time-consuming
- 2-5 yrs discovery; ~13.2 yrs development

Cost-intensive
- $1.1bn (1995-2000) - $1.7bn (2002-2002)[1]
- Total R&D costs for Discovery[2]

Design of Drug candidates: An iterative process

The design of new ligands is carried out as step by step procedure

The state of the art design process is based in large part, on a good understanding of molecular recognition of protein-ligand complexes relying upon analogies to other systems and using advanced computerized molecular design programs.

Steps in structure based drug design

The steps used in structure based drug design for designing new lead compounds are

- Obtaining 3D structure of protein
- Active site identification
- Ligand-receptor fit analysis
- Design of new leads

Beginning the Design Phase
Once the phase of analysis is complete, the design phase can start
One has to identify candidate scaffolds with appropriate substituent's that can ensure enhanced interactions with selected sites of the protein
In the case of the optimization of a known series, the information is used to design new analogs.
Eight golden rules in receptor-based ligand design
The important considerations for receptor-based ligand design can be summarized into the following eight rules:
1. Coordinate to key anchoring sites
2. Exploit hydrophobic interactions
3. Exploit hydrogen bonding capabilities
4. Exploit electrostatic interactions
5. Favor bioactive form & avoid energy strain
6. Optimize vdW Contacts and avoid bumps
7. Structural water molecules and solvation
8. Consider entropic effect

Rule 1: Coordinate to Key Anchoring Sites

- When working with target proteins, first one has to consider the proper anchorage of the ligand to key elements of the catalytic site
- This anchorage not only positions the ligand in the active site but also counteracts the effect of de-solvating the two components when binding occurs. This is very important energetically.

Rule 2: Exploit Hydrophobic Interactions

- With hydrophobic pockets, placing a hydrophobic surface of the ligand in hydrophobic sites of the target protein provides an important driving force in complex formation because it reduces non-polar surface areas exposed to water
- Although individually small, the total contribution of hydrophobic forces to drug-receptor interactions is substantial
- Empirical data suggests that the free energy contribution due to hydrophobic forces is approximately 2.9 kJ/mol per methylene group and 8.4 kJ/mol for a benzene ring
- Unlike hydrogen bonds, the hydrophobic interactions are not directional

Rule 3: Exploit Hydrogen Bonding Capabilities

- Unsatisfied hydrogen bond donors and acceptors are rarely seen in proteins and protein-ligand complexes because this would be highly energetically unfavorable
- A carbonyl oxygen is optimally satisfied when it accepts two different hydrogen bonds with C=O --- H angles close to 120°. However hydrogen bonds to carbonyl oxygen atoms with a C=O --- H angle close to 180° form the basis for β-sheet formation and are quite favorable. The average N-H --- O angle is about 155° (with 90% lying between 140° and 180°).
- Almost all protein groups are capable of forming hydrogen bonds like this. Where groups are not explicitly hydrogen bonded, they are probably solvated.

Rule 4: Exploit Electrostatic Interactions

- The optimization of ligand-protein electrostatics can be achieved by placing a positive charge in close vicinity to an enzyme negative charge

Rule 5: Favor Bioactive Form & Avoid Energy Strain

- Conformational energy calculations are performed on each design idea in order to determine the internal penalty required for the new ligand to attain its bioactive binding conformation inside the protein. The internal energy that is required for the small molecule to reach its binding conformation is energy lost in binding.
- Restricting the conformation space of an inhibitor can be beneficial to binding when the conformation is biased towards the bioactive conformer.

Rule 6: Optimize VDW Contacts and Avoid Bumps

- Attractive van der Waals interactions occur over a short distance range and attraction decreases as 1/r6. As a result, optimization of attractive van der Waals interactions occurs as the shape of the protein binding site and the shape of the ligand match well.
- Calculations of steric fit are difficult because of possible flexing motions of the protein backbone and especially the residue side chains

Rule 7: Structural Water Molecules and Solvation

- Inhibitor design strategies have great potential when they target the displacement of water molecules tightly bound to the protein by incorporating elements of the water molecule within the inhibitor.
- When polar charged groups are considered in the design of a ligand, one should leave some room for other water molecules to solvate the charged center (except possibly when a salt bridge is formed).

Rule 8: Consider Entropic Effect

- A flexible molecule has a better chance of finding an optimal fit into a receptor, but this is achieved at the cost of large conformational entropy
- Sufficient conformational rigidity is essential to ensure that the loss of entropy upon ligand binding is acceptable
- A rigid molecule has little conformational entropy but is unlikely to fit optimally into the receptor

- An analysis of the contributions of various functional groups to protein-ligand binding demonstrates that each freely rotating bond in a ligand reduces binding free energy by about 2.9 kJ/mol
- Making a flexible molecule more rigid will lead to enhanced activity if the right conformation is maintained
- Example of Successful Structure-Based Design
- The use of the crystallographic structure of the HIV-1 protease in drug design represents one of the more impressive success stories in the structure-based drug design field. Structure-based design studies has resulted in the identification of distinct classes of inhibitors and several successful drug candidates have emerged from these studies and are used in the control of AIDS.
- The HIV-1 protease plays a crucial part in the life cycle of the HIV virus. Inhibitor drugs block the action of the protease and the virus perishes because it is unable to mature into its infectious form.
- The HIV-1 protease is a small dimer enzyme comprising two identical folded 99 amino-acid chains A and B

Ligand-Based Computer-Aided Drug Design

The ligand-based computer-aided drug discovery (LBDD) approach involves the analysis of ligands known to interact with a target of interest. These methods use a set of reference structures collected from compounds known to interact with the target of interest and analyse their 2D or 3D structures. The overall goal is to represent these compounds in such a way that the physicochemical properties most important for their desired interactions are retained, whereas extraneous information not relevant to the interactions is discarded. It is considered as an indirect approach to the drug discovery in that it does not necessitate knowledge of the structure of the target of interest. The two fundamental approaches of LBDD are (1) selection of compounds based on chemical similarity to known actives using some similarity measure or (2) the construction of a quantitative structure activity relationship (QSAR) model that predicts biological activity from chemical structure. The methods are applied for in silico screening for novel compounds possessing the biological activity of interest, hit-to-lead and lead-to drug optimization, and also for the optimization of DMPK/ADMET properties. LBDD is based on the similar property principle which states that molecules that are structurally similar are likely to have similar properties. LBDD approaches in contrast to SBDD approaches can also be applied when the structure of the biological target is unknown. Additionally, active compounds identified by ligand based virtual high-throughput screening (LB-vHTS) methods are often more potent than those identified in SB-vHTS.

Molecular Descriptors

Molecular descriptors can include properties such as molecular weight, geometry, volume, surface areas, ring content, rotatable bonds, interatomic distances, bond distances, atom types, planar and nonplanar systems, molecular walk counts, electronegativities, polarizabilities, symmetry, atom distribution, topological charge indices, functional group composition, aromaticity indices, solvation properties, and many others. These descriptors are generated through knowledge-based, graph-theoretical methods, molecular mechanical, or quantum-mechanical tools and are classified according to the Chapter 1 Computer Aided Drug Design: An Overview 16 "dimensionality" of the chemical representation from which they are computed: 1- dimensional (1D), scalar physicochemical properties such as molecular weight; 2D, molecular constitution-derived descriptors; 2.5D, molecular configuration-derived descriptors; 3D, molecular conformation-derived descriptors. These different levels of complexity, however, are overlapping with the more complex descriptors, often incorporating information from the simpler ones.

Molecular Fingerprint and Similarity Searches

Molecular fingerprint-based techniques attempt to represent molecules in such a way as to allow rapid structural comparison in an effort to identify structurally similar molecules or to cluster collections based on structural similarity. These methods are fewer hypotheses driven and less computationally expensive than pharmacophore mapping or QSAR models. They rely entirely on chemical structure and omit compound with known biological activity, making the approach more qualitative in nature than other LBDD approaches. Additionally, fingerprint-based methods consider all parts of the molecule equally and avoid focusing only on parts of a molecule that are thought to be most important for activity. This is less error prone to overfitting and requires smaller datasets to begin with. Fingerprint methods may be used to search databases for compounds similar in structure to a lead query, providing an extended collection of compounds that can be tested for improved activity over the lead. In many situations, 2D similarity searches of databases are performed using chemotype information from first generation hits, leading to modifications that can be evaluated computationally or ordered for in vitro testing.

Quantitative Structure-Activity Relationship Models

Quantitative structure-activity relationship (QSAR) models describe the mathematical relation between structural attributes and target response of a set of chemicals. Classic QSAR is known as the Hansch-Fujita approach and involves the correlation of various electronic, hydrophobic, and steric features with biological activity. In the 1960s, Hansch and others began to establish QSAR models using various molecular descriptors to physical, chemical, and biological properties focused on providing computational estimates for the bioactivity of molecules. In 1964, Free and Wilson developed a mathematical model relating the presence of various chemical substituents to biological activity (each type of chemical group was assigned an activity contribution), and the two methods were later combined to create the Hansch/ Free-Wilson method. A model is then generated to identify the relationship between those descriptors and their experimental activity, maximizing the predictive power. Finally, the model is applied to predict activity for a library of test compounds that were encoded with the same descriptors. Success of QSAR, therefore, depends not only on the quality of the initial set of active/inactive compounds but also on the choice of descriptors and the ability to generate the appropriate mathematical relationship. One of the most important considerations regarding this method is the fact that all models generated will be dependent on the sampling space of the initial set of compounds with known activity, the chemical diversity. In other words, divergent scaffolds or functional groups not represented within this "training" set of compounds will not be represented in the final model, and any potential hits within the library to be screened that contain these groups will likely be missed. Therefore, it is advantageous to cover a wide chemical space within the training set.

3D-QSAR

Comparative field molecular analysis (CoMFA) is a 3D-QSAR technique that aligns molecules and extracts aligned features that can be related to biological activity. This method focuses on the alignment of molecular interaction fields rather than the features of each individual atom. CoMFA was established over 20 years ago as a standard technique for constructing 3D models in the absence of direct structural data of the target. In this method, molecules are aligned based on their 3D structures on a grid and the values of steric (van der Waals interactions) and electrostatic potential energies (Coulombic interactions) are calculated at each grid point. A comparative molecular similarity index (CoMSIA) is an important extension to CoMFA. In CoMSIA, the molecular field includes hydrophobic and hydrogen-bonding terms in addition to the steric and coulombic contributions. Similarity indices are calculated instead of interaction energies by comparing each ligand with a common probe and Gaussian-type functions are used to avoid extreme values.

Multidimensional QSAR

4D and 5D Descriptors Multidimensional QSAR (mQSAR) seeks to quantify all energy contributions of ligand binding including removal of solvent molecules, loss of conformational entropy, and binding pocket adaptation. 4D-QSAR is an extension of 3D-QSAR that treats each molecule as an ensemble of different conformations, orientations, tautomers, stereoisomers, and protonation states. The fourth dimension in 4D-QSAR refers to the ensemble sampling of spatial features of each molecule. A receptor-independent (RI) 4D-QSAR method was proposed by Hopfinger in 1997. This method begins by placing all molecules into a grid and assigning interaction pharmacophore elements to each atom in the molecule (polar, nonpolar, hydrogen bond donor, etc.). Molecular dynamics simulations are used to generate a Boltzmann weighted conformational ensemble of each molecule within the grid. Trial alignments are performed within the grid across the different molecules, and descriptors are defined based on occupancy frequencies within each of these alignments. These descriptors are called grid cell occupancy descriptors. A conformational ensemble of each compound is used to generate the grid cell occupancy descriptors rather than a single conformation. 5D-QSAR has been developed to account for local changes in the binding site that contribute to an induced fit model of ligand binding. In a method developed by Vedani and Dobler, induced fit is simulated by mapping a "mean envelope" for all ligands in a training set on to an "inner envelope" for each individual molecule. Their method involves several protocols for evaluating induced-fit models including a linear scale based on the adaptation of topology, adaptations based on property fields, energy minimization, and lipophilicity potential. By using this information, the energetic cost for adaptation of the ligand to the binding site geometry is calculated. Vedani from the Biographics Laboratory developed a receptor modeling concept, Quasar, based on 6D-QSAR that explicitly allows for the simulation of induced fit. Quasar concept, previously 3,4,5D extended to six dimensions allows for the simultaneous consideration of different solvation models which can be achieved explicitly by mapping parts of the surface area with solvent properties (position and size are optimized by the genetic algorithm).

Pharmacophore Mapping

In 1998, the International Union of Pure and Applied Chemistry (IUPAC) formally defined a pharmacophore as "the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response". In terms of drug activity, it is the spatial arrangement of functional groups that a compound or drug must contain to evoke a desired biological response. Therefore, an effective pharmacophore will contain information about functional groups that interact with the target, as well as information regarding the type of noncovalent interactions and interatomic distances between these functional groups/interactions. A pharmacophore model of the target binding site summarizes steric and electronic features needed for optimal interaction of a ligand with a target. Most common properties that are used to define pharmacophores are hydrogen bond acceptors, hydrogen bond donors, basic groups, acidic groups, partial charge, aliphatic hydrophobic moieties, and aromatic hydrophobic moieties. Pharmacophore features have been used extensively in drug discovery for virtual screening, de novo design, and lead optimization. A pharmacophore model of the target binding site can be used to virtually screen a compound library for putative hits. Apart from querying database for active compounds, pharmacophore models can also be used by de novo design algorithms to guide the design of new compounds. Structure-based pharmacophore methods are developed based on an analysis of the target binding site or based on a target-ligand complex structure. Ligand Scout uses protein-ligand complex data to map interactions between ligand and target. A knowledge based rule set obtained from the PDB is used to automatically detect and classify interactions into hydrogen bonds, charge transfers, and lipophilic regions. The algorithm creates regularly spaced grids around the ligand and the surrounding residues. Probe atoms that represent a hydrogen bond donor, a hydrogen bond acceptor, and a hydrophobic group are used to scan the grids. An empirical scoring function, SCORE, is used to describe the binding constant between probe atoms and the target. SCORE includes terms to account for van der Waals interactions, metal-ligand bonding, hydrogen bonding, and desolvation effects upon binding. A pharmacophore model is developed by rescoring the grids followed by clustering and sorting to extract features essential for protein-ligand interaction. The most common software packages used for ligand based pharmacophore generation include Phase, MOE, Catalyst, DISCO, and GASP.

ADME

- Nine of every ten new drugs fail in clinical testing.
- A drug in phase III testing testing has 32% chance of failure failure.
- Even in Phase I, 37% fail.
- Most drugs fail in phase II.

Absorption - route of drug delivery – Where absorbed

Distribution - where does the drug go, where does it need to go and what are the implications

Metabolism - this will occur and could impact several variables – Could be used to your advantage - Prodrugs

Excretion – how is the drug eliminated

Pharmacokinetics is concerned with the variation in drug concentration with time as a result of absorption, metabolism, distribution and excretion – Drug dose, route of administration, administration, rate and extent of absorption, absorption, distribution distribution rate (particularly to site of action) and rate of elimination – Pharmacokinetics may be simply defined as what the body does to the drug – Pharmacodynamics defined as what the drug does to the body.

Drug Delivery

Oral- by far the most common route. The passage of drug from the gut into the blood is influenced by biologic and physicochemical properties.

Sublingual (buccal) - Certain drugs are best given beneath the tongue or retained in the cheek pouch and are absorbed from these regions into the local circulation.

Rectal -The administration of suppositories is usually reserved reserved for situations situations in which oral administration administration is difficult. This route is more frequently used in small children.

Intravenous injection – Used when a rapid clinical response is necessary, e.g., an acute asthmatic episode. – Achieve relatively precise drug concentrations in the plasma, since bioavailability is not a concern.

Intra-arterial injection – Used in certain special situations, notably with anticancer drugs in an effort to deliver a high concentration of drug to a particular tissue. Typically, the injected artery leads directly to the target organ.

Intrathecal injection – The blood-brain barrier limits the entry of many drugs into cerebrospinal fluid. life-threatening, antibiotics, antifungals and anticancer drugs are given via lumbar puncture and injection into the subarachnoid space.

Intramuscular injection – Drugs may be injected into the arm, thigh or buttocks.

Subcutaneous injection – Some drugs, notably insulin, are routinely administered SC. Drug absorption is generally slower SC than IM, due to poorer vascularity.

Inhalation – Volatile anesthetics, as well as many drugs which affect pulmonary function, are administered as aerosols. Drugs administered via this route are not subject to first-pass liver metabolism.

Topical application – Eye, intravaginal, intranasal, skin. – Alleviation of local symptoms

Drug Absorption and Biological Factors

Membrane structure and function - The cell membrane is a semi- permeable lipoid sieve containing numerous aqueous channels, as well as a variety of specialized carrier molecules.

Passive diffusion is probably the most important absorptive mechanism. Lipid-soluble drugs dissolve in the membrane, and are driven through by a concentration gradient across the membrane.

Carrier-mediated facilitated transport occurs for some drugs, particularly those which are analogs of endogenous compounds for which there already exist specific membrane carrier systems. – For example, methotrexate, an anticancer drug which is structurally similar to folic acid, is actively transported by the folate membrane transport system.

Oral Drug Absorption

The blood supply draining the gut passes through the liver before reaching the systemic circulation. – First-pass effect may reduce the amount of drug reaching the target tissue.

Drug binding – Many drugs will bind strongly to proteins in the blood or to food substances in the gut. – Plasma protein binding will increase the rate of passive absorption by maintaining the concentration gradient of free drug.

Food effects – Absorption can be reduced by the presence of food in the gut – Absorption can be enhanced by food (bile secretion) – Some drugs are irritating and should be administered with meals to reduce adverse effects.

Distribution

Once in the blood, drugs are simultaneously distributed throughout the body and eliminated.– Distribution is much more rapid than elimination, accomplished via the circulation, and influenced by regional blood flow.

Compartments – Central Compartment- The central compartment includes the well-perfused organs and tissues (heart, blood, liver, brain and kidney) with which drug equilibrates rapidly. – Peripheral Compartment(s)- The peripheral compartment(s) include(s) those organs (e.g., adipose and skeletal muscle) which are less well- perfused, and with which drug therefore equilibrates more slowly. – Special compartments -The cerebrospinal fluid (CSF) and central nervous system (CNS) is restricted by the structure of the capillaries and pericapillary glial cells. – Drugs also have relatively poor access to pericardial fluid, bronchial secretions and fluid in the middle ear.

Metabolism

- Phase I and Phase II metabolism – Most products of drug metabolism are less active than the parent compound.
- Metabolites may be responsible for toxic, mutagenic, teratogenic or carcinogenic effects – For example, example, acetaminophen acetaminophen hepatotoxicity hepatotoxicity is due to a minor metabolite which reacts with liver proteins.
- Metabolism of so-called prodrugs, metabolites are actually the active therapeutic compounds – Cyclophosphamide, an inert compound which is metabolized by the liver into a highly active anticancer drug.

Receptor docking

- Computational techniques assist one in searching drug target and in designing drug in silico, but it takes long time and money. In order to design a new drug one need to follow the following path.
- **Identify Target Disease**: One needs to know all about the disease and existing or traditional remedies.
- It is also important to look at very similar afflictions and their known treatments.
- Target identification alone is not sufficient in order to achieve a successful treatment of a disease.

- A real drug needs to be developed. This drug must influence the target protein in such a way that it does not interfere with normal metabolism. One way to achieve this is to block activity of the protein with a small molecule.
- Bioinformatics methods have been developed to virtually screen the target for compounds that bind and inhibit the protein.
- Another possibility is to find other proteins that regulate the activity of the target by binding and forming a complex.

Study Interesting Compounds

- One needs to identify and study the lead compounds that have some activity against a disease. These may be only marginally useful and may have severe side effects.
- These compounds provide a starting point for refinement of the chemical structures.

Detect the Molecular Bases for Disease

- If it is known that a drug must bind to a particular spot on a particular protein or nucleotide then a drug can be tailor made to bind at that site.
- This is often modeled computationally using any of several different techniques. Traditionally, the primary way of determining what compounds would be tested computationally was provided by the researchers' understanding of molecular interactions.
- A second method is the brute force testing of large numbers of compounds from a database of available structures.

Rational drug design techniques

- These techniques attempt to reproduce the researchers' understanding of how to choose likely compounds built into a software package that is capable of modeling a very large number of compounds in an automated way.
- Many different algorithms have been used for this type of testing, many of which were adapted from artificial intelligence applications.

- The complexity of biological systems makes it very difficult to determine the structures of large biomolecules. Ideally experimentally determined (x-ray or NMR) structure is desired, but biomolecules are very difficult to crystallize.

Refinement of compounds

- Once you got a number of lead compounds have been found, computational and laboratory techniques have been very successful in refining the molecular structures to give a greater drug activity and fewer side effects.
- This is done both in the laboratory and computationally by examining the molecular structures to determine which aspects are responsible for both the drug activity and the side effects.

Solubility of Molecule

- One need to check whether the target molecule is water soluble or readily soluble in fatty tissue will affect what part of the body it becomes concentrated in.
- The ability to get a drug to the correct part of the body is an important factor in its potency.
- Ideally there is a continual exchange of information between the researchers doing QSAR studies, synthesis and testing.
- These techniques are frequently used and often very successful since they do not rely on knowing the biological basis of the disease which can be very difficult to determine.

Drug Testing

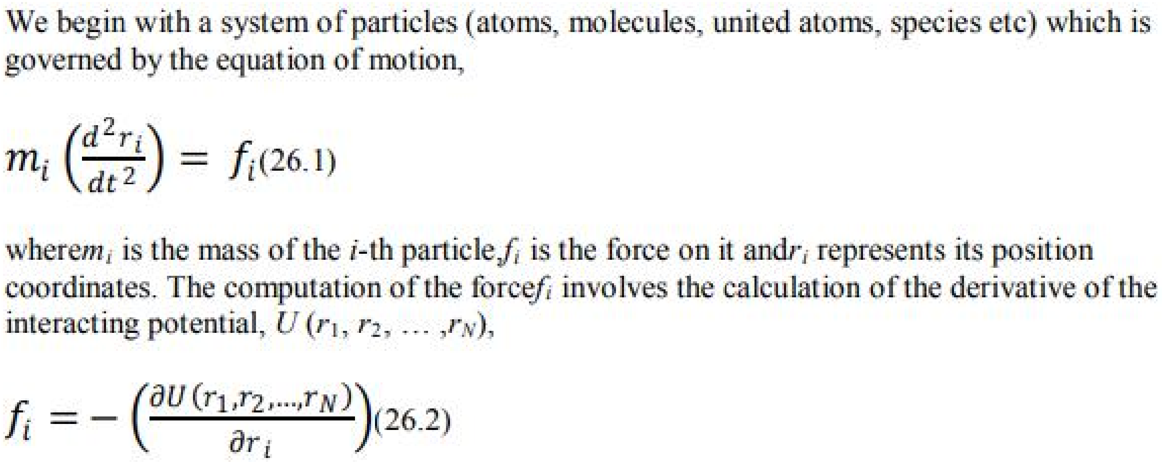
- Once a drug has been shown to be effective by an initial assay technique, much more testing must be done before it can be given to human patients.
- Animal testing is the primary type of testing at this stage.
- Eventually, the compounds, which are deemed suitable at this stage, are sent on to clinical trials.
- In the clinical trials, additional side effects may be found and human dosages are determined.

- Two proteins whose complex was determined by Protein-Protein-Docking. Docking of a small inhibitor to the urease protein.

Introduction to Molecular Dynamics Simulation

With the advent of the computers, chemists, physicists and material scientists had begun (since 1950s) to exploit the power of the computers for probing the properties of materials through simulations. Almost all the materials of the physical world can be probed for their properties by designing appropriate simulation algorithms. These include atomic and molecular systems, biomolecules, complex materials, nuclear materials, life processes and the like. One can construct simulation schemes for the dynamics of the molecules present in the materials; i.e., the Molecular Dynamics (MD), where the constituents of the system are allowed to interact according to known laws of physics, over a period of time. Through the numerical solutions of the equations of motion (often described by the laws of Newtonian mechanics), one obtains the trajectories (position coordinates and/or velocities) of all the constituents of the system, under the influence of the interacting potential. These trajectories are then analyzed in order to extract the desired properties such as pressure, stress, diffusion, viscosity, surface tension, dielectric constant, order parameter, autocorrelation functions, fluctuations, conformational changes etc. Since molecular systems generally consist of a vast number of particles, it is impossible to find the properties of such complex systems analytically. The MD simulation exercises circumvent this problem by using the numerical solutions of the equations of motion. Thus the MD simulation technique presents an interface between laboratory experiments and the theory. This often leads to the realization that ‘computer simulations’ are actually ‘computer experiments’.

The MD technique



We begin with a system of particles (atoms, molecules, united atoms, species etc) which is governed by the equation of motion,

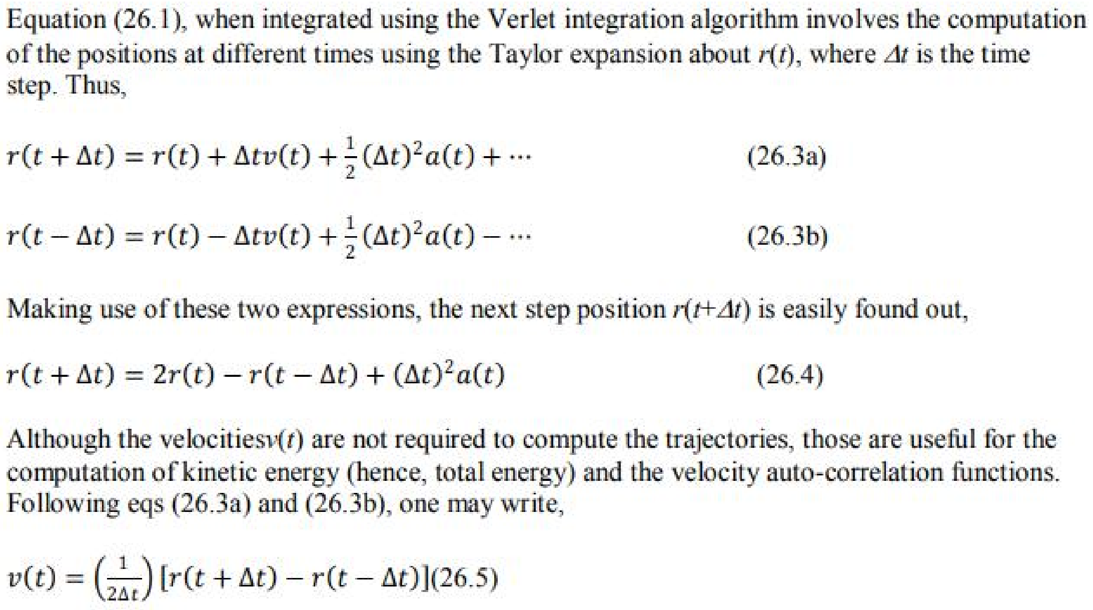
$$m_i\left(\frac{d^2r_i}{dt^2}\right) = f_i(26.1)$$

wherem_i is the mass of the i-th particle,f_i is the force on it andr_i represents its position coordinates. The computation of the forcef_i involves the calculation of the derivative of the interacting potential, $U\,(r_1, r_2, \ldots, r_N)$,

$$f_i = -\left(\frac{\partial U\,(r_1,r_2,\ldots,r_N)}{\partial r_i}\right)(26.2)$$

In each of the time step of the simulation, one needs to compute the force fi and using this force, the position rigets updated. In order to solve the second order differential equation as in eq. (26.1), there are several numerical schemes available. These are based on finite difference methods and the integration algorithms include Gear predictor-corrector algorithm, Verlet algorithm and the Toxvaerd algorithm. The Verlet algorithm and its several variations are the most widely used by the practitioners of the trade and we describe this algorithm below.

The Verlet Algorithm



Equation (26.1), when integrated using the Verlet integration algorithm involves the computation of the positions at different times using the Taylor expansion about *r*(*t*), where *Δt* is the time step. Thus,

$$r(t+\Delta t) = r(t) + \Delta t v(t) + \frac{1}{2}(\Delta t)^2 a(t) + \cdots \qquad (26.3a)$$

$$r(t-\Delta t) = r(t) - \Delta t v(t) + \frac{1}{2}(\Delta t)^2 a(t) - \cdots \qquad (26.3b)$$

Making use of these two expressions, the next step position *r*(*t*+*Δt*) is easily found out,

$$r(t+\Delta t) = 2r(t) - r(t-\Delta t) + (\Delta t)^2 a(t) \qquad (26.4)$$

Although the velocitiesv(*t*) are not required to compute the trajectories, those are useful for the computation of kinetic energy (hence, total energy) and the velocity auto-correlation functions. Following eqs (26.3a) and (26.3b), one may write,

$$v(t) = \left(\frac{1}{2\Delta t}\right)[r(t+\Delta t) - r(t-\Delta t)](26.5)$$

Possible Questions

1. Write a short notes on Molecular docking

2. Write about the ADME properties of a drug

3. Write the Lipinski's rule of five

4. Give short notes on active site prediction

5. Enumerate the steps involved in designing a therapeutic drug

6. Write the applications of bioinformatics in biotechnology and agriculture

7. What is QSAR? Add notes on its importance in novel drug designing

8. Write short notes on receptor docking

9. Briefly discuss about drug designing concepts

10. what are the essential properties required for a potent lead molecules.

11. What are the steps involved in discovering a new drug? Explain the role of in silico methods for drug design.

KAPRAGAM ACADEM YOF HIGHER EDUCATION

DEPARTMENT OF BIOCHEMISTRY
I- M.Sc Biochemistry
18BCP204-Bioinformatics

S. No	Unit V	Questions	Option I	Option II	Option III	Option IV	Answer
1	5	The term Bioinformatics was coined by	J D Watson	Margaret Dayhoff	Pualine Hogeweg	Frederic Sanger	**Margaret Dayhoff**
2	5	Application of Bioinformatics include	Data storage and management	Drug designing	Understand relationships between organisms	All of the above	**All of the above**
3	5	A compound that has desirable properties to become a drug is called	lead	find	fit drug	fit compound	**Lead**
4	5	What is meant by a lead compound in medicinal chemistry	A drug containging the element lead	A leading drug in a particular area of medicine	A compound that acts as the starting point for drug design and development	A drug which is normally the first to be prescribed for a particular ailment	**A compound that acts as the starting point for drug design and development**
5	5	which of the following needs to be established before the search for a lead compound takes place	the pharmacophore	structure activity relationships	a bioassay	patents	**a bioassay**
6	5	What is the term used for the automated in vitro testing of large numbers of compounds using genetically modified cells	robotic testing	high throughput screening	multiscreening	nanotechnology	**high throughput screening**
7	5	Which of the following statements is false with respect to NMR screening to detect drug-target interactions	The procedure relies on small molecules (drugs) having shorter relaxation times than large molecules (targets)	The procedure can be used on mixtures of compounds	the method can detect weak binding	the method can identify small molecules binding to different regions of the same binding site	**the method can detect weak binding**
8	5	There are several sources and methods of discovering new compounds. Which of the following is most likely to lead to the discovery of a complex structure quite unlike any other previously discovered	Combinatorial chemistry	database mining	screening plant extracts	me too drugs	**Screening plant extracts**
9	5	What is the term used for drugs that are similar in structure to a known drug and which are used for the same purpose	copycat drugs	me too drugs	derivative drugs	analogue drugs	**me too drugs**
10	5	What is the term used to small molecules that bind to different regions of a binding site	epimers	isomers	isotopes	epitopes	**isotopes**
11	5	Identify the kind of interactions that are typically involved in binding a drug to the binding site of protein	predominantly van der waals interactions	perdominantly ionic bonds	predominantly hydrogen bonds	a combination of all of the above	**a combination of all of the above**
12	5	Identify which of the following amino acids has a side chain that may be important in binding a drug by ionic bonding	aspartate	glycine	serine	valine	**aspartate**
13	5	Which of the following statements best describes pharmacodynamics	The study of how drugs reach their target in the body and how the levels of a drug in the blood are affected by absorption, distribution, metabolism and excretion	The study of how drugs can be designed using molecular modelling based on a drug's pharmacophore	the study of how a drug interacts with its target binding site at the molecular level	the study of which functional groups are important in binding a drug to its target binding site and the identification of a pharmacophore	**the study of how a drug interacts with its target binding site at the molecular level**
14	5	Which of the following statements best describes pharmockinetics	The study of how drugs reach their target in the body and how the levels of a drug in the blood are affected by various factos	The study of how drug can be designed using molecular, modelling based on a drug's pharmacophore	The study of how a drug interacts with its target binding site at the molecular level	The study of which functional groups are important in binding a drug to its target binding site and the identification of a pharmacophore	**The study of how drugs reach their target in the body and how the levels of a drug in the blood are affected by various factos**
15	5	Which of the following characteristics is detrimental to oral activity	stability to digestive enzymes	susceptibility to metabolic enzymes	stability to stomach acids	solubility in both aqueous and fatty environments	**susceptibility to metabolic enzymes**
16	5	Which of the following isone of the rules in Lipinski's rule of five	a molecular weight equal to 500	no more than five hydrogen bond acceptor groups	no more than 10 hydrogen bond donor groups	a calculated logP value less than +5	**a calculated logP value less than +5**
17	5	Which of the following statements is the closest description of Phase I metabolism	Reactions which add a polar molecules to a functional group already present on a drug or one of its metabolites	Reactions which occur in the blood supply	Reactions which add a polar functional group to a drug	Reactions which occur in the gut wall	**Reactions which add a polar functional group to a drug**
18	5	Which of the following statements is the closest description of Phase II metabolism	Reactions which add a polar molecules to a functional group already present on a drug or one of its metabolites	Reactions which occur in the blood supply	Reactions which add a polar functional group to a drug	Reactions which occur in the gut wall	**Reactions which add a polar molecules to a functional group already present on a drug or one of its metabolites**
19	5	Which of the following statements is not true about cytochrome P450 enzymes	they contain haem and magnesium	they belong to a general class of enzymes called monooxygenases	there are over 30 different cytochrome P450 enzymes	variation in cytochrome P450 enzyme profile between individuals can explain individual variation in drug susceptibility	**they contain haem and magnesium**
20	5	Which of the following groups is least susceptible to cytochrome P450 enzymes	terminal methyl groups	allylic carbons	benzylic carbon atoms	quaternary carbon atoms	**quaternary carbon atoms**
21	5	Alkenes and aromatic groups can be metabolised to diols. Which enzymes are involved	cytochrome P450 enzymes	epoxide hydrolase	both of the above	neither of the above	**both of the above**
22	5	Which of the following enzymes is not involved in catalyzing a phase I metabolic raction	flavin containing monooxygenases	monoamine oxidases	glucuronyltransferase	esterases	**glucuronyltransferase**
23	5	Which of the following reactions is not a Phase I metabolic transformation	reduction of ketones	conjugation to alcohols	oxidation of alkyl groups	ester hydrolysis	**conjugation to alcohols**
24	5	Which of the following terms refers to the molecular modelling computational method that uses equations obeying the laws of classical physics	Qunatum mechanics	Molecular calculations	Molecular mechanics	Quantum thoery	**Molecular mechanics**
25	5	Which of the following terms refers to the molecular modelling computational method that uses quantum physiscs	Quantum mechanics	Molecular calculations	Molecular mechanics	Quantum theory	**Quantum mechanics**
26	5	Which of the following needs to be known before two drugs can be overlaid to compare their structure	The pharmacophore of each drug	the active conformation of each drug	Both of the above	Neither of the above	**Both of the above**
27	5	Which of the following statements is true	The most stable conformation of a drug is also the active conformation	the active conformation is the most reactive conformation of structure	The active conformation is the conformation adopted by a drug when it binds to its target binding site	The active conformation can be determined by conformational analysis	**The active conformation is the conformation adopted by a drug when it binds to its target binding site**

28	5	What is meant by docking	The process by which two different structures are compared by molecular modelling	The process by which a lead compound is simplified by removing excess functional groups	The process by which drugs are fitted into their target binding sites using molecular modelling	The process by which a pharmacophore is identified	**The process by which drugs are fitted into their target binding sites using molecular modelling**
29	5	Which of the following statements is true in de novo drug desing	The design of rigid molecules is superior to flexible ones	Molecules should be designed to fit as snugly as possible into the target binding site	Molecules that have to adopt an unstable conformation in order to bind should be rejected	Desolvation energies can be ignored since they are likely to be the same for different molecules having the same pharmacophore	**Molecules that have to adopt an unstable conformation in order to bind should be rejected**
30	5	Which of the following software programmes is used for automated de novo drug desing	DOCK	LUDI	CHEM3D	CoMFA	**DOCK**
31	5	What is meant de novo drug desing	The deisng of rigid molecules is superior to flexible ones	Molecules that have to adopt an unstable conformation in order to bind should be rejected	Molecules should be designed to fit as snugly as possible into the target binding site	Desolvation energies can be ignored since they are likely to be the same for different molecules having the same pharmacophore	**Molecules that have to adopt an unstable conformation in order to bind should be rejected**
32	5	CADD stands for	Computer Aided Drug Design	Computer Asisted Drug Design	Computer Aided Drug Discovery	Computer Asisted Drug Discovery	**Computer Aided Drug Design**
33	5	Identification of lead molecules based on the receptor features is	Pharmacophore drug design	QSAR	Structure based drug design	Homology modeling	**Structure based drug design**
34	5	Identificationof leadmolecule based on the charge propensity of the receptor binding site is	Pharmacophore drug design	QSAR	Structure based drug design	de novo drug designing	**de novo drug designing**
35	5	Nuclear magnetic Resonance for protein structure in published in	1980	1986	1985	1970	**1980**
36	5	Lead molecules screening from the database based of pharmocoporic features is	Pharmacophore drug design	de novo drug design	Structure based drug design	Homology modeling	**Pharmacophore drug design**
37	5	What does the symbol P represent in a QSAR equation	pH	Plasma concentration	partition coefficient	prodrg	**partition coefficient**
38	5	What is the symbol πin a QSAR equation	The hydrophobicity of the molecule	The electronic effect of a substituent	The substituent hydrophobicity constant	A measure of the steric properties for a substituent	**The substituent hydrophobicity constant**
39	5	What does MR represent in a QSAR equation	Molar refractivity is a steric factor	Molar refractivity is an electronic factor	Molar refractivity is a hydrophobic factor	Molar refractivity is a stereoelectronic factor	**Molar refractivity is a steric factor**
40	5	What does a negative value of σ signify for a substituent	It a eletron donating	It is electron withdrawing	It is neutral	It is hydrophobic	**It a eletron donating**
41	5	Which of the following statement is unture when comparing 3D QSAR with conventional QSAR	only drugs of the same structural class should be studied by 3D QSAR or QSAR	3D QSAR has a predictive quality unlike QSAR	Experimental parameters are not required by 3D QSAR, but are for QSAR	Results can be shown graphically in 3D QSAR but not with QSAR	**only drugs of the same structural class should be studied by 3D QSAR or QSAR**
42	5	What value does the regression coefficient have for a perfect fit	0.1	1	10	100	**1**
43	5	Which of the following statements is true	Drugs and drug targets generally have similar molecular weights	Drugs are generally smaller than drug targets	Drugs are generally larger than drug targets	There is not general rule regarding the relative size of drugs and their targets	**Drugs are generally smaller than drug targets**
44	5	What is meant by a binding site	The area of a macromolecular target that is occupied by a drug when it binds	The portion of the drug to which a drug target binds	The functional groups used by a drug in binding to a drug target	The bonds involved in binding a drug to its target	**The area of a macromolecular target that is occupied by a drug when it binds**
45	5	Which of the following binding interactions is likely to be the most important initial interaction when a drug enters a binding site	van der waals interactions	hydrogen bond	ionic	induced dipole-dipole interactions	**ionic**
46	5	Which of the following functional groups is most likely to participate in a dipole-dipole interaction	Aromatic ring	Ketone	Alcohol	Alkene	**Ketone**
47	5	Which of the following statements is not true about receptors	Most receptors are protein situated in the cell membrane	Receptors contain a hollow or cleft on their surface which is know as a binding site	Receptors bind chemical messenger such as neurotransmitters or hormones	Receptors catalyse reactions on chemical messengers	**Receptors catalyse reactions on chemical messengers**
48	5	Which of the following statements is not true regarding the binding site or a receptor	The binding site is normally a hollow or cleft in the surface of a receptor	The binding site is normally hydrophobic in nature	Chemical messengers fit into binding sites and bind to functional groups within the binding site	The binding site contains amino acids which are important to the binding process and a catalytic mechanism	**The binding site contains amino acids which are important to the binding process and a catalytic mechanism**
49	5	Which of the following statements is true regarding the DNA binding region of intracellular receptors	It contains five cysteine residues	Four cysteine residues are involved in binding two zinc ions	It identifies particular nucleotide sequences in DNA	The DNA binding region is known as having "thiol fingers"	**It identifies particular nucleotide sequences in DNA**
50	5	The interactions of ligands with proteins	usually result in the inactivation of the proteins	are relatively rare in biological systems	are usually irreversible	are usually transient	**are usually transient**
51	5	Which of the following statements about protein-ligand binding is correct	the Ka is equal to the concentration of ligand when all of the binding sites are occupied	the ka is independent of such conditions as salt concentration and pH	the larger the ka (association constant), the weaker the affinity	the larger the ka the faster is the binding	**the larger the ka the faster is the binding**
52	5	All allosteric interaction between a ligand and a protein is one in which	binding of a molecular to a binding site affects binding of additional molecules to the same site	binding of a molecule to a binding site affects binding properties of another site on the protein	binding of the ligand to the protein is covalent	multiple molecules of the same ligand can bind to the same binding site	**binding of a molecule to a binding site affects binding properties of another site on the protein**
53	5	An individual molecular structure within an antigen to which an individual antibody binds is as an	antigen	epitope	Fab region	Fc region	**epitope**
54	5	Which of the following parts of the IgG molecule are not involved in binding to an antigen	Fab	Fc	Heavy chain	Light chain	**Fc**
55	5	A prosthetic group of a protein is non-protein structure that is	a ligand of the protein	a part of the secondary structure of the protein	a substrate of the protein	permanently associated with the protein	**permanently associated with the protein**
56	5	Which of the following is not correct concerning cooperative binding of a ligand to a protein	It is usually a form of allosteric interaction	It is usually associated with proteins with multiple subunits	It rarely occurs in enzymes	It results in a nonlinear Hill Plot	**It rarely occurs in enzymes**
57	5	Which statement about the process of drug discovery is true	It only encompasses the non-clinicallaboratory and animal testing	It is the process which ascertains the effectiveness and safety of potentialdrug condidates	It is the process by which therapeutic compounds are formulated into medicines	It ensures there are no side-effects associated with the potential drug candidates	**It is the process which ascertains the effectiveness and safety of potentialdrug condidates**

58	5	What are adverse drug reactions (ADRs)	The synergistic effects that are seen when some drugs are administered concurrently	Responses to increased drug doses required to achieve the same physiological outcome	Unintended alternative physiological responses caused by the drug that cause harm to the patient	Harmful chemical interactions between two drugs that are used to treat the same clinical symptoms	**Unintended alternative physiological responses caused by the drug that cause harm to the patient**
59	5	In pharmacokinetics what does the acronym ADME stand for	Absorption, Distribution, Metabolism and Excretion	Administration, Differentation,Metabolism and Excretion	Absorption, Disintegration, Metabolism and Efficacy	Administration, Distribution,Metabolism and Efficcacy	**Absorption, Distribution, Metabolism and Excretion**
60	5	Which of the following is not a type of cellular receptor	Tyrosine kinas receptor	G-protein coupled receptor	Endocrine receptors	Intracellular/nuclear receptor	**Endocrine receptors**