M.Sc., Biochemistry **2018-2020**

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. EBRINDHA
SUBJECT NAME: BIOINFORMATICS
SUB.CODE:18BCP414A
SEMESTER: IV CLASS: II B.Sc (BC)

S.NO	NAME OF THE EXPERIMENT	SUPPORT MATERIALS
1.	**Biologi al databases and data r trieval** StructureSequence downloadretrieval(protein(proteinandandgene)DNA)fromfromNCBIPDB GCG,CLUSTAL,MolecularfileSwissformats-Prot,PIR-FASTA, GenBank, Genpept, Molecular viewer by visualization software	W1 W2 T1 (29-34) T2 (201-211, 195-196)
2.	BLAST**Sequence**suite**alignment**oftoolsfor pairwise alignment Multiple sequence alignment using CLUSTALW	W3W4
3.	**GeneratingPhylogetic**phylogenetic**analysis** tree using PHYLIP	W5
4.	Primary**otein** sequence**tructure**analyses**prediction**(Protparam)**andanalysis** TertiarySecondarystructurestructurepredictionprediction(SWISSMODEL)(GOR,nnPredict, SOPMA)	W6 W7-W9 W10
5.	Protein structure evaluation-Ramachandran map (PROCHECK)	W11 W12-W13
	Gene prediction**structure** using**prediction**GENSCAN**and** AND**analysis**GLIMMER	

REFERENCES	
T1	Westhead, D.R., Parish J.H., Twyman, R.M. , (2003). Bioinformatics, Published by Vinod Vasishtha for Viva Books Private Limited, New Delhi-110 002.
T2	Mani K. Vijayaraj N, (2002). Bioinformatics for Beginers, Kalaikathir Achchagam, Coimbatore.
W1	https://www.ncbi.nlm.nih.gov/
W2	https://www.rcsb.org/pdb/home/home.do
W3	https://blast.ncbi.nlm.nih.gov/Blast.cgi
W4	https://www.ebi.ac.uk/Tools/msa/clustalw2/
W5	http://evolution.genetics.washington.edu/phylip.html
W6	https://web.expasy.org/protparam/
W7	https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_gor4.html
W8	http://130.88.97.239/bioactivity/nnpredictfrm.html
W9	https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_sopma.html
W10	https://swissmodel.expasy.org/
W11	https://services.mbi.ucla.edu/SAVES/
W12	http://genes.mit.edu/GENSCAN.html
W13	http://www.cs.jhu.edu/~genomics/Glimmer/

UNIT-

ISYLLABUS

Definition, concepts of Bioinformatics: Objectives, 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, andacademia.

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.

Anotherproblemrelatestowhatwedowiththedata.Scientificdiscoveryisdriven byfalsifiabilityandimaginationandnotbypurelylogicalprocessesthatturnobservations intounderstanding.Datawillnotgenerateknowledgeifweuseinductiveprinciples.

Thegathering,archival,dissemination,modeling,andanalysisofbiologicaldatafalls within a relatively young field of scientific inquiry, currently known as 'bioinformatics', 'Bioinformatics was spurred by wide accessibility of computers with increased compute powerandbytheadventof*genomic*s.Genomicsmadeitpossibletoacquirenucleicacid 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,sequenceswerematchedtothosecodingforglobularproteinsofknownstructure (definedbycrystallography)andwereusedinhigh-throughputcombinatorialapproaches (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 informationderivedfromtheseendeavorshasgiventhefalseimpressionthatimagination 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 withinthecontextofthecomplexandeverchanginglivingworld.

Thesciencethatrelatestobioinformaticshasmanycomponents.Itusuallyrelatesto 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 biotechnologyandgenomics,bioinformaticsismovingfromappliedtobasicscience,from developingtoolstodevelopinghypotheses.

Definition

- Bioinformatics is the recording, annotation, storage, analysis, and searching/retrieval of nucleic acid sequence (genes and RNAs), protein sequence andstructuralinformation.Thisincludesdatabasesofthesequencesandstructural information as well methods to access, search, visualize and retrieve the information.
- Sequencedatacanbeusedtomakepredictionsofthefunctionsofnewlyidentified genes.
- Estimate evolutionary distance in phylogeny reconstruction, determine the active sites of enzymes, construct novel mutations and characterize alleles of genetic diseasestonamejustafewuses.
- Bioinformatics is the field of science in which biology, computer science, and informationtechnologymergeintoasingle 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 bediscerned.

There are three important sub-disciplines within bioinformatics involving computationalbiology:

- The development of new algorithms and statistics with which to assess relationshipsamongmembersoflargedatasets;
- The analysis and interpretation of various types of data including nucleotide and aminoacidsequences,proteindomains,andproteinstructures;and
- The development and implementation of tools that enable efficient access and managementofdifferenttypesofinformation.

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 Bioinformaticshistory.
- G. Mendel is known as the "Father of Genetics". He did experiment on the cross-fertilizationofdifferentcolorsofthesamespecies.
- Mendelillustratedthattheinheritanceoftraitscouldbemoreeasilyexplainedifit wascontrolledbyfactorspasseddownfromgeneration togeneration.
- The understanding of genetics has advanced remarkably in the last thirty years.In 1972,PaulbergmadethefirstrecombinantDNAmoleculeusingligase.
- InthatsameyearStanleyCohen,AnnieChangandHerbertBoyerproducedthefirst recombinant DNAorganism.
- In1973,twoimportantthingshappenedinthefieldofgenomics.
- 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 wasborn.

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 proteincrystallography.
- 1955-Thesequenceofthefirstproteintobeanalyzed,bovineinsulinisannounced by F.Sanger.
- 1969-TheARPANETiscreatedbylinkingcomputersatStandfordandUCLA.
- 1970 - The details of the Needleman-Wunsch algorithm for sequence comparison arepublished.
- 1972-ThefirstrecombinantDNAmoleculeiscreatedbyPaulBergandhisgroup.
- 1973 - The Brookhaven Protein DataBank is announced. Robert Metcalfe receives hisPh.DfromHarvardUniversity.HisthesisdescribesEthernet.
- 1974 - Vint Cerf and Robert Khan develop the concept of connecting networksof computersintoan"internet"anddeveloptheTransmissionControlProtocol(TCP).
- 1975 - Microsoft Corporation is founded by Bill Gates and Paul Allen. Two-dimensionalelectrophoresis,whereseparationofproteinsonSDSpolyacrylamide 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 theUS.
- 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, cloneconstruction,mappingandprimerdesign.
- 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 diseasegenesandtheirrelatedpathways.

Major events in Computational Methods and Computational Biology

- 1993 - CuraGen Corporation is formed in New Haven, CT. Affymetrix begins independentoperationsinSantaClara,California.
- 1994 - Netscape Communications Corporation founded and releases Navigator, the commercial version of NCSA's Mozilla. Gene Logic is formed in Maryland. The PRINTSdatabaseofproteinmotifsispublishedbyAttwoodandBeck.
- 1995 - *The Haemophilus* influenza genome (1.8) is sequenced. The *Mycoplasma genitalium* genome issequenced.
- 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 DNAchips.
- 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 EuropeanBioinformaticsInstitute(EBI),theGermanCancerResearchCenter

(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 PEBiosystems.
- 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. MolecularSimulationsInc.isacquiredbyPharmacopeia.
- 1999 - deCode genetics maps the gene linked to pre-eclampsia as a locus on chromosome2p13.
- 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.PharmacopeiaacquiresOxfordMolecularGroup.
- 2001-Thehumangenome(3,000Mbp)ispublished.

Genome sequencing

- 1990s did advances in sequencing technology make it feasible to sequence the entiregenomeofanythingmorecomplexthanabacterium.

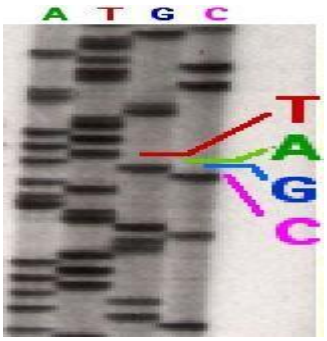
- DNA sequencing includes several methods and technologies that are used for determining the order of the nucleotide base adenine, guanine, cytosine, and thymineinamoleculeofDNA.
- Knowledge of DNA sequences has become indispensable for basic biological research, other research branches utilizing DNA 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.
- TherapidspeedofsequencingattainedwithmodernDNAsequencingtechnology has been instrumental in the sequencing of the human genome, in the Human GenomeProject.
- Related projects, often by scientific collaboration across continents, have generated thecompleteDNAsequencesofmanyanimal,plant,andmicrobialgenomes.
- ThefirstDNAsequenceswereobtainedintheearly1970sbyacademicresearchers usinglaboriousmethodsbasedontwo-dimensionalchromatography.Followingthe developmentofdye-basedsequencingmethodswithautomatedanalysis,
- DNAsequencinghasbecomeeasierandordersofmagnitudefaster.

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.
- AlthoughMaxamandGilbertpublishedtheirchemicalsequencingmethodtwoyears afterthe ground-breakingpaperofSangerandCoulsononplus-minussequencing.
- Maxam–Gilbertsequencingrapidlybecamemorepopular,sincepurifiedDNAcould be used directly, while the initial Sanger method required that each read start be clonedforproductionofsingle-strandedDNA.

- However,withtheimprovementofthechain-terminationmethod,Maxam-Gilbert sequencinghasfallenoutoffavorduetoitstechnicalcomplexityprohibitingitsuse in standard molecular biology kits, extensive use of hazardous chemicals, and difficulties withscale-up.
- The method requires radioactive labeling at one 5' end of the DNA and purification oftheDNAfragmenttobesequenced.
- Chemical treatment generates breaks at a small proportion of one or two of the four nucleotidebasesineachoffourreactions(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)aremethylatedusinghydrazine.
- The addition of salt (sodium chloride) to the hydrazine reaction inhibits the methylationofthyminefortheC-onlyreaction.
- 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 introduceonaverageonemodificationperDNAmolecule.Thusaseriesoflabeled 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 acrylamidegelsforsizeseparation.
- To visualize the fragments, the gel is exposed to X-ray film for autoradiography, yieldingaseriesofdarkbandseachcorrespondingtoaradiolabeledDNAfragment, fromwhichthesequencemaybeinferred.
- Chemical sequencing", this method led to the Methylation Interference Assay used tomapDNA-bindingsitesforDNA-bindingproteins.

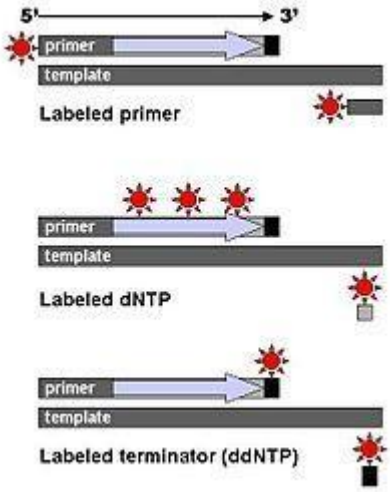
Chain-termination methods



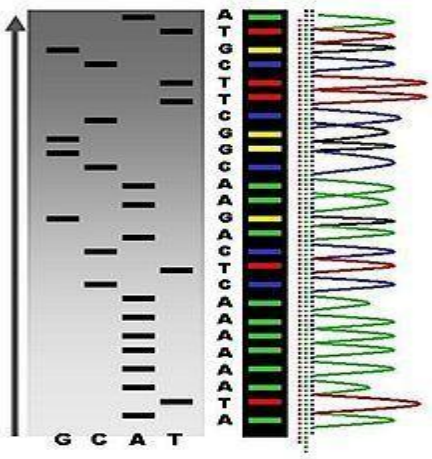
- Because the chain-terminator method is more efficient and uses fewer toxic chemicals and lower amounts of radioactivity than the method of Maxam and Gilbert,itrapidlybecamethemethodofchoice.
- The key principle of the Sanger method was the use of dideoxynucleotide triphosphates(ddNTPs)asDNAchainterminators.
- Theclassicalchain-terminationmethodrequiresasingle-strandedDNAtemplate,a DNAprimer,aDNApolymerase,normaldeoxynucleotidetriphosphates(dNTPs), andmodifiednucleotides(dideoxyNTPs)thatterminateDNAstrandelongation.
- These ddNTPs will also be radioactively or fluorescently labelled for detection in automated sequencingmachines.
- 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 varyinglength.
- 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 correspondtoDNAfragmentsofdifferentlengths.
- 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 usedtoread(frombottomtotop)theDNAsequence.



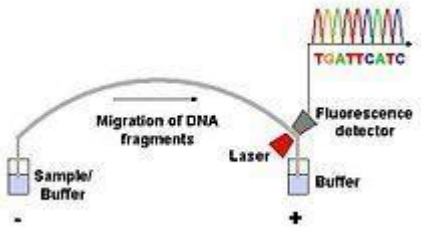
- DNA fragments are labeled with a radioactive or fluorescent tag on the primer, in thenewDNAstrandwithalabeleddNTP,orwithalabelecdddNTP.
- Technical variations of chain-termination sequencing include tagging with nucleotidescontainingradioactivephosphorusforradiolabeling,orusingaprimer labeledatthe5'endwithafluorescentdye.
- Dye-primersequencingfacilitatesreadinginanopticalsystemforfasterandmore economical analysis andautomation.
- The later development by Leroy Hood and coworkers of fluorescently labeled ddNTPsandprimerssetthestageforautomated,high-throughputDNAsequencing.



Sequence ladder by radioactive sequencing compared to fluorescent peaks

- Chain-terminationmethodshavegreatlysimplifiedDNAsequencing.
- Forexample,chain-termination-basedkitsarecommerciallyavailablethatcontain thereagentsneededforsequencing,pre-aliquotandreadytouse.
- Limitationsincludenon-specificbindingoftheprimertotheDNA,affectingaccurate read-out of the DNA sequence, and DNA secondary structures affecting the fidelity of thesequence.

Dye-terminator sequencing



Capillary electrophoresis

- Dye-terminatorsequencingutilizeslabelingofthechainterminatorddNTPs,which permitssequencinginasinglereaction,ratherthanfourreactionsasinthelabeled-primermethod.
- In dye-terminator sequencing, each of the four dideoxynucleotide chain terminators is labeled with fluorescent dyes, each of which emits light at differentwavelengths.

- Owing to its greater expediency and speed, dye-terminator sequencing is now the mainstay in automatedsequencing.
- Itslimitationsincludedyeeffectsduetodifferencesintheincorporationofthedye-labeledchainterminatorsintotheDNAfragment,resultinginunequalpeakheights and shapes in the electronic DNA sequence trace chromatogram after capillary electrophoresis
- ThisproblemhasbeenaddressedwiththeuseofmodifiedDNApolymeraseenzyme systems and dyes that minimize incorporation variability, as well as methods for eliminating "dyeblobs".
- The dye-terminator sequencing method, along with automated high-throughput DNA sequence analyzers, is now being used for the vast majority of sequencing projects.
- Veryfewgenesinacellareactuallyactiveinproteinproductionatanygiventime.
- Different sections of DNA may be dormant or active over the life of a cell, their expressiontriggeredbyothergenesandchangesinthecell'sinternalenvironment.
- 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 inmicrobiology.
- Chromosomes, which range in size from 50 million to 250 million bases, must first bebrokenintomuchshorterpieces(*subcloningstep*).
- 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 (*templatepreparationandsequencingreactionsteps*).
- Thefragmentsinasetareseparatedbygelelectrophoresis(*separationstep*).
- Newfluorescentdyesallowseparationofallfourfragmentsinasinglelaneonthe 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 firststep.
- Automatedsequencersanalyzetheresultingelectropherograms,andtheoutputisa four-colorchromatogramshowingpeaksthatrepresenteachofthefourDNAbases.
- Afterthebasesare"read,"computersareusedtoassembletheshortsequencesinto long continuous stretches that are analyzed for errors, gene-coding regions, and othercharacteristics.
- Finished sequences are submitted to major public sequence databases, such as GenBank.HumanGenomeProjectsequencedataarethusfreelyavailabletoanyone around theworld
- 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 geneticmaterialnecessaryforfullgenomesequencing.
- 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-containingcells.
- 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 storagecapacity.
- Fullgenomesequencingwouldhavebeennearlyimpossiblebeforetheadventofthe microprocessor,computers,andtheInformationAge.

- Unlikefullgenomesequencing,DNAprofilingonlydeterminesthelikelihoodthat 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.
- Almostalltrulycompletegenomesareofmicrobes;theterm"fullgenome"isthus sometimes used loosely to mean "greater than 95%". The remainder of thisarticle focusesonnearlycompletehumangenomes.
- Ingeneral,knowingthecompleteDNAsequenceofanindividual'sgenomedoesnot, onitsown,provideusefulclinicalinformation,butthismaychangeovertimeasa large number of scientific studies continue to be published detailing clear associationsbetweenspecificgeneticvariantsanddisease.
- 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-JinKim
- Therearecurrentlyover60nearlycompletehumangenomespubliclyavailable.
- Sequencing of nearly an entire human genome was first accomplished in 2000 partlythroughtheuseofshotgunsequencingtechnology.
- The Institute for Genomic Research (TIGR) to sequence the entire genome of the bacterium *Haemophilus influenzae* in 1995, and then by Celera Genomics to sequencetheentirefruitflygenomein2000.

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 comprisingtheDNAofaparticularplantoranimal.

- Since chromosomes consist of long, unbroken strands of DNA, a very convenient waytosequenceagenomewouldbetounraveleachchromosomeandreadoffthe basepairslikepunchtape.
- 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 fragmentsoverlap.
- VariousalignmentalgorithmstolookforoverlapsbetweenshortDNAfragments.

Finding the Genes

- Once a reliable DNA sequence has been established, there still remains the task of findingtheactualgenes(codingregions)embeddedwithintheDNAstrand.
- LargeproportionoftheDNAinagenomeisnon-coding.
- Findingthecodingregionsisanimportantstepingenomeanalysis,butitisnotthe end of theroad.
- Emergent behaviour is very hard to simulate, because there is no way to infer the simplerulesfromthecomplexbehaviourStill,computersgiveusawaytotrymany differentrulesetsandtesthypothesesaboutgeneinteraction.

Unexplored Territory

- Onlyaveryfeworganismshavehadtheirgenomefullysequenced
- Manytechnicalandcomputationalchallengesremainbeforesequencingbecomesan automaticprocess
- Some species are still very difficult for us to sequence, and much remains to be learnedabouttheroleandoriginofthatentirenon-codingDNA.

Human Genome Projects

Goals:

- Identifyalltheapproximate30,000genesinhumanDNA
- Determinethesequencesofthe3billionbasepairsthatmakeuphumanDNA

- Sequence the genomes of other model organisms including *Escherichia coli*, yeast (*Saccharomyces cerevisiae*), the fruit fly *Drosophila melanogaster*, the worm *Caenorhabditiselegans*andthelaboratorymouse,
- Storethisinformationindatabases
- Improvetoolsfordataanalysis
- Transferrelatedtechnologiestotheprivatesectorand
- Addresstheethical,legalandsocialissues(ELSI)thatmayarisefromtheproject.

Milestones:

- 1990:projectinitiatedasjointeffortofUSDepartmentofEnergyandtheNational Institutes ofHealth
- June2000:Completionofaworkingdraftoftheentirehumangenome
- February2001:Analysesoftheworkingdraftarepublished
- April 2003: HGP sequencing is completed and project is declared finished two years ahead ofschedule.

Whatdoesthedrafthumangenomesequencetellus? By thenumbers

- The human genome contains 3 billion chemical nucleotide bases (A, C, T, andG).
- The average gene consists of 3000 bases, but sizes vary greatly, with the largest knownhumangenebeingdystrophinat2.4millionbases.
- Thetotalnumberofgenesisestimatedataround20,000-25,000muchlowerthan previousestimatesofapproximately100,000.
- Almostall(99.9%)nucleotidebasesareexactlythesameinallpeople.
- Thefunctionsareunknownforover50%ofdiscoveredgenes.

How it's arranged

- The human genome's gene-dense "urban centers" are predominantly composed of theDNAbuildingblocksGandC.

- Incontrast,thegene-poor"deserts"arerichintheDNAbuildingblocksAandT.The GC and AT rich regions usually can be seen through a microscope as light and dark bands onchromosomes.
- Genes appear to be concentrated in random areas along the genome, with vast expansesofnon-codingDNAbetween.
- 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 CpGislandsarebelievedtohelpregulatedgeneactivity.
- Chromosome 1 has the most genes (2968), and the Y chromosome has thefewest (231).

Future Challenges:

What we still don't know

- Genenumber,exactlocations,andfunctions
- Generegulation
- DNA sequenceorganization
- Chromosomal structure andorganization
- Non-codingDNAtypes,amount,distribution,informationcontentandfunctions
- Coordinationofgeneexpression,proteinsynthesis,andpost-translationalevents
- Interactionofproteinsincomplexmolecularmachines
- Predictedvsexperimentallydeterminedgenefunction
- Evolutionary conservation amongorganisms
- Proteinconservation(Structureandfunction)
- Proteomes(totalproteincontentandfunction)inorganisms
- Correlation of SNPs (single-base DNA variations among individuals) with health anddisease
- Diseasesusceptibilitypredictionbasedongenesequencevariation
- Genesinvolvedincomplextraitsandmultigenediseases

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

Anticipated Benefits of Genome Research

- Molecularmedicine
- Improve diagnosis ofdisease
- Detectgeneticpredispositionstodisease
- Createdrugsbasedonmolecularinformation
- Usegenetherapyandcontrolsystemsasdrugs
- Design‘‘customdrugs’’basedonindividualgeneticprofiles

Microbial genomics

- Rapidlydetectandtreatpathogens(diseasecausingmicrobes)inclinicalpractice
- Developnewenergysources(biofuels)
- Monitorenvironmentstodetectpollutants
- Protectcitizenryfrombiologicalandchemicalwarface
- Cleanuptoxicwastesafelyandefficiently

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,EvolutionandHumanmigration
- Studyevolutionthroughgermlinemutationsinlineages
- Studymigrationofdifferentpopulationgroupsbasedonmaternalinheritance
- StudymutationsontheYchromosometotracelineageandmigrationofmales
- Compare breakpoints in the evolution of mutations with ages of populations and historicalevents.

DNA identification

- IdentifypotentialsuspectswhoseDNAmaymatchevidenceleftatcrimescenes

- Exoneratepersonswronglyaccusedofcrimes
- Identifycrimeandcatastrophevictims
- Establishpaternityandotherfamilyrelationships
- Identifyendangeredandprotectedspeciesasanaidtowildlifeofficials
- Detectbacteriaandotherorganismsthatmaypolluteair,water,soilandfood
- Matchorgandonorswithrecipientsintransplantprograms
- Determinepedigreeforseedorlivestockbreeds
- Authenticateconsumablessuchascaviarandwine

Agriculture, Livestock, Breeding, and Bioprocessing

- Growdisease,insect,anddroughtresistantcrops
- Breedhealthier,moreproductive,diseasesresistantfarmanimals
- Grow more nutritiousproduce
- Developbiopesticides
- Incorporateediblevaccinesincorporatedintofoodproducts
- Developnewenvironmentalcleanupusesforplantsliketobacco
- Cellulosisbiomassresearchforbioenergy

Anticipated benefits

- Improved diagnosis ofdisease
- Earlierdetectionofgeneticpredispositionstodisease
- Rational drugdesign
- Genetherapyandcontrolsystemsfordrugs
- Personalized, customdrugs

ELSI (Ethical, Legal, and Social Issues)

Privacy and confidentiality of genetic information

- Fairness in the use of genetic information by insurers, employers, courts, schools, adoptionagencies,andthemilitary,amongothers.
- Psychological impact, stigmatization and discrimination due to an individual's geneticdifferences.

- Reproductive issues including adequate and informed consent and use of genetic information in reproductive decisionmaking.
- 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 controlmeasures.
- Uncertainties associated with gene tests for susceptibilities and complex conditions (e.g.,heartdisease,diabetes,andalzheimer'sdisease).
- Fairnessinaccesstoadvancedgenomictechnologies.
- Conceptual and philosophical implications regarding human responsibility, free will vsgeneticdeterminismandconceptsofhealthanddisease.
- Healthandenvironmentalissuesconcerninggeneticallymodified(GM)foodsand microbes.
- Commercialization of products including property rights (patents, copyrights, and tradesecrets)andaccessibilityofdataandmaterials.

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

- Primarydatabase
- Secondarydatabase
- Compositedatabase

Primary Database

- Primary databases are produced with experimentally derived data such as nucleotidesequence,proteinsequenceormacromolecularstructure.
- Experimental results are submitted directly into the database by researchers, and thedataareessentiallyarchivalinnature.
- Once given a database accession number, the data in primary databases are never changed:theyformpartofthescientificrecord.

Secondary Database

- Secondary databases comprise data derived from the results of analyzing primary data.
- Secondary databases often draw upon information from numerous sources, includingotherdatabases(primaryandsecondary).
- They are highly curated, often using a complex combination of computational algorismandmanualanalysisandinterpretationtoderivenewknowledgefromthe public record ofscience.

	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 (Nucleotidesequence) Array Express Archieve and GEO (functional genomics data) Protein Data Bank (PDB coordinates of three dimensionalmacromolecular	InterPro (protein families, motifs and domains) UniProt Knowledgebase (sequence and functional information on proteins) Ensemble (variation, function, regulation and more layered onto whole genomesequences)

	structure	

Composite databases

- Collectionofvariousprimarydatabasesequences
- Renderssequencesearchinghighlyefficientasitsearchesmultipleresources
- Example:NRDB(nonredundantdatabase),OWL,MIPSX,SWISSPROT,TrEMBL

NucleicacidSequencedatabases

GenBank

- GenBank ® is the NIH genetic sequence database, an annotated collection of all publicly available DNAsequences
- GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive(ENA),andGenBankatNCBI.Thesethreeorganizationsexchangedataona dailybasis.
- AGenBankreleaseoccurseverytwomonthsandisavailablefromtheftpsite.
- The release notes for the current version of GenBank provide detailed information aboutthereleaseandnotificationsofupcomingchangestoGenBank.

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,TPAandPDB.

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 potentialvariation,andannotategenes.

dbGSS (Genome Survey Sequences)

- TheGSSdatabaseisacollectionofunannotatedshortsingle-readprimarilygenomic sequencesfromGenBankincludingrandomsurveysequencesclone-endsequences and exon-trappedsequences.

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 propertyrightsinalloraportionofthedatatheyhavesubmitted.
- NCBIisnotinapositiontoassessthevalidityofsuchclaims,andthereforecannot provide comment or unrestricted permission concerning the use, copying, or distributionoftheinformationcontainedinGenBank.

Confidentiality

- Some authors are concerned that the appearance of their data in GenBank prior to publicationwillcompromisetheirwork.
- GenBank will, upon request, withhold release of new submissions for a specified period oftime.
- 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 thespecifieddate,thesequencewillbereleaseduponpublication.
- Inordertopreventthedelayintheappearanceofpublishedsequencedata,weurge authorstoinformusoftheappearanceofthepublisheddata.
- 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 thistool.

Sequin, NCBI's stand-alone submission tool with wizards to guide the submission process isavailablebyFTPforuseonforMAC,PC,andUNIXplatforms.

EMBL

- European molecular BiologyLaboratory
- NucleicaciddatabasefromEBI(EuropeanBioinformaticsInstitute)
- ProducedincollaborationwithDDBJandGenBank
- Search engine – SRS (sequence RetrievalSystem)
- Keepingwiththetremendousgrowthinfieldofcomputationalbiology,aneedwas felt to establish an independent and parallel research institute that would act not justasamirrorhousingtheGenBanknucleotideresourcesofNCBI,butwouldalso develop matching databases and analysis tools. The European Molecular Biology Laboratory (EMBL) was thus established in 1974 and is now supported with fundingfrom20membersstatesoftheEuropeanUnion,IsraelandAustralia.EMBL currentlyoperatesfiveresearchinstitutesindifferentcountrieswithmaininstitute at Heidelberg,Germany.

The Five institutes of EMBL with their core research activities are

- EMBL Heidelberg(Germany)
- EMBL Grenoble (France)- StructuralBiology

- EMBL-EuropeanBioinformaticsInstitute(Hinxton,UK)-Bioinformatics
- EMBL Hamburg (Germany)-StructuralBiology
- EMBL Monterotondo (Italy)-MouseBiology

The broad goals of EMBL are

- Basic research in Molecularbiology
- Trainingmanpoweri.e.students,scientistandvisitors
- Developnewtools,technologiesandmethods
- Offerservicetotheresearchcommunity
- Transfertechnologytoindustryforcommercialization

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

- Biologicalontologies
- Literature
- Functional Genomics ormicroarray
- Nucleotides
- Pathways andnetworks
- Protein
- Proteomics
- Smallmolecules
- Structure

DDBJ

- DNA databank ofJapan
- Startedin1986incollaborationwithGenBank
- ProducedandmaintainedatNIG(NationalInstituteofGenetics)
- 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 andTechnology(MEXT).Lateronforitsefficientfunctioning,CenterforInformation Biology (CIB) was established at NIG in 1995. In 2004, NIG was made a member of Research Organization of Information andSystems.

- 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 tremendousincreaseinthenumberofsequencesubmittedtoDDBJ.

Roles of DDBJ

AsamemberofINSDC,primaryobjectiveofDDBJistocollectsequencedatafrom researchersalloverthe worldandtoissueauniqueaccessionnumberforeachentry.The datacollectedfromthesubmittersismadepublicallyavailableandanyonecanaccessthe datathroughdataretrievaltoolsavailableatDDBJ.EverydaydatasubmittedateitherDDBJ orEMBLorNCBIisexchanged,thereforeatanygiventimethesethreedatabasescontain samedata.Followingarethestepsalongwithsnapshotsshowingdataretrievalfrom DDBJ usinggetenetry.

- Open the homepage ofDDBJ
- Clickonthesearch/Analysislinkonthemenubar
- Click on getentrylink
- Typeintheaccessionnumberinthesearchboxandclickonsearch
- Desiredsequencewillberetrieved.

Software development

DDBJteamcontinuouslyfocusesondevelopingnewsoftwarewhichcanbeusedfor data analysis. For example, WINA (A window Analysis Program for the number of synonymousandnonsynonymousnucleotidesubstitutions)hasbeendevelopedbyDDBJ.It istoolwhichhelpsinvisualizingthedifferenceinaccumulationofbothsynonymousand nonsynonymous nucleotidesubstitutions.

Training courses

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

SwissProt

- Annotatedsequencedatabaseestablishedin1986
- Consistsofsequenceentriesofdifferentlieformats
- Similar format toEMBL

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 mostwidespreadproteinsequencecatalogcomprisingofEBI,SIBandPIR.

TherearefourmainfeaturesofSwiss-Prot

High qualityannotation

Itisachievedthroughmanuallycreatingtheproteinsequenceentries.Itisprocessed through 6stages.

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 retrievedfromliteraturedatabases.Thewholetextofeacharticleisscannedmanuallyand relevantinformationifgatheredandsupplementedtotheentry.

Family based curation: Putative homolog's are determined by reciprocal Blast searches and phylogenetic resources which are further evaluated, curated, annotated and propagatedacrosshomologousproteinstoensuredataconsistency.

Evidenceattribution: Allinformationincorporatedtothesequenceentryduringmanual annotation is linked to the original source so that users can trace back the origin of data and evaluateits.

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 minimalredundancy.

Integrationwithotherdatabases:Swiss-Protispresentlycross-referencedtomovethan50 specialized documentation files. Documentation file section provides an updated descriptivelistofalldocumentfiles.

PIR

- Protein InformationResource
- A division of National Biomedical Research Foundation (NBRF) inU.S
- OnecansearchforentriesordosequencesimilaritysearchatPIRsite.

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 mainaims.**

- To create an organized, non redundant, comprehensive protein database to study structural,functionalandevolutionaryrelationships.
- Togenerateinformationonbiologicaloriginofproteinsequences
- Tomakedatabaseeasilyaccessibleinpublicdomain

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

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.Withtheinventofhighthroughputsequencingtechniques,thereisanimmense 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,withouthavingtocompromisethequalityofSwiss-Prot.

TrEMBLsequencesareproducedattheEBIfromGenBankentriesandannotatedmostly 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.TrEMBLunreviewedentriesarekeptseparatedfromtheSwiss-Protmanually annotatedentriessoastomaintainthehighqualitydataoflater.

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 annotationinTrEMBLnexttothelevelofSwiss-Prot,thusimprovingthequalityofdata.

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 updateofdataiftheunderlyingdatasourcechanges.

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

SP-TrEMBL(Swiss-ProtTrEMBL)isacollectionofsequencesthatwillbefinallyupgraded toSwiss-Protaftertheirmanualannotationisfinished.

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 nucleicacids
- The NDB is supported by funds from the national science foundation and the department ofenergy
- TheNDBfollowsthedictionariesandformatsusedbytheworldwideproteindata bank
- Search the NDB byID
- EnteranNDBIDorPDBID
- Atlas,DepositData,DownloadData,Search,Education,Standards,Tools,Links
- TheNDBAtlasprovidessummaryinformationandimagesforeachstructureinthe database.TheAtlasisfirstdividedbyexperimentaltypeandthenbystructuretype.

Features include

- Image of the asymmetric and biological units, and crystal packing pictures for nucleicacidstructuresfromX-Raycrystallographicexperiments
- ImageoftheaverageandensemblestructureformNMRexperiments
- Linkstocoordinatefiles,experimentaldatafiles
- Tables of derived data, including torsion angles and hydrogen bonding classifications
- Special features for RNA structures, including images of secondary and tertiary structure

- EachpageinAtlasisgenerateddirectlyfromNDBusinganXMLtranslatorwhich formatsthedatacontainedineachNDBfiles.
- PicturespresentontheAtlaspagesweregeneratedbydifferentsoftware
- 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, oxygenisred,nitrogenisblue,andphosphatesareorange.
- The Atlas is first divided by how the structures were determined: by X-ray crystallographicorNMRexperiments.Galleryindexpages,whichincludeimagesfor eachstructureonthepage,andplaintextindexpagesareoffered.
- TheNDBprocessesdataforthecrystalstructuresofnucleicacids.Structurescanbe 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 otherinformationisenteredintoADITbytheauthor.
- Coordinatefilescanbedownloadedfromdownloadoption
- BasicdetailofDNAandRNAisgivenineducationoption
- X-plorparmetersandgeometriesaregiveninstandardoption

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 predictswhethergiventhreedimensionalproteinstructurecontainsaDNA-bindingHelix-turn-Helix (HTH) structuralmotif.

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

- StructuraldatafromthePDBcanbefreelyaccessedat
- It is very large global repository for processing and distribution of 3D macromolecularstructuredatasuchasprotein,nucleicacids
- DepositorstoPDBhavederivedthestructuresusingvarietyoftoolsandtechniques like X-ray crystal structure determination, NMR, cryoelectron microscopy and theoreticalmodeling.
- The database provides access at no charge on internet to structural data as well as methodstovisualizethestructureandtodownloadstructuralinformation.
- It is a primary data and databases derived from PDB are called secondary databases like SCOP andCATH.
- 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 morerecently theBioMagResBankattheUniversityofWisconsin-Madison.
- In1971WalterHamiltonofBNL(BrookhavenNationalLaboratory)agreedtosetup thedatabankatBrookhavanandthenhediedin1983.
- ThenTomKoeztletookoverdirectionofPDB

- 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 NMRandfewbyCryo-electronmicroscopy.Inthepast,numberofstructuresinPDB has grown nearlyexponentially.
- ThefileformatinitiallyusedbyPDBwascalledPDBfileformat.
- Around 1996, mmCIF (Macro Molecular Crystallographic Information File) started to phasedin.
- Thenin2005XMLversionofaboveformatcalledPDBMLwasdescribed.
- Thestructurefilecanbedownloadedinanyofthesethreeformats.
- Each structure published in PDB receives a four character alpha numeric identifier, its PDBID
- 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, SwissPDBViewer,startedBiochemandSirius.
- PDBWikiisawebsiteforcommunityannotationofPDBstructures.

Motif Database

Prosite

ItwasinitiatedandismaintainedbyAmosBairochandcolleagues,nowattheSwiss InstituteofBioinformatics.ItisbasedontheproteinssequencesinSWISS-PROT.Itaims atdescribingcharacteristicpatternsforsomedomainfamiliesusingregularexpressions, andcontainsabout1400patterns,rulesandprofile/matrices.Itisbeingmaintained,butit is fair to say that it has been superceded in practical terms by other search methods and databases,suchasPfam(mentionedbefore,anddiscussedlater).

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.
- Arule,ontheotherhand,isintendedtohighlightfeaturesinaproteinsequencethat doesnotnecessarilyhaveanythingtodowithaspecificproteinfamily.Forexample, potentialglycosylationsitesandphosphorylationsitescanbefoundinmanyprotein sequences,andhavelittletodowiththefamilyofaprotein.

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 PROSITEpatternintoaUNIX-typeregularexpression.

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;

DRQ00023,CEL1_AGABI,T;Q12714,GUN1_TRILO,T;P07981,GUN1_TRIRE,T;

DRP07982,GUN2_TRIRE,T;P43317,GUN5_TRIRE,T;P46236,GUNB_FUSOX,T;

DRP46239,GUNF_FUSOX,T;P45699,GUNK_FUSOX,T;P15828,GUX1_HUMGR,T;

DRQ06886,GUX1_PENJA,T;P13860,GUX1_PHACH,T;P00725,GUX1_TRIRE,T;

DRP19355,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.

Letusgothroughtheelementsinthepatterntoseewhattheymean:

- Eachnon-xletterdefinesoneparticulartypeofamino-acidresidueinthatposition inthepattern.Here,wemusthaveatripeptideCys-Gly-Glyinthebeginningofthe matchingsegmentofaproteinchain.Thedashcharacters'-'addnoinformationto thepattern,andareaddedtomakethepatternslightlyeasiertoread.
- The notation x(4,7) means that at least 4 and at most 7 residues of any type may occuratthisposition.Thiscorrespondstothenotation.{4,7}inaUNIX-typeregular expression.
- Thenotation[NHG]meansthesamethingasinaUNIX-typeregularexpression:in this position any of the residues within the brackets may be chosen. One and only onesuchresiduemustbeatthisposition.
- 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 andProareallowedinthisposition.

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 indexedinNLM'sMEDLINEdatabase,aswellasfromotherlifesciencejournalsandonline 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 otherNCBIresources,includingitsvariousmolecularbiologydatabases.

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 inlifesciences,withaconcentrationonbiomedicine.
- ThemajorityofjournalsselectedforMEDLINEarebasedontherecommendationof theLiteratureSelectionTechnicalReviewCommittee(LSTRC),anNIH-chartered advisory committee of external experts analogous to the committees that review NIHgrantapplications.Someadditionaljournalsandnewslettersareselectedbased onNLM-initiatedreviewsinareasthatarespecialprioritiesforNLMorotherNIH components(e.g.,historyofmedicine,healthservicesresearch,AIDS,toxicologyand environmental health, molecular biology, and complementary medicine). These reviewsgenerallyalsoinvolveconsultationwithanarrayofNIHandoutsideexperts or,insomecases,externalorganizationswithwhichNLMhasspecialcollaborative 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 withNLMMedicalSubjectHeadings(MeSH)andaddedtoMEDLINEorconvertedto out-of-scopestatus.
- CitationsthatprecedethedatethatajournalwasselectedforMEDLINEindexing.
- Some OLDMEDLINE citations that have not yet been updated with current vocabularyandconvertedtoMEDLINEstatus.
- 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 PubMedCentralandreceiveaqualitativereviewbyNLM.

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*.

TheEntrezProgrammingUtilities,E-Utilities,andtheCubby(My NCBIsubsequently replacedtheCubby)alsowerereleased.

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 theirpurpose.

GMOD's database related componentsare:

Chado

ChadoisthemodulardatabaseschemaofGMOD. Chadoisaboutorganizingyour datainadatabasesothatyoucanmanageitandcanconnectotherGMODcomponentstoit (eitherdirectlyorviadataexports).WhensomeonespeaksoftheGMODSchematheyare speaking aboutChado.

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 canyieldlargeamountsofinformationaswellassequencesetsthatcanbeusedforfurther analysis. Two of the major interfaces allowing you to search the metadata of sequences (andcarryoutotheranalysistasksaswell)aretheSequenceRetrievalService(SRS)atthe EBIandEntrezattheNCBI.Intheassociatedlectureandpractical,wewillfocusonusing SRS.WewilllookatEntrezinthecontextofblastsearchingofsequencedatabaseslaterin thecourse.

Possible questions

1. DefineBioinformatics?Explaintheobjectives,milestonesandgoalsofBioinformatics
2. Whatarebiologicaldatabases?Discussindetail
3. Writeshortnotesonapplicationofbioinformaticsin
    a) Molecularmedicine
    b) Microbialgenomics
    c) Agriculture
    d) Biotechnology
4. Discussindetailabouthumangenomeprojectaddanoteonitscurrentdevelopments.
5. Explainbibliographicdatabasesanditsuses.
6. Explainindetailsaboutthehistoricalbackgroundof Bioinformatics
7. WriteashortnotesonEthical,Legalandsocialissuesinhumangenomeproject.
8. Definebiologicaldatabases?Distinguishprimaryandsecondarydatabases.
9. Discuss in brief aboutPROSITE.
10. Explaininbriefaboutnucleicacidsequencedatabases.
11. Explainindetailaboutorganismspecificdatabases.

KAPRAGAM ACADEM YOF HIGHER EDUCATION

DEPARTMENT OF BIOCHEMISTRY

II- B.Sc Biochemistry

18BCU204-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-

IISYLLAB

US

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

SequencealignmentisawayofarrangingthesequencesofDNA,RNA,orproteinto identify regions of similarity that may be a consequence of functional, structural, or evolutionaryrelationshipsbetweenthesequences.

Alignedsequencesofnucleotideoraminoacidresiduesaretypicallyrepresentedasrows withinamatrix.Gapsareinsertedbetweentheresiduessothatidenticalorsimilar charactersarealignedinsuccessivecolumns.

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

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

Asequencealignment,producedbyClustalW,oftwohumanzincfingerproteins,identified ontheleftbyGenBankaccessionnumber.

Single letters: aminoacids.

Red: small, hydrophobic, aromatic, not Y.

Blue: acidic. Magenta: basic.

Green: hydroxyl, amine, amide, basic.

Gray:others."*":identical.":":conservedsubstitutions(samecolourgroup)."."":semi-conservedsubstitution(similarshapes).

Definition of sequence alignment

- Sequencealignmentistheprocedureofcomparingtwo(pair-wisealignment)or moremultiplesequencesbysearchingforaseriesofindividualcharactersor patternsthatareinthesameorderinthesequences.
- Therearetwotypesofalignment:localandglobal.Inglobalalignment,an attempt is made to align the entire sequence. If two sequences have approximatelythesamelengthandarequitesimilar,theyaresuitableforthe globalalignment.
- Localalignmentconcentratesonfindingstretchesofsequenceswithhighlevelof 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 evolutionaryinformation.
- Sequencesthatareverymuchalikemayhavesimilarsecondaryand3D structure,similarfunctionandlikelyacommonancestralsequence.Itis extremely unlikely that such sequences obtained similarity by chance. For DNA moleculeswith*n*nucleotidessuchprobabilityisverylow*P=4-n*.Forproteins theprobabilityevenmuchlower*P=20–n*,where*n*isanumberofaminoacid residues

- Largescalegenomestudiesrevealed existenceofhorizontaltransferofgenes andothersequencesbetweenspecies,whichmaycausesimilaritybetweensome sequencesinverydistantspecies

Methods of Sequence Alignment

- Dot Matrixanalysis
- DynamicProgramming(DP)algorithm
- Word(or)K-tuplemethods

Dot matrix analysis

- Comparingfortwosequence
- Onesequence(A)topofthematrix
- Otheronesequence(B)downleftside
- Anyregionofsimilarityisrevealedbyadiagonalrowofdots.
- Five cleardiagonals
- DiagonalsareobtainedbyaligninggenomicandcDNA.
- Fivediagonalsrepresentthefiveexonsofthegenewhichwasconfirmedfromthe annotatedentryofthegene.

Sequence alignment program is to align the two sequences

- Toproducehighestscoreascoringmatrixisusedtoaddpointstothescoreforeach matchandsubtractthemforeachmismatch.
- Matrixes are used for nucleic acid alignment to involve fairly simple match/mismatch scoringschemes.

Parameters used for sequence alignment

1. scoringmatrix
2. Substitutionmatrices
3. Gappenalty

Scoring matrices

- It is critical to have reasonable scoring schemes accepted by the scientific communityforDNAandproteinsandfordifferenttypesofalignments

- Thewealthofinformationaccumulatedinthegene/proteinbankswasutilisedwith dynamic programming procedure to create such matrices for scoring matches and separatelypenaltiesforgapsintroductionandextensions
- MatricesforDNAarerathersimilarastherareonlytwooptionspurine& pyrimidineandmatch&mismatch
- Proteinsaremuchmorecomplexandthenumberofoptionissignificant
- PAMandothermatricesarerepresentedinlogoddsscores,whichistheratioof chanceofaminoacidsubstitutionduetoessentialbiologicalreasontothechanceof randomsubstitution
- TherearemanydifferentPAMs,whicharerepresentingdifferentevolutionary scenarios.
- PAM250representsalevelof250%ofchangesexpectedin2500MY
- PAM is more suitable for studying quite distant proteins, BLOSUM is for more conservedproteinsofdomains

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	

- Aminoacidsaregroupedaccordingtotothechemistryofthesidegroup:(C) sulfhydryl,(STPAG)-smallhydrophilic,(NDEQ)acid,acidamideandhydrophilic, (HRK)basic,(MILV)smallhydrophobic,and(FYW)aromatic.

- Logoddsvalues:+10meansthatancestorprobabilityisgreater,0meansthatthe probabilitiesareequal,-4meansthatthechangeisrandom.Thustheprobabilityof alignmentYY/YYis10+10=20,whereasYY/TPis–3-5=-8,arareandunexpected between homologoussequences.

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	

IdeologyofBLOSUMissimilarbutitiscalculatedfromaverydifferentandmuch largersetofproteins,whicharemuchmoresimilarandcreateblocksofproteinswitha similarpattern.

Differences between PAM and BLOSUM

1. PAMmatricesarebasedonanexplicitevolutionarymodel(i.e.replacementsare countedonthebranchesofaphylogenetictree),whereastheBLOSUMmatricesare basedonanimplicitmodelofevolution.
2. ThePAMmatricesarebasedonmutationsobservedthroughoutaglobalalignment, this includes both highly conserved and highly mutable regions. The BLOSUM matricesarebasedonlyonhighlyconservedregionsinseriesofalignments forbiddentocontaingaps.
3. Themethodusedtocountthereplacementsisdifferent:unlikethePAMmatrix,the BLOSUMprocedureusesgroupsofsequenceswithinwhichnotallmutationsare counted thesame.
4. Higher numbers in the PAM matrix naming scheme denote largerevolutionary distance,whilelargernumbersintheBLOSUMmatrixnamingschemedenotehigher sequencesimilarityandthereforesmallerevolutionarydistance.Example:PAM150

isusedformoredistantsequencesthanPAM100;BLOSUM62isusedforcloser sequences thanBlosum50.

Substitution matrices

- 210scorepossibilitiesforanypossibleproteinpair.
- 20X20matrix,wherethediagonalgives100%matchbetweentheaminoacids.
- Maindiagonalareofidentical20aminoacidscoresandoneachsideofdiagonal190 scoringthataresimilarobtain210scoringtermsfor20aminoacidcombinations.
- Pairofaminoacidistermedaslog–oddsvaluesandthesehavebeenscaledand roundedtothenearestintegerforcomputationalefficiencyknownasscorematrix or substitutionmatrix.
- ThelateMargaretDayhoffpioneerinproteindatabasingandcomparison
- Dayhoff, MDM (Mutation Data Matrix] (or) PAM (Point (or) Percent Accepted Mutation).
- PAMonesuchmajoraminoacidscoringorsubstitutionmatrix
- BLOSUMseriesofmatriceswerecreatedbySteveHenikoffandcolleagues.
- MatricesareusedBLOSUM80,62,40and30.
- PAMmatricesusedarePAM120,160,250and350matrices.

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

- Gapisanymaximalconsecutiverunofspacesinasinglestringofagivenalignment.

- Gaphelpstocreatealignmentsthatbetterconformtounderlyingbiologicalmodels andmorecloselyfitpatternsthatoneexceptstofinalinmeaningfulalignment.
- No.ofcontinuousgapsandnotonlythenumberofspaceswhencalculatingan alignmentmark.

Example

X=attc--ga-tggacc Y=a--cgtgatt---cc

- Tourgapscontainingatotalofeightspaces
- 7matches,nomismatch.
- No.ofgapsinthealignmentwillbedenotedas#gaps

Pairwise Sequence Alignment:

- Thetwosequencesarehomologous,i.e.theyhaveevolvedfromacommon ancestor.
- Differencesbetweenthemareduetoonlytwokindsofevents,namely insertion
- deletions (*indels*) and substitutions (change of single elements of the sequence-
- aminoacidsifthesequenceisaproteinandnucleicacid,ifthesequenceis DNA).

Two types pairwise sequence alignment

- Needlemen-WunschAlogorithm(or)GlobalAlignment
- Smith-Waterman(or)LocalAlignment

Needleman–Wunsch algorithm

- The **Needleman–Wunsch algorithm** performs a global alignment on two sequences(called*A*and*B*here).
- Itiscommonlyusedinbioinformaticstoalignproteinornucleotidesequences.
- Thealgorithmwaspublishedin1970bySaulB.NeedlemanandChristianD. Wunsch.
- TheNeedleman–Wunschalgorithmisanexampleofdynamicprogramming,and was the first application of dynamic programming to biological sequence comparison.

A modern presentation

Scoresforalignedcharactersarespecifiedbyasimilaritymatrix.Here,$S(a,b)$isthe similarityofcharacters*a*and*b*.Itusesalineargappenalty,herecalled*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 thealignment:

AGACTAGTTAC CGA—GACGT

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

Tofindthealignmentwiththehighestscore,atwo-dimensionalarray(ormatrix) *F*isallocated.Theentryinrow*i*andcolumn*j*isdenotedherebyF_{ij}.Thereisonecolumn foreachcharacterinsequence*A*,andonerowforeachcharacterinsequence*B*.Thus,ifwe 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 doublingtherunningtime.

Dotplots

- Themostintuitiverepresentationofthecomparisonbetweentwosequencesuses dot-plots.
- Onesequenceisrepresentedoneachaxisandsignificantmatchingregionsare distributedalongdiagonalsinthematrix.

Exercise: Making a dotplot

unix % **dottup**

DNA sequence dot plot

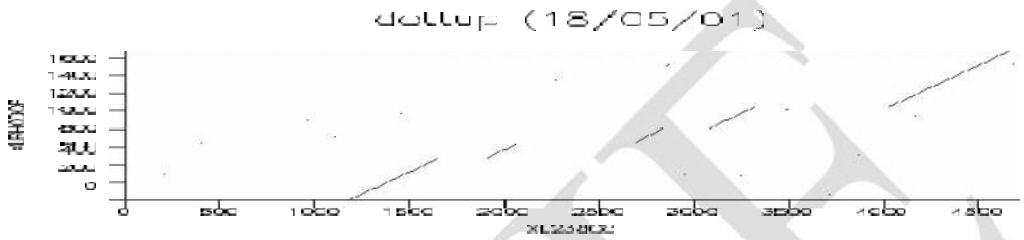
Input sequence: **embl:xl23808**

Second sequence:**embl:xlrhodop**

Wordsize[4]:**10**

Graph type[x11]:

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



- Thediagonallinesrepresentareaswherethetwosequencesalignwell.Youcansee thattherearefivecleardiagonals.
- AligninggenomicandcDNA-thesefivediagonalsrepresentthefiveexonsofthe gene!IfyoulookattheoriginalEMBLentryforthegenomicsequenceusingSRS,you willseethattheannotatedentrysaysthattherearefiveexonsinthisgene.Soour resultsareinagreement.
- Thesettingswehaveusedforthisexamplearethosethatgivethebestresults. dottuplooksforexactmatchesbetweensequences.
- Asweexpecttheexonregionsfromthegenomicsequencetoexactlymatchthe cDNAsequencewecanuselongerwordlengthsasweshouldstillgetexactmatches.
- ThisgivesaverycleanplotIfyouweretomatchthecDNAsequenceagainstthatof arelatedsequence,e.g.therhodopsinfrommouse(embl:m55171)thenyou wouldn'texpectlongexactmatchessoshoulduseashorterwordlength.

Smith–Waterman algorithm

- TheSmith–Watermanalgorithmisawell-knownalgorithmforperforminglocal sequencealignment;thatis,fordeterminingsimilarregionsbetweentwonucleotide or proteinsequences.
- Instead of looking at the total sequence, the Smith–Waterman algorithm compares segmentsofallpossiblelengthsandoptimizesthesimilaritymeasure

Algorithm Explanation

A matrix *H* is built as follows:

if$a_i=b_j w(a_i,b_j)=w$(match)orif$a_i!=b_j w(a_i,b_j)=w$(mismatch) Where:

a,*b* = Strings over the Alphabet Σ

m = length(*a*)

n = length(*b*)

H(*i*,*j*)-isthemaximumSimilarity-Scorebetweenasuffixofa[1...i]andasuffixof b[1...j]

, '-' is the gap-scoring scheme

Example

Sequence 1 = ACACACTA

Sequence 2 =AGCACACA

w(gap) =0

w(match) = +2

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

Toobtaintheoptimumlocalalignment,westartwiththehighestvalueinthematrix (i,j).Then,wegobackwardstooneofpositions(i-1,j),(i,j-1),and(i-1,j-1)dependingonthe directionofmovementusedtoconstructthematrix.Wekeeptheprocessuntilwereacha matrixcellwithzerovalue,orthevalueinposition(0,0).

Intheexample,thehighestvaluecorrespondstothecellinposition(8,8).Thewalk backcorrespondsto(8,8),(7,7),(7,6),(6,5),(5,4),(4,3),(3,2),(2,1),(1,1),and(0,0),

Oncewe'vefinished,wereconstructthealignmentasfollows:Startingwiththelast value,wereach(i,j)usingthepreviously-calculatedpath.Adiagonaljumpimpliesthereis analignment(eitheramatchoramismatch).Atop-downjumpimpliesthereisadeletion. Aleft-rightjumpimpliesthereisaninsertion.

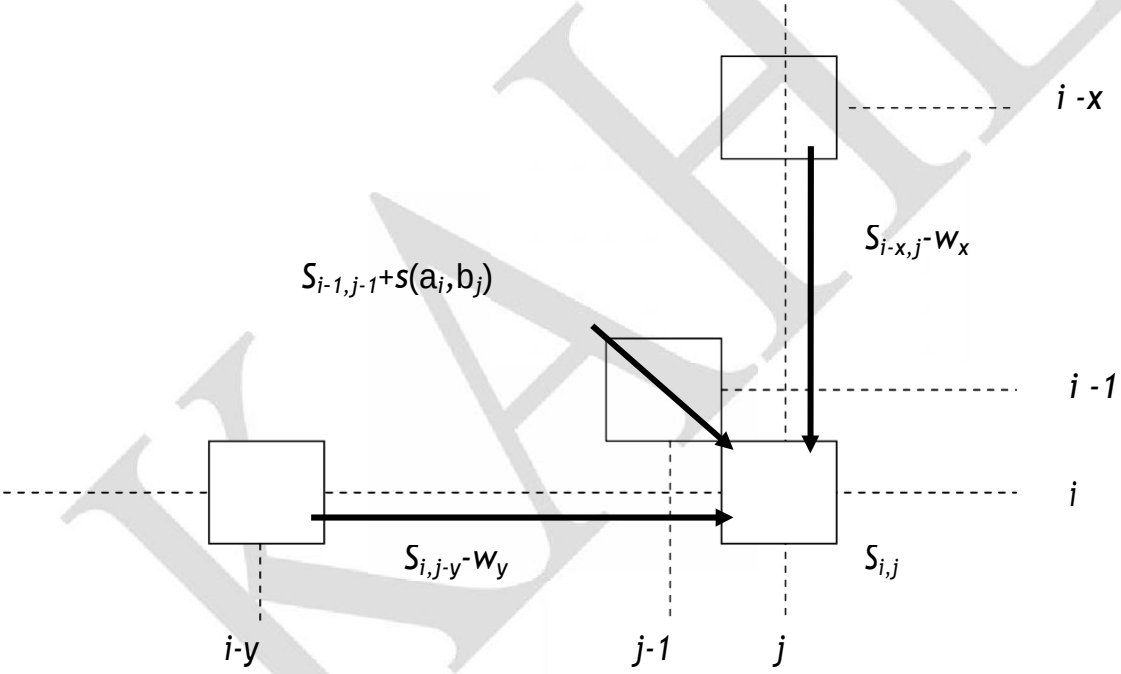
Description of the dynamic programming algorithm

- Considerbuildingthisalignmentinsteps,startingfromtheinitialmatch(V/V)and thensequentiallyaddinganewpairuntilthealignmentiscomplete,ateachstage

choosingapairfromallthepossiblematchesthatprovidesthehighestscoreforthe alignmentuptothatpoint.

- Ifthefullalignmenthasthehighestpossible(oroptimal)score,thentheold alignmentfromwhichitwasderived(A)byadditionofthealignedY/Ypairmust alsohavebeenoptimaluptothatpointinthealignment.
- Inthismanner,thealignmentcanbetracedbacktothefirstalignedpairthatwas alsoanoptimalalignment.
- Theexample,whichwehaveconsidered,illustrates3choices:1.Matchthenext character(s)inthefollowingposition(s);2.Matchthenextcharacter(s)toagapin theuppersequence;3.Addagapinthelowersequence.

Formal description of dynamic programming algorithm



- Thisdiagramindicatesthemovesthatarepossibletoreachacertainposition(*i*,*j*) startingfromthepreviousrowandcolumnatposition(*i-1*,*j-1*)orfromanyposition inthesameroworcolumn
- Diagonalmovewithnogappenaltiesormovefromanyotherpositionfromcolumn *j* or row *i*, with a gap penalty that depends on the size of the gap

Formal description of dynamic programming algorithm

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

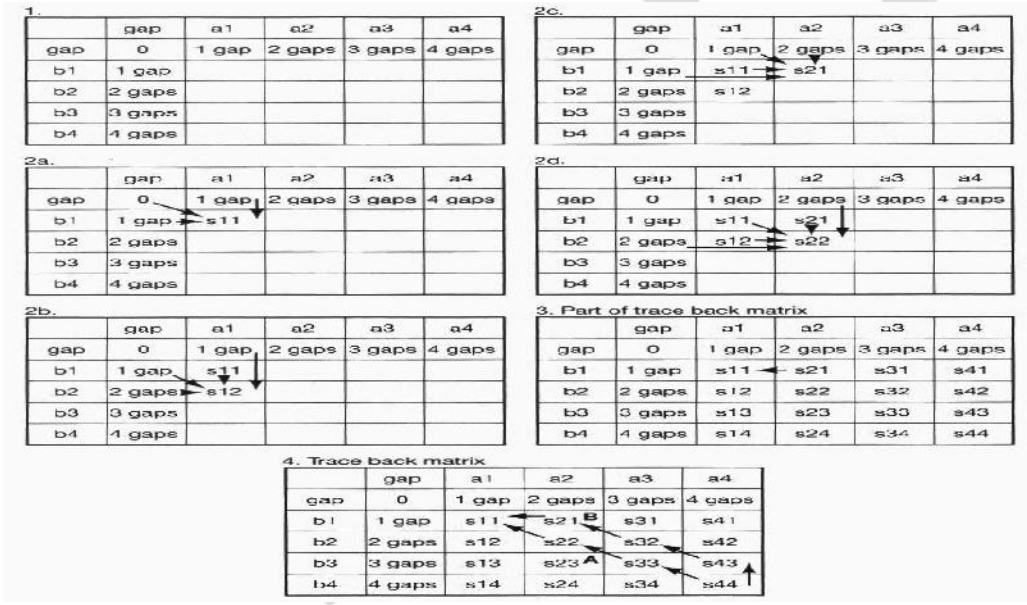
$$S_{ij} = \max \{ S_{i-1,j-1} + s(a_i b_j),$$

$$\max(S_{i-x,j} - w_x),$$

$$x \geq 1$$

where*Sij*isthescoreatpositionat*i*insequence**a**and*j*insequence**b**,s(a*i*b*j*)is scoreforaligningthecharacteratpositions*i*and*j*,*wx*isthepenaltyforagapoflength*x*in sequence**a**,and*wx*isthepenaltyforagapoflength*y*insequence**b**.

NotethatS*ij*isatypeofrunningbestscoreasthealgorithmmovesthrougheveryposition in thematrix



Alignment A:a1a2a3a4

b1b2b3b4

Alignment B: a1 a2 a3 a4-

b1 - b2 b3 b4

Thehighestscoringmatrixpositionislocated(inthiscases44)andthentracedback asfaraspossible,generatingthepathshown.

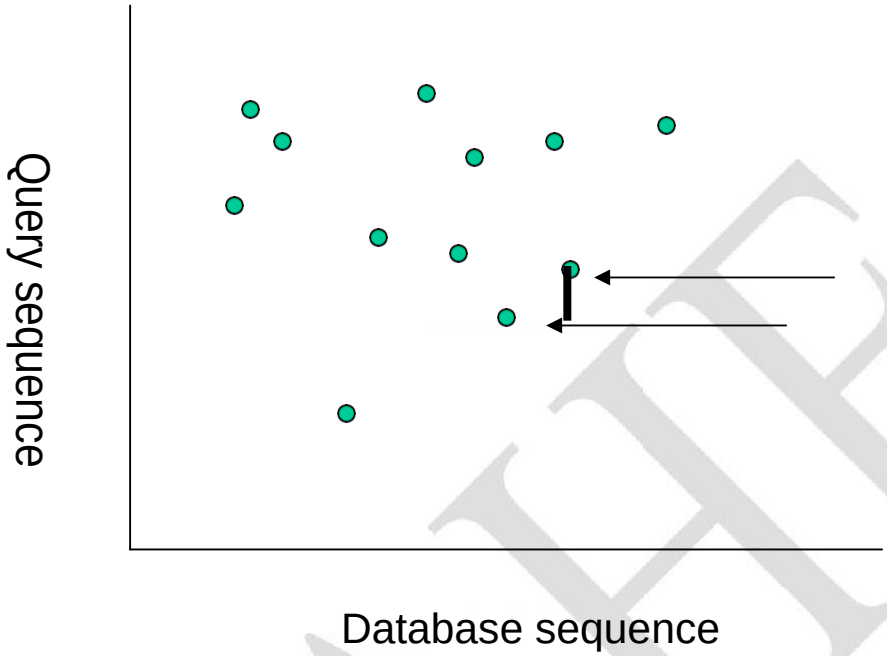
BLAST

- BLAST(**B**asic**L**ocal**A**lignment**S**earch**T**ool)comesunderthecategoryofhomology and similaritytools.
- Itisasetofsearchprogramsdesignedfortheplatformandisusedto performfastsimilaritysearchesregardlessofwhetherthequeryisforproteinor DNA.
- Comparisonofnucleotidesequencesinadatabasecanbeperformed.Alsoaprotein databasecanbesearchedtofindamatchagainstthequeriedproteinsequence.
- NCBIhasalsointroducedthenewqueuingsystemtoBLAST(QBLAST)thatallows userstoretrieveresultsattheirconvenienceandformattheirresultsmultipletimes withdifferentformattingoptions.

BLAST procedure

- ThestepsusedbytheBLASTalgorithm:
- Theseqisoptionallyfilteredtoremovelow-complexityregions(AGAGAG…)
- Alistofwordsofcertainlengthismade
- Usingsubstitutionscoresmatrixes(likePAMorBLOSUM62)thequeryseq.words areevaluatedformatcheswithanyDBseq.andthesescores(log)areadded
- Acutoffscore(*T*)isselectedtoreducenumberofmatchestothemostsignificant ones.
- Theaboveprocedureisrepeatedforeachwordinthequeryseq.
- The remaining high-scoring words are organised into efficient search treeand rapidlycomparedtotheDBseq.
- Ifagoodmatchisfoundthenanalignmentisextendedfromthematchareainboth directionsasfarasthescorecontinuetogrow.InthelatestversionofBLASTmore time-efficientmethodisused

- Theessenceofthismethodisfindingadiagonalconnectingungappedalignments and extendingthem



- Thenextstepistodeterminethosehighscoringpairs(HSP)ofseq.,whichhave scoregreaterthanacutoffscore(*S*).Sisdeterminedempiricallybyexamininga rangeofscoresfoundbycomparingrandomseq.andbychoosingavaluethatis significantlygreater.
- ThenBLASTdeterminesstatisticalsignificanceofeachHSPscore.Theprobability*p* ofobservingascoreSequaltoorgreaterthan*x*isgivenbytheequation:$p(S \geq x) = 1 - \exp(-e^{-\lambda(x-u)})$,whereu=[log(Km'n')]/λandKandλareparametersthatare calculatedbyBLASTforaminoacidornucleotidesubstitutionscoringmatrix,n'is effectivelengthofthequeryseq.andm'iseffectivelengthofthedatabaseseq.
- Onthenextstepastatisticalassessmentsismadeinthecaseiftwoormore HSP regionsarefoundandcertainmatchingpairsareputindescendingorderinthe outputfileasfarastheirsimilarity/scoreisconcerned.

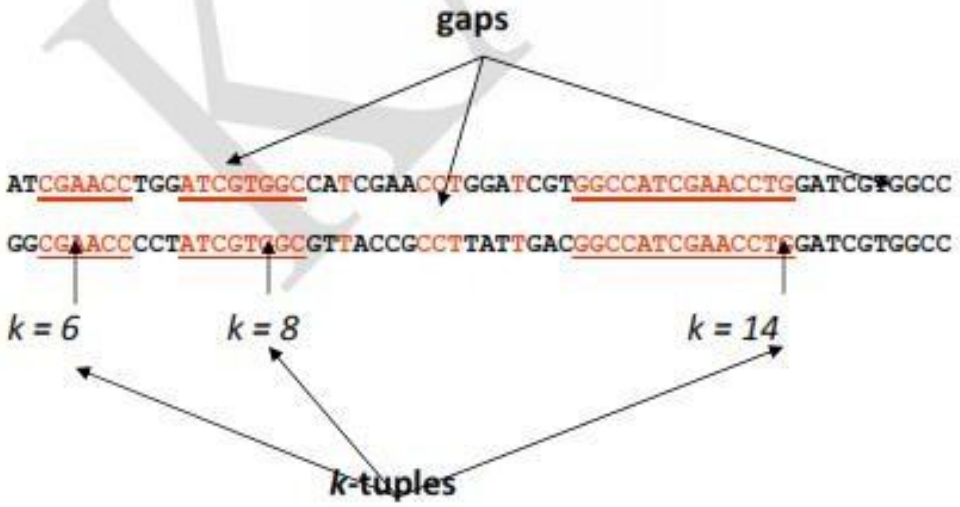
Depending on the type of sequences to compare, there are different programs:

- **blastp**comparesanaminoacidquerysequenceagainstaproteinsequencedatabase

- **blastn** comparesanucleotidequerysequenceagainstanucleotidesequence database
- **blastx**comparesanucleotidequerysequencetranslatedinallreadingframes againstaproteinsequencedatabase
- **tblastn**comparesaproteinquerysequenceagainstanucleotidesequencedatabase dynamicallytranslatedinallreadingframes
- **tblastx**comparesthesix-frametranslationsofanucleotidequerysequenceagainst thesix-frametranslationsofanucleotidesequencedatabase

FASTA

- FASTAisaprogramforrapidalignmentofpairsofproteinandDNAsequences.
- Comparisonofallnucleotidesoraminoacidsisnotanoption,evenforpowerful
- computers,FASTAinsteadsearchesformatchingsequencepatterns("words")called *k*-tuples. These patterns comprise *k* consecutive matches in the compared sequences.
- Using*k*-tuplesFASTAbuildsalocalalignment.
- FinallyFASTAscoresthisalignmentandoutputalistofsequencessimilartoa queryinthedescendingorder.



FASTAperformsthefollowingstatisticaltasks:1.TheaveragescoreforDBseq.of thesamelengthisdetermined,2.Theaveragescoreisplottedagainstthelogofaverage seq.lengthineachlengthrange,3.Thepointsarethenfittedtoastraightlinebylinear regression,4.Azscore,thenumberofstandarddeviationsfromfittedline,iscalculatedfor eachscore,5.Lowscoringseq.areremoved.6.AstatisticalcomparisonwithZdistribution follows,whichallowstocalculate*E*()value.If*E*()=0,andzscoreishightwosequences areidentical,whenEishigherthenathresholdlevel,noclearsimilarityisobserved**.**

Methods used by FASTA to locate sequence similarity:

A. Rapidlocationof10bestmatchingregionsineachpair.ForDNAseq.*k*=4-6,forprotein *k* = 1-2. The highest-density matches identified.

B. The highest-density regions are evaluated using special scoring matrixes (next lecture) andthebestinitialregions(INIT1)arefound(*-thebest).

C. LongerregionsofidentityofscoreINITNaregeneratedbyjoiningINITwithscores higherthanacertainthreshold,whichincludepositivescoresforsimilarityandnegative forgaps.Optimizationprocedurefollows.

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

- Oftenappliedtoproteins
- Proteinsthataresimilarinsequenceareoftensimilarinstructureandfunction.
- Sequencechangesmorerapidlyinevolutionthandoesstructureandfunction

Overview of Methods

- Dynamicprogramming–toocomputationallyexpensivetodoacompletesearch; usesheuristics
- Progressive–startswithpair-wisealignmentofmostsimilarsequences;addsto that
- Iterative–makeaninitialalignmentofgroupsofsequences,addstothese(e.g. geneticalgorithms)
- Locally conservedpatterns
- Statisticalandprobabilisticmethods

Dynamic Programming

- Computationalcomplexity–evenworsethanforpair-wisealignmentbecause we'refindingallthepathsthroughann-dimensionalhyperspace(Wecan picturethisin2or3dimensions.)
- Canalignabout7relativelyshort(200-300)proteinsequencesinareasonable amountoftime;notmuchbeyondthat.
- Let'spicturethisin3dimensions(pp.146-157inbook).Itgeneralizeston.
- Considerthepair-wisealignmentsofeachpairofsequences.
- Createaphylogenetictreefromthesescores.
- Consideramultiplesequencealignmentbuiltfromthephylogenetictree.
- Thesealignmentscircumscribeaspaceinwhichtosearchforagood(butnot necessarilyoptimal)alignmentofallnsequences.
- Createaphylogenetictreebasedonpair-wisealignments(Pairsofsequences thathavethebestscoresarepairedfirstinthetree.)
- Doa"first-cut"msabyincrementallydoingpair-wisealignmentsintheorderof "alikeness" of sequences as indicated by the tree. Most alike sequences aligned first.
- Usethepair-wisealignmentsandthe"first-cut"msatocircumscribeaspace withinwhichtodoafullmsathatsearchesthroughthissolutionspace.

- Thescoreforagivenalignmentofallthesequencesisthesumofthescoresfor eachpair,whereeachofthepair-wisescoresismultipliedbyaweighte indicatinghowfarthepair-wisescorediffersfromthefirst-cutmsaalignment score.
- Doesnotguaranteeanoptimalalignmentofallthesequencesinthegroup.
- Doesgetanoptimalalignmentwithinthespacechosen.

Phylogenetic Tree

- Dynamicprogrammingusesaphylogenetictreetobuilda"first-cut"msa
- Thetreeshowshowproteincouldhaveevolvedfromsharedoriginsover evolutionarytime.
- Seepage143in*Bioinformatics*byMount.
- Chapter6goesintodetailonthis.

Progressive Methods

- Similartodynamicprogrammingmethodinthatitusesthefirststep(i.e.,it createsaphylogenetictree,alignsthemost-alikepair,andincrementallyadds sequencestothealignmentinorderof"alikeness"asindicatedbythetree.)
- DiffersfromdynamicprogrammingmethodforMSAinthatitdoesn'trefinethe "first-cut"MSAbydoingafullsearchthroughthereducedsearchspace.(Thisis thecomputationallyexpensivepartofDPMSAinthat,eventhoughwe'vecut downthesearchspace,it'sstillbigwhenwehavemanysequencestoalign.)
- Generallyproceedsasfollows:
  - Chooseastartingpairofsequencesandalignthem
  - Aligneachnextsequencetothosealreadyaligned,oneatatime
- Heuristicmethod–doesn'tguaranteeanoptimalalignment

ClustalW

- Basedonphylogeneticanalysis
- Aphylogenetictreeiscreatedusingapairwisedistancematrixandnearest- neighboralgorithm

- The most closely-related pairs of sequences are aligned using dynamic programming
- Eachofthealignmentsisanalyzedandaprofileofitiscreated
- Alignmentprofilesarealignedprogressivelyforatotalalignment
- WinClustalWreferstoaweightingofscoresdependingonhowfarasequenceis fromtherootonthephylogenetictree.

Problems with Progressive Method

- Highlysensitivetothechoiceofinitialpairtoalign.Iftheyaren'tverysimilar,it throws everythingoff.
- It'snottrivialtocomeupwithasuitablescoringmatrixorgappenaties.

Iterative Methods for Multiple Sequence Alignment

- Get analignment.
- Refineit.
- Repeatuntilonemsadoesn'tchangesignificantlyfromthenext.
- Anexampleisgeneticalgorithmapproach

Genetic Algorithms

- Ageneralproblemsolvingmethodmodeledonevolutionarychange.
- Createasetofcandidatesolutionstoyourproblem,andcausethesesolutionsto evolveandbecomemoreandmorefitoverrepeatedgenerations.
- Usesurvivalofthefittest,mutation,andcrossovertoguideevolution.

Evolutionary Change in Genetic Algorithms

- survivalofthefittest–thebestsolutionssurviveandreproducetothenext generation
- mutation–somesolutionsmutateinrandomways(buttheymustalwaysremain viablesolutions)
- crossover–solutions"exchangeparts"

Laying Out the Problem

- Whatwouldacandidatesolutionlooklikeinamultiplesequencealignment program?(anMSAof~20proteins)
- Howmanycandidatesolutionsshouldtherebe?(~100)

Evolving to a Next Generation

- Whichcandidatesolutionsshouldsurvivetothenextgeneration?
  - First,takethetophalfbasedonbestsumofpairsscores
  - Thenrandomlyselectsecondhalf,givingmorechancetoanMSA'sbeing selectedinproportiontohowgooditsscoreis.

Phylogenetic tree

- A**phylogenetictree**or**evolutionarytree**isabranchingdiagramor"tree"showing theinferredevolutionaryrelationshipsamongvariousbiologicalspeciesorother entitiesbaseduponsimilaritiesanddifferencesintheirphysicaland/orgenetic characteristics.
- Thetaxajoinedtogetherinthetreeareimpliedtohavedescendedfromacommon ancestor.
- Ina**rooted**phylogenetictree,eachnodewithdescendantsrepresentstheinferred mostrecentcommonancestorofthedescendantsandtheedgelengthsinsome treesmaybeinterpretedastimeestimates.
- Eachnodeiscalledataxonomicunit.
- Internalnodesaregenerallycalledhypotheticaltaxonomicunits(HTUs)asthey cannotbedirectlyobserved.
- Trees are useful in fields of biology such as systematics and comparative phylogenetics.

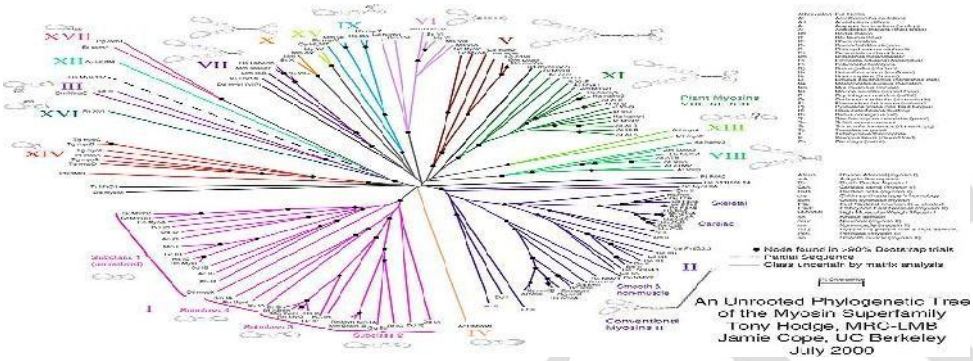
Types

A rooted phylogenetic tree

Arootedtreeisusedtomakeinferencesaboutthemostcommonancestorofthe leavesorbranchesofthetree.Mostcommonlytherootisreferredtoasan"outgroup"

Unrooted tree:

Anunrootedtreeisusedtomakeanillustrationabouttheleavesorbranches,but notmakeassumptionregardingacommonancestor.



Totalrootedtreesandtotalunrootedtrees,where*n*representsthenumberofleafnodes.

Amonglabeledbifurcatingtrees,thenumberofunrootedtreeswith*n*leavesisequaltothe numberofrootedtreeswith*n*−1leaves.

dendrogram isabroadtermforthediagrammaticrepresentationofaphylogenetictree.

A**cladogram** isatreeformedusingcladisticmethods.Thistypeoftreeonlyrepresentsa branchingpattern,i.e.,itsbranchlengthsdonotrepresenttime.

A**phylogram**isaphylogenetictreethatexplicitlyrepresentsnumberofcharacterchanges throughitsbranchlengths.

A**chronogram**isaphylogenetictreethatexplicitlyrepresentsevolutionarytimethrough its branchlengths.

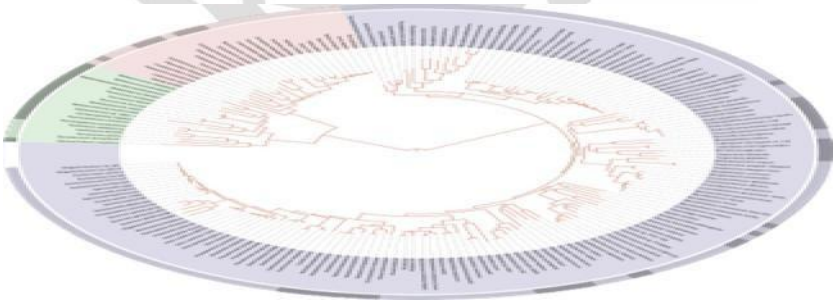
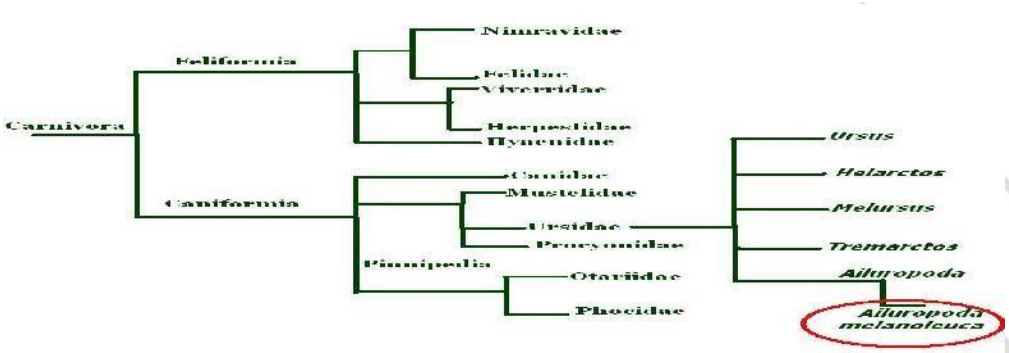
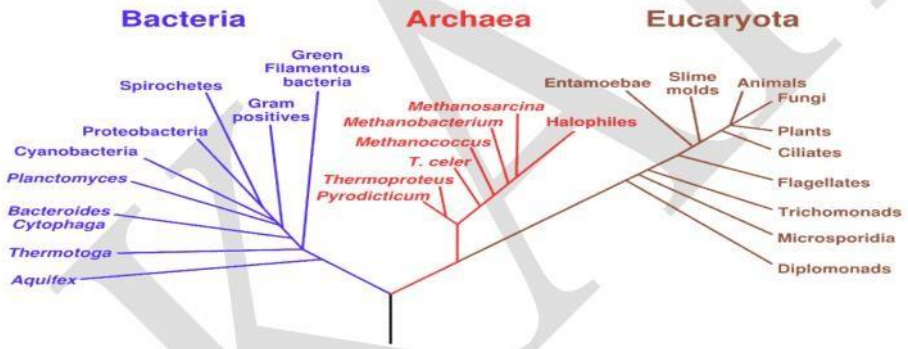


Fig. 2: A highly resolved, automatically generated Tree Of Life, **based on completely sequenced genomes.**

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.



Aphylogenetictree,showinghowEukaryotaandArchaeaaremorecloselyrelatedtoeach otherthantoBacteria,basedonCavalier-Smith'stheoryofbacterialevolution.



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 canalsobebuiltusingT-theory.

Limitations

- Itisimportanttorememberthattreesdohavelimitations.Forexample,treesare meanttoprovideinsightintoaresearchquestionandnotintendedtorepresentan entire specieshistory.
- Severalfactors,likegenetransfers,mayaffecttheoutputplacedintoatree.
- AllknowledgeoflimitationsrelatedtoDNAdegradationovertimemustbe considered,especiallyinthecaseofevolutionarytreesaimedatancientorextinct 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 thatthetreeisadditiveandthatallnodesareequallydistancefromtheroot.Sincea

"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 directlycalculated.

Steps for building a tree

1. Construct distancematrix
2. Cluster the two shortest distance OTUs into an internalnodes
3. Recalculate the distancematrix
4. Repeat the process until all OTUs are grouped in a singlecluster

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 commonancestor.

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 changessummedalongthebranches,becomesthemostparsimonioustreeandistaken

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 multiplealignment
- 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 finaltree

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 phylogenetictree.

PHYLIP

PHYLIPcomeswithanextensivesetofdocumentationfiles.Theseincludethemain documentationfile(thisone),whichyoushouldreadfairlycompletely.Inadditionthere arefilesforgroupsofprograms,includingonesforthemolecularsequenceprograms,the distancematrixprograms,thegenefrequencyandcontinuouscharactersprograms,the discretecharactersprograms,andthetreedrawingprograms.

Clique

Findsthelargestcliqueofmutuallycompatiblecharacters,andthephylogenywhichthey recommend,fordiscretecharacterdatawithtwostates.Thelargestclique(orallcliques withinagivensizerangeofthelargestone)arefoundbyaveryfastbranchandbound searchmethod.Themethoddoesnotallowformissingdata.ForsuchcasestheT (Threshold)optionofParsorMixmaybeausefulalternative.Compatibilitymethodsare particularusefulwhensomecharactersareofpoorqualityandtherestofgoodquality,but whenitisnotknowninadvancewhichonesarewhich.

Consense

Computesconsensustreesbythemajority-ruleconsensustreemethod,whichalsoallows onetoeasilyfindthestrictconsensustree.IsnotabletocomputetheAdamsconsensus tree.Treesareinputinatreefileinstandardnested-parenthesisnotation,whichis producedbymanyofthetreeestimationprogramsinthepackage.Thisprogramcanbe usedasthefinalstepindoingbootstrapanalysesformanyofthemethodsinthepackage. **Contml**

Estimatesphylogeniesfromgenefrequencydatabymaximumlikelihoodunderamodel inwhichalldivergenceisduetogeneticdriftintheabsenceofnewmutations.Doesnot

assumeamolecularclock.AnalternativemethodofanalyzingthisdataistocomputeNei's geneticdistanceanduseoneofthedistancematrixprograms.Thisprogramcanalsodo maximumlikelihoodanalysisofcontinuouscharactersthatevolvebyaBrownianMotion model,butitassumesthatthecharactersevolveatequalratesandinanuncorrelated fashion,sothatitdoesnottakeintoaccounttheusualcorrelationsofcharacters.

Contrast

Readsatreefromatreefile,andadatasetwithcontinuouscharactersdata,andproduces the independent contrasts for those characters, for use in any multivariate statistics package. Will also produce covariances, regressions and correlations between characters forthosecontrasts.Canalsocorrectforwithin-speciessamplingvariationwhenindividual phenotypesareavailablewithinapopulation.

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 whensitesvarygreatlyintheirratesofevolution,butwedonotknowinadvancewhich arethelessreliableones.

Dnadist

Computes four different distances between species from nucleic acid sequences. The distancescanthenbeusedinthedistancematrixprograms.Thedistancesarethe JukesCantorformula,onebasedonKimura's2-parametermethod,theF84modelusedin Dnaml,andtheLogDetdistance.Thedistancescanalsobecorrectedforgammadistributed andgamma-plus-invariant-sites-distributedratesofchangeindifferentsites.Ratesof evolution can vary among sites in a prespecified way, and also according to a Hidden Markovmodel.

Dnainvar

For nucleic acid sequence data on four species, computes Lake's andCavender's phylogeneticinvariants,whichtestalternativetreetopologies.Theprogramalsotabulates

thefrequenciesofoccurrenceofthedifferentnucleotidepatterns.Lake'sinvariantsarethe methodwhichhecalls"evolutionaryparsimony".

Dnaml

Estimates phylogenies from nucleotide sequences by maximum likelihood. Themodel employedallowsforunequalexpectedfrequenciesofthefournucleotides,forunequal ratesoftransitionsandtransversions,andfordifferent(prespecified)ratesofchangein differentcategoriesofsites,andalsouseofaHiddenMarkovmodelofrates,withthe programinferringwhichsiteshavewhichrates.Thisalsoallowsgamma-distributionand gamma-plus-invariantsitesdistributionsofratesacrosssites.

Dnamlk

SameasDnamlbutassumesamolecularclock.Theuseofthetwoprogramstogether permitsalikelihoodratiotestofthemolecularclockhypothesistobemade.

Dnamove

Interactive construction of phylogenies from nucleic acid sequences, with their evaluation byparsimonyandcompatibilityandthedisplayofreconstructedancestralbases.Thiscan beusedtofindparsimonyorcompatibilityestimatesbyhand.

Dnapars

Estimates phylogenies by the parsimony method using nucleic acid sequences. Allows use thefullIUBambiguitycodes,andestimatesancestralnucleotidestates.Gapstreatedasa fifthnucleotidestate.Itcanalsodotransversionparsimony.Cancopewithmultifurcations, reconstructancestralstates,use0/1characterweights,andinferbranchlengths.

Dnapenny

Finds all most parsimonious phylogenies for nucleic acid sequences by branch-and-bound search.Thismaynotbepractical(dependingonthedata)formorethan10-11speciesor so.

Dollop

Estimates phylogenies by the Dollo or polymorphism parsimony criteria for discrete characterdatawithtwostates(0and1).Alsoreconstructsancestralstatesandallows weightingofcharacters.Dolloparsimonyisparticularlyappropriateforrestrictionsites

data; with ancestor states specified as unknown it may be appropriate for restriction fragments data.

Dolmove

Interactiveconstructionofphylogeniesfromdiscretecharacterdatawithtwostates(0and 1) using the Dollo or polymorphism parsimony criteria. Evaluates parsimonyand compatibility criteria for those phylogenies and displays reconstructed states throughout thetree.Thiscanbeusedtofindparsimonyorcompatibilityestimatesbyhand.

Dolpenny

Findsallmostparsimoniousphylogeniesfordiscrete-characterdatawithtwostates,for the Dollo or polymorphism parsimony criteria using the branch-and-bound method of exactsearch.Maybeimpractical(dependingonthedata)formorethan10-11species.

Drawgram

Plotsrootedphylogenies,cladograms,circulartreesandphenogramsinawidevarietyof user-controllable formats. The program is interactive. It has an interface in the Java languagewhichgivesitacloselysimilarmenuonallthreemajoroperatingsystems.Final outputcanbetoafileformattedforoneofthedrawingprograms,foraray-tracingor VRMLbrowser,oroneatcanbesenttoalaserprinter(suchasPostscriptorPCL compatibleprinters),ongraphicsscreensorterminals,onpenplottersorondotmatrix printerscapableofgraphics.Manyoftheseformatsarehistoricsowenolongerhave hardwaretotestthem.Ifyoufindaproblempleasereportit.

Drawtree

SimilartoDrawgrambutplotsunrootedphylogenies.ItalsohasaJavainterfacefor previews.

Factor

Takes discrete multistate data with character state trees and produces the corresponding datasetwithtwostates(0and1).WrittenbyChristopherMeacham.Thisprogramwas formerlyusedtoaccomodatemultistatecharactersinMix,butthisislessnecessarynow thatParsisavailable.

Fitch

Estimatesphylogeniesfromdistancematrixdataunderthe"additivetreemodel"according to which the distances are expected to equal the sums of branch lengths betweenthe species.UsestheFitch-Margoliashcriterionandsomerelatedleastsquarescriteria,orthe MinimumEvolutiondistancematrixmethod.Doesnotassumeanevolutionaryclock.This programwillbeusefulwithdistancescomputedfrommolecularsequences,restriction sitesorfragmentsdistances,withDNAhybridizationmeasurements,andwithgenetic distancescomputedfromgenefrequencies.

Gendist

Computes one of three different genetic distance formulas from gene frequency data. The formulasareNei'sgeneticdistance,theCavalli-Sforzachordmeasure,andthegenetic distanceofReynoldset.al.Theformerisappropriatefordatainwhichnewmutations occurinaninfiniteisoallelesneutralmutationmodel,thelattertwoforamodelwithout mutationandwithpuregeneticdrift.Thedistancesarewrittentoafileinaformat appropriateforinputtothedistancematrixprograms.

Kitsch

Estimatesphylogeniesfromdistancematrixdataunderthe"ultrametric"modelwhichis thesameastheadditivetreemodelexceptthatanevolutionaryclockisassumed.The Fitch-Margoliashcriterionandotherleastsquarescriteria,ortheMinimumEvolution criterionarepossible.Thisprogramwillbeusefulwithdistancescomputedfrommolecular sequences,restrictionsitesorfragmentsdistances,withdistancesfromDNA hybridization measurements,andwithgeneticdistancescomputedfromgenefrequencies.

Mix

Estimatesphylogeniesbysomeparsimonymethodsfordiscretecharacterdatawithtwo states(0and1).AllowsuseoftheWagnerparsimonymethod,theCamin-Sokalparsimony method, or arbitrary mixtures of these. Also reconstructs ancestral states and allows weightingofcharacters(doesnotinferbranchlengths).

Move

Interactiveconstructionofphylogeniesfromdiscretecharacterdatawithtwostates(0and 1). Evaluates parsimony and compatibility criteria for those phylogenies anddisplays reconstructed states throughout the tree. This can be used to find parsimony or compatibilityestimatesbyhand.

Neighbor

AnimplementationbyMaryKuhnerandJohnYamatoofSaitouandNei's"NeighborJoining Method,"andoftheUPGMA(AverageLinkageclustering)method.NeighborJoiningisa distancematrixmethodproducinganunrootedtreewithouttheassumptionofaclock. UPGMAdoesassumeaclock.Thebranchlengthsarenotoptimizedbytheleastsquares criterionbutthemethodsareveryfastandthuscanhandlemuchlargerdatasets.

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

Findsallmostparsimoniousphylogeniesfordiscrete-characterdatawithtwostates,for the Wagner, Camin-Sokal, and mixed parsimony criteria using the branch-and-bound methodofexactsearch.Maybeimpractical(dependingonthedata)formorethan10-11species.

Proml

Estimatesphylogeniesfromproteinaminoacidsequencesbymaximumlikelihood.The PAM,JTT,orPMBmodelscanbeemployed,andalsouseofaHiddenMarkovmodelof rates,withtheprograminferringwhichsiteshavewhichrates.Thisalsoallowsgamma distribution and gamma-plus-invariant sites distributions of rates across sites. It also allowsdifferentratesofchangeatknownsites.

Promlk

Same as Proml but assumes a molecular clock. The use of the two programstogether permitsalikelihoodratiotestofthemolecularclockhypothesistobemade.

Protdist

Computes a distance measure for protein sequences, using maximum likelihood estimates basedontheDayhoffPAMmatrix,theJTTmatrixmodel,thePBMmodel,Kimura's1983 approximationtothese,oramodelbasedonthegeneticcodeplusaconstraintonchanging toadifferentcategoryofaminoacid.Thedistancescanalsobecorrectedforgamma- distributed and gamma-plus-invariant-sites-distributed rates of change in different sites. Ratesofevolutioncanvaryamongsitesinaprespecifiedway,andalsoaccordingtoa HiddenMarkovmodel.Theprogramcanalsomakeatableofpercentagesimilarityamong sequences.Thedistancescanbeusedinthedistancematrixprograms.

Protpars

Estimatesphylogeniesfromproteinsequences(inputusingthestandardone-lettercode foraminoacids)usingtheparsimonymethod,inavariantwhichcountsonlythose nucleotidechangesthatchangetheaminoacid,ontheassumptionthatsilentchangesare moreeasilyaccomplished.percentagesimilarityamongsequences.

Restdist

Distancescalculatedfromrestrictionsitesdataorrestrictionfragmentsdata.The restrictionsitesoptionistheonetousetoalsomakedistancesforRAPDsorAFLPs.

Restml

Estimationofphylogeniesbymaximumlikelihoodusingrestrictionsitesdata(not restrictionfragmentsbutpresence/absenceofindividualsites).ItemploystheJukesCantor symmetrical model of nucleotide change, which does not allow for differences of rate betweentransitionsandtransversions.Thisprogramisveryslow.

Retree

Readsinatree(withbranchlengthsifnecessary)andallowsyoutorerootthetree,toflip branches,tochangespeciesnamesandbranchlengths,andthenwritetheresultout.Can beusedtoconvertbetweenrootedandunrootedtrees,andtowritethetreeintoa preliminaryversionofanewXMLtreefileformatwhichisunderdevelopmentandwhich isdescribedintheRetreedocumentationwebpage.

Seqboot

Readsinadataset,andproducesmultipledatasetsfromitbybootstrapresampling.Since mostprogramsinthecurrentversionofthepackageallowprocessingofmultipledatasets, thiscanbeusedtogetherwiththeconsensustreeprogramConsensetodobootstrap(or delete-half-jackknife)analyseswithmostofthemethodsinthispackage.Thisprogramalso allowstheArchie/Faithtechniqueofpermutationofspecieswithincharacters.Itcanalso rewriteadatasettoconvertitfrombetweenthePHYLIPInterleavedandSequentialforms, andintoapreliminaryversionofanewXMLsequencealignmentformatwhichisunder developmentandwhichisdescribedintheSeqbootdocumentationwebpage.

Threshml

Readsatreefromatreefile,andadatasetwithdiscrete0/1characters.Usingthe thresholdmodelofquantitativegenetics,theprogramrunsaMarkovChainMonteCarlo (MCMC)samplertosampletheunderlyingcontinuouscharacters(theliabilities)thatcause thediscretecharacters.Thecovariancesoftheliabilitiesareestimated,aswellasthe transformationfromtheliabilitiestounderlyingindependentlyevolvingcharacters.

Treedist

ComputestheBranchScoredistancebetweentrees,whichallowsfordifferencesintree topologyandwhichalsomakesuseofbranchlengths.Alsocomputesanotherdistanceby RobinsonandFouldsthatusesbranchlengths,andtheSymmetricDifferencedistance between trees, which allows for differences in tree topology but does not use branch lengths.

Possible questions

1. Whatissequencealignment?Explainindetail.
2. Explainthestepsinconstructingaphylogenetictree.
3. Differentiatethetblastnandtblastx
4. IllustrateaboutthePHYLIPpackage
5. Writeaboutthefollowing
    a. BLAST
    b. FASTA

c. ClustalW

d. RasMol

6. Writeindetailaboutdifferenttypesof BLASTandtheirsignificance
7. Whatisphylogenticstree?Explainabouttheirtypesandterminologies.
8. Differentiatelocalandglobalalignment.
9. Mentionthesignificanceandmethodologyinvolvedinmultiplealignment.
10. Writeshortnotesontreebuildingmethods.
11. Definemultiplesequencealignment?WriteshortnotesonClustalW.

KAPRAGAM ACADEM YOF HIGHER EDUCATION

DEPARTMENT OF BIOCHEMISTRY

II- B.Sc Biochemistry

18BCU204-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	ploryphyletic	**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-

IIISYLLAB

US

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:fromtheGreekwordPROTEUOwhichmeans"tobefirst(inrankor influence)"
- Whyareproteinsimportanttous:

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

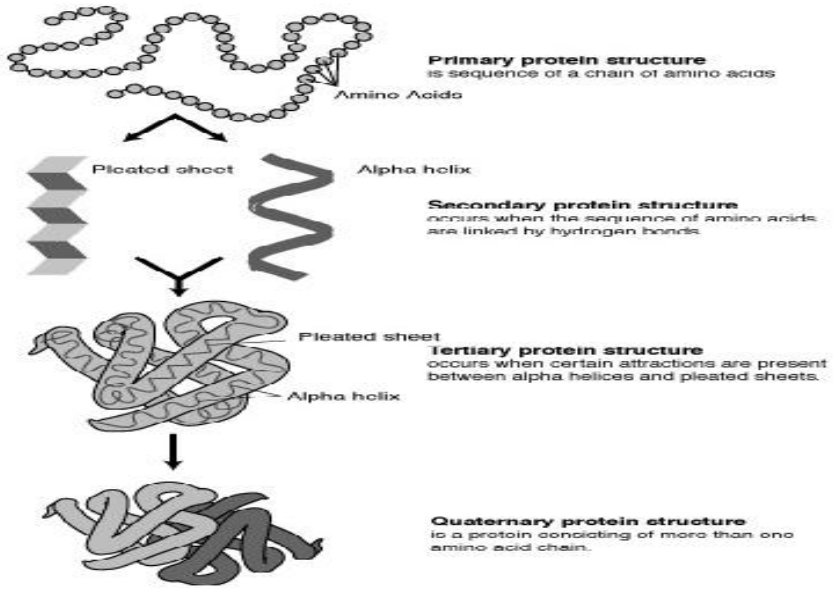
  Enzyme–actsasabiologicalcatalyst

  Storageandtransport–Haemoglobin

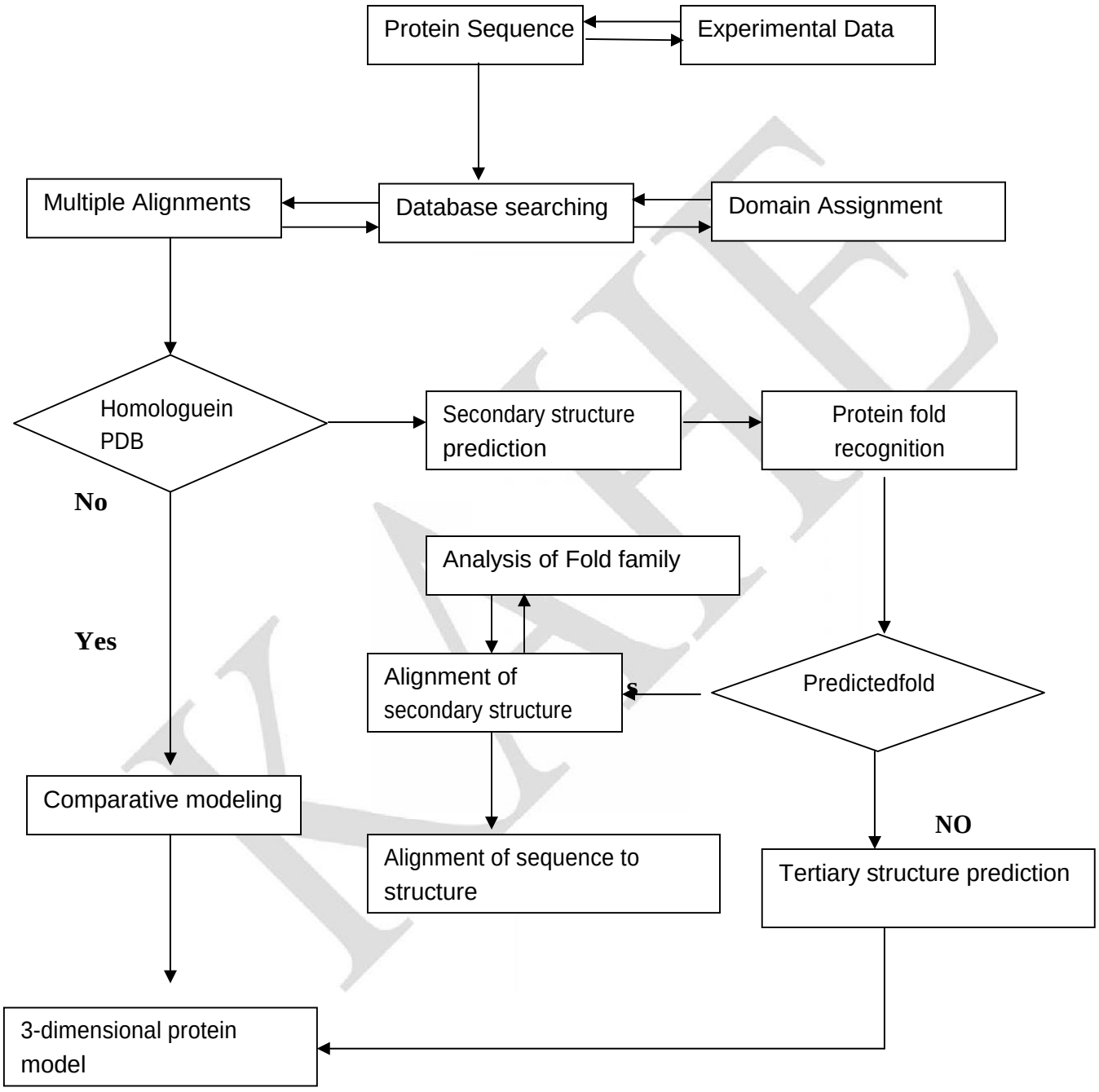
  Antibodies

  Hormones – Insulin

Four levels of protein structure



Flow chart showing steps involved in prediction protein structure



Protein Secondary Structure

- Secondarystructureisthetermproteinchemistgivetothearrangementofthe peptidebackboneinspace.Itisproducedbyhydrogenbondingsbetween aminoacids
- TheassignmentoftheSScategoriestotheexperimentallydeterminedthree-dimensional(3D)structureofproteinsisanon-trivialprocessandistypically performedbywidelyusedDSSPprogram
- PROTEIN SECONDARY STRUCTURE consists of : protein sequence and its hydrogenbondingpatternscalledSScategories
- Databases for protein sequences are expanding rapidly due to the genome sequencingprojectsandthegapbetweenthenumberofdeterminedprotein structures(PSS–proteinsecondarystructures)andthenumberofknown proteinsequencesinpublic
- Proteindatabanks(PDB)isgrowingbigger.
- PSSP (Protein Secondary Structure Prediction) research is trying to breachthis gap.

Early methods for Secondary Structure Prediction

Chou and Fasman

- Startbycomputingaminoacidspropensitiestobelongtoagiventypeof secondarystructure:

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

Propensities>1meanthattheresiduetypeIislikelytobefoundinthe

Correspondingsecondarystructuretype.

Predicting Alpha helices

- findnucleationsite:4outof6contiguousresidueswithP(a)>1
- extension:extendhelixinbothdirectionsuntilasetof4contiguous residues hasan

average P(a) < 1 (breaker)

- ifaverageP(a)overwholeregionis>1,itispredictedtobehelical
- a Spiralshape.

Predicting Beta strands

- findnucleationsite:3outof5contiguousresidueswithP(b)>1
- extension:extendstrandinbothdirectionsuntilasetof4contiguous
- residueshasanaverageP(b)<1(breaker)
- ifaverageP(b)overwholeregionis>1,itispredictedtobeastrand

Predicting turns

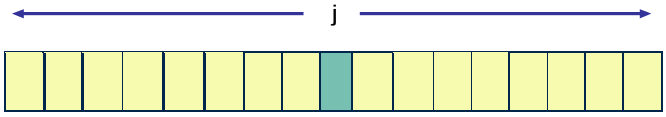
- foreachtetrapeptidestartingatresidueI,compute:
- Pturn(averagepropensityoverall4residues)
- F =f(i)*f(i+1)*f(i+2)*f(i+3)
- ifPturn>PaandPturn>PbandPturn>1andF>0.000075
- tetrapeptideisconsideredaturn.

Random Coils

- predictionofsecondarystructureofproteinthoughdifficultbutisimportantfor mainly tworeasons.
- Functionalpropertiesofproteinsdependupontheir3Dstructure.
- Duetorelationshipsbetweenthewaysaminoacidsequencearrangementandtheir correspondingstructure.
- Nodefiniteruleoranalgorithmwhichcharacterizesthisrelationships.

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

- BothChouandFasmanandGORhavebeenassessedandtheiraccuracyisestimated to beQ3=60-65%.
- (initially,higherscoreswerereported,buttheexperimentssettomeasureQ3were flawed,asthetestcasesincludedproteinsusedtoderivethepropensities!)

Protein tertiary structure prediction Methods

Thebiologicalroleofaproteinisdeterminedbyitsfunction,whichisinturnlargely determinedbyitsstructure.Thusthereisenormousbenefitinknowingthethree dimensionalstructuresofalltheproteins.Althoughmoreandmorestructuresare determinedexperimentallyatanacceleratedrate,itissimplynotpossibletodetermineall the protein structures from experiments. As more and more protein sequences are determined,thereispressingneedforpredictingproteinstructurescomputationally. Decadesofintenseresearchinthisareabroughtabouthugeprogressinourabilityto predictproteinstructuresfromsequencesonly.

Theproteinstructurepredictionmethodscanbebroadlydividedintothreecategories:

- Homologymodeling,
- Threadingorfoldrecognition,and

- AbInitio

Essentially,theclassificationreflectsthedegreetowhichdifferentmethodsutilizethe informationcontentavailablefromtheknownstructuredatabase.

Comparative homology modeling

So far protein prediction methods based on homology have been the most successful.Homologymodelingisbasedonthenotionthatnewproteinsevolvegradually fromexistingonesbyaminoacidsubstitution,addition,and/ordeletionandthatthe3D structures and functions are often strongly conserved during this process. Many proteins thussharesimilarfunctionsandstructuresandthereareusuallystrongsequence similaritiesamongthestructurallysimilarproteins.Strongsequencesimilarityoften indicates strong structure similarity, although the opposite is not necessarily true. Homology modeling tries to identify structures similar to the target protein through sequencecomparison.Thequalityofhomologymodelingdependsonwhethertheseexists oneormoreproteinstructuresintheproteinstructuredatabasesthatshowsignificant sequencesimilaritytothetargetsequence.

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

- Identifyoneormoresuitablestructuraltemplatesfromtheknownproteinstructure databases;
- Alignthetargetsequencetothestructuraltemplate;
- Buildthebackbonefromthealignment,includingtheloopregionandanyregion thatissignificantlydifferentfromthetemplate;and
- Place theside-chains.

Thefirsttwosteps,identificationofstructuraltemplatesandalignmentofthetarget sequence onto the parent structures, are usually related. Sequence comparison methods determinesequencesimilaritybyaligningthesequencesoptimally.Thealignedresiduals ofthestructuretemplatesareusedtoconstructthestructuralmodelinthesecondstep. Thequalityofthesequencecomparisonthusnotonlydetermineswhetherasuitable structuraltemplatecanbefoundbutalsothequalityofthealignmentbetweenthetarget

sequenceandtheparentstructure,whichinturndeterminestheaccuracyofthestructural model.Ofcriticalimportanceistheabilityforthesequencecomparisontodetectremote homologuesandtocorrectlyalignthetargetsequencetoandparentstructure.Inthe followingIdiscussthevarioussequencecomparisonmethodsinrelationtohomology modelingandtheirrangeofapplicability,accuracyandshortcomings.Forcomparative modeling, local sequence comparison methods are usually used since the sequence similarityismostlikelyoversegmentsofthetwosequences.Thelocalsequence comparisoncaneitherbepairwiseorprofilebased.Pairwisecomparisons,suchasthe widelyusedBLASTintheearlydays,candetectsequencesimilaritiesbetterthan30%.A numberoftoolshavealsobeendevelopedtodetectweakhomologyrelationships. Methods like profile and HMM use a statistical profile of a protein family. Tofurther increase the chance of detecting remote homologues, PSI-BLAST and SAM-T98 build the profileorHMMbysearchingthedatabaseiterativelyuntilnonewhitsarefound.Methods suchasPSI-BLASTencodetheinformationaboutawholeproteinfamilyforthetarget sequenceinamodeltoincreasethechanceofdetectingremotehomologies.Tofurther increasethedetectionsensitivity,thesequencesinthestructuredatabasecanalsobe encodedinprofiles.Thisformsthebasisoftheprofile-profilebasedcomparisonmethods. Withlowsequenceidentities((<20%),profile-profilemethodsclearlyoutperformthe other two kinds of methods: profile-profile methods identified more than 90%of homologous pairs, determined from structure-structure similarity comparison, with sequenceidentitybetterthan10%andanimpressive38%evenforcaseswithsequence identitiesbetween5%and9%.

Thestructuremodelsareconstructedfromtheresidualsofthestructuretemplate thatarealignedtothetargetsequenceinthesequencecomparison.Thequalityofthis alignmentthusiscriticalfortheaccuracyachievable.Thealignedresiduesfromsequence comparisonaregenerallydifferentfromthatfromstructure-structurecomparisonthough, especiallywhenthesequenceidentityislow.Toassesstheabilityofthesequence comparisonmethodstoalignthesequencescorrectly,itisinstructivetocomparethe sequence-sequencealignmenttothestructure-structurealignmentofthesamepairof

proteins. To determine how well the different similarity search methods can detect remotehomologiesandassesstheirabilityincorrectlyaligningthesequences,compared varioussequencealignmentmethodstotheCEstructurealignmentoftheSCOPprotein structures.Forsequenceidentitieslessthan30%,profile-basedcomparisonmethods, suchasPSI-BLASTandprofile-profilecomparison,areallobviouslybetterthanthepair wiseBLASTmethod.Forexample,at10-15%sequenceidentity,BLASTalignsonly20% correctlywhilePSIBLASTandprofile-profilecomparisoncancorrectlyalign40%and48% respectively.Thisalsoindicatesthatthereisstilllargeroomforimprovementincorrectly aligningthetargetsequencetothetargetstructure.Oneindicationoftheaccuracyof comparativemodelingisthesequenceidentitybetweenthetargetandthetemplate.Itis believedthatiftwoproteinsequenceshave50%orhighersequenceidentity,thenthe RMSDofthealignablepotionbetweenthetwostructureswillnormallybelessthan1.In the so-called "twilight zone", with sequence identity between 20%~30%, 95% of the sequenceswiththislevelofidentityhavedifferentstructuresthough.Whenastructure templatecanindeedbefoundwithintheknownproteinstructuredatabasesinsuchcases, thebackboneRMSDcanbeexpectedtobenobetterthan2.Structurallysimilarproteins canhavelowsequenceidentitiesinthe8~10%rangeandcanstillbeidentifiedwith sensitiveprofile-profilebasedcomparison,buttheRMSDcanbeaslargeas3~6.Theerror largelycomesfromthemisalignmentfromsequencecomparison.Atsuchlowsequence identity,comparisonmethodthatcandetecttheremotehomologyaswellasalignthe sequencesclosetotheoptimalfromstructurealignmentwillbedesirable.

Threading or fold recognition:

For evolutionally remotely related proteins, even if the sequence similarity is difficulttodetectwithsequencecomparisonmethods,therecouldstillbeidentifiable structuralsimilarity.Structurealignmentshavebeenshowntobeabletoidentify homologous protein pairs with sequence similarities less than 10%. When sequence comparisonbasedmethodsarenolongersensitiveenoughtorecognizethecorrectfoldfor thetargetsequence,foldrecognitionorthreadingcanstillbeusedtoassignthecorrectfold tothetargetsequence.Threadingorfoldrecognitionisthemethodbywhichalibraryof

uniqueorrepresentativestructuresissearchedforstructureanalogstothetarget sequence,andisbasedonthetheorythattheremaybeonlyalimitednumberofdistinct proteinfolds.Forexample,inanearlypaper,Chothiapostulatedthatthenumberofunique proteinfoldswouldbeontheorderofonlyabout1000uniqueproteinfolds.Inanother estimation, the number of distinct domains and folds were placed around 7000. Even thoughthenumberofnewstructuressolvedhasbeenincreasingatanacceleratedrate (closeto3000structuressolvedin2002),theproportionofnewfolds,asdeterminedby theCEalgorithm(http://cl.sdsc.edu/ce.html),tothetotalnumberofnewstructuressolved inagivenyeardecreasedfromanaverageofca.30%inthe80'ssteadilydowntoonlyca. 8%inyear2001(http://www.rcsb.org/pdb/holdings.html).Itisreasonabletoexpectthat asmoreandmoreproteinstructuresaredeterminedexperimentally,wewillbeabletofind closestructureanaloguesinthedatabasesofknownstructuresforalmostanyprotein sequenceinthenearfuture.

Threadingorfoldrecognitioninvolvessimilarstepsascomparativemodeling.The differenceisinthefoldidentificationstep.Firstofall,astructurelibraryneedstobe defined.Thelibrarycanincludewholechains,domains,orevenconservedproteincores. Oncethelibraryisdefined,thetargetsequencewillbefittedtoeachlibraryentryanda energyfunctionisusedtoevaluatethefitbetweenthetargetsequenceandthelibrary entriestodeterminethebestpossibletemplates.Dependingonthealgorithmstoalignthe targetsequencewiththefoldsandtheenergyfunctionstodeterminethebestfits,the threadingmethodscanroughlybedividedintofourclasses.

- Theearliestthreadingmethodsusedtheenvironmentofeachresidueinthe structureastheenergyfunctionanddynamicalprogrammingtoevaluatethefitand thealignment.
- Instead of using overly simplified residual environment as the energyfunction, statisticallyderivedpairwiseinteractionpotentialsbetweenresiduepairsoratom pairscanbeusedtoevaluatethebestpossiblefitsbetweenthetargetsequenceand libraryfolds.Inthismethod,forefficientoptimalalignmentbetweenthetarget sequenceandthefolds,thepotentialforresidualisobtainedbysummingoverall

thepairwisepotentialsinvolvingi,andthen"double dynamicalprogramming" methodcanbeused.

- Thethirdkindofmethodsdoesnotuseanyexplicitenergyfunctionatall.Instead, secondarystructuresandaccessibilityofeachresiduearepredictedfirstandthe targetsequenceandlibraryfoldsareencodedintostringsforthepurposeof sequence-structurealignment.
- Finally,sequencesimilarityandthreadingcanbecombinedforfoldrecognition.For large-scalegenomewiseproteinstructureprediction,sequencesimilaritycanbe firstusedfortheinitialalignmentsandthealignmentscanbeevaluatedby threadingmethods.

Thethreadingmethodsarelimitedbythehighcomputationalcostsinceeachentry inthewholelibraryofthousandsofpossiblefoldsneedstobealignedinallpossibleways toselectthefold(s).Anothermajorbottleneckistheenergyfunctionusedforthe evaluation of the alignment. As these functions are drastically simplified for efficient evaluation,itisnotreasonabletoexpecttobeabletofindthecorrectfoldsinallcaseswith asingleformofenergyfunction.Nevertheless,withthecurrentfunctions,itispossibleto reducethethousandsofpossiblefoldstoonlyafew.Similartothecomparativemodeling case,forsequencesimilaritiesatproteinfamilylevel,threadingcanproducealignments thatareaccurateto1to3_,orinthecasewithlowsequencesimilarityatthesuper-family level,alignmentattherangeof3to6_canstillbeexpected.Asmoreproteinstructuresare determined and sequence comparison methods improve, more and more targetsequences foldassignmentcanbeachievedbycomparativemodelingthough.Worthmentioningisthe threadingprogramPROSPECT,whichperformedbestinitscategoryintheCASP4 competition.WhatisuniquetoPROSPECTisthatitisdesignedtofindthegloballyoptimal 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 theconformationsearchspacethatisshownnottocontaintheoptimalalignment.Inseveral casesthathavesequenceidentityaslowas17%,perfectsequence-structurealignmentis stillachievedforthealignablepotionsbetweenthetargetandtemplatestructures.Evenin

casesthatnofoldtemplatesexistforthetargetsequence,importantfeaturesofthe structurearestillrecognizedthroughthreadingthetargetsequencetothestructures.

Protein folding

Aproteinisapolymeroffixedlength,compositionandstructure,madeupbya combinationofthe20naturallyoccurringaminoacids.With20aminoacidsitispossibleto generate20differentchainsof200aminoacidseach.Onlyasmallfractionisactuallyused bylivingorganisms.Outofallthepossiblesequencesofaminoacids,proteinshavebeen selectedbyevolutionover10^6-10^7ofyearstoperformaspecificbiologicaltask.

The role of protein folding

3Dstructureofeachproteinarisesfromfoldingontoaspecificuniqueconformationwitha particularfunction.

Aproteinmadeintheribosomeasapolypeptidechainmustevolvetoreachitsmorestable conformation-thiscantakebetween1toseveralminutes.

From sequence to 3D structure

3Dstructureofeachproteinarisesfromfoldingintoaspecificuniqueconformationwitha particularfunction.

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-mercaptoethanolisreducingagentthatbreaksdisulphidebridgesinproteins Ureadisruptsnon-covalentinteractions

Denaturation leads to complete inactivation of RNase

Dialysiswasusedtoremoveureaandairtore-oxidiseprotein

RNaserecoversallactivity

Correct tertiary structure of RNase backbone was recovered

RightSHgroupsmusthavebeenadjacenttoeachotherpriortore-oxidationuponcorrect refolding of backbone because disulphide bridges formed spontaneously with the correct combinationofCysaminoacids.

Anfinsen'sDogma

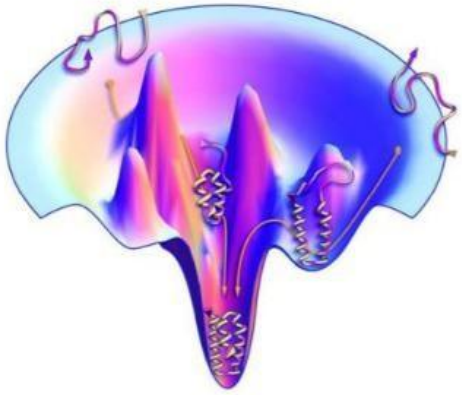
Nativestructureisdeterminedonlybytheaminoacidsequenceofaprotein,atleastfour globularproteins:

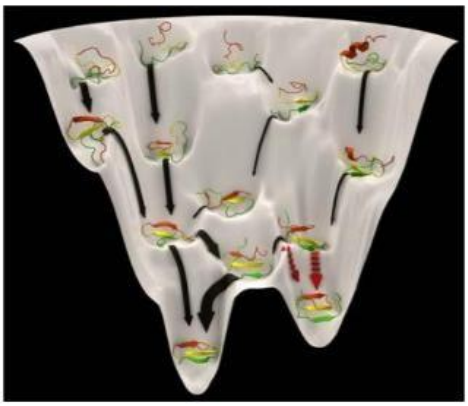
Uniqueness:nativestructureisthethermodynamicallymoststable(favoured)andthus unique

Stability:Smallchangesinsurroundingenvironmentdonotaffectfreeenergyminimum configuration

Kineticaccessibility:pathintheconformationalfreeenergysurfacemustbesmooth.

Freeenergylandscapeofproteinfolding





The Levinthal Paradox

Ifwetakeaprotein150aminoacidslongandassumethatifonlyhastwomainchain torsionaldegreesoffreedomperaminoacid.

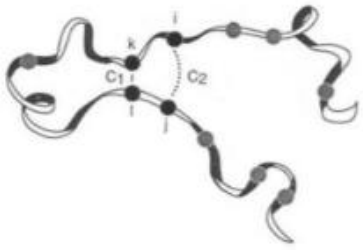
Proteins do not randomly search all possible conformations until they reach the most stable structure.

Cooperatively in protein folding

Aproteinwillreachanoptimalconformationwithoutactuallyundertakingaglobal conformationalsearch.

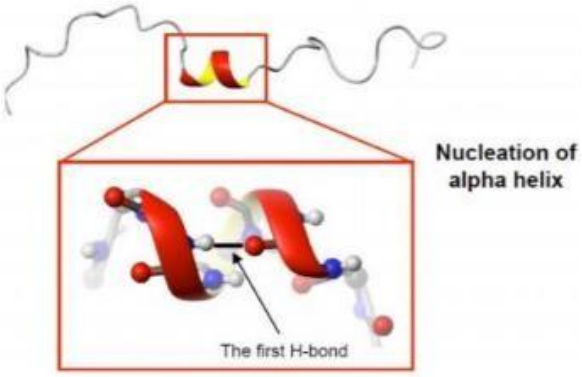
Depends on physic-chemical conditions: pH, temperature, ionic strength, redox potential

CooperativityisessentialprobabilityofformingcontactC2ismuchhigherifC1isformed thatintheabsenceofC1.

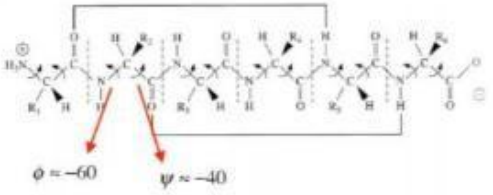


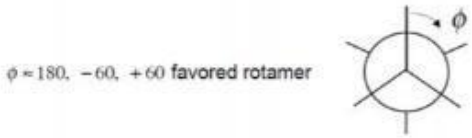
TheCoil-HelixTransition

TheParadigmforcooperativityinproteinfolding

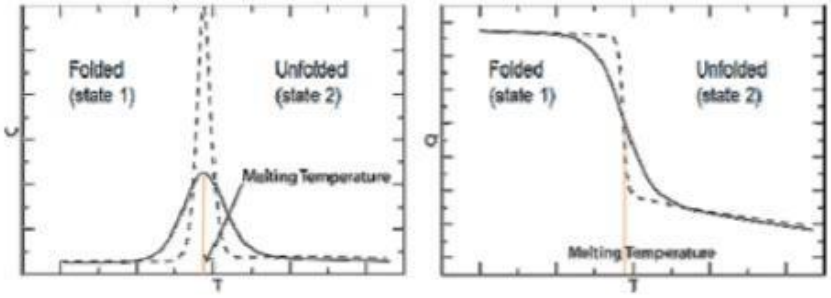


Initiation of a helical turn is much harder than appending another residue to a helical segment, due to the higher entropy cost





Cooperative transitions have a sigmoidal profile



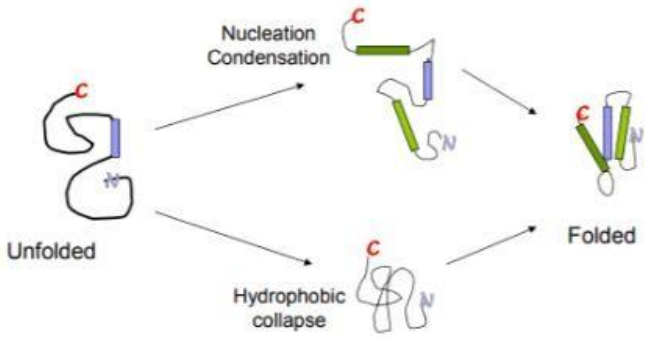
Melting temperature at which unfolding occurs coincidence with a rapid decline in the proportion of folded protein with respect to unfolded protein.

Mechanisms of Protein folding

Two over all mechanisms have been proposed:

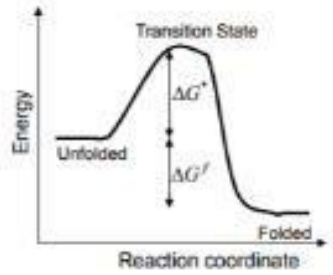
Nucleation-condensation: Somesecondarystructuremotifsareformedandactas templatefortheformationoftertiarystructure

Hydrophobiccollapse: Hydrophobicinteractionsproduceacompactstructure(molten globule)thatsubsequentlyfoldsintoitsfinalstate.



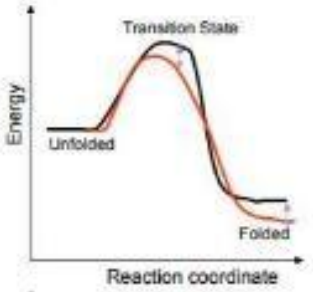
Characterisation of Folding States

Allresiduesinaproteinaremutatedto Alaonebyoneandthestabilitytowards denaturationandfoldingrateofeachmutantwithrespecttothewildtypesequencecanbe measured. The relative free energies of the folded, unfolded and transition state are calculated.



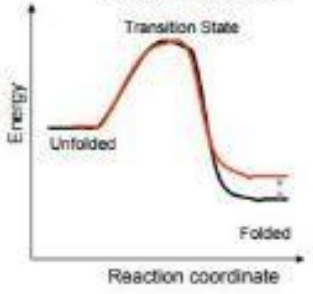
Φ values for each amino acid *i* are calculated as

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



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

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.



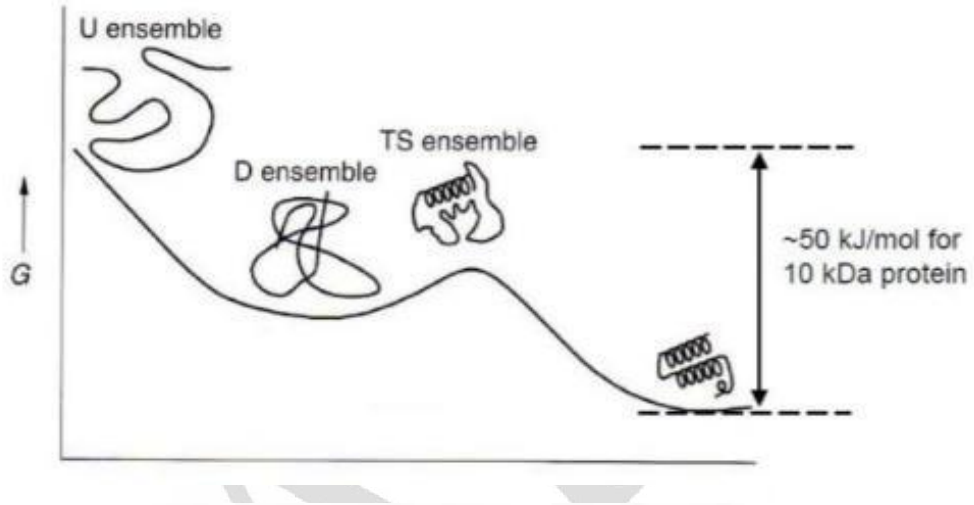
A Φ value close to 0 indicates that the mutation has no effect on the transition state. This can be interpreted as the residue being in an unfolded conformation at the transition state.

φvaluesofsmallproteinindicatethatα-helicesareoftenformedbeforethetransition statewhiletheformationofβ-sheetsisrate-limiting.

Proteinwithhighα-helixcontentfoldmuchfasterthanproteinswithlargeβ-sheetcontent, although there are some exceptions These finding suggest anucleation-condensation mechanism,atleastforsmallproteins.

More on Thermodynamics of Folding

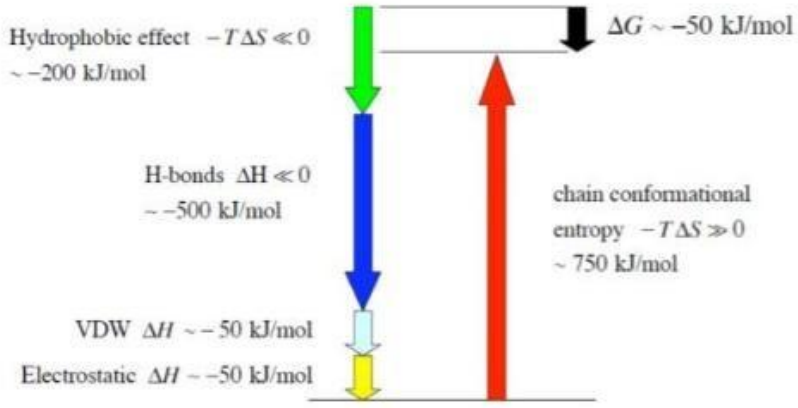
Simplifies free energy profile of protein folding:



Dissecting the Free Energy of Folding

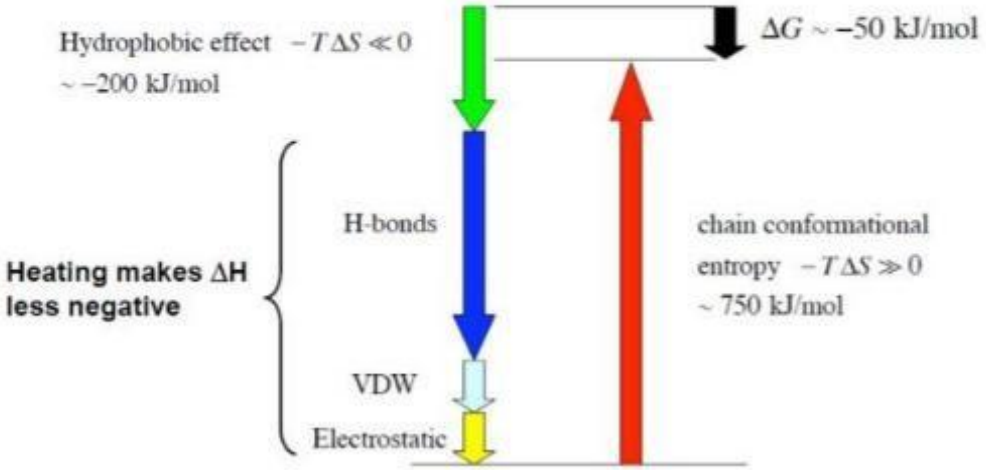
$$\text{Unfolded} \overset{\Delta G}{\rightleftharpoons} \text{Folded}$$

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



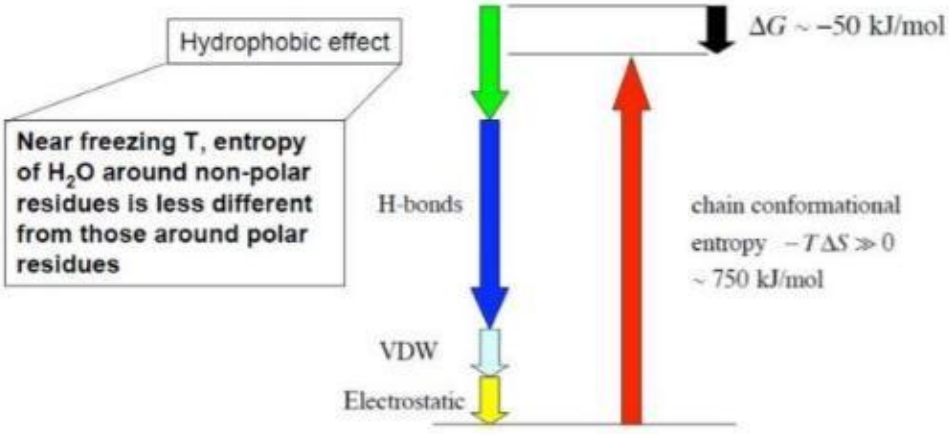
Heat Denaturation Breaks Interactions

$$\Delta G = \Delta H - T\Delta S < 0$$



Cold Denaturation Reduces Hydrophobic Effect

$$\Delta G = \Delta H - T\Delta S < 0$$



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

- Moleculesareusedtwo-dimensional(2D)structureand3Dstructure.
- Mostlythemoleculeswithinteractingthreedimensions.
- No.oftoolsareavailableforeg.Rotate,flipandotherwisemanipulatevirtual molecularmodelsofchemicalsandmacromolecules.
- Smallmoleculesin3-Ddownloadandinstallonyourcomputer.
- ScientificwebsitesareJMol,Java-basedviewerforrenderingmoleculesin3-D.
- Chimeistheprogramusedmostforviewingsmallmoleculesfromwebsites.
- Macromoleculesin3Ddownloadandinstall.

CHIME

- Chimeisafreedownloadable
- ItschemicalstructurevisualizationPlug-inwindowsandMacintosh.
- Itallowsviewchemicalstructurefromwithinpopularwebbrowsers Javaapplets and Javaapplications**.**
- Chimealreadybeinstalledfortheirgraphicstoworkproperly.
- Rasmol,Chimeshowsmoleculeswithinawebpage.

Cn3D

- Cn3Disavisualizationtoolformacromolecules.
- Toview3-dimensionalstructurefromNCBI'sEntrez→it'saretrievalservice.
- Cn3Disabletocorrelatestructureandsequenceinformation.
    - **Example**:findtheresiduesinacrystalstructurethatcorrespondtoknown
    - diseasemutations.
    - =powerfulannotationsandeditingfeatures.
- =rightclickonthemoleculetoseetheviewingoptions.

RASMOL

- It's freeprogram

- Developed by Roger A Sayle (1993) University of Edinburgh's, Biocomputing ResearchunitandtheBiomolecularstructureDepartmentatGlaxoResearchand DevelopmentGreenFord,UK.
- Rasmolderivedfrom**Raster**(thearrayofpixelsonacomputerscreen)molecules.
- Moleculargraphicsprogramforvisualizationofproteins,nucleicacidsandsmall molecules.
- Powerful program aimed at display, teaching and generation of publication quality image.
- RasmolreadsinmolecularCo-ordinatefilesinformatslikeBrookhavenProtein Databanks(PDB).
- Differentpartsofthemoleculesdisplayedandcoloredindependentlyrestofthe moleculeorshowindifferentrepresentationssimultaneously.
- Molecule may be shown Wire frame, cylinder(deriding), stick bonds, alpha-carbon trace, space filling (CPK) spheres, macromolecular ribbons (either smooth shaded solidribbonsorparallelstrands)hydrogenbondinganddotsurface.
- Molecule displayed may be rotated, translated, zoomed, z-clipped (slabbed) interactivelyusingeitherthemouse,thescrollbars,commandlineoranattached dialsbox.
- Modelcanberotatedaboutthex,yandzaxesinteractivelysothatallpartsofthe moleculecanbestudied.
- Smallermoleculeorlayerandinadditionitispossibletoexpandtheviewing windowuptothefullsizeofthescreen.
- Largerpicturemoreelaboratethemodel,thelongerittakesthecomputerto calculatetheappearanceofthedrawing.

Color schemes are available

CPK

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

Group colouring

- Chaincoloredwithcoloroftherainbow.
- Blue –N-terminus
- Red –C-terminus
- Usefulforfollowingthefoldfromoneendofthechaintotheother.
- Shapelyandaminocolours
- Backbone–Palegraysidechainatomsareallgivenacolourdependsuponthesize andthepolarityofthesidechain.
- Oxygencontainingsidechain(acids,amidesandthehydroxy-aminoacidsSerand Thr.
- VariousshadeofredandbasicsidechainBlue(Arg,Lus,His)
- HydrophobicaminoacidsaremostlyGrey;IleisdarkgreenandValaPaleMagenta.
- Sulphurcontainingaminoacids(Cys,Met)havemuddyyellowcolors.
- TrpisyellowandGrey,White

Residue colour list

A(ala)–Palegreen
C(cys)–Sandyyellow
D(Asp)–darkmagenta
E(glu)–red
F(phe)–grey
G(gly)–White
H(his)–slateblue
I(ile)–darkgreen
K(lys)–royalblue
W(trp)–yellow
L(leu)-grey
M(met)–palebrown
N (asn) -salmon
P(pro)-grey
Q(gln)–flashpink
R(arg)–navyblue
S(ser)-tomato
T(thr)–orangered
V(val)–palemagenta
Y(tyr)–claygrey

- Rasmolpreparedbylistofcommandsfromascriptfile
- Rasmolworkswellforbothsmallmoleculesandforlargeonessuchasproteins, DNA,RNA.

Protein Explorer a Rasmol derivatives:

- Proteinexplorer(PE)enablestoexplorethe3Dstructureofanymacromolecule.
- Proteins,DNA,RAN,carbohydratesandcomplexessuchasbetweentranscriptional regulatoryexplorers.
- ItisnotcompatiblewithInternetExplorer
- Fireboxfreeandisrecommendedforproteinexplorer.

Biomodel – 3:

- DevelopedbyAngelHerraez,LecturerinBiochemistryandmolecularBiologyatthe UniversityofAlcaladeHenares(Spain)
- VersionV3
- UseJ_{MOL},Javaapplettoshowmanipulatesthemolecularmodels.

3D – Chemical libraries

- Use chimeplug-in

3-D Virtual ChemistryLibrary

- Moleculardatabasehasabout150moleculesdividedintosixmaingroups.
- Simplemolecule
- Polymers
- Senses
- Medical
- Horrible moleculeand
- Interestingmolecules
- Iadditiontostructureitalsohasphysicaldata,historyandreactivityofthe molecules.

3D Macromolecular structures

Using Cn3D

Entrez molecular modeling Databases

- Itcontains3-Dmacromolecularstructure
- Includingproteinsandpolynucleotide

- MMDBcontainover40,000structuresandlinkedtotherestoftheNCBIdatabase
- Includingsequencesbibliographiccitations,taxonomicclassificationsandsequence and structureneighbors.

Possible Questions

1. Writeanaccountonproteinstructurepredictionwithsuitableexamples.
2. HowwillyouvisualizethemoleculesusingRasMolttol.
3. Explainthestepsinvolvedinhomologymodelling.
4. Discussindetailaboutthethreadingmethod.
5. Definecomparativemodelling?Explainthestepsinvolvedinmodellingthethree dimensionalstructureoftheprotein.
6. Writeanaccountonproteinfolding.
7. Howwillyouconstructa3Dproteinmodelandvalidateit?Explaininastepwisemanner.
8. HowwillyouvisualizethemoleculesusingDeepViewtool
9. Writeanaccountonproteinstructurepredictionwithsuitableexamples.
10. Discussinsomeofcommandsusedtovisualizethemoleculesinthevisualization software's.

KAPRAGAM ACADEM YOF HIGHER EDUCATION

DEPARTMENT OF BIOCHEMISTRY

II- B.Sc Biochemistry

18BCU204-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-

IVSYLLAB

US

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

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

- Mappingprocedures.
- Physiological and pathologicalinformation.
- Experimental data (isoelectric point, molecular weight, amino acid composition, peptidemasses).
- Bibliographicalreferences.
- imagesshowingtheexperimentallydeterminedlocationoftheprotein,aswellasa theoretical region computed from the sequence protein, indicating where the proteinmightbefoundinthegel.
- 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 ofvertebrates.

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.

AsmanyoftheprogramsattheGeneMarkwebsitesharesimilarinterfaces,weuse here the prokaryotic GeneMark.hmm program as an exemplar and discuss program-specificdifferencesbelow,whereappropriate.

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 suchastheMycoplasmageneticcode.

GeneMark.hmmreportsallpredictedgenesinaformatthatincludesthestrandthe generesideson,itsboundaries,lengthinnucleotidesandgeneclass.Classindicateswhich ofthetwoMarkovchainmodelsusedinGeneMark.hmm,TypicalorAtypicalgenemodel, provided the higher likelihood for the gene sequence. Genes of the Typical class exhibit codonusagepatternsspecifictothemajorityofgenesinthegivenspecies,whileAtypical class genes may not follow such patterns and frequently contain significant numbers of laterallytransferredgenes.Thenucleotidesequencesofpredictedgenesandtranslated

proteinsequencesareavailableasanoutputtofacilitatefurtheranalysis,suchasBLAST searching. An option to generate Gene Mark predictions in parallel with the GeneMark.hmm analysis provides important additional information. In this case, Gene Markissetuptousemodelsderivedfromthesametrainingdataasmodelsforthecurrent run ofGeneMark.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. Differencesfrequentlyindicatesequenceerrorsanddeviationsingeneorganization,very shortgenes,genefragments,geneoverlaps,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 representingdifferentclassesofgenes.Here,thecodingpotentialgraphobtainedusingthe Typicalgenemodel,derivedbyGeneMarkS,isdenotedbyasolidblackline,andthecoding potentialgraphobtainedusingtheAtypicalgenemodel(derivedbyaheuristicapproach)is denoted by a dotted line. The GeneMark graph also includes indications of frameshift positions(alsolistedinthetextreport),whichareoftensequencingerrorsbutinrarecases arenaturalandbiologicallyveryinteresting.

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 generalizedViterbi

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'RBSProb','RBSSite'and'RBSSeq').Inadditiontothelistofpredictedgenes, 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 thethreshold.

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 thegenome,suchaspathogenicityislands.

Ifmodels(includingRBSmodels)havetobecomputed*denovo*forananonymous DNAsequencewithlengthoftheorderof1Mborlonger,theGeneMarkSprogramcanbe used. This program needs significantly more computational resources; thus, its output is providedviaemail.AmodifiedversionofGeneMarkStunedfortheanalysisofvirusesof eukaryotic hosts creates a model for the Kozak consensus sequence instead of a two-component RBSmodel.

TheeukaryoticversionofGeneMark.hmmiscurrentlyavailablefortheanalysisof 11 eukaryotic genomes: *Homo sapiens*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Chlamydomonas reinhardtii*, *Drosophila melanogaster*, *Gallusgallus*, *Hordeum vulgare*, *Mus musculus*, *Oryza sativa*, *Triticum aestivum* and *Zea mays*. From theprediction

accuracy tables given at the website (http://opal.biology.gatech.edu/GeneMark/plant_accuracy.html), it follows that thelatest 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 genomesequencingconsortiums,whichhaveusedtheprogramintensively.Theanalysisof cDNAandESTsequencesfromeukaryotes,whichtypicallycontainnointrons,isfacilitated byaspecialversionofGeneMarkcalledGeneMark.SPL.Interestingly,eukaryoticgenomes 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 suchas*Saccharomycescerevisiae*.Currently,thisinterfaceemploysversionsofprokaryoticGene Mark and GeneMark.hmm augmented with Kozak start site models instead of the prokaryotic RBSmodel.

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 currentlybeingexploredtoaddressthisissue.

Patter Recognition

Every accumulation of data in its raw form holds obscure patterns. Pattern recognitiondealswiththescienceoftransformingandclassifyingentitiesonthebasisof these patterns. It is a vast field as it deals with data from diverse sources. Data can be of singledimensionalnatureasincaseofstockexchangesandsound,two-dimensionalasin case of images, and even multidimensional. It has many applications, for example, in medical science, it provides origins for computer-aided diagnosis (CAD) whichsupports medical practitioners in interpretations and finding of diseases. It has other typical applications:automaticspeechrecognition;recognitionoftextinvariouscategories;and automaticrecognitionofhumanfaces.Moreover,thegeneticandproteinstructureinliving organismsformintrinsicpatterns.Datacollectedfromthedecompositionoftheseproteins help to identify them and hence to classify the protein. The ultimate objective is to make machinesideallyasintelligentashumansinrecognizingsuchpatternswhichhelptoform automatedsystemsforconductionofroutinematters.

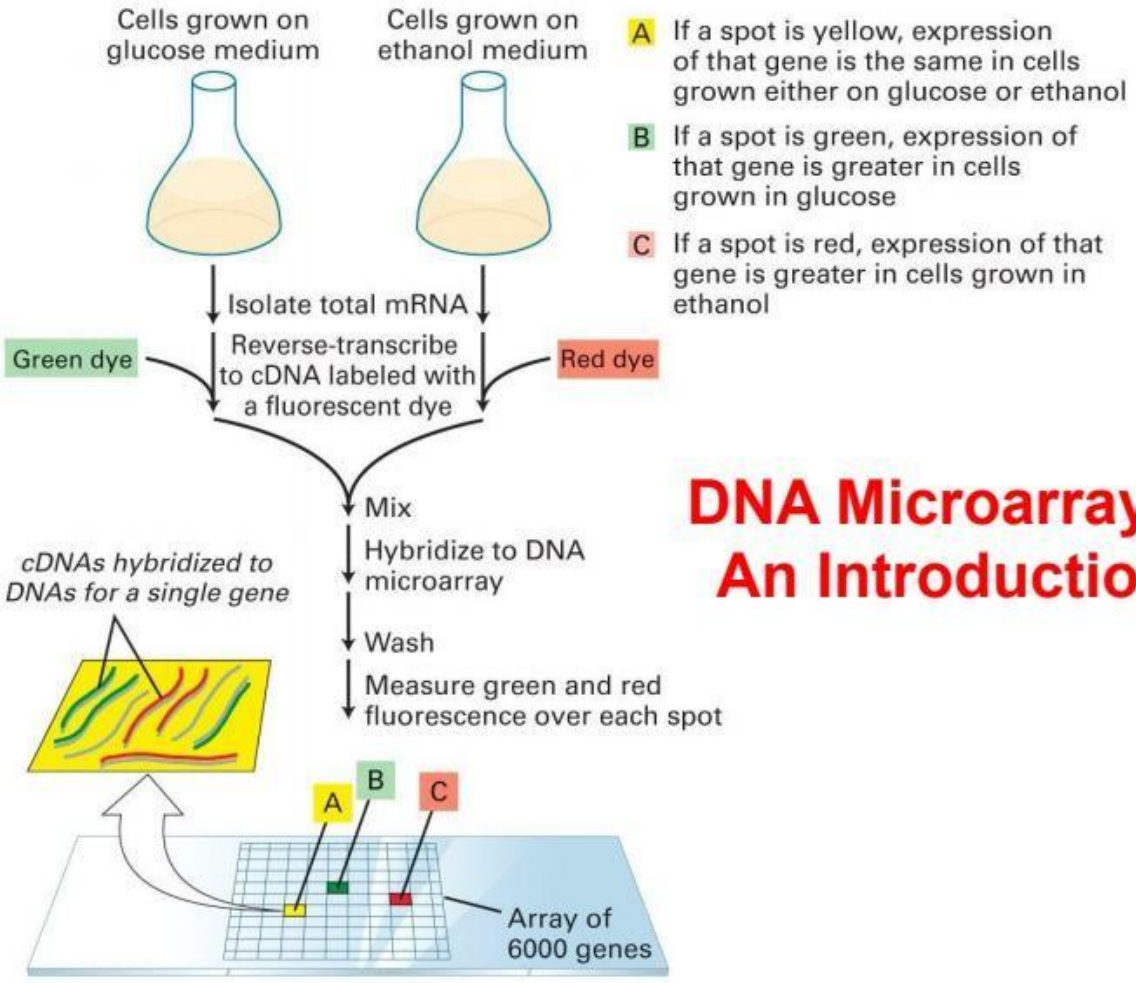
Bioinformatics deals with development of algorithms and software for understandingthebiologicaldata.Foranalyzingandinterpretationofthebiologicaldata,

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 machinelearning.

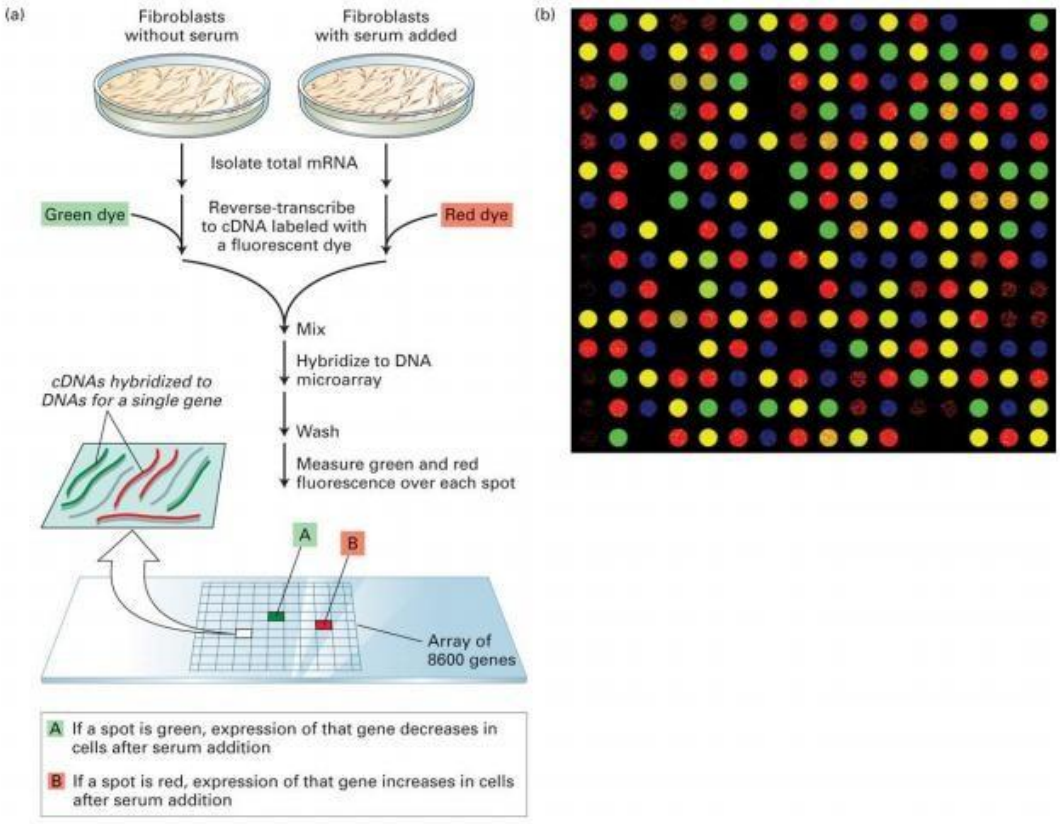
Bioinformatics not just deals with application of pattern recognition for protein classification but it also incorporates use of computational intelligence in protein sequencing, gene expression, comparative genomics, mutation, disease genetics, and molecular interactive networks.

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 multinomialclassifier.

DNA Microarray



DNAmicroarrayanalysiscanrevealdifferencesingeneexpressioninfibroblastsunder different experimentalconditions



Possible Question

1. WriteindetailaboutDNAMicroarray
2. Writeanaccountongeneprediction
3. Application of DNAmicroarray
4. Enumeratetheimportanceofpatternrecognition
5. WriteanoteonSAGEanalysis
6. Discussthegenepredictiontools
7. Compate Genscan vs Genemark
8. Writenoteonoligomersanditsimportanceinmicroarray.
9. Writeindetailaboutdifferentialgeneexpressionanalysis.
10. Explaintheimportanceofgenepredictionmethodsinwholegenomesequencing
11. WriteshortnotesonGeneMark.

KAPRAGAM ACADEM YOF HIGHER EDUCATION

DEPARTMENT OF BIOCHEMISTRY

II- B.Sc Biochemistry

18BCU204-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-

VSYLLAB

US

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 warehousingandanalyzingtheDNAsequences.
- In Bioinformatics knowledge of many branches are required like biology, mathematics, computer science, laws of physics & chemistry, and of course soundknowledgeofITtoanalyzebiotechdata.
- Bioinformaticsisnotlimitedtothecomputingdata,butinrealityitcanbeusedto solvemanybiologicalproblemsandfindouthowlivingthingsworks.

Bioinformatics is being used in following fields:

- Molecularmedicine
- Personalisedmedicine
- Preventativemedicine
- Genetherapy
- Drugdevelopment
- Microbial genomeapplications
- Wastecleanup
- Climate changeStudies
- Alternative energysources
- Biotechnology

Molecular medicine

- Thehumangenomewillhaveprofoundeffectsonthefieldsofbiomedicalresearch and clinicalmedicine.

- Everydiseasehasageneticcomponent.Thismaybeinherited(asisthecasewithan estimated3000-4000hereditarydiseaseincludingCysticFibrosisandHuntingtons disease)oraresultofthebody'sresponsetoanenvironmentalstresswhichcauses alterationsinthegenome(e.g.cancers,heartdisease,diabetes.).
- The completion of the human genome means that we can search for the genes directlyassociatedwithdifferentdiseasesandbegintounderstandthemolecular basisofthesediseasesmoreclearly.
- This new knowledge of the molecular mechanisms of disease will enable better treatments,curesandevenpreventativeteststobedeveloped.

Personalised medicine

- Clinicalmedicinewillbecomemorepersonalisedwiththedevelopmentofthefield ofpharmacogenomics.
- This is the study of how an individual's genetic inheritance affects the body's response todrugs.
- Atpresent,somedrugsfailtomakeittothemarketbecauseasmallpercentageof the clinical patient population show adverse affects to a drug due to sequence variantsintheirDNA.
- Asaresult,potentiallylifesavingdrugsnevermakesittothemarketplace.
- Today,doctorshavetousetrialanderrortofindthebestdrugtotreataparticular patient as those with the same clinical symptoms can show a wide range of responsestothesametreatment.
- Inthefuture,doctorswillbeabletoanalyseapatient'sgeneticprofileandprescribe thebestavailabledrugtherapyanddosagefromthebeginning.

Preventative medicine

- Withthespecificdetailsofthegeneticmechanismsofdiseasesbeingunraveled,the development of diagnostic tests to measure a persons susceptibility to different diseasesmaybecomeadistinctreality.

- Preventativeactionssuchaschangeoflifestyleorhavingtreatmentattheearliest possible stages when they are more likely to be successful, could result in huge advancesinourstruggletoconquerdisease.

Gene therapy

- In the not too distant future, the potential for using genes themselves to treat diseasemaybecomeareality.
- Gene therapy is the approach used to treat, cure or even prevent disease by changingtheexpressionofaperson'sgenes.
- Currently,thisfieldisinitsinfantilestagewithclinicaltrialsformanydifferent typesofcancerandotherdiseasesongoing.

Drug development

- Atpresentalldrugsonthemarkettargetonlyabout500proteins.
- Withanimprovedunderstandingofdiseasemechanismsandusingcomputational toolstoidentifyandvalidatenewdrugtargets,morespecificmedicinesthatacton thecause,notmerelythesymptoms,ofthediseasecanbedeveloped.
- Thesehighlyspecificdrugspromisetohavefewersideeffectsthanmanyoftoday's medicines.

Microbial genome applications

- Microorganismsareubiquitous,thatistheyarefoundeverywhere.
- Theyhavebeenfoundsurvivingandthrivinginextremesofheat,cold,radiation, salt,acidityandpressure.
- Theyarepresentintheenvironment,ourbodies,theair,foodandwater.
- Traditionally,usehasbeenmadeofavarietyofmicrobialpropertiesinthebaking, brewingandfoodindustries.
- The arrival of the complete genome sequences and their potential to provide a greaterinsightintothemicrobialworldanditscapacitiescouldhavebroadandfar reachingimplicationsforenvironment,health,energyandindustrialapplications.

- Forthesereasons,in1994,theUSDepartmentofEnergy(DOE)initiatedtheMGP (Microbial Genome Project) to sequence genomes of bacteria useful in energy production,environmentalcleanup,industrialprocessingandtoxicwastereduction.
- By studying the genetic material of these organisms, scientists can begin to understandthesemicrobesataveryfundamentallevelandisolatethegenesthat givethemtheiruniqueabilitiestosurviveunderextremeconditions.

Waste cleanup

- *Deinococcusradiodurans*isknownastheworld'stoughestbacteriaanditisthemost radiationresistantorganismknown.
- Scientists are interested in this organism because of its potential usefulness in cleaningupwastesitesthatcontainradiationandtoxicchemicals.

Climate change Studies

- Increasinglevelsofcarbondioxideemission,mainlythroughtheexpandinguseof fossilfuelsforenergy,arethoughttocontributetoglobalclimatechange.
- Recently,theDOE(DepartmentofEnergy,USA)launchedaprogramtodecrease atmosphericcarbondioxidelevels.Onemethodofdoingsoistostudythegenomes ofmicrobesthatusecarbondioxideastheirsolecarbonsource.

Alternative energy sources

Scientistsarestudyingthegenomeofthemicrobe*Chlorobiumtepidum*whichhas anunusualcapacityforgeneratingenergyfromlight

Biotechnology

- The archaeon *Archaeoglobus fulgidus* and the bacterium *Thermotoga maritima* have potential for practical applications in industry and government-funded environmentalremediation.
- These microorganisms thrive in water temperatures above the boiling point and thereforemayprovidetheDOE,theDepartmentofDefence,andprivatecompanies withheat-stableenzymessuitableforuseinindustrialprocesses.

- Otherindustriallyusefulmicrobesinclude,*Corynebacteriumglutamic*umwhichisof high industrial interest as a research object because it is used by the chemical industryforthebiotechnologicalproductionoftheaminoacidlysine.
- Thesubstanceisemployedasasourceofproteininanimalnutrition.
- Lysineisoneoftheessentialaminoacidsinanimalnutrition.
- Biotechnologically produced lysine is added to feed concentrates as a sourceof protein,andisanalternativetosoybeansormeatandbonemeal.
- *Xanthomonas campestris* pv. is grown commercially to produce the exopolysaccharide xanthan gum, which is used as a viscosifying and stabilizing agentinmanyindustries.
- *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 manufacturingdairyproductslikebuttermilk,yogurtandcheese.
- Thisbacterium,*Lactococcuslactis*ssp.,isalsousedtopreparepickledvegetables, beer,wine,somebreadandsausagesandotherfermentedfoods.
- Researchersanticipatethatunderstandingthephysiologyandgeneticmake-upof this bacterium will prove invaluable for food manufacturers as well as the pharmaceuticalindustry,whichisexploringthecapacityof*L.lactis*toserveasa vehiclefordeliveringdrugs.

Antibiotic resistance

- Scientistshavebeenexaminingthegenomeof*Enterococcusfaecalis*-aleadingcause ofbacterialinfectionamonghospitalpatients.
- Theyhavediscoveredavirulenceregionmadeupofanumberofantibiotic-resistant genesthatmaycontributetothebacterium'stransformationfromaharmlessgut bacteriatoamenacinginvader.
- Thediscoveryoftheregion,knownasapathogenicityisland,couldprovideuseful markersfordetectingpathogenicstrainsandhelptoestablishcontrolstoprevent thespreadofinfectioninwards.

Forensic analysis of microbes

Scientistsusedtheirgenomictoolstohelpdistinguishbetweenthestrainof*Bacillus anthryacis* that was used in the summer of 2001 terrorist attack in Florida with that of closelyrelatedanthraxstrains

The reality of bioweapon creation

- Scientistshaverecentlybuiltthevirus poliomyelitisusingentirelyartificialmeans.
- Theydidthisusinggenomicdataavailableontheinternetandmaterialsfroma mail-order chemicalsupply.
- TheresearchwasfinancedbytheUSDepartmentofDefenseaspartofabiowarfare responseprogramtoprovetotheworldtherealityofbioweapons.
- Theresearchersalsohopetheirworkwilldiscourageofficialsfromeverrelaxing programs ofimmunisation.
- Thisprojecthasbeenmetwithverymixedfeelings

Evolutionary studies

Thesequencingofgenomesfromallthreedomainsoflife,eukaryota,bacteriaand archaeameansthatevolutionarystudiescanbeperformedinaquesttodeterminethetree oflifeandthelastuniversalcommonancestor.

Crop improvement

- Comparativegeneticsoftheplantgenomeshasshownthattheorganisationoftheir genes has remained more conserved over evolutionary time than was previously believed.
- Thesefindingssuggestthatinformationobtainedfromthemodelcropsystemscan beusedtosuggestimprovementstootherfoodcrops.
- AtpresentthecompletegenomesofArabidopsisthaliana(watercress)andOryza sativa(rice)areavailable.

Insect resistance

- Genes from *Bacillus thuringiensis* that can control a number of serious pests have beensuccessfullytransferredtocotton,maizeandpotatoes.

- This new ability of the plants to resist insect attack means that the amount of insecticidesbeingusedcanbereducedandhencethenutritionalqualityofthecrops isincreased.

Improve nutritional quality

- Scientistshaverecentlysucceededintransferringgenesintoricetoincreaselevels ofVitaminA,ironandothermicronutrients.
- Thisworkcouldhaveaprofoundimpactinreducingoccurrencesofblindnessand anaemiacausedbydeficienciesinVitaminAandironrespectively.
- Scientistshaveinsertedagenefromyeastintothetomato,andtheresultisaplant whosefruitstayslongeronthevineandhasanextendedshelflife

Development of Drought resistance varieties

- Progresshasbeenmadeindevelopingcerealvarietiesthathaveagreatertolerance forsoilalkalinity,freealuminumandirontoxicities.
- Thesevarietieswillallowagriculturetosucceedinpoorersoilareas,thusadding morelandtotheglobalproductionbase.
- Researchisalsoinprogresstoproducecropvarietiescapableoftoleratingreduced waterconditions

Veterinary Science

Sequencingprojectsofmanyfarmanimalsincludingcows,pigsandsheeparenow wellunderwayinthehopethatabetterunderstandingofthebiologyoftheseorganisms willhavehugeimpactsforimprovingtheproductionandhealthoflivestockandultimately havebenefitsforhumannutrition.

Comparative Studies

- Analysingandcomparingthegeneticmaterialofdifferentspeciesisanimportant methodforstudyingthefunctionsofgenes,themechanismsofinheriteddiseases and speciesevolution.
- Bioinformatics tools can be used to make comparisons between the numbers, locationsandbiochemicalfunctionsofgenesindifferentorganisms.

- 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 includingshortlifespans,rapidreproduction,beingeasytohandle,inexpensiveand theycanbemanipulatedatthegeneticlevel.
- Anexampleofahumanmodelorganismisthemouse.
- Mouseandhumanareverycloselyrelated(>98%)andforthemostpartweseea onetoonecorrespondencebetweengenesinthetwospecies.
- Manipulationofthemouseatthemolecularlevelandgenomecomparisonsbetween thetwospeciescanandisrevealingdetailedinformationonthefunctionsofhuman genes,theevolutionaryrelationshipbetweenthetwospeciesandthemolecular mechanismsofmanyhumandiseases.

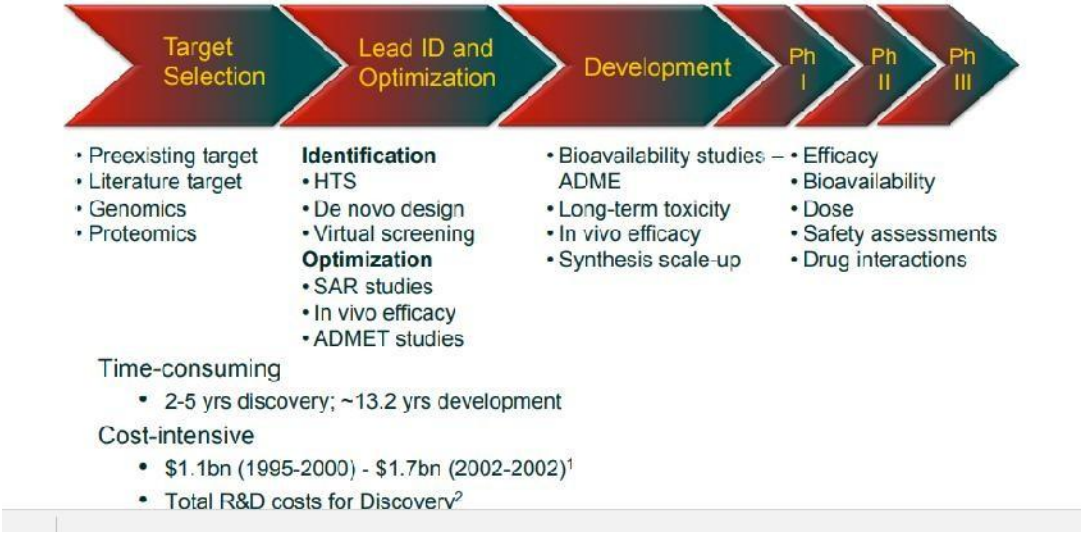
Computer aided drug design

Computer-aided drug design, often called structure based drug design involves using the biochemical information of ligand-receptor interaction in order to postulate ligandrefinements.Forexample,ifweknowthebindingsitethestericcomplementarityof theligandcouldbeimprovedtoincreasetheaffinityforitsreceptor.Indeed,usingthe crystalstructureofthecomplexwecantargetregionsoftheligandthatfitpoorlywithin theactivesiteandpostulatechemicalmodificationsthatlowertheenergeticpotentialby making more negative van der Waals terms, thus improving complementarity with the receptor.Inasimilarfashion,functionalgroupsontheligandcanbechangedinorderto augmentelectrostaticcomplementaritywiththereceptor.Whenatargetisselectedforthe designofnewleadcompoundsthreedifferentsituationscanbefacedregardingtheamount ofinformationofthesystemthatisavailable:

1) Thestructureofthereceptoriswellknownandthebioactiveconformation oftheligandisnotknown,
2) Onlythebioactiveconformationoftheligandisknownand

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

Thebestpossiblestartingpointisan X-raycrystalstructureofthetargetsite.Ifthe molecularmodelofthebindingsiteispreciseenough,onecanapplydockingalgorithms thatsimulatethebindingofdrugstotherespectivereceptorsite,likeAutodock.24Inthe firststeptheprogramcreatesanegativeimageofthetargetsitethroughtheuseofseveral atomprobesthatdetermineaffinitypotentialsforeachatomtypeinthesubstratemolecule atdifferentpointsinagrid,placetheputativeligandsintothesiteandfinallytheyevaluate thequalityofthefit.Theprogramwilltryasetofdifferentconformersoftheligandin ordertoobtainthebestdispositionoftheatomsofthemoleculeformaximizingthescoring functionthatquantifiesligandreceptorinteraction.Adifferentstrategyforobtainingnew leadcompoundsthroughrationaldrugdesignisthedenovodesignofligandswiththeuse ofabuilderprogram,likeLigbuilder.25Thisprogramalsodeterminestheshapeandthe electrostaticpropertiesofthebindingsitecavitythroughtheuseofseveralatomprobes andthenitcombinesfromalibraryofchemicalfragmentsthosethatbetterfillthecavity basedonstericandelectrostaticcomplementarity.



Design of Drug candidates: An iterative process

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

Thestateoftheartdesignprocessisbasedinlargepart,onagoodunderstandingof molecularrecognitionofprotein-ligandcomplexesrelyinguponanalogiestoothersystems andusingadvancedcomputerizedmoleculardesignprograms.

Steps in structure based drug design

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

- Obtaining3Dstructureofprotein
- Active siteidentification
- Ligand-receptor fitanalysis
- Designofnewleads

BeginningtheDesignPhase

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

Inthecaseoftheoptimizationofaknownseries,theinformationisusedtodesignnew analogs.

Eight golden rules in receptor-based ligand design

Theimportantconsiderationsforreceptor-basedliganddesigncanbesummarizedintothe following eightrules:

1. Coordinatetokeyanchoringsites
2. Exploit hydrophobicinteractions
3. Exploithydrogenbondingcapabilities
4. Exploit electrostaticinteractions
5. Favorbioactiveform&avoidenergystrain
6. OptimizevdWContactsandavoidbumps
7. Structural water molecules andsolvation
8. Consider entropiceffect

Rule 1: Coordinate to Key Anchoring Sites

- Whenworkingwithtargetproteins,firstonehastoconsiderthepropereanchorage oftheligandtokeyelementsofthecatalyticsite
- Thisanchoragenotonlypositionstheligandintheactivesitebutalsocounteracts the effect of de-solvating the two components when binding occurs. This isvery importantenergetically.

Rule2:ExploitHydrophobicInteractions

- With hydrophobic pockets, placing a hydrophobic surface of the ligand in hydrophobic sites of the target protein provides an important driving force in complexformationbecauseitreducesnon-polarsurfaceareasexposedtowater
- Althoughindividuallysmall,thetotalcontributionofhydrophobicforcestodrug-receptorinteractionsissubstantial
- Empiricaldatasuggeststhatthefreeenergycontributionduetohydrophobicforces isapproximately2.9kJ/molpermethylenegroupand8.4kJ/molforabenzenering
- Unlikehydrogenbonds,thehydrophobicinteractionsarenotdirectional

Rule 3: Exploit Hydrogen Bonding Capabilities

- Unsatisfied hydrogen bond donors and acceptors are rarely seen in proteins and protein-ligandcomplexesbecausethiswouldbehighlyenergeticallyunfavorable
- 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 ß-sheet formationandarequitefavorable.TheaverageN-H---Oangleisabout155°(with 90%lyingbetween140°and180°).
- Almostallproteingroupsarecapableofforminghydrogenbondslikethis.Where groupsarenotexplicitlyhydrogenbonded,theyareprobablysolvated.

Rule 4: Exploit Electrostatic Interactions

- The optimization of ligand-protein electrostatics can be achieved by placing a positivechargeinclosevicinitytoanenzymenegativecharge

Rule 5: Favor Bioactive Form & Avoid Energy Strain

- Conformationalenergycalculationsareperformedoneachdesignideainorderto determinetheinternalpenaltyrequiredforthenewligandtoattainitsbioactive bindingconformationinsidetheprotein.Theinternalenergythatisrequiredforthe smallmoleculetoreachitsbindingconformationisenergylostinbinding.
- Restrictingtheconformationspaceofaninhibitorcanbebeneficialtobindingwhen theconformationisbiasedtowardsthebioactiveconformer.

Rule 6: Optimize VDW Contacts and Avoid Bumps

- Attractive van der Waals interactions occur over a short distance range and attractiondecreasesas1/r6.Asaresult,optimizationofattractivevanderWaals interactionsoccursastheshapeoftheproteinbindingsiteandtheshapeofthe ligand matchwell.
- Calculations of steric fit are difficult because of possible flexing motions of the proteinbackboneandespeciallytheresiduesidechains

Rule 7: Structural Water Molecules and Solvation

- Inhibitordesignstrategieshavegreatpotentialwhentheytargetthedisplacement ofwatermoleculestightlyboundtotheproteinbyincorporatingelementsofthe watermoleculewithintheinhibitor.
- When polar charged groups are considered in the design of a ligand, one should leavesomeroomforotherwatermoleculestosolvatethechargedcenter(except possiblywhenasaltbridgeisformed).

Rule 8: Consider Entropic Effect

- Aflexiblemoleculehasabetterchanceoffindinganoptimalfitintoareceptor,but thisisachievedatthecostoflargeconformationalentropy
- Sufficientconformationalrigidityisessentialtoensurethatthelossofentropyupon ligandbindingisacceptable
- Arigidmoleculehaslittleconformationalentropybutisunlikelytofitoptimallyinto thereceptor

- An analysis of the contributions of various functional groups to protein-ligand bindingdemonstratesthateachfreelyrotatingbondinaligandreducesbindingfree energybyabout2.9kJ/mol
- Making a flexible molecule more rigid will lead to enhanced activity if the right conformation ismaintained
- ExampleofSuccessfulStructure-BasedDesign
- The use of the crystallographic structure of the HIV-1 protease in drug design representsoneofthemoreimpressivesuccessstoriesinthestructure-baseddrug design field. Structure-based design studies has resulted in the identification of distinctclassesofinhibitorsandseveralsuccessfuldrugcandidateshaveemerged fromthesestudiesandareusedinthecontrolofAIDS.
- TheHIV-1proteaseplaysacrucialpartinthelifecycleoftheHIVvirus.Inhibitor drugsblocktheactionoftheproteaseandthevirusperishesbecauseitisunableto matureintoitsinfectiousform.
- The HIV-1 protease is a small dimer enzyme comprising two identical folded 99 amino-acidchainsAandB

Ligand-Based Computer-Aided Drug Design

The ligand-based computer-aided drug discovery (LBDD) approach involves the analysisofligandsknowntointeractwithatargetofinterest.Thesemethodsuseasetof 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 compoundsinsuchawaythatthephysicochemicalpropertiesmostimportantfortheir desired interactions are retained, whereas extraneous information not relevant to the interactionsisdiscarded.Itisconsideredasanindirectapproachtothedrugdiscoveryin thatitdoesnotnecessitateknowledgeofthestructureofthetargetofinterest.Thetwo 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 quantitativestructureactivityrelationship(QSAR)modelthatpredictsbiologicalactivity

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,andalsofortheoptimizationofDMPK/ADMETproperties.LBDDisbasedon thesimilarpropertyprinciplewhichstatesthatmoleculesthatarestructurallysimilarare likely to have similar properties. LBDD approaches in contrast to SBDD approaches can alsobeappliedwhenthestructureofthebiologicaltargetisunknown.Additionally,active compounds identified by ligand based virtual high-throughput screening (LB-vHTS) methodsareoftenmorepotentthanthoseidentifiedinSB-vHTS.

Molecular Descriptors

Moleculardescriptorscanincludepropertiessuchasmolecularweight,geometry, volume,surfaceareas,ringcontent,rotatablebonds,interatomicdistances,bonddistances, atomtypes,planarandnonplanarsystems,molecularwalkcounts,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,orquantum-mechanicaltoolsandareclassifiedaccordingtotheChapter1 Computer Aided Drug Design: An Overview 16 “dimensionality” of the chemical representation from which they are computed: 1- dimensional (1D), scalar physicochemicalpropertiessuchasmolecularweight;2D,molecularconstitution-derived descriptors; 2.5D, molecular configuration-derived descriptors; 3D, molecular conformation-derived descriptors. These different levels of complexity, however, are overlappingwiththemorecomplexdescriptors,oftenincorporatinginformationfromthe simplerones.

Molecular Fingerprint and Similarity Searches

Molecularfingerprint-basedtechniquesattempttorepresentmoleculesinsucha way as to allow rapid structural comparison in an effort to identify structurally similar moleculesortoclustercollectionsbasedonstructuralsimilarity.Thesemethodsarefewer hypothesesdrivenandlesscomputationallyexpensivethanpharmacophoremappingor QSARmodels.Theyrelyentirelyonchemicalstructureandomitcompoundwithknown

biological activity, making the approach more qualitative in nature than other LBDD approaches.Additionally,fingerprint-basedmethodsconsiderallpartsofthemolecule equally and avoid focusing only on parts of a molecule that are thought to be most importantforactivity.Thisislesserrorpronetooverfittingandrequiressmallerdatasets tobeginwith.Fingerprintmethodsmaybeusedtosearchdatabasesforcompoundssimilar instructuretoaleadquery,providinganextendedcollectionofcompoundsthatcanbe testedforimprovedactivityoverthelead.Inmanysituations,2Dsimilaritysearchesof databasesareperformedusingchemotypeinformationfromfirstgenerationhits,leading tomodificationsthatcanbeevaluatedcomputationallyororderedforinvitrotesting.

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 correlationofvariouselectronic,hydrophobic,andstericfeatureswithbiologicalactivity. Inthe1960s,HanschandothersbegantoestablishQSARmodelsusingvariousmolecular descriptors to physical, chemical, and biological properties focused on providing computational estimates for the bioactivity of molecules. In 1964, Free and Wilson developedamathematicalmodelrelatingthepresenceofvariouschemicalsubstituentsto biologicalactivity(eachtypeofchemicalgroupwasassignedanactivitycontribution),and thetwomethodswerelatercombinedtocreatetheHansch/Free-Wilsonmethod.Amodel is then generated to identify the relationship between those descriptors and their experimentalactivity,maximizingthepredictivepower.Finally,themodelisappliedto predict activity for a library of test compounds that were encoded with the same descriptors.SuccessofQSAR,therefore,dependsnotonlyonthequalityoftheinitialsetof active/inactivecompoundsbutalsoonthechoiceofdescriptorsandtheabilitytogenerate 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 samplingspaceoftheinitialsetofcompoundswithknownactivity,thechemicaldiversity. Inotherwords,divergentscaffoldsorfunctionalgroupsnotrepresentedwithinthis

"training"setofcompoundswillnotberepresentedinthefinalmodel,andanypotential hits within the library to be screened that contain these groups will likely be missed. Therefore,itisadvantageoustocoverawidechemicalspacewithinthetrainingset.

3D-QSAR

Comparativefieldmolecularanalysis(CoMFA)isa3D-QSARtechniquethataligns molecules and extracts aligned features that can be related to biological activity. This methodfocusesonthealignmentofmolecularinteractionfieldsratherthanthefeaturesof eachindividualatom.CoMFAwasestablishedover20yearsagoasastandardtechniquefor constructing 3D models in the absence of direct structural data of the target. In this method,moleculesarealignedbasedontheir3Dstructuresonagridandthevaluesof steric (van der Waals interactions) and electrostatic potential energies (Coulombic interactions)arecalculatedateachgridpoint.Acomparativemolecularsimilarityindex (CoMSIA)isanimportantextensiontoCoMFA.InCoMSIA,themolecularfieldincludes 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 extremevalues.

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 conformationalentropy,andbindingpocketadaptation.4D-QSARisanextensionof3D-QSARthattreatseachmoleculeasanensembleofdifferentconformations,orientations, tautomers,stereoisomers,andprotonationstates.Thefourthdimensionin4D-QSARrefers totheensemblesamplingofspatialfeaturesofeachmolecule.Areceptor-independent(RI) 4D-QSARmethodwasproposedbyHopfingerin1997.Thismethodbeginsbyplacingall moleculesintoagridandassigninginteractionpharmacophoreelementstoeachatomin themolecule(polar,nonpolar,hydrogenbonddonor,etc.).Moleculardynamicssimulations areusedtogenerateaBoltzmannweightedconformationalensembleofeachmolecule withinthegrid.Trialalignmentsareperformedwithinthegridacrossthedifferent

molecules, and descriptors are defined based on occupancy frequencies within eachof these alignments. These descriptors are called grid cell occupancy descriptors. A conformationalensembleofeachcompoundisusedtogeneratethegridcelloccupancy descriptorsratherthanasingleconformation.5D-QSARhasbeendevelopedtoaccountfor localchangesinthebindingsitethatcontributetoaninducedfitmodelofligandbinding.In amethoddevelopedbyVedaniandDobler,inducedfitissimulatedbymappinga“mean envelope”forallligandsinatrainingsetontoan“innerenvelope”foreachindividual molecule. Their method involves several protocols for evaluating induced-fit models includingalinearscalebasedontheadaptationoftopology,adaptationsbasedonproperty fields, energy minimization, and lipophilicity potential. By using this information, the energeticcostforadaptationoftheligandtothebindingsitegeometryiscalculated.Vedani fromtheBiographicsLaboratorydevelopedareceptormodelingconcept,Quasar,basedon 6D-QSARthatexplicitlyallowsforthesimulationofinducedfit.Quasarconcept,previously 3,4,5Dextendedtosixdimensionsallowsforthesimultaneousconsiderationofdifferent solvationmodelswhichcanbeachievedexplicitlybymappingpartsofthesurfacearea withsolventproperties(positionandsizeareoptimizedbythegeneticalgorithm).

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,itisthespatialarrangementoffunctionalgroupsthatacompoundordrugmust containtoevokeadesiredbiologicalresponse.Therefore,aneffectivepharmacophorewill 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 bindingsitesummarizessstericandelectronicfeaturesneededforoptimalinteractionofa ligandwithatarget.Mostcommonpropertiesthatareusedtodefinepharmacophoresare hydrogenbondacceptors,hydrogenbonddonors,basicgroups,acidicgroups,partial

charge, aliphatic hydrophobic moieties, and aromatic hydrophobic moieties. Pharmacophore features have been used extensively in drug discovery for virtual screening,denovodesign,andleadoptimization.Apharmacophoremodelofthetarget binding site can be used to virtually screen a compound library for putative hits. Apart fromqueryingdatabaseforactivecompounds,pharmacophoremodelscanalsobeusedby de novo design algorithms to guide the design of new compounds. Structure-based pharmacophoremethodsaredevelopedbasedonananalysisofthetargetbindingsiteor basedonatarget-ligandcomplexstructure.LigandScoutusesprotein-ligandcomplexdata tomapinteractionsbetweenligandandtarget.Aknowledgebasedrulesetobtainedfrom the PDB is used to automatically detect and classify interactions into hydrogen bonds, charge transfers, and lipophilic regions. The algorithm creates regularly spaced grids aroundtheligandandthesurroundingresidues.Probeatomsthatrepresentahydrogen 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 betweenprobeatomsandthetarget.SCOREincludestermstoaccountforvanderWaals interactions, metal-ligand bonding, hydrogen bonding, and desolvation effects upon binding. A pharmacophore model is developed by rescoring the grids followed by clusteringandsortingtoextractfeaturesessentialforprotein-ligandinteraction.Themost common software packages used for ligand based pharmacophore generation include Phase,MOE,Catalyst,DISCO,andGASP.

ADME

- Nineofeverytennewdrugsfailinclinicaltesting.
- AdruginphaseIIItestingtestinghas32%chanceoffailurefailure.
- EveninPhaseI,37%fail.
- MostdrugsfailinphaseII.

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-thiswilloccurandcouldimpactseveralvariables–Couldbeusedtoyour 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 distributionrate(particularlytositeofaction)andrateofelimination–Pharmacokinetics maybesimplydefinedaswhatthebodydoestothedrug–Pharmacodynamicsdefinedas whatthedrugdoestothebody.

Drug Delivery

Oral-byfarthemostcommonroute.Thepassageofdrugfromthegutintothebloodis influencedbybiologicandphysicochemicalproperties.

Sublingual(buccal)-Certaindrugsarebestgivenbeneaththetongueorretainedinthe cheekpouchandareabsorbedfromtheseregionsintothelocalcirculation.

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.

Intravenousinjection–Usedwhenarapidclinicalresponseisnecessary,e.g.,anacute asthmaticepisode.–Achieverelativelyprecisedrugconcentrationsintheplasma,since bioavailabilityisnotaconcern.

Intra-arterialinjection–Usedincertainspecialsituations,notablywithanticancerdrugs inanefforttodeliverahighconcentrationofdrugtoaparticulartissue.Typically,the injectedarteryleadsdirectlytothetargetorgan.

Intrathecal injection – The blood-brain barrier limits the entry of many drugs into cerebrospinalfluid.life-threatening,antibiotics,antifungalsandanticancerdrugsaregiven vialumbarpunctureandinjectionintothesubarachnoidspace.

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

Subcutaneousinjection–Somedrugs,notablyinsulin,areroutinelyadministeredSC. DrugabsorptionisgenerallyslowerSCthanIM,duetopoorervascularity.

Inhalation–Volatileanesthetics,aswellasmanydrugswhichaffectpulmonaryfunction, areadministeredasaerosols.Drugsadministeredviathisroutearenotsubjecttofirst-pass livermetabolism.

Topicalapplication–Eye,intravaginal,intranasal,skin.–Alleviationoflocalsymptoms

Drug Absorption and Biological Factors

Membranestructureandfunction-Thecellmembraneisasemi-permeablelipoidsieve containing numerous aqueous channels, as well as a variety of specialized carrier molecules.

Passive diffusion is probably the most important absorptive mechanism. Lipid-soluble drugsdissolveinthemembrane,andaredriventhroughbyaconcentrationgradientacross themembrane.

Carrier-mediatedfacilitatedtransportoccursforsomedrugs,particularlythosewhich areanalogsofendogenouscompoundsforwhichtherealreadyexistspecificmembrane carrier systems. – For example, methotrexate, an anticancer drug which isstructurally similartofolicacid,isactivelytransportedbythefolatemembranetransportsystem.

Oral Drug Absorption

Thebloodsupplydrainingthegutpassesthroughtheliverbeforereachingthesystemic circulation.–First-passeffectmayreducetheamountofdrugreachingthetargettissue. **Drug binding** – Many drugs will bind strongly to proteins in the blood or to food substancesinthegut.–Plasmaproteinbindingwillincreasetherateofpassiveabsorption bymaintainingtheconcentrationgradientoffreedrug.

Foodeffects–Absorptioncanbereducedbythepresenceoffoodinthegut–Absorption can be enhanced by food (bile secretion) – Some drugs are irritating and should be administeredwithmealstoreduceadverseeffects.

Distribution

Once in the blood, drugs are simultaneously distributed throughout the body and eliminated.–Distributionismuchmorerapidthanelimination,accomplishedviathe circulation,andinfluencedbyregionalbloodflow.

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)thoseorgans(e.g.,adiposeandskeletalmuscle)whicharelesswell-perfused, and with which drug therefore equilibrates more slowly. – Special compartments -The cerebrospinalfluid(CSF)andcentralnervoussystem(CNS)isrestrictedbythestructureof the capillaries and pericapillary glial cells. – Drugs also have relatively poor access to pericardialfluid,bronchialsecretionsandfluidinthemiddleear.

Metabolism

- PhaseIandPhaseIImetabolism–Mostproductsofdrugmetabolismarelessactive thantheparentcompound.
- Metabolitesmayberesponsiblefortoxic,mutagenic,teratogenicorcarcinogenic effects – For example, example, acetaminophen acetaminophen hepatotoxicity hepatotoxicityisduetoaminormetabolitewhichreactswithliverproteins.
- Metabolismofso-calledprodrugs,metabolitesareactuallytheactivetherapeutic compounds–Cyclophosphamide,aninertcompoundwhichismetabolizedbythe liverintoahighlyactiveanticancerdrug.

Receptor docking

- Computationaltechniquesassistoneinsearchingdrugtargetandindesigningdrug insilico,butittakeslongtimeandmoney.Inordertodesignanewdrugoneneedto followthefollowingpath.
- **IdentifyTargetDisease**:Oneneedstoknowallaboutthediseaseandexistingor traditionalremedies.
- Itisalsoimportanttolookatverysimilarafflictionsandtheirknowntreatments.
- Target identification alone is not sufficient in order to achieve a successful treatmentofadisease.

- Arealdrugneedstobedeveloped.Thisdrugmustinfluencethetargetproteinin suchawaythatitdoesnotinterferewithnormalmetabolism.Onewaytoachieve thisistoblockactivityoftheproteinwithasmallmolecule.
- Bioinformatics methods have been developed to virtually screen the target for compoundsthatbindandinhibittheprotein.
- Anotherpossibilityistofindotherproteinsthatregulatetheactivityofthetargetby bindingandformingacomplex.

Study Interesting Compounds

- Oneneedstoidentifyandstudytheleadcompoundsthathavesomeactivityagainst adisease.Thesemaybeonlymarginallyusefulandmayhaveseveresideeffects.
- Thesecompoundsprovideastartingpointforrefinementofthechemicalstructures.

Detect the Molecular Bases for Disease

- Ifitisknownthatadrugmustbindtoaparticularspotonaparticularproteinor nucleotidethenadrugcanbetailormadetobindatthatsite.
- Thisisoftenmodeledcomputationallyusinganyofseveraldifferenttechniques. Traditionally, the primary way of determining what compounds would be tested computationally was provided by the researchers' understanding of molecular interactions.
- Asecondmethodisthebruteforcetestingoflargenumbersofcompoundsfroma databaseofavailablestructures.

Rational drug design techniques

- Thesetechniquesattempttoreproducetheresearchers'understandingofhowto chooselikelycompoundsbuiltintoasoftwarepackagethatiscapableofmodelinga verylargenumberofcompoundsinanautomatedway.
- Manydifferentalgorithmshavebeenusedforthistypeoftesting,manyofwhich wereadaptedfromartificialintelligenceapplications.

- The complexity of biological systems makes it very difficult to determine the structuresoflargebiomolecules.Ideallyexperimentallydetermined(x-rayorNMR) structureisdesired,butbiomoleculesareverydifficulttocrystallize.

Refinement of compounds

- Once you got a number of lead compounds have been found, computationaland laboratory techniques have been very successful in refining the molecular structurestogiveagreaterdrugactivityandfewersideeffects.
- Thisisdonebothinthelaboratoryandcomputationallybyexaminingthemolecular structurestodeterminewhichaspectsareresponsibleforboththedrugactivityand the sideeffects.

Solubility of Molecule

- Oneneedtocheckwhetherthetargetmoleculeiswatersolubleorreadilysolublein fattytissuewillaffectwhatpartofthebodyitbecomesconcentratedin.
- Theabilitytogetadrugtothecorrectpartofthebodyisanimportantfactorinits potency.
- Ideally there is a continual exchange of information between the researchersdoing QSARstudies,synthesisandtesting.
- Thesetechniquesarefrequentlyusedandoftenverysuccessfulsincetheydonotrely onknowingthebiologicalbasisofthediseasewhichcanbeverydifficulttodetermine.

Drug Testing

- Onceadrughasbeenshowntobeeffectivebyaninitialassaytechnique,muchmore testingmustbedonebeforeitcanbegiventohumanpatients.
- Animaltestingistheprimarytypeoftestingatthisstage.
- Eventually,thecompounds,whicharedeemedsuitableatthisstage,aresentonto clinicaltrials.
- Intheclinicaltrials,additionalsideeffectsmaybefoundandhumandosagesare determined.

- TwoproteinswhosecomplexwasdeterminedbyProtein-Protein-Docking.Docking ofasmallinhibitortotheureaseprotein.

Introduction to Molecular Dynamics Simulation

Withtheadventofthecomputers,chemists,physicistsandmaterialscientistshad begun(since1950s)toexploitthepowerofthecomputersforprobingthepropertiesof materialsthroughsimulations.Almostallthematerialsofthephysicalworldcanbeprobed fortheirpropertiesbydesigningappropriatesimulationalgorithms.Theseincludeatomic and molecular systems, biomolecules, complex materials, nuclear materials, life processes and the like. One can construct simulation schemes for the dynamics of the molecules presentinthematerials;i.e.,theMolecularDynamics(MD),wheretheconstituentsofthe systemareallowedtointeractaccordingtoknownlawsofphysics,overaperiodoftime. Throughthenumericalsolutionsoftheequationsofmotion(oftendescribedbythelawsof 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.Thesetrajectoriesarethenanalyzedinordertoextractthedesiredproperties suchaspressure,stress,diffusion,viscosity,surfacetension,dielectricconstant,order parameter, autocorrelation functions, fluctuations, conformational changes etc. Since molecularsystemsgenerallyconsistofavastnumberofparticles,itisimpossibletofind the properties of such complex systems analytically. The MD simulation exercises circumventthisproblembyusingthenumericalsolutionsoftheequationsofmotion.Thus theMDsimulationtechniquepresentsaninterfacebetweenlaboratoryexperimentsand thetheory.Thisoftenleadstotherealizationthat‘computersimulations’areactually ‘computerexperiments’.

The MD technique

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

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

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

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

Ineachofthetimestepofthesimulation,oneneedstocomputetheforcefiand usingthisforce,thepositionrigetsupdated.Inordertosolvethesecondorderdifferential equationasineq.(26.1),thereareseveralnumericalschemesavailable.Thesearebasedon finitedifferencemethodsandtheintegrationalgorithmsincludeGearpredictor-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 describethisalgorithmbelow.

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,

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

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

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

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

Although the velocities$v(t)$ 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,

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

Possible Questions

1. WriteashortnotesonMoleculardocking
2. WriteabouttheADMEpropertiesofadrug
3. WritetheLipinski'sruleoffive
4. Giveshortnotesonactivesiteprediction
5. Enumeratethestepsinvolvedindesigningatherapeuticdrug
6. Writetheapplicationsofbioinformaticsinbiotechnologyandagriculture
7. WhatisQSAR?Addnotesonitsimportanceinnoveldrugdesigning
8. Writeshortnotesonreceptordocking
9. Brieflydiscussaboutdrugdesigningconcepts
10. whataretheessentialpropertiesrequiredforapotentleadmolecules.
11. What are the steps involved in discovering a new drug? Explain the role of insilico methodsfordrugdesign.

KAPRAGAM ACADEM YOF HIGHER EDUCATION

DEPARTMENT OF BIOCHEMISTRY

II- B.Sc Biochemistry

18BCB204-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 π in 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**