

EXPERIMENTS

1. Introduction to different operating systems - UNIX, LINUX and Windows
2. Introduction to bioinformatics databases (any three): NCBI/PDB/DDBJ, Uniprot, PDB
3. Sequence retrieval using BLAST
4. Sequence alignment & phylogenetic analysis using clustalW & phylip
5. Picking out a given gene from genomes using Genscan or other softwares (promoter region identification, repeat in genome, ORF prediction). Gene finding tools (Glimmer, GENSCAN), Primer designing, Genscan/Genetool
6. Protein structure prediction: primary structure analysis, secondary structure prediction using psi- pred, homology modeling using Swissmodel. Molecular visualization using jmol, Protein structure model evaluation (PROCHECK)
7. Prediction of different features of a functional gene

SUGGESTED READINGS

1. Saxena Sanjay (2003) A First Course in Computers, Vikas Publishing House
2. Pradeep and Sinha Preeti (2007) Foundations of Computing, 4th ed., BPB Publications
3. Lesk M.A.(2008) Introduction to Bioinformatics . Oxford Publication, 3rd International Student Edition.
4. Rastogi S.C., Mendiratta N. and Rastogi P. (2007) Bioinformatics: methods and applications, genomics, proteomics and drug discovery, 2nd ed. Prentice Hall India Publication
5. Primrose and Twyman (2003) Principles of Genome Analysis & Genomics. Blackwell.

1. Introduction to different operating systems- Windows, Linux, Unix

AIM

To analyse and understand different operating systems such as Windows, Linux, Unix

Operating System

An operating system (OS) is a software program that manages the hardware and software resources of a computer.

- Linux is an operating system or a kernel. It is distributed under an open source license. Its functionality list is quite like UNIX.
- UNIX is called the mother of operating systems which laid out the foundation to Linux. Unix is designed mainly for mainframes and is in enterprises and universities. While Linux is fast becoming a household name for computer users, developers, and server environment. You may have to pay for a Unix kernel while in Linux it is free.
- But, the commands used on both the operating systems are usually the same. There is not much difference between UNIX and Linux. Though they might seem different, at the core, they are essentially the same.
- Since Linux is a clone of UNIX. So learning one is same as learning another.

RESULTS

2. Introduction to bioinformatics databases (any three): NCBI/PDB/DDBJ, Uniprot, PDB

Aim: To analyze various sequence and structural information provided in the NCBI, PDB and Uniprot databases.

Theory

NCBI (National Center for Biotechnology Information) acting as a resource database for molecular biology, computational biology, biochemistry and genetics information which aids in developing new technologies for managing the biological process that control health and diseases. Database is collection of information managed as a computer program in an organized manner. Biological databases are libraries of biological information collected from different literatures, experiments and other analysis which can easily be accessed, updated, and retrieved. Composite database integrates various other primary databases. NCBI established on November 4, 1988 as a part of National Library of Medicine (NLM) at the National Institute of Health (NIH). It can be accessed from the URL <http://www.ncbi.nlm.nih.gov/>. It maintains collaborations with several NIH institutes, industries and other government agencies.

The information managed in the Gene database is the results of curation and automated integration of data from NCBI's Reference Sequence project (RefSeq). It comprises of information about various species including their nomenclature, associated pathways, RefSeqs, phenotypes and links to genome. It can be accessed from the URL <http://www.ncbi.nlm.nih.gov/gene>. Gene database can be accessed by simply query the word, preferably the gene name or the disease names to the query box which will display the list of genes associated with the search. User can also search records with their GeneID, which is a unique identifier given by NCBI. The 'limits' feature allows the user to filter the search according to their needs.

There are different methods to query in NCBI gene. Queries can be searched on the following basis (Table 1).

Query genes by	Search text
Free text	Breast cancer
Gene name (symbol)	AKT1[sym]
Chromosome and symbol	(111[chr] OR 1[chr]) AND brca*[sym]
Partial name and multiple species	Neurotransporter[title]AND("Homo sapiens"[orgn]OR "Rattus norvegicus"[orgn])
Associate sequence accession number	M21213[accn]
Chromosome and species	X[CHR]AND human[ORGN]
Enzyme Commission[EC]number	2.5.1.6[EC]
Gene Ontology(GO)terms or identifiers	"molecular function"[GO]
Publication (PubMed ID)	1802527[PMID]

Table 1: Different method to query in gene database

Searching data

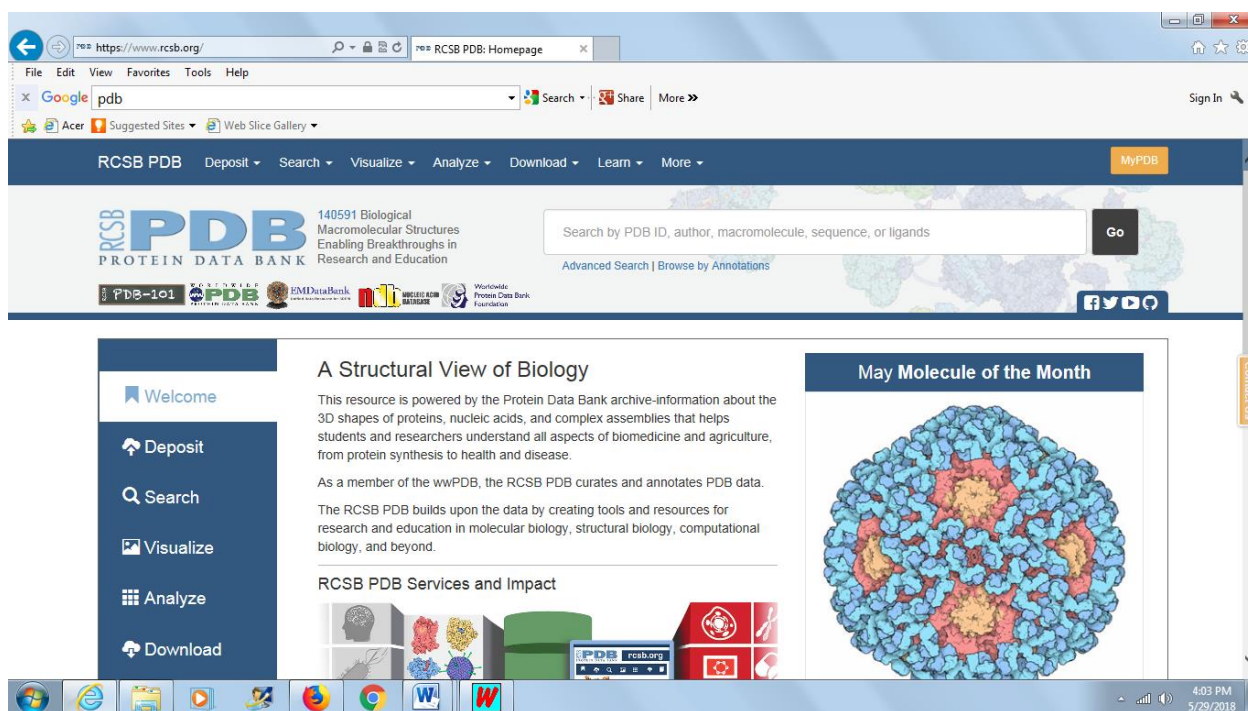
Mainly one can search for genes either by giving a particular disease condition and searching the different genes involved or collecting the gene names from the articles/ interaction experiments/ pathways and providing the names as such. The gene names are provided by the gene nomenclature committee specific for each organism. HUGO Gene Nomenclature Committee (HGNC) is the committee which gives unique gene names for the organism *Homo sapiens* (human). One can also search genes based on these names

RESULTS

PROTEIN DATA BANK

Procedure:

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



Class: III B.Sc Microbiology
COURSE NAME: BIOINFORMATICS PRACTICAL
Subject code: 16MBU512B

For instance, the following is the structural information provided in the PDB file of a cardiotoxin from *Naja naja atra*. The PDB ID of the protein is 2CRT.

```

HEADER          CARDIOTOXIN                                     12-
MAR-94    2CRT
TITLE          CARDIOTOXIN III FROM TAIWAN COBRA (NAJA NAJA
ATRA)
TITLE          2 DETERMINATION OF STRUCTURE IN SOLUTION AND
COMPARISON WITH
TITLE          3 SHORT NEUROTOXINS
COMPND        MOL_ID: 1;
COMPND        2 MOLECULE: CARDIOTOXIN III;
COMPND        3 CHAIN: A;
COMPND        4 ENGINEERED: YES
SOURCE        MOL_ID: 1;
SOURCE        2 ORGANISM_SCIENTIFIC: NAJA ATRA;
SOURCE        3 ORGANISM_COMMON: CHINESE COBRA;
SOURCE        4 ORGANISM_TAXID: 8656
KEYWDS        CARDIOTOXIN
EXPDTA        SOLUTION NMR
AUTHOR        R.BHASKARAN,C.C.HUANG,K.D.CHANG,C.YU
REVDAT        3   24-FEB-09 2CRT      1      VERSN
REVDAT        2   01-APR-03 2CRT      1      JRNL
REVDAT        1   01-NOV-94 2CRT      0
JRNL
JRNL
AUTH          R.BHASKARAN,C.C.HUANG,D.K.CHANG,C.YU
JRNL          TITL    CARDIOTOXIN III FROM THE TAIWAN
COBRA (NAJA NAJA
JRNL          TITL 2  ATRA). DETERMINATION OF STRUCTURE
IN SOLUTION AND
JRNL          TITL 3  COMPARISON WITH SHORT NEUROTOXINS.
JRNL          REF      J.MOL.BIOL.                      V.
235 1291 1994
JRNL          REFN                      ISSN 0022-2836
JRNL          PMID    8308891
JRNL          DOI      10.1006/JMBI.1994.1082
REMARK        1
REMARK        2
REMARK        2 RESOLUTION. NOT APPLICABLE.
REMARK        3
REMARK        3 REFINEMENT.
REMARK        3   PROGRAM      : X-PLOR
REMARK        3   AUTHORS      : BRUNGER

```

REMARK 3
 REMARK 3 OTHER REFINEMENT REMARKS: NULL
 REMARK 4
 REMARK 4 2CRT COMPLIES WITH FORMAT V. 3.15, 01-DEC-08
 REMARK 100
 REMARK 100 THIS ENTRY HAS BEEN PROCESSED BY BNL.
 REMARK 210
 REMARK 210 EXPERIMENTAL DETAILS
 REMARK 210 EXPERIMENT TYPE : NMR
 REMARK 210 TEMPERATURE (KELVIN) : NULL
 The following is atomic coordinates of the protein. It provides information pertaining to atom types, atom nomenclature, atom number, atom position and atom occupancy.

ATOM	1	N	LEU A	1	12.825	6.867	0.006
1.00	1.76		N				
ATOM	2	CA	LEU A	1	12.032	6.006	0.919
1.00	0.90		C				
ATOM	3	C	LEU A	1	11.849	4.631	0.268
1.00	0.73		C				
ATOM	4	O	LEU A	1	12.036	4.471	-0.924
1.00	0.90		O				
ATOM	5	CB	LEU A	1	10.658	6.639	1.165
1.00	1.16		C				
ATOM	6	CG	LEU A	1	10.813	8.094	1.633
1.00	1.61		C				
ATOM	7	CD1	LEU A	1	9.429	8.662	1.957
1.00	2.23		C				
ATOM	8	CD2	LEU A	1	11.676	8.150	2.900
1.00	2.57		C				
ATOM	9	H1	LEU A	1	13.292	6.273	-0.707
1.00	2.15		H				

Class: III B.Sc Microbiology

COURSE NAME: BIOINFORMATICS PRACTICAL

Subject code: 16MBU512B

ATOM	10	H2	LEU A	1	12.194	7.547	-0.470
1.00	2.36		H				
ATOM	11	H3	LEU A	1	13.547	7.380	0.551
1.00	2.21		H				

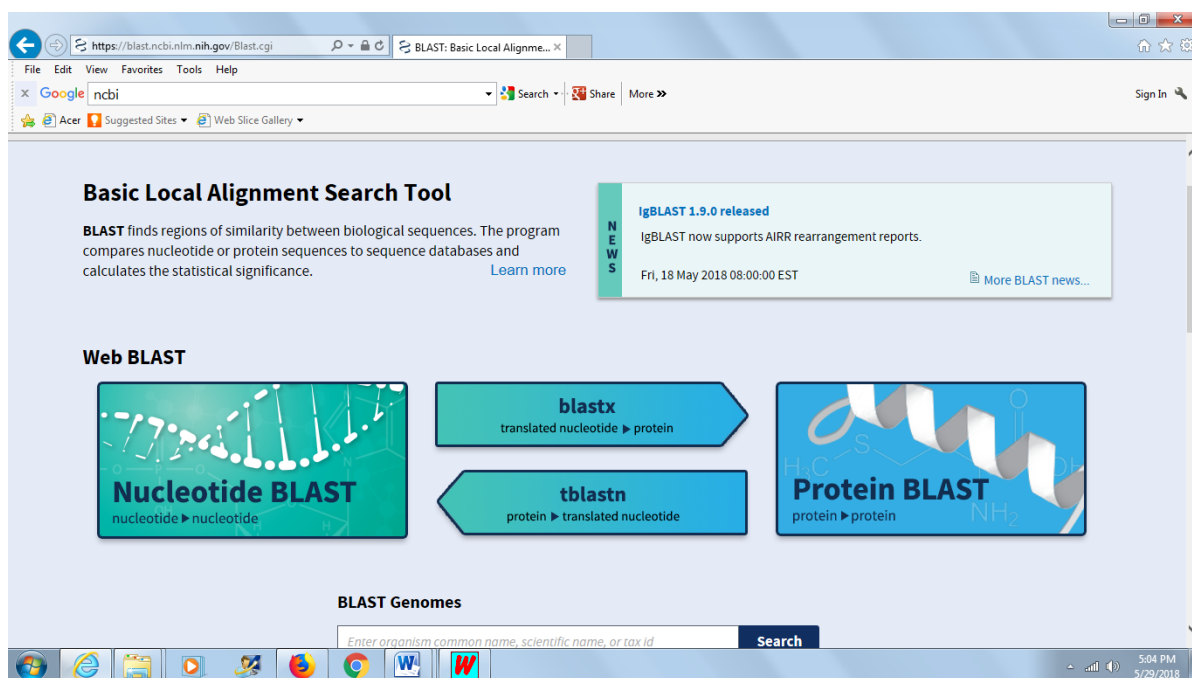
Results:

3. SEQUENCE RETRIEVAL USING BLAST

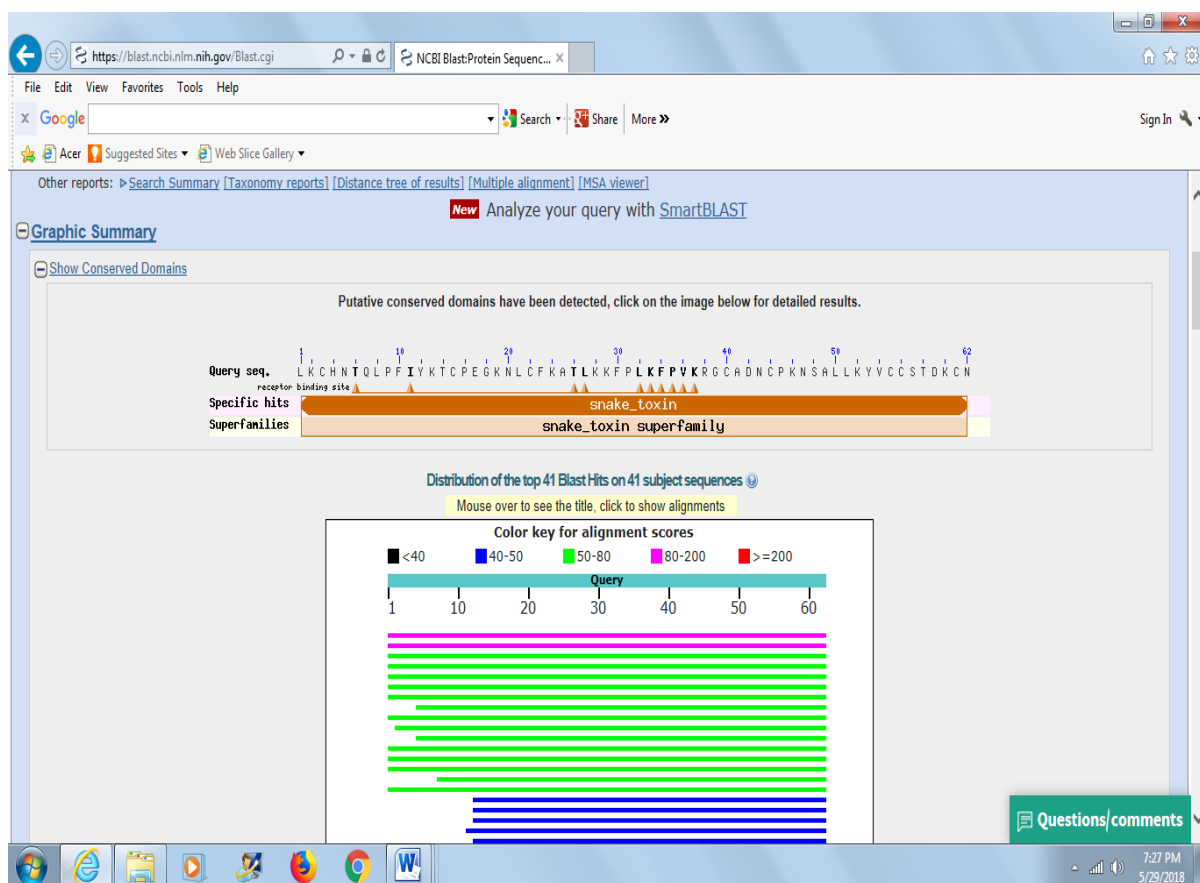
Aim: To retrieve similar sequences for a given protein/nucleotide primary structure using BLAST

Procedure

Retrieve a few numbers of protein/nucleotide sequences from NCBI/UniProt databases (<https://www.ncbi.nlm.nih.gov/>) and save the retrieved sequences in FASTA format and accession IDs for all the sequences as well. The 'sequence similarity search' can be executed using various BLAST algorithms.



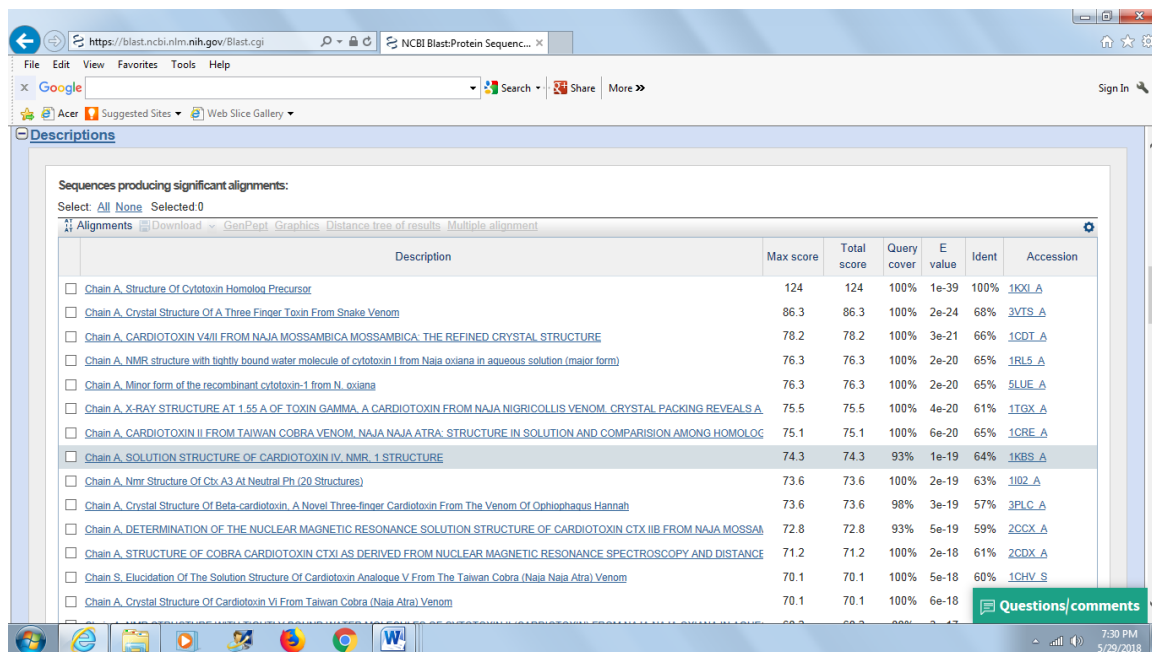
Using protein blast, similar sequences can be obtained either from primary or from structural databases. In general, for distant related and close related sequences, PSI-BLAST, BLASTP algorithms are preferred, respectively. For instance, a cardiotoxin (1CVO) was subjected to BLASTP similarity search and the outcomes were as shown herein below.



The sequences would be further analyzed on the basis of query coverage, total score, percentage of identities and E-values. The score values and other statistical

Class: III B.Sc Microbiology
COURSE NAME: BIOINFORMATICS PRACTICAL
Subject code: 16MBU512B

parameters for a few hits of the query sequences are depicted in the following illustration.



Sequences producing significant alignments:

Description	Max score	Total score	Query cover	E value	Ident	Accession
Chain A. Structure Of Cytotoxin Homolog Precursor	124	124	100%	1e-39	100%	1KXI_A
Chain A. Crystal Structure Of A Three Finger Toxin From Snake Venom	86.3	86.3	100%	2e-24	68%	3VTS_A
Chain A. CARDIOTOXIN V4/II FROM NAJA MOSSAMBICA MOSSAMBICA. THE REFINED CRYSTAL STRUCTURE	78.2	78.2	100%	3e-21	66%	1CDT_A
Chain A. NMR structure with tightly bound water molecule of cytotoxin I from Naja oxiana in aqueous solution (major form)	76.3	76.3	100%	2e-20	65%	1RL5_A
Chain A. Minor form of the recombinant cytotoxin-1 from N. oxiana	76.3	76.3	100%	2e-20	65%	5LUE_A
Chain A. X-RAY STRUCTURE AT 1.55 Å OF TOXIN GAMMA, A CARDIOTOXIN FROM NAJA NIGRICOLLIS VENOM. CRYSTAL PACKING REVEALS A	75.5	75.5	100%	4e-20	61%	1TGX_A
Chain A. CARDIOTOXIN II FROM TAIWAN COBRA VENOM, NAJA NAJA ATRA. STRUCTURE IN SOLUTION AND COMPARISON AMONG HOMOLOG	75.1	75.1	100%	6e-20	65%	1CRE_A
Chain A. SOLUTION STRUCTURE OF CARDIOTOXIN IV. NMR 1 STRUCTURE	74.3	74.3	93%	1e-19	64%	1KBS_A
Chain A. Nmr Structure Of Ctx A3 At Neutral Ph (20 Structures)	73.6	73.6	100%	2e-19	63%	1J02_A
Chain A. Crystal Structure Of Beta-cardiotoxin, A Novel Three-finger Cardiotoxin From The Venom Of Ophiophagus Hannah	73.6	73.6	98%	3e-19	57%	3PLC_A
Chain A. DETERMINATION OF THE NUCLEAR MAGNETIC RESONANCE SOLUTION STRUCTURE OF CARDIOTOXIN CTX IIB FROM NAJA MOSSAM	72.8	72.8	93%	5e-19	59%	2CCX_A
Chain A. STRUCTURE OF COBRA CARDIOTOXIN CTXI AS DERIVED FROM NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY AND DISTANCE	71.2	71.2	100%	2e-18	61%	2CDX_A
Chain S. Elucidation Of The Solution Structure Of Cardiotoxin Analogue V From The Taiwan Cobra (Naja Naja Atra) Venom	70.1	70.1	100%	5e-18	60%	1CHV_S
Chain A. Crystal Structure Of Cardiotoxin Vi From Taiwan Cobra (Naja Atra) Venom	70.1	70.1	100%	6e-18		

Similar types of data analyzes can be carried out for nucleotide sequences and as well for other BLAST algorithms.

Results:

4. Sequence alignment & phylogenetic analysis using clustalW & phylip

AIM

To align three or more sequences to find out structural and functional relationship between these sequences.

Key terms

Conserved regions: In biology, during the evolutionary time there may be some regions called group of bases or a sequence of nucleotides preserved as such in DNA, those sequences or a region, if seen in next generations called as Conserved regions.

Consensus Sequence: In a Nucleotide or an amino acid sequence, each base pair (an amino acid or a nucleotide) may occur more frequently at a particular region in different sequences of nature.

Theory

Sequence is a collection of nucleotides or amino acid residues which are connected with each other. Speaking biologically, a typical DNA/RNA sequence consist of nucleotides while a protein sequence consist of amino acids.

Sequencing is the process to determine the nucleotide or amino acid sequence of a DNA fragment or a protein. There are different experimental methods for sequencing, and the obtained sequence is submitted to different databases like NCBI, Genbank etc.

Methods of Sequencing:

Sequences stored in the database were obtained from different experimental methods. Most commonly used methods for DNA sequencing are Sanger Method and Maxam-Gilbert Method. Similarly Edman Degradation method and Mass Spectrometry technique are used for protein sequencing.

Sanger Method (dideoxy chain termination method): Here 4 test tubes are taken labelled with A, T, G and C. Into each of the test tubes, DNA has to be added in denatured form (single strands). Next a primer is to be added which anneals to one of

the strand in DNA template. The 3' end of the primer accommodates the dideoxy nucleotides [ddNTPs] (specific to each tube) as well as deoxy nucleotides randomly. When the ddNTP's gets attached to the growing chain, the chain terminates due to lack of 3'OH which forms the phospho diester bond with the next nucleotide. Thus small strands of DNA are formed. Electrophoresis is done and the sequence order can be obtained by analysing the bands in the gel based on the molecular weight. The primer or one of the nucleotides can be radioactively or fluorescently labeled also, so that the final product can be detected from the gel easily and the sequence can be inferred.

Maxam-Gilbert (Chemical degradation method): This method requires denatured DNA fragment whose 5' end is radioactively labeled. This fragment is then subjected to purification before proceeding for chemical treatment which results in a series of labeled fragments. Electrophoresis technique helps in arranging the fragments based on their molecular weight. To view the fragments, gel is exposed to X-ray film for autoradiography. A series of dark bands will appear, each corresponding to a radio labeled DNA fragment, from which the sequence can be inferred.

Edman Degradation reaction: The reaction finds the order of amino acids in a protein by cleaving each amino acid from the N-terminal without disturbing the bonds in the protein. After each cleavage, chromatography or electrophoresis is done to identify the amino acid.

Mass Spectrometry: It is used to determine the mass of particle, composition of molecule and for finding the chemical structures of molecules like peptides and other chemical compounds. Based on the mass to charge ratio, one can identify the amino acids in a protein.

Sequence Alignment:

When a new sequence is found, the structure and function can be easily predicted by doing sequence alignment. Since it is believed that, a sequence sharing a common ancestor would exhibit similar structure or function. Greater the sequence similarity, greater is the chance that they share similar structure or function.

Sequence alignment can be of two types i.e., comparing two (pair-wise) or more sequences (multiple) for a series of characters or patterns. Alignment of three or more biological nucleotides or protein sequences, simply defines multiple sequence alignment. The genes which are similar may be conserved among different species.

Take these identical or similar set of genes to perform multiple sequence alignment. Through this, we can easily identify the most evolutionarily conserved regions that play critical role in functionality of a specified gene.

During the evolutionary time, the genes may have got altered at sequence level, which results in alteration of function. Multiple sequence alignment can identify alterations in the function and the causes for that alteration at sequence level.

At protein level, information regarding structure and function of proteins can be obtained by multiple sequence alignment. It would be helpful in getting new domains or motifs with biological significance.

We can find many tools for multiple sequence alignment like MSA DIALIGN, CLUSTAL series, MAFT, MUSCLE, T-Coffee, BlastAlign, etc.

Table 1: Summary of multiple sequence alignment programs

Summary of MSA programs that we consider to be the best currently available		
Program	Advantages	Cautions
CLUSTALW	Uses less memory than other programs	Less accurate or scalable than modern programs
DIALIGN	Attempts to distinguish between alignable and non-alignable regions	Less accurate than CLUSTALW on global benchmarks
MAFFT, MUSCLE	Faster and more accurate than CLUSTALW; good trade-off of accuracy and computational cost. Options to run even faster, with lower average accuracy, for high-throughput applications.	For very large data sets (say, more than 1000 sequences) select time- and memory-saving options
PROBCONS	Highest accuracy score on several benchmarks	Computation time and memory usage is a limiting factor for large alignment problems (>100 sequences)
ProDA	Does not assume global alignability; allows repeated, shuffled and absent domains.	High computational cost and less accurate than CLUSTALW on global benchmarks
T-COFFEE	High accuracy and the ability to incorporate heterogeneous types of information	Computation time and memory usage is a limiting factor for large alignment problems (>100 sequences)

CLUSTALW / CLUSTAL Omega

Pair wise sequence alignment has been approached with dynamic programming between nucleotide or amino acid sequences. The same approach can be used for alignment of 'n' number of sequences. But this program is limited to pair wise, since there will be exponential increase in memory, number of steps with respect to number of sequences. Because of such limitations with dynamic programming, researchers came up with an approach called '*progressive method*' to align three or more sequences.

Progressive method was first suggested by Feng and Doolittle in 1987. It compares only a pair of sequences together at a time using the following steps:

Using the standard dynamic programming algorithm on each pair, we can calculate the $(N*(N-1))/2$ (N is total number of sequences) distances between the sequence pairs.

From the distance matrix obtained using the clustering algorithm, construct a guide tree.

From the tree obtained, align the first node to the second node. After fixing the alignment, add another sequence or the third node. Iterate the step until all the sequences are aligned. When a sequence is aligned to a group or when there is alignment in between the two groups of sequences, the alignment is performed that had the highest alignment score. The gap symbols in the alignment replaced with a neutral character. Where it helps to guide the alignment of sequence- alignment and alignment –alignment.

Working of Algorithm

Multiple sequence alignment can be done through different tools. CLUSTALW is one among the mostly accepted tool. Recently a newer version of CLUSTALW came out with the name CLUSTAL Omega which is available from the European Bioinformatics institute www.ebi.ac.uk/Tools/clustalw2/

Higgins D has written the first program of CLUSTAL, considering memory and time various CLUSTAL series of programs have came up and presently used version is CLUSTALW, which came up with dynamic programming and progressive alignment methods.

CLUSTALW uses the progressive algorithm, by adding the sequence one by one until all the sequences are completely aligned.

Steps for CLUSTAL algorithm

1. Calculate all possible pairwise alignments, record the score for each pair.
2. Calculate a guide tree based on the pairwise distances (algorithm: Neighbor Joining).
3. Find the two most closely related sequences
4. Align the sequences by progressive method
 - i. Calculate a consensus of this alignment
 - ii. Replace the two sequences with the consensus

Class: III B.Sc Microbiology

COURSE NAME: BIOINFORMATICS PRACTICAL

Subject code: 16MBU512B

- iii. Find the two next-most closely related sequences (one of these could be a previously determined consensus sequence).
- iv. Iterate until all sequences have been aligned
- v. Expand the consensus sequences with the (gapped) original sequences
Report the multiple sequence alignment

RESULTS

Phylogenetic analysis using phylip

AIM:

To find the evolutionary relationship between different organisms based on the time scale and to analyze the changes that occurred in an organisms using PHYLIP.

Key words:

Phylogenetic analysis: Analyze the evolutionary relationships between different organisms and this analysis would help to find out the changes that occurred in organisms during the evolution.

Boot Strapping: It is a way to test the reliability of Dataset.

Query: User can give input called as a query. This can be either a protein or nucleotide sequence.

Rooted tree: A tree which is having a special node as main node also called the root. A tree without root is treated as a free tree.

Tree topology: Tree topology refers to the arrangement of phylogenetic tree.

Theory :

PHYLIP is a complete phylogenetic analysis package which was developed by Joseph Felsenstein at University of Washington. PHYLIP is used to find the evolutionary relationships between different organisms. Some of the methods available in this package are maximum parsimony method, distance matrix and likelihood methods. The data is presented to the program from a text file, which is prepared by the user using common text editors such as word processor, etc. Some of the sequence analysis programs such as ClustalW can write data files in PHYLIP format. Most of the programs look for the input file called "infile" -- if they do not find this file, then they ask the user to type in the file name of the data file. Before starting the computation, the

program will ask the user to set options (optional) through a menu. Output is written into special files with names like outfile and outtree.

PHYLIP file format :

- The input files have information about the number of sequences, nucleic acids and amino acids.
- The sequence has 10 characters length. Spaces can be added to the end of the short sequences to make them long.
- Gaps can be represented as ‘-’.
- Missing data can be represented as ‘?’
- Spaces between the alignments are allowed usually after every 10 bases.

Example:

4 1061

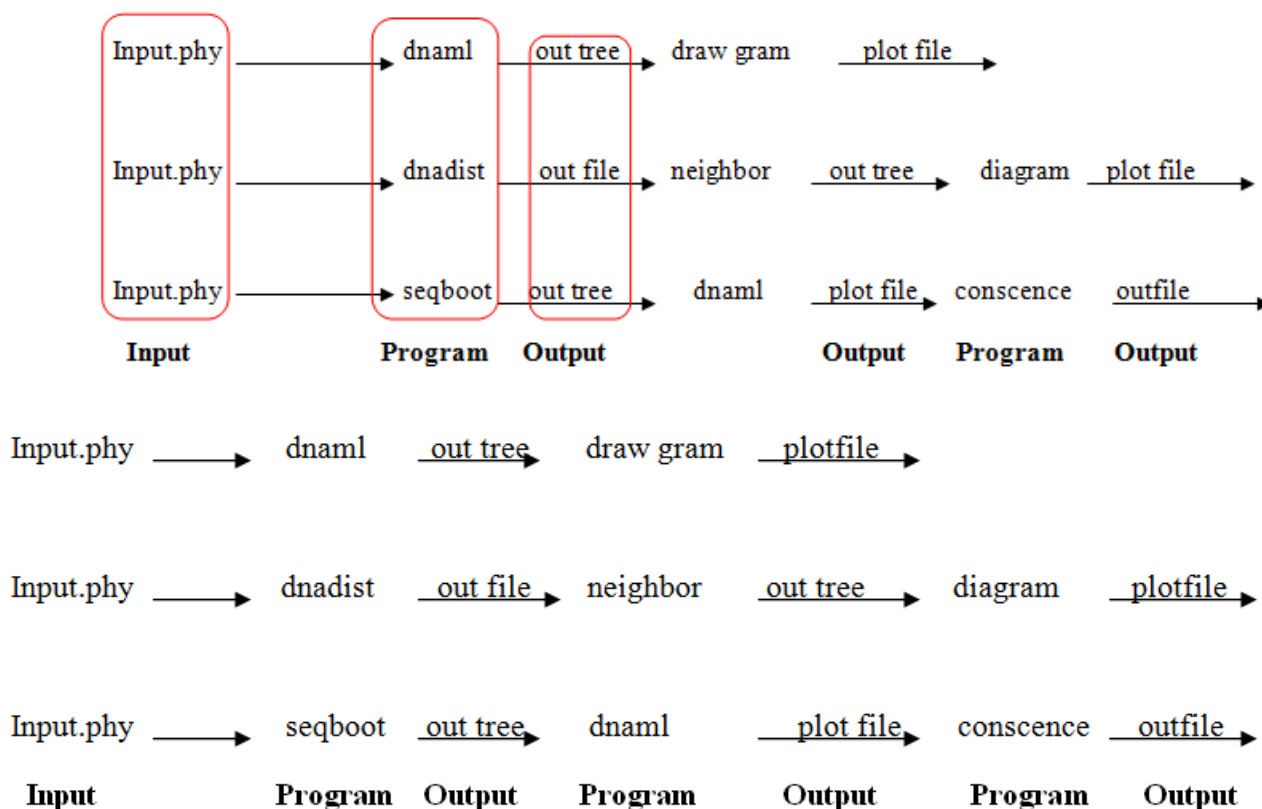
```
GGCCTGCTCT  GCCTG-----  CCCTGGCTTC AAGAGGG—C  AGTGCCTTCC
AGACGGAAAA AAAGGAAAAG  TCACGACATC CCCAA- --- C  AGCCCCTCCA
-----  ---?-----G  CCGTGGTA--  -----  ---- GATTTG
```

4 indicates number of species taken for phylogenetic analysis

1061 indicates number of characters.

PHYLIP program :

The PHYLIP programs have to be run in sequential manner, output of one program is used as input of another program. User has to know how to use these programs in a sequential manner. Simple examples to run PHYLIP programs are given in the below flowcharts.



There are three different ways to determine the nucleotide and amino acid sequence:

1. Maximum parsimony method
2. Distance methods
3. Maximum likelihood methods

Maximum parsimony method: It is a character-based method which infers a phylogenetic tree, by minimizing the total number of evolutionary steps or the total tree

length for a given set of data. It is also referred to as sequence based tree reconstruction method.

Distance methods: Evolutionary distances are calculated for all operational taxonomic units and build tree where distance between the operational taxonomic units matches these distances.

Maximum likelihood method: Refers a model of sequence evolution, finds the tree which gives the highest likelihood of the observed data.

Programs used in PHYLIP program:

The following are the methods available in PHYLIP program.

Dnapars: Estimates the phylogeny using parsimony method from nucleic acid sequence.

Dnamove: It is an interactive process used for construction of phylogeny from nucleic acid sequences using parsimony method.

Dnapenny: Estimates the parsimonious phylogeny for nucleic acid sequences which uses branch and bound theory.

Dnacomp: States the phylogeny of nucleic acids and searches for the largest sites which have uniquely evolved on the same tree.

Dnainvar: Computes the nucleic acid sequence which tests the alternative tree topologies. The programs tabulate (chart) the frequencies of occurrences of different nucleotide patterns.

Dnaml: Estimates the phylogenies from nucleotide sequences by maximum likelihood method without assuming molecular clock. Molecular clock defines to calculate timings of evolutionary events.

Dnamlk: It estimates the phylogeny using maximum likelihood method, it assumes the molecular clock.

Dnadist: Dnadist calculates the pair wise distances between the sequences. It also makes a table of percentage similarity among different sequences.

Seqboot: Reads a dataset, and produces multiple datasets by bootstrap resampling. Most of the programs in the current version allow processing of multiple datasets; this can be used together with the consensus tree program CONSENSE.

Concense: Computes consensus trees by the consensus tree method, which can allow one to easily find the consensus tree.

Protpars: Estimates the phylogenies from protein sequences which use parsimony method.

Protdist: It measures the distances of protein sequences using maximum likelihood method which is based on the PAM matrix, JTT model and PBM model. It can give the percentage of similarity among the sequences.

Promol: Estimates phylogeny from amino acid sequences by using maximum likelihood methods. The program allows us to find different changes at known sites. Proml is without a molecular clock.

Promlk: This estimates the phylogeny from amino acid sequence by using maximum likelihood method. It assumes a molecular clock. Molecular clock defines to calculate timings of evolutionary events.

Restml: Estimates the phylogeny using maximum likelihood method with restriction sites data. It does not allow the rate difference between the transitions and transversions.

Restdist: It estimates the phylogeny and calculates the distance from the restriction site data and restriction fragment data.

Fitch: Estimates phylogenies from distance matrix data under “additive tree model”. It uses fitch-Margoliash and some related least square criteria or the distance matrix method. It does not assume the evolutionary clock. The program computes the distance from molecular sequences, fragment distances, and genetic distances calculated from gene frequencies.

Kitsch: Estimates phylogenies from distance matrix data under “Ultrametric model” same as the additive tree model except the evolutionary clock is measured. It is similar to Fitch algorithm.

Neighbor: Neighbor joining is a distance matrix method which will produce an unrooted tree without the assumption of an evolutionary clock. This method is very fast, it can handle large data sets.

Dnadist : It’s a distance matrix method which can be used to find the distances between nucleic acid sequences. This can give the percentage similarity among the sequences.

Protdist: Computes distance between the protein sequences uses maximum likelihood method.

Restdist: Computes the distance calculated from restriction sites data and restriction fragment data.

Drawgram: It estimates the rooted phylogeny, cladograms, circular trees in a wide variety.

Drawtree: It estimates the unrooted phylogeny similar to Drawgram.

RESULTS

5. Picking out a given gene from genomes using Genscan or ORF prediction

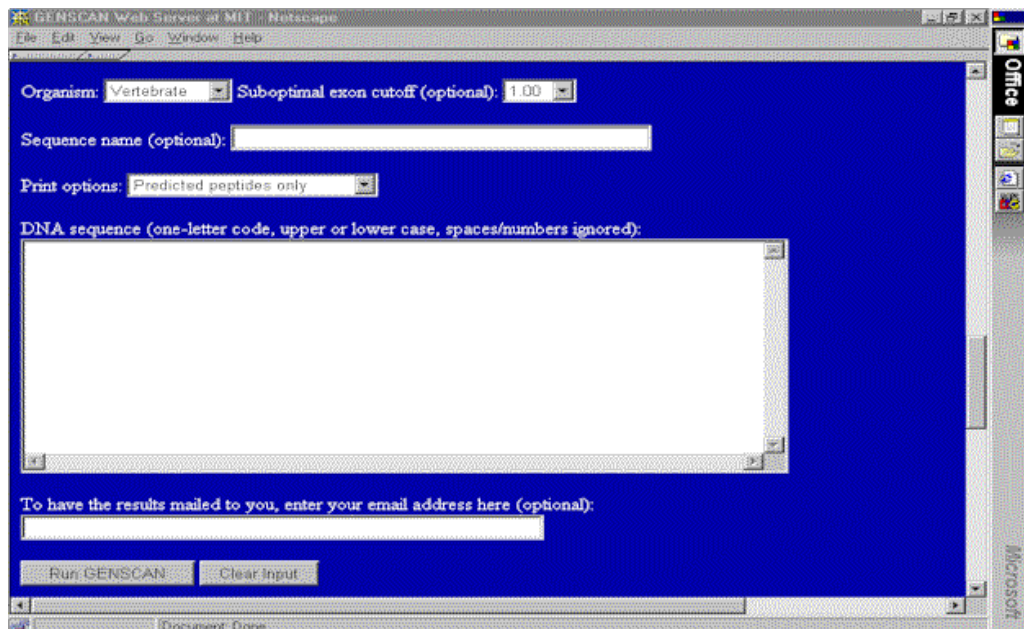
AIM

To identify a gene in a fragment of human genomic DNA using GENSCAN software.

PROCEDURE

GENSCAN - is particularly useful as it can handle upto 1Mb of sequence, programs such as Fgenes and Fgenesh are considered overall to be slightly more accurate at predictions overall, but will only take 10kb of sequence in any one analysis.

1. Open the DNA sequence file arabidopsis.txt in notepad and copy the sequence to the clipboard.
2. Access a Web-based version of GENSCAN.
3. You should see the screen shown below. Paste the sequence into the DNA sequence box



The program has certain parameters that need to be set before you run it. Select Arabidopsis from the dropdown selection of organisms. Everything else can be left at its default value.

4. Limitations are many. Splice sites may be a poor fit to the consensus used by the algorithm, or commonly the gene will use different splice acceptors and donors at different times.

For a human example:

- i) load the following genomic sequence from the human paraplegin gene
- ii) copy the sequence to the clipboard
- iii) enter the GENESCAN web site and paste in your sequence performing the search as before
- iv) count the number of exons predicted - you can view this as a pdf or a postscript file
- v) analyse the sequence from chromosome 16 containing the paraplegin gene sequence - use the "Find" on the Edit menu (CtrlF) to locate the description.
- vi) determine how many splice forms are actually found and how many exons each contains.

RESULTS

Class: III B.Sc Microbiology
COURSE NAME: BIOINFORMATICS PRACTICAL
Subject code: 16MBU512B

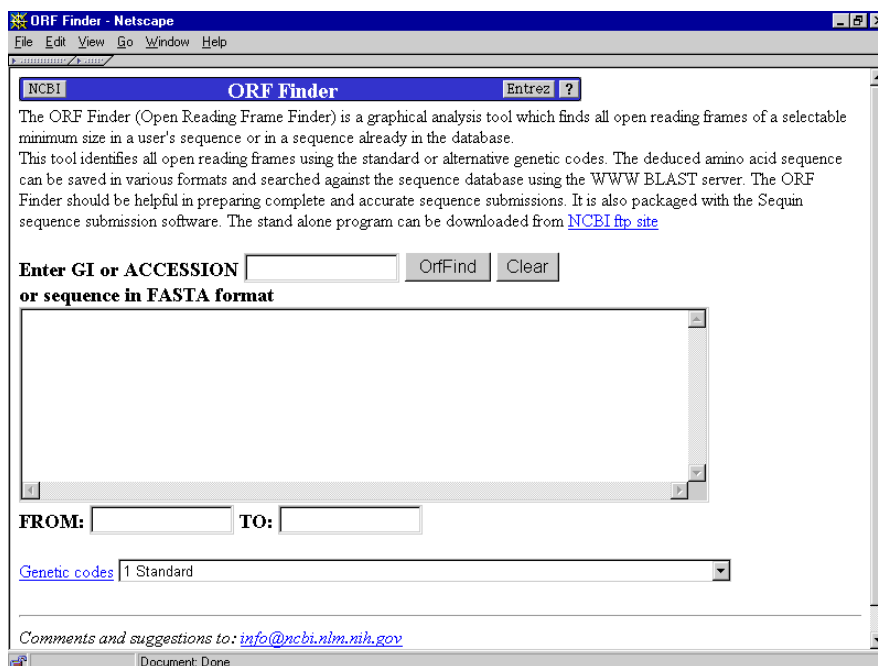
ORF PREDICTION

AIM

To identify ORFs using the ORF Finder software.

ORF Finder

- Open the DNA sequence bacteriophage.txt in notepad and copy the sequence to the clipboard.
- Access the Web-based version of ORF finder at the National Centre for Biotechnology Information <http://www.ncbi.nlm.nih.gov/gorf/gorf.html>.



1. Click in the text box and paste the sequence (ctrl V) from the clipboard
2. You must decide which option to use for the genetic code prior to carrying out the analysis. Click the drop down selection box and select the appropriate genetic code
3. Click the ORF Find button and you should soon see the results of your analysis in a screen similar to the one below

Class: III B.Sc Microbiology

COURSE NAME: BIOINFORMATICS PRACTICAL

Subject code: 16MBU512B

RESULTS

PRIMER DESIGNING

AIM

To find primers from a given nucleotide sequence

Theory

DNA (Deoxyribonucleic acid) is the genetic material that contains the genetic information for the development and maintaining all functions in living organisms. The information is stored as genetic codes using four types of nucleotides. They are adenine (A), guanine (G), cytosine(C) and thymine (T). In two strands of DNA, adenine always pair with thymine and guanine pair with cytosine. Each of these base pairs will bond with a sugar and phosphate molecule to form a nucleotide. The base pairing of DNA will result in a ladder shape structure of these strands which is called a double helix. DNA replication is a mandatory biological process to maintain life where a single set of DNA gives rise to two copies of DNA. The process of DNA replication is catalyzed by an enzyme named DNA polymerase. These enzymes are able to add nucleotides only to an existing DNA strands.

A primer can be defined as short nucleic acid sequences. It can act as a starting point for DNA synthesis. The polymerase enzyme starts adding nucleotides in the 3'-end of the primer. Process like DNA sequencing (to determine the exact order of nucleotides in a DNA) and polymerase chain reaction (or PCR, used to amplify DNA sequences) require DNA primers whereas for natural DNA replication short sequences of RNA is used as primer. Usually the length of the primer is 18 to 24 nucleotides.

Since primer has a very important role in most of the experiments in biochemistry and molecular biology, it is essential to choose these sequences correctly.

ApE is a tool which can be used to find primers from a given sequence. Some of the features of ApE are given below:

1. Runs in Linux and Windows OS
2. Able to highlight the text using pre-defined custom feature libraries which allows for quick searching and highlighting of all the available primers, quick annotation of the sequence, searching sequences to be annotated etc.

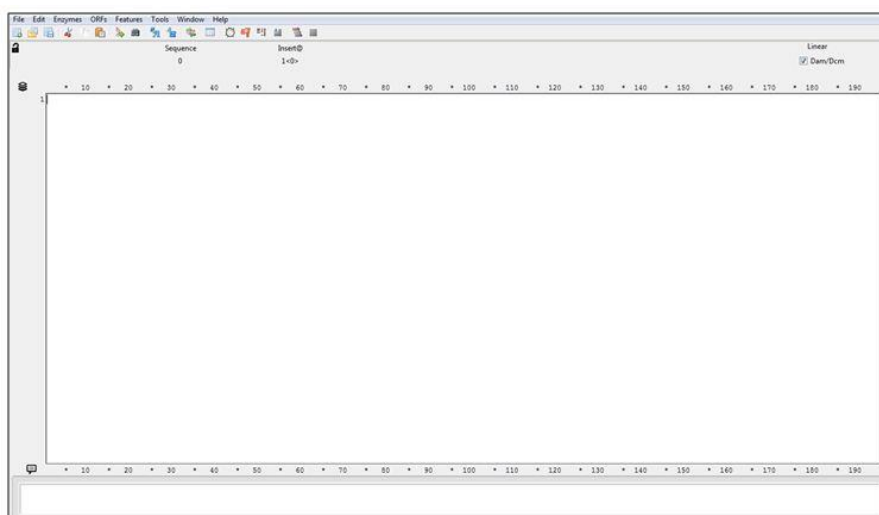
Class: III B.Sc Microbiology

COURSE NAME: BIOINFORMATICS PRACTICAL

Subject code: 16MBU512B

3. Searches for possible primers that matches with the length, GC content, Tm etc. which user specifies.
4. Reads file formats such as FASTA, Genbank, EMBL and formats from DNA Strider.
5. Save files in Genbank format or DNA Strider-compatible format.
6. Allows direct BLAST search in NCBI.
7. Highlights restriction sites from the input.
8. Shows Tm (melting point), GC content of the sequence, ORF in a sequence.
9. Translate the input sequence by means of an optional DNA alignment.
10. Draw and highlight the graphical maps from Genbank and EMBL files through its feature annotations.

Figure 1 shows the GUI of ApE.



6. PROTEIN STRUCTURE PREDICTION- primary structure analysis, secondary structure prediction using psi- pred, homology modeling using Swissmodel.

AIM

To understand the basis of homology modelling and 3D structure of protein using SWISS MODEL

PROCEDURE

SWISS-MODEL is a fully automated protein structure homology-modeling server, accessible via the ExPASy web server, or from the program DeepView (Swiss Pdb-Viewer). The purpose of this server is to make Protein Modeling accessible to all biochemists and molecular biologists worldwide. SWISS-MODEL (<http://swissmodel.expasy.org/>) is an automated system for modelling the 3D structure of a protein from its amino acid sequence using homology modelling techniques.

RESULTS

Secondary structure prediction

AIM

To visualize the secondary structure of a protein.

THEORY

Proteins are important biological macromolecules, usually considered as fundamental units of a cell which play a vitally important role in different cell functions. The function of a protein is very specific and dependent on the molecule to which it binds. The protein structure is classified into primary, secondary, tertiary and quaternary. These molecules, form a linear chain of amino acids initially, and then fold into secondary, tertiary and quaternary structures. The different secondary structures of a protein are alpha helices, beta pleated sheets and loops. During the due course of evolution, some regions of the protein remain conserved which are regarded as motifs, play a key role in determining the function of that particular protein. In simple terms, researchers say that the molecules having similar structure will have a similar function. To understand the structural biology, visualization of complex macromolecular structure is essential. PyMol is an open source, three dimensional visualization tool to view the macromolecular structures like proteins and nucleic acids.

All the proteins are made up of long chain of amino acids that fold into a 3-D shape. Amino acids are organic compounds that contain a hydrogen atom, a carbon, two functional groups and a side chain R group. There are almost 20 amino acids found in human body which vary in their R groups. Amino acids are linked to each other by peptide bond. A peptide bond is formed when the carboxyl group of one amino acid linked to the amino group of another molecule through a covalent bond.

The primary structure of a protein is made up of a linear sequence of amino acids. It is synthesized during the translation process of DNA to mRNA. DNA (Deoxyribonucleic acid) is the genetic material that contains all the genetic information for the development and maintaining all functions in all living organisms. The information is stored as genetic codes using four types of bases. They are adenine (A), guanine (G), cytosine(C) and thymine (T). RNA is differing from DNA only in 1 base pair i.e. in RNA, it is uracil (U) instead of thymine. mRNA

(messenger RNA) is a molecule of RNA which is forming from DNA transcription process. During the transcription process, DNA is transcribed to mRNA i.e. thymine is replaced by Uracil.

The intermolecular and intra-molecular hydrogen bonding between the amide groups in primary structure of protein form secondary structure. The attraction of hydrogen molecule towards electro negative atom (N, F, O etc) within same molecule is called intra-molecular hydrogen bonding and formed between two different molecules. Alpha helices and beta sheets are two important secondary structures in protein (Figure 1). Alpha helix is a right-handed conformation, beta sheets or strands which may be located parallel or anti parallel to each other (each strand).

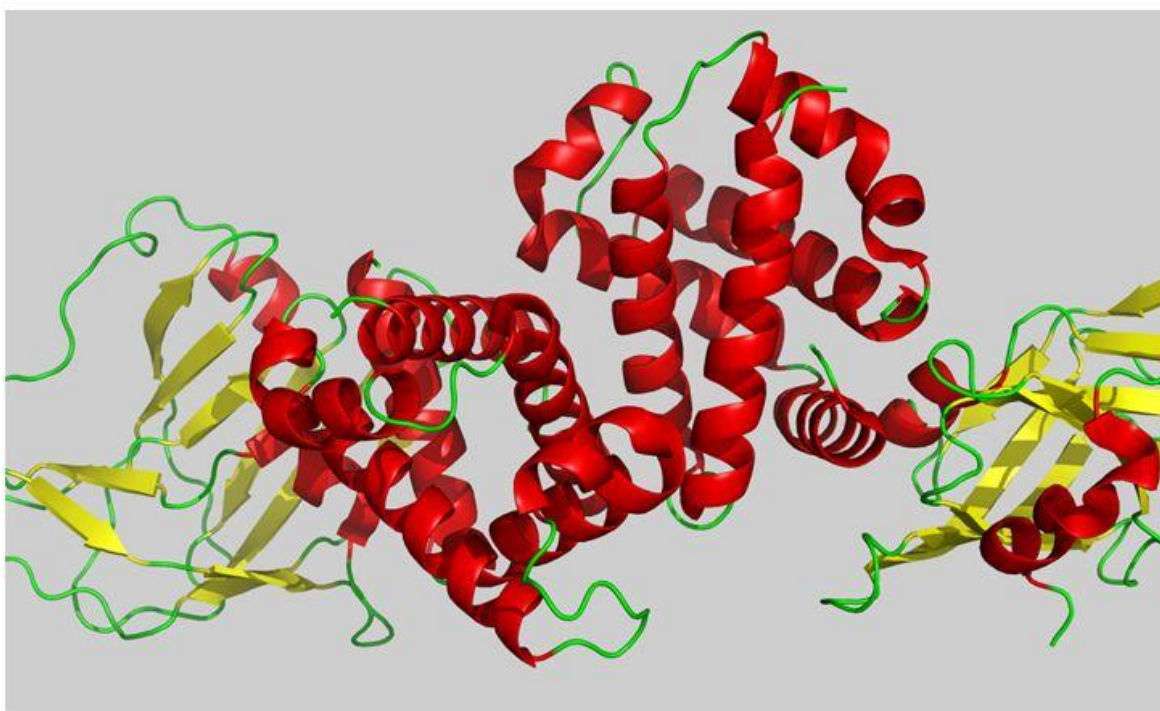


Figure 1: Example for Secondary structure of a protein - Human Alpha-Hemoglobin

PyMol

PyMOL is Copyrighted © software DeLano Scientific LLC, San Carlos, California (U.S.A.). It is free for all to use, modify, and redistribute. Warren Lyford

Delano, developed the interesting molecular visualization tool PyMol, which has been regularly used by crystallographers (who finds out the macromolecular structures through the technique crystallography). In many journal papers (research papers), we can see the structural images created using PyMol. Users can get high quality images and animations of biological macromolecules like proteins. PyMol is freely available, since it is an open source visualization tool with python (programming language) interpreter. This visualization software have inbuilt demonstration of what it does (Figure 2).

To see the demo, go to the standard menu bar, select the wizard menu, from the submenu select demo with representations.

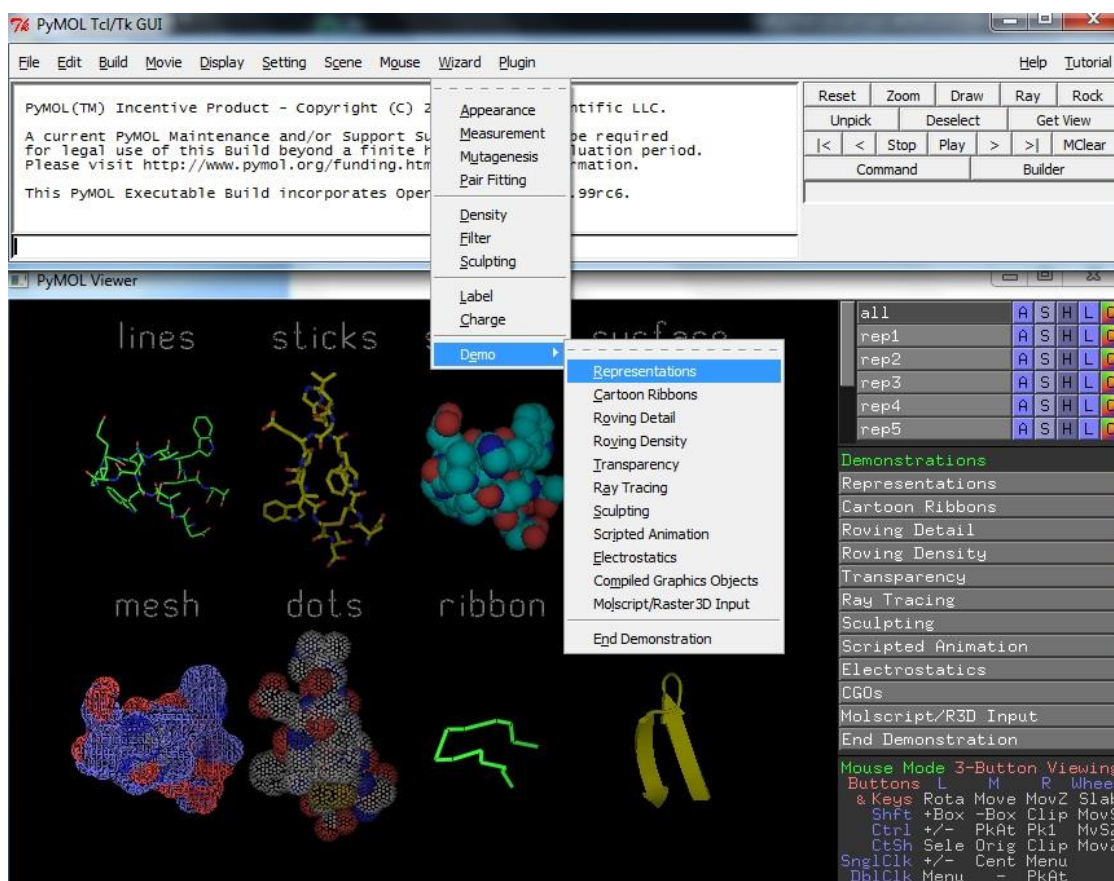


Figure 2: Screenshot to view the demo of representations

PyMol is used to visualize the .pdb files, which are mostly available from the protein databank. It contains structures extracted from techniques like x-ray

crystallography, NMR Spectroscopy. With the help of initial experimental data, the structural biologists use these techniques or methods to determine the location of each atom relative to each other in a molecule. The x-ray crystallography has x-ray diffraction pattern data, in NMR spectroscopy the scientist need the information of local conformation and distance between the atoms that are close to one another.

In x-ray crystallography the protein is purified and then crystallized, which is then treated to x-ray beams. These crystallized proteins are analyzed to get the distribution of electrons in proteins, by diffracting the x-ray beams into one or other characteristic pattern of spots. Once we get the distribution of electrons, which gives the map of electron density is interpreted to determine the location of each atom.

In NMR Spectroscopy they use, strong magnetic field to determine the protein structure. The protein is purified and subjected to strong magnetic field which is then probed with the radio waves. Thus we can observe the resonances, which can be analyzed to give the list of atomic nuclei that are close to each other. This list is used to build a model of protein that can show a location of each atom.

RESULTS

Predict the structure of a protein

AIM:

To predict the structure of a protein from its amino acid sequence with experimentally resolved structures of related proteins.

Theory:

It is the process of predicting a structure from sequence which should be comparable with the experimental results. The structure of a protein is determined by its amino acid sequence. These structures of protein can be obtained from X-ray crystallography, NMR spectroscopy or from theoretical methods using real experiments or by homology modeling. But real experiments failed to provide high resolution information for the majority of proteins and NMR and other analysis too failed due to high protein dimension. During the evolutionary process, the structure stays more conserved rather than a sequence. These protein which share similar sequence form identical structure and distantly related protein fold into similar structure.

1. Template recognition and initial alignment

The sequence of similarity can be searched using BLAST or Psi blast or fold recognition methods and align with the known structures in PDB. PDB which is the largest database contains only experimentally resolved structure. BLAST allows comparing a query sequence with a database such as PDB and identifying the best sequence which shares a high degree of similarity. The sequence of similarity of each line is summarised with its E-value (Expected value) which is closer to zero, have high degree of similarity. The E-value describes the number of hits one can “expect” when searching through a database of a particular size. The sequences which fall under safe zone are expected to be getting good structure than twilight zone and midnight zone. After identifying one or more possible template, alignment correction is performed. Sometimes it is difficult to align two sequences that have percentage identity which is low. Such cases, one can use other sequences from homologous proteins to solve this problem. Multiple Sequence Alignment programs such as CLUSTALW align sequences by insertions and deletions. Alignment correction is the critical step in homology modeling, otherwise which in turn creates a defective model.

2. Backbone generation

The backbone generation from the aligned regions can be done using modelling tools such as Modeller or CASP. The actual experimentally determined structures contain manual errors due to poor electron density in the map. Therefore a good model has to be chosen with less number of errors.

3. Loop Modeling

In most cases, alignment between model and template sequence contain gaps. By means of insertions and deletions with some conformational changes to the backbone it can be modelled, although it rarely happens to secondary structures. So it is safe to shift the insertion and deletions of the alignment, out of helices or strands and placing them in loops or coils. But this loop conformational change is difficult to predict due to many reasons like

1. Surface loops tend to be involved in crystal contacts, leading to a significant conformational change between template and target.
2. The interchange of the side chains can lead to change in the orientation and spatial arrangement especially when it is an interchange between small and a bulky group.
3. Proline and glycine are an exception when a Ramachandran plot is considered. Proline has a restriction in the plot due to its 5 membered ring whereas glycine has a hydrogen atom as its side chain which is very difficult to predict from the plot. This makes it difficult for detect mutations that have happened to loop residue from/to either glycine or proline.

There are two main ways to overcome this and model the loop region:

1. Knowledge based:
User can search PDB for known loops with endpoints that match the residues between loops that have to be inserted and simply copy the loop conformation.
2. Energy based:
The quality of a loop is determined with energy function and minimizes the function using Monte Carlo or molecular dynamics to find the best loop conformation.

4. Side Chain Modeling

Proteins that are structurally similar, have similar torsion angle about Ca-Cb bond (ψ angle) when comparing with side chain conformations. In such cases, copying conserved residues entirely from the template to the model will result in higher accuracy than copying the backbone or re-predicting side chains. Side chain conformations are partially knowledge based which uses libraries of rotamers extracted from high resolution X ray structures. To build a position-specific rotamer library, one can take high-resolution protein structures and collect all stretches of three to seven residues (method dependant) with a given amino acid at the center. Prediction accuracy is usually quite high for residues in the hydrophobic core, where more than 90% of all ψ angles fall with 20 σ of experimental values, it is much lower for surface residues, where the percentage is often lower than 50%.

There are two reasons for this:

1. Flexible side chains on the surface tend to adopt multiple conformations, which are additionally influenced by crystal contacts.
2. Energy functions used to score rotamers can easily handle hydrophobic packing in the core (Van der Waals interactions), but are not accurate enough to get complicated electrostatic interactions on the surface.

5. Model Optimization

Sometimes the rotamers are predicted based on incorrect backbone or incorrect prediction. Such cases modeling programs either restrain the atom positions and/or apply only a few hundred steps of energy minimization to get an accurate value. This accuracy can be achieved by 2 ways.

1. Quantum force field: To handle large molecules efficiently force field can be used, energies are therefore normally expressed as a function of the positions of the atomic nuclei only. Van der Waals forces are, for example, so difficult to treat, that they must often be completely omitted. While providing more accurate electrostatics, the overall precision achieved is still about the same as in the classical force fields.
2. Self-parametrizing force fields: The precision of a force field depends to a large extent on its parameters (e.g., Van der Waals radii, atomic charges). These

parameters are usually obtained from quantum chemical calculations on small molecules and fitting to experimental data, following elaborate rules (Wang, Cieplak, and Kollman, 2000). By applying the force field to proteins, one implicitly assumes that a peptide chain is just the sum of its individual small molecule building blocks—the amino acids. To increase the precision of the force field, the following steps can be used. Take initial parameters (for example, from an existing force field), change a parameter randomly, energy minimize models, see if the result improved, keep the new force field if yes, otherwise go back to the previous force field.

6. Model Validation

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

1. The percentage identity between the template and the target.

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

2. The number of errors in the template.

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

Modeller (Sali and Blundell 1993)

Modeller is a program for comparative protein structure modelling by satisfaction of spatial restraints. It can be described as “Modeling by satisfaction of restraints” uses a set of restraints derived from an alignment and the model is obtained by minimization of these restraints. These restraints can be from related protein structures or NMR experiments. User gives an alignment of sequences to be modelled with known

structures. Modeller calculates a model with all non hydrogen atoms. It also performs comparison of protein structures or sequences, clustering of proteins, searching of sequence databases.

RESULT