SYLLABUS

1. **Biological databases and data retrieval**
   - Sequence retrieval (protein and gene) from NCBI
   - Structure download (protein and DNA) from PDB
   - Molecular file formats-FASTA, GenBank, Genpept, GCG, CLUSTAL, Swiss-Prot, PIR

2. Sequence alignment
   - BLAST suite of tools for pairwise alignment
   - Multiple sequence alignment using CLUSTALW

3. Phylogenetic analysis
   - Generating phylogenetic tree using PHYLIP

4. Protein structure prediction and analysis
   - Primary sequence analyses (Protparam)
   - Secondary structure prediction (GOR, nnPredict, SOPMA)
   - Tertiary structure prediction (SWISSMODEL)
   - Protein structure evaluation-Ramachandran map (PROCHECK)

5. Gene structure prediction and analysis
   - Gene prediction using GENSCAN and GLIMMER

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. E.BRINDHA
SUBJECT NAME: BIOINFORMATICS PRACTICAL
SUB.CODE:18BCU414A
SEMESTER: IV CLASS: I B.Sc(BC)

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

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

SYLLABUS

1.**Biological databases and data retrieval**

Sequence retrieval (protein and gene) from NCBI

Structure download (protein and DNA) from PDB

Molecular file formats-FASTA, GenBank, Genpept, GCG, CLUSTAL, Swiss-Prot, PIR

2. Sequence alignment

BLAST suite of tools for pairwise alignment

Multiple sequence alignment using CLUSTALW

3. Phylogenetic analysis

Generating phylogenetic tree using PHYLIP

4. Protein structure prediction and analysis KAHE

Primary sequence analyses (Protparam)

Secondary structure prediction (GOR, nnPredict, SOPMA)

Tertiary structure prediction (SWISSMODEL)

Protein structure evaluation-Ramachandran map (PROCHECK)

5. Gene structure prediction and analysis

Gene prediction using GENSCAN and GLIMMER

Experiment No: 1

Biological Databases and data retrieval

Introduction

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

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

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

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

Sequence retrieval (protein) from NCBI

Aim

To retrieve the protein sequence of ' _____' from National Centre for Biotechnology Information (NCBI).

Resource URL

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

Procedure

1. Open NCBI homepage
2. In search column protein was selected
3. In the same column protein of interest was given
4. The protein sequences were retrieved using FASTA format.
5. Sequences were copied and pasted in the notepad.

Result

The proteinsequences of ' _____' was successfully retrieved.

Sequence retrieval (gene) from NCBI

Aim

To retrieve the ‘ _____’ gene sequence from National Centre for Biotechnology Information (NCBI).

Resource URL

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

Procedure

1. Open NCBI homepage
2. In search column gene was selected
3. In the same column gene of interest was given
4. The gene sequences were retrieved using FASTA format.
5. Sequences were copied and pasted in the notepad.

Result

The ‘ _____’ gene sequence was successfully retrieved.

Structure download (protein) from Protein Data Bank

Aim

To retrieve structural information about the protein "_____" from protein databank.

Description

Protein Data Bank (PDB) is a structural databank. The PDB achieve contains macromolecules structural data of proteins, nucleic acids, protein-nucleic acids complexes. PDB databank is freely available worldwide. PDB gives information regarding the sequence 3D structure of compounds using various methods. PDB search can be performed using the output from one search as input. A search can return a single structure or multiple structures.

Procedure

1. Go to http://www.rcsb.org/pdb/website
2. Enter protein name and click search.
3. Note down the title, compound, experiment method, classification, source, polymer chains, residues, atoms and chemical composition from summary information.
4. Click on geometry and note down the dihedral angles, common bond angles and bond length of the respective chains.
5. Click on the structural details and save the structural details.
6. Close the window.

Result

The structural information about the protein "_____" was thus retrieved from the PDB database.

Structure download (DNA) from Protein Data Bank

Aim

To retrieve structural information about the DNA_____ from protein databank.

Description

Protein Data Bank (PDB) is a structural databank. The PDB achieve contains macromolecules structural data of proteins, nucleic acids, protein-nucleic acids complexes. PDB databank is freely available worldwide. PDB gives information regarding the sequence 3D structure of compounds using various methods. PDB search can be performed using the output from one search as input. A search can return a single structure or multiple structures.

Procedure

1. Go to http://www.rcsb.org/pdb/website
2. Enter DNA name and click search.
3. Note down the title, DNA, base pairs from summary information.
4. Click on geometry and note down the common bond angles and bond length of the respective chains.
5. Click on the structural details and save the structural details.
6. Close the window.

Result

The structural information about the DNA _____ was thus retrieved from the PDB database.

Molecular file formats-FASTA, GenBank, Genpept, GCG, CLUSTAL, Swiss-Prot, PIR

Aim

To find the database file formats.

Procedure

FASTA

1. A sequence in FASTA format begins with a single-line description,
2. Followed by lines of sequence data
3. The Description line is distinguished from the sequence data by a greater than (□<□) symbol in the first column.
4. It is recommended that all line of text be shorter than 80 characters in length

```
>U03518 Aspergillus awamori internal transcribed spacer 1 (ITS1)
 AACCTGCGGAAGGATCATTACCGAGTGCGGGTCCTTTGGGCCCAACCTCCCATCCGTGTCTATTGTACCC
 TGTTGCTTCGGCGGGCCCGCCGCTTGTCGGCCGCCGGGGGGGCGCCTCTGCCCCCCGGGCCCGTGCCCGC
 CGGAGACCCCAACACGAACACTGTCTGAAAGCGTGCAGTCTGAGTTGATTGAATGCAATCAGTTAAAACT
 TTCAACAATGGATCTCTTGGTTCCGGC
```

GenBank/Genpept

The nucleotide (GenBank) and protein (Gen Pept) database entries are available from Entrez in this format.

1. Can contain several sequences
2. One sequence starts with: □LOCUS□
3. The sequence starts with: □ORIGIN□
4. The sequence ends with: □//□

```
LOCUS       AAU03518 237 bp DNA PLN 04-FEB-1995
DEFINITION Aspergillus awamori internal transcribed spacer 1 (ITS1) and
18S
           rRNA and 5.8S rRNA genes, partial sequence.
ACCESSION  U03518
BASE COUNT 41 a 77 c 67 g 52 t
ORIGIN
  1 aacctgcgga aggatcatta ccgagtgcgg gtcctttggg cccaacctcc catccgtgtc
 61 tattgtaccc tgttgcttcg gcgggcccgc cgcttgtcgg ccgccggggg ggcgcctctg
121 ccccccgggc ccgtgcccgc cggagacccc aacacgaaca ctgtctgaaa gcgtgcagtc
181 tgagttgatt gaatgcaatc agttaaaact ttcaacaatg gatctcttgg ttccggc
//
```

GCG

1. Exactly one sequence
2. Begins with annotation lines
3. Start of the sequence is marked by a line ending with ..
4. This line also contains the sequence identifier, the sequence length and a checksum

```
ID AA03518 standard; DNA; FUN; 237 BP.
XX
AC U03518;
XX
DE Aspergillus awamori internal transcribed spacer 1 (ITS1) and 18S
DE rRNA and 5.8S rRNA genes, partial sequence.
XX
SQ Sequence 237 BP; 41 A; 77 C; 67 G; 52 T; 0 other; AA03518 Length: 237 Check: 4514
..
   1 aacctgcgga aggatcatta ccgagtgcgg gtcctttggg cccaacctcc catccgtgtc
  61 tattgtaccc tgttgcttcg gcgggcccgc cgcttgtcgg ccgccggggg ggcgcctctg
 121 ccccccgggc ccgtgcccgc cggagacccc aacacgaaca ctgtctgaaa gcgtgcagtc
 181 tgagttgatt gaatgcaatc agttaaaact ttcaacaatg gatctcttgg ttccggc
```

SWISS-PROT

1. The first line of each sequence entry is the ID definition line which contains entry name, dataclass, molecule, division and sequence length
2. XX line contains no data, just a separator
3. The AC line lists the accession number
4. DE line gives description about the sequence
5. FT precise annotation for the sequence
6. Sequence information SQ in the first two spaces
7. The sequence information begins on the fifth line of the sequence entry
8. The last line of each sequence entry in the file is a terminator line which has the two characters // in the first two spaces.

```
ID AA03518 standard; DNA; FUN; 237 BP. XX AC U03518;
XX
AC U03518;
XX
DE Aspergillus awamori internal transcribed spacer 1 (ITS1) and 18S
DE rRNA and 5.8S rRNA genes, partial sequence.
DE rRNA and 5.8S rRNA genes, partial sequence.
RX MEDLINE; 94303342.
RX PUBMED; 8030378.
XX
FT rRNA     <1..20
FT          /product="18S ribosomal RNA"
FT misc RNA 21..205
FT          /standard_name="Internal transcribed spacer 1 (ITS1)"
FT rRNA     206..>237
FT          /product="5.8S ribosomal RNA"
SQ Sequence 237 BP; 41 A; 77 C; 67 G; 52 T; 0 other;
    aacctgcgga aggatcatta ccgagtgcgg gtcctttggg cccaacctcc catccgtgtc  60
    tattgtaccc tgttgcttcg gcgggcccgc cgcttgtcgg ccgccggggg ggcgcctctg 120
    ccccccgggc ccgtgcccgc cggagacccc aacacgaaca ctgtctgaaa gcgtgcagtc 180
    tgagttgatt gaatgcaatc agttaaaact ttcaacaatg gatctcttgg ttccggc    237
//
```

PIR

1. In this format, the first line begins with ‘>PI’; it denotes only for protein sequence.
2. ‘>Ni’; denotes only nucleic acid sequence.
3. The semi colon is followed by a code which is a unique sequence identifier.
4. For example >PI;5HIB-CAVPO,were 5HIB identifies the protein serotonin receptor IB, while CAVPO identifies its source as guinea-pig.
5. Then the sequence itself follows and is terminated by a asterisk(*).
6. Its conventional to give files in this format with an extension ‘.pir’ or ‘.seq’.

Molecular viewer by visualization software

Display Hydrogen Bonds

Aim

To display the hydrogen bonds for the selected protein.

Description

The RasMol hbonds command is used to represent the hydrogen bonding in the protein molecules backbone. The command hbonds on displays the selected bonds as dotted lines, and the hbonds off turns off their display. The color of the hbond objects may be changed by the colourhbond command. Initially, each hydrogen bond has the colors of its connected atoms.

Procedure

1. Load a protein structure file in the PDB format from protein data bank.
2. Open the PDB protein structure file in RasMol.
3. Render the molecule in wire frame display model.
4. Use hbonds command to display the hydrogen bonds of the protein.
5. View the result.
6. Close the molecule file in the graphics window.

Syntax

hbond<boolean>

hbond<value>

Command

Rasmol>hbonds on

Rasmol>color hbonds yellow

Result

The hydrogen bond for the selected protein was displayed.

Disulphide Bridges

Aim

To represent the disulphide bridges of the protein molecule □_____□ along the specified axis using RasMol.

Description

The RasMol □ssbond□ command is used to represent the disulphide bridges of the protein molecule as either dotted lines cylinders between the connected cysteine.

Procedure

1. Load a protein structure file in the PDB format from protein protein data bank.
2. Open the PDB protein structure file in RasMol.
3. Render the molecule in wire frame display model.
4. Use command □ssbond□ to display the disulphide bridges in the protein.
5. View the result.
6. Close the molecule file in the graphics window.

Syntax

ssbonds<Boolean>

ssbonds<value>

Command

rasmol>ssbonds on

rasmol>ssbonds 100

rasmol>color ssbonds yellow

Result

The disulphide bridges of the protein molecule □_____□ along the specified axis were thus represented using RasMol.

Strands

Aim

To display the protein molecule "_____" as "ribbon" of depth-cued curves passing along the backbone of the protein, using RasMol.

Description

The RasMol "strands" command display the currently loaded protein or nucleic acid as a smooth "ribbon" of depth-cued curves passing along the backbone of the protein. The ribbon is composed of as number of strands that run parallel to one another along the peptide plane of each residue. The ribbon is drawn between each amino acid whose alpha carbon is currently selected. The central and outermost strands may be coloured independently using the "colour ribbon 1" and "colour ribbon2" commands, respectively. The number of strands in the ribbon may be altered using the RasMol "set strands" command.

Procedure

1. Load a protein structure file in the PDB format from a protein data bank.
2. Open the PDB protein structure file in RasMol.
3. Render the molecule in wire frame display model.
4. Use command "strands" to display protein as ribbons.
5. View the result.
6. Close the molecule file in the graphics window.

SYNTAX

strands<Boolean>

strands<value>

COMMAND

rasmol>strands on

rasmol>color strands green

rasmol>color strands red

Result

The protein molecule □_____□ was displayed as ribbon of depth-called curves passing along the backbone of the protein RasMol.

LABEL

Aim

To Label the atoms, chain identifier and residues name of the protein molecule keratin using the RasMol.

Description

The RasMol label command allows an arbitrary formatted text string to be associated with each currently selected atom. The string may contain embedded expansion specifier which display properties of the atom being labeled. In expansion specifier consist of a% characters followed by a single alphabetic character specifying the property to be displayed.

Procedure

1. Load a protein structure file in the PDB format from protein data bank.
2. Open the PDB protein structure file in RasMol.
3. Render the molecule in wireframe display model.
4. Label the atoms, chain identifier and the residues name for the protein.
5. View the result.
6. Close the molecule file in the graphics window.

SYNTAX

label{<string>}

label<Boolean>

INPUT

The PDB structure of keratin is used as input.

COMMAND

rasmol>label%a

rasmol>label%n

rasmol>label%c

Result

The atoms, chain, identifier and residue name of the protein molecule _____ was labeled using RasMol.

RIBBON

Aim

To display the protein molecule "_____" as a smooth solid "Ribbon" surface passing along the backbone of the protein using RasMol.

Description

The RasMol ribbon" command display the protein molecule as a smooth solid "ribbon" surface along the backbone of the protein. The ribbon is drawn between each amino acid whose alpha carbon is currently selected.

Procedure

1. Load a protein structure file in the PDB format from a PDB.
2. Open the PDB protein structure file in RasMol.
3. Render the molecule in wireframe display model.
4. Use "ribbon" command to manipulate electron density map.
5. View the result.
6. Close the molecule file in the graphics window.

SYNTAX

ribbon{<Boolean>}

ribbon

COMMAND

rasmol> color ribbon blue

rasmol> color ribbon red

Result

The protein molecule "_____" was displayed as a smooth as a solid "ribbon" surface passing along the back bone of protein using RasMol

Experiment No: 2

Sequence alignment

BLAST suite of tools for pairwise alignment

Aim

To search homologous sequence for "_____" using Basic Local Alignment Search Tool (BLAST).

Resource URL

http://www.ncbi.nlm.nih.gov/blast/

Procedure

1. Open BLAST from NCBI home page.
2. Select either (blastp) or (blastn) option.
3. Now the protein- protein BLAST page appears.
4. Paste your query sequence in the search text box.
5. Click "BLAST" button from protein-protein BLAST page.
6. Click "format" button from the "formatting BLAST page.
7. Now the result page appears.
8. Click the color key from color key window.
9. The result corresponding to the color key is displayed.

Result

The homologous protein of albumin was successfully verified.

Multiple sequence alignment using Clustal-W

Aim

To perform multiple sequence alignment for the aminoacid sequence encoding for the protein "_____" from the different organisms using Clustal-W.

Description

Clustal-W is a general purpose multiple sequence alignment program for DNA or Protein. It produce biologically meaningful multiple sequence alignment of divergent sequence. It calculates the best match for the selected sequences and lines them up so that the identifies; similarities and differences can be seen. Evolutionary relationship can be seen via viewing phylogram.

Procedure

1. Enter into http://www.ebi.ac.uk/tools/clustalw.
2. Select Clustal-W from EBI tools.
3. Paste the input sequences in FastA format.
4. The title of the alignment was changed and the other options were left on default.
5. Click submit query.
6. Record the result and close the window.

Result

Thus the evolutionary relationships for the protein "_____" were thus predicted.

Experiment No: 3

Generating phylogenetic tree using PHYLIP

Aim

To find the evolutionary relationships between organisms and to analyze the changes occurring in these organisms during evolution using PHYLIP.

Description

PHYLIP is a complete phylogenetic analysis package which was developed by Joseph Felsestein 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.

Procedure

Align the multiple DNA sequences (output of the ClustalW) and save it in PHYLIP format as infile.phy. Start the program of Dnadist by clicking the icon and giving this infile as input.

All the PHYLIP programs are menu driven programs. Dnadist will calculate pairwise distances between the sequences. At first, Dnadist will ask whether the input file is there in the PHYLIP folder. If the file does not exist, it will ask you to give the correct file name. After giving the correct input, if needed it will ask to change any settings for the program by typing the first letter or number. If the changes are not required, by typing ‘Y’ it will start running the program. Output will return to the file as outfile, so that the output of this file can be used as input of another program.

Like Dnadist, Neighbor also gives sequence distance analysis. Output of Dnadist is given as input to Neighbor.

Branch lengths and tree are represented with the help of Neighbor joining method.

Unrooted trees are represented via Drawtree by giving outtree from the previous program as the input.

Result

Thus the phylogenetic tree was successfully generated.

Experiment No: 3

Protein structure prediction and analysis

Primary sequence analysis (ProtParam)

Aim

To predict the physico-chemical parameters for the protein "_____" using ExPasy proteomics tools.

Description

ExPASY is a tool which allows the computation of various physical and chemical parameters including the molecular weight; theoretical PI; amino acid composition, extinction, coefficient, estimated half life; instability index; aliphatic index and grand average of hydrophobicity.

Procedure

1. Go to http://www.expasy.ch/tools website.
2. Select "primary structure analysis" option in the ExPASY proteomics tools.
3. Paste the amino acid sequence of a protein.
4. Click the "perform" button to start prediction.
5. Close the ExPASY proteomics analysis tool window.

Result

Thus the physico-chemical parameters for the protein "_____" were thus predicted.

Secondary structure prediction (GOR, nnPredict, SOPMA)

GOR

Aim

To predict the secondary structure of protein sequence of "_____" using GOR.

Description

The secondary structure of protein refers to the different conformations that can be takenup by the polypeptide. The main types of secondary structure are alpha helices and beta pleatedstructures. They are stabilized mainly by hydrogen bonds between different amino acids in thepolypeptide chain. Most of the polypeptides are long enough to be folded into a series ofsecondary structures, one after another along the molecule. Bioinformatics proteomics server ExPASy helps.

Bioinformatics proteomics server ExPASy helps to predict the protein secondary structure using powerful secondary structure prediction tools such as GOR, SOPMA, NN Predict, Predict Protein, JPRED, and Chou-Fasman, PSI-Predetc

Procedure

1. Open the NCBI homepage and retrieve protein sequence of _____ and save it on the notepad.
2. Click the link https://npsa-prabi.ibcp.fr/NPSA/npsa_gor4.html
3. Paste the query sequence in text area without the FASTA description.
4. Analyse the various secondary structure (helices, strands, turns, coils, bent region etc)
5. Save and display the result

Result

The various secondary structure of given protein sequence _____ was successfully predicted using GOR. The results were analysed and displayed.

nnPredict

Aim

To predict the secondary structure of protein sequence of "_____" using GOR.

Description

The secondary structure of protein refers to the different conformations that can be takenup by the polypeptide. The main types of secondary structure are alpha helices and beta pleatedstructures. They are stabilized mainly by hydrogen bonds between different amino acids in thepolypeptide chain. Most of the polypeptides are long enough to be folded into a series ofsecondary structures, one after another along the molecule. Bioinformatics proteomics server ExPASy helps.

Bioinformatics proteomics server ExPASy helps to predict the protein secondary structure using powerful secondary structure prediction tools such as GOR, SOPMA, NN Predict, Predict Protein, JPRED, and Chou-Fasman, PSI-Predetc

Procedure

1. Open the NCBI homepage and retrieve protein sequence of _____ and save it on the notepad.
2. Click the link https://npsa-prabi.ibcp.fr/NPSA/npsa_gor4.html
3. Paste the query sequence in text area without the FASTA description.
4. Analyse the various secondary structure (helices, strands, turns, coils, bent region etc)
5. Save and display the result

Result

The various secondary structure of given protein sequence _____ was successfully predicted using GOR. The results were analysed and displayed.

nnPredict

Aim

To predict the secondary structure of protein sequence of "_____" using nnPredict.

Procedure

1. Open the NCBI homepage and retrieve protein sequence of _____ and save it on the notepad.
2. Click the link http://130.88.97.239/bioactivity/nnpredictfrm.html
3. Paste the query sequence in text area without the FASTA description.
4. Analyse the various secondary structure (helices, strands, turns, coils, bent region etc)
5. Save and display the result

Result

The various secondary structure of given protein sequence _____ was successfully predicted using nnPredict. The results were analysed and displayed.

SOPMA

Aim

To predict the secondary structure of protein using proteomic tool SOPMA.

Procedure

1. Open the web page of ExPASY
2. Fetch the FASTA format of the selected protein sequence
3. Select the secondary option, which displays in various available tools such as SOPMA
4. Open the web page of SOPMA by double clicking the link SOPMA
5. Paste the FASTA format obtained from the expasy in the text area provided.
6. Click the submit button, to display the result.

Result

The secondary structure of the □_____□ protein was successfully predicted using SOPMA.

Tertiary structure prediction (SWISSMODEL)

Aim

To perform homology modeling for the given protein.

Description

Homology modelling is a process by which a 3D model of a target sequence is built based on a homologue experimentally solved structure (X-ray crystallography and NMR spectroscopy). To perform homology modelling, one requires a target sequence, template structure and software to do modelling. The most powerful homology modelling server is called Swiss-Model Server. The graphic interface of Swiss-Model is called SPDBV (Swiss PDB Viewer/Deep View). SPDBV provides a good graphical environment for viewing and analyzing biomolecules like protein and nucleic acid. SPDBV also include advanced features like electrostatic potential, molecular surface and energy minimization. It is also used for mutation modelling and geometry analysis like bond length, bond angle etc A typical homology modelling involves the following steps:

1. Obtaining the target sequence.
2. Obtaining the template sequence.
3. Model building (Preparing the model).
4. Submitting the job to Swiss-Model Server.
5. Analyzing the accuracy of the model Model
6. Model Refinement.

Procedure

1. Protein sequence whose 3D structure needs to be built is obtained from Uniprot in .fasta format.
2. Open the link https://swissmodel.expasy.org/
3. Click on start modeling button
4. Paste or upload your protein of interest in the box provided
5. Provide the e-mail address and title in the appropriate column
6. Then search for the suitable template the model the protein sequence

7. The click on build model button

Result

The homology modeling of the target sequence of _____was successfully carried out using SWISSMODEL server.

Protein structure validation Ramachandran plot (PROCHECK)

Aim

To evaluate the stereo-chemical properties of the protein.

Description

The PROCHECK suite of programs provides a detailed check on the stereochemistry of a protein structure. Its outputs comprise a number of plots in PostScript format and a comprehensive residue-by-residue listing. These give an assessment of the overall quality of the structure as compared with well refined structures of the same resolution and also highlight regions that may need further investigation. The PROCHECK programs are useful for assessing the quality not only of protein structures in the process of being solved but also of existing structures and of those being modelled on known structures.

Procedure

1. Open the link http://services.mbi.ucla.edu/PROCHECK/
2. Upload your PDB file
3. The add the two numbers
4. Click run Procheck

Result

The stereo chemical properties of protein are analyzed.

Experiment No: 5

Gene structure prediction using Genscan and Glimmer

GENSCAN

Aim

To predict the putative genes from the given genomic sequences using GENESCAN

Description

Genscan identifies complete exon/intron structure of genes in genomic DNA. The features of the program include the capacity to predict multiple genes in a sequence and to deal with partial as well as complete genes and to predict consistent sets of genes occurring on either one or both the strands of DNA.

Procedure

1. Go to (http://genes.mit.edu/genscan.html) website.
2. Paste the DNA sequence and then click runscan.
3. Record the result of the predicted internal and terminal exons introns and intergenic regions.
4. Close the window.

Result

The initial internal and terminal exons and intergenic regions in the DNA sequence encoding for the protein _____ were thus predicted using Genscan.

Glimmer

Aim

To predict the genes in the microbial DNA.

Description

Glimmer is a system for finding genes in microbial DNA, especially the genomes of bacteria, archaea, and viruses. Glimmer (Gene Locator and Interpolated Markov ModelER) uses interpolated Markov models (IMMs) to identify the coding regions and distinguish them from non-coding DNA. The IMM approach is described in our original Nucleic Acids Research paper on Glimmer 1.0 and in our subsequent paper on Glimmer 2.0. The IMM is a combination of Markov models from 1st through 8th-order, where the order used is determined by the amount of data available to train the model. In addition, the positions used as context for the model need not immediately precede the predicted position but are determined by a decision procedure based on the predictive power of each position in the training data set (which we term an Interpolated Context Model or ICM). The models for coding sequence are 3-periodic non-homogenous Markov models. Improvements made in version 3 of Glimmer are described in the third Glimmer paper.

Glimmer was the primary microbial gene finder used at The Institute for Genomic Research (TIGR), where it was first developed, and has been used to annotate the genomes of 100s of bacterial and archaeal species from TIGR and other labs. Glimmer3 predictions are available for all NCBI RefSeq bacterial genomes at their ftp site.

Possible Viva Question

1. What you meant by biological database?
2. List out the types of biological database
3. Expand NCBI
4. What do you meant by evolutionary relationship?
5. What are the components of BLAST output?
6. What are the various types of BLAST?
7. What is the different between similarity searching via BLAST and FASTA?
8. What do you mean by global alignment?
9. What do you mean by local alignment?
10. What are homologous sequences?
11. What are the main steps of phylogenetic analysis?
12. Expand UPGMA
13. What do you mean by bootstrapping?
14. What are secondary structures?
15. Expand SOPMA
16. What are structural databases?
17. What are the salient features of PDB files?
18. What PDB ID indicates?
19. What is the various visualization tools used to visualize the structure of macromolecules?
20. What are various methods to represent the structures of macromolecules?
21. What do you mean by homology modeling?
22. List the steps required for homology modeling
23. What are the parameters required for the selection of good templates?
24. What are various tools employed for the refinement of homology models?
25. Expand PIR
26. Define node