Semester II

18BCP212 **PRACTICAL – IV** **4H-2C**

BIOLOGICAL DATABASES AND ANALYSIS

Instruction hours/week: L:0 T:0 P:4 **Marks:** Internal: **40** External: **60** Total: **100**

End Semester Exam: 3 Hours

Course objectives

To make the students

- To provide hands on experience on various biological databases and to learn the retrieval of data from the biological databases
- To make them learn about pair wise and multiple sequence analysis.
- To learn and apply the statistical approaches and models for phylogenetic analysis and tree reconstruction.
- To learn them protein prediction methods and its validation.

Course outcomes (CO's)

The students shall be able to

1. The course will enable students to use various biological databases and understand their importance functions in the biological system.
2. The course will enable students to use computational approaches for pair wise, multiple and phylogenetic analysis.
3. They would be well aware to predict the physio-chemical properties, protein structure and validation using computer based labs.
4. They would be able to solve the biological problems using various computational tools and techniques.

1. Biological Databanks Sequence databases, Structure Databases, Specialized databases
2. Data base file formats.
3. Data retrieval tools and methods (PUBMED, ENTREZ, SRS)
4. Sequence Similarity searching (NCBI- BLAST, FASTA)
5. Protein sequence analysis (ExPASY proteomics tools)
6. Multiple sequence alignment (Clustal-W)
7. Gene structure and function prediction (Using ORF Finder, Genscan, GeneMark)
8. Molecular Phylogeny (PHYLIP)
9. Sequence Analysis using EMBOSS
10. Protein structure visualization – RASMOL (Menu function and Command line entries), Deep View.

SUGGESTED READINGS

1. Lesk, A.M., (2014). Introduction to Bioinformatics, 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, 3rd edition, John Wiley & Sons, New York.
4. Mount, D.W., (2013). Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbour Laboratory Press, New York.

KARPAGAM ACADEMY OF HIGHER EDUCATION

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

Coimbatore – 641 021.

LECTURE PLAN
DEPARTMENT OF BIOCHEMISTRY

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

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

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

SYLLABUS

1. Biological Databanks Sequence databases, Structure Databases, Specialized Databases
2. Data base file formats
3. Data retrieval tools and methods (PUBMED, ENTREZ, SRS
4. Sequence similarity searching (NCBI-BLAST, FASTA)
5. Protein sequence analysis (ExPASYproteomics tools)
6. Multiple sequence alignment (Clustal-W)
7. Gene structure and function prediction (using ORF Finder, Genscan, GeneMark)
8. Molecular Phylogeny (PHYLIP)
9. Sequence analysis using EMBOSS
10. Protein structure visualization – RASMOL (Menu function and Command line entries), Deep View.

KAHE

Experiment No: 1

Biological Databanks, Sequence Databases, Structure Databases, Specialized Databases

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.

GENBANK

Aim

To retrieve the nucleotide sequence encoding for the protein "_____" from the nucleotide database of NCBI in Genbank format.

Description

GenBank is the NIH genetic sequence database, an annotated collection of all publically available DNA sequence. Genbank nucleotide database is maintained by the National Centre for Biotechnology Information (NCBI) which is a part of National Institute of Health (NIH), a federal agency of the government. It was established on November 4, 1922 as a national resource for Molecular Biology, NCBI creates public database, conduct research in computational biology, develops software tools for analyzing genome data and disseminates biomedical information. The search and retrieval system used in NCBI is Entrez which provides the users with integrated access to sequence, mapping, taxonomy and structural data. A powerful and unique feature of Entrez is the ability to retrieve related sequences, and references.

Procedure

1. Enter into http://www.ncbi.nlm.nih.gov/website.
2. Select nucleotide from the all database drop down list.
3. Enter the protein name and click on go to search for the nucleotide sequence information.
4. Data would be retrieved in Genbank format.
5. Note down the accession number, locus BP, molecule type, definition, source organism, organism classification, author title, journal, university version.
6. Save the page.
7. Close the window.

Result

The nucleotide sequence encoding for the protein "_____" was retrieved from the nucleotide database NCBI in Genbank format.

SWISS PROT

Aim

To retrieve amino acid sequence encoding for the protein "_____" from SWISS-PROT Database.

Description

SWISS-PROT is a curated protein sequence database which strives to provide a high level of annotation such as description, function of protein, domain structure, post-translational modification, variants etc. It has minimal level of redundancy and high level of integration with other database.

SWISS-PROT is a protein knowledge database established in 1986 and maintain collaboratively, since 1987, by the Department of Medicinal Biochemistry of University of Geneve and EMBL Data library. In SWISS-PROT, 2 classes of data can be distinguished: core data and the annotation.

Procedure

1. Enter into http://www.expasy.org\sprot\website.
2. Type the protein name in the box and click go.
3. Select any one of the hits from the entries.
4. Save the result.
5. Close the window.

Result

Thus the amino-acid sequence encoding for the protein "_____" was retrieved from SWISS-PROT database.

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.

Specialised Database

Aim

To retrieve the information about the '_____' of human from KEGG database.

Description

Kyto encyclopedia of genes and genomes is a collection of online database dealing with genomes pathway and biological chemicals. The pathway database record network of molecular interaction in the cells and variants of them specific to particular organism. It was initiated by the Japanese Human Genome Program in 1955. KEGG is considered to be a computer representation of the biological system. It consist of genetic building block of genes and proteins (KEGG GENES); chemical building blocks of both endogenous and exogenous substances (KEGG LCGAND); molecular diagrams of interaction and reaction network and hierarchies and relationship of various biological objects KEGG provides a reference knowledge base for linking genomes to biological system and to environment by the processes of pathway mapping.

Procedure

1. Enter the http://www.genome.adjp/keg/pathway.html.website
2. Select' KEGG PATHWAY'
3. Select' metabolism' pathway
4. Select the pentose and glucoronate pathway in carbohydrate metabolism
5. Select any organism from the drop down list
6. Note down the pathway
7. Close the window

Result

The information about the '_____' of human was thus retrieved from KEGG database

Experiment No: 2

Database file formats

Aim

To find the database file formats.

Procedure

There are three types of file format representing protein and nucleic acid sequences, NBRF/PIR, FASTA and GDE. In multiple sequence alignments, the following formats are applicable. MSF, PHYLIP and all. In structure wise PDB format is being used.

1. **NBRF/PIR:** In this format, the first line begins with ">PI"; it denotes only for protein sequence. ">Ni"; denotes only nucleic acid sequence. The semi colon is followed by a code which is a unique sequence identifier. For example >PI; 5HIB-CAVPO, were 5HIB identifies the protein serotonin receptor IB, while CAVPO identifies its source as guinea-pig. Then the sequence itself follows and is terminated by a asterisk(*). Its conventional to give files in this format with an extension '.pir' or '.seq'.
2. **FASTA format:** In this format, the first line begins with ">" but there is no designation of protein or nucleic acid sequence. The code is entered next and is followed by comments, although its conventional to delimit the comments with a "1" symbol. For example: >5HIB_CAVPO 008892/guinea pig serotonin receptor. One point to note about fasta files is that they allow lower-case letters for the amino acids. Files in this format commonly have the extension ".fasta".
3. **GDE format:** The GDE format is essentially similar to FASTA format, but the ">" symbol in the first line is replaced by "%". Files in this format have the extension '.gde'.
4. **FILES FOR ALIGNED SEQUENCES:** The output from sequence alignment program can be anyone of a number of formats. In order to achieve the alignments, gaps must be introduced and this are represented either by hyphens of dots. Multiple sequence format (MSF) is used by several software tools.
5. **PHYLIP:** Phylogenetic Interference package is the output format of the software.

6. **ALL Format:** Clustal-w/x [F1] has its own ALL format
7. **FILES OF STRUCTURAL DATA:** These are text files using a format devised by the Protein Data Bank (PDB) files. Such files contain orthogonal atomic coordinates together with annotations, comments and experimental details. The most important aspect of PDB files is the 'ATOM', lines are laid out in columns of characters, not columns of words.

KAHE

Experiment No: 3

Data retrieval tools and methods (PUBMED, ENTRZ, SRS)

AIM

To know about the different data retrieval tools and methods available.

PROCEDURE

A Large amount of biological information is available over the world wide web (WWW). But the data are widely distributed and it is therefore necessary for the scientist to have efficient mechanisms for data retrieval. Efficient mechanisms of data retrieval, one approach is to use standard search engines to find relevant web pages. However, it is sometimes difficult to find the desired information using this method. Alternatively, there are a number of dedicated data retrieval tools that can be used to access information of molecular biology. The most widely used of these are Entrez, PUBMED and SRS (Sequence Retrieval System). Each of these tools allows text based searching of a number of a linked data base as well as sequence searching.

Pubmed

Aim

To analyze scientific article related to research

Resource URL

www.ncbi.nlm.nih.gov/pubmed/

Procedure

1. Open the NCBI home page
2. Select the option all database in PUBMED
3. Type any one the disease
4. You can visualize article

Result

The research article regarding '_____' has been searched and analyzed by using PUBMED.

KAHE

ENTREZ

It is a worldwide web based data retrieval system developed by a NCBI (National Centre for Biotechnology Information), which integrates information held in all NCBI databases. These databases include nucleotide sequences, protein sequences, macromolecular structure and whole genomes. Other resources linked to NCBI can also be searched using this tool. This includes Online Mendelian Inheritance in Man (OMIM) and the literature data base through Pubmed. In total, this tool links to all databases which are being listed in the following table.

Category	Database
Nucleic acid sequence	EntreZ nucleotides, sequences obtained from GenBank, Refseq and PDB
Protein sequences	EntreZ protein sequences obtained from SWISS PROT, PIR, PRF, PDB and translations from annotated coding regions in GenBank and Refseq
3D Structure	EntreZ Molecular Modeling Database(MMDB)
Genomes	Complete genome assemblies from many sources
Popset	From GenBank set of DNA sequences, that have been collected to analyze the evolutionary relatedness of a population
OMIM	Online Mendelian Inheritance in Men
Taxonomy	NCBI Taxonomy database
Books	Bookshelf
Probeset	Gene expression Omnibus (GEO)
3D Domains	Domain from the Molecular Modeling database
Literature	PubMed

Procedure

1. Go to NCBI-ENTREZ site (www.ncbi.nlm.nih.gov/entrez)
2. Choose "Nucleotide" or "Protein" database
3. Enter the search term in the search box
4. Click on "Go"
5. From the list of hits obtained choose the sequence that are of interest by ticking
6. In place of 'Summary' choose "FASTA"
7. Enter the "Text" is chosen (default)
8. Click on "Send to"
9. Copy the sequences obtained on to word pad or note pad application and save the file

Result

A nucleotide sequence was retrieved from GenBank and a protein sequence was retrieved from the protein sequence database at NCBI using ENTREZ.

KAHE

SRS

SRS is a retrieval tool developed by the European Bioinformatics Institute that integrates over 80 Molecular Biology databases. It is a open source software that can be downloaded and installed locally. The result is that the databases that are available to the end user are not restricted by the activities of the curators. Several large sites including SWISS-PROT use SRS as a standard. The main ExPASY site, one of the most useful bioinformatics gateway is an example of SRS in action. The databases covered by the SRS are listed below:

SRS description	Examples
Literature	MEDLINE
Sequence	EMBL, EMBLNEW, SWISS-PROT, SP-TREMBL, PEMTREMBL, TREMBLNEW, ENSEMBL, PATNET-PRT, USPO-PRT, IMGTLIGM, IMGTHLA
Interpro and Related	INTERPRO, IPRMATCHES, IPRMATHES_ENSEMBL, PROSITE, PROSITEDOC, BLOCKS, PFAMA, PFAMB, PFMHMM, PRODOM, PFAMSEED
Sequence Related	UTR, UTRSITE, TAXONOMY, GENETICCODE, EPD, HTG-QSCORE, CPGISLAND, EMESLIB, EMBLALIGN
Transfac	TFSITE, TFFACTOR ,TFCELL, TFCLASS, TFMATRIX, TFGENE
Userowned Databanks	USERDNA,USERPROTEIN
Application Results	FASTA, FASTX, FASTY, NFASTA, BLASTP, BLASTN, CLUSTALW, NCLUSTALW, PPSEARCH, RESTRICTIONMAP

Protein 3 D Structures	PDB, DSSP, HSSP, FSSP, PDBFINDER, RESID
Genome	MOUSE2HUMAN, LOCUSLINK, HGNC, HSAGENES
Mapping	RHDB, PHDBNEW, LOCUSLINK, HGNC, HSAGENES
Mutations	OMIMALLELE, MUTRES, SWISSCHANGE, EMBLCHANGE, MUTRESSTATUS, OMIMOFFSET, OMIM, HUMUT, HUMAN-MITBASE, P53LINK
Locus Specific Mutations	41 entries omitted here
SNP	MITSNP, dbSNP, submitter, dbSNP assay, dbSBPSBP, HGBASE, HGBASE_SUBMITTER
Metabolic pathway	LENZYME, LCOMPOUND, PATHWAY, ENZYME, EMP, MPW, UPATHWAY, UREACTION, UENZYME, UCOMPOUND, UNIMAGEMAP, REBASE, SRSFAQ, BIOCATAL
Other System	PRIMASTATUS

Experiment No: 4

Sequence similarity searching (NCBI-BLAST, FASTA)

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.

Sequence similarity search for a pair of sequences using FASTA

AIM

To find the homologous protein sequence of '_____' sequence using FASTA.

Resource URL

http://www.ebi.ac.uk/fasta33/

Procedure

1. Open FASTA from EBI homepage.
2. Now the EMBL-EBI page appears.
3. Select 'protein from the 'DATABASES' list.
4. Paste the sequence in the text box provided, or click 'Browse' Button to load the file.
5. Then, click 'Run FASTA3' button.
6. Now the result page appears.
7. Click 'Show Alignments' button.
8. Now the result is displayed.

Result

The homologous sequence of '_____' protein was successfully analyzed.

KAHE

Experiment No: 5

Protein sequence analysis (ExPASY Proteomics tools)

Aim

To predict the physico-chemical parameters for the protein "insulin" 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.

Experiment No: 6

Multiple sequence alignment (Clustal-W)

Aim

To perform multiple sequence alignment for the amino acid 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: 7

Gene Structure and function prediction using (ORF finder, Genscan, GeneMark)

ORF finder

Aim

To identify the functional sites present in the genome of '_____' using ORF finder.

Procedure

The ORF finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence or a in a sequence already in the database. This tool is hosted in NCBI. The tool identifies all open reading frames using the standard or alternative genetic codes. The deduced amino acid sequence can be saved in various formats and searched against the sequence database using the WWW BLAST server.

ORF Finder

1. Open the NCBI homepage and retrieve the sequence of
2. Open the ORF Finder homepage
   http://www.ncbi.nlm.nih.gov/projects.gorf/
3. Paste the query sequence in the text area and click ORF find button
4. Analyze the six frame translations, coding regions, start codon, stop codon etc.
5. Display all graphical output and interpret it.

Result

The different functional sites present in the query sequence '_____' and _____ were successfully identified, analyzed using ORF finder and displayed.

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.

GENEMARK

Aim

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

Gene finding using Genemark hmm:

1. Go to gene mark program (eg: http://opal.biology.gatch.edu/genemark/).
2. Scroll down to eukaryotic genomic sequence analysis and click on 'here'.
3. Give a title (optional but helpful).
4. Paste sequence saved in notepad file (from biology work bench).
5. Select species (eg:H.sapiens).
6. Unclick 'generate PDF graphics'.
7. Click on 'translate predicted genes into protein'.
8. Click on 'start gene mark.hmm'.
9. Note result of predicted genes/exons.
10. Copy and paste into notepad.
11. Scroll down to Genemark .hmm protein translations.
12. Copy and paste the resultant protein sequences into notepad.
13. You may also generate a postscript graphics and have it mailed to you.

RESULT:

Using gene finding program Genemark genes (exons) were predicted. The predicted gene sequences (exons) were also theoretically translated into protein sequence.

Experiment No: 8

Molecular Phylogeny (PHYLIP)

Aim

To predict the phylogenetic relationship between the sequences.

Description

PHYLIP (the PHYLogeny Inference Package) is one of the most used phylogeny inference tool. It is a actually package of 35 programs, carrying out various type of inference algorithms, e.g. parsimony, distance matrix, and likelihood methods, including bootstrapping and consensus trees. Data types that can be handled include molecular sequences, gene frequencies, restriction sites and fragments, distance matrices, and discrete characters.

PHYLIP is probably the most widely distributed phylogeny package. It is the third most frequently cited phylogeny package, after PAUP* and MrBayes, and ahead of MEGA. PHYLIP is also the oldest widely distributed package. It has been in distribution since October, 1980, and has over 28,000 registered users. It is regularly updated.

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.

KAHE

Experiment No: 10

Protein structure visualization – RASMOL (Menu function and command line entries), Deep View

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 molecule's 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 'colour hbond' 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.

KAHE

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

DEEP View

Aim

To visualize the 3-Dimensional structure of the molecules.

Description

Deep View is a friendly but powerful molecular graphics program. It is designed for full compatibility with computing tools available from the Expert Protein Analysis System, or ExPASy, Molecular Biology Server in Geneva, Switzerland. While Deep View is simple to use for viewing structures and creating vivid illustrations, it also shines as an analytical tool. Deep View allows you to build models from scratch, simply by giving an amino-acid sequence. Deep View can find hydrogen bonds within proteins and between proteins and ligands. It allows you to examine electron-density maps from crystallographic structure determination, to judge the quality of maps and models, and to identify many common types of problems in protein models. It allows you to view several or many models simultaneously and superimpose them to compare their structures and sequences. It computes electrostatic potentials and molecular surfaces, and carries out energy minimization. For proteins of known sequence but unknown structure, Deep View submits amino acid sequences to ExPASy to find homologous proteins, onto which you can subsequently align your sequence to build a preliminary three-dimensional model. Then Deep View submits your alignment to ExPASy, where the SWISS-MODEL server builds a final model, called a homology model, and returns it directly to Deep View.

Possible Viva Question

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

28. What are the parameters required for the selection of good templates?
29. What are various tools employed for the refinement of homology models?
30. Expand PIR
31. Define node

KAHE