II M.SC BIOCHEMISTRY

BIOINFORMATICS ((P16BC42)

SEMESTER IV

Prepared by

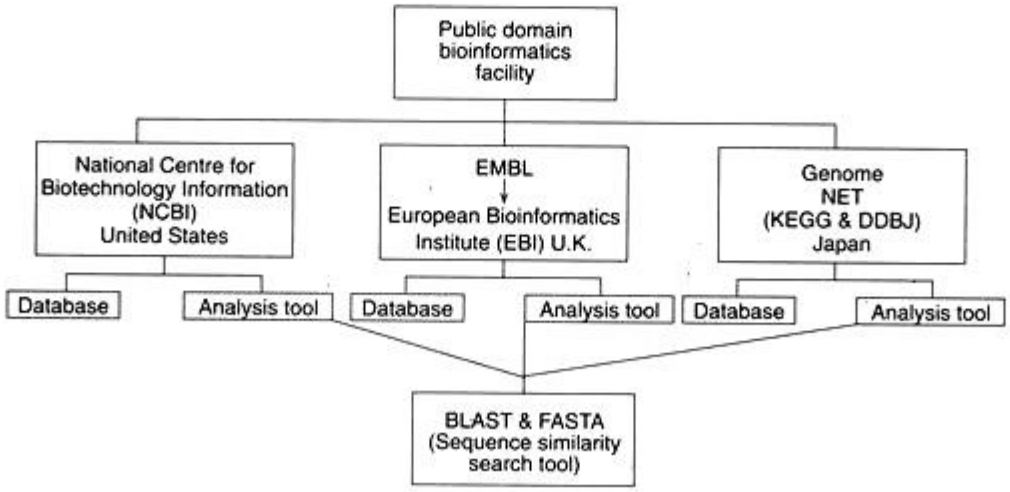
Dr.T.Ananthi,
Assistant Professor,
PG and Research Department of Biochemistry,
S.T.E.T. Women's College,
Sundarakkottai, Mannargudi

SHORT QUESTIONS AND ANSWERS

1. Definition of Bioinformatics

Bioinformatics is currently defined as the study of information content and information flow in biological systems and processes. It serves as the bridge between observations (data) in diverse biologically-related disciplines and the derivations of understanding (information) about how the systems or processes function and subsequently the application (knowledge).

Major public domain bioinformatics facilities are

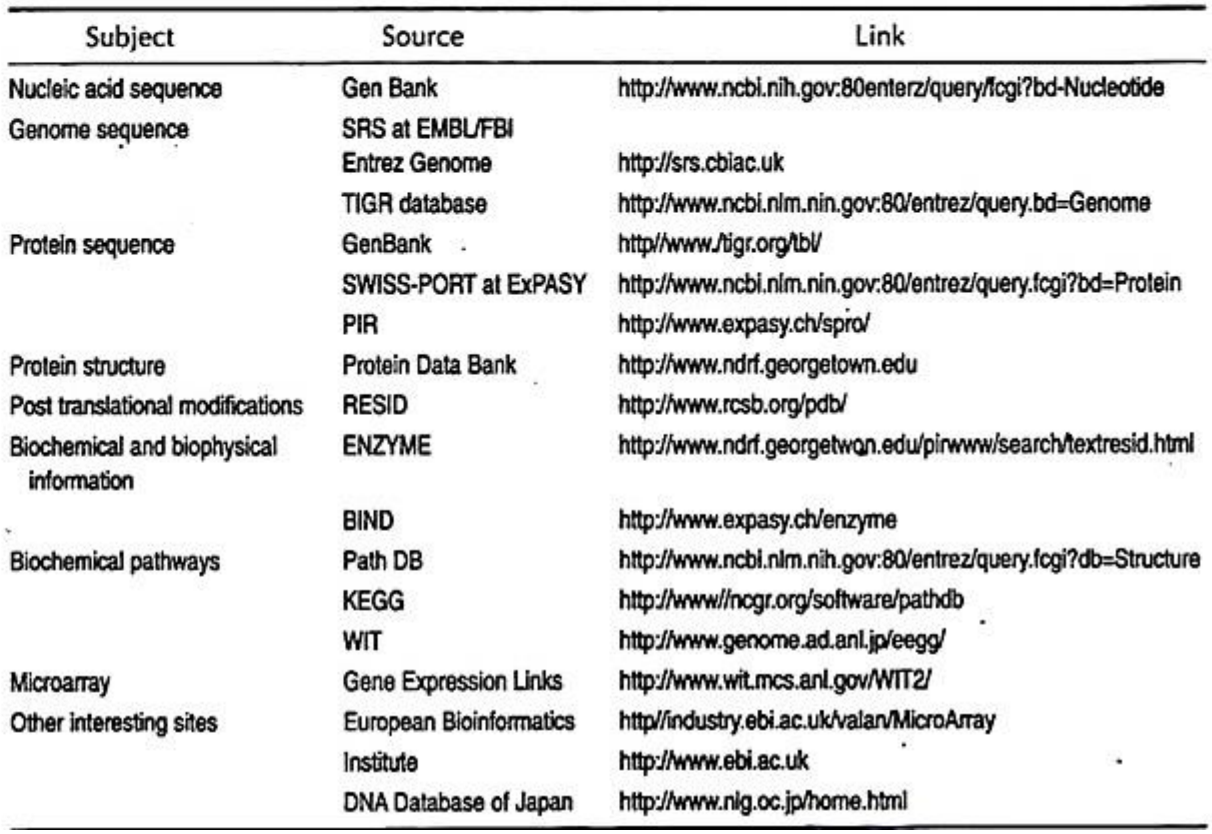

Public domain bioinformatics facility
National Centre for Biotechnology Information (NCBI) United States
EMBL
European Bioinformatics Institute (EBI) U.K.
Genome NET (KEGG & DDBJ) Japan
Database
Analysis tool
Database
Analysis tool
Database
Analysis tool
BLAST & FASTA (Sequence similarity search tool)

(a)NCBI – National Centre for Biotechnology Information, USA.

(b) EBI – European Bioinformatics Institute, UK.

(c) SIB – Swiss Institute of Bioinformatics, Switzerland.

2. Some important websites commonly used for bioinformatics



Subject	Source	Link
Nucleic acid sequence	Gen Bank	http://www.ncbi.nih.gov:80enterz/query/fcgi?bd-Nucleotide
Genome sequence	SRS at EMBL/FBI	
	Entrez Genome	http://srs.cbiac.uk
	TIGR database	http://www.ncbi.nlm.nin.gov:80/entrez/query.bd=Genome
Protein sequence	GenBank	http//www./tigr.org/tbl/
	SWISS-PORT at ExPASY	http://www.ncbi.nlm.nin.gov:80/entrez/query.fcgi?bd=Protein
	PIR	http://www.expasy.ch/spro/
Protein structure	Protein Data Bank	http://www.ndrf.georgetown.edu
Post translational modifications	RESID	http://www.rcsb.org/pdb/
Biochemical and biophysical information	ENZYME	http://www.ndrf.georgetwon.edu/pirwww/search/textresid.html
	BIND	http://www.expasy.ch/enzyme
Biochemical pathways	Path DB	http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Structure
	KEGG	http://www//ncgr.org/software/pathdb
	WIT	http://www.genome.ad.anl.jp/eegg/
Microarray	Gene Expression Links	http://www.wit.mcs.anl.gov/WIT2/
Other interesting sites	European Bioinformatics	http//industry.ebi.ac.uk/valan/MicroArray
	Institute	http://www.ebi.ac.uk
	DNA Database of Japan	http://www.nig.oc.jp/home.html

3. Search engine

A search engine is a type of utility or tools, which provide facility to retrieve information from different databases. In general life we use many search engines such as Goggle, Rediff and Yahoo but for bioinformatics there are mainly two search engines BLAST and FASTA.

4. **Databases**

Database is the combination of same type of information or files that are collectively called as database.

The major protein databases are

- PDB, SWISS-PROT, PROSITE, ExPASy, PIR, PRINTS, BLOCKS, PRODOM, Pfam, Inter Pro.

The major nucleic acid databases are:

- Gen Bank, DDBJ, Ref Seq, dbEST, NDB, CSD, EMBL.

5. **Genomics**

The complete genetic content of an organism is genome, and the DMA obtained is called genomic DNA. This genomic DNA of prokaryote contains all the coding region and can be sequenced, whereas the DNA of eukaryotes includes both intron and exon sequences (coding sequence) as well as non-coding regulatory sequences such as promoter, and enhancer sequences.

The subject genomics is the complete analysis of the entire genome of a chosen organism which involves the study of physical structure of the organism’s genome or the genetic makeup of an organism to know the number of genes present and the type of genes, i.e., to study the function of different genes.

Significance of Genomics:

All the information’s require input in probability theory, database management and manipulation, and computer science.

This will help in:

(a) Identification of open reading frame sequences,
(b) Gene splicing sites (introns),
(c) Gene annotation (inter-genomic comparisons) and
(d) Determination of sequence patterns of regulatory sites and gene regulations.

6. **Proteomics**

The entire protein component of a given organism is called ‘proteome’, the term coined by Wasinger in 1995. A proteome is a quantitatively expressed protein of a genome that provides information on the gene products that are translated, amount of products and any post translational modifications.

Proteomics is an emerging area of research in the post-genomic era, which involves identifying the structures and functions of all proteins of a proteome. It is sometimes also treated as structural based functional genomics.

Scope of Proteomics:

Proteomics deals with significant problems like:

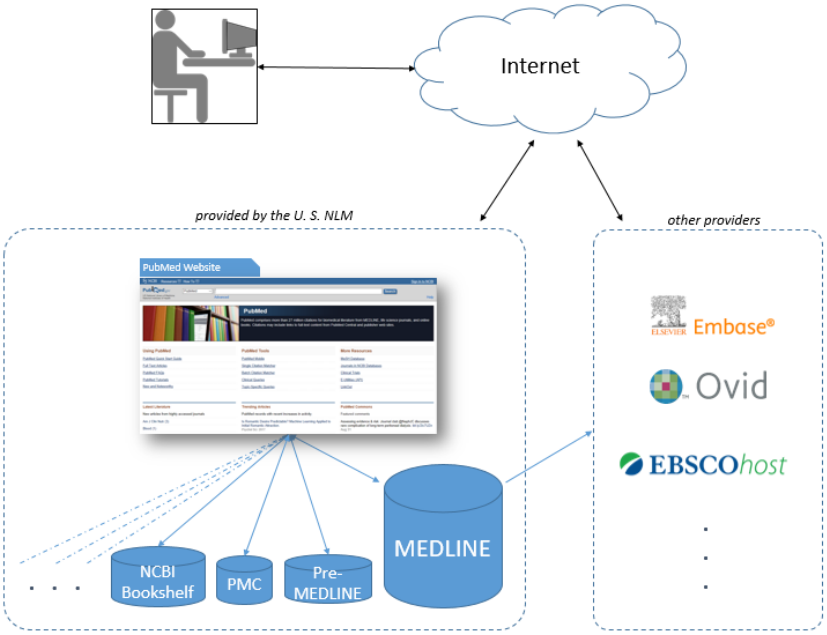
(a) Identification of functional domains in protein sequences.
(b) Single, multiple protein alignment (homology).
(c) Determining sequence-structure, sequence-function relationships (structural bioinformatics).
(d) Discovery of protein pattern and providing the framework for the analysis of signalling networks.

7. **SCOP**

The Structural Classification of Proteins (**SCOP**) database is a largely manual classification of protein structural domains based on similarities of their structures and amino acid sequences. A motivation for this classification is to determine the evolutionary relationship between proteins.

8. **PubMed**

PubMed is a free search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. The United States National Library of Medicine (NLM) at the National Institutes of Health maintain the database as part of the Entrez system of information retrieval.



9. **NCBI** (National Center for Biotechnology Information)

The NCBI houses a series of databases relevant to biotechnology and biomedicine and is an important resource for bioinformatics tools and services. Major databases include GenBank for DNA sequences and PubMed, a bibliographic database for the biomedical literature.

10. Swiss Prot database

SWISS-PROT is a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases.

11. TrEMBL

TrEMBL is a computer-annotated protein sequence database. ′ UniProtKB/TrEMBL contains the translations of all coding sequences (CDS) present in the EMBL/GenBank/DDBJ Nucleotide Sequence Databases and also protein sequences extracted from the literature or submitted to UniProtKB/Swiss-Prot.

12. RasMol

Rasmol is most popular 3D molecular graphics viewer. It is particularly good at viewing and rotating protein molecules, although it also works perfectly with smaller molecules. Rasmol is a molecular visualization software raster display of molecules Raster=display of pixels on monitor pixel=one dot on monitor Protein Explorer, Chime, Jmol, pymol are other similar softwares used. To use Rasmol, your computer need Rasmol software and PDB data file Rasmol Software is available for Windows and Linux. PDB files for various molecules are available on Internet

13. TIGR

The Institute for Genomic Research (TIGR) The Institute for Genomic Research (TIGR) is a non-profit research institute located in Rockville, Maryland. The primary interest of TIGR is the sequencing of the genomes and the subsequent analysis of the sequences in prokaryotic and eukaryotic organisms.

14. Unix

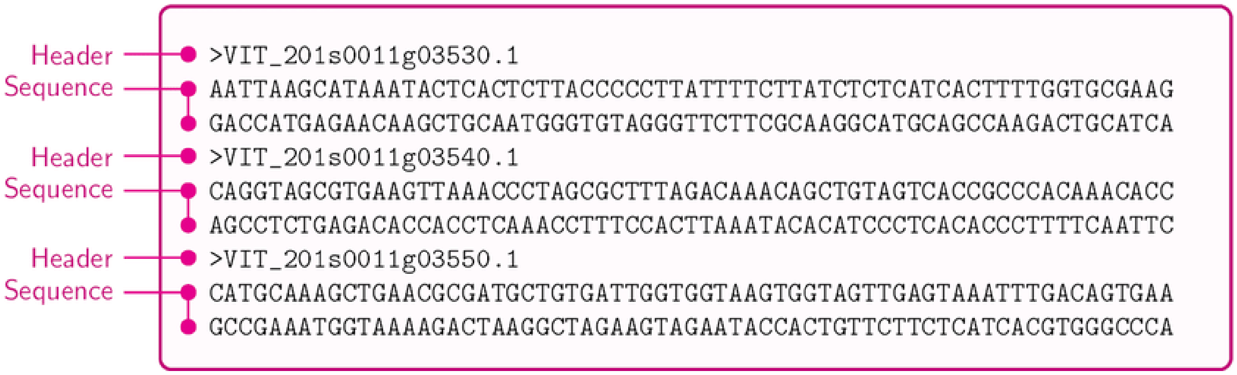
- Multiuser operating system
- Linux: free version of this type of operating system
  - Red Hat Enterprise Linux, Ubuntu, and CentOS
- Used on high-end workstations, database servers, web servers and managing shared resources
- Standard features include: Security, reliability, scalability
  - Can supports 100s of users at a time

15. FASTA (FAST-All)

FASTA (pronounced FAST-Aye) stands for FAST-All, reflecting the fact that it can be used for a fast protein comparison or a fast nucleotide comparison. This program achieves a high level of sensitivity for similarity searching at high speed. This is achieved by performing optimised searches for local alignments using a substitution matrix. The high speed of this program is achieved by using the observed pattern of word hits to identify potential matches before attempting the more time consuming optimised search. The trade-off between speed and sensitivity is controlled by the ktup parameter, which specifies the size of the word.

Increasing the ktup decreases the number of background hits. Not every word hit is investigated but instead initially looks for segment's containing several nearby hits.

In **bioinformatics** and biochemistry, the **FASTA** format is a text-based format for representing either nucleotide sequences or amino acid (protein) sequences, in which nucleotides or amino acids are represented using single-letter codes. The format also allows for sequence names and comments to precede the sequences. FASTA is pronounced "**fast A**", and stands for "**FAST-All**", because it works with any alphabet, an extension of the original "FAST-P" (protein) and "FAST-N" (nucleotide) alignment tools.



```
>VIT_201s0011g03530.1
AATTAAGCATAAATACTCACTCTTACCCCCTTATTTTCTTATCTCTCATCACTTTTGGTGCGAAG
GACCATGAGAACAAGCTGCAATGGGTGTAGGGTTCTTCGCAAGGCATGCAGCCAAGACTGCATCA
>VIT_201s0011g03540.1
CAGGTAGCGTGAAGTTAAACCCTAGCGCTTTAGACAAACAGCTGTAGTCACCGCCCACAAACACC
AGCCTCTGAGACACCACCTCAAACCTTTCCACTTAAATACACATCCCTCACACCCTTTTCAATTC
>VIT_201s0011g03550.1
CATGCAAAGCTGAACGCGATGCTGTGATTGGTGGTAAGTGGTAGTTGAGTAAATTTGACAGTGAA
GCCGAAATGGTAAAAGACTAAGGCTAGAAGTAGAATACCACTGTTCTTCTCATCACGTGGGCCCA
```

16.**BLAST** (**The Basic Local Alignment Search Tool)**

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

Types

- BLASTn- compares one or more nucleotide query sequences to a subject nucleotide sequence or a database of nucleotide sequences.
- BLAST (blastp)- This program, given a protein query, returns **the** most similar protein sequences from **the** protein database that **the** user specifies.
- BLASTgp- Position-Specific Iterative

17.Pairwise Alignment

Aligns your query sequence and database matches in pairs. Matches are connected with a "|" symbol. Mismatches are opposed with a spce. Gaps are introduced with a "-" symbol. e.g.

18.NCBI-Blast2:

BLAST stands for Basic Local Alignment Search Tool.The emphasis of this tool is to find regions of sequence similarity, which will yield functional and evolutionary clues about the structure and function of your novel sequence. WU-BLAST 2.0 and NCBI BLAST2 are distinctly different software packages, although they have a common lineage for some portions of their code, so the two packages do their work differently and obtain different results and offer different features.

19.**Pfam**

Pfam is a database of curated protein families, each of which is defined by two alignments and a profile hidden Markov model (HMM). Profile HMMs are probabilistic models used for the statistical inference of homology (1,2) built from an aligned set of curator-defined family-representative sequences.

20.KEGG Database

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a database resource that integrates genomic, chemical and systemic functional information. In particular, gene catalogs from completely sequenced genomes are linked to higher-level systemic functions of the cell, the organism and the ecosystem.

Major efforts have been undertaken to manually create a knowledge base for such systemic functions by capturing and organizing experimental knowledge in computable forms; namely, in the forms of molecular networks called KEGG pathway maps, BRITE functional hierarchies and KEGG modules. Continuous efforts have also been made to develop and improve the cross-species annotation procedure for linking genomes to the molecular networks through the KEGG Orthology (KO) system.

As the result, KEGG is widely used as a reference knowledge base for integration and interpretation of large-scale datasets generated by genome sequencing and other high-throughput experimental technologies. In addition to maintaining the aspects to support basic research, KEGG is being expanded towards more practical applications integrating human diseases, drugs and other health-related substances.

KEGG is developed by Kanehisa Laboratories.

21. CASP

Critical Assessment of protein Structure Prediction, or CASP, is a community-wide, worldwide experiment for protein structure prediction taking place every two years since 1994.[1] CASP provides research groups with an opportunity to objectively test their structure prediction methods and delivers an independent assessment of the state of the art in protein structure modeling to the research community and software users. The primary goal of CASP is to help advance the methods of identifying protein three-dimensional structure from its amino acid sequence, many view the experiment more as a “world championship” in this field of science. More than 100 research groups from all over the world participate in CASP on a regular basis and it is not uncommon for entire groups to suspend their other research for months while they focus on getting their servers ready for the experiment and on performing the detailed predictions.

22. Rosetta

The Rosetta software suite includes algorithms for computational modeling and analysis of protein structures. It has enabled notable scientific advances in computational biology, including de novo protein design, enzyme design, ligand docking, and structure prediction of biological macromolecules and macromolecular complexes.

Uses

- Understanding macromolecular interactions
- Designing custom molecules
- Developing efficient ways to search conformation and sequence space

- Finding a broadly useful energy functions for various biomolecular representations

23.CYTOSCAPE

Cytoscape is an open source bioinformatics software platform for visualizing molecular interaction networks and integrating with gene expression profiles and other state data. Additional features are available as plugins. Plugins are available for network and molecular profiling analyses, new layouts, additional file format support and connection with databases and searching in large networks. Plugins may be developed using the Cytoscape open Java software architecture by anyone and plugin community development is encouraged. Cytoscape also has a JavaScript-centric sister project named Cytoscape.js that can be used to analyse and visualise graphs in JavaScript environments, like a browser. Cytoscape was originally created at the Institute of Systems Biology in Seattle in 2002. Now, it is developed by an international consortium of open source developers. Cytoscape was initially made public in July, 2002 (v0.8); the second release (v0.9) was in November, 2002, and v1.0 was released in March 2003.

24.**DNA Data Bank of Japan** (**DDBJ)**

The **DNA Data Bank of Japan** (**DDBJ**) is a biological database that collects DNA sequences.[1][2] It is located at the National Institute of Genetics (NIG) in the Shizuoka prefecture of Japan. It is also a member of the International Nucleotide Sequence Database Collaboration or INSDC. It exchanges its data with European Molecular Biology Laboratory at the European Bioinformatics Institute and with GenBank at the National Center for Biotechnology Information on a daily basis. Thus these three databanks contain the same data at any given time.

DDBJ began data bank activities in 1986 at NIG and remains the only nucleotide sequence data bank in Asia. Although DDBJ mainly receives its data from Japanese researchers, it can accept data from contributors from any other country. DDBJ is primarily funded by the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT). DDBJ has an international advisory committee which consists of nine members, 3 members each from Europe, US, and Japan. This committee advises DDBJ about its maintenance, management and future plans once a year. Apart from this DDBJ also has an international collaborative committee which advises on various technical issues related to international collaboration and consists of working-level participants.

25. **European Bioinformatics Institute** (**EMBL-EBI**)

The European Bioinformatics Institute (EMBL-EBI) is an International Governmental Organization (IGO) which, as part of the European Molecular Biology Laboratory (EMBL) family, focuses on research and services in bioinformatics. It is located on the Wellcome Genome Campus in Hinxton near Cambridge, and employs over 600 full-time equivalent (FTE) staff, Institute leaders such as Rolf Apweiler, Alex Bateman, Ewan Birney, and Guy Cochrane, an adviser on the National Genomics Data Center Scientific Advisory Board, serve as part of the international research network of the BIG Data Center at the Beijing Institute of Genomics.

Other bioinformatics organisations

- National Center for Biotechnology Information, United States National Library of Medicine
- National Institute of Genetics (DNA Data Bank of Japan)
- Swiss Institute of Bioinformatics (Expasy)
- Australia Bioinformatics Resource

26. **Dynamic Programming**:

Dynamic programing is solving complex prblems by breaking them into a simpler subproblems. Dynamic programming in bioinformatics Dynamic programming is widely used in bioinformatics for the tasks such as sequence alignment, protein folding, RNA structure prediction and protein-DNA binding.

27. **Gene cluster**

A **gene cluster** is a group of two or more **genes** found within an organism's DNA that encode similar polypeptides, or proteins, which collectively share a generalized function and are often located within a few thousand base pairs of each other.

28.ClustalW

ClustalW produces multiple alignments of protein sequences, such tools are important tools in studying sequences. The basic information they provide is identification of conserved sequence regions. This is very useful in designing experiments to test and modify the function of specific proteins, in predicting the function and structure of proteins, and in identifying new members of protein families. Sequences can be aligned across their entire length (global alignment) or only in certain regions (local alignment). This is true for pairwise and multiple alignments. Global alignments need to use gaps (representing insertions/deletions) while local alignments can avoid them, aligning regions between gaps. ClustalW is a fully automatic program for global multiple alignment of DNA and protein sequences. The alignment is progressive and considers the sequence redundancy. Trees can also be calculated from multiple alignments. The program has some adjustable parameters with reasonable defaults.

29.Functional genomics

The use of genomic information to delineate protein structure, function, pathways and networks. Function may be determined by "knocking out" or "knocking in" expressed genes in model organisms such as worm, fruitfly, yeast or mouse.

30.Hidden Markov Model

A joint statistical model for an ordered sequence of variables. The result of stochastically perturbing the variables in a Markov chain (the original variables are thus "hidden"), where the Markov chain has discrete variables which select the "state" of the HMM at each step. The perturbed values can be continuous and are the "outputs" of the HMM. A Hidden Markov Model is equivalently a coupled mixture model where the joint distribution over states is a Markov chain. Hidden Markov models are valuable in bioinformatics because they allow a search or alignment algorithm to be trained using unaligned or unweighted input sequences; and because they allow position-dependent scoring parameters such as gap penalties, thus more accurately modelling the consequences of evolutionary events on sequence families.

31.Homology and Similarity Searching

Given a newly sequenced gene, there are two main approaches to the prediction of structure and function from the amino acid sequence. Homology methods are the most powerful and are based on the detection of significant extended sequence similarity to a protein of known structure, or of a sequence pattern characteristic of a protein family. Statistical methods are less successful but more general and are based on the derivation of structural preference values for single residues, pairs of residues, short oligopeptides or short sequence patterns. The transfer of structure/function information to a potentially homologous protein is straightforward when the sequence similarity is high and extended in length, but the assessment of the structural significance of sequence similarity can be difficult when sequence similarity is weak or restricted to a short region.

32.MEDLINE

MEDLINE is a bibliographic database covering the fields of medicine, nursing, dentistry, veterinary medicine, the health care system, and the pre-clinical sciences. MEDLINE searches are available using the EBI´s SRS server.

33.Multiple sequence alignment

A Multiple Alignment of k sequences is a rectangular array, consisting of characters taken from the alphabet A, that satisfies the following conditions: There are exactly k rows; ignoring the gap character, row number i is exactly the sequence sI; and each column contains at least one character different from "-". In practice multiple sequence alignments include a cost/weight function, that defines the penalty for the insertion of gaps (the "-" character) and weights identities and conservative substitutions accordingly. Multiple alignment algorithms attempt to create the optimal alignment defined as the one with the lowest cost/weight score.

34.DNA microarrays

The deposition of oligonucleotides or cDNAs onto an inert substrate such as glass or silicon. Thousands of molecules may be organized spatially into a high - density matrix. These DNA chips may be probed to allow expression monitoring of many thousands of genes simultaneously. Uses include study of polymorphisms in genes, de novo sequencing, or molecular diagnosis of disease.

35. Substitution matrix

A model of protein evolution at the sequence level resulting in the development of a set of widely used substitution matrices. These are frequently called Dayhoff, MDM (Mutation Data Matrix), BLOSUM or PAM (Percent Accepted Mutation) matrices. They are derived from global alignments of closely related sequences. Matrices for greater evolutionary distances are extrapolated from those for lesser ones.

Reference

https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470904640.app3

en.wikipedia.org › wiki › CAS