IV SEMESTER

PAPER – Bioinformatics and Biostatistics

UNIT - II

TOPIC – Databases and protein database

SOURCE – INTERNET

NAME OF INSTRUCTOR – G.N.V. SATISH

DATABASES

Data Bases: Bioinformatics data bases are organised collections of biological data (sequences, structure, functions, pathways etc).

These data bases are classified primarily as primary data base (raw data like Genbank), secondary data base (curated, derived information like PROSITE), and composite data base (integrated like NCBI).

The other way of classification based on data type like Sequence (Genbank, Uniprot), structure (PDB), functional (Gene ontology), and literature (PubMed).

Primary Data Base: It can also be called archival data base since it archives the experimental results submitted by the scientists. The primary data base is populated with experimentally derived data like Genome sequence, Macromolecular structure etc. The data entered here remains untouched (no modifications are performed over the data). It contains unique data obtained from the laboratory, and these data are made accessible to normal users without any change. The data are given accession numbers when they are entered into the data base. The same data can later be retrieved by using the accession number.

Example: Examples of primary data base – Nucleic acid data bases are Genbank and DDBJ and Protein data bases are PDB, Swiss Prot etc.

Secondary data Base :- The data stored are the analysed results of the primary data base. Computational algorithms and informative data stored inside the secondary base.

The data in the secondary base are highly curated (processing the data before it is presented in a data base). A secondary data base is better and contains more valuable knowledge compared to the primary data base.

Examples: Examples of Secondary data bases are as follows:-

1) Interpro (protein families, motifs and domains)

2) Uniprot knowledge base (sequence and functional information on protein)

Composite data Base: The data entered in these types of data bases are first compared and then filtered based on desired criteria.

The initial data are taken from the primary data base and then they are merged based on certain conditions.

It helps in searching sequences rapidly.

Examples : Swiss Prot + TREMBL

Protein Data Bases

- Protein Data Bases are a type of biological data base that are collections of information about proteins.
- The information contain in a protein data base includes the amino acid sequence, The domain structure, the biological function of the protein, its 3-Dimensional structure and its interactions with other protein.

The protein sequence data bases are useful to common user, based on the type of information stored, protein data base can be classified into several categories they are:

(1) Protein sequence data base

(2) Protein structure data base

(3) Protein-Protein interaction data base

(4) Protein-pathway and protein-protein data base

(5) Metabolic path way data base

(1) Protein sequence data base: The protein sequence data base contains amino acid sequence of proteins and related information. The amino acid sequence of a protein is important because it determines the protein's 3-Dimensional structure and function, as well as its identity.

Some of the most popular protein sequence data bases are:

(1) PIR (Protein Information Resource) is a popular protein sequence data base that provides information, and accurately annotated protein sequence.

(2) SWISS-PROT (It is a protein sequence data base that provides high levels of annotations, including information on the protein's function, domain structure, post translational modifications and variants).
Swiss-prot is jointly managed by the SIB (Swiss Institute of Bioinformatics) and the EBI (European Bioinformatics Institute)

(3) TREMBL (TREMBL is a computer annotated supplement of Swiss-prot).
It contains all the translations of EMBL (European molecular Biology Laboratory), nucleotide sequence entries that have not yet been integrated into Swiss-prot.

(2) Protein Structure Data Base –

Protein structure data bases are collection of information related to the three-dimensional structure and secondary structures of proteins.
There are several examples of protein structure data bases. Some are:

(a) **PDB**: PDB (protein Data Bank) is a world wide repository of 3D structure data on large molecules such as proteins, nucleic acids, and other biological macro molecules.

3. It stores 3-dimensional structural models of macromolecules obtained through three frequently used experimental methods - X-ray crystallography, nuclear magnetic resonance spectroscopy (NMR) and Electron Microscopy (3DEM).

(b) SCOP - SCOP (structural classification of proteins) is a Protein structure data Base that organizes proteins based on their secondary structure Properties.

- SCOP categorizes proteins into different levels based on their evolutionary relationship and structural similarities.
- Proteins with high Sequence identity or similar structure and function are grouped into families and families with similar structure but low sequence identity are placed into super families.

(C) CATH - CATH is a data base that categorizes Protein domains into hierarchical levels based on their folding pattern

(3) Protein-Protein interaction data base:-

Protein-Protein interaction data base are collection of information on the interactions between proteins. These data bases provide valuable information on the relationships between different proteins and their functions in biological systems

Examples:- BIND (Biomolecular Interaction network database)
DIP (Database of Interacting proteins)
MINT (Molecular Interactions)

(4) Protein Pattern and Profile Data base:

Protein Pattern and profile data bases contain information on motifs found in the sequences. Sequence motifs corresponding to structural or functional protein features. So the use of protein sequence pattern & profiles is a valuable tool in determining the function of proteins.

Examples:- PROSITE

(5) Metabolic pathway databases: Metabolic pathway data bases contain information about enzymes, biochemical reactions and metabolic pathways.

Examples :- ENZYME - It is a database that stores information on enzyme nomenclature
KEGG (Kyoto Encyclopedia of genes and genomes) is a comprehensive data base that maps out molecular and cellular pathways showing interactions between genes and molecules.