Quadrant II – Transcript and Related Materials

Programme: Bachelor of Science (Third Year)

Subject: Botany

Paper Code: BOD-102

Paper Title: Research Methodology, Bioinformatics and Biostatistics

Unit: IV

Module Name: Biological Databases and their Classification Format

Module No: 20

Name of the Presenter: Ms. Shreeveni S. Tari

Biological Databases and Their Classification Format

A biological database is a large, organized body of persistent data, associated with computerized software designed to update, query, and retrieve components of the data stored within the system.

The chief objective of the development of a database is to organize data in a set of structured records to enable easy retrieval of information.

Need for databases

- For storing and communicating large data sets
- To make biological data available to the scientist
- To make biological data available in computer readable form

Use

Biological databases are used for Storing, Maintaining, Entering data, Searching, Sorting, Retrieving, Presenting or displaying.

Properties

- Easy searching
- Easy to understand
- Cross referenced and connected to other databases
- Easy retrieval of data

Biological databases and their classification format

There are various criteria which can be utilised for classification of biological databases which are as follows :

- Data type – based on type of data that is stored in the database
- Data content – based on the content that is stored
- Data source – based on the data source
- Organism / Specialised – based on the data stored of a particular organism or a particular area
- Maintainer status – based on the curator for maintaining the databases
- Data access – based on the accessibility offered for usage of the data in the databases
- Database design – based on the pattern utilised for organisation of the database

One of the popular classification is based on the data source, wherein there are three major classes :

1. Primary Database

Also called as Archival Database.

Archives of raw sequence or structural data submitted by the scientific community.

Populated with experimentally derived data such as nucleotide sequence, protein sequence or macromolecular structure.

Examples

- GenBank, DNA Data Bank of Japan (DDBJ) and The European Molecular Biology Laboratory (EMBL) - nucleotide sequence
- Protein Data Bank (PDB) – three dimensional structures of biological macromolecules (the database archives atomic coordinates of macromolecules)
- Array Express Archive at EMBL-EBI and GEO at NCBI- functional genomics data

2. Secondary Databases

Secondary databases comprises computationally processed sequence information from primary databases.

The amount of computational processing work varies greatly among the secondary databases, some are simple archives of translated sequence data from identified open reading frame in DNA, whereas others provide additional annotation and information related to higher levels of information regarding structure and function.

Secondary databases often draw upon information from numerous sources, including other primary databases, controlled vocabularies and the scientific literature.

They are highly curated, often using a complex combination of computational algorithms and manual analysis and interpretation to derive new knowledge from the public record of science.

Examples :

- SWISS-PROT – sequence annotation including structure, function and protein family assignment
- UniProt Knowledgebase - sequence and functional information on proteins
- SCOP – classification of the protein structural domains
- CATH – classification of protein structures

3. Integrated / Composite Databases

Data sources having both primary and secondary data characteristics.

Integrated databases offers one stop centre for knowledge extraction.

These databases are more like consortiums managing and integrating sources of information to provide a unified access to the users.

Example :

- InterPro is an integrated documentation resource for protein families, domains and functional sites, which amalgamates the efforts of the PROSITE, PRINTS, Pfam and ProDom database projects. Each InterPro entry includes a functional description, annotation, literature references and links back to the relevant member database(s).

Databases in Various Categories

- Literature

  PubMed : scientific and medical abstracts and citations

- Health

OMIM (Online Mendelian Inheritance in Man) : information about the genes and genetic disorders

- Nucleotide sequences
  Nucleotide : DNA and RNA sequences

- Genomes
  Genome : genome sequencing projects
  dbSNP : short genetic variations

- Genes
  Protein : protein sequences
  UniProt : protein sequences and related information

- Chemicals
  PubChem compound : chemical information with structures, information and links

- Pathways
  BioSystems : molecular pathways with links to genes and proteins
  KEGG Pathway : information on biological pathways

- Organisms
  FlyBase – Genes and genome of Drosophila sp.
  Saccharomyces Genome Database (SGD) – Genes and genome of Saccharomyces sp.

- Taxonomic – information related to biological taxa