Course: PGPathshala-Biophysics
Paper 14: Bioinformatics
Module 2: BIOLOGICAL DATABASES – I
Content Writer : Dr. K. Saraboji, Sastra University

Biological Databases – I, focuses on the Introduction to biological databases, primary and secondary sequence databases of nucleotides and proteins.

Biological Databases – II, focuses on primary and secondary databases of biomolecular structures, human genes and diseases databases, microarray data and gene expression databases, metabolic pathways and metabolomic databases, specialized databases and literature databases

---

1.0 INTRODUCTION TO BIOLOGICAL DATABASES:

1.1 Background:

Now biology becomes increasingly turned into a data-rich science, so the need for strong and communicating large datasets has grown tremendously (e.g. Nucleotide and protein sequences, three-dimensional structures from X-ray crystallography and NMR). A biological database is a collection of data that is organized so that it contents can easily be accessed, managed and updated. Biological databases play a fundamental role in bioscience particularly in bioinformatics. They offer scientists the opportunity to access sequence and structure data for tens of thousands of sequences from a broad range of organisms. Biological databases represent an invaluable resource in support of biological research.

Objectives:

- ✓ **Need for the Biological Databases:**
- ✓ **Classification of biological database**
- ✓ **Types of Biological databases and its diversity**
- ✓ **Major objectives of biological databases:**
  - ***Availability of biological data to scientific community***
  - ***Availability of biological data in computer-readable form***

1.2 Classification of Biological Databases:

Sequence and structural databases are further can be classified into (i) primary, (ii) secondary and (iii) composite databases.

(i) Primary database: Consisting of data derived experimentally such as nucleotide, protein sequences and three dimensional structures alone.

Examples of these include UniProtKB for protein sequences, GenBank & DDBJ for Genome sequences and the Protein Data Bank for protein structures.

(ii) Secondary databases: Contains data that are derived from the analysis or treatment of primary data such as secondary structures, hydrophobicity plots, conserved sequence, signature sequence and domain are stored in secondary databases.

Secondary structure database contains detailed information of the PDB entry in an organized way. Example: Structural classification of protein class, fold, superfamily, etc.

Most of the secondary database created and hosted by various researchers at their individual laboratories. Example: SCOP-developed at Cambridge University, CATH-developed at University College of London, BMCD-developed at NIST, USA.

(iii) Composite databases: This merges a variety of different primary database sources, which avoids the need to search multiple resources. Different composite database use different combinations of primary database and different criteria in their search algorithm.
Example: The nucleotide and protein databases hosted at the National Center for Biotechnology Information (NCBI), provides OMIM (Online Mendelian Inheritance in Man) an online comprehensive, authoritative compendium of human genes and genetic phenotypes.

1.3 Types of Biological Databases and its Diversity:

Biological databases can be broadly classified into two categories

(i) Sequence databases: Contains nucleic acid and protein sequences information

(ii) Structure databases: Three dimensional structures of proteins, nucleic acids and macromolecular complexes.

These databases are important tools in assisting scientists to analyze and explain a host of biological phenomena from the structure of biomolecules and their interaction, to the whole metabolism of organisms and to understanding the evolution of species. This knowledge helps to facilitate the fight against diseases, assists in the development of medications, predicting certain genetic diseases and in discovering basic relationships among species in the history of life.

Sequences and structures are only among the several different types of data required in the practice of the modern biology. Other important data types includes metabolic pathway networks and molecular interactions, mutations and polymorphisms in molecular sequences and structures as well as organelle structure and tissue type, genetic maps, physicochemical data, gene and mRNA expression profiles, two dimensional gel electrophoresis images of protein expression.

Thus, biological databases are developed for diverse purposes, encompass various types of data at heterogeneous coverage and are curated at different levels with different methods, so that there are accordingly several different criteria applicable to database classification.

The two major objectives of biological databases include:

(i) *Availability of biological data to scientific community:* To store, organize and share data in a structured and searchable manner with the aim to facilitate data retrieval and visualization.

(ii) *Availability of biological data in computer-readable form:* To maintain the data in the common formats and to provide web application programming interfaces for computers to exchange and integrate data from various database resources in an automated manner.

1.4 Current Status:

The Database Issue of the journal "Nucleic Acids Research" is freely available, and categorizes many of the publicly available online databases related to biology and bioinformatics. According to a report of 21st Nucleic Acids Research Database Issue, published in 2014, there are 1552 databases that are publicly accessible online [ref] and the recent 22nd Nucleic Acids Research Database Issue reports the addition of 58 new molecular biology databases, and the updates on 115 existing databases. *(Nucleic Acids Research, 2015, Vol. 43, Database issue D1–D5)*

1.5. Types of Biological Data:

The biological data obtained from the nucleotide to the networks level results the diverse classes of biological databases which includes (i) nucleic acid sequence and structure, (ii) transcriptional regulation/gene expression patterns, (iii) protein sequence and structure, (iv) motifs and domains, (v) protein-protein interactions, (vi) metabolic and signaling pathways, (vii) metabolites, enzymes, protein modification, (viii) viruses, bacteria, protozoa and fungi, (ix) partial and whole genome sequences, (x) genomic variation, diseases and drugs, (xi) plant databases (xii) other molecular biology databases, etc.

2.0 PRIMARY SEQUENCE REPOSITORIES:

2.1. Primary Nucleotide Sequence Databases:

The following three databases are the major repositories for nucleotide sequence data from all organisms.

- NCBI (GenBank, USA) - *http://www.ncbi.nlm.nih.gov/genbank*
- EMBL-EBI (ENA - European Nucleotide Archive) - *http://www.ebi.ac.uk/ena*

- DNA Data Bank of Japan (DDBJ) - *http://www.ddbj.nig.ac.jp*

The entries in the GenBank, EMBL and DDBJ databases are synchronized on a daily basis, and the accession numbers are managed in a consistent manner between these three centers. They include sequences submitted directly by scientists and genome sequencing group, and sequences taken from literature and patents. There is comparatively minimum error checking and there is a fair amount of redundancy.

The International Nucleotide Sequence Database Collaboration (INSDC) is a long-standing foundational initiative that operates between NCBI, EMBL-EBI and DDBJ. They collaborate with Sequence Read Archive (SRA), a part of the INSDC, which archives raw sequencing data and alignment information from high-throughput sequencing instruments. This collaboration facilitates the available spectrum of data raw reads, through alignments and assemblies to functional annotation, enriched with contextual information relating to samples and experimental configurations.

2.2. Primary Protein Sequence Databases:

The major primary protein sequence databases are,

(i) PIR-PSD (Protein Information Resource) at the National Biomedical Research Foundation (NBRF) – (*http://pir.georgetown.edu/pirwww*)

(ii) SWISS-PROT at the Swiss Institute of Bioinformatics (SIB), Switzerland.

One of the most significant developments with regard to protein sequence databases is the merge of PIR-PSD, Swiss-Prot, TrEMBL databases into a single resource, UniProt (*http://www.uniprot.org*).

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.), with a minimal level of redundancy and high level of integration with other databases. SWISS-PROT is an annotated protein sequence database, which was created as the collaborative effort of the Swiss Institute of Bioinformatics (SIB) and the European Molecular Biology Laboratory (EMBL-EBI)

TrEMBL (for Translated EMBL) is a computer-annotated protein sequence database, derived from the translation of all coding sequences in the DDBJ/EMBL/ GenBank nucleotide sequence database that are not yet included in Swiss-Prot.

The sequence entries SWISS-PROT follows as closely as possible to that of the EMBL Nucleotide Sequence Database. The format differences between the Swiss-Prot and EMBL databases are available at: *http://web.expasy.org/docs/userman.html#diffEMBL*

PIR-International Protein Sequence Database (PIR-PSD), is the world's first database of classified and functionally annotated protein sequences. PIR-PSD was developed and distributed by the Protein Information Resource in collaboration with MIPS (Munich Information Center for Protein Sequences) and JIPID (Japan International Protein

Information Database), PIR-PSD has been the most comprehensive and expertly-curated protein sequence database in the public domain.

A unique characteristic feature of the PIR-PSD is its superfamily based classification of protein sequences. Further, the sequence in PIR-PSD is also classified based on homology domain and sequence motifs. Homology domains may correspond to evolutionary building blocks, while sequence motifs represent functional sites or conserved regions. The classification approach allows a more complete understanding of sequence-structure-function relationship.

In 2002, PIR joined EBI (European Bioinformatics Institute) and SIB (Swiss Institute of Bioinformatics) to form the UniProt consortium. PIR-PSD sequences and annotations have been integrated into UniProt Knowledgebase. Bi-directional cross-references between UniProt (UniProt Knowledgebase and/or UniParc) and PIR-PSD are established to allow easy tracking of former PIR-PSD entries. PIR-PSD unique sequences, reference citations, and experimentally-verified data can now be found in the relevant UniProt records.

The **UniProt** consortium comprises the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB), and the Protein Information Resource (PIR) to host the large resource of bioinformatics databases and services. The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

UniProt curators extract biological information from the literature and perform numerous computational analyses. UniProt was launched in December 2003 and comprises three databases:

- The UniProt Knowledgebase (UniProtKB) - the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. The UniProt Knowledgebase consists of two sections:

    (a) "UniProtKB/Swiss-Prot" - manually-annotated records with information extracted from literature and curator-evaluated computational analysis,

    (b) "UniProtKB/TrEMBL" computationally analyzed records that needs full manual annotation.

- UniProt Archive (UniParc) - sequence archive which contains all protein sequences from the main publicly available protein sequence databases, updates on daily basis.

- UniProt Reference Clusters (UniRef) provide clustered sets of sequences at several resolutions, which provide non-redundant views on top of the UniProt Knowledgebase and UniParc.

UniProt has been mainly supported by the National Institutes of Health, USA (NIH) grants. UniProt acts as a central hub for biomolecular information archived in more than 50 cross-referenced databases. It provides cross-references to external data collections such as the

underlying DNA sequence entries in the DDBJ/EMBL/GenBank nucleotide sequence databases, 2D PAGE and 3D protein structure databases, various protein domain and family characterization databases, post-translational modification databases, species-specific data collections, variant databases and disease databases.

3.0 SECONDARY SEQUENCE REPOSITORIES:

3.1. Secondary or Derived Nucleotide Sequence Databases:

A secondary database contains derived information from the analysis or treatment of primary databases such as GenBank or EMBL. They are value addition in terms of annotation, software, presentation of the information and the cross-references. The other type of secondary databases provides only information gathered from the sequence databases but not present the sequence at all. To date several secondary nucleotide sequence databases were developed which contain more information or links compared to primary ones, or have a different organization of the data to better suit some specific purpose.

Entrez is NCBI's primary text search and retrieval system that integrates the molecular (including DNA and protein sequence, structure, gene, genome, genetic variation and gene expression) and literature databases.

The following are some of the examples for the databases that contain subsets of the EMBL/GenBank databases:

The Nucleotide database (*http://www.ncbi.nlm.nih.gov/nuccore*) is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB.

UniGene (*http://www.ncbi.nlm.nih.gov/unigene*) process the GenBank sequence data into a non-redundant set of gene-oriented clusters. It provides automatically generated nonredundant clusters of transcript sequences; each cluster representing a distinct transcription locus. UniGene clusters also provide information on protein similarities, gene expression, cDNA clone reagents, and genomic location.

Entrez Gene *(http://www.ncbi.nlm.nih.gov/gene)* supplies gene-specific connections in the nexus of map, sequence, expression, structure, function, citation, and homology data which integrates information from a wide range of species. Entrez Gene focuses on the genomes that have been completely sequenced.

EST database (*http://www.ncbi.nlm.nih.gov/nucest*) is collection of short single-read transcript sequences from GenBank which provide a resource to evaluate gene expression, find potential variation, and annotate genes.

Ensembl (*http://www.ensembl.org/index.html*) is a joint project between EMBL-EBI and the Wellcome Trust Sanger Institute (WTSI) for the automatic annotation of eukaryotic genomes. The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Dfam (*http://dfam.org*) is a collection of repetitive DNA (transposable elements or interspersed repeats) element sequence alignments; hidden Markov models (HMMs) and matches lists for complete Eukaryote genomes.

ECRbase (*http://ecrbase.dcode.org*) offers Evolutionary Conserved Regions (ECRs), promoters, and transcription factor binding sites in vertebrate genomes, includes human, rhesus macaque, dog, opossum, rat, mouse, chicken, frog and zebrafish genomes.

CUTG (*http://www.kazusa.or.jp/codon*) is a comprehensive database for codon usage for each full-length protein gene; for each organism codon usage is calculated using the nucleotide sequence obtained from GenBank sequence database.

RefSeq (*http://www.ncbi.nlm.nih.gov/RefSeq)* provides a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins. RefSeq provides a stable reference for genome annotation, gene identification and characterization, mutation and polymorphism analysis, expression studies, and comparative analyses.

dbSNP (*http://www.ncbi.nlm.nih.gov/snp*) is a database for single nucleotide polymorphisms (SNPs), microsatellites, and small-scale insertions and deletions

MethBank (*http://dnamethylome.org*) is a DNA methylome programming database, integrates the genome-wide single-base nucleotide methylomes of gametes and early embryos at multiple diverse stages in different model organisms. Currently MethBank incorporates large cohorts of gamete and early embryo methylomes for *Danio Rerio* and *Mus musculus*.

GENEOME DATABASES: Specifically some secondary nucleotide sequence databases focuses on the genome sequences of various species including humans to annotate and analyze them and provide public access. These databases may hold many species genomes, or a single model organism genome. Examples of such genome databases are listed below:

Genome (*http://www.ncbi.nlm.nih.gov/genome*) database organizes information on sequence, map, chromosomes, assemblies, and annotations from the whole genomes of over 1000 species, includes bacteria, archaea, and eukaryota, as well as many viruses, phages, viroids, plasmids, and organelles.

The Saccharomyces Genome Database (SGD) (*http://www.yeastgenome.org*) is the database of the molecular biology and genetics of the yeast *Saccharomyces cerevisiae*.

EBI Genomes (*www.ebi.ac.uk/genomes/*) provides access and statistics for the completed genomes, and information about ongoing projects.

JGI Genome Portal of the DOE-Joint Genome Institute (JGI) (*http://genome.jgi.doe.gov*) provides databases of many eukaryote and microbial genomes. Further it provides several specialized analytical capabilities to manage and interpret complex genomic data sets

Vertebrate Genome Annotation database (VEGA) (*http://vega.sanger.ac.uk/index.html*) – is a repository for high-quality gene models produced by the manual annotation of vertebrate genomes.

FlyBase (*http://flybase.org*) is a database of Drosophila genes and genomes. Various sources including large-scale genome projects to the research literature provides different data types such as mutant phenotypes, molecular characterization of mutant alleles, cytological maps, wild-type expression patterns, sequence-level gene models and molecular classification of gene product functions.

UCSC Genome Browser (*https://genome.ucsc.edu*) contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides portals to Encyclopedia of DNA Elements (ENCODE) data and to the Neanderthal project.

PlantGDB (*http://www.plantgdb.org*) is a database of molecular sequence data for all plant species with significant sequencing efforts. The database organizes EST sequences into contigs that represent tentative unique genes.

Further, *Entrez Gene, Ensembl* databases are also severed as potential providers of genome specific details (already mentioned in previous section).

3.2. Secondary or Derived Protein Sequence Databases:

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, etc. The following are some of the examples of such derived protein sequence databases developed for different purposes:

PROSITE (*http://prosite.expasy.org*) is a database of protein domains, families and functional sites. It is based on the observation that, proteins or protein domains belonging to a particular family generally share functional attributes and are derived from a common ancestor. PROSITE currently contains patterns and profiles specific for more than a thousand protein families or domains.

Pfam (*http://pfam.xfam.org*) is a large collection of curated protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs). It is a widely used database of protein families, containing 16230 manually curated entries in the current release (Pfam 28.0).

PRIDE - The PRoteomics IDEntifications (*http://www.ebi.ac.uk/pride/archive*) is a database of protein and peptide identifications that have been described in the scientific literature, including post-translational modifications and supporting spectral evidence.

InterPro (*http://www.ebi.ac.uk/interpro*) is a database that provides functional analysis of protein sequences by classifying them into families and predicting the presence of domains

and important sites. To classify proteins in this way, InterPro uses predictive models, known as signatures, provided by several different member databases (InterPro consortium).

PRINTS (*http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS*) is a compendium of protein fingerprints, provides both a detailed annotation resource for protein families, and a diagnostic tool for newly determined sequences. Fingerprint is a group of conserved motifs which can encode protein folds and functionalities better than single motifs; such fingerprints are taken from a multiple sequence alignment used to characterize a protein family. Further its diagnostic power is refined by iterative scanning of a SWISS-PROT/TrEMBL composite.

4.0 Summary

Sequence databases are the important resource in support of biological research.

- ✓ The databases and the analysis tools facilitates much more details about a particular molecule / biological phenomena.
- ✓ Huge number of both general and specialized databases are available for a task- *For best results we often need to access multiple databases.*
- ✓ The present challenge is to handle huge volume of data (e.g. HGP) is improve the database design, develop new software to access them and sharing - ***The era of Big Data Biology***
- ✓ Biological databases includes incomplete information, data spread over multiple databases, data redundancy, errors, update frequency - *needs special attention*