# BIOLOGICAL DATABASE AND ITS CLASSIFICATION

***-BY DR. R. KAGYUNG***

A biological database is a collection of data that is organized so that its contents can easily be accessed, managed, and updated. The activity of preparing a database can be divided into:

- Collection of data in a form which can be easily accessed
- Making it available to a multi-user system

Databases can be classified in to following categories in general,

1. PRIMARY DATABASES: A primary database contains information of the sequence or structure alone, e.g. Swiss-Prot & PIR for protein sequences, GenBank & DDBJ for Genome sequences and the Protein Databank (PDB) for protein structures.

2. SECONDARY 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. 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, turns and helices. These also contain information on conserved secondary structure motifs of a particular protein. Some of the secondary databases created and hosted by various researchers at their individual laboratories include SCOP, developed at Cambridge University; CATH developed at University College of London, PROSITE of Swiss Institute of Bioinformatics, eMOTIF at Stanford. 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 51 residues 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.

3.COMPOSITE DATABASE: A composite database combines information from various primary databases for convenient searching of the desired information without querying to all these primary databases. Composite database makes searching much simpler because information from different resources is gathered in a single database. The composite database has its own format and different strategies are used to create them taking data from various primary resources. To create a good composite database, the entries should be taken from validated and well annotated primary resources. OWL, MISPX, NRDB, are examples of composite database.

Modern biological databases comprise not only data, but also sophisticated query facilities and bioinformatic data analysis tools; hence, the term "bioinformatic databases" is often used.

Biological databases can be broadly classified in to three categories,

1. **Sequence databases**
2. **Structure databases and**
3. **Pathway databases**

Sequence databases are applicable to both nucleic acid sequences and protein sequences, whereas structure databases are applicable to only Proteins.

### 1. Sequence databases

Nucleotide and protein sequence databases represent the most widely used and some of the best established biological databases. These databases serve as repositories for wet lab results and the primary source for experimental results. Major public data banks which takes care of the DNA and protein sequences are GenBank in USA, EMBL (European Molecular Biology Laboratory) in Europe and DDBJ (DNA DataBank) in Japan.

a. **GenBank:** The GenBank nucleotide database is maintained by the National Center for Biotechnology Information (NCBI), which is part of the National Institute of Health (NIH), a federal agency of the US government.

b. **EMBL:** The EMBL nucleotide sequence database is maintained by the European Bioinformatics Institute (EBI) in Hinxton and

c. **DDBJ:** DNA Data Bank of Japan Is a biological database that collects DNA sequences submitted by researchers. It is run by the National Institute of Genetics, Japan.

d. **Ensembl:** The Ensembl database is a repository of stable, automatically annotated human genome sequences. Ensembl annotates and predicts new genes, with annotation from the InterPro protein family databases and with additional annotations from databases of genetic disease-OMIM, expression-SAGE and gene family.

e. **SGD:** The Saccharomyces Genome Database (SGD) is a scientific database of the molecular biology and genetics of the yeast Saccharomyces cerevisiae.

f. **dbEST:** dbEST is a division of GenBank that contains sequence data and other information on short, "single-pass" cDNA sequences, or Expressed Sequence Tags (ESTs), generated from randomly selected library clones. Expressed

Sequence Tags (ESTs) are currently the most widely sequenced nucleotide commodity in the terms of number of sequences and total nucleotide count.

g. **PIR:** The Protein Information Resource (PIR) is an integrated public bioinformatics resource that supports genomic and proteomic research and scientific studies. PIR has provided many protein databases and analysis tools to the scientific community, including the PIR-International Protein Sequence Database (PSD) of functionally annotated protein sequences. The PIR-PSD, originally created as the Atlas of Protein Sequence and Structure edited by Margaret Dayhoff, contained protein sequences that were highly annotated with functional, structural, bibliographic, and sequence data.

h. **Swiss-Prot:** Swiss-Prot is a protein sequence and knowledge database. It is well known for its minimal redundancy, high quality of annotation, use of standardized nomenclature, and links to specialized databases. As Swiss-Prot is a protein sequence database, its repository contains the amino acid sequence, the protein name and description, taxonomic data, and citation information.

i. **TrEMBL:** The European Bioinformatics Institute, collaborating with Swiss-Prot, introduced another database, TrEMBL (translation of EMBL nucleotide sequence database). This database consists of computer annotated entries derived from the translation of all coding sequences in the nucleotide databases.

j. **UniProt:** UniProt database is organized into three layers. The UniProt Archive (Uni-Parc) stores the stable, nonredundant, corpus of publicly available protein sequence data. The UniProt Knowledge base (UniProtKB) consists of accurate protein sequences with functional annotation. Finally, the UniProt Reference Cluster (UniRef) datasets provide nonredundant reference clusters based primarily on UniProtKB. UniProt also offers users multiple tools, including searches against the individual contributing databases, BLAST and multiple sequence alignment, proteomic tools, and bibliographic searches.

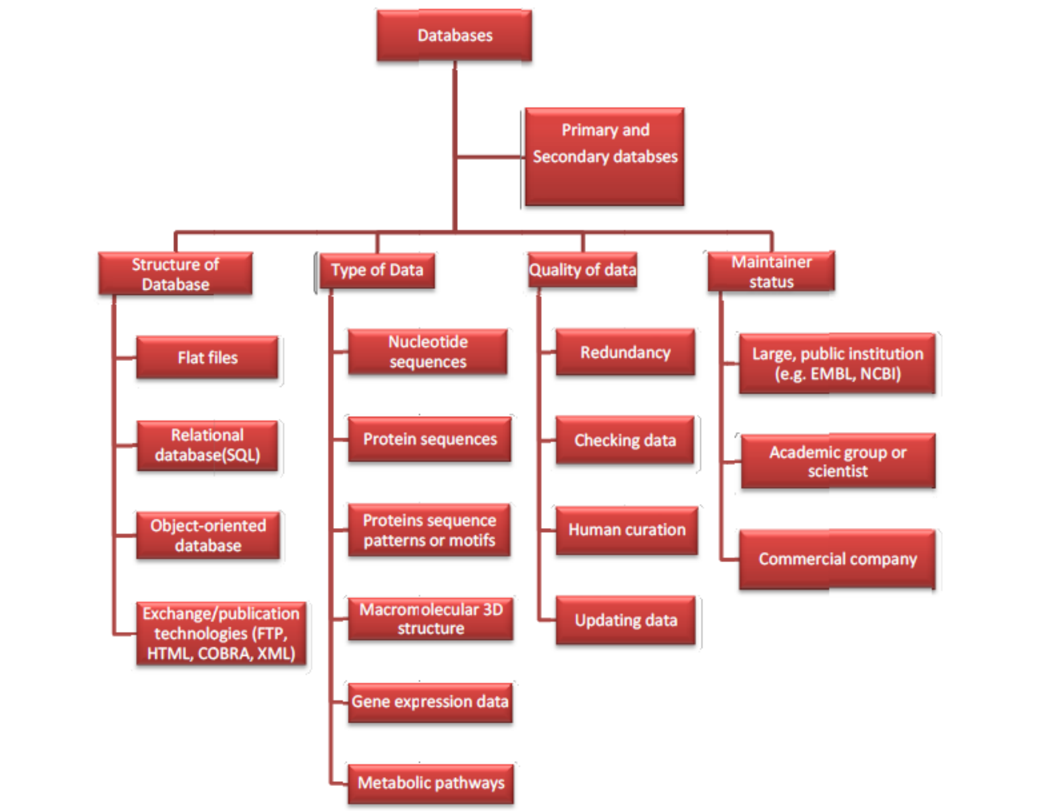

**Figure 1:** Schematic representation of Database.

## 2. Structure databases

Knowledge of protein structures and of molecular interactions is key to understanding protein functions and complex regulatory mechanisms underlying many biological processes.

a. **Protein Data Bank:** The PDB (Protein Data Bank) is the single worldwide archive of Structural data of Biological macromolecules, established in Brookhaven National Laboratories in 1971. It contains Structural information of the macromolecules determined by X-ray crystallographic, NMR methods. PDB is maintained by the Research Collaboratory for Structural Bioinformatics (RCSB). It allows the user to view data both in plain text and through a molecular viewer using Jmol.

**b. SCOP:** The SCOP (Structural Classification of Proteins) database was started by Alexey Murzin in 1994. Its purpose is to classify protein 3D structures in a hierarchical scheme of structural classes.

**c. CATH:** The CATH database (Class, architecure, topology, homologous superfamily) is a hierarchical classification of protein domain structures, which clusters proteins at four major structural levels.

**d. NDB**: Nucleic Acid Database, also curated by RCSB and similar to the PDB and the Cambridge Structural Database is a repository for nucleic acid structures. It gives users access to tools for extracting information from nucleic acid structures and distributes data and software.

**Pathway databases**

Development of metabolic databases derived from the comparative study of metabolic pathways will cater the industrial needs in more efficient manner to further the growth of systems biotechnology. Some examples of the pathway databses are KEGG, BRENDA, Biocyc.

**Table 1:** Summary of Nucleotide sequence databases.

| Database | URL | Features |
|---|---|---|
| GenBank [31] | http://www.ncbi.nlm.nih.gov/ | NIH's archival genetic sequence database |
| EMBL | http://www.ebi.ac.uk/embl/ | EBI's archival genetic sequence database |
| DDBJ | http://www.ddbj.nig.ac.jp/ | NIG's archival genetic sequence database |
| SGD | http://www.yeastgenome.org/ | A repository for baker's yeast genome and biological data |
| EBI genomes | http://www.ebi.ac.uk/genomes/ | It provides access and statistics for the completed genomes [32] |
| Ensembl | http://www.ensembl.org/ | Database that maintains automatic annotation on selected eukaryotic genomes [33] |
| UniGene | http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene | Each UniGene cluster contains sequences that represent a unique gene, as well as related information. |
| dbEST | http://www.ncbi.nlm.nih.gov/dbEST/ | Division of GenBank that contains expression tag sequence data |

**Table 2:** Summary of Protein sequence databases.

| Database | URL | Features |
|---|---|---|
| Swiss-Prot/TrEMBL | http://www.expasy.org/sprot/ | Description of the function of a protein, its domains structure, post-translational modifications etc, |
| UniProt [48] | http://www.pir.uniprot.org/ | Central repository for PIR, Swiss-Prot, and TrEMBL |
| PIR | http://pir.georgetown.edu/ | It strives to be comprehensive, well-organized, accurate, and consistently annotated. |
| Pfam | pfam.sanger.ac.uk/ | Database of protein families defined as domains [49] |
| PROSITE | www.expasy.ch/prosite/ | Database of protein families and domains |

**Table 3:** Summary of Structure databases.

| Database | URL | Feature |
|---|---|---|
| PDB | www.rcsb.org/pdb/ | Protein structure repository that provides tools for analyzing these structures |
| SCOP | scop.mrc-lmb.cam.ac.uk/scop/ | Classification of protein 3D structures in a hierarchical scheme of structural Classes |
| CATH | www.cathdb.info | Hierarchical classification of protein domain structure |
| NDB | http://ndbserver.rutgers.edu/ | Database housing nucleic acid structural information |

**Table 4:** Summary of Pathway databases.

| Database | URL | Feature |
|---|---|---|
| KEGG | http://www.genome.jp/kegg/ | Protein structure repository that provides tools for analyzing these structures |
| BioCyc | http://www.biocyc.org/ | Classification of protein 3D structures in a hierarchical scheme of structural classes |
| BRENDA | http://www.brenda-enzymes.org/ | Hierarchical classification of protein domain structure |
| EMP | http://emp.mcs.anl.gov/ | Database of Enzymes and Metabolic pathways public server |
| BRITE | http://www.genome.jp/kegg/brite.html | Biomolecular Relations in Information, Transmission and Expression |

**a. KEGG:** The Kyoto Encyclopedia of Genes and Genomes (KEGG) is the primary resource for the Japanese GenomeNet service that attempts to define the relationships between the functional meanings and utilities of the cell or the organism and its genome information. KEGG contains three databases: PATHWAY, GENES, and LIGAND. The PATHWAY database stores computerized knowledge on molecular interaction networks. The GENES database contains data concerning sequences of genes and proteins generated by the genome projects. The LIGAND database holds information about the chemical compounds and chemical reactions that are relevant to cellular processes.

**b. BRENDA:** It is the main collection of enzyme functional data available to the scientific community. It is maintained and developed at the Institute of Biochemistry and Bioinformatics at the Technical University of Braunschweig, Germany.

**c. BioCyc:** The BioCyc Database Collection is a compilation of pathway and genome information for different organisms. It includes two other databases, EcoCyc, which describes Escherichia coli K-12; MetaCyc, which describes pathways for more than 300 organisms.

As biology has increasingly turned into a data rich science, the need for storing and communicating large datasets has grown immensely. The examples are the nucleotide sequences, the protein sequences, and the 3D structural data produced by X-ray crystallography and NMR. Biological databases are an important tool in assisting scientists to understand 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 facilitate the fight against diseases, assists in the development of medications and in discovering basic relationships amongst species in the history of life.