2

BIOLOGICAL DATA AND DATABASE

Biological Data

The data that are collected from biological world are called biological data. For example, DNA sequence data, population data, genetical data, ecological data etc. However, bioinformatics deals with biomolecule's related data collected from scientific experiments, published literatures and computational analyses. With the advent of new techniques and computational tools there is an exponential increase of information on biomolecules. Bioinformaticists store these biological data (data on DNA, RNA and protein) in digitalize form i.e. in database. For world wide data submission and management, access, exchange through internet bioinformaticists constructed databases where data on biomolecules are stored in systematically. Till date vast range of databases have been constructed by the bioinformaticists and made accessible through internet for the researchers. The databases include DNA, RNA and protein sequence data, structural information, gene expression data, molecular interaction data, mutation data, phenotypic data, metabolic pathways information, taxonomic information of biological organism etc.

Biological Database

The Biological data can be broadly classified as:

Biological Databases	**Information they contain**
1. Bibliographic databases	Literature
2. Taxonomic databases	Classification
3. Nucleic acid databases	DNA information
4. Genomic databases	Gene level information
5. Protein databases	Protein information
6. Protein families, domains and functional sites	Classification of proteins and identifying domains
7. Enzymes/ metabolic pathways	Metabolic pathways

Biological databases act as repository of biological information for utilization by researchers. The biological database is a collection of "entry" which is the unit of the data. Each entry includes nucleotide sequence data, protein sequence and 3D structure data, the biological nature such as gene function and other property of the sequence etc. and the information of submitters, references, source organisms. Biological databases are coherent, consistent and designed for a specific purpose to store a set of clearly defined data in an organized manner. The contents of database can easily be accessed, managed and updated. The contents of database can be analyzed with the help of defined algorithm. The biological databases are of various types which can be outlined as follows –

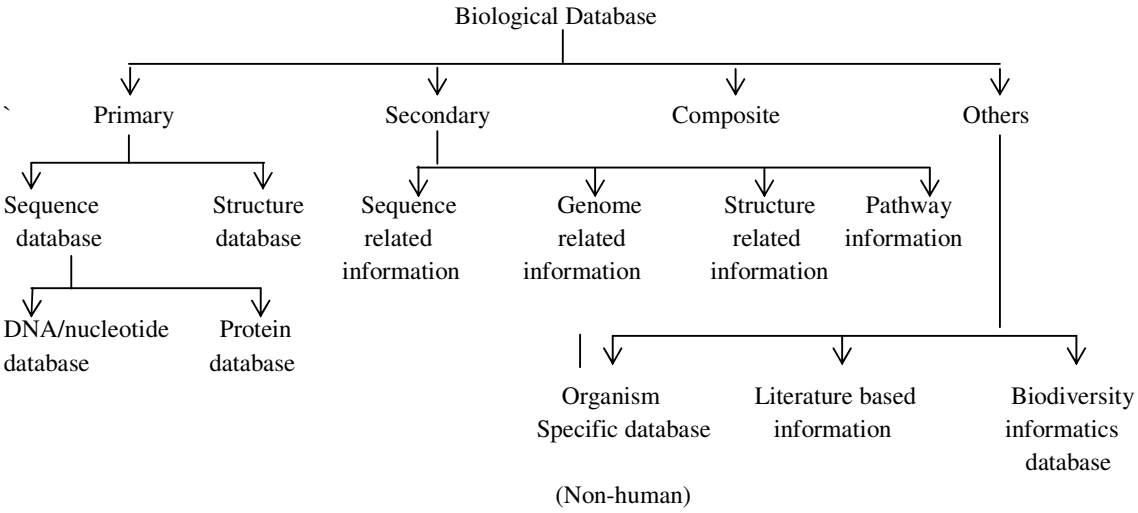

Fig. 2.1: Classification of biological databases.

A. **Primary Database**

Primary databases are repository of raw sequenced and annotated data that derived experimentally such as nucleotide sequences, three dimensional structures of protein and signifies important properties of each sequence. These primary databases can be accessed freely through internet over World Wide Web (W.W.W). Primary databases can be further classified as given below, –

1. Sequence database: Sequence database stores information on sequence of DNA/nucleotide and protein.

i. DNA/nucleotide database: DNA/nucleotide databases stores data on DNA/nucleotide sequence. Each database maintains an own set of submission and retrieval tools, but they exchange data daily so that all the databases should contain the same set of sequences. Some important examples of DNA/nucleotide databases are given below, –

GenBank

GenBank sequence database is an open access and annotated collection of nucleotide sequences and their protein translations including mRNA sequences with coding regions, segments of genomic DNA with a single gene or multiple genes, and ribosomal RNA gene clusters. GenBank is produced and maintained by the National Centre for Biotechnology Information (NCBI) as part of the International collaboration with EMBL Data Library from the EBI and the DNA Data Bank of Japan (DDBJ). Individual laboratory can submit sequence data or large scale sequencing centre can submit bulk submission directly to the GenBank by using Banklt or Sequin. The Banklt is a web-based form and Sequin is a stand-alone software tool developed by the NCBI for submitting and updating sequence to the GenBank, EMBL and DDBJ databases. After sequence submission the GenBank staffs assigns an Accession Number to the newly entered sequence

and performs quality assurance checks. Then the newly submitted sequence is released to the database. Data that are stored in GenBank can be retrieved by Entrez or by downloading File Transfer Protocol (FTP). The GenBank is a collection of information on Expressed Sequence Tag (EST), Sequence Tagged Site (STS), Genome Survey Sequence (GSS), and High-Throughput Genome Sequence (HTGS) and complete microbial genome sequences. Information of GenBank can be accessed through the server http://www.ncbi.nlm.nih.gov/genbank/. There are several ways to search and retrieve data from GenBank as given under –

- Search GenBank for sequence identifiers and annotations with Entrez Nucleotide , which is divided into three divisions: CoreNucleotide (the main collection), dbEST (Expressed Sequence Tags), and dbGSS (Genome Survey Sequences).
- Search and align GenBank sequences to a query sequence using BLAST.
- Search, link, and download sequences programmatically using NCBI e-utilities .

DNA Data Bank of Japan (DDBJ)

DDB is a kind of nucleotide sequence data bank that receives nucleotide sequence from researchers and assigns an accession number to data submitters. DDBJ collects sequence data mainly from Japanese researchers, however, they also receive data and assign accession number to researchers of any other countries. DDBJ began data bank activities in 1986 at National Institute of Genetics (NIG). Currently, DDBJ is in operation at NIG in Mishima, Japan.

Main activities of DDBJ are – i) being a member of INSDC, DDBJ collects nucleotide sequence data from researcher, assigns an accession number to the data submitters exchanges the collected data with EMBL-Bank and GenBank on a daily basis, ii) DDBJ manage bioinformatics tools for data submission and retrieval, iii) DDBJ develops tools for analysis of biological data and iv) organizes Bioinformatics Training Course in Japanese to teach how to analyze biological data. Information of DDBJ can be accessed through the server http://**www.ddbj.nig.ac.jp**.

European Molecular Biology Laboratory- European Bioinformatics Institute (EMBL-EBI)

European Bioinformatics Institute (EBI) is part of **European Molecular Biology Laboratory (EMBL)**. EMBL-EBI now known as EMBL-Bank and was established in 1980 at the EMBL in Heidelberg, Germany. It was the world's first nucleotide sequence database. EMBL-EBI provides freely available data from life science experiments, performs basic research in computational biology and offers an extensive user training programme for the researchers. EMBL-EBI stores data on DNA and RNA (genes, genomes and variation), gene expression (RNA, protein and metabolite expression), protein (sequence, families and motifs), structure (molecular and cellular structures), systems (reaction, interaction, pathways), chemical biology (chemogenomics and metabolomics), ontologies (taxonomies and controlled vocabularies) and literature (scientific publications and patents). EMBL-EBI can be accessed through the server http://www.ebi.ac.uk.

EnsEMBL

Ensembl is a joint project between EBI, EMBL and the Welcome Trust Sanger Institute to develop a software system that produces and maintains automatically annotation on selected eukaryotic genomes. EnsEMBL was stated in 1999 with an aim to automatically annotate the genome, integrate this annotation with other available biological data and release the information to the researchers via the web. EnsEMBL produces genome databases for vertebrates and other eukaryotic species and makes this information freely available online. EnsEMBL can be freely accessed through the server http://www.asia.ensembl.org**.** Various research projects around the world contribute DNA sequence and their assemblies data to the EnsEMBL. This database emphasizes on two areas of comparative genomics – the creation of gene trees using representative proteins from each gene in a species, and the alignment of DNA sequences to infer synteny, conservation, etc. The Ensembl Variation database stores data on the regions of genome that differ between individual genomes, associated disease and phenotype information. Ensembl Regulation stores data on the mechanisms of gene regulation in human and mouse cells, transcriptional and post-transcriptional mechanisms.

ii. **Protein database:** Primary protein sequence databases store information on protein sequence. Some primary protein sequence databases are briefly describe hereunder –

Protein Information Resource (PIR)

PIR was developed by the National Biomedical Research Foundation (NBRF) in 1984 to assist researchers in the identification and interpretation of protein sequence information. It is an integrated public resource of protein informatics that supports genomic and proteomic research and scientific discovery. PIR has three distinct sections – PIR1 contains fully classified and annotated entries, PIR2 contains preliminary entries that has not been thoroughly reviewed and contain redundancy, PIR3 contain unverified entries and PIR4 has one of the following categories – i. conceptual translations of art factual sequences, ii. conceptual translations of sequences that are not transcribed or translated, iii. protein sequences or conceptual translations that are genetically engineered, iv. sequences that are not genetically encoded and not produced on ribosomes. PIR maintains the Protein Sequence Database (PSD) that stores over 283 000 sequences. For over four decades PIR has been providing protein databases and analysis tools those are freely accessible to the researchers including the Protein Sequence Database (PSD). The PIR has a bibliography system for literature searching, mapping, and user submission. PIR also maintains a (Non-redundant Reference) NREF database, and iProClass, an integrated database of protein family, function and structure information. PIR-NREF protein sequences information. Currently it is consisting of more than 1 000 000 entries from PIR-PSD, SWISS-PROT, TrEMBL, RefSeq, GenPept, and PDB. The PIR web site is http://www.pir.georgetown.edu.

UniProtKB/Swiss-Prot

UniProtKB/SwissProt is the manually annotated and reviewed section of the UniProt knowledgebase (UniProtKB). It is an annotated and non-redundant (means less identical sequences are present in the database) protein sequence database. Swiss-Prot was established in 1986 and maintained collaboratively by the EMBL Outstation (EBI) and the Swiss Institute of Bioinformatics (SIB). It provides information on domain structure of protein, its function, post-translational modification, variants etc. The Swiss-Prot database distinguishes itself from other protein sequence databases by three distinct criteria – i. annotation, ii. minimal redundancy and iii. integration with other databases. In 1996, a computer-annotated supplement to SWISS-PORT was created and named as TrEMBL (Translation of EMBL nucleotide sequence database). TrEMBL consists of computer-annotated entries derived from the translation of all coding sequence (CDSs) in the EMBL database except for CDSs already present in Swiss-Prot (Bairoch and Apweiler, 2000). The server for SwissProt and TrEMBL are www.ebi.ac.uk/uniprot and www.ebi.ac.uk respectively.

Protein Sequence Database (PSD)

PIR-International Protein Sequence Database (PIR-PSD), the world's first database of classified and functionally annotated protein sequences that grew out of the Atlas of Protein Sequence and Structure (1965-1978) edited by Margaret Dayhoff. It is produced 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 for over 20 years. The URL for PSD is http:// www.pir.georgetown.edu.

2. **Structure database:** Structure databases are created to store structure information of biological macromolecules e.g. nucleic acid and protein. Some important structure databases are given here under –

Protein Data Bank (PDB)

Protein Data Bank was established at Brookhaven National Laboratories (BNL) in 1971. PDB contains 3D structure of protein that is established by x-ray crystallographic and nuclear magnetic resonance (NMR) studies and is maintained by Research Collaboratory for Structural Bioinformatics (RCSB) at Rutgers University. As on December 24, 2013 there are 96596 structures of proteins available at PDB whose provide information on atomic co-ordinate of amino acids in protein, protein fragments or protein bound to substrate or inhibitors. Protein structure data can be deposited in the PDB using a web-based AutoDep Input Tool (ADIT). Molecular structure of protein of PDB can be displayed by molecular graphics program such as Resnol, Chine CnED (Bansal, 2009). PDB's URL is http://www.rcsb.org/pdb/.

Nucleic Acid Database (NDB)

Nucleic Acid Database is a comprehensive database of 3D structures of nucleic acids. The goal of the Nucleic Acid Database Project (NDB) is to store and distribute structural information about nucleic acids. The NDB was founded in 1992 by Helen M. Berman and Wilma K. Olson of Rutgers University and David Beveridge of Wesleyan University.

Cambridge Crystallographic Data Centre/Cambridge Structural Database (CCDB/CSD)

Cambridge Crystallographic Data Centre (CCDC) is a crystallographic non-profit organisation that is situated in Cambridge, England. It compiles and distributes the Cambridge Structural Database (CSD). CSD is a database of small molecules. It stores experimentally determined organic and metal-organic crystal structural data. CSD's URL is www.ccdc.cam.ac.uk.

B. Secondary database

The secondary database contains the data of analyses of the primary sequence data in form of regular pattern, fingerprints, blocks, profiles or Hidden Markov Model. Thus, the secondary database contains the fruit of analyses of the sequences stored in the primary database and contains information about conserved sequence, signature sequence, active site residues of protein families arrived by multiple sequence alignment of a set of related proteins. ProSite, Profiles, Prints, Pfam, REBase etc. are some examples of secondary database. Secondary database falls under four categories – sequence related database, genome related database, structure related database and pathway database.

1. Sequence related database

ProSite

ProSite is a secondary protein database and contains information of protein families, domains, functional sites as well as amino acid pattern and profile of protein. These information are manually curated by a team of the Swiss Institute of Bioinformatics. ProSite was created by Amos Bairoch in 1988. ProSite offers tools for protein sequence analysis and motif detection. The database ProRule builds on the domain description of ProSite (Sigrist *et al*, 2005). It provides information about structurally and functionally critical amino acid. The URL for ProSite is http://www.prosite.expasy.org.

Pfam

Pfam is a database of protein families that provides multiple sequence alignment and Hidden Markov Models (HMM) for protein domains of all types of organisms. Information on multiple sequence alignment is generated by using HMM (Finn *et al*, 2008; Finn *et al*, 2006 and Bateman *et al*, 2005). Pfam can be searched for multiple sequence

alignment, protein domain structure, to examine species distribution, to follow other databases and to view known protein structure. The Pfam has two components. Pfam-A stores manually curated high quality entries. For each entry a protein sequence alignment and a HMM is stored. Automatically generated lower quality entries are stored in Pfam-B. Entries in Pfam-A do not cover all known proteins, in that case lower quality Pfam-B families can be used.

ExplorEnz

ExplorEnz is an open access, manually curated and peer reviewed enzyme database. ExplorEnz is designed, developed and maintained by Andrew McDonald at School of Biochemistry and Immunology, Trinity College, Dublin, Ireland to access the data of the International Union of Biochemistry and Molecular Biology (IUBMB) enzyme nomenclature list.

REBase

REBase is the restriction enzyme database. It is maintained by Richard J. Roberts since before 1980 (Roberts, 1980). It stores information about restriction enzymes and related proteins, published and unpublished references, recognition and cleavage sites, isoschizomers, commercial availability, methylation sensitivity, crystal, genome, sequence data, DNA methyltransferases, homing endonucleases, nicking enzymes, specificity subunits and control proteins. The official website for REBase is http://www.rebase.neb.com.

2. Genome related database

Online Mendelian Inheritance in Man (OMIM)

OMIM is a database that stores informations of human genes and genetic disorders, genetic variation in human and pictures. It also provides references for further research and tools for genomic analysis of a catalogued gene (Hamosh, 2004). Mendelian Inheritance in Man (MIM) was started in the early 1960s and is available in form of book. The online version i.e. OMIM has been available since 1987 (McKusick, 1993). The information in OMIM is collected and processed under the leadership of Dr. Victor A. McKusick at Johns Hopkins University, assisted by a team of science writers and editors. A six digit number is assigned against every disease and gene of which the first number classifies the method of inheritance. The initial digit 1 signifies autosomal dominant trait; 2, signifies autosomal recessive trait; 3, signifies X-linked trait; a number symbol (#) before an entry number indicates that it is a descriptive entry; a plus sign (+) before an entry number indicates that the entry contains the description of a gene of known sequence and a phenotype; a percent sign (%) before an entry number indicates that the entry describes a confirmed Mendelian phenotype or phenotypic locus for which the underlying molecular basis is not known; no symbol before an entry number indicates a description of a phenotype for which the Mendelian basis has not been clearly established and a caret (^) before an entry number means the entry no longer exists because it was removed from the database or moved to another entry (http://www.omim.org/help/faq#1.2).

Plant Transcription Factor Database (PlnTFDB)

PlnTFDB is a public database that identifies and stores all plant genes involved in transcriptional control. The Plant Transcription Factor Database (PlnTFDB; http://plntfdb.bio.uni-potsdam.de/v3.0/) is an integrative database that provides complete sets of transcription factors (TFs) and other transcriptional regulators (TRs) in plant species whose genomes have been completely sequenced and annotated. The complete sets of 84 families of TFs and TRs from 19 species ranging from unicellular red and green algae to angiosperms are included in PlnTFDB, representing >1.6 billion years of evolution of gene regulatory networks. TF or TR gene entries include information of expressed sequence tags, 3D protein structures of homologous proteins, domain architecture and cross-links to other computational resources online. Moreover, the different species in PlnTFDB are linked to each other by means of orthologous genes facilitating cross-species comparisons. (Paulino, 2010).

3. Structure related database

Database of Secondary Structure Assignments (DSSP)

The DSSP program was designed by Wolfgang Kabsch and Chris Sander for standardization of secondary structure assignment. Basically DSSP is a database of secondary structure assignments for all protein entries in the Protein Data Bank (PDB). DSSP does not predict secondary structure.

Homology-derived Secondary Structure of Proteins (HSSP)

HSSP was designed by Chris Sander and Reinhard Schneider. For each know protein structure the database contains the aligned sequence, secondary structure, sequence variability and sequence profile (Sander and Schneider, 1991). HSSP also implies tertiary structure of protein.

Families of Structurally Similar Proteins (FSSP)

FSSP is a database of structurally similiar proteins generated using the "Distance-matrix Alignment" (DALI) algorithm. The database is helpful for the comparison of protein structures.

Dali database

Dali is the secondary structure database, based on comparison of 3D protein structures of Protein Data Bank (PDB).

4. Pathway information database

Pathway information database provide information on molecular interaction in biological process, information about genes and proteins and information about the chemical compounds and their reactions.

C. Composite database

Composite databases store data of different primary databases, thus obviates the need to search multiple primary databases for nucleotide sequence, protein sequence, protein structure etc. Examples of some composite databases are –

1. nrdb (nonredundant database) combines and stores sequences from GenBank (CDS translations), PDB, Swiss-Prot, PIR, and PRF.
2. INSD (International Nucleotide Sequence Database) stores nucleotide sequences of EMBL, GenBank, and DDBJ.
3. UniProt (universal protein sequence database) is a collection of protein sequences from PIR-PSD, Swiss-Prot, and TrEMBL.

D. Others

Other databases includes organism specific database, literature based database and biodiversity informatics database. Organism specific databases provide information on various organisms other than human like- virus, bacteria, fungi, invertebrates, Drosophila, beetle, silkworm etc. Organism specific databases store information on genome of respective organism, genome map, gene expression, genomic mutation of Drosophila, morphological features of Yeast, growth requirements of bacterial pathogen etc. Literature based database contain scientific articles in form of abstract or full paper published by researchers in different journals. There are several high quality literature based databases but the most popular is PubMed. Biodiversity informatics databases contain information on biodiversity of different organisms.

Biological database retrieval system

There are three important data retrieval systems: Entrez (at NCBI), Sequence Retrieval System, SRS (at EBI) and DBGET/LinkDB (at Japan). These retrieval systems not only return matches to a query, but also provide additional important information in related databases.

ENTREZ

Entrez is a molecular biology database and retrieval system developed by the National Center for Biotechnology Information (NCBI). It can be accessed through http://www.ncbi.nlm.nih.gov/Entrez/. The Entrez system provides access to Nucleotide sequence databases – GenBank/DDBJ/EBI; Protein sequence databases - Swiss-Prot, PIR, PRF, PDB, and translated protein sequences from DNA sequence databases; Genome and chromosome mapping data; Molecular Modeling 3-D structures Database; Literature database, PubMed - provides access to MEDLINE and pre-MEDLINE articles; Taxonomy database - allows retrieval of DNA and protein sequences for different taxonomic groups; Specialized Databases – OMIM, dbSNP, UniSTS, etc.

The most valuable feature of Entrez is- by exploiting the concept of ‘neighbouring’ it provides access to related articles of linked databases. For example, in a nucleotide sequence page, one may find cross-referencing links to the translated protein sequence, genome mapping data or to the related PubMed literature information and to protein structures if

available (Xiong, 2006). Another useful feature in Entrez is – one can retrieve large sets of data on the basis of some criterion and can download them to a local computer.

Sequence Retrieval System (SRS)

The SRS is a network browser for databases in molecular biology. It can be accessed through http://srs.ebi.ac.uk/. SRS searches 80 biological databases developed at the European Bioinformatics Institute (EBI) at Hinxton, UK. It provides access to sequence and sequence related, metabolic pathways, transcription factors, application results (e.g., BLAST), protein 3D-structure, genome, mapping, mutations, and locus-specific mutations databases.

Advantage of SRS is that it rapidly searches query, allowing users to retrieve, link and access entries from all the interconnected resources.

DBGET/LinkDB

It is an integrated bioinformatics database retrieval system at GenomeNet, developed by the Institute for Chemical Research, Kyoto University, and the Human Genome Center of the University of Tokyo. It can be accessed through http://www.genome.ad.jp/dbget/. DBGET provides access to about 20 molecular biology databases, which can be queried one at a time. After querying one of database, DBGET provides links to associate information in addition to the list of results. A unique feature of DBGET is its connection with the Kyoto Encyclopedia of Genes and Genomes (KEGG) database - a database of metabolic and regulatory pathways.

***