Notes on Biological Databases

Edited by: Dr. Rameshori Yumnam

Biological Databases

- These are the databases consisting of biological data like protein sequencing, molecular structure, DNA sequences, etc in an organized form.
- Several computer tools are there to manipulate the biological data like an update, delete, insert, etc. Scientists, researchers from all over the world enter their experiment data and results in a biological database so that it is available to a wider audience.
- Biological databases are free to use and contain a huge collection of a variety of biological data.

There are two main functions of biological databases:

1. **Make biological data available to scientists.**

As much as possible of a particular type of information should be available in one single place (book, site, database). Published data may be difficult to find or access, and collecting it from the literature is very time-consuming. And not all data is actually published explicitly in an article (genome sequences!).

2. **To make biological data available in computer-readable form.**

Since analysis of biological data almost always involves computers, having the data in computer-readable form (rather than printed on paper) is a necessary first step.

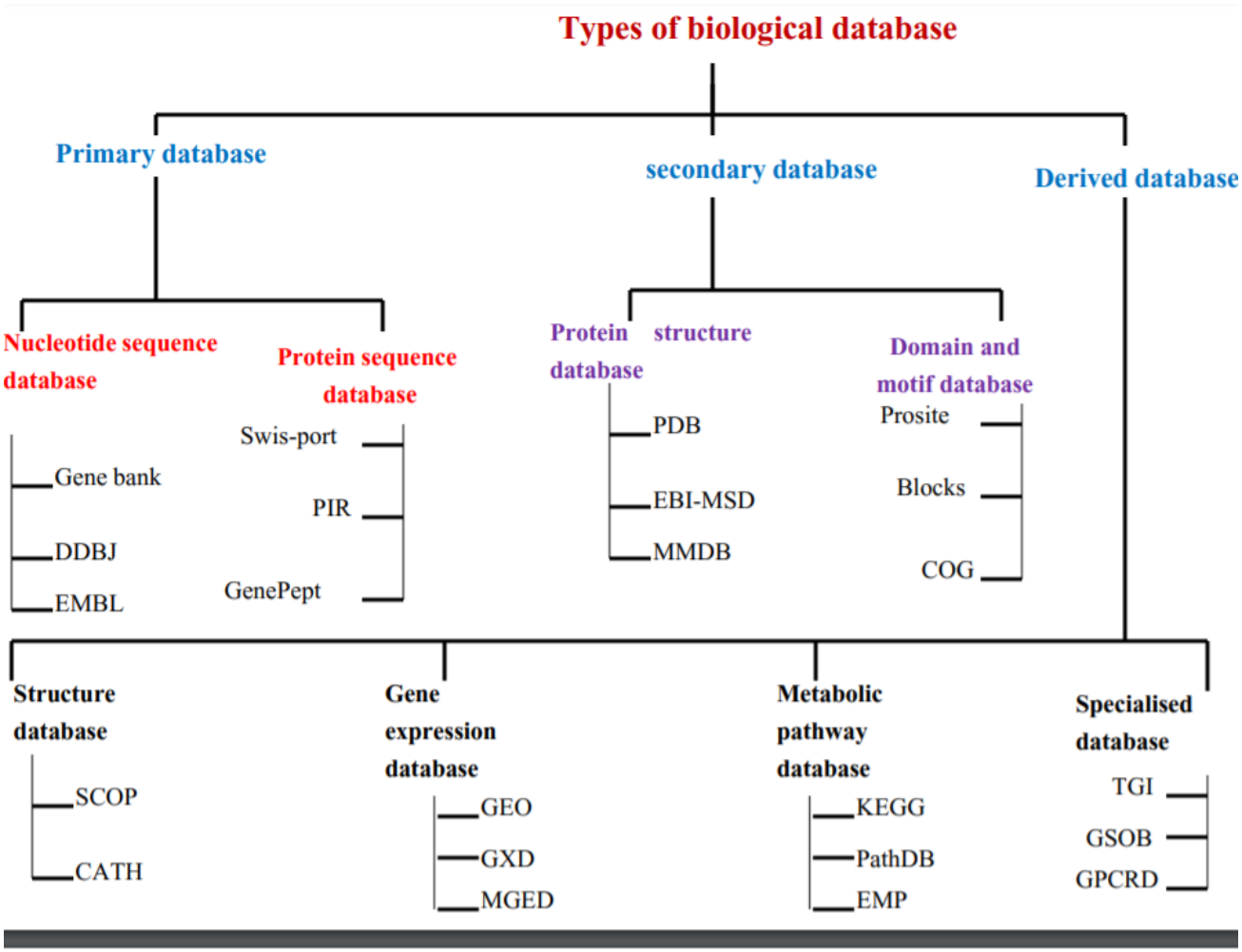
One of the first biological sequence databases was probably the book "Atlas of Protein Sequences and Structures" by Margaret Dayhoff and colleagues, first published in 1965. It contained the protein sequences determined at the time, and new editions of the book were published well into the 1970s. Its data became the foundation for the PIR database.

The computer became the storage medium of choice as soon as it was accessible to ordinary scientists. Databases were distributed on tape, and later on various kinds of disks. When universities and academic institutes were connected to the Internet or its precursors (national computer networks), it is easy to understand why it became the medium of choice. And it is even easier to see why the World Wide Web (WWW, based on the Internet protocol HTTP) since the beginning of the 1990s is the standard method of communication and access for nearly all biological databases.

As biology has increasingly turned into a data-rich science, the need for storing and communicating large datasets has grown tremendously. The obvious examples are the nucleotide sequences, the protein sequences, and the 3D structural data produced by X-ray crystallography and macromolecular NMR. An new field of science dealing with issues,

challenges and new possibilities created by these databases has emerged: bioinformatics. Other types of data that are or will soon be available in databases are metabolic pathways, gene expression data (microarrays), protein-protein interactions and other types of data relating to biological function and processes.

One very important issue is the frequency and type of errors among the entries of a database. Naturally, this depends strongly on the type of data, and whether the database is curated (modified by a defined group of people) or not. For the sequence databases, the errors may be either in the sequence itself (misprint, wrong on entry, genuine experimental error...) or in the annotation (mistaken features, errors in references,...). In the 3D structure database (PDB), structures have been deposited which were later discovered to contain severe errors. The error handling policy differs considerably between databases. If one bases new experiments or analysis on the data in a particular database, then the implications of its particular error-handling policy need to be considered.



Types of Biological Databases:

There are basically 3 types of biological databases are as follows.

1. **Primary databases :**

- It can also be called an archival database since it archives the experimental results submitted by the scientists. The primary database is populated with experimentally derived data like genome sequence, macromolecular structure, etc. The data entered here remains uncurated(no modifications are performed over the data).
- It obtains 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 database. The same data can later be retrieved using the accession number. Accession number identifies each data uniquely and it never changes.

Examples –

- Examples of Primary database- Nucleic Acid Databases are GenBank and DDBJ
- Protein Databases are PDB,SwissProt,PIR,TrEMBL,Metacyc, etc.

2. Secondary Database :

- The data stored in these types of databases are the analyzed result of the primary database. Computational algorithms are applied to the primary database and meaningful and informative data is stored inside the secondary database.
- The data here are highly curated(processing the data before it is presented in the database). A secondary database is better and contains more valuable knowledge compared to the primary database.

Examples –

Examples of Secondary databases are as follows.

- InterPro (protein families, motifs, and domains)
- UniProt Knowledgebase (sequence and functional information on proteins)

3. Composite Databases :

- The data entered in these types of databases are first compared and then filtered based on desired criteria.
- The initial data are taken from the primary database, and then they are merged together based on certain conditions.
- It helps in searching sequences rapidly. Composite Databases contain non-redundant data.

Examples –

Examples of Composite Databases are as follows.

- Composite Databases -OWL,NRD and Swissport +TREMBL

The different types of databases based on different properties

One may characterize the available biological databases by several different properties. Here is a list to help you think about the various properties a particular database may have.

- Type of data

    - nucleotide sequences
    - protein sequences
    - proteins sequence patterns or motifs
    - macromolecular 3D structure
    - gene expression data
    - metabolic pathways
- Data entry and quality control
    - Scientists (teams) deposit data directly
    - Appointed curators add and update data
    - Are erroneous data removed or marked?
    - Type and degree of error checking
    - Consistency, redundancy, conflicts, updates
- Primary or derived data
    - Primary databases: experimental results directly into database
    - Secondary databases: results of analysis of primary databases
    - Aggregate of many databases
        - Links to other data items
        - Combination of data
        - Consolidation of data
- Technical design
    - Flat-files
    - Relational database (SQL)
    - Object-oriented database (e.g. CORBA, XML)
- Maintainer status
    - Large, public institution (e.g. EMBL, NCBI)
    - Quasi-academic institute (e.g. Swiss Institute of Bioinformatics, TIGR)
    - Academic group or scientist
    - Commercial company
- Availability
    - Publicly available, no restrictions
    - Available, but with copyright
    - Accessible, but not downloadable
    - Academic, but not freely available
    - Proprietary, commercial; possibly free for academics

Accession codes vs identifiers

Many databases in bioinformatics (SWISS-PROT, EMBL, GenBank, Pfam) use a system where an entry can be identified in two different ways. Basically, it has two names:

- Identifier
- Accession code (or number)

The question how to deal with changed, updated and deleted entries in databases is a very tricky problem, and the policies for how accession codes and identifiers are changed or kept constant are not completely consistent between databases or even over time for one single database.

The exact definition of what the identifier and accession code are supposed to denote varies between the different databases, but the basic idea is the following.

Identifier

An **identifier** ("locus" in GenBank, "entry name" in SWISS-PROT) is a string of letters and digits that generally is interpretable in some meaningful way by a human, for instance as a recognizable abbreviation of the full protein or gene name.

SWISS-PROT uses a system where the entry name consists of two parts: the first denotes the protein and the second part denotes the species it is found in. For example, **KRAF_HUMAN** is the entry name for the Raf-1 oncogene from Homo sapiens.

An identifier **can usually change**. For example, the database curators may decide that the identifier for an entry no longer is appropriate. However, this does not happen very often. In fact, it happens so rarely that it's not really a big problem.

Accession code (number)

An **accession code** (or number) is a number (possibly with a few characters in front) that uniquely identifies an entry in its database. For example, the accession code for **KRAF_HUMAN** in SWISS-PROT is **P04049**.

The main conceptual difference from the identifier is that it is supposed to be stable: any given accession code will, as soon as it has been issued, always refer to that entry, or its ancestors. It is often called the **primary key** for the entry. The accession code, once issued, must always point to its entry, even after large changes have been made to the entry. This means that in discussions about specific database entries (e.g. an article about a specific protein), one should always give the accession code for the entry in the relevant database.

In the case where **two entries are merged into one single**, then the new entry will have **both accession codes**, where one will be the **primary** and the other the **secondary** accession code. When an entry is **split into two**, both new entries will get new accession codes, but will also have the old accession code as secondary codes.

Nucleotide sequence databases

Primary nucleotide sequence databases

The databases EMBL, GenBank, and DDBJ are the **three primary nucleotide sequence databases**: They include sequences submitted directly by scientists and genome sequencing group, and sequences taken from literature and patents. There is comparatively little error checking and there is a fair amount of redundancy.

The entries in the EMBL, GenBank and DDBJ databases are **synchronized** on a daily basis, and the accession numbers are managed in a consistent manner between these three centers.

The nucleotide databases have reached such large sizes that they are available in **subdivisions** that allow searches or downloads that are more limited, and hence less time-consuming. For example, GenBank has currently 17 divisions.

There are **no legal restrictions** on the use of the data in these databases. However, there are some patented sequences in the databases.

EMBL www.ebi.ac.uk/embl/

The EMBL (European Molecular Biology Laboratory) nucleotide sequence database is maintained by the European Bioinformatics Institute (EBI) in Hinxton, Cambridge, UK. Its size is given below, in total number of bases, and total number of records. Note its speed of increase since one year. For the current numbers, the EMBL DB statistics page.

Date	# records	# bases
30 Oct 2001	13,771,247	14,745,640,065
16 Oct 2000	9,156,113	10,333,087,560

It can be accessed and searched through the SRS system at EBI, or one can download the entire database as flat files. An example of what an entry looks like is given for the human raf oncogene protein, ID: HSRAFR.

GenBank www.ncbi.nlm.nih.gov/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.

It can be accessed and searched through the Entrez system at NCBI, or one can download the entire database as flat files. An example of what an entry looks like is given for the human raf oncogene protein, ID: HSRAFR.

DDBJ www.ddbj.nig.ac.jp

The DNA Data Bank of Japan began as a collaboration with EMBL and GenBank. It is run by the National Institute of Genetics. One can search for entries by accession number, and little else.

Other nucleotide sequence databases

The following databases contain subsets of the EMBL/GenBank databases. Some also contain more information or links than the primary ones, or have a different organization of the data to better some specific purpose. However, the nucleotide sequences themselves

should always be available in the EMBL/GenBank databases. In this sense, the databases below are secondary databases.

UniGene http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene

The UniGene system attempts to process the GenBank sequence data into a non-redundant set of gene-oriented clusters. Each UniGene cluster contains sequences that represent a unique gene, as well as related information such as the tissue types in which the gene has been expressed and map location.

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

EBI Genomes www.ebi.ac.uk/genomes/

This web site provides access and statistics for the completed genomes, and information about ongoing projects.

Genome Biology www.ncbi.nlm.nih.gov/Genomes/

The Genome Biology site at NCBI contains information about the available complete genomes.

Ensembl www.ensembl.org

Ensembl is a joint project between EMBL-EBI and the Sanger Centre to develop a software system which produces and maintains automatic annotation on eukaryotic genomes.

Protein sequence databases

The two protein sequence databases SWISS-PROT and PIR are different from the nucleotide databases in that they are both **curated**. This means that groups of designated curators (scientists) prepare the entries from literature and/or contacts with external experts.

SWISS-PROT, TrEMBL www.expasy.ch/sprot/

SWISS-PROT is a protein sequence database which strives to provide a **high level of annotations** (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases.

It was started in 1986 by Amos Bairoch in the Department of Medical Biochemistry at the University of Geneva. This database is generally considered one of the best protein sequence databases in terms of the quality of the annotation. Its size is given in the table below.

TrEMBL is a **computer-annotated supplement** of SWISS-PROT that contains all the translations of EMBL nucleotide sequence entries not yet integrated in SWISS-PROT. The procedure that is used to produce it was developed by Rolf Apweiler. The annotation of an entry in TrEMBL has not (yet) reached the standards required for inclusion into SWISS-PROT proper. Its size is given in the table below.

	SWISS-PROT		**TrEMBL**	
Date	**Release**	**# entries**	**Release**	**# entries**
24 Oct 2001	40.1	101,737	18.0	484,388
2 Oct 2000	39.7	88,757	14.17	300,152

SWISS-PROT and TrEMBL are developed by the SWISS-PROT groups at Swiss Institute of Bioinformatics (SIB) and at EBI. The databases can be accessed and searched through the the SRS system at ExPASy, or one can download the entire database as one single flat file. An example of what an entry looks like is given for the human raf oncogene protein, ID KRAF_HUMAN.

The SWISS-PROT database has **some legal restrictions**: the entries themselves are copyrighted, but freely accessible and usable by academic researchers. Commercial companies must buy a license fee from SIB.

PIR pir.georgetown.edu

The Protein Information Resource (PIR) is a division of the National Biomedical Research Foundation (NBRF) in the US. It is involved in a collaboration with the Munich Information Center for Protein Sequences (MIPS) and the Japanese International Protein Sequence Database (JIPID). The PIR-PSD (Protein Sequence Database) release 70.01 (22 Oct 2000) contains 254,293 entries.

PIR grew out of Margaret Dayhoff's work in the middle of the 1960s. It strives to be **comprehensive**, well-organized, accurate, and consistently annotated. However, it is generally believed that it does not reach the level of completeness in the entry annotation as does SWISS-PROT. Although SWISS-PROT and PIR overlap extensively, there are still many sequences which can be found in only one of them.

One can search for entries or do sequence similarity searches at the PIR site. The database can also be downloaded as a set of falt files. An example of what an entry looks like is given for the human raf-1 oncogene protein, ID TVHUF6.

PIR also produces the **NRL-3D**, which is a database of sequences extracted from the three-dimensional structures in the Protein Databank (PDB) (see also the following page in this lecture. The NRL_3D database makes the sequence information in PDB available for

similarity searches and retrieval and provides cross-reference information for use with the other PIR Protein Sequence Databases.

It appears that the PIR web site, and possibly also the underlying database, has improved considerably since one year ago. This means that if one is interested in protein sequences, there is now even more reason to check out PIR; SWISS-PROT is not the only game in town.

Sequence motif databases

Pfam pfam.sanger.ac.uk/, pfam.cgb.ki.se

Pfam is a database of protein families defined as domains (contiguous segments of entire protein sequences). For each domain, it contains a multiple alignment of a set of defining sequences (the seeds) and the other sequences in SWISS-PROT and TrEMBL that can be matched to that alignment.

The database was started in 1996 and is maintained by a consortium of scientists, among them Erik Sonnhammer (CGB, KI, Sweden), Sean Eddy (WashU, St Louis USA), Richard Durbin, Alan Bateman and Ewan Birney (Sanger Centre, UK). Release 6.6 (Sep 2001) contains 3,071 families.

The alignments can be converted into hidden Markov models (HMM), which can be used to search for domains in a query protein sequence. The software HMMER (by Sean Eddy) is the computational foundation for Pfam. The domain structure of protein sequences in SWISS-PROT and TrEMBL are available directly from the Pfam web sites, and it is also possible to search for domains in other sequences using servers at the web sites.

The Pfam database can be searched, or used to identify domains in a sequence, or downloaded from the websites above. An example of an alignment is given for the Raf-like Ras-binding domain (Pfam name RBD, accession code PF02196).

The Pfam database is licensed under the GNU General Public License, which basically makes it available to anyone, but imposes the restriction that derivative works (new databases, modifications) must be made available in source form.

PROSITE www.expasy.ch/prosite/

PROSITE is a database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family (if any) a new sequence belongs.

It was started by Amos Bairoch, is part of SWISS-PROT and is maintained in the same way as SWISS-PROT. The basis of it are regular expressions describing characteristic subsequences of specific protein families or domains. PROSITE has been extended to contain also some profiles, which can be described as probability patterns for specific protein sequence families.

The site above can be used to search by keyword or other text in the entries, to search for a pattern in a sequence, or to search for proteins in SWISS-PROT that match a pattern. An example of a PROSITE regular expression is given for the Ras GTPase-activating proteins signature pattern (RAS_GTPASE_ACTIV_1, accession code PS00509).

Macromolecular 3D structure databases

PDB www.rcsb.org/pdb/

The PDB is the **main primary database for 3D structures** of biological macromolecules determined by X-ray crystallography and NMR. Structural biologists usually deposit their structures in the PDB on publication, and some scientific journals require this before accepting a paper. It also accepts the experimental data used to determine the structures (X-ray structure factors and NMR restraints) and homology models. As of 23 Oct 2001 the PDB contained 16,358 entries, the majority of which (12,304) are X-ray structures.

The Protein Data Bank (PDB) was established in the 1970s at the Brookhaven Lab on Long Island, New York State, US. In 1999, the management was moved to the Research Collaboratory for Structural Bionformatics (RCSB, a joint organisation between Rutgers University, San Diego Supercomputer Center and NIST).

The PDB entries contain the **atomic coordinates**, and some structural parameters connected with the atoms (B-factors, occupancies), or computed from the structures (secondary structure). The PDB entries contain some annotation, but it is not as comprehensive as in SWISS-PROT. Fortunately, there are cross-links between the databases in both file formats. Here is an example of an entry is the the Ras-binding domain of the human Raf-1 oncogene in the traditional PDB format and in the mmCIF format.

There are **no legal restrictions** on the use of the data in the PDB.

SCOP scop.mrc-lmb.cam.ac.uk/scop/

The SCOP (Structural Classification of Proteins) database was started by Alexey Murzin in 1994 (Lab of Molecular Biology, MRC, Cambridge, UK). Its purpose is to **classify protein 3D structures in a hierarchical scheme of structural classes**. It is maintained by experts ("by hand"), and all protein structures in the PDB are classified, and it is updated as new structures are deposited in the PDB.

This is **a typical secondary database**; it is based on data in a primary database (in this case the PDB), but adds information through analysis and/or organisation, in this case the classification of protein 3d structures into a hierarchical scheme of folds, superfamilies and families.

CATH www.cathdb.info

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

levels. Although the aim is very similar to SCOP, the scheme it uses is different, and the philosophy and practical details of producing the classification differ considerably. For instance, a larger fraction of the decisions made when classifying a new protein 3D structure is made automatically by software. It was started by Christine Orengo in Janet Thornton's lab (University College London) in 1996.

Other relevant databases

GeneCards www.genecards.org

GeneCards is a database of human genes, their products and their involvement in diseases. It offers concise information about the functions of all human genes that have an approved symbol, as well as selected others. It is a typical example of a **secondary** database, which contains many links to other databases, and attempts to consolidate the information that is available for a specific class of entity, in this case human genes.

GeneLynx www.genelynx.org

GeneLynx is a database of Web links for human genes. It contains pointers to a large number of other databases. This is also a typical secondary database. It is maintained by Boris Lenhard and Wyeth Wasserman at CGB, KI, Sweden.

KEGG www.genome.ad.jp/kegg/

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is an effort to computerize current knowledge of molecular and cellular biology in terms of the information pathways that consist of interacting molecules or genes and to provide links from the gene catalogs produced by genome sequencing projects.

Amos' WWW links page www.expasy.org/links.html

A page of many links to biological databases and/or web sites (formerly known as Amos' links page).

Systems for searching, indexing and cross-referencing

The usefulness of a database can be increased enormously if it is easy to find entries that satisfy certain search criteria. Some examples of searches that a scientist might want to do:

- All entries with the keyword "GTPase".
- The entries which have a given literature reference (by author or article).
- All proteins with the keyword "ribosomal" from human (organism).

The databases themselves may contain this information, but some software systems must be used to actually perform this kind of search. There are different ways of designing such systems, and two examples are mentioned here.

SRS

The Sequence Retrieval System (SRS) developed by Thure Etzold is a system for **integrating heterogenous databases**. It is based on premade indexes of the items (words, entries, data fields, text,...) found in a set of documents (database files). Apart from the database files themselves, the indexing procedure requires a grammar (Icarus) that describes what different words in the data files mean, how they are to be indexed, and how they cross-reference to other items in other databases. SRS is a web-oriented system located on a server which is accessed through HTML pages and CGI scripts.

SRS started as an academic project, but is now a commercial system which used to be developed and marketed by LION Bioscience AG. Its current status is unknown.

EBI runs an SRS service which can be used by anyone. It indexes a large number of databases, and it also provides a well-defined web interface which allows programs or web sites to create links that query SRS at EBI.

Entrez

The Entrez system developed and accessible at the NCBI Entrez site. Similar to the SRS system, it provides search facilities for a large number of databases, and provides links between them. It provides a well-defined web interface which allows programs or web sites to define links that will query Entrez.

However, it appears that the Entrez system is not available to set up at one's own server. It is purely a system for accessing and searching the databases at NCBI.

Uses of biological Databases:

- It helps the researchers to study the available data and form a new thesis, anti-virus, helpful bacteria, medicines, etc.
- It helps scientists to understand the concepts of biological phenomena.
- The database acts as a storage of information.
- It helps remove the redundancy of data.
- Databases are used to store and organize data in such a way that information can be retrieved easily via a variety of search criteria.
- It allows knowledge discovery, which refers to the identification of connections between pieces of information that were not known when the information was first entered. This facilitates the discovery of new biological insights from raw data.
- Secondary databases have become the molecular biologist’s reference library over the past decade or so, providing a wealth of information on just about any gene or gene product that has been investigated by the research community.
- It helps to solve cases where many users want to access the same entries of data.
- Allows the indexing of data.