Biological Databases & Data Banks

1. Biological databases: why?

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 in an article.
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, which is now part of the UNIPROT 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) and other types of data relating to biological function and processes.

One very important issue is the frequency and type of **errors** that the entries in a database have. 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 needs to use any particular database heavily, then the implications of its particular policy need to be considered.

2. The different types of databases

One may characterize the available biological databases by several different properties:

- 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 (CORBA, XML)
- Maintainer status
  - Large, public institution (EMBL, NCBI)
  - Quasi-academic institute (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

3. 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; essentially 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, **RAF1_HUMAN** is the entry name for the Raf-1 oncogene from Homo sapiens.

An identifier **can change**. For example, the database curators may decide that the identifier for an entry no longer is appropriate. This does not happen very often.

Accession code (number)

An **accession code** (or number) is a number (possibly with a few characters in front) that uniquely identifies an entry. For example, the accession code for **RAF1_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 be possible to find again, even after large changes have been made to the entry.

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.

4. Nucleotide sequence data banks

Primary nucleotide sequence data banks

The data banks ENA (the name name for the EMBL data bank), GenBank, and DDBJ are the **three primary nucleotide sequence data banks (or databases depending on your point of view)**: 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 ENA, GenBank and DDBJ databases are **synchronized** on a regular basis, and the accession numbers are managed in a consistent manner between these three centres.

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 patented sequences in the databases.

ENA www.ebi.ac.uk/ena/

The EMBL (European Molecular Biology Laboratory) nucleotide sequence database is maintained by the European Bioinformatics Institute (EBI) in Hinxton, Cambridge, UK. As of Dec-2024, it contains over 500 billion sequence records with a total of 31 trillion bases; for more up-to-date statistics see the ENA statistics page.

It can be accessed and searched through at the 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: RAF1.

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: RAF1.

DDBJ www.ddbj.nig.ac.jp

The DNA Data Bank of Japan began as a collaboration with EMBL (ENA) and GenBank. It is run by the National Institute of Genetics. Searching facilities are not quite as advanced as EMBL and NCBI's portals.

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 www.ncbi.nlm.nih.gov/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 genome-www.stanford.edu/Saccharomyces/

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.

5. Protein sequence databases

There is now just a single unified protein sequence database called UniProt.

SWISS-PROT, TrEMBL, UniProt **www.uniprot.org**

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. In Dec 2014, it contained just over 547,000 entries.

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 Feb 2015 release contained just over 90,000,000 entries. The annotation of an entry in TrEMBL has not (yet) reached the standards required for inclusion into SWISS-PROT proper, although the SWISS-PROT data bank is now focussed more on complete annotation of model organisms such as Human, mouse, E. coli and so on.

SWISS-PROT and TrEMBL are developed by the SWISS-PROT groups at Swiss Institute of Bioinformatics (SIB) and at EBI. The databases can be accesses 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 RAF1_HUMAN.

At one point, the SWISS-PROT database had some legal restrictions: the entries themselves were copyrighted, but freely accessible and usable by academic researchers. Commercial companies had to buy a license from SIB to use SWISS-PROT. Now SWISS-PROT is licensed under a Creative Commons license, with entries now copyrighted by the publicly-funded UniProt Consortium.

The UniProt Consortium in fact make several related resources available to the public:

UniProtKB/Swiss-Prot (see above)

UniProtKB/TrEMBL (see above)

UniProt Archive (UniParc) is a comprehensive non-redundant database, which contains all the protein sequences from the main, publicly available protein sequence databases. Identical sequences found in different databases are merged, regardless of whether they are from the same or different species, and each sequence is given a stable and unique identifier (UPI), making it possible to identify the same protein from different source databases. Essentially, UniParc aims at being the most comprehensive collection of known protein sequences, but with little or no additional annotation – it just collects the sequences in one location.

Currently UniParc contains protein sequences from the following publicly available databases:

INSDC EMBL-Bank/DDBJ/GenBank nucleotide sequence databases

Ensembl

European Patent Office (EPO)

FlyBase

H-Invitational Database (H-Inv)

International Protein Index (IPI)

Japan Patent Office (JPO)

Protein Information Resource (PIR-PSD)

Protein Data Bank (PDB)

Protein Research Foundation (PRF)

RefSeq

Saccharomyces Genome Database (SGD)

The Arabidopsis Information Resource (TAIR)

TROME

US Patent Office (USPTO)

UniProtKB/Swiss-Prot

UniProtKB/Swiss-Prot protein isoforms

UniProtKB/TrEMBL

Vertebrate and Genome Annotation Database (VEGA)

WormBase

UniRef - UniRef50, UniRef90 and UniRef100

The UniProt Reference Clusters (UniRef) consist of three databases of clustered sets of protein sequences from UniProtKB and selected UniParc records. UniRef100 combines identical sequences and sequence fragments (from any organism) into a single UniRef entry. The sequence of a representative protein, the accession numbers of all the merged entries and links to the corresponding UniProtKB and UniParc records are displayed. UniRef100 sequences are then further clustered according to sequence identity (90% and 50%) to make the less redundant UniRef90 and UniRef50 respectively. Clustering sequences like this can significantly reduce the storage requirements, but more importantly can allow much faster sequence searches.

MGnify

The UniProt Metagenomic and Environmental Sequences (UniMES) database was a repository specifically developed for metagenomic and environmental data. Sequences in this database may not even have a known source organism because they may have been sequenced from samples taken from the soil or from the sea, for example, where the exact mixture of different organisms present in each sample may be unknown. Sequences in this database are often not complete sequences due to the randomness of the sample collection and extraction process. It has now been replaced by the EBI Metagenomics portal MGnify.

6. Sequence motif databases

Pfam http://pfam.xfam.org

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 (CGR, KI, Sweden), Sean Eddy (WashU, St Louis USA), Richard Durbin, Alan Bateman and Ewan Birney (Sanger Centre, UK). Release 37.2 (Feb 2025) contains 24076 families. It is now maintained as part of the InterPro service.

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 http://prosite.expasy.org/

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).

7. Macromolecular 3D structure databases

PDB www.rcsb.org

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 19-Sep-2000 the PDB contained 13,379 entries, the majority of which (10,990) 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 for 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.

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.

SCOP (SCOPe) scop.berkeley.edu

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 regularly electronically updated as new structures are deposited in the PDB under the name SCOPe.

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.

8. Other relevant databases

GeneCards http://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.

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. Although still a very useful resource, KEGG is now a partially pay-walled system, which restricts its use in open-access projects.

9. Systems for searching, indexing and cross-referencing

The usefulness of a database is greatly enhanced when it provides efficient search, indexing, and cross-referencing capabilities. Some common queries that researchers may need to perform include:

- Finding all entries containing the keyword "GTPase."
- Retrieving database entries that reference a specific research article or author.
- Identifying all ribosomal proteins in a specific organism.

Many biological databases have integrated search functions, but dedicated tools have been developed to facilitate complex queries across multiple datasets. Two of the most widely used systems for this purpose are **BioMart** and **Entrez**.

BioMart

BioMart is an open-source, federated database system designed to provide unified access to a variety of biological data resources. It is particularly useful for large-scale queries, such as retrieving gene annotations, protein information, and comparative genomics data.

BioMart is heavily used in the **Ensembl** database and is also integrated into other bioinformatics platforms. Users can perform queries via a web interface or programmatically through APIs, enabling batch retrieval of data. The system allows users to:

- Retrieve gene and protein annotations based on specific criteria (e.g., gene name, function, or chromosomal location).
- Access genome-wide datasets across multiple species.
- Download large datasets in user-defined formats.

Entrez

The **Entrez** system, developed by the **NCBI**, provides powerful search capabilities across multiple databases, including **GenBank, PubMed, UniProt, and Genome databases**. Entrez is widely used for literature searches, sequence retrieval, and cross-referencing data between different biological resources. The platform enables researchers to find relevant information through keyword searches, gene identifiers, and sequence similarity queries.

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.