BIOLOGICAL DATABASES

Current biological databases use all three types of database structures: flat files, relational, and object oriented. Despite the obvious drawbacks of using flat files in database management, many biological databases still use this format. The justification for this is that this system involves minimum amount of database design and the search output can be easily understood by working biologists.

Based on their contents, biological databases can be roughly divided into three categories: primary databases, secondary databases, and specialized databases. *Primary databases* contain original biological data. They are archives of raw sequence or structural data submitted by the scientific community. GenBank and Protein Data Bank (PDB) are examples of primary databases. *Secondary databases* contain computationally processed or manually curated information, based on original information from primary databases. Translated protein sequence databases containing functional annotation belong to this category. Examples are SWISS-Prot and Protein Information Resources (PIR). Specialized databases are those that cater to a particular research interest. For example, Flybase, HIV sequence database, and Ribosomal Database Project are databases that specialize in a particular organism or a particular type of data. Table (2.1).

TABLE 2.1. Major Biological Databases Available Via the World Wide Web

Databases and Retrieval Systems	Brief Summary of Content	URL
AceDB	Genome database for *Caenorhabditis elegans*	www.acedb.org
DDBJ	Primary nucleotide sequence database in Japan	www.ddbj.nig.ac.jp
EMBL	Primary nucleotide sequence database in Europe	www.ebi.ac.uk/embl/index.html
Entrez	NCBI portal for a variety of biological databases	www.ncbi.nlm.nih.gov/gquery/gquery.fcgi
ExPASY	Proteomics database	http://us.expasy.org/
FlyBase	A database of the *Drosophila* genome	http://flybase.bio.indiana.edu/
FSSP	Protein secondary structures	www.bioinfo.biocenter.helsinki.fi:8080/dali/index.html
GenBank	Primary nucleotide sequence database in NCBI	www.ncbi.nlm.nih.gov/Genbank
HIV databases	HIV sequence data and related immunologic information	www.hiv.lanl.gov/content/index
Microarray gene expression database	DNA microarray data and analysis tools	www.ebi.ac.uk/microarray
OMIM	Genetic information of human diseases	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
PIR	Annotated protein sequences	http://pir.georgetown.edu/pirwww/pirhome3.shtml
PubMed	Biomedical literature information	www.ncbi.nlm.nih.gov/PubMed
Ribosomal database project	Ribosomal RNA sequences and phylogenetic trees derived from the sequences	http://rdp.cme.msu.edu/html
SRS	General sequence retrieval system	http://srs6.ebi.ac.uk
SWISS-Prot	Curated protein sequence database	www.ebi.ac.uk/swissprot/access.html
TAIR	Arabidopsis information	www.arabidopsis.org

Primary Databases

There are three major public sequence databases that store raw nucleic acid sequence data produced and submitted by researchers worldwide: GenBank, the European Molecular Biology Laboratory (EMBL) database and the DNA Data Bank of Japan (DDBJ), which are all freely available on the Internet. Most of the data in the databases are contributed directly by authors with a minimal level of annotation. A small number of sequences, especially those published in the 1980s, were entered manually from published literature by database management staff. Presently, sequence submission to either GenBank, EMBL, or DDBJ is a precondition for publication in most scientific journals to ensure the fundamental molecular data to be made freely available. These three public databases closely collaborate and exchange new data daily. They together constitute the International Nucleotide Sequence Database Collaboration. This means that by connecting to any one of the three databases, one should have access to the same nucleotide sequence data. Although the three databases all contain the same sets of raw data, each of the individual databases has a slightly different kind of format to represent the data. Fortunately, for the three-dimensional structures of biological macromolecules, there is only one centralized database, the PDB. This database archives atomic coordinates of macromolecules (both proteins and nucleic acids) determined by x-ray crystallography and NMR. It uses a flat file format to represent protein name, authors, experimental details, secondary structure, cofactors, and atomic coordinates. The web interface of PDB also provides viewing tools for simple image manipulation.

Secondary Databases

Sequence annotation information in the primary database is often minimal. To turn the raw sequence information into more sophisticated biological knowledge, much post processing of the sequence information is needed. This begs the need for secondary databases, which contain computationally processed sequence information derived from the primary databases. The amount of computational processing work varies greatly among the secondary databases; some are simple archives of translated sequence data from identified open reading frames in DNA, whereas others provide additional annotation and information related to higher levels of information regarding structure and functions.

A prominent example of secondary databases is SWISS-PROT, which provides detailed sequence annotation that includes structure, function, and protein family assignment. The sequence data are mainly derived from TrEMBL, a database of translated nucleic acid sequences stored in the EMBL database. The annotation of each entry is carefully curated by human experts and thus is of good quality. The protein annotation includes function, domain structure, catalytic sites, cofactor binding,

Post translational modification, metabolic pathway information, disease association, and similarity with other sequences. Much of this information is obtained from scientific literature and entered by database curators. The annotation provides significant added value to each original sequence record. The data record also provides cross referencing links to other online resources of interest. Other features such as very low redundancy and high level of integration with other primary and secondary databases make SWISS-PROT very popular among biologists.

A recent effort to combine SWISS-PROT, TrEMBL, and PIR led to the creation of the UniProt database, which has larger coverage than any one of the three databases while at the same time maintaining the original SWISS-PROT feature of low redundancy, cross-references, and a high quality of annotation.

Specialized Databases

Specialized databases normally serve a specific research community or focus on a particular organism. The content of these databases may be sequences or other types of information. The sequences in these databases may overlap with a primary database, but may also have new data submitted directly by authors. Because they are often curated by experts in the field, they may have unique organizations and additional annotations associated with the sequences. Many genome databases that are taxonomic specific fall within this category. Examples include Flybase,WormBase, AceDB, and TAIR (Table 2.1). In addition, there are also specialized databases that contain original data derived from functional analysis. For example, GenBank EST database and Microarray Gene Expression Database at the European Bioinformatics Institute (EBI) are some of the gene expression databases available.

INFORMATION RETRIEVAL FROM BIOLOGICAL DATABASES

As mentioned, a major goal in developing databases is to provide efficient and user friendly access to the data stored. There are a number of retrieval systems for biological data. The most popular retrieval systems for biological databases are Entrez and Sequence Retrieval Systems (SRS) that provide access to multiple databases for retrieval of integrated search results. To perform complex queries in a database often requires the use of Boolean operators. This is to join a series of keywords using logical terms such as AND, OR, and NOT to indicate relationships between the keywords used in a search. *AND* means that the search result must contain both words; *OR* means to search for results containing either word or both; *NOT* excludes results containing either one of the words. In addition, one can use parentheses ( ) to define a concept if multiple words and relationships are involved, so that the computer knows which part of the search to execute first. Items contained within parentheses are executed first. Quotes can be used to specify a phrase. Most search engines of public biological databases use some form of this Boolean logic.