C4 MISCELLANEOUS DATABASES

Key Notes

Database resources

There are many types of database available to researchers in the field of biology. These include primary sequence databases for the storage of raw experimental data, secondary databases that contain information on sequence patterns and motifs, and organism-specific databases tailored for researchers working on a particular species. Other miscellaneous databases are discussed in this topic.

Specialized sequence databases

A number of databases have been developed for the storage and analysis of particular types of sequence, for example rRNA and tRNA sequences, introns, promoters and other regulatory elements.

OMIM

OMIM is the Online Mendelian Inheritance in Man database, a powerful resource for the study of human genetics and human molecular biology. Each OMIM entry has a full text summary of information known about a particular gene or trait, with links to primary sequence databases and other human genetics resources.

Incyte and UniGene

Incyte is a commercial database belonging to the LifeSeq® Foundation, providing gene sequences and transcripts with expert annotation. It is designed specifically for drug discovery research. UniGene is an experimental facility for the clustering of GenBank sequences and related to EST data. Currently six vertebrate and five plant species are covered by UniGene.

Structural databases

The primary resource for protein structural data is the PDB, which contains data derived from X-ray crystallography and NMR studies. Another structural database, the MMDB can be accessed at the NCBI web site using Entrez.

Proteins and higher-order functions

Many databases have been set up to store information on particular types of proteins, such as receptors, signal transduction components and enzymes. The compilation of data on different types of proteins, their functions and interactions, makes it possible to deduce higher-order functional networks in the cell, such as biochemical pathways, signal transduction systems and regulatory hierarchies. An example of such a combined database is KEGG.

Literature databases

Literature databases store scientific articles and allow various fields (title, authors, keywords, abstract) to be searched using text strings. Among the most widely used literature resources on the Internet are MEDLINE and PubMED, which cover the scientific literature from the 1960s up to the present day.

Related topics

- Annotated sequence databases (C2)
- Genome and organism-specific databases (C3)
- Data retrieval with Entrez and DBGET/LinkDB (D1)
- Data retrieval with SRS (Sequence Retrieval System) (D2)
- Protein families and pattern databases (F2)

Database resources

Databases are essentially large storage devices for scientific and other data. They can be searched and cross-referenced either over the Internet or using downloaded versions on local computers or computer networks. Specific types of database are discussed in different topics throughout this book. For example, the three primary nucleic acid databases [GenBank, the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database and the DNA Databank of Japan] are discussed in Topic C2. These are called **primary databases** because they store raw sequence data. Similarly, SWISS-PROT and TrEMBL are the major primary databases for the storage of protein sequences. There are also secondary databases of protein families and sequence patterns, such as PROSITE, PRINTS and BLOCKS (Topic F2). These are called **secondary databases** because the sequences they contain are not raw data, but have been derived from the data in the primary databases. There are also many organism-specific databases containing information, links and resources dedicated to particular species (Topic C3). In this topic, we discuss some of the remaining database resources available, which can be grouped under the description **miscellaneous databases**. Note that the journal *Nucleic Acids Research* devotes its first issue every year to articles describing new databases and updates to existing ones. These articles can be accessed online at the following URL: http://www.nar.oupjournals.org/

Specialized sequence databases

The primary sequence databases are unbiased as to the type of sequence data they contain. However, a number of more specialized databases have been developed with particular types of nucleic acid or protein sequence in mind. For example, there are databases specifically for rRNA (ribosomal RNA) and tRNA (transfer RNA) sequences, for example the database of 5S rRNA sequences (http://biobases.ibch.poznan.pl/5SData/) and the database of small subunit rRNA sequences (http://rrna.uia.ac.be/ssu/). Further examples include databases for promoter sequences and other transcriptional regulatory elements, databases for regulatory elements in the noncoding region of mRNAs (messenger RNAs) and InBase, a database of inteins, which are small peptides that are spliced out of some microbial proteins (http://www.neb.com/neb/inteins.html).

OMIM

OMIM (Online Mendelian Inheritance in Man) is a comprehensive database of human genes and genetic disorders maintained by the National Center for Biotechnology Information (NCBI) and can be accessed at the following URL: http://www.ncbi.nlm.nih.gov/omim or through Entrez (Topic D1). Each OMIM entry has a full text summary of a gene or genetically determined phenotype and has numerous links to other databases such as the primary sequence databases, SWISS-PROT, PubMed references, general and locus-specific mutation databases, gene nomenclature databases and mapviewer. OMIM is an excellent starting point to find information on human genetics. An example of an OMIM file (*Fig. 1*) refers to the same gene/protein represented by *Figs 1* and *3* in Topic C2.

Incyte and UniGene

Incyte is an example of a **commercial database**. Unlike the **public databases** discussed above, which can be accessed freely by anyone using the WWW, commercial databases require subscription, as they are often the result of a single company's research and investment. **Incyte** is an integrated database of DNA sequences, transcripts, extensive annotations, expression data and access

```
1: *602902
BASIC TRANSCRIPTION ELEMENT-BINDING PROTEIN 1; BTEB1

Alternative titles; symbols
BTEB

Gene map locus 9q13

TEXT
The GC box is a common regulatory DNA element of eukaryotic genes. The
promoter region of rat CYP1A1 (108330) contains a single GC box within a
basic transcriptional element (BTE) required for constitutive expression of
the gene. By screening a liver library for the ability to bind BTE, Imataka
et al. (1992) isolated rat cDNAs encoding Sp1 (189906) and a protein that
they designated BTEB(BTE binding protein). Sequence analysis revealed that,
like Sp1, BTEB contains 3 consecutive zinc finger motifs. In transient
transfection experiments both BTEB and Sp1 stimulated promoters with
repeated GC boxes. However, the CYP1A1 promoter with only 1 GC box was
activated by Sp1 and repressed by BTEB. Ohe et al. (1993) used a rat BTEB
cDNA to screen a human placenta library and isolated cDNAs encoding BTEB1.
The sequences of the predicted 244-amino acid rat and human proteins are 98%
identical. Imataka et al. (1992) and Ohe et al. (1993) found that the mRNAs
encoding BTEB and BTEB1 contain a GC-rich leader sequence in the 5-prime
untranslated region that has the potential to form stem-loop structures and
that may control translation.

By analysis of a somatic cell hybrid panel and by fluorescence in situ
hybridization, Ohe et al. (1993) mapped the BTEB1 gene to 9q13.


REFERENCES
1. Imataka, H.; Sogawa, K.; Yasumoto, K.; Kikuchi, Y.; Sasano, K.;
Kobayashi, A.; Hayami, M.; Fujii-Kuriyama, Y. : Two regulatory proteins that
bind to the basic transcription element (BTE), a GC box sequence in the
promoter region of the rat P-4501A1 gene. EMBO J. 11: 3663-3671, 1992.
PubMed ID : 1356762

2. Ohe, N.; Yamasaki, Y.; Sogawa, K.; Inazawa, J.; Ariyama, T.; Oshimura,
M.; Fujii-Kuriyama, Y. : Chromosomal localization and cDNA sequence of human
BTEB, a GC box binding protein. Somat. Cell Molec. Genet. 19: 499-503, 1993.
PubMed ID : 8291025


CREATION DATE
Rebekah S. Rasooly : 7/29/1998

EDIT HISTORY
alopez : 7/29/1998


Copyright (c) 2000 Johns Hopkins University
```

Fig. 1. The OMIM file for the human gene BTEB.

to cDNA (copy DNA) clones for experimental studies. It is the property of the LifeSeq® Foundation and subscription information can be found at the following URL: http://www.incyte.com.

UniGene is another resource for genome research. In the words of its developers: 'UniGene is an experimental system for automatically partitioning GenBank sequences 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 its map location.' UniGene incorporates about 10^5 expressed sequence tags (ESTs; Topic B1) and is used by experimenters to design probes and reagents for gene mapping and expression analysis. The organisms included in UniGene were chosen on the basis of the availability of large amounts of EST data and to give a reasonable coverage of the vertebrate and plant kingdoms with examples of closely and distantly related species. These include human, mouse, cow, rat, zebrafish, *Xenopus*, wheat, rice, barley, maize and *Arabidopsis*.

Structural databases

Structural databases store data on protein (and nucleic acid) structure. The primary resource for protein structure data is the **Protein Data Bank (PDB)** available at the following URL: http://www.pdb.org/. This is the single worldwide archive of structural data and is maintained by the **Research Collaboratory for Structural Bioinformatics (RCSB)**, at Rutgers University. The associated **Nucleic Acid Data Bank (NDB)** is also maintained there. Data from both X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy studies (Topic B2) can be deposited in the PDB using a web-based interface called the **AutoDep Input Tool (ADIT)**. The data are extensively checked and verified by human curators before acceptance. An equivalent European database is the **Macromolecular Structure Database (MSD)** maintained by the European Bioinformatics Institute. The RSCB and MSD databases contain the same data.

There are also other structural databases such as Entrez's **Molecular Modeling Database (MMDB)** which aims to provide information on sequence and structure neighbors, links between the scientific literature and 3D structures, and sequence and structure visualization.

Proteins and higher-order functions

One important aim of bioinformatics is to use biological data to understand the higher-level functions of the cell, that is, biochemical pathways, regulatory networks, signal transduction pathways, and how these influence cell and organism behavior. A number of databases have been established with this goal in mind. Several databases have been designed, for example, to provide information on the functional annotation of proteins. Such databases include **PIR (Protein Information Resource)**, which can be accessed at http://pir.georgetown.edu/. Another valuable resource is the **Kyoto Encyclopedia of Genes and Genomes (KEGG)**, which is the primary resource of the Japanese GenomeNet service. KEGG is available at the following URL: http://www.genome.ad.jp/kegg/. The main database integrates a number of subsidiaries including **PATHWAY** (which stores data on molecular pathways and complexes), **GENES** (which stores functional information about genes and their products) and **LIGAND** (which stores data about chemical compounds and reactions occurring in the cell). Together, these data can be used for functional annotation and the grouping of genes and proteins into common pathways, networks and hierarchies.

There are also many databases that provide information on specific aspects of protein function. For example, **DIP (the Database of Interacting Proteins)** and **BIND (Biomolecular Interaction Network Database)** provide functional annotations of proteins on the basis of their interactions with each other and with other ligands. There are also databases for particular types of protein, for example ReBase, a database of restriction endonucleases and their target sites;

TRANSFAC, a database of transcription factors; Sentra, a database of signal transduction proteins; and NUREBASE, a database of nuclear receptors.

Literature databases

A literature database contains the abstracts and, in some cases, the full text and figures of published scientific articles. Such databases can be searched using text strings to find words in the title, abstract, keywords or main text, or by author or author's institution. One of the earliest comprehensive online library resources was **Medline**, which has been incorporated into a large resource called **PubMed** maintained by the NCBI. They are integrated into the Entrez suite of databases described in Topic D1. Other such resources include the **Web of Science**, which requires institutional subscription, and **BioMedNet** (http://www.bmn.com), which provides access to thousands of review articles published in the popular *Trends* and *Current Opinion* journals. One of the best features of BioMedNet is that reviews can be downloaded as .pdf files and stored on the computer like a personal library.