

# Databases of biological information

The new wealth of biological data generated by ongoing genome projects is being used to develop database tools for biologists. This basic biological information can then be interpreted from many viewpoints – from molecular interactions to interactions among organisms.

Bioinformatics has emerged as a new branch of biology as a result of the advancement of experimental technologies in molecular biology, which generate a vast amount of data, as exemplified by **high-throughput DNA sequencing** technology. Consequently, a primary role of bioinformatics has been to organize and manage the data and to assist experimental projects. However, the next major task of bioinformatics is to make biological sense out of the data. At the moment, the data are largely sequence and 3-D structural information but, in the near future, other types of data will become available, such as gene expression profiles generated by **DNA chip technologies**. There are two aspects of bioinformatics that are relevant to data interpretation on a massive scale: developing new algorithms and software, and computerizing data and knowledge. For example, although it is important to develop rapid and sensitive methods for database searching using sequences, interpretation of the results would not be possible without the reference biological information associated with the database sequences. In fact, the fraction of unknown sequences that match other unknown sequences is increasing as more complete genome sequences are stored in the databases. However, the increasing amount of data does not necessarily mean an increasing amount of knowledge in biology. This is where more effort needs to be exerted to organize databases of biological information, as well as to design new systematic experiments on biological functions.

## Sequence–function relationships

In contrast to the well-established databases of literature, sequences and 3-D coordinates, databases of biological functions are still in an early stage of development. There are many variations on how to represent and utilize biological knowledge in computerized databases. For example, some databases concentrate on specific molecules or specific functions and provide highly detailed information, while others try to cover a broad range of biology with less detailed information. Some databases are designed for humans to read and interpret, while

### Minoru Kanehisa

Institute for Chemical Research,  
Kyoto University, Uji,  
Kyoto 611-0011, Japan.

kanehisa@kuicr.kyoto-u.ac.jp  
<http://kanehisa.kuicr.kyoto-u.ac.jp/>

others aim at computerized reasoning in the form of a **deductive database**. In some cases, the biological information is generated by computer analysis of other databases; in other cases, it is obtained from the literature by manual means. A list of selected databases that attempt to

organize biological information from different perspectives is provided in the URLs box.

Functional data that relate to sequence information are stored in the features tables of the sequence databases. However, the information is not well standardized because the annotations are made by different authors. A notable exception is **SWISS-PROT**, which is the best-curated protein sequence database. PROSITE is derived from SWISS-PROT and contains conserved sequence patterns, or **sequence motifs**, associated with specific functions. There are other motif libraries, notably BLOCKS and PRINTS, as well as libraries containing longer sequence patterns of protein domain structures, such as Pfam and ProDom (see the article by Kay Hofmann on pp. 18–21). One of the main differences is how the sequence information is represented, namely in **regular expression** of text patterns, **multiple alignments, profiles** or hidden Markov models (**HMMs**). The motif libraries are like dictionaries of foreign languages in which the meanings of words (subsequences) are written primarily for humans to understand.

## Classification of molecules

The local similarities of sequences, as represented in sequence motifs, reflect specific functions of proteins. By contrast, a broader functional hierarchy can be constructed by the global similarity of protein sequences and by the global similarity of 3-D protein structures. The URLs box shows some of the representative databases: PIR contains a **superfamily** classification of protein sequences, and SCOP and CATH are hierarchical classifications of protein 3-D structures. It is often the case that two proteins without apparent sequence similarity share a common 3-D **fold**, thus representing the same functional category. This type of information can be obtained, for example, from SCOP, in which

proteins are hierarchically classified into four levels: classes, folds, superfamilies and families.

In contrast to the general-purpose databases that classify the entire set of protein sequences or protein 3-D structures, there are numerous databases that specialize in specific categories of molecules. The URLs box shows only a few of these excellent resources: TRANSFAC for transcription factors, GCRDb for G-protein-coupled receptors, and PKR for protein kinases. As most of these specialized databases are developed and maintained by individual researchers or individual research groups, the quality and quantity vary significantly. Proweb provides a good guide to the sites of specific protein families that are currently available<sup>1</sup>.

### Wiring diagrams of molecules

The databases mentioned above contain biological information about individual molecules that can be considered the building blocks of life. However, an intricate network of interacting molecules forms the living cell. Without knowledge of such interactions, or wiring diagrams of molecules, an understanding of cellular functions would not be possible.

The databases of **biochemical pathways** in the URLs box contain information about networks of interactions, especially for metabolic pathways – KEGG and WIT are general-purpose databases for many species, EcoCyc is *Escherichia coli* specific, and UM-BBD is a specialized database for secondary biodegradation pathways in bacteria. LIGAND is tightly coupled with KEGG and contains information about enzymatic reactions and chemical compounds. A reaction or an interaction represents a unit of wiring, but interaction databases, other than those for enzymatic reactions, are not yet well developed.

Pathway databases are becoming increasingly important in order systematically to process the data generated by genome projects<sup>2-4</sup>. In KEGG, for example, the catalog of genes can be mapped onto the reference pathways by sequence similarity to check whether functional units are correctly formed. In the same way that the sequence databases are rapidly expanding as a result of whole genome sequencing of a number of organisms, the pathway databases are bound to expand following systematic experiments in **functional genomics** (see Box 1).

### Comparison and classification of organisms

The availability of complete genome sequences has made it possible to perform comparative analyses of entire complements of genes from different species<sup>5</sup>. The results of the analyses are represented as groups of **orthologous genes**, which are stored in databases such as COG and MBGD (see the URLs box), as well as in the ortholog group tables of KEGG. Because the data in these databases reflect multiple alignments of genes, sequence similarity searching against them is more reliable than searching against primary sequence databases. However, the number of

## Box 1. How to reconstruct an organism

An ultimate form of reductionism in molecular biology is the genome project, which identifies the complete set of genes by experimental technologies (see Fig. 1). This poses a new grand-challenge problem in bioinformatics<sup>6</sup>. Given a complete genomic sequence, is it possible to reconstruct a functioning system of the biological organism? The answer would be 'no' if only the complete genome sequence is provided. However, the prospects become brighter if, in addition, systematic experimental data are available for protein interactions (identified by yeast two-hybrid systems) and gene interactions (identified by expression profiling after gene disruption or overexpression). Thus, we are given the complete set of genes with sequence information, incomplete sets of protein and gene interactions, and incomplete reference knowledge of molecular wiring diagrams stored in databases such as KEGG. A general protocol of reconstruction would be first to make use of the reference wiring diagrams and build up partial subsystems, and then to fill in the gaps and obtain an overall picture of the system. Success will largely depend on the quality and completeness of the reference database, which should contain all relevant experiments in genetics, biochemistry and molecular and cellular biology, as well as the quality and quantity of the interaction experiments. Bioinformatics will play a key role in developing the reference database, designing systematic experiments and processing data to predict molecular wiring.

gene families that can be identified in this way is not very large at present.

The classification of organisms on the basis of morphological and molecular evidence is a difficult and controversial problem. The URLs box lists two sites containing such information – the NCBI Taxonomy database stores the classification of organisms adopted in the GenBank database and the Tree of Life summarizes phylogenetic relationships and the characteristics of organisms. In view of the expanding knowledge gathered by **comparative genomics**, the databases of organism classifications should benefit considerably.

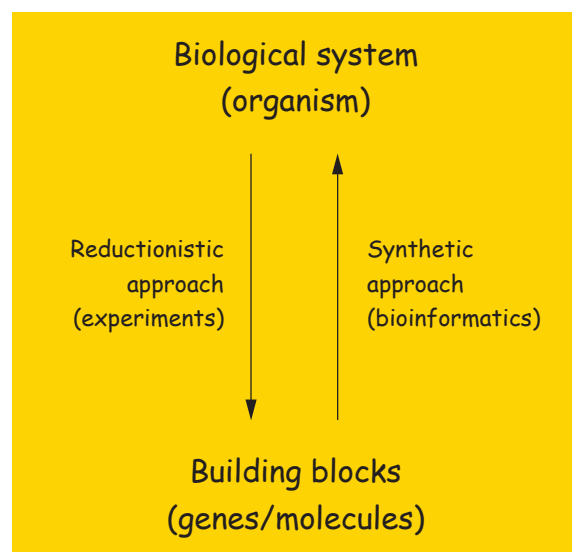


Fig. 1. The reductionistic and synthetic approaches in biology.



URLS...

### Sequence annotations

#### SWISS-PROT

<http://expasy.hcuge.ch/sprot/>

### Sequence motifs

#### PROSITE

<http://expasy.hcuge.ch/sprot/prosite.html>

#### Blocks

<http://www.blocks.fhcrc.org/>

#### PRINTS

<http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/>

#### Pfam

<http://www.sanger.ac.uk/Pfam/>

<http://pfam.wustl.edu/>

#### ProDom

<http://protein.toulouse.inra.fr/prodom.html>

### Sequence classifications

#### PIR Superfamily

<http://www-nbrf.georgetown.edu/pir/>

### 3-D fold classifications

#### SCOP

<http://scop.mrc-lmb.cam.ac.uk/scop/>

#### CATH

<http://www.biochem.ucl.ac.uk/bsm/cath/>

### Specific molecules

#### TRANSFAC

<http://transfac.gbf-braunschweig.de/TRANSFAC/>

#### GCRDb

<http://www.gcrdb.uthscsa.edu/>

#### PKR

<http://www.sdsc.edu/kinases/>

### Proweb

<http://www.proweb.org/>

### Biochemical pathways

#### KEGG

<http://www.genome.ad.jp/kegg/>

#### LIGAND

<http://www.genome.ad.jp/dbget/ligand.html>

#### WIT

<http://www.cme.msu.edu/WIT/>

#### EcoCyc

<http://ecocyc.PangeaSystems.com/ecocyc/>

#### UM-BBD

<http://www.labmed.umn.edu/umbbd/>

### Orthologous genes

#### COG

<http://www.ncbi.nlm.nih.gov/COG/>

#### MBGD

<http://mbgd.genome.ad.jp/>

### Organism classifications

#### NCBI Taxonomy

<http://www.ncbi.nlm.nih.gov/Taxonomy/>

#### Tree of Life

<http://phylogeny.arizona.edu/tree/phylogeny.html>

### Medical resources

#### OMIM

<http://www.ncbi.nlm.nih.gov/Omim/>

#### HGMD

<http://www.uwcm.ac.uk/uwcm/mg/hgmd0.html>

#### Merck Manual

<http://www.merck.com/pubs/mmanual/html/sectoc.htm>

### Future directions

Bioinformatics has become an essential tool not only for basic research but also for applied research in biomedical sciences. The URLs box shows selected medical databases: OMIM provides a catalog of human genes and genetic disorders; HGMD provides mutation data of human disease genes; and the Merck Manual provides a comprehensive guide to clinical information. In general, medical databases are like textbooks for humans to search and read, and they are not yet well integrated in the web of existing molecular biology databases. However, the situation will change as genome sequencing identifies more disease genes, and **oligonucleotide chips** accumulate their mutation patterns.

Bioinformatics was born from the marriage between biological science in the age of massive-scale

sequence data production and computer science. Although the marriage has been successful, advances in genomics will begin to generate new types of data on interactions and mutations. At the same time, there is a huge body of knowledge on molecular interactions and pathways that exists only in the literature or in the minds of experts. Significant progress is therefore required in the development of databases of biological information.

### References

- 1 Henikoff, S. *et al.* (1997) *Science* 278, 609–614
- 2 Kanehisa, M. (1997) *Trends Genet.* 13, 375–376
- 3 Karp, P.D. (1998) *Trends Biochem. Sci.* 23, 114–116
- 4 Galperin, M.Y. and Brenner, S.E. (1998) *Trends Genet.* 14, 332–333
- 5 Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) *Science* 278, 631–637
- 6 Kanehisa, M. (1998) *Bioinformatics* 14, 309

## Trends Guide to Bioinformatics

### Extra Copy Sales

This special supplement from *Elsevier Trends Journals* provides an essential introduction to the cutting-edge field of bioinformatics. If you require extra copies of the *Trends Guide to Bioinformatics*, either for teaching or distribution, please contact:

Thelma Reid (t.reid@elsevier.co.uk)  
Elsevier Trends Journals, 68 Hills Road, Cambridge, UK CB2 1LA.  
Tel: +44 1223 311114 • Fax: +44 1223 321410