## Bioinformatics

### Value-added databases

Databases provide the necessary organizational infrastructure for bioinformatics. Generally, databases can be considered either primary or secondary. Primary databases essentially contain experimental data that are usually compiled automatically with little or no curation, interpretation and annotation. The nucleotide sequence databases GenBank (http://www.ncbi.nlm.nih.gov), EMBL (http://www.ebi.ac.uk) and DDBJ (http://www.ddbj.nig.ac.jp) are all well known primary databases. These primary databases provide not only a repository and archive for all sequence data but also the essential raw data on which other analyses and inferences are made.

The primary databases are concerned with capturing and organizing, rather than interpreting, sequence data. Consequently, this has lead to the development of secondary databases that interpret primary data and 'add value' to the primary databases. The protein sequence databases Swiss Prot (http://www.expasy.ch – for which commercial users now have to pay a license to access) and Protein Information Resource (PIR; http://www-nbrf.georgetown.edu/pir/) are well known secondary databases.

There are also more-specialized sequence and non-sequence secondary databases being developed. This is an attempt to make sense of the mountain of data, to put data into some sort of context, to integrate and relate data from different sources, to validate existing data and to make data more useful to the researcher. This specialized collection of sequence and non-sequence information compiled by experts, makes it more of an 'information base' or 'knowledge base' rather than a 'data base'.

Examples of specialized sequence and non-sequence 'knowledge bases' are given in Box 1. The 'value' in these databases includes not only their trustworthy information content but also their presentation. Many of these databases are easily accessible (via the Internet), user-friendly, relatively easy to navigate (through hypertext links), highly interactive (using Java applets) and graphical.

### New knowledge bases

Within the past year some of the 'knowledge bases' developed include the following:

*ACUTS (http://pbil.univ-lyon1.fr/acuts/ACUTS.html)*

The database of Ancient Conserved UnTranslated Sequences (ACUTS) seeks to provide information on new regulatory elements in untranslated regions of protein-coding genes (5′flanks, 5′UTRs, introns, 3′UTRs and 3′flanks) from vertebrates, insects and nematodes. It is compiled by Laurent Duret and colleagues at Laboratoire de Biométrie, Génétique et Biologie des Populations (Lyon, France).

*GeneMap98 (http://www.ncbi.nlm.nih.gov/genemap98/)*

GeneMap98 is a database map of 30,181 human genes generated from the Human Genome Project. Its focus is to help increase understanding of the genetic basis of disease and to provide a framework for sequencing projects by highlighting key markers (gene-rich regions) of the chromosomes. It is the result of collaborative efforts of an international consortium of scientists and is maintained at the National Center for Biotechnology Information (NCBI), Bethesda, MD, USA.

*HGBase (http://hgbase.interactiva.de/)*

HGBase is a new polymorphism database that lists human intra-genic (promoter to transcription end point) DNA sequence polymorphisms. HGBase is essentially a catalogue of intra-genic sequence variants found in 'normal' individuals. HGBase contains all types of gene-based variation including single nucleotide polymorphisms (SNPs), functionally consequential polymorphisms (promoter and non-silent codon changes) and other polymorphisms (such as intron sequence differences). It has been constructed by the Department of Genetics and Pathology at Uppsala University (Sweden) and Interactiva Biotechnologie (Ulm, Germany).

*mitoDat database (http://www lecb.ncifcrf.gov/mitoDat/>)*

The mitoDat database collects information on the genes and the expressed proteins of human and mouse mitochondria and is a subset of the larger Reference Gene database (RefGene) at the NCBI. It is being compiled by Steven J. Zullo at the Laboratory of Biochemical Genetics National Institutes of Health (Bethesda, MD, USA).

*Protein Kinase Resource (http://www.sdsc.edu/Kinases/pkr/)*

The Protein Kinase Resource (PKR) database integrates molecular and cellular information on the protein kinase family of enzymes. The PKR is a collaborative project involving many protein kinase researchers and computational biologists and is maintained at the San Diego Supercomputer Center, CA, USA with funding from the National Science Foundation.

Pseudomonas aeruginosa *genes (http://www.bit.uq.edu.au/pseudomonas/)*

This database provides information on ~3800 genes in *Pseudomonas aeruginosa* PA01 that have been identified as having similarity to known genes. This interactive site is maintained by the Centre of Molecular and Cellular Biochemistry at the University of Queensland, Australia.


1359-6446/99/$ – see front matter  PII: S1359-6446(98)01285-9 

*Metalloprotein Database and Browser (MDB) (http://metallo.scripps.edu/index.html)*

The MDB contains quantitative information, derived from three dimensional structures (X-ray or NMR), on all the metal-containing sites in metalloproteins. The Web site provides the basic dataset and includes tools for geometrical and functional queries of the database. Currently, the dataset comprises 6261 metal sites from 2246 X-ray and NMR structures.

*Protein Function and Metabolic Pathways (PFMP) database*

Currently under development at the EBI this database is something to watch out for when it eventually comes online. It will be an object-oriented database of information on protein function and biochemical pathways. In particular, it hopes to relate genome and protein sequence data to existing knowledge of protein function and metabolic pathways.

*dbSNP (http://www.ncbi.nlm.nih.gov/SNP/)*

This is a database of single nucleotide polymorphisms (SNPs). SNPs are single nucleotide variations in the genome and are becoming increasingly popular as a way to help guide the sequencing efforts of the human genome and the discovery of genes involved in major diseases. The database is the result of a collaborative effort between the National Human Genome Research Institute and The National Center for Biotechnology Information.

## Databases of databases

More databases are being developed and existing databases are being revised. Consequently, it is important, although difficult, to remain up-to-date. Fortunately, there are 'databases of databases' that can be used to stay abreast of database progress and development. A good general list of databases and other molecular biology resources can be found at the well known (though a little out-of-date) 'Pedro's Biomolecular Research Tools' Web site (http://www.public.iastate.edu/~pedro/research_tools.html). The following sites also provide recent lists of databases:

- Australian National University Bioinformatics (http://life.anu.edu.au/)
- Hopkins Bioinformatics Home Page (http://www.gdb.org/hopkins.html)
- UK MRC HGMP Resource Centre (http://www.hgmp.mrc.ac.uk/)
- Harvard Biological Laboratories – Genome Research (http://golgi.harvard.edu/)
- Genome Databases List (gopher://genome-gopher.stanford.edu/1/topic/genome_db)
- Listing of Molecular Biology Databases – LIMB (gopher:// gopher.nih.gov/11/molbio/other)

Additionally, the January edition of *Nucleic Acids Research* annually publishes an extensive list and full articles of databases. It is envisaged that from

**Box 1. Knowledge bases**

**Specialized sequence 'knowledge bases'**

Genomes
- Yeast Protein Database (http://quest7.protease.com/YPDhome.html)
- EcoCyc (http://ecocyc.panbio.com/ecocyc/)
- FlyBase (http://flybase.bio.indiana.edu/)
- Caenorhabditiselegans (http://www.sanger.ac.uk/worm/C.elegans_Home.html)

Protein families
- ProWeb (http://howard.fhcrc.org/kinesin/ProWeb.html)
- SCOP (http://www.bio.cam.ac.uk/scop/)

Protein motifs and patterns
- Prosite (http://expasy.hcuge.ch/sprot/prosite.html)
- Prints (http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/)

Protein three-dimensional structure
- Protein Data Bank – PDB (http://pdb.pdb.bnl.gov)

**Specialized non-sequence 'knowledge bases'**

Genetics
- OMIM (http://www.ncbi.nlm.nih.gov/Omim)

Proteomes
- Swiss 2D (http://www.expasy.ch/ch2d/)

Enzymes
- Enzyme (http://www.expasy.ch/sprot/enzyme.html)

Metabolic pathways
- KEGG – Kyoto Encyclopedia of Genes and Genomes (http://www.genome.ad.jp/kegg/kegg2.html)
- PUMA – Phylogeny Metabolism Alignments (http://www.mcs.anl.gov/home/compbio/PUMA/Production/puma_graphics.html)


1359-6446/99/$ – see front matter  PII: S1359-6446(98)01286-0 

the beginning of 1999 this list will be available on-line as a Web-based table. *Trends in Biochemical Sciences* and other popular 'Trends' journals also publish articles on new and revised databases as they arise.

*Steve Bottomley*
*School of Biomedical Sciences*
*Curtin University of Technology*
*Perth, Western Australia*
*e-mail: ibottoml@info.curtin.edu.au*

## About the *Profiles* section in *Monitor*...

*Profiles* gives a commentary on promising lines of research, new technologies and progress in therapeutic areas. We welcome offers of contributions for this series. Articles should provide an accurate summary of the essential facts to givie a perspective; brief outlines of proposed articles should be addressed to the *Monitor* Editor.

Articles for publication in *Monitor* are subject to peer review and occasionally may be rejected or, as is more often the case, authors may be asked to revise their contribution. The *Monitor* Editor also reserves the right to edit articles after acceptance.

Monitor Editor: Andrew Lloyd,
School of Pharmacy and
Biomolecular Sciences,
University of Brighton,
Cockcroft Building,
Moulsecoomb, Brighton,
UK BN2 4GJ.
tel: +44 1273 642049,
fax: +44 1273 679333,
e-mail: a.w.lloyd@brighton.ac.uk

## Erratum

In the December issue, the review article entitled *Antimalarial drug discovery: development of inhibitors of dihydrofolate reductase active in drug resistance* by David C. Warhurst [*Drug Discovery Today* (1988) 3, 538–546] showed Fig. 1 reproduced in black and white. This was incorrect and should have appeared as below. We apologise to the author and the readers of this journal for this error.

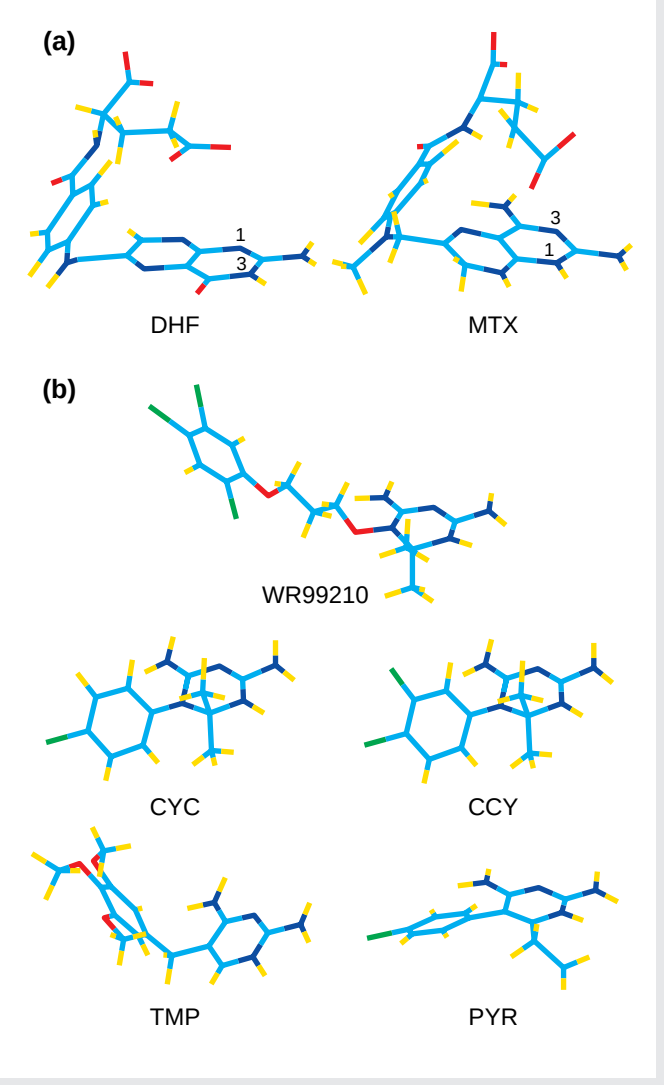


***Figure 1.*** *(a) Active-site conformations of dihydrofolate (DHF) and of methotrexate (MTX). (b) Optimized structures of dihydrotriazines WR99210, cycloguanil (CYC), chlorcycloguanil (CCY) and of pyrimidines trimethoprim (TMP) and pyrimethamine (PYR). Colours show: carbon, light blue; nitrogen, dark blue; hydrogen, yellow; oxygen, red; chlorine, green.*