UDC 004.738.5:004.65:577.1
ISSN 1330-9862

review

(FTB-1226)

Macromolecular Databases – A Background of Bioinformatics

Dubravko Jelić[1], Tibor Toth[2] and Donatella Verbanac[1]*

[1]High-Throughput Screening Unit, PLIVA d.d. Research Division, Applied Research, Prilaz baruna Filipovića 25, HR-10 000 Zagreb, Croatia

[2]Research Information Center, PLIVA d.d. Research Division, Applied Research, Prilaz baruna Filipovića 25, HR-10 000 Zagreb, Croatia

Key words: macromolecular databases, proteins, nucleic acids, bioinformatics, internet

CONTENTS

* Corresponding author; Phone: +385 1 3721 518; Fax: ++385 1 3721 570; E-mail: Dubravko.Jelic@pliva.hr

A. INTRODUCTION

During the last couple of years, informatics and information systems on the Internet and elsewhere have been enormously developed and improved. But, although there are many search options, it is still sometimes very difficult to find information and data needed. We realize that the Internet is a huge world-depository of knowledge from all fields of science, technology and medicine and that it is impossible to comprehend all databases from this very wide area, but we decided to try to make an overview of the biggest, most important and most available bioinformatic databases about nucleic acids and proteins.

When we consider macromolecular databases, or sources of bioinformatic knowledge about proteins and nucleic acids, inside this area are databases of nucleotide and protein sequences, three-dimensional (3D) structure databases, bibliographic databases, biological or different specialized databases and tools. There are many online resources listing databases valuable for different research areas, such as »The Molecular Biology Database Collection: 2002 update« on http://nar.oupjournals.org/content/vol30/issue1/, where key databases of value to the biological community are listed [A. D. Baxevanis, *Nucl. Acids Res. 30* (2002) 1-12], or »Database resources of the NCBI: 2002 update« on http://www.ncbi.nlm.nih.gov/Database/index.html, where data analysis and retrieval resources that operate on the data in GenBank and a variety of other biological data are provided [D. L. Wheeler *et al., Nucl. Acids Res. 30*(1) (2002) 13-16]. Within Swiss Institute for Bioinformatics (SIB) there is ExPaSy proteomic server (http://www.expasy.org/), and this is a powerful data source for all kinds of protein databases, software and tools. Amos' WWW page with links on ExPaSy proteomic server contains almost exclusively pointers to information sources for life scientists with an interest in biological macromolecules. List of links to protein sequence, 3D structure and 2D-gel analytical tools are provided. There are more than **1 000** links within this page [http://us.expasy.org/alinks.html; on 16th June, 2003]. On ExPaSy there is also BioHunt Molecular Biology Finder [http://us.expasy.org/BioHunt/]. Furthemore, SRS (*Sequence Retrieval System;* http://srs.ddbj.nig.ac.jp/index-e.html) is a server enabled by LION Bioscience. This is a network browser for Databanks in Molecular Biology, powered by DDBJ (Japan), where many sequence, structure and application databases are listed. BiolRes also gives a list of biological and chemical datasources available at [http://psyche.uthct.edu/ous/BiolRes.html]. Within these sources, links are given to different databases specific to some particular areas of research.

Macromolecular databases can be divided in two basic groups following macromolecular architecture: in the primary structure databases, which comprise primary structures – nucleotide and amino acid sequences and in the secondary, tertiary and quaternary structure databases, which comprehend information and knowledge about interactions of amino acid residues in proteins or nucleotide residues within nucleic acid, as well as chain and domain folding for different nucleic acids and proteins. Nucleotides or amino acids can be close to

each other in linear sequence (α-helixes, β-sheets, loops, *etc.*), but very often they can be far apart in the linear sequence and close to each other in three-dimensional view, with important interaction between them (disulphide, electrostatic, hydrogen bonds, *etc.*), possibly crucial for activity. Also, we cannot forget proteins, which have more than one polypeptide chain (subunit). Between subunits there are quaternary interactions, as well as between protein domains or different proteins. Both of these groups are important and researchers should not be focused on sequence databases only, because specific three-dimensional binding and transmission of structural changes are the basis of most biochemical and metabolic pathways and actions.

B. PRIMARY STRUCTURE DATABASES

B.1. NUCLEOTIDE SEQUENCE DATABASES

Nucleotide sequence databases collect and edit sequences of nucleic acids (DNA and RNA). There are three biggest nucleotide sequence databases on the Internet, available to majority researchers in the world, and many smaller specialized databases. Each of these three biggest nucleotide groups collects a portion of the total sequence data reported worldwide, and all new and updated database entries are exchanged between the groups on a daily basis. **EMBL** cover the United Kingdom and the rest of the Europe area, **GenBank,** as part of the NCBI (*National Center for Biotechnology Information*), collects data from the USA, and **DDBJ** Database is oriented on Japanese area.

B.1.1. EMBL
(http://www.ebi.ac.uk/embl/index.html)

EMBL (*European Molecular Biology Laboratory*) is a nucleotide sequence database of European Bioinformatic Institute (EBI) from the UK. EBI is a center for research and development of bioinformatic tools, where numerous biological databases are located, including nucleic acids, protein sequences and macromolecular structures. EMBL is the biggest and most important European DNA and RNA data source [G. Stoesser *et al.*, *Nucl. Acids Res. 30*(1) (2002) 21–26]. This nucleotide database is European member of international three-group collaboration DDBJ/EMBL/GenBank (Fig. 1). Filling the data is enabled through direct access of the researchers and genome sequencing project teams or from European Patent Office (EPO). Different tools (*Blitz*, *Fasta*, *BLAST*, *etc.*) are available for sequence similarity searching within this database. Within the last few years, the size of EMBL Database has increased several times, and today there are more than **26.2** million entries and almost **40.4** billion nucleotides inside from more than **75 000** species [http://www3.ebi.ac.uk/Services/DBStats/; on 16th June 2003].

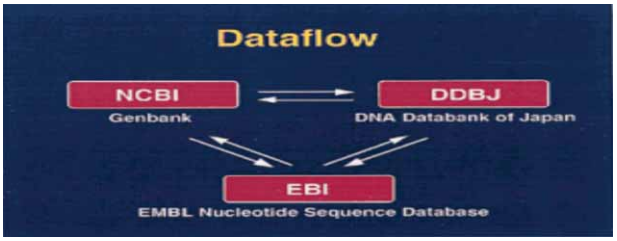


Fig. 1. Nucleotide Sequence Databases international collaboration

B.1.2. NCBI
(http://www.ncbi.nlm.nih.gov)

NCBI (*National Center for Biotechnology Information*) is an organization established in 1988 which creates public databases. This is a resource for molecular biology research information, development of software tools, analyzing genome data and better understanding of biomedical information and molecular processes affecting human health and disease. NCBI is a Division of National Library of Medicine (NLM) at National Institutes

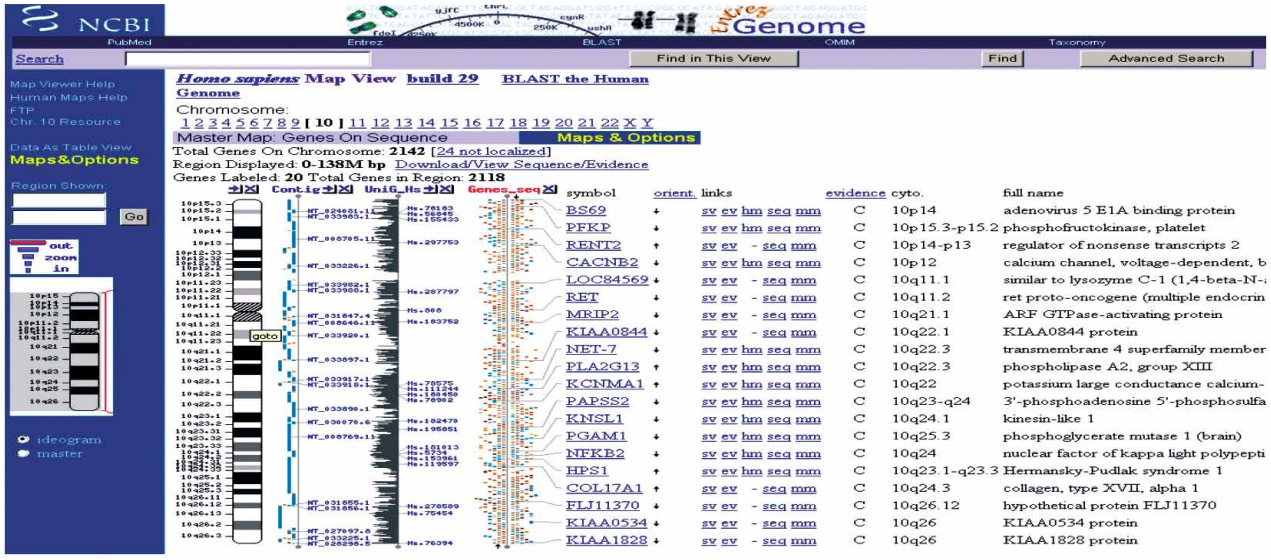


Fig. 2. NCBI Map View: Genes On Sequence (Human Genome; 10th Chromosome)
(http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/ hum_srch?chr=hum_chr.inf&query)

of Health (NIH). NCBI has an automatic system for storing and analysis of molecular biology, biochemistry and genetic knowledge data.

NCBI is one of the biggest bioinformatic centers in the USA that comprises a huge network of protein and nucleotide databases, as well as total overview of human genome and genomes of other organisms. One part of human genome map can be seen in Fig. 2.

B.1.2.1. Entrez / PubMed

Inside NCBI network there are numerous databases integrated by **Entrez** system, which enables the search and listing of information from NCBI databases. These databases cover nucleotide sequences, protein sequences, macromolecular structures and known genomes or part of genomes (Fig. 3). Inside this system there is also **PubMed** [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed], scientific overview of bibliographic news and education. Through PubMed we can find MEDLINE, a database with over **12** million citations from biomedical journals (Fig. 3) [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed; on 16th June 2003].

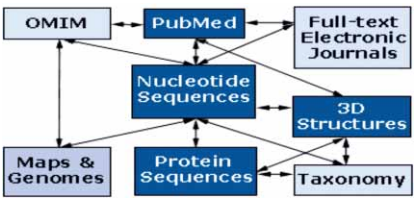


Fig. 3. Entrez overview of NCBI Databases (http://www.ncbi. nlm.nih.gov/Database/index.html)

B.1.2.2. GenBank / dbSTS / dbEST / dbGSS

GenBank is nucleotide database of **NCBI** (*National Center for Biotechnology Information*). This is one of three biggest collections of all DNA sequences available to public [D. A. Benson *et al.*, *Nucl. Acids Res. 30* (1) (2002) 17–20]. GenBank International Nucleotide Database is growing exponentially (Fig. 4) and the major part of imports is enabled through direct admissions from the authors with **BankIT** form (over the Internet) or with **Sequin** software.

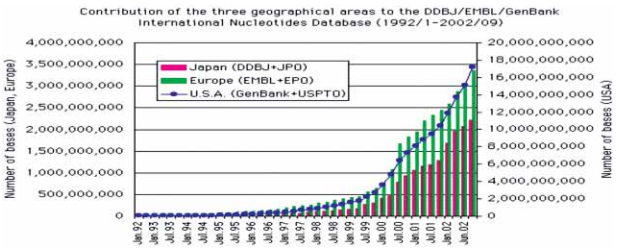


Fig. 4. Growth of GenBank/EMBL/DDBJ nucleotide Databases (http://www.ncbi.nlm.nih.gov/Genbank/index.html)

Within the nucleotide database, GenBank contains also many search tools options, such as BLAST set of programs, which enable similarity search of nucleotide and protein sequences. There are also several cDNA sequence databases within GenBank. **dbSTS** (*Sequence Tagged Sites*; http://www.ncbi.nlm.nih.gov/dbSTS/index.html) [M. Olson *et al.*, *Science*, *245*(4925) (1989) 1434–1435] and **dbEST** (*Expressed Sequence Tags*; http://www.ncbi.nlm.nih.gov/dbEST/index.html) [M. D. Adams *et al.*, *Science*, *252*(5013) (1991) 1651–1656] are cDNA databases, and **dbGSS** (*Genome Survey Sequences*; http://www.ncbi.nlm.nih.gov/dbGSS/index.html) is a database similar to dbEST, but with pure genome sequences, and not cDNA (mRNA). dbGSS contains detailed information about contributors, experimental conditions and genetic map locations.

B.1.3. DDBJ

(http://www.ddbj.nig.ac.jp/)

DDBJ (*DNA Databank of Japan*) is Japanese nucleotide database created in 1986 within the Center for Information Biology on National Institute of Genetics (NIG). DDBJ collects the data mainly from Japanese scientists, but since they are in the three-part collaboration with European and American scientists, data and information are exchanged with them on daily basis. The scheme and functionality of DDBJ Nucleotide Sequence Database is shown in Fig. 5.

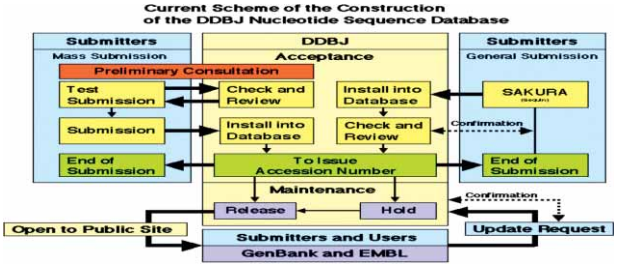


Fig. 5. Data construction of DDBJ Database (http://www.ddbj. nig.ac.jp/images/DataFlow-e.gif)

There are several homology search options in DDBJ Nucleotide Database such as FASTA, BLAST (*Basic Local Alignment Search Tool*), PSI-BLAST (*Position Specific Iterated BLAST*), SSEARCH (*Smith-Waterman algorithm*) and S&W SEARCH (*high speed S-W algorithm*). Given data can be analyzed by multiple alignments and tree-making techniques with CLUSTALW program. Within DDBJ there is also a complete genome analysis service (DDBJ/CIB *Human Genomics Studio*; http://studio.nig.ac.jp/index.html) [T. Imanishi *et al.*, http://studio.nig.ac.jp/].

B.1.3.1. GIB

(http://gib.genes.nig.ac.jp/)

In **GIB** database (*Genome Information Broker*; http://gib.genes.nig.ac.jp/) there is a lot of information about genomes sequenced up to date. Here are stored data about the whole list of species. At the moment, there are **6** Eukaryota, **109** Bacteria and **16** Archea organisms [http://gib.genes.nig.ac.jp; on 16th June 2003].

B.1.3.2 GTOP

(http://spock.genes.nig.ac.jp/%7Egenome/gtop.html)

GTOP (*Genome to Protein structure and function*) is a database that comprises the analysis of proteins identi-

```
RN   [5]
RP   1-1859
RX   MEDLINE; 91322517.
RA   Oxtoby E., Dunn M.A., Pancoro A., Hughes M.A.;
RT   "Nucleotide and derived amino acid sequence of the cyanogenic
RT   beta-glucosidase (linamarase) from white clover (Trifolium repens L.).";
RL   Plant Mol. Biol. 17:209-219(1991).
XX
RN   [6]
RP   1-1859
RA   Hughes M.A.;
RT   ;
RL   Submitted (19-NOV-1990) to the EMBL/GenBank/DDBJ databases.
RL   M.A. Hughes, UNIVERSITY OF NEWCASTLE UPON TYNE, MEDICAL SCHOOL, NEW CASTLE
RL   UPON TYNE, NE2 4HH, UK
XX
DR   MENDEL; 11000; Trirp;1162;11000.
DR   SWISS-PROT; P26204; BGLS_TRIRP.
XX
FH   Key                  Location/Qualifiers
FH
FT   source               1..1859
FT                        /db_xref="taxon:3899"
FT                        /organism="Trifolium repens"
FT                        /tissue_type="leaves"
FT                        /clone_lib="lambda gt10"
FT                        /clone="TRE361"
FT   CDS                  14..1495
FT                        /db_xref="SWISS-PROT:P26204"
FT                        /note="non-cyanogenic"
FT                        /EC_number="3.2.1.21"
FT                        /product="beta-glucosidase"
FT                        /protein_id="CAA40058.1"
FT                        /translation="MDFIVAIFALFVISSFTITSTNAVEASTLLDIGNLSRSSFPRGFI
FT                        FGAGSSAYQFEGAVNEGGRGPSIWDTFTHKYPEKIRDGSNADITVDQYHRYKEDVGIMK
FT                        DQNMDSYRFSISWPRILPKGKLSGGINHEGIKYYNNLINELLANGIQPFVTLFHWDLPQ
FT                        VLEDEYGGFLNSGVINDFRDYTDLCFKEFGDRVRYWSTLNEPWVFSNSGYALGTNAPGR
FT                        CSASNVAKPGDSGTGPYIVTHNQILAHAEAVHVYKTKYQAYQKGKIGITLVSNWLMPLD
FT                        DNSIPDIKAAERSLDFQFGLFMEQLTTGDYSKSMRRIVKNRLPKFSKFESSLVNGSFDF
FT                        IGINYYSSSYISNAPSHGNAKPSYSTNPMTNISFEKHGIPLGPRAASIWIYVYPYMFIQ
FT                        EDFEIFCYILKINITILQFSITENGMNEFNDATLPVEEALLNTYRIDYYYRHLYYIRSA
FT                        IRAGSNVKGFYAWSFLDCNEWFAGFTVRFGLNFVD"
FT   mRNA                 1..1859
FT                        /evidence=EXPERIMENTAL
XX
SQ   Sequence 1859 BP; 609 A; 314 C; 355 G; 581 T; 0 other;
     aaacaaaacca aatatggatt ttattgtagc catatttgct ctgtttgtta ttagctcatt        60
     cacaattact tccacaaaatg cagttgaaagc ttctactctt cttgacatag gtaacctgag       120
     tcggagcagt tttcctcgtcg gcttcaatct tggtgctgga tctttcaagcat accaaattga      180
     aggtgcagta aacgaaggcc gtagaggacc aagtatttgg gatacctcca cccataaata      240
     tccagaaaaaa atagggatg gaagcaatgc agacatcacg gttgaccaat atcaccgcta      300
     caaggaagat gttgggatta tgaagggatca aaatatggat tcgtatagat tctcaatctc      360
```

Fig. 6. Part of the sequence in EMBL Nucleotide Database (http://srs6.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-e+[EMBL-ID:'TRBG361'])

fied in different genome projects. Receiving the data from GTOP is enabled through sequence homology comparison with PSI-BLAST and predicting protein families as well as three-dimensional protein structures.

All sequences stored within any of these databases have similar form. Every new sequence is added to some group (division), and they are assigned with a three-letter code: Humans-HUM, Plants-PLN, Fungi-FUN, Viruses-VRL, *etc*. Each record in the database has a unique Key (ID), which is unchangeable and can be cited in scientific literature, and an Accession Number (AC). Other fields represent sequence origin, date, source, references and at the end pure sequence. Almost all programs handling sequences need them in one of the standard formats, such as FASTA format (*.txt type of file, where every sequence has header beginning with symbol > followed by short descriptions, and sequence itself is in the next row) [G. Stoesser *et al.*, *Nucl. Acids Res. 30*(1) (2002) 21–26]. Example of a part of one new added sequence in nucleotide database is presented in Fig. 6.

B.1.4. OTHER DATABASES

There are also a lot of specialized nucleotide databases that cover particular areas such as Ribosomes, Mitochondria, Eukaryote promotors, Imunoglobulins, Vectors, Telomers, *etc*.

B.1.4.1. AsDB
(http://www.hgc.ims.u-tokyo.ac.jp/~knakai/asdb.html)

AsDB is Aberant Splicing database which contains a collection of mutation causing genetic diseases. At the moment, there are **90** genes and **209** mutations in the database [K. Nakai, H. Sakamoto, *Gene*, *141* (1994) 171–177].

B.1.4.2. ACUTS
(http://pbil.univ-lyon1.fr/acuts/ACUTS.html)

ACUTS (*Ancient Conserved UnTranslated Sequences*) database identifies new regulatory elements in ancient conserved untranslated regions of protein-coding genes. The approach to the database is based on comparative sequence analysis for the identification of conserved evolutionary regulatory elements [L. Duret, P. Bucher, *Curr. Opin. Struct. Biol. 7* (1997) 399–406].

B.1.4.3. EPD
(http://www.epd.isb-sib.ch/)

EPD is a database of Eukaryote promotors developed and maintained by bioinformatic group on ISREC (*Swiss Institute for Experimental Cancer Research*) within the Swiss Institute for Bioinformatics (SIB). EPD is structured in a way that facilitates dynamic extraction of biologically meaningful promoter subsets for comparative sequence analysis using different criteria [V. Praz *et al.*, *Nucl. Acids Res. 30*(1) (2002) 322-324].

B.1.4.4. IMGT
(http://imgt.cines.fr)

IMGT (*International ImMunoGeneTics Database*) is a high-quality information system specializing in Immunoglobulins (IG), T cell receptors (TR) and Major Histocompatibility Complex (MHC) molecules of all vertebrate species, created in 1989 by Marie-Paule Lefranc. IMGT consists of sequence databases, genome and structure databases, Web resources and interactive tools. The IMGT server provides common access to all immunogenetics data [M. P. Lefranc, *Nucl. Acids Res. 29*(1) (2001) 207–209].

B.1.4.5. HOVERGEN / HOVERPROT / HOVERNUCL
(http://pbil.univ-lyon1.fr/databases/hovergen.html)

HOVERGEN (*Homologous Vertebrate Genes Database*) is a database of homologous vertebrate genes which enables choosing homologous genes and then vizualizing the same genes by multiple alignments and phylogenetic trees [L. Duret et *al.*, *Nucl. Acids Res. 22* (1994) 2360–2365]. It is especially useful for comparative sequence analysis and molecular evolution studies. On this server there are two databases: **HOVERPROT** (contains protein se-

quences) and **HOVERNUCL** (contains nucleotide sequences).

B.1.4.6. ISIS
(http://www.introns.com)

ISIS (*Intron Sequence and Information database*) is the first database of intron sequences and information about them (Fig. 7). It is a database where characteristics of particular genes or introns, and information about possible evolution and function are stored. Searching through the ISIS is possible by several different criteria, such as taxonomy searching or searching using common elements within introns. The purpose of this database is to enable better understanding of the role of introns in eukaryote biology [L. Croft *et al.*, *Nat. Genet. 24*(4) (2000) 340–341].

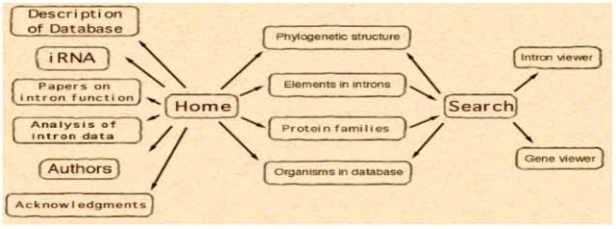


Fig. 7. Scheme of the structure of ISIS Database (http://www. introns.com)

B.1.4.7. GenLink
(http://www.genlink.wustl.edu/index.html)

GenLink is very important multimedia database resource for human genetics and telomere research. The purpose of this database is to facilitate the integration of physical and genetic linkage data (published and unpublished) to produce unified maps of the human genome. This is enabled by linkage mapping information and software tools.

B.1.4.7.1. GenotypesDB
(http://www.genlink.wustl.edu/gtypes/index.html)

GenotypesDB is genotype database within GenLink. All public information about genotypes is stored here.

B.1.4.7.2. TelDB / PtelDB
(http://www.genlink.wustl.edu/teldb/index.html)

TelDB is a database with information about telomeres. It is also part of the GenLink. Since telomeres are specialized nucleoprotein complexes that represent the physical ends of linear eukaryotic chromosomes, they have important functions, primarily in the protection, replication, and stabilization of the chromosome ends. TelDB contains over **2 100** citations about telomeres from **290** journals and covers over **120** species [http://www.genlink.wustl.edu/teldb/teldb.html; on 16th June 2003]. New part of this database is **PtelDB**, Protein Telomeres Database, where information about genes of telomeres and their protein products can be found.

B.1.4.8. TRADAT
(http://www.itba.mi.cnr.it/tradat/)

TRADAT (*TRAnscription Databases and Analysis Tools*) is a database, which is the result of European Commission Project, with a purpose to develop and provide tools for the interpretation of genomic DNA sequences with special emphasis on regulatory regions.

B.1.4.9. MPDB
(http://www.biotech.ist.unige.it/interlab/mpdb.html)

MPDB (*Molecular Probe Data Base*) contains information about **4 300** synthetic oligonucleotides with sequences up to 100 nucleotides long from **821** different genes [http://www.biotech.ist.unige.it/interlab/mpdb.html; on 16th June, 2003]. Among them, 691 are human genes, and 112 are viral. These are mainly literature data. MPDB database is available through some SRS servers (*eg*. Cancer Institute, Genoa, Italy) and the access is limited.

B.1.4.10. VectorDB
(http://life.nthu.edu.tw/~g854202/Vecdtb.html)

VectorDB is database of vector sequences Anderson Unicom group which contains descriptions and information about numerous vector sequences that are used most frequently in molecular biology. Information for more than **2 600** vectors can be found through the search options [http://life.nthu.edu.tw/~g854202/Vecdtb.html; on 16th June, 2003]. Available vectors in this Database are Phage, Plasmid, Phagemid, Phasmid, Cosmid, Virus and YAC vectors.

B.1.4.11. RNA-Specific Databases

B.1.4.11.1. rRNA WWW Server
(http://oberon.rug.ac.be:8080/rRNA/)

rRNA WWW Server (*rRNA-Database of Ribosomal Subunit Sequences*) is a Web-site of Ribosomal RNA. This is the European rRNA database created at the University of Antwerpen (Belgium) in 1983. This database compiles all complete or nearly complete SSU (small subunit) and LSU (large subunit) ribosomal RNA sequences [J. Wuyts *et al.*, *Nucl. Acids Res. 29*(1) (2001) 175–177] and [J. Wuyts *et al.*, *Nucl. Acids Res. 30* (2002) 183–185].

B.1.4.11.2. RDP (http://rdp.cme.msu.edu/html/)

RDP (*Ribosomal Database Project*) is a project of the Center for Microbial Ecology Michigan State University, USA. The RDP provides ribosome related data services to the scientific community, including online data analysis, rRNA derived phylogenetic trees, and aligned and annotated rRNA sequences [B. L. Maidak *et al.*, *Nucl. Acids Res. 29*(1) (2001) 173–174]. The database contains more than **71 000** 16S rRNA sequences [http://rdp.cme.msu.edu/html/; on 16th June, 2003]. RDP project is connected mainly with three biggest nucleotide databases: GenBank, EMBL and DDBJ.

B.1.4.11.3. 5S Ribosomal RNA DB
(http://rose.man.poznan.pl/5SData/)

5S ribosomal RNA Database contains data about all cytoplasmatic 5S rRNA, their genes and most organelle ribosomes. There are **1 985** primary structures 5S rRNA and 5S rDNA inside the Database [M. Szymanski, *et al.*, *Nucl. Acids Res. 28*(1) (2000) 166–167]. Nucleotide sequences are divided by the taxonomy of source organisms.

B.1.4.11.4. RISSC (http://ulises.umh.es/RISSC/)

RISSC (*Ribosomal Internal Spacer Sequence Collection*) is the database of ribosomal 16S–23S spacer sequences. Ribosomal spacers have proven to be useful tools for typing and identifying closely related prokaryotes due to their high variability in size and sequence. RISSC genes are used for molecular linking of the microbes on the taxonomic level of species. RISSC works with sequence data collected from the EMBL-GenBank-DDBJ databases, and they use the same formats.

B.1.4.11.5. GtRDB (http://rna.wustl.edu/GtRDB/)

GtRDB (*The Genomic tRNA Database*) is genomic tRNA database that identifies 99–100 % of tRNA genes in DNA sequences. This is enabled using *tRNAscan-SE* program [T. M. Lowe, S. R. Eddy, *Nucl. Acids Res. 25* (1997) 955–964]. The program is made within the search server of the University in St. Louis, USA [http://www.genetics.wustl.edu/eddy/tRNAscan-SE/].

B.1.4.11.6. tRNA and tRNA Gene Sequences
(http://www.uni-bayreuth.de/departments/biochemie/trna/)

This is a compilation of **tRNA sequences** and their **genes coding tRNA** published in literature. Current Genomic tRNA Compilation consists of about **4 900** tRNA gene sequences from **87** organisms covering archea, bacteria, higher and lower eukaryotes [http://www.uni-bayreuth.de/departments/biochemie/trna/; on 16th June, 2003]. The database also includes the tRNA gene sequences collected in GtRDB [T. M. Lowe, S. R. Eddy, *Nucl. Acids Res. 25* (1997) 955–964] as well as those from the additional complete genomes found in DNA databases. Sequences are presented as MS Excel® files.

B.1.4.11.7. Mamit-tRNA
(http://mamit-trna.u-strasbg.fr/)

Mamit-tRNA is a compilation of genes covering primary and secondary structural features of mammalian mitochondria tRNAs. At the moment, it contains **679** tRNA gene sequences from **31** fully sequenced mammalian mitochondria genomes. These are classified into **22** families according to the amino acid specificity as defined by the anticodon triplets [http://mamit-trna.u-strasbg.fr/Summary.html; on 16th June, 2003]. Mamit-tRNA Database is divided in sequences, 2D structures and tables of organisms [M. Helm *et al.*, *RNA*, *6* (2000) 1356–1379].

B.1.4.11.8. tmRDB
(http://psyche.uthct.edu/dbs/tmRDB/tmRDB.html)

TmRDB (*tmRNA Database*) is a database from the University of Texas (Health Center Tyler), which provides aligned, annotated and phylogenetically ordered sequences related to the structure and function of tmRNA (before known as »10S RNA«). tmRNA have properties of both tRNA and mRNA combined in one molecule, and sequence similarity is searched by comparative sequence analysis [B. Knudsen *et al.*, *Nucl. Acids Res. 29* (2001) 171–172].

B.1.4.11.9. uRNADB
(http://psyche.uthct.edu/dbs/uRNADB/uRNADB.html)

uRNADB is another nucleotide database from the University of Texas, which contains structures as well as functions of different uRNA [C. Zwieb, *Nucl. Acids Res. 24* (1996) 76–79]. The data within this database are free for research purposes only, but it has not been updated since May 1997.

B.1.4.11.10. UTRdb
(http://bighost.area.ba.cnr.it/BIG/UTRHome/)

UTRdb (*UnTranslated Regions db*) is the Internet datasource for 5′ and 3′ untranslated regions of eukaryotic mRNA molecules [G. Pesole *et al.*, *Nucl. Acids Res. 30*(1) (2002) 335–340].

B.1.4.11.11. MITOMAP
(http://www.mitomap.org/)

MITOMAP is a database of human mitochondrial genomes, which also contains additional information about polymorphisms and mutations of human mitochondrial DNA (mtDNA). Searching through the database is enabled by entering a gene, disease, enzyme, or keyword information. Inside the database there are translational tables of aminoacids, references and tools for sequence manipulations [A. M. Kogelnik *et al.*, *Nucl. Acids Res. 26*(1) (1998) 112–115].

B.1.4.11.12. RNA Modification Database
(http://medlib.med.utah.edu/RNAmods/)

RNA Modification Database is a comprehensive list of post-transcriptionally modified nucleosides from RNA. The database consists of all RNA-derived ribonucleosides of known structure, including those from the established sequence positions, as well as those detected or characterized from hydrolysates of RNA. The information provided permits access to the modified nucleoside literature providing both computer-searchable Chemical Abstracts registry numbers and key literature citations [P. A. Limbach, *Nucl. Acids Res. 22* (1994) 2183–2196]. A total of **96** modified nucleosides, for which structures have been assigned, have been reported in RNA [http://medlib.med.utah.edu/RNAmods/; on 16th June, 2003].

B.1.4.11.13. Subviral RNA Database
(http://132.210.163.235/subviral/home.cgi)

Subviral RNA Database is a compilation of subviral RNA sequences. It is the database of the smallest known selfreplicable RNA: viroid and quasiviroid RNA [M. Pelchat *et al.*, *Nucl. Acids Res. 28*(1) (2000) 179–180]. Here their predicted secondary structures can be visualized with RnaViz [http://silk.uia.ac.be/rnaviz/] or RNAdraw [http://rnadraw.base8.se/] programs. At the moment,

there are **1 300** sequences in the database [http://132.210.163.235/subviral/home.cgi; on 16th June, 2003].

B.1.4.11.14. Small RNA database
(http://mbcr.bcm.tmc.edu/smallRNA/smallrna.html)

Small RNA database is a database of those parts of RNA that are not directly involved in protein synthesis. The size of these molecules is usually 75–400 nucleotides, but some of them are up to 1 000 nucleotides long.

B.1.4.11.15. Noncoding RNAs Database

Noncoding RNAs Database (http://biobases.ibch.poznan.pl/ncRNA) contains nucleotide sequences of eubacteria, archea and different eucaryotic sequences (stress induced transcripts, noncoding transcripts from imprinted genes, protein function modulators, nervous system RNAs, RNAs involved in RNA/protein localization and other types of noncoding RNAs).

B.1.4.11.16. CRW
(http://www.rna.icmb.utexas.edu/)

CRW (*Comparative RNA Web Site*) is publicly available service within the University of Texas (USA), but access is limited with login and password (for commercial usage permission of author is needed). This service gives information about collected RNA sequences and structures, comparative structure models, nucleotide frequency, conservation information, phylogenetic structure analysis and interpretation, visualization and downloading of the data [J. J. Cannone *et al.*, *BioMed Central Bioinformatics*, *3* (2002) 2].

B.1.4.12. Gencarta Database
(http://www.gencarta.com/hello/)

Gencarta is commercially available database designed and created by Compugen [http://www.cgen.com]. Within the database there are genes predicted by computational methods, information about chromosoms, known mRNA, transcripts and the alignment of genomes. Additionaly, there is information about expression profiles, functional anaysis, domain information and detailed reports about known and predicted proteins and their homologies (Fig. 8).

B.1.4.13. GENESEQ
(http://www.biocenter.helsinki.fi/bi/margus/KURSSI/db/dbgeneseq.html)

GENESEQ computer data bank consists of protein and nucleotide sequences collected from published patent documents around the world. GENESEQ Database is developed in cooperation with Derwent Publications and it consists of two parts: N-GENESEQ (nucleic acid part of GENESEQ) and A-GENESEQ (protein part of GENESEQ).

B.1.4.14. ArrayExpress
(http://www.ebi.ac.uk/arrayexpress/)

ArrayExpress is a public repository for microarray based gene expression data. Searching is performed by entering keywords, by description of particular array, or using a query for protocol, experiment or sample. The data about experimental conditions, protocols and biosources (nucleic acids for hybridization) are stored here. At the moment, ArrayExpress contains **31** experiments, **38** arrays and **243** protocols [http://www.ebi.ac.uk/arrayexpress/; on 16th June, 2003]. There is a possibility of sending the own new data into the database by **MIAMExpress** tool or through **Expression Profiler** tool, accessible over the Internet. These tools enable clustering of data and other types of data analysis.

B.1.4.15. PEDANT
(http://pedant.gsf.de)

PEDANT genome database of MIPS (*Munich Information center for Protein Sequences*) provides exhaustive automatic analysis of genomic sequences by a large va-

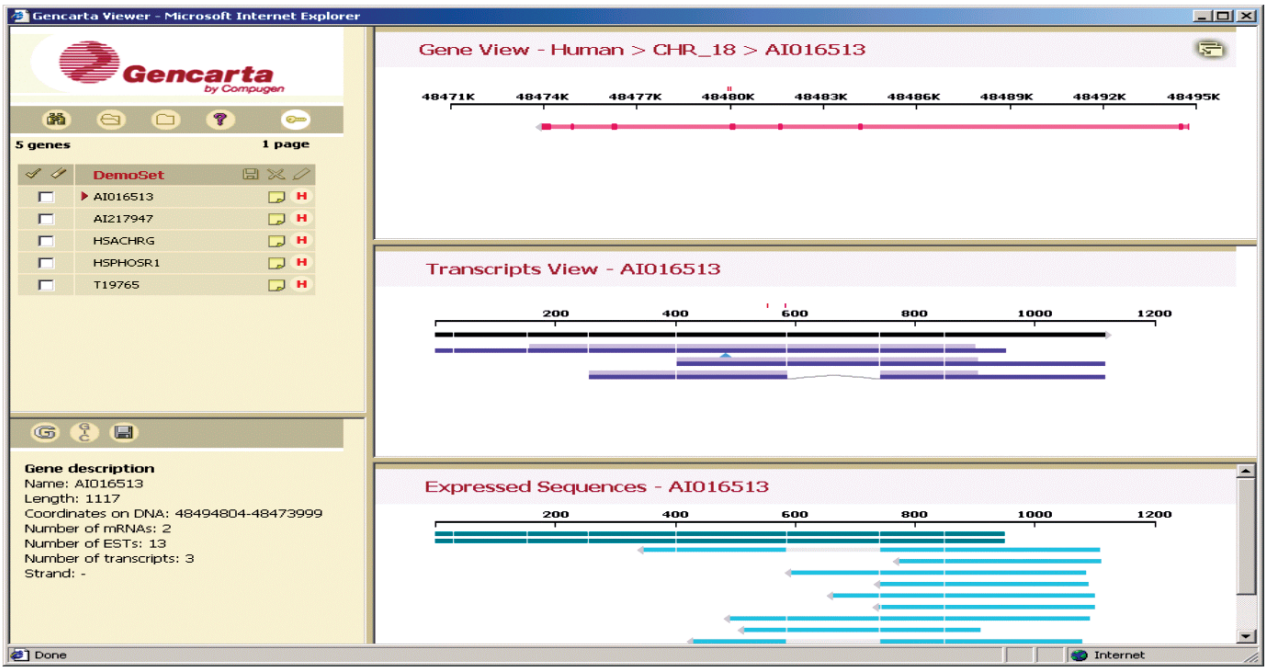


Fig. 8. A view of Gencarta Database (http://www.gencarta.com/demo/gencarta.html)

riety of established bioinformatics tools through a comprehensive Web-based user interface [D. Frishman *et al.*, *Nucl. Acids Res. 31*(1) (2003) 207–211]. Only academic and non-commercial users have permission to access this server without a commercial license.

B.2. AMINOACID SEQUENCE DATABASES

Searching protein databases using sequence similarity criteria is more sensitive if compared with nucleotide (DNA) sequence databases search, because of much bigger capacity of protein information (20 amino acids *vs.* only 4 nucleotides). From the point of view of genetic code, some amino acids have more than one nucleotide three-letter code. Therefore, on the protein level some compared sequences can be identical, but they can be different if we take a look into a corresponding DNA sequence.

B.2.1. SWISS-PROT / ExPASy
(http://www.ebi.ac.uk/swissprot/index.html)

SWISS-PROT is a database of protein sequences established in 1986. It contains high level of different protein information such as description of protein function, its domain structure, post-translational modification and other useful information. A simple way to access the SWISS-PROT Database is using **SRS** (*Sequence Retrieval System*) [http://srs6.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-page+top], or using **ExPASy Server** in Geneva, [http://us.expasy.org/sprot/]. Within the database it is possible to search by using keywords, taxonomy, different descriptives or identification numbers, authors of citations [Boeckmann *et al.*, *Nucl. Acids Res. 31* (2003) 365–370]. SWISS-PROT is maintained collaboratively by the Swiss Institute for Bioinformatics (SIB) and the European Bioinformatics Institute (EBI). Database contains **127 863** sequence entries, comprising **47.0** million amino acids abstracted from **105 755** references [http://us.expasy.org/sprot/relnotes/relstat.html; on 16th June, 2003].

B.2.2. TrEMBL
(http://www.ebi.ac.uk/trembl/index.html)

Within the biggest world nucleotide databases: EMBL, GenBank and DDBJ, there is direct access to **TrEMBL** (Translated EMBL) Protein Sequence Database. This database contains translations of all coding sequences existing in EMBL nucleotide database, which are not yet integrated in SWISS-PROT Database. TrEMBL represents link between nucleotide and protein sequence databases. It is divided into two main sections: **SP-TrEMBL** (SWISS-PROT TrEMBL; containing data that are part of SWISS-PROT database) and **REM-TrEMBL** (REMaining TrEMBL; containing data that are not included in SWISS-PROT) [A. Bairoch and R. Apweiler, *Nucl. Acids Res. 28* (2000) 45–48]. TrEMBL database contains **857 951** sequences with **266.5** million amino acids [http://www2.ebi.ac.uk/swissprot/sptr_stats/index.html; on 16th June, 2003].

B.2.3. PIR / PSD / iProClass / RESID
(http://www-nbrf.georgetown.edu/pir/)

PIR (*Protein Information Resource*) is a public protein information resource established in 1984 as division of the National Biomedical Research Foundation (NBRF) with the purpose of supporting genomic and proteomic research and scientifis discovery.

In collaboration with Munich Information Center for Proteine Sequences (MIPS) and Japanese International Protein Information Database (JIPID), International Protein Sequence Database **PSD** is created within PIR [H. Cathy *et al.*, *Nucl. Acids Res. 30* (2002) 35–37]. This is one of the major protein sequence databases, which contains **283 308** entries [http://www-nbrf.georgetown.edu/pirwww/search/textpsd.shtml; on 16th June, 2003].

To improve assignation of proteins and better exploring of experimental data, a system for bibliographyc and literature categorization of data has been developed: **iProClass** Database, which includes classification of protein families, superfamilies, domains and motifs, structural and functional characteristics of proteins [C. Wu *et al.*, *Nucl. Acids Res. 29* (2001) 52–54]. It currently consists of non-redundant PIR and SWISS-PROT/TrEMBL proteins organized in more than **36 200** PIR superfamilies, **145 300** families, **5 720** domains, **1 300** motifs, **280** post-translational modification sites, and links to over **50** biological databases. At the moment there are more than **1** million entries inside [http://pir.georgetown.edu/iproclass/; on 16th June, 2003].

Within PIR there is also **RESID** Database, which represents comprehensive collection of protein modifications including amino-terminal, carboxy-terminal modifications, peptide chain and post-translational modifications. At the moment, this database contains **333** entries [http://www-nbrf.georgetown.edu/pirwww/dbinfo/resid.html; on 16th June, 2003].

B.2.4. GenPept
(http://bioinfo.huji.ac.il/databases/genpept.shtml)

GenPept is a database of translated sequences coding proteins. GenPept entries are derived from entries in the GenBank nucleotide sequence data bank. They contain minimal annotation, primarily extracted from the corresponding GenBank entries. GenPept Database is made within Bioinformatic server of the Faculty of Medicine from the Hebrew University Jerusalem.

B.2.5. PROTOMAP
(http://protomap.cornell.edu/)

PROTOMAP is an automatic hierarchical classification of all SWISS-PROT and TrEMBL proteins. Protein sequences are classified in well defined groups, and most of them correlate with natural biological families and superfamilies. [G. Yona *et al.*, *Nucl. Acids Res. 28* (2000) 49–55]. At the moment, there are **365 174** protein sequences within PROTOMAP [http://protomap.cornell.edu/; on 16th June, 2003].

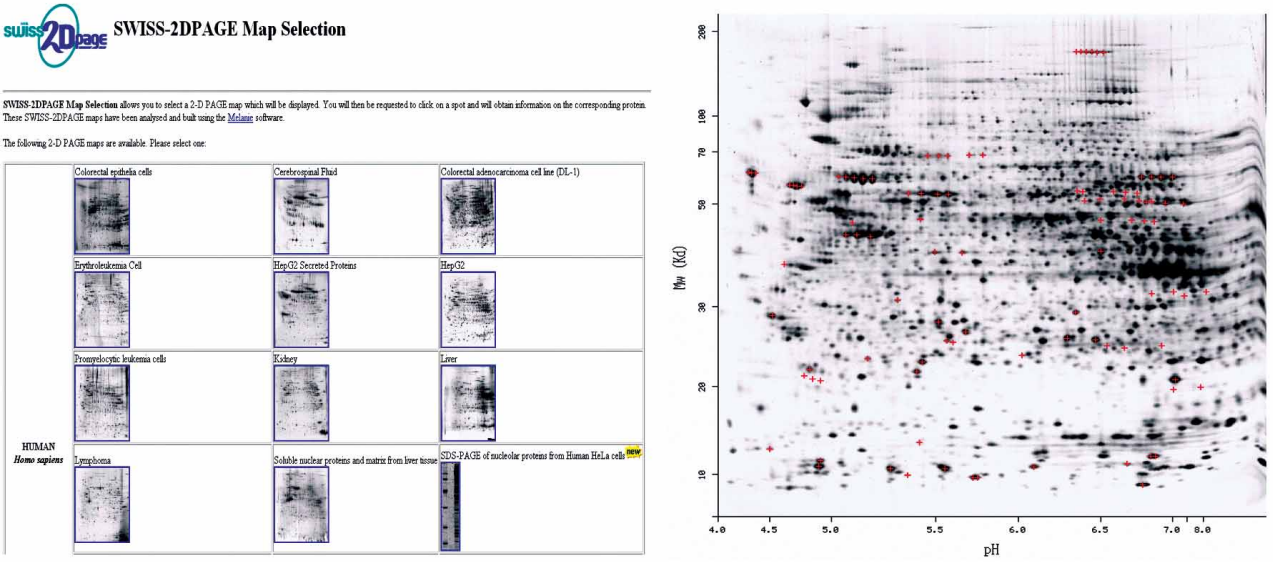


Fig. 9. SWISS-2DPAGE Reference Maps (left: 2D-PAGE Map Selection; right: human HEP-G2 [J. C. Sanchez *et al.*, *Electrophoresis*, *16* (7) (1995) 1131–1151]) (http://us.expasy.org/cgi-bin/map1) (Copyright Swiss Institute of Bioinformatics, Geneva, Switzerland)

B.2.6. PROSITE
(http://us.expasy.org/prosite/)

PROSITE is a database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help identify reliably to which known protein family (if any) a new sequence belongs [L. Falquet *et al.*, *Nucl. Acids Res. 30* (2002) 235–238].

B.2.7. SWISS-2DPAGE
(http://us.expasy.org/ch2d/)

SWISS-2DPAGE is a two-dimensional polyacrylamide gel electrophoresis database established in 1993, which contains data on proteins identified on various 2-D PAGE and SDS-PAGE reference maps. You can locate these proteins on the 2-D PAGE maps or display the region of a 2-D PAGE map where one might expect to find a protein from SWISS-PROT [C. Hoogland *et al.*, *Nucl. Acids Res. 28*(1) (2000) 286–288]. Besides the keywords, author and text, this database is also searchable with a click on the particular spot on the gel (Fig. 9). The reference maps from many human and mouse biological samples, as well as from some other organisms, are available within the database. There are **1 042** entries in **34** reference maps in the SWISS-2DPAGE [http://us.expasy.org/ch2d/; on 16th June, 2003].

B.2.8. ENZYME
(http://us.expasy.org/enzyme/)

ENZYME is an enzyme nomenclature database. This is a repository of information relative to the nomenclature of enzymes and it is primarily based on the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB) and it describes each type of characterized enzyme for which an EC (Enzyme Commission) number has been provided. Since enzymes are proteins, which are precisely classified upon their function, searching within ENZYME is possible by using EC number, defined class, description or alternative names, chemical compounds or cofactors [A. Bairoch, *Nucl. Acids Res. 21*(13) (1993) 3155–3156]. There are **4 159** entries within the database [http://us.expasy.org/enzyme/; on 16th June, 2003].

B.2.9. BRENDA
(http://www.brenda.uni-koeln.de/)

BRENDA is the main collection of enzyme functional data extracted directly from the primary literature. It is available free of charge for academic, non-profit users, and as an in-house database for commercial users (requires a license). This comprehensive enzyme information system is maintained and developed at the institute of Biochemistry at the University of Cologne. Enzymes are divided by nomenclature (EC number, enzyme name, synonyms, CAS number, *etc.*), reactions and specificity (substrates, inhibitors, cofactors, *etc.*), functional parameters (K_m value, turnover number, specific activity, pH, temperature, *etc.*) and isolation organism information.

B.2.10. SeqAnalRef
(http://us.expasy.org/seqanalref/)

SeqAnalRef is a sequence analysis bibliographic reference database from the field of mathematical and computer analysis of biomolecular sequences initiated by Amos Bairoch in 1995. SeqAnalRef Database has not been updated since 1996, although it is still available [http://us.expasy.org/seqanalref/].

B.2.11. AARS
(http://rose.man.poznan.pl/aars/index.html)

AARS (*Aminoacyl-tRNA Synthetases Database*) is a database with information about aminoacid sequences

of aminoacyl-tRNA synthetases. Sequences in the database are available through a list of aminoacids or through a list of organisms [M. Szymanski *et al.*, *Nucl. Acids Res. 29* (2001) 288–290].

B.2.12. Pfam
(http://www.sanger.ac.uk/Software/Pfam/index.shtml)

Pfam (*Protein families database of alignments and HMMs*) is a large collection of multiple sequence alignments and Hidden Markov Models covering many common protein domains and families. For each family in Pfam there is a possibility to look at multiple alignments, to view protein domain architectures, to examine species distribution, to follow links to other databases, and to view known protein structures or domain organization of proteins [A. Baternan *et al.*, *Nucl. Acids Res. 30*(1) (2002) 276–280]. 75 % of protein sequences have at least one match to Pfam, and at the moment this database consists of **5 724** protein families [http://www.sanger.ac.uk/Software/Pfam/index.shtml; on 16th June, 2003].

B.2.13. PRINTS
(http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/)

PRINTS is a database of protein *fingerprints*. A »fingerprint« is a group of conserved motifs used to characterize a protein family. SWISS-PROT and TrEMBL Databases are scanned for motifs. Usually the motifs do not overlap, but are separated along a sequence, although they may be contiguous in 3D-space. Fingerprints can encode protein folds and functionalities more flexibly and powerfully than single motifs [T. K. Attwood *et al.*, *Nucl. Acids Res. 30*(1) (2002) 239–241]. PRINTS Database can be accessed by accession number, PRINTS code, database code, text, sequence, title, number of motifs and by author. At the moment, there are **1 800** fingerprints, encoding **10 931** single motifs in the database [http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/printscontents.html; on 16th June, 2003].

B.2.14. ProDom
(http://prodes.toulouse.inra.fr/prodom/2002.1/html/home.php)

ProDom is a comprehensive set of protein domain families automatically generated from the SWISS-PROT and TrEMBL sequence databases [F. Servant *et al.*, *Briefings in Bioinformatics*, *3*(3) (2002) 246–251]. It contains automatic clustering of homologous domains, which is a rational way of organizing protein sequence data. An interactive graphical interface was designed to allow for easy navigation between schematic domain arrangements, multiple alignments, phylogenetic trees, SWISS-PROT entries, PROSITE patterns, Pfam families and 3-D structures in the PDB [F. Corpet *et al.*, *Nucl. Acids Res. 28*(1) (2000) 267–269].

B.2.15. SMART
(http://smart.embl-heidelberg.de/)

SMART (*Simple Modular Architecture Research Tool*) is a database that allows the identification and annotation of genetically mobile domains and the analysis of domain architectures. Different domain families found in signaling, extra cellular and chromatin-associated proteins are detectable. These domains are extensively annotated with respect to functional class, tertiary structures and functionally important residues [I. Letunic *et al.*, *Nucl. Acids Res. 30*(1) (2002) 242–244]. At the moment, **660** domain families are present in the database [http://smart.embl-heidelberg.de/; on 16th June, 2003].

B.2.16. TIGRFAMs
(http://www.tigr.org/TIGRFAMs/index.shtml)

TIGRFAMs (*TIGR protein FAMilies*) are a collection of protein families based on Hidden Markov Models (HMMs). The database is designed to support the automated functional identification of proteins by sequence homology. It provides the information best suited for automatic assignment of specific functions to proteins from large-scale genome sequencing projects.

B.2.17. InterPro
(http://www.ebi.ac.uk/interpro)

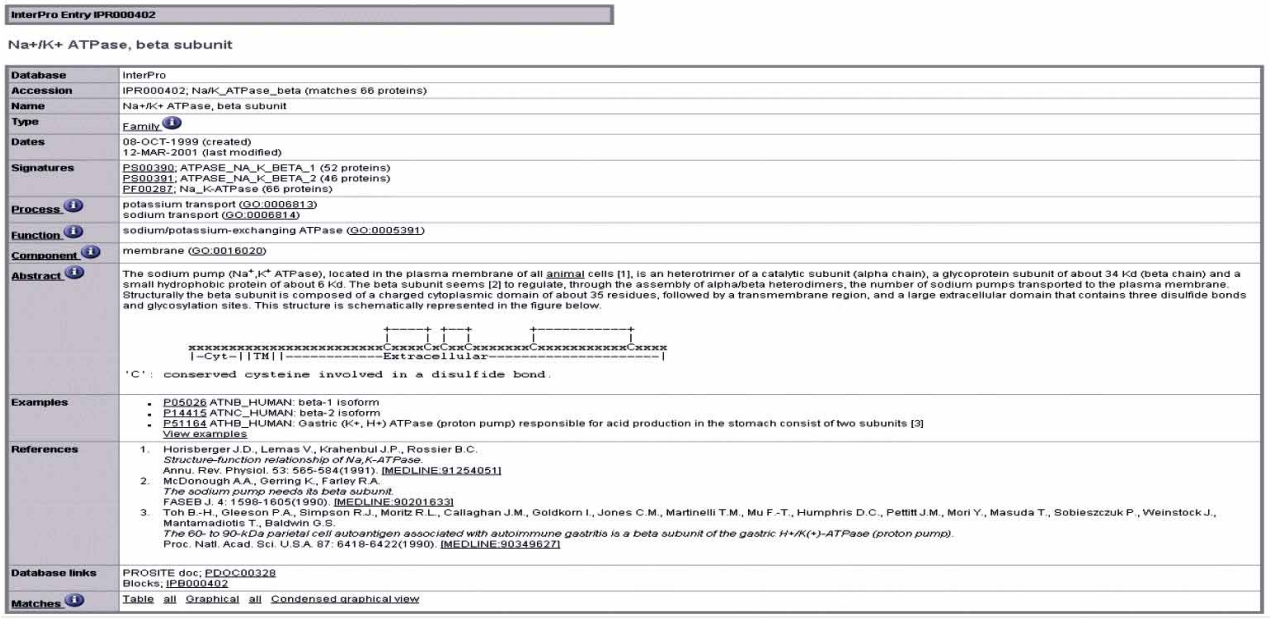
InterPro (*Integrated resource of Protein Families, Domains and Sites*) is a database organized in 1999 by the following partners: EBI, SIB, University of Manchester, Sanger Institute, GENE-IT, CNRS/INRA, LION bioscience AG and the University of Bergen. Since there are many secondary protein structure databases, and they have different formats and nomenclature, these institutions unified PROSITE, PRINTS, ProDom, Pfam, SMART and TIGRFAMs into *InterPro* [N. J. Mulder *et al.*, *Nucl. Acids Res. 31*(1) (2003) 315–318]. An example of the record in InterPro database is shown in Fig. 10. At the moment, InterPro contains **7 785** entries, representing **1 744** domains, **5 877** families, **147** repeats, and **17** post-translational modification sites [http:// www.ebi.ac.uk/interpro/; on 16th June, 2003].

B.2.18. CluSTr
(http://www.ebi.ac.uk/clustr/index.html)

CluSTr (*Clusters of SWISS-PROT+TrEMBL proteins*) Database offers an automatic classification of SWISS-PROT and TrEMBL proteins into groups of related proteins. Clustering is based on the analysis of all pairwise comparisons between protein sequences [E. V. Kriventseva *et al.*, *Nucl. Acids Res. 29*(1) (2001) 33–36]. Clusters for **5** complete eukaryote proteomes (*Homo sapiens* is one of them) and **53** prokaryote proteomes are presented in the database [http://www.ebi.ac.uk/clustr/index.html; on 16th June, 2003].

B.2.19. BLOCKS
(http://www.blocks.fhcrc.org/)

BLOCKS is a server for biological sequence analysis at the *Fred Hutchinson Cancer Research Center* in Seattle, USA. BLOCKS are multiply aligned ungapped segments corresponding to the most highly conserved regions of



InterPro Entry IPR000402

Na+/K+ ATPase, beta subunit

Database	InterPro
Accession	IPR000402; Na/K_ATPase_beta (matches 66 proteins)
Name	Na+/K+ ATPase, beta subunit
Type	Family
Dates	08-OCT-1999 (created) 12-MAR-2001 (last modified)
Signatures	PS00390; ATPASE_NA_K_BETA_1 (52 proteins) PS00391; ATPASE_NA_K_BETA_2 (46 proteins) PF00287; Na_K-ATPase (66 proteins)
Process	potassium transport (GO:0006813) sodium transport (GO:0006814)
Function	sodium/potassium-exchanging ATPase (GO:0005391)
Component	membrane (GO:0016020)

Abstract	The sodium pump (Na^+,K^+ ATPase), located in the plasma membrane of all animal cells [1], is an heterotrimer of a catalytic subunit (alpha chain), a glycoprotein subunit of about 34 Kd (beta chain) and a small hydrophobic protein of about 6 Kd. The beta subunit seems [2] to regulate, through the assembly of alpha/beta heterodimers, the number of sodium pumps transported to the plasma membrane. Structurally the beta subunit is composed of a charged cytoplasmic domain of about 35 residues, followed by a transmembrane region, and a large extracellular domain that contains three disulfide bonds and glycosylation sites. This structure is schematically represented in the figure below.
	`xxxxxxxxxxxxxxxxxxxxxxxxCxxxxCxCxxCxxxxxxxCxxxxxxxxxxxxCxxxx`
	`|-Cyt-||TM||------------Extracellular----------------------|`
	'C': conserved cysteine involved in a disulfide bond.

Examples	P05026 ATNB_HUMAN: beta-1 isoform P14415 ATNC_HUMAN: beta-2 isoform P51164 ATHB_HUMAN: Gastric (K+, H+) ATPase (proton pump) responsible for acid production in the stomach consist of two subunits [3] View examples
References	1. Horisberger J.D., Lemas V., Krahenbul J.P., Rossier B.C. *Structure-function relationship of Na,K-ATPase.* Annu. Rev. Physiol. 53: 565-584(1991). [MEDLINE:91254051] 2. McDonough A.A., Gerring K., Farley R.A. *The sodium pump needs its beta subunit.* FASEB J. 4: 1598-1605(1990). [MEDLINE:90201633] 3. Toh B.-H., Gleeson P.A., Simpson R.J., Moritz R.L., Callaghan J.M., Goldkorn I., Jones C.M., Martinelli T.M., Mu F.-T., Humphris D.C., Pettitt J.M., Mori Y., Masuda T., Sobieszczuk P., Weinstock J., Mantamadiotis T., Baldwin G.S. *The 60- to 90-kDa parietal cell autoantigen associated with autoimmune gastritis is a beta subunit of the gastric H+/K(+)-ATPase (proton pump).* Proc. Natl. Acad. Sci. U.S.A. 87: 6418-6422(1990). [MEDLINE:90349627]
Database links	PROSITE doc; PDOC00328 Blocks; IPB000402
Matches	Table all Graphical all Condensed graphical view

Fig. 10. Example of the InterPro record (http://www.ebi.ac.uk/interpro/)

proteins documented in InterPro database. *Block Searcher*, *Get Blocks* and *Block Maker* are tools for detection and verification of protein sequence homology. They compare a protein or DNA sequence to a database of protein blocks, retrieve blocks, and create new blocks, respectively [S. Henikoff *et al.*, *Bioinformatics*, *15*(6) (1999) 471–479]. Currently, BLOCKS database consists of **8 656** blocks representing **2 101** groups documented in InterPro [http://blocks.fhcrc.org/blocks/blocks_release.html; on 16th June, 2003].

B.2.20. SBASE

(http://hydra.icgeb.trieste.it/~kristian/SBASE/)

SBASE is a library of protein domains and sequences collected from the literature, from protein sequence databases and from genomic databases. Similarity search on SBASE is based on the BLAST program. The results of the BLAST search are processed to give domain similarities. The database is made in Padriciano (Trieste, Italy) at International Center for genetical Engineering and Biotechnology (ICGEB), and contains at the moment **338 655** sequences [http://hydra.icgeb.trieste.it/~kristian/SBASE/sbase.php?sec=info&sub=summary; on 16th June, 2003].

B.2.21. PMD

(http://pmd.ddbj.nig.ac.jp/)

PMD (*Protein Mutant Database*) is a database created in Japanese National Institute of Genetics, where information about all amino-acid mutations on specific positions of the proteins (functional or/and structural) are stored. PMD covers natural as well as artificial mutants, including random and site-directed ones, for all proteins except members of the globin and immunoglobulin families. For those proteins with known 3-D structures, this structure is presented here with mutation sites assigned with different colors [T. Kawabata *et al.*, *Nucl. Acids Res. 27* (1999) 355–357]. Number of entries available in the database is **29 365,** and number of mutants is **154 388** [http://pmd.ddbj.nig.ac.jp/~pmd/pmdstat.html; on 16th June, 2003].

B.2.22. ProTherm

(http://www.rtc.riken.go.jp/jouhou/protherm/protherm.html)

ProTherm (*Thermodynamic Database for Proteins and Mutants*) is a collection of numerical data of thermodynamic parameters such as Gibbs free energy change, enthalpy change, heat capacity change, transition temperature *etc.* for wild type and mutant proteins, which are important for understanding the structure and stability of proteins. It also contains information about secondary structure and accessibility of wild type residues, experimental conditions (pH, temperature, buffer, ion and protein concentration), measurements and methods used for each data, and activity information (K_m and K_{cat}) [M. M. Gromiha *et al. Nucl. Acids Res. 30*(1) (2002) 301–302]. The database contains **13 605** entries and **615** proteins [http://www.rtc.riken.go.jp/cgi-bin/jouhou/protherm/pp_stat.pl; on 16th June, 2003].

B.2.23. BindingDB

(http://www.bindingdb.org/bind/index.jsp)

BindingDB is a public database of measured binding affinities for biomolecules, genetically or chemically modified biomolecules, and synthetic compounds. Currently it contains data generated by isothermal titration calorimetry (ITC) and enzyme inhibition [X. Chen *et al.*, *J Combi Chem High-Throughput Screen*, *4* (2001) 719–725]. Searching through the database is enabled by entering authors names, keywords, inhibition constants, IC_{50}, reactants and changing of enthalpy and entropy.

B.2.24. Calcium-Binding Proteins Data Library
(http://structbio.vanderbilt.edu/cabp_database/)

EF-Hand CaBP-DL (*EF-Hand Calcium-Binding Proteins Data Library*) is a growing collection of published sequences, structural, functional and evolutionary information about proteins (and their mutants) that are targets of EF-hand CaBPs and about their roles in cellular signal transduction.

B.2.25. MDB
(http://metallo.scripps.edu/)

MDB (*Metalloprotein Database*) contains quantitative information on all the metal-containing sites available from structures in the PDB distribution. This database contains geometrical and molecular information that allows the classification and search of particular combinations of site characteristics. Metalloproteins (chemical combinations of protein atoms C, N, O, H, S with metal ions such as Fe, Ca, Cu and Zn) represent one third of all proteins and they have important roles in organism [J. M. Castagnetto *et al.*, *Nucl. Acids. Res. 30*(1) (2002) 379–382]. At the moment, the database contains information about **6 634** PDB proteins that have **24 568** sites (comprising a total of **133 909** ligands) and contained one or more different metal atoms [http://metallo.scripps.edu/index.html#about_mdb; on 16th June, 2003].

B.2.26. Histone Sequence Database
(http://genome.nhgri.nih.gov/histones/)

Histone Sequence Database is a protein database, which covers histone fold-containing sequences derived from sequence-similarity searches of public databases [S. Sullivan *et al.*, *Nucl. Acids Res. 30* (1) (2002) 341–342].

B.2.27. PKR
(http://pkr.sdsc.edu/html/index.shtml)

PKR (*Protein Kinase Resource*) is a web-accessible database of information about protein kinase family of enzymes. Within this database there are also tools for structural and computational analysis, as well as links toward similar resources and databases. PKR is a database, which integrates molecular and cell information [C. M. Smith *et al.*, *Trends in Biochemical Sciences*, *22*(11) (1997) 444–446].

B.2.28. SciFinder
(http://www.cas.org/SCIFINDER/blast.html)

SciFinder is a comprehensive server of CAS (*Chemical Abstracts Service*), one of the American Chemical Society Divisions. CAS is an organisation founded in 1907, with the goal to collect, analyze and distribute all chemistry, biochemistry or molecular biology-linked articles from more than **40 000** scientific journals, patents, conferences and other documents. Online access is enabled to more than **22** million abstracts, over **21** million organic and inorganic substances, and over **28** million sequences through the SciFinder [http://www.cas.org/SCIFINDER/scicover2.html; on 16th June, 2003]. Searching of majority of nucleotide and protein sequences is possible using BLAST search technology.

C. THREE-DIMENSIONAL (3D) STRUCTURE DATABASES

Besides noumerous protein and nucleotide databases based mainly on sequences or parts of the sequences, there are also many three-dimensional structure databases accessible on the Internet.

C.1. NUCLEOTIDE 3D STRUCTURE DATABASES

C.1.1. NDB
(http://ndbserver.rutgers.edu/NDB/ndb.html)

NDB (*Nucleic Acid Database*) is a three-dimensional database, which collects and distributes structural information about nucleic acids. NDB was founded in 1991, when it contained structures determined by X-ray crystallography. Its related database is **NMR Nucleic Acids**, which collects nucleic acid structures determined by NMR. For every structure within NDB, users can see *NDB Atlas* document, list all coordinates for asymmetric or biological unit in PDB format, list coordinates in *mmCIF* format or see structure using *RasMol* tool. NDB contains an atlas of nucleic acid structures, an archive that contains standards of nucleic acids, information about them, software, coordinates and experimental information as well as depository of nucleic acid crystal structures information (data are imported using *AutoDep Input* tool).

C.1.2. MSD / EMD
(http://www.ebi.ac.uk/msd/)

MSD (*Macromolecular Structure Database*) is a European project for collection, management and distribution of the data about macromolecular structures integrating current database and informatics technologies with a solid core of expertise in structural biology (X-ray crystallography, NMR spectroscopy, electron diffraction and bioinformatics). MSD collaborate closely with the Research Collaboratory for Structural Bioinformatics (RCSB; http://www.rcsb.org/pdb/), who maintain the Protein Data Bank (PDB) in the USA. Structures deposited through the *AutoDep* tool are processed at the EBI and are then passed to the RCSB. [H. Boutselakis *et al.*, *Nucl. Acids Res. 31*(1) (2003) 458–462]. Within MSD database is also **EMD** (*Electron Microscopy Data Base: 3D-EM Macromolecular Structure Database*; http://www.ebi.ac.uk/msd/MSDProjects/IIMS3D_EMdep.html). EMD allows the management, organization and dissemination of data on the structures of biological macromolecules solved by three-dimensional electron microscopy.

C.1.3. Gene3D
(http://www.biochem.ucl.ac.uk/bsm/cath_new/Gene3D/)

Gene3D Gene3D is a database of precalculated structural assignments for genes within whole genomes. The

data are obtained using PSI-BLAST and other protocols [D. Buchan *et al.*, *Nucl. Acids Res. 31*(1) (2003) 469–73] and they are currently stored for **66** genomes available from the GenBank public database (mainly CATH domain structure database) [http://www.biochem.ucl.ac.uk/bsm/cath_new/Gene3D/; on 16th June, 2003].

C.2. PROTEIN 3D STRUCTURE DATABASES

C.2.1. PDB
(http://www.rcsb.org/pdb/)

PDB (*Protein Data Bank*) is a unique worldwide repository for the processing and distribution of 3D biological macromolecular structure data [H. M. Berman *et al.*, *Nucl. Acids Res. 28* (2000) 235–242]. PDB was established at Brookhaven National Laboratories (BNL) in 1971 as an archive for biological macromolecular crystal structures. The archive started with seven structures, and in 1980s and especially 1990s, with the development of new crystalographyc technologies (X-ray diffraction, NMR, cryoelectron microscopy, homology modelling), the number of deposited structures started to grow dramatically. Since 1998 PDB has been maintained by RCSB (*Research Collaboratory for Structural Bioinformatics*). Within one week more than 50 new structures are added in PDB, and inside of this database there are **21 248** structures at the moment [http://www.rcsb.org/pdb/; on 16th June, 2003]. The growth of PDB can be seen in Figs. 11 and 12.

Deposition of data in PDB is enabled through the *AutoDepInput* tool [ADIT; http://pdb.rutgers.ecu/adit/]. Results of a query for some protein can be seen in different formats (HTML, TXT, PDB, mmCIF, *etc.*).

C.2.2. SWISS-3DIMAGE
(http://us.expasy.org/sw3d/)

SWISS-3DIMAGE is an image database, where high quality pictures of biological macromolecules with known three-dimensional structure can be found. The database contains mostly images of experimentally elucidated structures, but also provides views of well-accepted theoretical protein models (Fig. 13). The images are provided in several useful formats; both mono and stereo pictures are generally available [M. C. Peitsch *et al.*, *Trends in Biochemical Sciences, 20* (1995) 82–84].

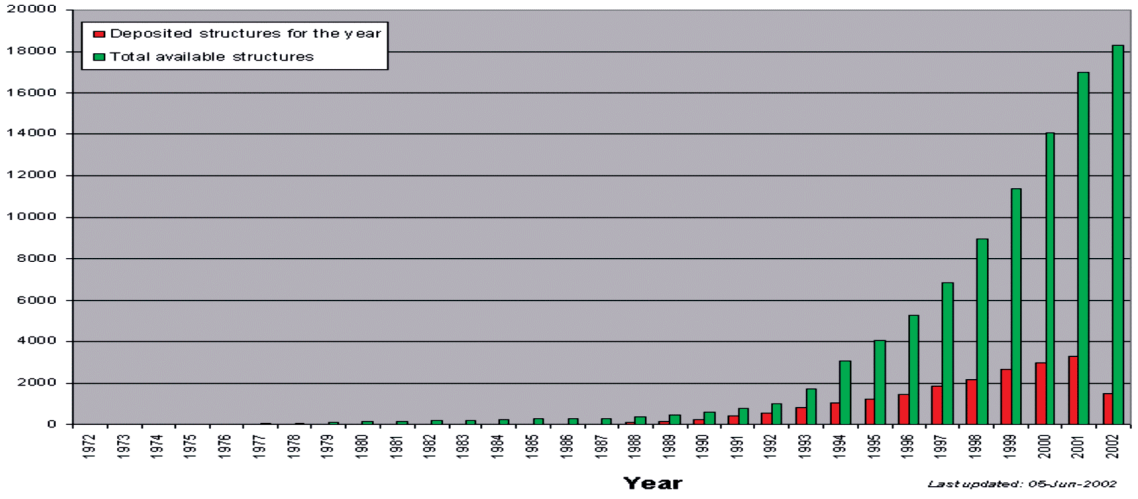


Fig. 11. Growth of PDB content (http://www.rcsb.org/pdb/holdings.html)

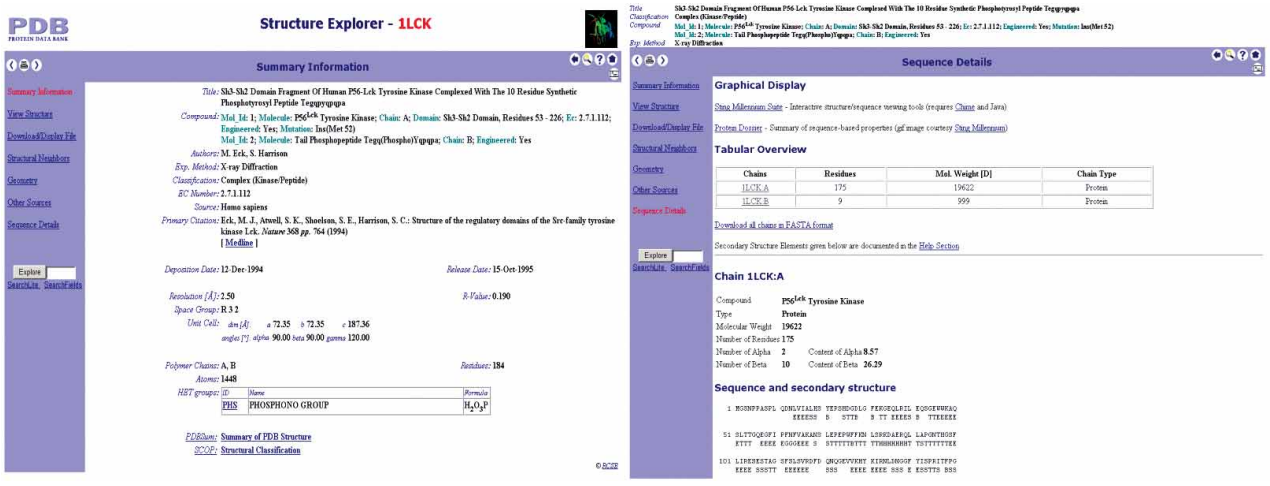


Fig. 12. Entries in PDB (http://www.rcsb.org/pdb/)

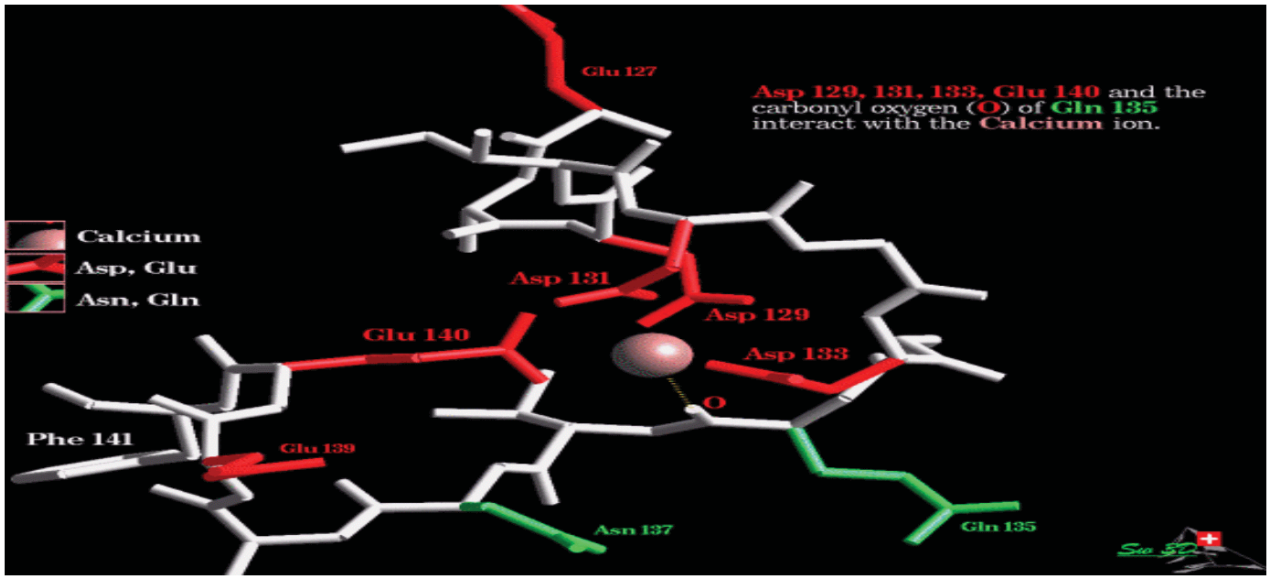


Fig. 13. Example from SWISS-3DIMAGE Database (ftp://us.expasy.org/databases/swiss-3dimage/IMAGES/GIF/S3D00271.gif)

C.2.3. SWISS MODEL REPOSITORY
(http://www.expasy.org/swissmod/SM_3DCrunch_Search.html)

SWISS MODEL REPOSITORY is a database of protein models generated automatically by the SWISS-MODEL server for comparative protein modeling from the University of Basel (Switzerland) [N. Guex and M. C. Peitsch, *Electrophoresis, 18* (1997) 2714–2723; M. C. Peitsch et al., *Pharmacogenomics, 1* (2000) 257–266] [http://swissmodel.expasy.org/]. The Swiss-Model Repository is continuously updated when new sequences or structures become available.

C.2.4. MMDB
(http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml)

MMDB (*Molecular Modeling DataBase)* is a structural database of NCBI, containing experimentally determined three-dimensional biomolecular structures. Most 3D-structure data are obtained by X-ray crystallography and NMR-spectroscopy. MMDB provide information on the biological function, on mechanisms linked to the function, and on the evolutionary history of and relationships between macromolecules [Y. Wang *et al., Nucl. Acids Res. 30*(1) (2002) 249–252].

C.2.5. BIND
(http://www.bind.ca/)

BIND (*Biomolecular Interaction Network Database*) is a database designed to store full descriptions of interactions, molecular complexes and pathways. Here are incorporated virtually all components of molecular mechanisms including interactions between any two molecules composed of proteins, nucleic acids and small molecules. Chemical reactions, photochemical activation and conformational changes can also be described. The database can be used to map pathways across taxonomic branches and to generate information for kinetic simulations [G. D. Bader *et al., Nucl. Acids Res. 29*(1) (2001) 242–245]. At the moment, there are **16 644** interactions, **1 306** complexes and **8** pathways in the database [http://www.bind.ca/index.phtml?page=about; on 16th June, 2003].

C.2.6. DIP
(http://dip.doe-mbi.ucla.edu/)

DIP (*Database of Interacting Proteins*) database catalogs experimentally determined interactions between proteins. It combines information from a variety of sources to create a single, consistent set of protein-protein interactions [I. Xenarios *et al., Nucl. Acids Res. 30*(1) (2002) 303–305]. DIP contains **18 494** interactions, **7 141** proteins and **104** organisms [http://dip.doe-mbi.ucla.edu/dip/Stat.cgi; on 16th June, 2003].

C.2.7. MINT
(http://cbm.bio.uniroma2.it/mint/)

MINT is a database designed to store functional interactions between biological molecules (proteins, RNA, DNA). MINT contains **4 568** interactions, **782** of which are indirect or genetic interactions [A. Zanzoni *et al., FEBS Letters, 513* (2002) 135–140].

C.2.8. BioMagResBank
(http://www.bmrb.wisc.edu/)

BioMagResBank (BMRB) is the publicly-accessible depository for NMR results from peptides, proteins, and nucleic acids recognized by the International Society of Magnetic Resonance and by the IUPAC-IUBMB-IUPAB *Inter-Union Task Group on the Standardization of Data Bases of Protein and Nucleic Acid Structures Determined by NMR Spectroscopy*. The BMRB database contains also amino acid sequence information and data describing the source of

the protein and the conditions used to study the protein [B. R. Seavey *et al.*, *J. Biomol. NMR, 1* (1991) 217–236]. BMRB is in collaboration with the Protein Data Bank (PDB, Brookhaven National Laboratories) and Nucleic Acid Data Bank (NDB, Rutgers University) and at the moment contains over **94 000** unique chemical shifts including more than **6 000** ^{13}C shifts and **4 000** ^{15}N shifts [http://www.bmrb.wisc.edu/index.html; on 16th June, 2003].

C.2.9. ModBase

(http://pipe.rockefeller.edu/modbase-cgi/index.cgi)

ModBase (*Database of Comparative Protein Structure Models*) is a queriable database of annotated protein structure models. The models are derived by *ModPipe*, an automated modeling pipeline relying on the programs *PSI-BLAST* and *MODELLER* [R. Sánchez *et al.*, *Nucl. Acids Res. 28* (2000) 250–253]. The database also includes fold assignments and alignments on which the models were based. ModBase contains theoretically calculated models, which may contain significant errors, not experimentally determined structures. The database contains **837 698** reliable models for domains in **415 937** proteins [http://alto.rockefeller.edu/modbase-cgi/index.cgi; on 16th June, 2003].

C.2.10. HSSP

(http://www.sander.ebi.ac.uk/hssp/)

HSSP (*Homology derived Secondary Structure of Proteins*) is a database of known three-dimensional structures of proteins derived by sequence homology methods. For each known protein 3D structure from the Protein Data Bank [http://www.rcsb.org/pdb/] the derived database contains the aligned sequences, secondary structure, sequence variability and sequence profile. Tertiary structures of the aligned sequences are implied, but not modeled explicitly [C. Sander and R. Schneider, *Proteins, Structure, Function & Genetics*, *9* (1991) 56–68]. At the moment, there are **19 745** entries in the database [http://srs.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-page+LibInfo+-id+7LA1T1KVT7h+-lib+HSSP; on 16th June, 2003].

C.2.11. CATH

(http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html)

CATH is a hierarchical classification database of protein domain structures, which clusters protein at four major levels, Class (C), Architecture (A), Topology (T) and Homologous superfamily (H) [C. A. Orengo *et al.*, *Structure*, *5*(8) (1997) 1093–1108].

C.2.12. SCOP / DBAli

(http://scop.mrc-lmb.cam.ac.uk/scop/)

SCOP (*Structural Classification of Proteins*) is a database where structural classification of proteins can be found. This database provides a detailed and comprehensive description of the structural and evolutionary relationships of the proteins from PDB. It also provides links to coordinates, images of the structure, interactive viewers, sequence data and literature references for each entry [A. G. Murzin *et al.*, *J. Mol. Biol. 247* (1995) 536–540]. Within SCOP, there are **18 946** PDB proteins and **49 497** Domains [http://scop.mrc-lmb.cam.ac.uk/scop/count.html; on 16th June, 2003].

Highly connected with a SCOP is also **DBAli** Database (*Database of protein structure alignments*; http://guitar.rockefeller.edu/DBAli/), which contains more than **34 000** pairwise structural alignments generated using different alignment methods [M. A. Marti-Renom, *Bioinformatics*, *17* (2001) 746–747; http://guitar.rockefeller.edu/DBAli/; on 16th June, 2003].

C.2.13. TOPS

(http://www3.ebi.ac.uk/tops/)

TOPS cartoons are a schematic view of protein three-dimensional structures in two dimensions, and are used for understanding and manual comparison of protein folds. Users can search the patterns selected from a library of motifs or, alternatively, they can define their own search patterns [D. Gilbert *et al.*, *Bioinformatics*, *15*(4) (1999) 317–326].

C.2.14. MolMovDB

(http://molmovdb.mbb.yale.edu/molmovdb/)

MolMovDB (*Database of Macromolecular Movements*) is a database which describes the motions that occur in proteins and other macromolecules, particularly using movies. These domain movements are explained in terms of the repertoire of low-energy conformation changes that are known to occur in proteins [M. Gerstein and W. Krebs, *Nucl. Acids Res. 26*(18) (1998) 4280–4290].

C.2.15. MUTAPROT

(http://bioinfo.weizmann.ac.il/cgi-bin/MutaProt/mutations0.cgi)

MUTAPROT is a database with comparison of PDB files which differ by point mutations. It is created in Weizmann Institute of Science, and there is a full list of protein pairs different in only one or two amino acids. Furthemore, amino acid neighbors and atomic contacts of the mutated residues can be identified here, which can be visualized using *CHIME* (tool for 3D visualization) [Eyal *et al.*, *Bioinformatics*, *17* (2001) 381–382]. There are **17 668** mutations at the moment in the database [http://bioinfo.weizmann.ac.il/cgi-bin/MutaProt/mutations0.cgi; on 16th June, 2003].

C.2.16. DSDBASE

(http://www.ncbs.res.in/~faculty/mini/dsdbase/dsdbase.html)

DSDBASE is a database of disulphide bonds in proteins that provides information on native disulphides and those that are stereochemically possible between pairs of residues in a protein. The modeling of disulphides has been achieved by using the program called MODIP (*Modeling of Disulphide bonds in Protein*)

[R. Sowdhamini *et al.*, *Protein Engineering*, *3* (1989) 95–103].

C.2.17. PhosphoBase
(http://www.cbs.dtu.dk/databases/PhosphoBase/)

PhosphoBase is a revised database of phosphorylation sites in proteins provided by the Center for Biological Sequence Analysis (CBS). Information about phosphorylated residues (the position of phosphorylated serines, threonines, or tyrosines) and relevant kinetic parameters are presented here, as well as data about peptide phosphorylation by a variety of protein kinases [A. Kreegipuu *et al.*, *Nucl. Acids Res. 27*(1) (1999) 237–239]. The data are collected from literature and compiled into a common format. The current version has more than **400** protein entries containing more than **1 400** individual phosphorylation sites [http://www.cbs.dtu.dk/databases/PhosphoBase/relnotes.html; on 16th June, 2003].

C.2.18. GLYCOPROTEIN DATABASES

C.2.18.1. GlycoSuiteDB
(http://www.glycosuite.com/)

GlycoSuiteDB is a relational database of glycan structures, which contains most published O-linked glycans, and N-linked glycans in the literature from years 1990–2002. For each structure, information is available concerning the glycan type, linkage and anomeric configuration, mass and composition. Detailed information is provided on native and recombinant sources, including tissue and/or cell type, cell line, strain and disease state. Where known, the proteins to which the glycan structures are attached are described, and cross-references to SWISS-PROT/TrEMBL are given if applicable. The database annotations include literature references, which are linked to PubMed [C. A. Cooper *et al.*, *Nucl. Acids Res. 29*(1) (2001) 332–335]. At the moment, there are more than **8 100** entries in the database [http://www.glycosuite.com/; on 16th June, 2003]. GlycoSuiteDB is free for non-profit organizations, and commercially available through license agreement.

C.2.18.2. O-GlycBase
(http://www.cbs.dtu.dk/databases/OGLYCBASE/)

O-GlycBase is a revised database of O- and C-glycosylated proteins provided by Center for Biological Sequence Analysis (CBS) of Technical University in Denmark. The criteria for inclusion are at least one experimentally verified O- or C-glycosylation site. Each entry contains information about the glycan involved, the species, sequence, a literature reference and http-linked cross-references to other databases [R. Gupta *et al.*, *Nucl. Acids Res. 27*(1) (1999) 370–372]. According to the latest release, there are **242** proteins and **2 413** verified O-glycosylation sites in the Database [http://www.cbs.dtu.dk/ databases/OGLYCBASE/Changes.html; on 16th June, 2003].

C.2.19. TTD
(http://xin.cz3.nus.edu.sg/group/ttd/ttd.asp)

TTD (*Therapeutic Target Database*) provides information about the known and newly proposed therapeutic protein and nucleic acid targets, the targeted disease, pathway information and the corresponding drugs/ligands directed at each of these targets. In this database there are links to relevant databases that contain information about the function, sequence, 3D structure, ligand binding properties, enzyme nomenclature and related literatures of each target [X. Chen *et al.*, *Nucl. Acids Res. 30* (2002) 412–415]. This database currently contains **433** targets and **809** drugs/ligands [http://xin.cz3.nus.edu.sg/group/ttd/ttd.asp; on 16th June, 2003].

D. CONCLUSION

Bioinformatic infrastructure is one of the most important conditions for successful biomolecular or molecular biology and biochemical research. These fields of research have become impossible without high-quality information about macromolecular structures. Completed and updated genomic and protein databases are the key for this research.

The quality of a database depends on the quality of the data it contains. Evidently, it is almost impossible that the biggest databases and data depositories are maintained by smaller groups of people: databases are better and more complete if greater number of people and groups are included in creating, filling and maintaining the database. One of the best examples of successful collaboration of databases is certanly EMBL – GenBank – DDBJ worldwide database collaboration. The organisers of smaller specialized databases must achieve some critical mass of information, which can move the process of functionalization of their databases, as well as their real benefit in scientific or commercial point of view.

Finally, it is definitely a great challenge to find new and better ways to analyze and evaluate molecular biology, biochemical and biomedicinal data. Bioinformatic knowledge about macromolecular databases, developed methods and tools for processing of this huge amount of available data stored within WWW-accessible databases is basic and necessary part of all scientific and applied research and development efforts, for better understanding of human genetic profile and its role in human health and diseases.

Acknowledgements

Many thanks to Prof. Dr. Ivana Weygand Đurašević and Dr. Boris Mildner for their great help and critical assessment of this work. We would like to express our gratitude to the editors of databases for their courtesy and permission to introduce the database logos, pictures, tables and diagrams within this paper.