MIPS: a database for genomes and protein sequences

H. W. Mewes*, D. Frishman, C. Gruber, B. Geier, D. Haase, A. Kaps, K. Lemcke, G. Mannhaupt, F. Pfeiffer, C. Schüller, S. Stocker and B. Weil

GSF-Forschungszentrum für Umwelt und Gesundheit, Munich Information Center for Protein Sequences, am Max-Planck-Institut für Biochemie, Am Klopferspitz 18, D-82152 Martinsried, Germany

Received October 7, 1999; Revised and Accepted October 27, 1999

ABSTRACT

The Munich Information Center for Protein Sequences (MIPS-GSF), Martinsried, near Munich, Germany, continues its longstanding tradition to develop and maintain high quality curated genome databases. In addition, efforts have been intensified to cover the wealth of complete genome sequences in a systematic, comprehensive form. Bioinformatics, supporting national as well as European sequencing and functional analysis projects, has resulted in several up-to-date genome-oriented databases. This report describes growing databases reflecting the progress of sequencing the *Arabidopsis thaliana* (MATDB) and *Neurospora crassa* genomes (MNCDB), the yeast genome database (MYGD) extended by functional analysis data, the database of annotated human EST-clusters (HIB) and the database of the complete cDNA sequences from the DHGP (German Human Genome Project). It also contains information on the up-to-date database of complete genomes (PEDANT), the classification of protein sequences (ProtFam) and the collection of protein sequence data within the framework of the PIR-International Protein Sequence Database. These databases can be accessed through the MIPS WWW server (http://www. mips.biochem.mpg.de).

DESCRIPTION

The database of the *Arabidopsis thaliana* genome (MATDB)

In recent years the unobtrusive crucifer plant *A.thaliana* has become a well-established model plant due to its low demands, short life cycle and tightly packaged genome. Because of small evolutionary distance among flowering plants, the data and experimental results obtained from *Arabidopsis* are also valid for other plants. Therefore, the molecular and genetic repertoire of *Arabidopsis* is considered a fundamental toolbox for plant genomes.

The genome size of *A.thaliana* is estimated at 120 Mb. The gene density is high compared to other plant genomes, e.g. on chromosome 4, one gene is encoded per 4.2 kb on average. Overall, approximately 25 000 genes are expected to be encoded by the *Arabidopsis* genome. A large portion of them (~90%) code for proteins which have not yet been characterized from *Arabidopsis* and >50% do not have close homologues in other organisms.

The overall goal of the international *Arabidopsis* Genome Initiative (AGI) is to finish the sequence of the whole *Arabidopsis* genome by mid-2000 (1). Within this effort, sequencing and analysis of *Arabidopsis* chromosome 4 has recently been completed by a collaborative approach of the European ESSA consortium, Cold Spring Harbor Laboratory and Washington University. MIPS is responsible for data acquisition, analysis and the compilation of the chromosome sequence. Sequences were subjected to an extensive analysis using several advanced gene prediction and gene modeling algorithms. The analysis therefore combines several lines of evidence and merges intrinsic and extrinsic analysis data for the crucial process of gene prediction. Data submitted by the collaborating groups were integrated into MATDB. All 3800 genes encoded on chromosome 4 have been characterised by a variety of bioinformatics methods and manually assigned to supervised functional categories using the PEDANT analysis software. Beside the extraction of genes, tRNAs, transposons and repetitive regions are also being analyzed and annotated accordingly.

Depending on the specific interest and query character, we provide distinct paths in which users can navigate to the required information. MATDB allows the data to be browsed by functional motifs or categories or for the presence of signal sequences needed for transit into chloroplasts or mitochondria. Map-oriented queries lead to graphical chromosome overviews containing physical and genetic distances as well as positions of specific markers. This allows the user to navigate in a top-down motion to sub-regions, single clones and, finally, to specific genes.

The yeast genome database (MYGD)

Based upon the genomic structure of *Saccharomyces cerevisiae*, the MIPS Yeast Genome Database (MYGD) (2) provides information about open reading frames (ORFs), RNA-genes and other genetic elements. In addition to features of a specific ORF or element, such as automatically annotated functional properties, homologies, and structures, MYGD displays genetic, biochemical and cell biological knowledge extracted from the literature. Relevant citations and corresponding abstracts are concurrently integrated into the MIPS reference database. Data from several systematic, functional analysis projects, co-coordinated by MIPS (EUROFAN I, SCDEGEN)

*To whom correspondence should be addressed. Tel: +49 89 857 8 2657; Fax: +49 89 857 8 2655; Email: mewes@mips.embuct.org

(3), have been opened to the public in 1999, and the results presented on project specific WWW pages.

A retrieval system, providing information on different mutant phenotypes characterised by the systematic functional analysis projects, has been implemented. MYGD supplies a synopsis of functional descriptions of genetic elements and proteins. Yeast genes are categorised by function, protein complexes, protein classes, mutant phenotypes, interaction patterns and their subcellular localisation. Via the general MYGD search tools, using gene names, systematic codes, accession numbers or free text, detailed information on any particular yeast protein or genetic element can be obtained. Beside tables concerning special topics, the MYGD web pages also offer models for physiological and genetic pathways as well as selected reviews provided by members of the yeast scientific community. MIPS has compiled a number of catalogues which use controlled vocabularies and provide information on the genetic and physiological context of proteins. Compared to last year, the MIPS Complex-Catalogue was extended by another 38 novel complexes. The MIPS Functional Categories Catalogue is now comprised of more than twice the amount of sub- and subsub-categories (total 400) and allows precise standardised functional descriptions of a gene of interest.

MIPS *Neurospora crassa* database

Fungi represent a diverse group of eukaryotes, placed into their own kingdom by modern taxonomists just as plants and animals are. The largest single taxonomic unit of fungi is called Ascomycetes, while the best-studied fungus is the yeast *S.cerevisiae*. However, yeast is not an adequate paradigm even for filamentous fungi who have more genes, larger genomes and more or different developmental, catabolic or anabolic capabilities in a wider ecological range. Yeasts, for example, do not form secondary metabolites. In addition to *S.cerevisiae* and *Schizosaccharomyces pombe* the two filamentous species *N.crassa* and *Aspergillus nidulans* are important genetic models.

Neurospora crassa has served as a model organism for more than 50 years making detailed genetic maps and hundreds of mutants available (4). In a collaboration between US and German laboratories, the entire genome, consisting of the seven chromosomes ranging in size between 4.0 and 10.9 Mb, will be sequenced. The total genome size is ~43 Mb.

MIPS is responsible for analyzing the data, gene modelling and annotating predicted proteins and other genetic features of chromosomes II (4.6 Mb) and V (9.2 Mb) within the German *N.crassa* sequencing project. Annotation of gene products is performed by the PEDANT software (see below). A comprehensive database will be established similar to our approach in the yeast and *Arabidopsis* genome databases.

The Human Information Base (HIB) and the cDNA database of the DHGP

The functional analysis of human genes is currently reflected more by a large (1.5 Mio., Sept. 99) and growing number (up to 40 000 per week) of EST sequences than on cDNA or genomic DNA. By contrast, less than 10 000 annotated human proteins are found in the protein sequence database (PIR-International).

HIB is a database of automatically annotated human gene clusters including a functional classification of human proteins based on systematic homology and pattern analysis. An important prerequisite for this work is a set of data that fulfill high quality criteria. A reliable source of data is the UniGene set (5) from which, after the quality checks, such as the removal of contaminants, non-redundant clusters are formed representing putative human transcripts. These clusters, however, are not systematically assembled or furthermore characterized in UniGene. In the HIB database, each cluster was assembled using the CAP3 program (6). Higher base-quality values for cDNAs or complete cds than for EST sequences were assigned. The current version of HIB contains 64 056 entries assembled from the 50 458 UniGene clusters with at least two members.

The longest ORF of each assembled EST cluster is automatically submitted to PEDANT analysis for the functional and structural characterization of the predicted protein sequence (see below). The data is visualized using a WWW oriented graphical user interface. Similar to MYGD, the database can be accessed by selecting a variety of categories, e.g. keywords, superfamilies, PROSITE patterns, PFAM domains and structural classifications.

Selected views allow for representation of interesting aspects and an easy interpretation of the results. Homologies of the clusters in relation to various other species for instance are represented in tabular form. As expected, the rate of significant matches to other mammalian species is nearly identical (*Bos taurus* 45%, rat 43%, mouse 42%; E-values <1e–35), whereas more distant eukaryotic species display only between 6 and 10% closely related proteins (*C.elegans*, *S.cerevisiae*, *A.thaliana*). For prokaryotes, e.g. *Escherichia coli*, this rate drops to only 2%.

The main goal of the German cDNA-Project, as a part of the German Human Genome Project (DHGP), is the isolation, analysis and application of novel full-length cDNAs. All EST and complete cDNA data are stored in an object oriented database. The user-interface of this database provides a platform-independent way to access the data. Clones which have been completely sequenced and annotated pass several client specific steps until they are released for publication to the public databases.

The PEDANT genome analysis server

At the time of writing, PEDANT (7) contains functional assignments and structure predictions for over 140 000 ORFs from 26 completely sequenced and 25 unfinished genomic sequences. Report pages for individual proteins contain a rich set of automatically generated links to a number of external databases, including the yeast functional catalogue at MIPS, the KEGG metabolic pathway database, the SCOP classification of protein domains, the NCBI Web site, and several protein motif collections (PFAM, PROSITE, BLOCKS) (for references see the present issue of NAR). An advanced DNA viewer, accessible from each report page gives a graphical overview of the contig analysed, shows the location of genetic elements (genes, exons, tRNAs, etc.), locates restriction sites, start and stop codons, allows to zoom to a particular region of interest, inspects the DNA sequence and offers a six-frame translation. The protein viewer visualizes structure predictions and similarity matches found by various search methods.

A new genome comparison page allows queries to be conducted across all genomes analysed by PEDANT. For example, the user can select one PFAM domain from the list of all such domains found in all 51 genomes, and the list of ORFs containing such domains will be produced with the links to the corresponding pages of individual genomes.

Protein sequence homology database (ProtFam)

The ProtFam project is a curated database of homology clusters (protein superfamilies, protein families and homology domains) (8). Classification results are directly copied to the entries of the PIR-International Protein Sequence Database (described elsewhere in this volume).

Homologous proteins with identical domain architecture are classified into protein superfamilies. Highly homologous superfamily members are further clustered into protein families. For family classification, an arbitrary cut-off of 50% sequence identity is used. Regions of local homology within otherwise unrelated proteins are annotated as homology domains. The domain sequence of a protein is represented as a domain feature annotation in the PIR-International Protein Sequence Database, each domain and its representatives are entries in the HOMOL protein sequence database. The latter database contains 32 000 entries as of September 1999.

For every homology cluster (family, superfamily, domain) we provide an integrated view of the biological information, e.g. protein names, EC numbers or keywords. As a powerful tool, approximately 20 000 multiple sequence alignments are on display (4500 superfamily alignments, 15 000 family alignments, 374 domain alignments). For every multiple sequence alignment, access is provided to sequence-based biological information (e.g. domains, sequence motifs, active sites, post-translational modifications).

Collection of protein sequence data within the framework of PIR-International

Support of data input and annotation for the protein sequence database is provided by a groupware system. PrIAn (Protein Input and Annotation), a work-flow system, allowing for the fully automated or semi-automated input of new protein sequences, has been introduced. PrIAn processing starts with relevant information from EBI nucleic acid sequence database entries (9). Then each coding region can be annotated in a manual procedure. This procedure permits the inspection of the original EBI entry and, through a link to the MIPS FastA database, biological information present in homologous entries. Due to the implementation of the PrIAn data input, the PIR-International Protein Sequence database has grown to 142 000 entries as of September 1999.

The MIPS section of the Protein Sequence Database has been migrated to a suite of object-oriented database components based on the commercial OODBMS ObjectStore™. The underlying object model of each database component represents parts of the biological or organizational aspects of data management. Citations of different types are stored in a literature database. This database is a central service at MIPS that it is used by several projects simultaneously. Finally, annotations and canonical sequences are managed in a third database component, the Annotation Database.

The architecture of the software system is based on the layer pattern, realising different levels of abstraction. Databases are located at the bottom. The management of persistent storage and objects is realised by another layer. On top, access to a database component is provided by an interface layer containing services provided by servers. To achieve database interoperability, a generic communication layer is part of the MIPS infrastructure. As communication technology, CORBA and a proprietary RPC mechanism, are used with common programming languages.

MIPS internet resources

All projects described are accessible through the internet. Up-to-date descriptions providing detailed material as well as links to direct access to the individual project pages are summarised in Table 1.

Table 1.

Project description	**WWW link**
Project overview	http://www.mips.biochem.mpg.de/desc
The database of the *Arabidopsis thaliana* genome (MATDB)	http://www.mips.biochem.mpg.de/desc/thal
The yeast genome database (MYGD)	http://www.mips.biochem.mpg.de/desc/yeast
The *Neurospora crassa* database (MNCDB)	http://www.mips.biochem.mpg.de/desc/neurospora
The Human Information Base (HIB)	http://www.mips.biochem.mpg.de/desc/human
cDNA database of the DHGP	http://www.mips.biochem.mpg.de/desc/cDNA
Protein sequence homology database (ProtFam)	http://www.mips.biochem.mpg.de/desc/protfam
Resource	**WWW link**
Arabidposis thaliana (MATDB)	http://www.mips.biochem.mpg.de/desc/thal
Neurospora crassa (MNCDB)	http://www.mips.biochem.mpg.de/desc/yeast
Yeast genome databases (MYGD)	http://www.mips.biochem.mpg.de/desc/neurospora
Annotated human EST-clusters (HIB)	http://www.mips.biochem.mpg.de/proj/human
Database of complete cDNAs (DHGP)	http://www.mips.biochem.mpg.de/desc/cDNA
Complete genomes (PEDANT server)	http://pedant.mips.biochem.mpg.de/
Protein sequence database (PIR-International)	http://www.mips.biochem.mpg.de/proj/pir_int
Protein sequence homology database (ProtFam)	http://www.mips.biochem.mpg.de/proj/protfam

How to contact MIPS

Munich Information Center for Protein Sequences, GSF-Forschungszentrum, Max-Planck-Institute for Biochemistry, D-82152 Martinsried, Germany; Tel: +49 89 8578 2656; Fax: +49 8578 2655; Email: w.mewes@gsf.de

ACKNOWLEDGEMENTS

This work was supported by the Federal Ministry of Education, Science, Research and Technology (BMBF, FKZ 03311670, 01KW9703/7), the Max-Planck-Society and the European Commission (BIO4-CT96-0110, 0338, 0558).

REFERENCES

1. Bevan,M., Bancroft,I., Mewes,H.W., Martienssen,R. and McCombie,R. (1999) *Bioessays*, **21**, 110–120.
2. Dolinski,K., Ball,C.A., Chervitz,S.A., Dwight,S.S., Harris,M.A., Roberts,S., Roe,T., Cherry,J.M. and Botstein,D. (1998) *Yeast*, **14**, 1453–1469.
3. Oliver,S.G. (1997) *Curr. Opin. Genet. Dev.*, **7**, 405–409.
4. Radford,A. and Parish,J.H. (1997) *Fungal. Genet. Biol.*, **21**, 258–266.
5. Schuler,G.D. (1997) *J. Mol. Med.*, **75**, 694–698.
6. Huang,X. (1996) *Genomics*, **33**, 21–31.
7. Frishman,D. and Mewes,H.W. (1997) *Trends Genet.*, **13**, 415–416.
8. Barker,W.C., Pfeiffer,F. and George,D.G. (1996) *Methods Enzymol.*, **266**, 59–71.
9. Stoesser,G., Tuli,M.A., Lopez,R. and Sterk,P. (1999) *Nucleic Acids Res.*, **27**, 18–24. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 19–23.