Nucleic Acids Research, 2004, Vol. 32, Database issue ***D41–D44***
DOI: 10.1093/nar/gkh092

MIPS: analysis and annotation of proteins from whole genomes

H. W. Mewes[1,2,*], C. Amid[1], R. Arnold[1], D. Frishman[2], U. Güldener[1], G. Mannhaupt[2], M. Münsterkötter[1], P. Pagel[1], N. Strack[2], V. Stümpflen[1], J. Warfsmann[1] and A. Ruepp[1]

[1]Institute for Bioinformatics (MIPS), GSF National Research Center for Environment and Health, Ingolstaedter Landstrasse 1, D-85764 Neuherberg, Germany and [2]Technische Universität München, Chair of Genome Oriented Bioinformatics, Center of Life and Food Science, D-85350 Freising-Weihenstephan, Germany

Received September 15, 2003; Revised and Accepted October 7, 2003

ABSTRACT

The Munich Information Center for Protein Sequences (MIPS-GSF), Neuherberg, Germany, provides protein sequence-related information based on whole-genome analysis. The main focus of the work is directed toward the systematic organization of sequence-related attributes as gathered by a variety of algorithms, primary information from experimental data together with information compiled from the scientific literature. MIPS maintains automatically generated and manually annotated genome-specific databases, develops systematic classification schemes for the functional annotation of protein sequences and provides tools for the comprehensive analysis of protein sequences. This report updates the information on the yeast genome (CYGD), the *Neurospora crassa* genome (MNCDB), the database of complete cDNAs (German Human Genome Project, NGFN), the database of mammalian protein–protein interactions (MPPI), the database of FASTA homologies (SIMAP), and the interface for the fast retrieval of protein-associated information (QUIPOS). The *Arabidopsis thaliana* database, the rice database, the plant EST databases (MATDB, MOsDB, SPUTNIK), as well as the databases for the comprehensive set of genomes (PEDANT genomes) are described elsewhere in the 2003 and 2004 NAR database issues, respectively. All databases described, and the detailed descriptions of our projects can be accessed through the MIPS web server (http://mips.gsf.de).

INTRODUCTION

MIPS develops and maintains automatically generated and manually annotated genome-specific databases, develops systematic classification schemes for the functional annotation of protein sequences, and provides tools for the comprehensive analysis of protein squences. An overview of the general organization and annotation of genome-related information at MIPS is shown in Figure 1.

FUNGAL MODEL ORGANISMS: THE COMPREHENSIVE YEAST GENOME DATABASE (CYGD) AND THE MIPS *NEUROSPORA CRASSA* DATABASE (MNCDB)

The MIPS yeast genome database is developed and maintained by a group of European databases and yeast laboratories forming a decentralized network of expertise in order to provide detailed information on protein-coding sequences and other genetic elements. Representing the best investigated eukaryote, the database is organized by complementary classifiers with the aim of allowing for the interpretation of functional relations between genes and their corresponding proteins. For instance, the Functional Catalogue (FunCat), providing a systematic classification of protein function is intensively used to project functional data such as expression profiles onto known or probable functional units. Manual FunCat classifications are available not only for yeast, but also for other MIPS-curated genomes such as *Arabidopsis thaliana*, *N.crassa* and the human genome (1). In addition, the set of proteins represented in the PEDANT genomes database was assigned to FunCat classes; the complete list of the assignments is accessible (see Table 1). Within each functional class, the sequences have been clustered into disjoint homology-based subsets.

The Catalogue of Protein–Protein Interactions, the Protein Complex Catalogue and the Protein Localization Catalogues allow information related to the interaction of proteins in yeast to be obtained. More than 10 600 protein–protein interaction records (~9100 physical, ~1500 genetic) were compiled from published large-scale experiments and the literature. The annotated protein complexes (>1000) can be split into ~87 000 putative binary interactions. The vast majority of the records are documented by PubMed reference IDs and by information on the nature of the experimental evidence, which correlates with the confidence of the assignment used in probabilistic computations.

Detailed information on transport proteins [Yeast Transport Protein DB (2)], transcription factors and their binding sites

*To whom correspondence should be addressed at Institute for Bioinformatics (MIPS), GSF National Research Center for Environment and Health, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany. Tel: +49 89 3187 3580; Fax: +49 89 3187 3585; Email: w.mewes@gsf.de

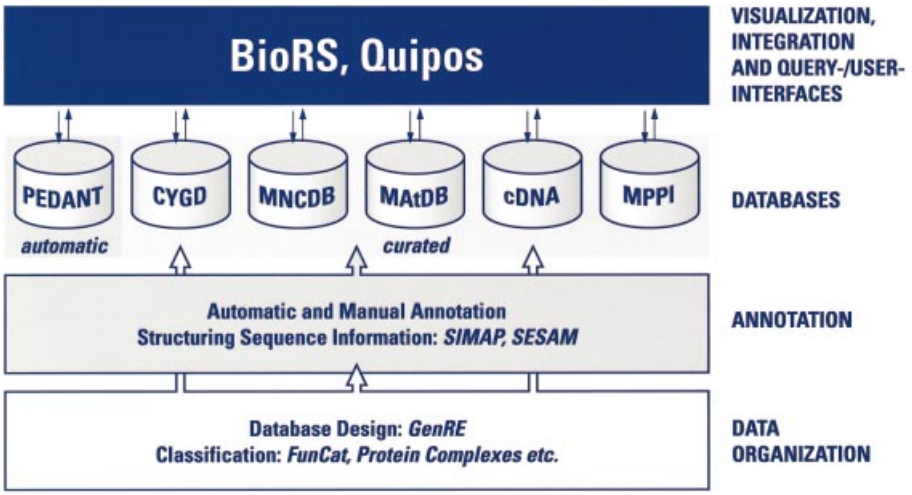

Figure 1. MIPS schema for the organization and annotation of genome-related information.

Table 1. URL addresses for MIPS database resources

Project description	
Project overview	http://mips.gsf.de/desc
Arabidopsis thaliana genome (MATDB)	http://mips.gsf.de/desc/thal/
Comprehensive Yeast Genome Database (CYGD)	http://mips.gsf.de/desc/yeast/
MIPS *Neurospora crassa* Database (MNCDB)	http://mips.gsf.de/desc/neurospora/
Database of Human cDNAs (DHGP)	http://mips.gsf.de/desc/cDNA/
GABI: Genomanalyse im Biologischen System Pflanze	http://mips.gsf.de/desc/gabi/
Helmholtz Network for Bioinformatics	http://mips.gsf.de/projects/hnb
Resource	
Yeast Genome	http://mips.gsf.de/genre/proj/yeast/index.jsp
Neurospora crassa	http://mips.gsf.de/proj/neurospora/
Database of human cDNAs (DHGP)	http://mips.gsf.de/projects/cdna
Complete Genomes (PEDANT server)	http://pedant.gsf.de/
MPPI: Mammalian Protein–Protein Interactions	http://mips.gsf.de/proj/mppi/
Protein Sequence Database (PIR-International)	http://mips.gsf.de/proj/protseqdb/
Helmholtz Network for Bioinformatics	http://mips.gsf.de/proj/hnb/
QUIPOS	http://mips.gsf.de/proj/hnb/quipos/
SIMAP	http://mips.gsf.de/proj/simap/
GenRE: Genome Research Environment	http://mips.gsf.de/genre/proj/
Arabidopsis thaliana	http://mips.gsf.de/proj/thal/
MOsDB	http://mips.gsf.de/proj/rice/
SPUTNIK	http://mips.gsf.de/proj/sputnik/
GABI: Genomanalyse im Biologischen System Pflanze	http://mips.gsf.de/proj/gabi/

[TRANSFAC (3)] and metabolic pathways are either part of the core yeast database or can be retrieved using the BioRS data integration system. To be able to represent complex data of fungal genomes, we use the Genome Research Environment (GenRE) as our annotation data structure. GenRE allows combination of information on different classes of genetic elements and their relations, such as protein–protein interactions or common regulatory features; it provides annotation as well as flexible data retrieval interfaces.

Related proteins from other species can be retrieved using the precomputed SIMAP database (see below) but also using the integrated SESAM tool [Seed Extraction Sequence Analysis Method (4)]. SESAM was developed to achieve better selectivity and sensitivity for the characterization of proteins on a large scale without being dependent on secondary data collections, such as InterPro (5). The selectivity and sensitivity particularly addresses the challenging 'twilight zone' of <30% overall pairwise sequence identity. SESAM does not require the manual adjustment of parameters and copes well with different cases of highly conserved as well as distantly related homologues. A subsequent clustering step starts from SESAM seed-based alignments and leads to 'SESAM feature clusters'.

In CYGD, manually annotated genomes are interlinked via BioRS with the PEDANT analysis of recently published full genomes as well as the 13 hemiascomycetous yeasts, generated by the Génolevures I project (6). An up-to-date compilation of the *Saccharomyces cerevisiae* introns and the analysis of introns in seven related species can be accessed through the 'Hemiascomycetous Yeast Spliceosomal Introns' view (7).

Comparative analysis of the *S.cerevisiae* chromosomes is enabled by a graphical display of the fungal orthologues. The integrated complete genomes include: *Schizosaccharomyces pombe*, *Candida albicans*, *Saccharomyces bayanus*, *Saccharomyces castellii*, *Saccharomyces kluyveri*, *Saccharomyces kudriavzevii*, *Saccharomyces mikatae*, *Saccharomyces paradoxus* [Whitehead Genome Center (http://www-genome.wi.mit.edu/) and George Washington University, St Louis (http://www.genetics.wustl.edu/)], *Candida glabrata*, *Debaryomyces hansenii*, *Kluyveromyces lactis*, *Yarrowia lipolytica* (Génolevures II, http://cbi.labri.u-bordeaux.fr/Genolevures), as well as the genomes annotated at MIPS: *N.crassa* (MNCDB), *Magnaporthe grisea*, *Aspergillus nidulans*, *Fusarium graminearum* (FGDB) and *Ustilago maydis*. Further genomes will be added to enable a comprehensive comparative fungal data resource.

The recently annotated genome of the filamentous fungus *N.crassa* is based on data from the German *Neurospora* Sequencing Project (Chromosomes II and V) (8) and the whole genome sequence, assembled by the Whitehead Genome Center, Cambridge, MA in 2002 (9). In a collaborative effort with the Whitehead group, the MIPS group has annotated the complete genome including manually supervised gene modeling and functional classification of the proteins encoded. The genome of ~40 Mb encodes ~10 000 proteins automatically predicted by the program FGENESH (http://softberry.com), specifically trained for *Neurospora*. The manual inspection of the gene models included intrinsic and extrinsic information such as comparison with known proteins and ESTs as well as splicing consensus signals. Protein sequences were subsequently submitted to the comprehensive analysis of the functional and structural attributes. All information is available at the *Neurospora* project page (Table 1).

THE MIPS HUMAN CDNA DATABASE

With the draft sequence of the human genome in hand, attention has focused on the identification of the complete set of its genes and gene function, respectively. In order to complete this task, combinations of *ab initio* gene predictions, mapping of full length cDNAs with the genomic sequence and comparative genome analysis are widely applied methods.

Since no approach to the prediction of human genes from the genome has returned satisfactory results, sequencing full length cDNAs provides an essential source of information, in particular since it allows elucidation of the structure of alternative splice variants which are thought to be a basis for the complexity of human and other higher eukaryotes. cDNA clones can also represent non-coding RNAs and may contain regulatory elements.

The German cDNA consortium started its work in 1997 as part of the German Human Genome Project (DHGP/NGFN) to release completely sequenced novel cDNAs from various human tissues (10). During this period, several cDNA libraries from so far uncharacterized tissues have been constructed and sequenced. From these libraries 182 543 ESTs (102 396 966 bp) have been generated and analysed from independent clones, and 9380 complete cDNAs (31 187 876 bp) have been identified. At MIPS, the sequence data are automatically analysed and subsequently subjected to several steps of manual annotation and curation, including their functional classification. All sequences are submitted to the public DNA data repositories; the sequences and their annotations are accessible via the MIPS website (Table 1).

Comparative analysis of the human genome data and closely related species, in particular the great apes, additionally improves the quality of predicted genes and allows the discovery of yet unidentified genes. At the beginning of 2003 analysis of the transcriptome of orangutan (*Pongo pygmeus*) as a model organism was initiated by the German cDNA consortium. So far, 27 813 ESTs (14 439 916 bp) and 578 completely sequenced cDNAs (1 548 893 bp) derived and sequenced from different tissues of *P.pygmeus* have been stored in the MIPS database. Comparative sequence analysis from at least three primate species (human, chimpanzee and orangutan) can provide insights into human evolution and will help to find the genetic changes in the human lineage that count for unusual traits such as bipedalism and large brain (11).

The German cDNA data set has been made available to the H-Invitational initiative organized by JBIRC (Japan Biological Information Research Center) and DDBJ (DNA Data Bank of Japan) to assemble a single transcriptome database to overcome present inconsistencies between various databases such as diverse nomenclature or insufficient annotation and to remove redundancies from the data set.

MPPI: A DATABASE OF MAMMALIAN PROTEIN–PROTEIN INTERACTIONS

Protein–protein interactions (PPIs) represent a pivotal aspect of protein function. Almost every cellular process relies on transient or permanent physical binding of two or more proteins in order to accomplish its. The importance of protein–protein interactions is reflected by the recent popularity of experimental techniques such as co-immunoprecipitation, the yeast two-hybrid system, large-scale co-purification and identification of binding partners by mass spectroscopy. Accordingly, comprehensive databases of PPI in *S.cerevisiae* (see CYGD above) have proved to be invaluable resources for various predictive methods applied to experimental data (12). Although yeast is a well established model organism, not all interactions in higher eukaryotes have equivalent counterparts in unicellular model systems. Although current databases include some information on PPI in mammals the vast majority of data comes from microorganisms.

In order to fill this critical gap we have started a database of high-quality protein interaction data from mammals. Expert curators are building the database by harvesting experimental evidence about PPI from the publicly available literature. In contrast to high-throughput data the results of carefully performed individual experiments are considered to be more reliable, especially if multiple independent evidence is presented. Currently, our database contains ~1600 entries of experimental evidence for PPI, which have been integrated in the mouse database of the PEDANT genome information system. A comprehensive user interface for the database is under development. We are in the process of complementing the manually curated data with data from external sources such as mammalian high-throughput experiments. Despite being at an early stage, this database currently represents, according to our knowledge, the single largest publicly available collection of high-quality PPI data in mammals.

SIMAP: A DATABASE OF HOMOLOGY SCORES BASED ON PRECALCULATED EXHAUSTIVE SIMILARITY SEARCHES

Pairwise similarity comparison remains the most powerful tool in genome analysis. Individual searches for homologues do not allow structuring of the sequence universe. Also, since most of the searches are performed with known query sequences, similarity scores stored in an up-to-date all-against-all matrix constitute a very valuable data collection for the systematic analysis of genomes. This set can be used easily to explore interesting genome features such as neighbourhood relations, taxonomic distribution of protein families, etc. Postprocessing steps are easily applied to extract conserved domain patterns or to identify inconsistencies in genome annotation.

SIMAP (SImilarity MAtrix of Protein Sequences) provides a precalculated all-against-all comparison of the protein sets of over 200 fully sequenced genomes. The similarity searches were carried out using the FASTA (13) package. Several tools are available to analyse large sets of sequences, including the generation of subsets (clusters) through iterative queries. For instance, Markov-Random-Field clustering methods such as MCL (14) can be applied to detect protein families. Subclusters can be subjected to SESAM for the generation of conserved sequence patterns or using standard software to generate multiple alignments [POA2 (15)] or to build Hidden Markov Models (HMMER2.3, University of St Louis). Results can be filtered using various parameters such as taxonomic assignments or sequences carrying certain features such as domains of the SCOP database. SIMAP is being incrementally updated.

QUIPOS: QUERY INTERFACE TO PROTEIN SEQUENCE DATA

The assumption that most sequence queries are performed with sequences that are already part of the database or very closely related to known and annotated sequences is well justified. QUIPOS was introduced as an interface of information present in MIPS databases as well as a PEDANT-like tool (16) for on-the-fly protein sequence analyses. To transfer information from well characterized proteins to their close relatives, already known and annotated sequences have to be identified. Most strategies to find related proteins employ similarity searches against a database of annotated sequences; however, in most cases such sensitive but time-consuming searches can be skipped as a close relative is found.

QUIPOS has implemented its own Fast Similarity Sequence Search method (F3S). Closely similar sequences are detected and pairwise aligned to the query sequence by F3S in a few seconds. In a further step, selected information coming from MIPS in-house databases, corresponding to the best hit is mapped to the query. In case no closely related sequence is found, a PEDANT session starts a multi-threaded sequence analysis workflow. In its current version, QUIPOS displays primary information such as statistical evaluation of protein properties, best BLAST hits, multiple alignment to homologous sequences and the presence of sequence domains or domain patterns. Additionally, secondary structures and trans-membrane segments are predicted using standard algorithms, and a 3D structure (17) is assigned to the sequence whenever possible.

ACKNOWLEDGEMENTS

This work was supported by the Federal Ministry of Education, Science, Research and Technology (BMBF: BFAM: 031U112C; NGFN: 01KW9710; HNB: 01SF9985), the Deutsche Forschungsgemeinschaft (MNCDB), and the European Commission (BIO4-CT98-0549, QLRI-1999-01333).

REFERENCES

1. Schueller,C. and Fritz,A. (2002) An enhanced human-genome database. *Genet. Eng.*, **22**, 38.
2. Van Belle,D. and Andre,B. (2001) A genomic view of yeast membrane transporters. *Curr. Opin. Cell Biol.*, **13**, 389–398.
3. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
4. Strack,N. and Mewes,H.W. (1999) SESAM: Seed Extraction Sequence Analysis Method. Giegerich, R. and Wingender, E. *Proceedings of the German Conference on Bioinformatics GCB '99*. Computer Science and Biology, Braunschweig, Bielefeld. 4-6-0099, pp. 59–65.
5. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
6. Souciet,J., Aigle,M., Artiguenave,F., Blandin,G., Bolotin-Fukuhara,M., Bon,B., Brottier,P., Casaregola,S., de Montigny,J., Dujon,B. *et al.* (2000) Genomic exploration of the hemiascomycetous yeasts: 1. A set of yeast species for molecular evolution studies. *FEBS Lett.*, **487**, 3–12.
7. Bon,E., Casaregola,S., Blandin,G., Llorente,B., Neuveglise,C., Munsterkotter,M., Guldener,U., Mewes,H.W., Van Helden,J., Dujon,B. *et al.* (2003) Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns. *Nucleic Acids Res.*, **31**, 1121–1135.
8. Schulte,U., Becker,I., Mewes,H.W. and Mannhaupt,G. (2002) Large scale analysis of sequences from *Neurospora crassa*. *J. Biotechnol.*, **94**, 3–13.
9. Galagan,J.E., Calvo,S.E., Borkovich,K.A., Selker,E.U., Read,N.D., Jaffe,D., FitzHugh,W., Ma,L.J., Smirnov,S., Purcell,S. *et al.* (2003) The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature*, **422**, 859–868.
10. Wiemann,S., Weil,B., Wellenreuther,R., Gassenhuber,H., Glassl,S., Ansorge,W., Bocher,M., Blöcker,H., Bauersachs,S., Blum,H. *et al.* (2001) Toward a catalog of human genes and proteins: sequencing and analysis of 500 novel complete protein coding human cDNAs. *Genome Res.*, **11**, 422–435.
11. Olson,M.V. and Varki,A. (2003) Sequencing the chimpanzee genome: insights into human evolution and disease. *Nature Rev. Genet.*, **4**, 20–28.
12. Vazquez,A., Flammini,A., Maritan,A. and Vespignani,A. (2003) Global protein function prediction from protein–protein interaction networks. *Nat. Biotechnol.*, **21**, 697–700.
13. Pearson,W.R. (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.*, **132**, 185–219.
14. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
15. Lee,C., Grasso,C. and Sharlow,M.F. (2002) Multiple sequence alignment using partial order graphs. *Bioinformatics*, **18**, 452–464.
16. Frishman,D., Albermann,K., Hani,J., Heumann,K., Metanomski,A., Zollner,A. and Mewes,H.W. (2001) Functional and structural genomics using PEDANT. *Bioinformatics*, **17**, 144–157.
17. Westbrook,J., Feng,Z., Jain,S., Bhat,T.N., Thanki,N., Ravichandran,V., Gilliland,G.L., Bluhm,W., Weissig,H., Greer,D.S. *et al.* (2002) The Protein Data Bank: unifying the archive. *Nucleic Acids Res.*, **30**, 245–248.