Nucleic Acids Research, 2006, Vol. 34, Database issue ***D169–D172***
doi:10.1093/nar/gkj148

MIPS: analysis and annotation of proteins from whole genomes in 2005

H. W. Mewes[1,2,*], D. Frishman[2], K. F. X. Mayer[1], M. Münsterkötter[1], O. Noubibou[1], P. Pagel[1], T. Rattei[2], M. Oesterheld[1], A. Ruepp[1] and V. Stümpflen[1]

[1]Institute for Bioinformatics (MIPS), GSF National Research Center for Environment and Health, Ingolstädter 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, 2005; Revised and Accepted October 28, 2005

ABSTRACT

The Munich Information Center for Protein Sequences (MIPS at the GSF), Neuherberg, Germany, provides resources related to genome information. Manually curated databases for several reference organisms are maintained. Several of these databases are described elsewhere in this and other recent NAR database issues. In a complementary effort, a comprehensive set of >400 genomes automatically annotated with the PEDANT system are maintained. The main goal of our current work on creating and maintaining genome databases is to extend gene centered information to information on interactions within a generic comprehensive framework. We have concentrated our efforts along three lines (i) the development of suitable comprehensive data structures and database technology, communication and query tools to include a wide range of different types of information enabling the representation of complex information such as functional modules or networks Genome Research Environment System, (ii) the development of databases covering computable information such as the basic evolutionary relations among all genes, namely SIMAP, the sequence similarity matrix and the CABiNet network analysis framework and (iii) the compilation and manual annotation of information related to interactions such as protein–protein interactions or other types of relations (e.g. MPCDB, MPPI, CYGD). All databases described and the detailed descriptions of our projects can be accessed through the MIPS WWW server (http://mips.gsf.de).

FROM GENE-CENTRIC TO NETWORK-BASED REPRESENTATION OF GENOME INFORMATION

Since the creation of the very first genome databases in 1992 (1), data structures as well as the information content of genome databases underwent little change. Essentially the concept of genome databases is gene centered and the sequence associated information does not reach beyond its individual functional properties such as EC numbers or the annotation of certain classification types such as functional categories [e.g. FunCat (2) or GeneOntology (3)].

An important contribution of novel experimental techniques during the last few years has been the computational analysis of modules and networks based on the combination of independent types of information (4,5). To understand complex cellular processes, it is necessary to uncover the functional context of any individual gene. As a consequence, there is an urgent need to build information resources that enable the integration of different types of data as well as the quantitative evaluation of their reliability at inferring function as part of the annotation process. An essential but so far unsolved challenge for the annotation of gene properties within the functional context is the very dynamic change of the underlying data. In contrast to genome sequence data which are complete for hundreds of species and for which robust and mathematically rigorous algorithms are available, interaction information for most organisms is highly incomplete and the transfer of information between species is not straightforward [e.g. (6)].

From the user's point of view, a number of basic requirements have to be met for genome databases to provide complex functional information. The gene centered view has to be extended from the single-gene to encompass a network perspective and metabolic, regulatory and interactive dimensions have to be included. In addition, to enable an interaction-based comprehensible view, the so called 'giant' networks need to be

*To whom correspondence should be addressed. Tel: +49 89 3187 3580; Fax: +49 89 3187 3585; Email: w.mewes@gsf.de

Table 1. URL addresses for MIPS database resources

Project description	Link
Project overview	http://mips.gsf.de/projects/
Arabidopsis thaliana genome (MATDB)	http://mips.gsf.de/projects/plants/thal
CABiNet: Comprehensive Network Analysis	http://mips.gsf.de/genre/proj/CABiNet/
Complete Genomes (PEDANT server)	http://pedant.gsf.de/
Comprehensive Yeast Genome Database (CYGD)	http://mips.gsf.de/projects/fungi/yeast/
Database of Human cDNAs (DHGP)	http://mips.gsf.de/proj/cdna/Human_cDNA/
FunCat: Functional Catalogue of Proteins	http://mips.gsf.de/projects/funcat/
GABI: Genome analysis in plants	http://mips.gsf.de/projects/plants/projects_html#gabi
GenRE: Genome Research Environment	http://mips.gsf.de/genre/proj/
MIPS *Neurospora crassa* Database (MNCDB)	http://mips.gsf.de/projects/fungi/neurospora/
MosDB: Rice Genome Database	http://mips.gsf.de/projects/plants/rice/
MPPI: Mammalian Protein–Protein Interactions	http://mips.gsf.de/proj/mppi/
SIMAP : Similarity Matrix of Proteins	http://mips.gsf.de/genre/proj/simap/
The Lotus Genome Database (*Lotus japonica*)	http://mips.gsf.de/proj/plants/lotus/
MPCDB Mammalian Protein Complex Data Base	http://mips.gsf.de/genre/proj/mpcdb/
URMELDB (European Medicago and Legume Database (Medicago)	http://mips.gsf.de/projects/plants/medicago/
FGDB: *Fusarium graminearum* genome database	http://mips.gsf.de/genre/proj/fusarium/
MUMDB: *Ustillago maydis* genome database	http://mips.gsf.de/genre/proj/ustilago/
MPACT: Representation of interaction data at MIPS	http://mips.gsf.de/genre/proj/mpact/

subclustered to reflect their underlying modular structure and the edges of the functional interaction graphs must be quantitatively labeled. Both views need to be accessible through browsers as well as through suitable computational interfaces. Obviously, the current state of genome databases does not fulfill these requirements. In this paper, we will describe the current state of our developments to achieve these long-term goals for a limited number of model organisms (fungi including yeast, *Arabidopsis thaliana* and other plant models, and the mouse genome).

GENRE, THE GENERIC MODEL FOR COMPLEX GENOME INFORMATION

Genome information is traditionally stored in databases containing entries as instances of predefined rigid data structures (i.e. formats). However, the development of concepts to cope with complex data structures within a database becomes practically unmanageable as soon as several independent data sources have to be covered. Therefore information pointing to the same biological objects is distributed over a large number of independent and often syntactically incompatible databases (e.g. nucleic acids, proteins, protein interactions, metabolic and regulatory networks and the like). While passive integration of these databases is feasible through database indexing and integration of flat files or web resources [e.g. PubMed (7)], it does not allow for any semantic integration required for comprehensive annotation purposes (Table 1).

The Munich Information Center for Protein Sequences (MIPS) Genome Research Environment System (GenRE) provides a flexible technology to cope with the needs of biological data representation. It is a J2EE based multi-tier architecture, implemented with established software design patterns. Seamless integration of distributed information resources (databases and applications) is realized with Enterprise Java Beans (EJBs) capable of retrieving information in XML format for straightforward web publishing including expression based queries similar to PubMed.

Internally, GenRE is based on three different types of objects and components. Components of the first type are responsible for the access of applications and databases. These EJBs provide a uniform interface while hiding the data resource dependent access mechanisms in the data integration tier. Databases for example are typically accessed via Hibernate, an object-relational mapping tool, whereas applications are often directly accessed. Input and output are commonly XML documents and data objects. Data objects, which represent the second type within GenRE abstract biological entities such as genes, proteins or even complexes at a semantic level. In this layered approach, the data object level is unambiguously separated from the underlying data sources. These objects are used for semantic integration into a third type of component. They are realized as EJBs and are responsible for any further information processing. This allows the association of any biochemical entities (e.g. RNA, drugs, etc.) with either an entity describing binary relationships—e.g. protein interactions from yeast two-hybrid experiments—with many to many relationships, e.g. functional assignments using the MIPS FunCat (2).

Hence GenRE does not only allow for the flexible creation of different object types needed to include various types of 'omics' data, but is also capable of incorporating relations between instances from different data sources. Even complex data models suitable for handling biological networks together with functional annotation of the distinct nodes are realized. In combination with integrated applications like SIMAP components for comparative proteomics can be realized.

The MIPS protein–protein interaction resource (MPact) (8) is illustrative of the advantage of our approach to extend the single-protein view into a network perspective. The data model allows extensions for interactions of proteins with other biochemical entities (e.g. RNA, drugs, etc.). Interacting objects can not only be associated with each other to represent for example complexes, but also with external information describing the corresponding experiments (e.g. yeast two-hybrid, co-immunoprecipitation or mass spectrometry data). In the same way information about the evidence of interactions and various protein annotations such as functions, motifs and

cellular localization are associated with the interacting object. Owing to the object-oriented approach any instances of the interacting object (notably a protein) can be furthermore associated with the corresponding entry e.g. in a genome database.

Our implementation allows two different approaches to query the repository. On the one hand a gene-centric or protein-centric query is possible where distinct interacting objects can be retrieved within a specific context. Since the proteins and interactions are 'decorated' with annotation information, it is possible to query for specific attributes of proteins (e.g. functions) or the interactions (e.g. evidence). On the other hand, network-centric queries can be performed. It is possible to query both for the nodes (the proteins) and the edges (the interactions) of the graph. Based on functional annotation a traversal of the network graph is possible. This traversal can be used to quickly scan the network for false-positive interactions between proteins whose functions and/or localizations differ completely or to assign new functions to proteins without functions which interact specifically within a certain functional context. Furthermore extraction of sub-networks based on any associated context (function, localization, experiment) is possible. It is relevant to point out that our approach enables seamless context dependent views starting from single genes and ending with complete networks. MIPS databases are implemented in the GenRE environment.

SIMAP AND CABINET: COMPUTATIONAL METHODS TO GENERATE INFORMATION FOR GENOME DATABASES

Pairwise similarity comparison of every protein against the set of all known proteins is an indispensable step in any annotation process. Many biological and evolutionary questions are related to the structure of the sequence space and its partitioning into substructures represented by the all-against-all similarity relations. However, individual searches for homologs do not allow structuring the sequence universe. The MIPS Similarity Matrix of Proteins (9), currently contains a matrix of all-against-all comparisons of more than 4 million proteins from >400 organisms including all UNIPROT sequences. They have been generated by exhaustive sequence similarity searches using FASTA (10). SIMAP is continuously updated to keep up with the rapidly increasing amount of data. The linked sequence feature database containing information on protein domains as well as predicted transmembrane regions, signal peptides and protein localization supports efficient post-processing of homology lists into sub-clusters of homogeneous sequence properties. These steps are also essential to identify inconsistencies in genome annotation. SIMAP serves as an example of a system offering highly dynamic information in the form of a persistent database to be explored systematically at high performance. SIMAP is also used as an annotation tool and generates similarity input information for the PEDANT databases (11).

In contrast to the straightforward similarity matrix, the compilation of biological interaction information requires methods dealing with mostly incomplete but also inherently heterogeneous sets of data. To provide a system for comprehensive network analysis, we have developed CABiNet (Comprehensive Analysis of Biomolecular Networks) within the framework of GenRE. CABiNet offers a set of methods for network statistics, integration, analysis and clustering applied to interaction data administered by GenRE as well as to user-submitted network data. CABInet allows the user to browse and query across any subset of the generated networks and clusters. As with all methods in GenRE, the software design of CABiNet allows an easy adoption of new functions into the system. To address the issue of inconsistent use of protein identifiers in different networks, a GenRE component to resolve protein identifiers and aliases in all major sequence databases is employed by CABiNet.

THE MAMMALIAN PROTEIN COMPLEX DATABASE

High quality data collections are needed as a reference set for testing computational methods in network analysis. In the case of protein–protein interaction data, any collection from various sources including high-throughput experiments contains a considerable number of false-positive and false-negative results (12). Thus, for the analysis of protein networks there is a need for gold standards (12) in order to validate the quality of data resources or as a reference for testing the reliability of computational methods. The cooperation and interaction of proteins is most unambiguously found in protein complexes where several proteins simultaneously act together to perform a single reaction. For the analysis of protein networks in lower eukaryotes the protein complexes from yeast have become a gold standard in the field (8).

Analysis of the even more elaborate protein networks in mammals requires manually curated resources in its own right. Information in Mammalian Protein Complex Database (MPCDB) is extracted from scientific literature describing individual experiments; data stemming from high-throughput experiments has not been incorporated. Information about protein complexes includes gene names of the members as well as protein names, cross-references and literature references. If a protein complex has been analysed from other mammals the orthologous mouse proteins are presented. In addition, the type of experiment that was used to analyse the protein complex is given. Evidence is structured according to the MIPS evidence catalogue originally developed for yeast protein complex and protein–protein interaction data and subsequently adopted for higher organisms. This is in line with the requirements for PSI-MI compliant annotation (8). Manual annotation of the respective proteins is performed in the Mouse functional Genome Database (MfunGD). Here, protein function annotation is performed with the FunCat (2) annotation scheme, which in a comparison of several individual features showing the highest predictive power for the analysis of protein networks (13). Whereas within the protein complex data in BIND (14) the vast majority of protein complexes contain less than three different proteins, the protein complexes within MPCDB contain 4.6 different proteins on average. Currently, MPCDB contains 122 protein complexes with a total of 643 proteins. High molecular weight protein complexes like proteasomes and the eukaryotic chaperonin TRiC perform central functions in cells like protein degradation and protein folding, respectively. These

two protein complexes are examples that so far, are not present in any manually curated publicly available database of protein complexes. Hence, the MPCDB dataset provide scientists with a reliable data resource for the analysis of protein networks and functional modules in mammals.

SUMMARY

The MIPS data resource aims to extend the scope of its curated genome databases towards functional context information. These databases include model systems for mammals (*Mus musculus*, in progress), fungi (yeast, CYGD; *Neurospora crassa*, MNDB; *Ustillago maydis*, MUMDB; *Fusarium graminearum*, FGDB), plants (*A.thaliana*, MatDB; *Oryza sativa*, MOsDB; Maize Genome Sequencing Project, MGSP; European Medicago and Legume Database; URMELDB, *Lotus japonica*), microorganisms (Chlamydiae, Listeriae), the comprehensive automatic annotation of genomes in the PEDANT database (11). In addition, we have compiled manually curated reference protein–protein interaction datasets for mammalia (MPPI) and yeast (MPact) as well as the database for mammalian protein complexes (MPCDB). Proteins in these databases are functionally classified using the well established functional catalogue FunCat (2).

ACKNOWLEDGEMENTS

This work was supported by the Federal Ministry of Education, Science, Research and Technology (BMBF: BFAM: 031U112C, GABI: 0312270/4, NGFN: 01KW9710) and the Deutsche Forschungs-gemeinschaft (Bioinformatics Munich, BIM). Funding to pay the Open Access publication charges for this article was provided by the GSF—National Research Center for Environment and Health.

Conflict of interest statement. None declared.

REFERENCES

1. Oliver,S.G., van der Aart,Q.J., Agostoni-Carbone,M.L., Aigle,M., Alberghina,L., Alexandraki,D., Antoine,G., Anwar,R., Ballesta,J.P., Benit,P. *et al*. (1992) The complete DNA sequence of yeast chromosome III. *Nature*, **357**, 38–46.
2. Ruepp,A., Zollner,A., Maier,D., Albermann,K., Hani,J., Mokrejs,M., Tetko,I., Güldener,U., Mannhaupt,G., Münsterkötter,M. *et al*. (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res*., **32**, 5539–5545.
3. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet*., **25**, 25–29.
4. von Mering,C., Huynen,M., Jaeggi,D., Schmidt,S., Bork,P. and Snel,B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res*., **31**, 258–261.
5. Kelley,R. and Ideker,T. (2005) Systematic interpretation of genetic interactions using protein networks. *Nat. Biotechnol*., **23**, 561–566.
6. Pagel,P., Mewes,H.W. and Frishman,D. (2004) Conservation of protein–protein interactions—lessons from ascomycota. *Trends Genet*., **20**, 72–76.
7. Goetz,T. and von der Lieth,C.W. (2005) PubFinder: a tool for improving retrieval rate of relevant PubMed abstracts. *Nucleic Acids Res*., **33**, W774–W778.
8. Güldener,U., Münsterkötter,M., Oesterheld,M., Pagel,P., Ruepp,A., Mewes,H.-W. and Stuempflen,V. (2006) MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res*., **34**, D436–D441.
9. Arnold,R., Rattei,T., Tischler,P., Lindner,D., Stuempflen,V. and Mewes,H.W. SIMAP: The similarity matrix of proteins. *Bioinformatics*, **21** Suppl. 2, ii42–ii46.
10. Pearson,W.R. (1994) Using the FASTA program to search protein and DNA sequence databases. *Methods Mol. Biol*., **24**, 307–331.
11. Riley,M.L., Schmidt,T., Wagner,C., Mewes,H.W. and Frishman,D. (2005) The PEDANT genome database in 2005. *Nucleic Acids Res*., **33**, D308–D310.
12. Jansen,R., Lan,N., Qian,J. and Gerstein,M. (2002) Integration of genomic datasets to predict protein complexes in yeast. *J. Struct. Funct. Genomics*, **2**, 71–81.
13. Lu,L.J., Xia,Y., Paccanaro,A., Yu,H. and Gerstein,M. (2005) Assessing the limits of genomic data integration for predicting protein networks. *Genome Res*., **15**, 945–953.
14. Bader,G.D., Betel,D. and Hogue,C.W. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*., **31**, 248–250.