44–48 *Nucleic Acids Research, 1999, Vol. 27, No. 1*

MIPS: a database for genomes and protein sequences

H. W. Mewes*, K. Heumann, A. Kaps, K. Mayer, F. Pfeiffer, S. Stocker and D. Frishman

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

Received October 2, 1998; Revised October 8, 1998; Accepted October 23, 1998

ABSTRACT

The Munich Information Center for Protein Sequences (MIPS-GSF), Martinsried near Munich, Germany, develops and maintains genome oriented databases. It is commonplace that the amount of sequence data available increases rapidly, but not the capacity of qualified manual annotation at the sequence databases. Therefore, our strategy aims to cope with the data stream by the comprehensive application of analysis tools to sequences of complete genomes, the systematic classification of protein sequences and the active support of sequence analysis and functional genomics projects. This report describes the systematic and up-to-date analysis of genomes (PEDANT), a comprehensive database of the yeast genome (MYGD), a database reflecting the progress in sequencing the *Arabidopsis thaliana* genome (MATD), the database of assembled, annotated human EST clusters (MEST), and the collection of protein sequence data within the framework of the PIR-International Protein Sequence Database (described elsewhere in this volume). MIPS provides access through its WWW server (http://www.mips.biochem.mpg.de) to a spectrum of generic databases, including the above mentioned as well as a database of protein families (PROTFAM), the MITOP database, and the all-against-all FASTA database.

DESCRIPTION

The PEDANT genome analysis server

The PEDANT (Protein Extraction, Description, and ANalysis Tool; http://pedant.mips.biochem.mpg.de) genome browser provides a comprehensive and up-to-date analysis of publicly available genomic sequences (1). The software system allows the user to conduct first-pass automatic genome annotation utilizing a whole spectrum of sequence analysis and structure prediction techniques, such as similarity searches, motif analysis, automatic attribution of ORFs to functional categories, secondary structure and membrane region prediction, detection of low complexity and coiled-coil regions, etc. The set of bioinformatics methods used in PEDANT for functional and structural characterization of proteins has been extended. Recently added methods include Hidden Markov Model searches against the database of protein domains (2), prediction of signal peptides in eukaryotic, gram-positive, and gram-negative organisms (3), similarity-based delineation of enzyme EC numbers as well as the calculation of protein molecular weight and pI (F.Lindberg, unpublished).

PEDANT results are split into two sections. One section presents the completed genomes. It is based on the prediction of gene products as supplied by the authors of the corresponding publication. The second section includes unfinished and experimental sequences. Two eukaryotic genomes available in this section are (i) the *Schizosaccharomyces pombe* partial sequence which is being processed in collaboration with the Sanger Center and (ii) the FCA contig of the *Arabidopsis thaliana* genome analysed at MIPS as part of the EU funded ESSA project (see below). Other sequences in this section are unfinished bacterial genomic fragments stemming from different sequencing projects that are currently underway at the Sanger Center, The Institute for Genomic Research, University of Chicago, University of California at Berkeley, Stanford University, Genome Therapeutics Inc., PathoGenesis Corporation and other sequencing centers. Prediction of genes in unfinished bacterial genomic sequences is performed by the ORPHEUS software (4) in a completely automatic fashion.

The total amount of data managed by the PEDANT system has passed 10 gigabytes. In order to ensure fast access to individual results and to allow for frequent updates, the data are stored using the freely available MySQL relational database management system.

The MIPS Yeast Genome Database (MYGD)

Based upon the genomic structure of the first eukaryotic genome sequenced (5,6), the MIPS Yeast Genome Database aims at a comprehensive presentation of information about all open reading frames (ORFs), RNA-genes, and DNA-elements of *Saccharomyces cerevisiae*.

A main challenge of MYGD is the integration of current results from the yeast literature and functional analysis experiments in order to supplement and corroborate the information gained by automatic annotation procedures. This allows for determination of possible sequence deviations as well as the annotation of genetic element features which are not automatically detected, such as snRNAs (80 snRNAs at MIPS as of September 1998) or ORFs smaller than 100 amino acids (133 small ORFs as of September 1998; 7). Moreover, the combined presentation of genetic, biochemical, and cell biological information extracted from the relevant publications enables MYGD to supply a functional description of genetic elements and proteins.

Detailed information on a particular yeast protein or genetic element can be obtained via queries using gene names, systematic

*To whom correspondence should be addressed. Tel: +49 89 8578 2657; Fax: +49 89 8578 2655; Email: mewes@mips.biochem.mpg.de

codes, accession numbers or free text. MYGD harbors 19 413 citations in total, 8141 different publications build the reference lists, and at least one link to a reference is found in the datasets of 4432 ORFs. Manual annotation applies a standardized terminology as much as possible, in addition free text descriptions are provided by internal or external experts, such as the 'Phenotype-', 'Overexpression-' or 'Suppression-Notes'. For example, the ORF YKL145w (http://www.mips.biochem.mpg.de/YKL145) was attributed according to aspects of its function, its protein class, its structural location in a protein complex, and the nature of its interaction with CDC28/RAD23. Links to related 'Schemes' [*] and 'Tables' of MYGD give additional functional information (http://www.mips.biochem.mpg.de/yeast_tables). Moreover, links to COGs at NCBI (at MIPS provided for 1523 ORFs as of September 1998; 8), to PIR-families and superfamilies (9), to related human ESTs, as well as cross-links to other data collections (e.g., YPD; 10) contribute to a comprehensive view regarding this protein and its cellular role.

Users of the database are often driven by special interests for genes sharing common attributes like the ABC transporter family (http://www.mips.biochem.mpg.de/yeast_classes), transcription factors and other functional categories. In addition to the protein related entries, MYGD provides a number of tables and graphics which comprise information about all ORFs sharing a common attribute. Graphics and reviews provided by members of the yeast community are presented. For example, tables list essential genes (626 ORFs) versus genes dispensable for cell viability (1704 ORFs). YTA proteins are graphically illustrated (http://www.mips.biochem.mpg.de/YTA), and alignments of the yeast centromers presented. The interaction tables compile protein–protein interactions; at this time, 2952 physical and genetic interactions between genes/gene products are annotated in MYGD, characterizing 1060 ORFs.

The approach to describe each genetic element individually is limited. Even the homology based classification into families and superfamilies (see below) is unable to account for the complex classification schemes required to describe the functional or cellular context of a protein. The reason for a similar behaviour of proteins that causes an equivalent phenotype in the corresponding deletion strains may be obscure, i.e., cannot be inferred from the sequence-associated information directly. Categorization of entries is achieved by using a standardized terminology to describe attributes of genetic elements. MIPS has compiled a number of such catalogues to provide information on the genetic and physiological context of the proteins. Table 1 summarizes information regarding numbers of main- and sub-categories, and the number of ORFs which have been assigned. The catalogues allow the user to browse yeast proteins according to their affiliation to a functional category, a protein complex, a protein class, a mutant phenotype or a specific subcellular localization. As a starting point in building a compendium of pathways, current models for physiological and genetic pathways were collected and integrated with information of MYGD.

MIPS *Arabidopsis thaliana* Database (MATD)

The unobtrusive crucifer plant *A.thaliana* is widely accepted as a model to study a broad spectrum of plant biology features. Plant scientists appreciate its ease to grow, the short life cycle and the tightly packaged genome. Due to the close relationship among higher plants, the molecular and genetic repertoire of *Arabidopsis* is thought to represent a basic toolbox for plant genomes.

The estimated size of the *Arabidopsis* genome is 100 Mb, eight times larger than that of the first fully sequenced and analyzed eukaryotic genome of *S.cerevisiae*. The gene density is high in contrast to many other plant genomes. From the analysis of long genomic segments an average size of 4.5 kb per gene can be deduced. Approximately 21 000 genes are expected to be encoded and expressed by the *Arabidopsis* genome. A large portion of them (~95%) code for proteins which have not been characterized so far in *Arabidopsis* and more than 50% do not have close homologues in other organisms (11).

To elucidate the information encoded in the *Arabidopsis* genome the *Arabidopsis* Genome Initiative (AGI) was set up. The AGI is a collaboration of six sequencing groups from Europe, the US and Japan. The overall goal is to sequence and analyze the whole *Arabidopsis* genome by the year 2001. The European consortium (ESSA, European Scientists Sequencing *Arabidopsis*) intends to sequence and analyze ~22 Mb. MIPS acts as the bioinformatic co-ordinator of the ESSA project. So far 7.5 Mb of contiguous sequence data from *Arabidopsis* chromosome IV have been sequenced and analyzed. Approximately 1700 genes have been extracted. Gene modeling is performed through a combination of several advanced gene prediction algorithms as well as homology searches against up-to-date databases. Extrinsic and intrinsic evidence is subsequently merged using a customized procedure presented in Figure 1. Extracted genes are extensively characterized using the PEDANT (see above) software and features as PROSITE patterns (12) and 3D structure prediction are assigned to the respective genes. Beside extraction of genes tRNAs, snRNAs and repetitive regions are being analyzed and annotated accordingly.

Table 1. MIPS yeast protein catalogues (as of September 1998)

Yeast protein catalogues	Main categories	Sub-categories	No. of proteins assigned
Functional Categories (FunCat)	16	182	3476[a]
Protein Complexes(CompCat)	58	208	1048
Protein Classes(ClassCat)	22	159	1014
Phenotypes (PhenCat)	11	166	in preparation (1384)
Subcellular Localizations (SubcellCat)	15	26	2210
PROSITE Motifs (15)	493	–	2173
EC numbers	6	579	972

All ORFs of the different catalogues are directly linked to the respective MYGD entries, giving access to annotated genetic, biochemical and structural data.
[a]This figure represents the number of ORFs with defined function attributed. Due to the FunCat category 'unclassified', the catalogue actually comprises all annotated ORFs.

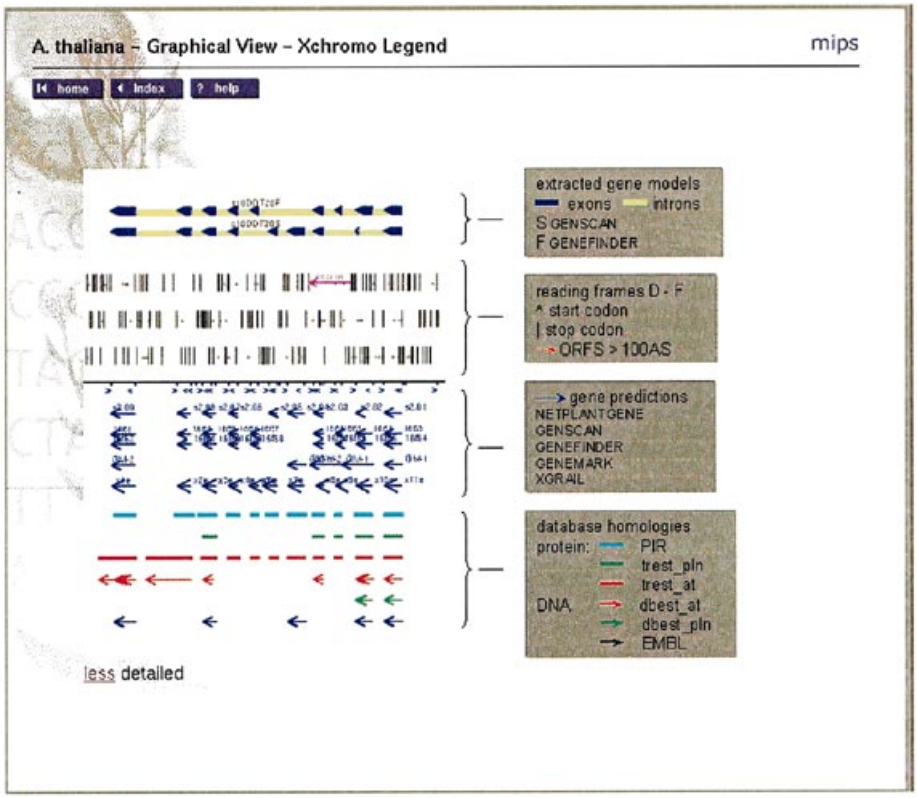

Figure 1. Xchromo summary of analysis results obtained by geneprediction and genemodelling algorithms as well as detection of homology based analysis; combination of extrinsic and intrinsic data (http://www.mips.biochem.mpg.de/gene_model).

Similar to our approach to provide a comprehensive database of the yeast genome, MATD provides distinct paths along which the user can navigate to explore the database information. The standard approach of a search mask returns the annotation related to a specific gene, clone or accession number. In addition we provide dynamically generated lists of genes belonging to certain functional categories or classes. These tables give a first overview on genes belonging to distinct functional or structural categories and are a convenient starting point to explore the genome with respect to a given field of interest.

For map oriented queries, specific regions of the genome are displayed as contigs, mapped to the chromosome. The user navigates in a top-down way from a chromosome overview via subregions to single clones and finally to specific genes (http://www.mips.biochem.mpg.de/arabi/overview.html). The chromosome is subdivided into several zones spanning physical regions from 400 kb to 3 Mb. These regions are delimited by specific markers; markers within the regions are displayed as landmarks.

Contig information on sequenced regions is given either in tabular or graphical format. Users can retrieve the corresponding DNA as well as the protein sequence data. The annotation includes the classification into functional categories whenever a reliable assignment can be done. FASTA similarity scores including pairwise sequence alignments are given. Newly identified *Arabidopsis* proteins are compared to MATDB and all entries of the PIR International Database. The resulting homology database is updated nightly.

In addition to the above described queries based on chromosomal location or sequence attributes, users can submit their sequences to BLAST searches against all ESSA sequences and all *Arabidopsis* entries in GenBank. Integration of all available *Arabidopsis* sequence and protein data to provide a uniform presentation of the whole *Arabidopsis* genome is in progress.

The entry point for all information related to the *A.thaliana* genome is at http://www.mips.biochem.mpg.de/arabi/

The database of human transcripts (MEST)

The sequence of the human genome on the genomic level cannot serve as a reliable source of information on the primary structure of the expressed transcripts due to the yet unsolved problem of a model for RNA splicing. A more direct albeit computationally challenging approach is the reconstruction of spliced DNA transcript sequences from short overlapping RNA-fragments (expressed sequence tags, ESTs). Several putative transcript data sets are available either as unassembled clusters of ESTs and cDNA sequences (13) or as databases of assembled, putative transcripts (TIGR, http://www.tigr.org/tdb/hgi/hgi.html ; Stack, http://ziggy.sanbi.ac.za/stack/abstract.html). These inferred databases have major drawbacks: no systematic analysis of clusters is performed (e.g., functional assignment by homology), no evaluation or estimate of the quality is available, and updates are performed at unspecified intervals. A rigorous comparison between these datasets is very difficult despite of the fact that they are all derived from the same public sources.

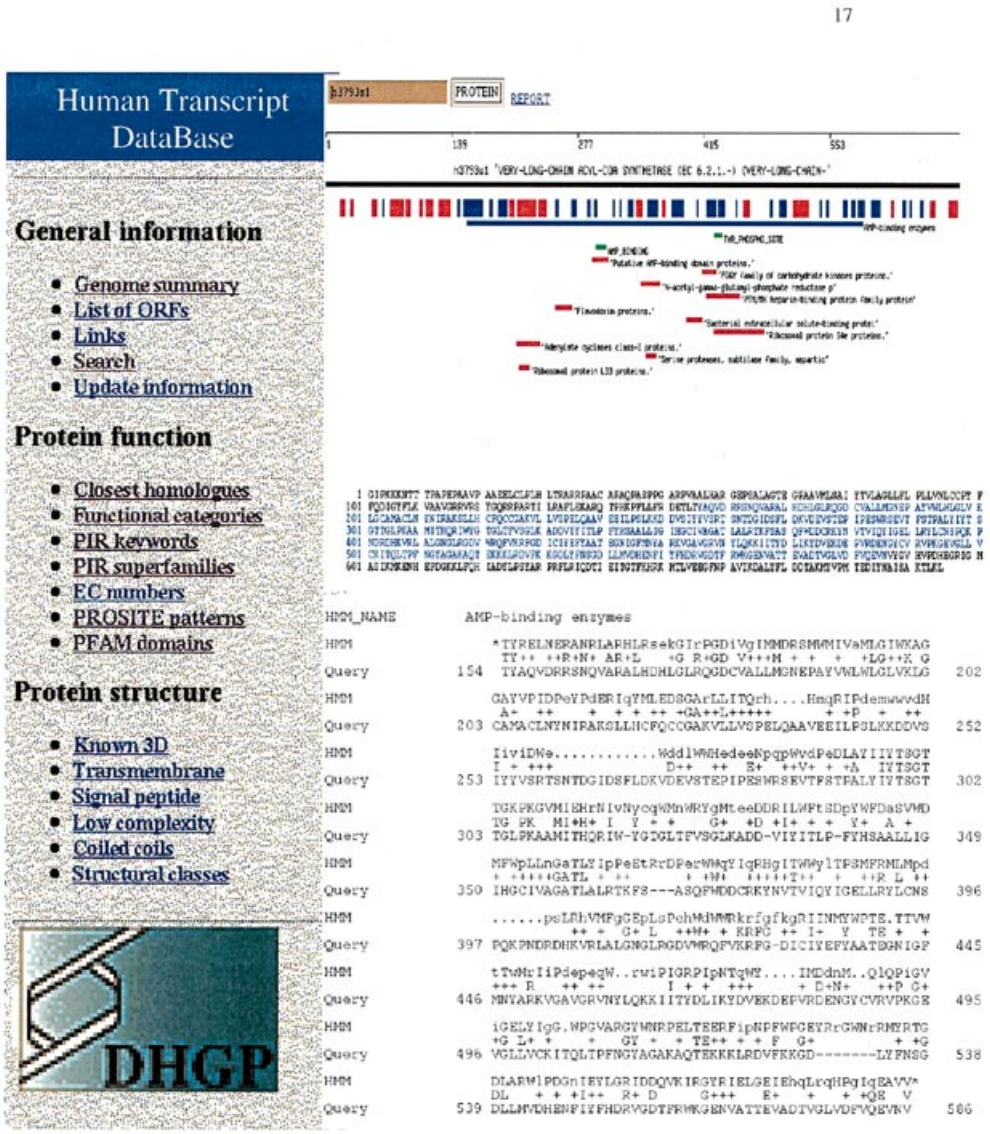

Figure 2. Display of a MEST human transcript database entry (http://www.mips.biochem.mpg.de/PEDANT/human). The left frame allows the selection of functional groups in the database according to the classification scheme provided by PEDANT. The right frame displays the graphical map of annotated features for the selected transcript. Starting from the top the map shows the secondary structure prediction: blue indicates β-sheets, and red indicates α-helices). Below, selected features are displayed as bars with a short description: PFAM domains (3; blue: the appropriate alignment is shown in the lower part of the frame), significant PROSITE patterns (12; green) and BLIMPS (17; red).

The MIPS-DHGP human EST database MEST applies a systematic annotation process to all putative transcripts. In a first approach we have selected the Uni-Gene-clusters (http://www.tigr.org/tdb/hgi/hgi.html) and assembled them using the CAP2 program (14). At the time of writing 46 000 contigs have been generated (http://www.mips.biochem.mpg.de/PEDANT/human). Each cluster in MEST was characterized using the sequence analysis suite PEDANT (1). An annotation map for all putative transcripts was created in addition (Fig. 2). The map allows for visual inspection of multiple features assigned to each sequence. Sequence features are displayed as a graphical map. Users may inspect any feature from the map in detail. This approach makes it easy to verify the consistency of the overall functional characterization of the putative gene product.

The human genome project will head for a detailed, verified annotation of every single gene identified. Information will be generated starting with sometimes unreliable fragmentary information. Nevertheless, careful cluster generation and verification by comparison with well defined genes from model organisms often provides reliable information for further experimental work (e.g., expression profiling). We have experienced that our combination of symbolic representation of sequence features with a comprehensive graphical display allows for an easy validation for manual annotation. Visual inspection of sequence features substantially improves the possibility to ensure high quality annotation and interpretation of automatically generated data.

For rapidly growing databases such as the map of human transcripts the ability to supervise automatic assignments efficiently becomes crucial to ensure a consistent standard of annotation while keeping the database up to date. MEST provides a powerful tool for the rapid retrieval of information as well as the detailed manual annotation of human genes.

Protein Sequence Database and family classification of protein sequences

MIPS is responsible for the collection and annotation of protein sequence data from European sources for the PIR-International Protein Sequence Database. The EBI nucleic acid sequence database serves as a major source of new sequence data. Details of the PIR-International Protein Sequence Database are published elsewhere in this volume. To improve the efficiency of the annotation process, a re-design of the Protein Sequence Database at MIPS has been performed. The design of the database structure is principally different from the flat-file oriented final product that is no longer appropriate for processing increasing amounts of complex biological data. To realize a concept for data management on high level of abstraction we use object-oriented analysis and design methodologies for handling complex heterogeneous information entities. The software system of the new database is based on the layer architectural pattern (15).

The Protein Sequence Database is being migrated to an object-oriented database based on a commercial OODBMS (http://www.odi.com). Following the fundamental principle to break down the complexity into smaller, homogenous structures a set of independent component databases has been implemented. Each component is part of a class hierarchy that can be assembled with a high degree of flexibility (e.g., the information on a particular genome can be extracted from components in protein, nucleic, reference and other databases). This flexibility allows for dynamically increasing the global functionality of the complete system. However, relations between objects residing in different component databases have to be modeled as well.

For realizing the access and data transfer between components of the system the Client/Server model is used. CORBA (Common Object Request Broker Architecture; http://www.omg.org) is used as middleware system decreasing the complexity of developing distributed applications by applying a standard for inter-process communication allowing for database interoperability. Independent from the interactive WWW browser technology, the community will be able to access the Protein Sequence Database using CORBA as far as the IDL (Interface Definition Language) definition has been published.

All protein sequences are classified into protein families and superfamilies and organized in the PROT-FAM database (9). As of September 1998, the PIR-International Protein Sequence Database contains ~117 000 entries, classified into one of the 54 000 protein families. 43 000 families contain a single sequence (37%) as there is no homologue with >50% sequence identity in the database. 6000 families contain two sequences (12 000 entries, 10%) and ~6500 families contain three or more sequences (57 000 entries, 49%). Approximately 1% fragments and 3% very short sequences cannot be classified. For families with two or more members alignments have been computed using the multiple alignment program PILEUP (Genetics Computer Group, 575 Science Drive, Madison, WI 53711, USA). Nearly 70% of all database entries are classified into one of the ~6600 superfamilies. For ~3800 superfamilies which are derived from more than one family and contain more than one sequence, multiple alignments have been built. The PROT-FAM database as the source for the protein classification of the PIR-International Protein Sequence Database is accepted as the standard for protein sequence classification (16).

MIPS extracts all homology domains annotated as a domain feature from the PIR-International Protein Sequence Database into a specific database called HOMDOM. This database is used to identify yet unannotated occurrences of homology domains and to generate the corresponding features. Currently, 28 000 individual domain features are annotated for the 361 distinct homology domains represented in the PIR-International Protein Sequence Database. For each homology domain, a multiple alignment is computed. It is noteworthy that the alignment is restricted to the sequence of the homology domain itself and excludes the neighboring, non-homologous sequences. The MIPS WWW site gives access to the PROT-FAM project with ~16 500 multiple sequence alignments at the level of the protein family (12 500), protein superfamily (3800) and homology domain (361).

How to contact MIPS

Munich Information Center for Protein Sequences, GSF-Forschungszentrum, Max-Planck-Institute for Biochemistry, D-8152 Martinsried, Germany; Tel. +49 89 8578 2656; FAX +49 8578 2655; Email: w.mewes@gsf.de

ACKNOWLEDGMENTS

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 Frishman,D. and Mewes,H.W. (1997) *Trends Genet.*, **13**, 415–416.
2 Sonnhammer,E.L., Eddy,S.R., Birney,E., Bateman,A. and Durbin,R. (1998) *Nucleic Acids Res.*, **26**, 320–322.
3 Nielsen,H., Engelbrecht,J., Brunak,S. and von Heijne,G. (1997) *Protein Engng*, **10**, 1–6.
4 Frishman,D., Mironov,A., Mewes,H.W. and Gelfand,M. (1998) *Nucleic Acids Res.*, **26**, 2941–2947.
5 Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M., Louis,E.J., Mewes,H.W., Murakami,Y., Philippsen,P., Tettelin,H. and Oliver,S.G. (1996) *Science*, **274**, 546–567.
6 Mewes,H.W., Albermann,K., Bahr,M., Frishman,D., Gleissner,A., Hani,J., Heumann,K., Kleine,K., Maierl,A., Oliver,S.G., Pfeiffer,F. and Zollner,A. (1997) *Nature*, **387** (Suppl.), 7–65.
7 Velculescu,V.E., Zhang,L., Zhou,W., Vogelstein,J., Basrai,M.A., Bassett,D.E.,Jr, Hieter,P., Vogelstein,B. and Kinzler,K.W. (1997) *Cell*, **88**, 243–251.
8 Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) *Science*, **278**, 631–637.
9 Barker,W.C., Pfeiffer,F. and George,D.G. (1996) *Methods Enzymol.*, **266**, 59–71.
10 Hodges,P.E., Payne,W.E. and Garrels,J.I. (1998) *Nucleic Acids Res.*, **26**, 68–72.
11 Bevan,M., Bancroft,I., Bent,E., Love,K., Goodman,H., Dean,C., Bergkamp,R., Dirkse,W., Vanstaveren,M., Stiekema,W., Drost,L. *et al.* (1998) *Nature*, **391**, 485–488.
12 Bairoch,A., Bucher,P. and Hofmann,K. (1997) *Nucleic Acids Res.*, **25**, 217–221.
13 Schuler,G.D., Boguski,M.S., Stewart,E.A., Stein,L.D., Gyapay,G., Rice,K., White,R.E., Rodriguez-Tome,P., Agarwal,A., Bajorek,E. *et al.* (1996) *Science*, **274**, 540–546.
14 Huang,X. (1996) *Genomics*, **33**, 21–31.
15 Buschmann,F., Meunier,R., Rohnert,H., Sommerlad,P. and Stal,M. (1996) *Pattern-oriented Software Architecture*. John Wiley and Sons, Chichester, UK.
16 Gracy,J. and Argos,P. (1998) *Bioinformatics*, **14**, 164–173.
17 Wallace,J.C. and Henikoff,S. (1992) *Comput. Applic. Biosci.*, **8**, 249–254.