Nucleic Acids Research, 1998, Vol. 26, No. 1 33–37

MIPS: a database for protein sequences and complete genomes

H. W. Mewes*, J. Hani, F. Pfeiffer and D. Frishman

Munich Information Center for Protein Sequences (MIPS/GSF) am Max-Planck-Institut für Biochemie, Am Klopferspitz 18, D-82152 Martinsried, Germany

Received October 6, 1997; Accepted October 8, 1997

ABSTRACT

The MIPS group [Munich Information Center for Protein Sequences of the German National Center for Environment and Health (GSF)] at the Max-Planck-Institute for Biochemistry, Martinsried near Munich, Germany, is involved in a number of data collection activities, including a comprehensive database of the yeast genome, a database reflecting the progress in sequencing the *Arabidopsis thaliana* genome, the systematic analysis of other small genomes and the collection of protein sequence data within the framework of the PIR-International Protein Sequence Database (described elsewhere in this volume). Through its WWW server (http://www.mips.biochem.mpg.de) MIPS provides access to a variety of generic databases, including a database of protein families as well as automatically generated data by the systematic application of sequence analysis algorithms. The yeast genome sequence and its related information was also compiled on CD-ROM to provide dynamic interactive access to the 16 chromosomes of the first eukaryotic genome unraveled.

DESCRIPTION

The yeast genome database

The complete sequence of the nuclear yeast genome consisting of 16 chromosomes with more than 13 million base pairs has been recently published as the result of a world-wide collaboration (1,2). MIPS has served as the informatics coordinator for the European part of the project. The MIPS yeast genome database reflects an effort to integrate the available information related to the *Saccharomyces cerevisiae* genome into a comprehensive database. The database is the primary source of the complete DNA sequence of the yeast genome. The DNA sequences of both nuclear and mitochondrial genomes (3) were analyzed by MIPS with regard to putative protein-coding open reading frames (ORFs), RNA genes and DNA elements (Tables 1 and 2).

6306 ORFs have been extracted from the sequence of the nuclear yeast genome. 122 of them are smaller than 100 amino acids and 217 contain introns. The analysis of the mitochondrial genome results in 33 ORFs, of which four are smaller than 100 amino acids; three ORFs contain introns (Table 1).

Based on FASTA sequence homology searches (4) all yeast ORFs have been classified into six categories. FASTA scores between 200 and 1/3 of the FASTA score of the protein when aligned with itself were defined as ‘similarity’. A FASTA score higher than 1/3 of selfscore was defined as ‘strong similarity’. ORFs were defined as questionable by a procedure similar to the one described by Termier and Kalogeropoulos (5). The content of the yeast database is summarized in Table 1. After subtraction of all questionable ORFs and all TY-ORFs (retrotransposon related ORFs) a total of 5802 predicted yeast proteins remains. About 20% of the yeast proteins are expected to be essential for viability. So far 518 essential genes (9%) have been identified and included in the database. The prediction of protein transmembrane regions was performed with the ALOM program (6) and then evaluated manually according to Goffeau *et al*. (7). A total of 2523 proteins can be predicted to contain between one and up to 20 transmembrane spans.

The extraction of ORFs was followed by automatic procedures for calculating the basic protein features and searching for protein motifs. For each genetic entry links to other databases [PIR (8), EMBL (9), YPD (10)] are provided.

The automatic basic annotation of the yeast genome is supplemented by an intensive manual annotation of all genetic elements including the verification and/or correction of automatically annotated features and the linkage of all genetic elements with up-to-date biological information. The information is extracted from the relevant literature. Additional information from other databases is scrutinized, e.g. YPD (10), SGD (11), SWISS-PROT (12) and functional analysis projects [e.g. EuroFan (13)].

Besides ORF identification the characterization of all known RNAs was done. In the yeast genome a total of 299 tRNAs can be identified by similarity searches with the previously known tRNAs (Table 2).

All nuclear encoded ribosomal RNAs are clustered in one region on chromosome XII. Only two of the 100–200 repeats of the rDNA cluster have been sequenced and are shown in the database. On the right side of the rDNA cluster there are several copies of a 3.6 kb repeat that includes the ASP3 gene and a 5S rDNA gene. The precise number of copies of this 3.6 kb repeat in the genome is not known. The mitochondrial 21S rRNA has two exons. The second exon is divided into two parts because the sequence of the circular mitochondrial genome starts in the second exon.

*To whom correspondence should be addressed. Tel: +49 8578 2656; Fax: +49 89 8578 2655; Email: mewes@mips.biochem.mpg.de

Table 1. Overview of the yeast ORFs

Yeast ORFs	Nuclear	Mitochondrial
Extracted ORFs	6306	33
Questionable ORFs	431	2
TY ORFs	104	–
Predicted Yeast Proteins	5771	31
Intron containing yeast ORFs	217	3
Essential genes	518	–
Small ORFs (<100 aa)	122	4
Transmembrane proteins	2503	20

Table 2. Overview of all RNA- and DNA-elements identified in the yeast genome

Yeast RNAs and DNA elements	Nuclear	Mitochondrial
Yeast tRNAs (tRNAs)	275	24
rRNA (5S, 5.8S, 9S, 15S, 18S, 21S, 25S)	100 –200	3
Small nuclear RNAs (snRNAs)	51	–
SRP-associated RNA	1	–
RNAse P RNA	1	–
Telomerase template RNA	1	–
LTR elements (TY-associated)	104	–
LTR elemtents (solo LTRs)	103	–
LTR elements (remnant)	161	–

Table 3. MIPS yeast protein catalogues

Yeast protein catalogues	Number of categories	Number of protein entries
Functional Catalogue (FunCat)	12 main, 168 sub	3392
Protein complexes (CompCat)	207	907
Protein Classes (ClassCat)	21 main, 164 sub	952
PROSITE Motifs Catalogue (15)	572	2075
EC number catalogue (32)	466	945

The entries of each catalogue are directly linked to the annotated database, giving access to all annotated genetic, biochemical and structural data. In addition direct access to protein sequences, DNA sequences and links to related databases via the annotated database are implemented.

All known small nuclear RNAs (snRNAs) have been annotated according to the literature and subdivided into two groups: six snRNAs for pre-mRNA splicing and 45 snRNAs for pre-rRNA processing. Three other RNA genes are described: the SRP-associated RNA, the RNAse P RNA and the telomerase template RNA.

All known DNA elements have been identified by similarity searches: the centromeres, containing the coordinates of the elements CDEI, CDEII and CDEIII; predicted ARS elements according to an 11 bp consensus sequence; the telomeric and subtelomeric regions that are divided into repeat units $(TG1\text{-}3)_n$, STR-D,C,B,A elements, core X elements and Y′ elements according to Louis (14). A total of 368 LTR elements were found that could be grouped into delta-, sigma-, tau- and TY5-LTRs. 104 LTRs belong to the 52 yeast retrotransposons (TYs) and 264 LTRs have no TY in their neighborhood. These consist of 103 LTRs with a well conserved sequence (solo LTRs) and 161 LTRs (remnant LTRs) with a degenerated sequence. Solo LTRs represent positions at which a transposition must have occurred in the near past, in contrast remnant LTRs represent positions at which a transposition must have occurred a long time ago.

The annotation of the genetic elements of the mitochondrial genome also includes the eight origins of replication.

To provide information on the biochemical and physiological context of protein function MIPS has compiled a functional catalogue (1). The Yeast Functional Catalogue lists all the ORFs that can be related to well understood functions (Table 3). Beside the functional catalogue a number of other catalogues specialized on different features of the yeast genome have been compiled (see Table 3).

All the data in the MIPS yeast database are accessible via the World Wide Web under http://www.mips.biochem.mpg.de/yeast/ . Users can perform the following operations:

- receive detailed information on a yeast gene or protein by searching with accession numbers, systematic codes, or gene names
- receive up-to-date genetic, biochemical, physiological and structural information for each ORF
- search for human homologues (ESTs)
- visualize whole chromosomes or selected regions to inspect genetic elements, such as ORFs, tRNAs etc.
- download nucleic-acid or protein sequence data
- inspect the yeast genome for gene redundancy
- browse tables of all yeast RNA elements
- browse tables with special features
- browse yeast genes according to their functional classification
- browse yeast protein complexes and protein families catalogues
- browse all yeast ORFs with EC numbers and PROSITE (15) motifs
- inspect up-to-date sequence homologies and alignments (FASTA database)

A version of the MIPS yeast database is available on a CD-ROM for subscribers of Nature and Science magazines. This tool enables individual access to the data independent of network resources. The standard WWW browser technology can be used for local data processing as the program was written in JAVA programming language and all documents are stored in HTML format. The CD-ROM can be installed on Windows95, WindowsNT and Power-Macintosh operating systems.

The *Arabidopsis* genome database

Sequencing the genome of a model plant, *Arabidopsis thaliana*, is ongoing in a world-wide coordinated effort. MIPS serves as the informatics node for the EU activities funded by the European Commission. The ESSA (European Scientists Sequencing ARABIDOPSIS) projects have completed a 2 Mb contig of the FCA region and a 0.4 Mb contig around the APETALA2 gene, both on chromosome IV. These two regions represent ~2% of the 100 Mb low copy region of the nuclear genome. Based on the observed mean gene density of one gene every 4.8 kb, the total complement of *Arabidopsis* protein-coding genes can be calculated as

~21 000. The putative cellular roles of 54% of the predicted proteins were established by sequence similarity to proteins from other plants and other organisms. The project aims to sequence at least 5 Mb of the 13.5 Mb low-copy regions of the long arm of chromosome IV. In collaboration with large scale sequencing projects in the USA and Japan, summarized in the AGI (Arabisopsis Genome Initiative), the completion of the genome should be achieved by the year 2004. Progress of the *A.thaliana* sequencing project can be monitored through the MIPS WWW server (16) and latest, unannotated BAC sequences can be downloaded.

Automatic annotation of protein sequences by PEDANT

With the advent of the genome sequencing era, the gap between available sequence information and capabilities of the not-so-numerous annotation teams in the major protein sequence databases is widening quickly. While ~250 000 gene products are now available in the public databases, <100 000 protein sequences have been manually annotated so far. Millions of ESTs resulting from various cDNA sequencing projects will never be subjected to the conventional annotation procedures. Although manual analysis remains the only alternative in the 'twilight zone' of sequence similarity, for the majority of newly determined gene products rapid and automatic functional assignment is possible.

PEDANT (Protein Extraction, Description and ANalysis Tool) (17) is a software system that utilizes modern bioinformatics methods to provide complete functional and structural characterization of protein sequence sets—from individual sequences to complete genomes. Currently implemented features allow conducting similarity searches (4,18) against the protein sequence databank, specialized motif collections (15,19,20), and sequences with known secondary structures (21), predicting the location of secondary structural elements (22,23) and transmembrane helices (6,24) as well as low complexity (25) and coiled-coil (26) regions, attributing each sequence to known superfamilies and assigning the protein to one or several pre-defined functional categories derived by careful manual analysis of complete genomic sequences (27). New methods will be easily added.

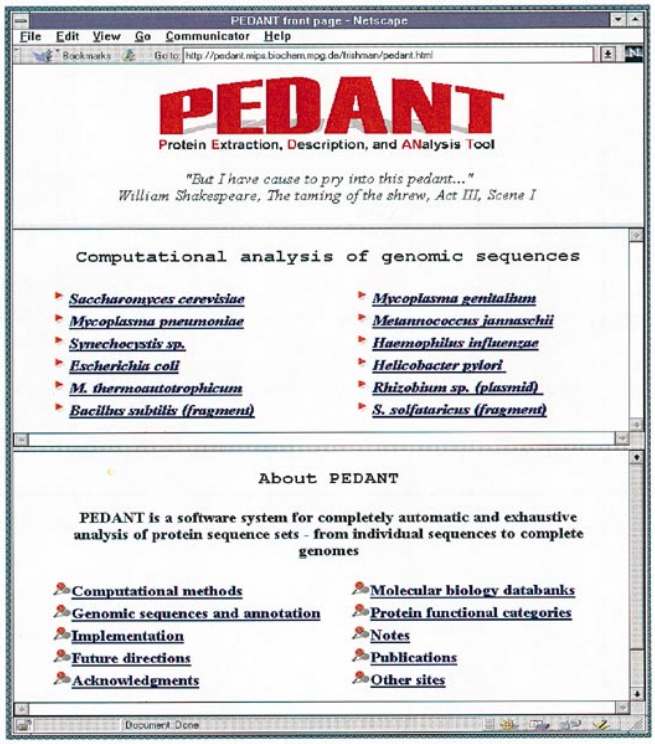

Figure 1. PEDANT front page with the list of genomic sequences analyzed so far.

The PEDANT WWW server (28) (Fig. 1) provides the results of the computational analysis of all publicly released complete genomic sequences. Exhaustive information about structural and functional features of each encoded gene product can be obtained through PEDANT genome browser (Table 4, Fig. 2) (see also 29).

Table 4. Current status of automatic genome annotation by PEDANT

	SC	*MG*	*MP*	*MJ*	*SYN*	*HI*	*EC*	*HP*	*MT*
Total number of ORFs	6303	468	677	1735	3168	1680	4277	1590	1871
ORFs with homologues in other organisms (%)	54.0	67.7	84.3	46.2	55.3	81.1	84.9	69	61.9
ORFs with available 3D information (%)	8.3	12.2	8.9	7.1	10.4	12.1	11.6	10.6	10.6
Total number of PROSITE motifs	225	147	155	190	125	429	583	235	214
Total number of superfamily assignments	1002	226	256	440	754	713	1019	677	551
Total number of PIR keywords	554	185	242	351	476	428	514	442	415
ORFs with assigned functional classes (%)	52.6	87.7	73.3	70.2	31.6	78.0	44.8	46.5	52.3
ORFs with transmembrane regions (%):									
one and more	31.4	32.1	33.4	24.4	33.6	31.5	34.8	32.7	28.9
two and more	14.2	15.6	15.5	11.9	15.8	15.1	18.1	14.3	15.0
three and more	8.7	10.9	10.6	8.9	11.2	12.3	14.6	10.5	10.8
ORFs containing coiled-coil regions (%)	6.8	7.5	6.6	5.2	3.5	3.5	8.9	3.6	
ORFs with >20% low complexity sequence (%)	8.5	6.1	7.2	3.9	3.5	3.5	5.2	4.7	
ORFs with available multiple alignment (%)	43	78	75	41	54	76	82	36	29.7
ORFs with significant BLOCKS hits (%)	17.4	27.7	19.5	12	15.3	25	22.5	16.9	13.1

Abbreviations and references: *SC*, *S.cerevisiae* (1), *MG*, *M.genitalium* (33), *MP*, *M.pneumoniae* (34), *MJ*, *M.jannaschii* (35), *SYN*, *Synechocystis* sp. (36), *HI*, *H.influenzae* (37), *EC*, *E.coli* (38), *HP*, *H.pylori* (39), *MT*, *M.thermoautotrophicum* (40).

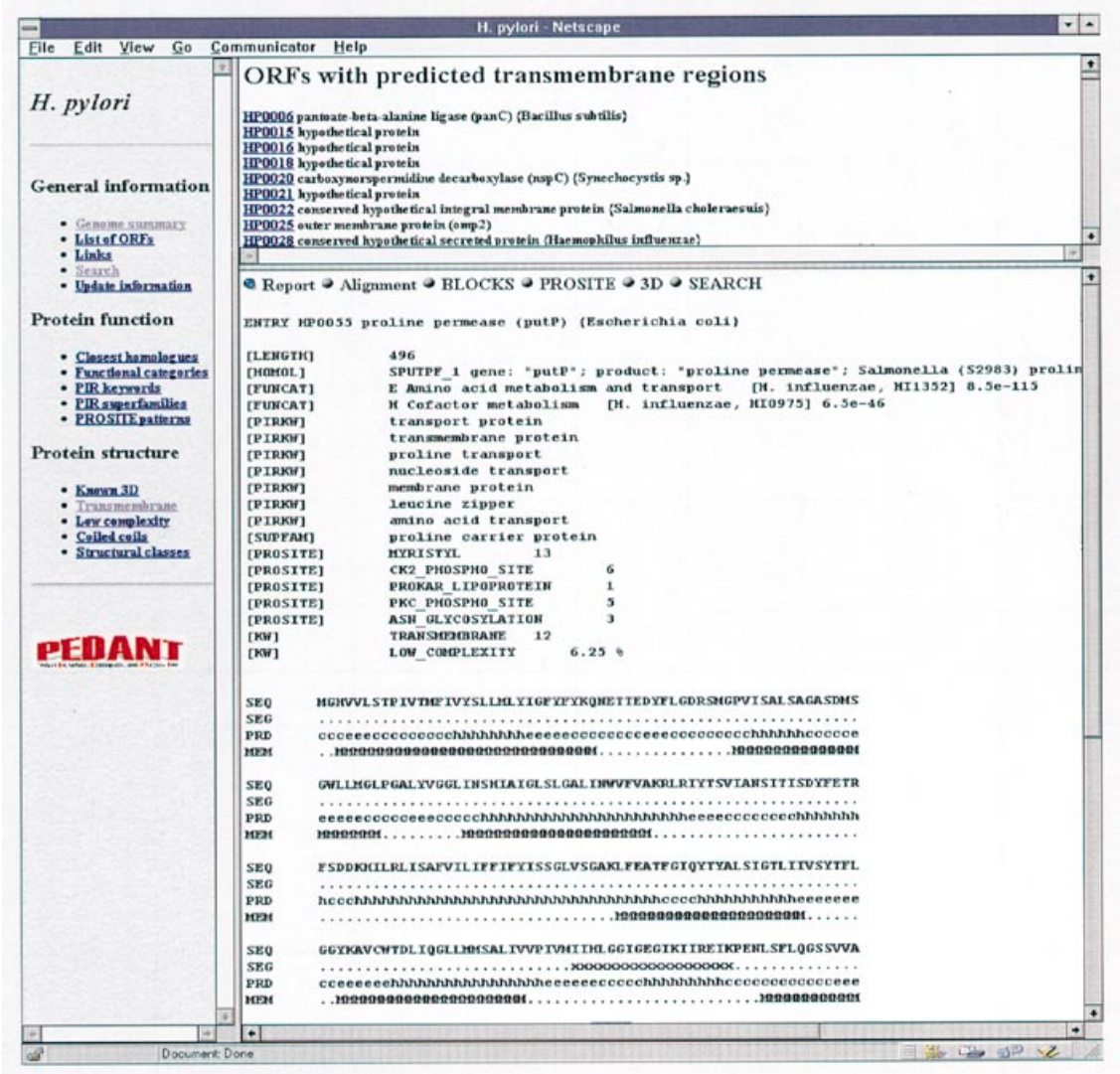

Figure 2. Example of a PEDANT page devoted to a particular genome. The page consist of three areas. The left vertical area is used to select ORFs belonging to a given category (e.g., all proteins with predicted transmembrane regions). The upper right area will then contain the list of ORFs. Integrated report about the ORF of interest appears in the bottom right area. Along with the annotation provided by the original authors (field ENTRY), it contains information about close homologues found by PEDANT (field HOMOL), functional category (field FUNCAT), relevant PIR keywords and superfamilies (fields PIRKW and SUPFAM), PROSITE patterns (field PROSITE), and specific PEDANT keywords (field KW) as well as containing links to more detailed information—multiple sequence alignments, available three-dimensional structural information, detailed description of the found patterns, BLAST and FASTA search results. Protein sequence is shown together with structural features such as predicted transmembrane regions (line MEM) and secondary structure (line PRD), low complexity regions (line SEG), etc.

MIPS anticipates continuing the collection and automatic analysis of all sequenced genomes. While PEDANT provides a rapid first depth analysis, manual inspection, functional classification and individual assignment to the MIPS protein superfamilies will be performed. Links to a diversity of catalogs as well as to databases containing related information will be provided.

Data collection and processing of protein sequences

MIPS is responsible for collecting protein sequence data from European sources for the Protein Sequence Database of PIR-International. The EBI nucleic acid sequence database is used as a major source to retrieve new sequence data. Details of the PIR-International protein sequence database are published elsewhere in this volume.

To cluster evolutionary related sequences, sequences are classified into protein families and superfamilies organized in the PROT-FAM database (30). The classification of all sequences in the PIR-International database into protein families has been completed.

As of September 1997 the PIR-International Protein Sequence Database contains ~98 000 entries. All entries have been classified into one of the ~40 000 protein families. 29 000 families contain a single sequence (30%) as there is no homologue with >50% sequence identity in the database. 5000 families contain two sequences (10%), and ~6000 families contain three or more sequences (54 000 entries; 55%). Approximately 1.5% fragments and 4% very short sequences (<20 amino acid residues) cannot be classified.

For families with two or more members alignments have been computed using the multiple alignment program PILEUP (31). Currently, we continue to cluster protein families into superfamilies. As of September 1997, nearly 60% of all database entries are classified into one of the ~5000 superfamilies. For ~2300 superfamilies which contain two or more sequences, multiple alignments have been built.

MIPS extracts all homology domains annotated as domain feature from the Protein Sequence Database into a specific homology domain sequence database called HOMDOM. This database is used to identify yet unannotated occurrences of homology domains and to generate the corresponding features. Currently, 21 000 individual domain features are annotated for the 316 distinct homology domains.

The MIPS WWW site gives access to the PROT-FAM project with nearly 13 500 multiple sequence alignments at the level of the protein family (11 000 alignments), protein superfamily (2300 alignments) or homology domain (316 alignments).

The MIPS WWW services

The MIPS WWW server is designed to facilitate the access to generic data and functions by integrating services. A layered software architecture was developed to interface the WWW browser to a versatile, comprehensive resource. The server is powered by a DEC 4100 Unix system. Data and services on the WWW server focus on resources uniquely supplied by MIPS including access to individual tables and features of the PEDANT protein analysis system.

How to contact MIPS: Munich Information Center for Protein Sequences of the National Research Center for Environment and Health at the Max-Planck-Institute for Biochemistry, D-82152 Martinsried, Germany. Tel: +49 89 8578 2656; Fax: +49 8578 2655; Email: mailto:mips@gsf.de.

ACKNOWLEDGMENTS

MIPS is supported by the Federal Ministry of Education, Science, Research and Technology (BMBF, FKZ 03311670, 01KW9703), the National Center for Environment and Health (GSF), the Max-Planck-Society and the European Commission (BIO4-CT96-0110, 0338, 0558).

REFERENCES

1 Mewes,H.W., Albermann,K., Bähr,M., Frishman,D., Gleissner,A., Hani,J., Heumann,K., Kleine,K., Maierl,A., Oliver,S.G., *et al.* (1997) *Nature*, **387**, (Suppl.) 7–65.
2 Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M., *et al.* (1996) *Science*, **274**, 546–567.
3 de Zamaroczy,M. and Bernardi,G. (1986) *Gene*, **47**, 155–177.
4 Pearson,W.R. (1990) *Methods Enzymol.*, **183**, 63–98.
5 Termier,M. and Kalogeropoulos,A. (1996) *Yeast*, **12**, 369–384.
6 Klein,P., Kanehisa,M. and Delisi,C. (1985) *Biochim. Biophys. Acta*, **815**, 468–476.
7 Goffeau,A., Nakai,K., Slonimski,P. and Risler,J.-L. (1993) *FEBS Lett.*, **325**, 112–117.
8 George,D.G., Dodson,R.J., Garavelli,J.S., Haft,D.H., Hunt,L.T., Marzec,C.R., Orcutt,B.C., Sidman,K.E., Srinivasarao,G.Y., Yeh,L.S.L., *et al.* (1997) *Nucleic Acids Res.*, **25**, 24–27 [see also this issue (1998) *Nucleic Acids Res.* **26**, 27–32].
9 Stoesser,G., Sterk,P., Tuli,M.A., Stoehr,P.J. and Cameron,G.N. (1997) *Nucleic Acids Res.*, **25**, 7–13 [see also this issue (1998) *Nucleic Acids Res.* **26**, 8–15].
10 Payne,W.E. and Garrels,J.I. (1997) *Nucleic Acids Res.*, **25**, 57–62 [see also this issue (1998) *Nucleic Acids Res.* **26**, 68–72].
11 http://genome-www.stanford.edu/Saccharomyces/
12 Bairoch,A. and Apweiler,R. (1997) *Nucleic Acids Res.*, **25**, 31–36 [see also this issue (1998) *Nucleic Acids Res.* **26**, 38–42].
13 Oliver,S. (1996) *Trends Genet.*, **12**, 241–242.
14 Louis,E.J. (1995) *Yeast*, **11**, 1553–1573.
15 Bairoch,A., Bucher,P. and Hofmann,K. (1997) *Nucleic Acids Res.*, **25**, 217–221.
16 http://mips.biochem.mpg.de/mips/athaliana
17 Frishman,D. and Mewes,H.W. (1997) *Trends Genet.*, **13**, 415–416.
18 Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
19 Kolakowski,L.F.,Jr, Leunissen,J.A.M. and Smith,J.E. (1992) *Biotechniques*, **13**, 919–921.
20 Wallace,J.C. and Henikoff,S., (1992) *Comput. Appl. Biosci.*, **8**, 249–254.
21 Frishman,D. and Argos,P. (1995) *Proteins*, **23**, 566–579.
22 Frishman,D. and Argos,P. (1996) *Protein Engng.*, **9**, 133–142.
23 Frishman,D. and Argos,P. (1997) *Proteins*, **27**, 329–335.
24 Persson,B. and Argos,P. (1994) *J. Mol. Biol.*, **237**, 182–192.
25 Wootton,J.C. and Federhen,S. (1993) *Comp. Chem.*, **17**, 149–163.
26 Lupas,A., Van Dyke,M. and Stock,J. (1991) *Science*, **252**, 1162–1164.
27 Koonin,E.V., Mushegian,A.R., Galperin,M.Y. and Walker,D.R. (1997) *Mol. Microbiol.*, **25**, 619–637.
28 http://pedant.mips.biochem.mpg.de/frishman/pedant.html
29 Frishman,D. and Mewes,H.W. (1997) *Nature Struct. Biol.*, **4**, 626–628.
30 Mewes,H.W., Albermann,K., Heumann,K., Liebl,S. and Pfeiffer,F. (1997) *Nucleic Acids Res.*, **25**, 28–30 [see also this issue (1998) *Nucleic Acids Res.* **26**, 33–37].
31 Genetics Computer Group, Univ. Research Park, 575 Science Drive, Madison, WI 53711.
32 http://www.expasy.ch/sprot/enzyme.html
33 Fraser,C.M., Gocayne,J.D., White,O., Adams,M.D., Clayton,R.A., Fleischmann,R.D., Bult,C.J., Keravage,A.R., Sutton,G., Kelley,J.H. *et al.* (1995) *Science*, **270**, 397–403.
34 Himmelreich,R., Plagens,H., Hilbert,H., Reiner,B. and Herrmann,R. (1997) *Nucleic Acids Res.*, **25**, 701–712.
35 Bult,C.J., White,O., Olsen,G.J., Zhou,L., Fleischmann,R.D., Sutton,G.G., Blake,J.A., FitzGerald,L.H., Clayton,R.A., Goacayne,J.D. *et al.* (1996) *Science*, **273**, 1058–1073.
36 Kaneko,T., Sato,S., Kotani,H., Tanaka,A., Asamizu,E., Nakamura,Y., Hiyajima,N., Hirosawa,M., Sugiura,M., Sasamoto,S. *et al.* (1996) *DNA Res.*, **3**, 109–136.
37 Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M. *et al.* (1995) *Science*, **269**, 496–512.
38 Blattner,F.R., Plunkett,III,G., Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al.* (1997) *Science*, **277**, 1453–1462.
39 Tomb,J.F., White,O., Kerlavage,A.R., Clayton,R.A., Sutton,G.G., Fleischmann,R.D., Ketchunm,K.A., Klenk,H.P., Gill,S., Dougherty,B.A. *et al.* (1997) *Nature*, **388**, 539–547.
40 http://www.cric.com/htdocs/sequences/methanobacter/abstract.html