Nucleic Acids Research, 1996, Vol. 24, No. 1 17–20

The PIR-International Protein Sequence Database

David G. George*, Winona C. Barker, Hans-Werner Mewes[1], Friedhelm Pfeiffer[1] and Akira Tsugita[2]

Protein Information Resource, National Biomedical Research Foundation, Washington, DC 20007, USA, [1]Martinsried Institute for Protein Sequences, Max Planck Institute for Biochemistry, Martinsried, Germany and [2]Japan International Protein Information Database, Science University of Tokyo, Noda, Japan

Received October 5, 1995; Accepted October 6, 1995

ABSTRACT

From its origin the Protein Sequence Database has been designed to support research and has focused on comprehensive coverage, quality control and organization of the data in accordance with biological principles. Since 1988 the database has been maintained collaboratively within the framework of PIR-International, an association of macromolecular sequence data collection centers dedicated to fostering international cooperation as an essential element in the development of scientific databases. The database is widely distributed and is available on the World Wide Web, via ftp, email server, on CD-ROM and magnetic media. It is widely redistributed and incorporated into many other protein sequence data compilations, including SWISS-PROT and the Entrez system of the NCBI.

INTRODUCTION

The Protein Sequence Database [originated by M. O. Dayhoff in 1965 (1–3)] has been maintained since 1988 by PIR-International (4–7). The participating centers include the Protein Information Resource (PIR) at the National Biomedical Research Foundation (NBRF) in the USA, the Martinsried Institute for Protein Sequences (MIPS) at the Max Planck Institute for Biochemistry in Germany and the Japan International Protein Information Database (JIPID) at the Science University of Tokyo, Japan. PIR-International is unique in successfully overcoming the political, social and economic hurdles of organizing a global effort to maintain a single integrated scientific database in full collaboration.

The database contains information concerning all naturally occurring, wild-type proteins whose primary structure (the sequence) is known. A major goal of the database project is to provide comprehensive, non-redundant data uniquely organized by homology and taxonomy. In addition to sequence data, the database contains information (called annotation) concerning: (i) the name and classification of the protein and the organism in which it naturally occurs; (ii) references to the primary literature, including information concerning the sequence determination; (iii) the function and general characteristics of the protein, including gene expression, post-translational processing and activation; (iv) sites and regions of biological interest within the sequence. The database is also unique in maintaining consistency of annotation, with restricted vocabularies employed for features and keywords. Data are accumulated from the published literature, by submissions to PIR-International and by translation of nucleic acid sequences submitted to GenBank (8), the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database (9) and the DNA Data Base of Japan (DDBJ) (10). These data include those deposited with the genome sequence database of the National Center for Genome Resources (http://www.ncgr.org/gsdb/). Entries in the database are cross-referenced to these source databases. In addition, cross-references are included to the genome database (GDB) (11), the yeast gene name LISTA database (12) and MEDLINE. Work is currently underway to cross-reference to the *Drosophila* genome database (FlyBase) (13), the Brookhaven Protein Data Bank (PDB) (14) and the Complex Carbohydrate Structure Database (CCSD) of the international CarbBank project (15).

The PIR-International Protein Sequence Database is widely redistributed; it is integrated into other public data sets, including those assembled by the National Center for Biotechnology Information (NCBI) (16–17) and SWISS-PROT (9), the Protein Sequence Database distributed by the European Biotechnology Institute (EBI) of the EMBL. The database is also distributed by many vendors in conjunction with software packages. Although users may find these software data packages convenient, they should be aware that the database supplied may not be the latest release and may not include all of the information available in the original. (The nodes of PIR-International reserve any rights on their intellectual properties and are not responsible for the versions of the database supplied by any secondary sources.)

RECENT CHANGES IN ARCHITECTURE

A primary goal of the database project is to maintain a comprehensive data set, covering all naturally occurring protein sequences, that is complete with respect to the data available in the public databases and is non-redundant. Non-redundancy is achieved without loss of information. This new architecture and

* To whom correspondence should be addressed

the methodology for processing data have recently been described (18).

Conceptually the database consists of three primary components: the Literature, Source Sequence and Canonical Sequence Components. The Literature Component contains full citation information for all sources of information in the database and is linked to MEDLINE abstracts via MEDLINE MUIDs. Each citation is uniquely identified by a PIR-International Reference Number. Each reported sequence is stored in its originally published form in the Source Sequence Component and is assigned a PIR-International Accession Number that uniquely identifies it. These unmerged source sequences are cross-referenced to the corresponding nucleic acid sequence when present in the GenBank, EMBL or DDBJ nucleic acids sequence databases. In the near future cross-references will be added to link directly to CDS IDs, providing a stable link between the conceptual translation and the corresponding coding region specification. The Canonical Sequence Component (which encompasses the PIR1, PIR2 and PIR3 data sections) is constructed by assembling source sequences that represent the same molecule into merged entries, which display a single canonical sequence and instructions for regeneration of each original source sequence. Hence, all information concerning the various reports of the sequence is stored in a compact, non-redundant form, while remaining directly accessible to users of the database. This architecture allows the competing goals of completeness and non-redundancy to be addressed seamlessly.

Canonical sequences within the database are organized by placement numbers reflecting their similarity to other sequences in the database known to be homologous. Secondarily, the data are organized by species and by protein type (proteins having the same name). Originally, the separation of the database into data sections PIR1, PIR2 and PIR3 reflected the level of data processing and classification. Incompletely processed data are now designated by the status preliminary, listed within the reference portion of the entry, and may occur in any section of the database. In a sense no entry is ever completely processed: if new information becomes available it is merged into the entry. Further, entries are continually monitored and revised as appropriate to reflect the most current biological understanding of the data. The status preliminary generally indicates that the paper reporting the sequence has not been fully analyzed.

The partitioning of the entries into sections PIR1, PIR2 and PIR3 has been retained for ease of physical access to the data in the file distribution form (each file is limited to 64 Kb entries), but this has no other significance. Entry codes are unique across all sections (PIR1–PIR4). For convenience classified entries, found in PIR1, are ordered by placement number (and by species and protein type within placement classes); non-classifed entries are ordered by species and taxonomy. These partitionings will be adjusted as appropriate in each release, therefore, section location is not a stable attribute of the entries.

Section PIR4 has recently been introduced. It is not a regular component of the database, but has been created to make available sequences that are not naturally occurring and/or naturally expressed. This includes conceptual translations of pseudogenes and other non-expressed potential genomic coding regions, engineered and chemically synthesized sequences, and sequences of natural polypeptides that are not ribosomally synthesized. No effort will be made to collect these data; however, they are often accumulated during routine data processing, in which case they will be stored and made available in PIR4. Sequences of these types that occur within PDB entries will be accumulated comprehensively and cross-referenced to PDB.

These and all subsequent changes in the database format are described in the *PIR Technical Development Bulletin*, distributed according to an email distribution list. Contact: PIRMAIL@nbrf.georgetown.edu for additional information.

FAMILY AND DOMAIN ALIGNMENTS

Homology domains are regions of homology shared by otherwise unrelated proteins or sequences repeated in a single protein. In the Protein Sequence Database the superfamily concept is applied both to homology domains within sequences and to the conceptually complete sequence, which is considered a special type of domain, termed ‘homeomorphic’. The names of homology domain superfamilies assigned to a given sequence are listed in the Superfamily record. When a name has been assigned to the homeomorphic superfamily, it appears first in the list (19–20).

More than 250 different homology domains are annotated as features and as superfamily names in the PIR-International Protein Sequence Database. For each of these homology domains MIPS generates a multiple alignment using the Genetics Computer Group program PILEUP (http://www.gcg.com). The alignment contains all domains annotated as features. The alignment is restricted to the domain itself and excludes neighboring sequences. All database entries are systematically classified at MIPS into ‘homeomorphic protein families’. Sequences are clustered into the same protein family when they are homologous from the N- to the C-terminus and have ≥50% sequence identity. Protein families are further clustered into ‘homeomorphic protein superfamilies’ when they are considered to be homologous. For each protein family with more than one member a multiple alignment is generated using the PILEUP program. Currently ~6000 such multiple alignments have been assembled. The alignments are enriched with a consensus sequence and a display of all features annotated in the database entries. The family and domain alignments are available from MIPS (http://www.mips.biochem.mpg.de/). Families that have been clustered into the same superfamily can be easily retrieved.

OTHER DATA SETS

Yeast Sequences is a non-redundant data set generated from worldwide systematic efforts to sequence the complete genome of *Saccharomyces cerevisiae*. The data set compiled at MIPS consists of annotated nucleic acid and protein sequences, including the sequences of chromosomes I, II, III, V, VIII, IX and XI. The most current version of the data set is available from the MIPS World Wide Web (WWW) site. The data are also distributed on the Yeast Sequences CD-ROM produced by MIPS, which includes data accessible through the ATLAS program, FASTA similarity scores and the MIPS family and domain alignment database. The chromosomal data are represented in a form that can be accessed directly by ACeDB for UNIX platforms.

RESID is a database of protein structure modifications produced by PIR. Due to the large and steadily increasing number of protein structure modifications that require standardized annotation in the PIR-International Protein Sequence Database, the RESID database was introduced in 1995 to assist users and

annotators in interpreting features annotations for covalent binding sites, modified sites and cross-links. The RESID database describes features annotated in the Protein Sequence Database and is accessible through the ATLAS program. For a feature in the Protein Sequence Database a corresponding RESID database entry provides systematic chemical names, frequently observed alternate names, Chemical Abstracts registry numbers, atomic formulae and weights, and original amino acids that may have the modification. In addition to providing supplemental documentation on protein structure modifications, the database also provides a means of predicting atomic weights for modified peptides and fragments from the Protein Sequence Database. Release 3.00 contains 204 entries.

The NRL_3D Sequence–Structure Database is produced by PIR from sequence and annotation information extracted from the crystallographic structures of PDB. This database makes the sequence information of PDB available for similarity searches and retrieval and provides cross-reference information for use with the PIR Protein Sequence Database. Release 20.0 of the NRL_3D database (September 1995) contained 6063 entries.

PIR-ALN is a database of protein alignments produced by PIR. Alignments are of sequences in the same family (<55% different from each other) or of sequences representing various families within a superfamily or of sequence segments corresponding to the same homology domain in different proteins. Release 9.0 of the PIR-ALN database (September 1995) contained 1519 entries, 257 of which are homology domain alignments. These effectively define the homology domains so that they can be consistently represented in the entry annotations.

The PATCHX (6) database is assembled by MIPS from a collection of other public domain sequence databases and includes protein sequences not identical with or contained within sequences in the PIR-International Protein Sequence Database. When PATCHX is used with the PIR-International database, together they provide the most complete collection of protein sequence data currently available in the public domain.

ECOLI, the *Escherichia coli* K12 Genomic Database (21), compiled by scientists at JIPID and NBRF, is a comprehensive, non-redundant, fully merged and annotated database containing sequence information from the major data collections (GenBank, EMBL and DDBJ) plus information entered directly from published reports. Protein coding regions are directly cross-referenced to the PIR-International Protein Sequence Database and features are formated to allow direct translation by computer. Release 3.3 of the ECOLI database (August 1995) contained 585 entries.

ACCESS TO PIR-INTERNATIONAL VIA THE WWW AND FTP

The most recent versions of the PIR-International data are accessible directly from MIPS through a developing WWW-based system (http://www.mips.biochem.mpg.de/) designed to provide full retrieval and database searching and analysis services. A similar system will be made available from PIR shortly. The services are organized within a client–server architecture designed to provide services from a cluster of heterogeneous computer systems (22). These services include: (i) access to the most recent versions of all data sets, including the yeast chromosomal data and multiple-sequence alignments assembled from the output of FLASH (23) database searches; (ii) access to a database of FASTA search results (24); (iii) run-time searches for sequence similarities or amino acid patterns based on the HPT-homology indexing structure (25); (iv) an ATLAS-based alert utility to inform users of the availability of new information concerning user-selected classes of proteins.

The PIR-International database may also be found at various other sites, such as the University of Houston Gene-Server (ftp.bchs.uh.du), The Johns Hopkins University (http://www.gdb.org/Dan/proteins/pir.html) and the NCBI (http://www.ncbi.nlm.nih.gov/). These sites also provide ftp access.

ACCESS TO PIR-INTERNATIONAL VIA AN EMAIL SERVER

The NBRF network request server responds to over 25 database query and general service commands. Most of the database query commands are implemented with calls to the ATLAS program. These commands provide simultaneous access to the PIR-International, NRL_3D, PATCHX, PIR-ALN, GenBank and GBNEW databases. The sequence searching command, SEARCH, is implemented through a version of the FASTA program (22), with output routines modified for network transmission. This command also performs another unique function: nucleotide sequences are translated in six reading frames and each polypeptide translation is submitted to a FASTA search against the Protein Sequence Databases. The PIR taxonomic database can be searched with the TAXONOMY command. Complete instructions for the NBRF file server can be obtained by sending an email message containing the command HELP (in the body of the message, not on the subject line) to FILESERV@nbrf.georgetown.edu.

THE ATLAS OF PROTEIN AND GENOMIC SEQUENCES CD-ROM

The Atlas of Protein and Genomic Sequences CD-ROM contains the ATLAS Information Retrieval System, the FASTA program for similarity searches, the PIR-International Protein Sequence Database, the NRL_3D database, the PIR-ALN Protein Alignment Database, the RESID Database of Residue Modifications, the PATCHX database and the ECOLI *Escherichia coli* K12 Genomic Database.

The ATLAS program is a fully integrated multidatabase access program that allows simultaneous access to multiple databases. Although designed primarily to handle macromolecular sequence databases, it can operate on textual databases. The program employs a multidatabase, multifield index structure. This design provides a framework that allows simultaneous retrieval from any selected set of databases and any combination of fields within those databases. ATLAS provides a user friendly environment where entries from selected databases can be linked dynamically for simultaneous retrieval on biological annotations and bibliographic information, such as protein names, superfamily names, homology domains, organism names, gene names, keywords, feature descriptions, authors' names, etc. The ATLAS program also enables selected sets of sequences to be searched directly for exact subsequences or for patterns.

The *Atlas of Protein and Genomic Sequences User's Guide* is updated with each release and included on the CD-ROM in both PostScript and plain text versions; it can also be obtained separately in printed form.

The CCSD database and its associated CarbBank software are available on the ATLAS CD-ROM. Scott Doubet of CarbBank provides the CCSD, software, and documentation to PIR-International for distribution. CarbBank is the computer management system for the CCSD database files and currently runs on PC- or MS-DOS IBM-compatible microcomputers and has a menu-driven user interface. A Windows™ version is expected shortly. An installation manual and tutorial for CarbBank can be printed from the CD-ROM.

The CD-ROM is formated in accordance with the ISO 9660 standard and can be read from any computer system supporting this standard. The ATLAS program currently runs on PC-DOS, VAX/VMS, OpenVMS Alpha AXP, OSF/1 Alpha AXP, DEC ULTRIX (RISC), SunOS, SGI/IRIX and Macintosh systems. The program is written in the C computer language and complies with the ANSI standard.

DATA DISTRIBUTION ON MAGNETIC TAPES

The PIR-International Protein Sequence Database and associated information are also available on a variety of magnetic media.

HOW TO OBTAIN PIR-INTERNATIONAL DATABASES AND SOFTWARE

For information on currently available database releases or other services contact the PIR Technical Services Coordinator, National Biomedical Research Foundation, 3900 Reservoir Road, NW, Washington, DC 20007, USA (tel. +1 202 687-2121; fax +1 202 687-1662; email PIRMAIL@nbrf.georgetown.edu). In Europe contact the Martinsried Institute for Protein Sequences, Max Planck Institute for Biochemistry, D-82152 Martinsried, Germany (tel. +49 89 8578 2657; fax +49 89 8578 2655; email mewes@mips.embnet.org). In Asia or Australia please contact the Japan International Protein Information Database, Science University of Tokyo, 2669 Yamazaki, Noda 278, Japan
tel. +81 471 239778; fax +81 471 221544; email TSUGITA@JPNSUT31.BITNET.

ACKNOWLEDGEMENTS

PIR-International staff members with principal responsibility for distributed data sets include Friedhelm Pfeiffer (PATCHX), John S.Garavelli (NRL_3D and RESID), Geetha Srinivasarao and Lai-Su Yeh (PIR-ALN) and T.Kunisawa (ECOLI). This publication was supported in part by grant P41 LM05798 from the National Library of Medicine. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Library of Medicine. Development of the ATLAS CD-ROM was partially supported by a grant from Digital Equipment Corporation. MIPS is supported by the Max-Planck-Gesellschaft, the Forschungszentrum für Umwelt und Gesundheit (GSF) and the European Economic Community BRIDGE Programme grants BIOT-CT-0167 and 0172.

REFERENCES

1 Dayhoff,M.O., Eck,R.V., Chang,M.A. and Sochard,M.R. (1965) *Atlas of Protein Sequence and Structure.* National Biomedical Research Foundation, Silver Spring, MD.
2 Dayhoff,M.O. (1972) *Atlas of Protein Sequence and Structure,* Vol. 5. National Biomedical Research Foundation, Washington, DC.
3 Dayhoff,M.O. (1979) *Atlas of Protein Sequence and Structure,* Vol. 5, Suppl. 3. National Biomedical Research Foundation, Washington, DC.
4 Keil,B. (1989) In Colwell,R.R. (ed.), *Biomolecular Data: A Resource in Transition.* Oxford University Press, New York, NY, pp. 27–32.
5 Mewes,H.W., George,D.G., Barker,W.C. and Tsugita,A. (1989) In Wittmann-Liebold,B. (ed.), *Methods in Protein Sequence Analysis.* Springer-Verlag, Berlin, Germany, pp. 357–360.
6 Barker,W.C., George,D.G., Mewes,H.-W., Pfeiffer,F. and Tsugita,A. (1993) *Nucleic Acids Res.*, **21**, 3089–3092.
7 George,D.G., Barker,W.C., Mewes,H.-W., Pfeiffer,F. and Tsugita,A. (1994) *Nucleic Acids Res.*, **22**, 3569–3573.
8 Benson,D., Boguski,M., Lipman,D.J. and Ostell,J. (1994) *Nucleic Acids Res.*, **22**, 3441–3444.
9 Emmert,D.B., Stoehr,P.J., Stoesser,G. and Cameron,G.N. (1994) *Nucleic Acids Res.*, **22**, 3445–3449.
10 Tateno,Y., Ugawa,Y., Yamazaki,Y., Hayashida,H., Saitou,N. and Gojobori,T. (1991) *CODATA Bull.*, **23** (4), 74–75.
11 Fasman,K.H., Cuticchia,A.J. and Kingsbury,D.T. (1994) *Nucleic Acids Res.*, **22**, 3462–3469.
12 Linder,P., Doelz,R., Mosse,M.-O., Lazowska,J. and Slonimski,P.P. (1993) *Nucleic Acids Res.*, **21**, 3001–3002.
13 Merriam,J., Ashburner,M., Hartl,D.L. and Kafatos,F. (1991) *Science*, **254**, 221–225.
14 Abola,E.E., Bernstein,F.C., Bryant,S.H., Koetzle,T.F. and Weng,J. (1987) In Allen,F.H., Bergerhoff,G. and Sievers,R. (eds), *Crystallographic Databases—Information Content, Software Systems, Scientific Applications.* Data Commission of the International Union of Crystallography, Cambridge, UK, pp. 107–132.
15 Doubet,S. (1991) *CODATA Bull.*, **23**(4), 56–58.
16 Benson,D., Boguski,M., Lipman,D.J. and Ostell,J. (1990) *Genomics*, **6**, 389–391.
17 Benson,D. (1991) *CODATA Bull.*, **23** (4), 76–78.
18 George,D.G., Hunt,L.T. and Barker,W.C. (1996) In Doolittle,R.F. (ed.), *Computer Methods for Macromolecular Sequence Analysis.* Academic Press, Orlando, FL, in press.
19 Barker,W.C., Pfeiffer,F. and George,D.G. (1996) In Doolittle,R.F. (ed.), *Computer Methods for Macromolecular Sequence Analysis.* Academic Press, Orlando, FL, in press.
20 Barker,W.C., Pfeiffer,F. and George,D.G. (1995) In Atassi,M.Z. and Appella,E. (eds), *Methods in Protein Structure Analysis.* Plenum Publishing, New York, NY, pp. 473–481.
21 Kunisawa,T., Nakamura,M., Watanabe,H., Otsuka,J., Tsugita,A., Yeh,L.-S.L., George,D.G. and Barker,W.C. (1990) *Protein Sequence Data Anal.*, **3**, 157–162.
22 Heumann,K., Harris,C., Kaps,A., Liebl,S., Maierl,A., Pfeiffer,F., Mewes,H.W. (1996) manuscript in preparation.
23 Califano,A. and Rigoutsos,I. (1993) In Hunter,L., Searls,D. and Shavlik,J. (eds), *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology (ISMB-93).* The AAAI Press, Menlo Park, CA, pp. 56–64.
24 Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 2444–2448.
25 Mewes,H.-W. and Heumann,K. (1995) In Galil,Z. and Ukkonen,E. (eds), *Combinatorical Pattern Matching: Lecture Notes in Computer Science 937*. Springer Verlag, Berlin, Germany, pp. 261–285.