Nucleic Acids Research, 1999, Vol. 27, No. 1 39–43

The PIR-International Protein Sequence Database

Winona C. Barker*, John S. Garavelli, Peter B. McGarvey, Christopher R. Marzec, Bruce C. Orcutt, Geetha Y. Srinivasarao, Lai-Su L. Yeh, Robert S. Ledley, Hans-Werner Mewes[1], Friedhelm Pfeiffer[1], Akira Tsugita[2] and Cathy Wu[3]

Protein Information Resource, National Biomedical Research Foundation, Washington, DC 20007, USA, [1]GSF-Forschungszentrum für Umwelt und Gesundheit, Munich Information Center for Protein Sequences, am Max-Planck-Institut für Biochemie, am Klopferspitz 18, D-82152 Martinsried, Germany, [2]Japan International Protein Information Database, Science University of Tokyo, Noda, Japan and [3]Department of Epidemiology/Biomathematics, University of Texas Health Center at Tyler, Tyler, TX, USA

Received October 1, 1998; Accepted October 7, 1998

ABSTRACT

The Protein Information Resource (PIR; http://www-nbrf.georgetown.edu/pir/) supports research on molecular evolution, functional genomics, and computational biology by maintaining a comprehensive, non-redundant, well-organized and freely available protein sequence database. Since 1988 the database has been maintained collaboratively by PIR-International, an international association of data collection centers cooperating to develop this resource during a period of explosive growth in new sequence data and new computer technologies. The PIR Protein Sequence Database entries are classified into superfamilies, families and homology domains, for which sequence alignments are available. Full-scale family classification supports comparative genomics research, aids sequence annotation, assists database organization and improves database integrity. The PIR WWW server supports direct on-line sequence similarity searches, information retrieval, and knowledge discovery by providing the Protein Sequence Database and other supplementary databases. Sequence entries are extensively cross-referenced and hypertext-linked to major nucleic acid, literature, genome, structure, sequence alignment and family databases. The weekly release of the Protein Sequence Database can be accessed through the PIR Web site. The quarterly release of the database is freely available from our anonymous FTP server and is also available on CD-ROM with the accompanying ATLAS database search program.

THE PROTEIN INFORMATION RESOURCE

The Protein Information Resource (PIR) was established in 1984 by the National Biomedical Research Foundation (NBRF) as a resource to assist in the identification and interpretation of protein sequence information (1). The PIR database evolved from the original NBRF Protein Sequence Database, developed over a 20 year period by the late Margaret O. Dayhoff and published as the 'Atlas of Protein Sequence and Structure' (2,3). PIR-International is a collaboration established in 1988 between the NBRF, the Munich Information Center for Protein Sequences (MIPS), and the Japan International Protein Information Database (JIPID) to collect and publish what is now the oldest database of biomolecular sequence, source, bibliographic and feature information. The mission of PIR-International remains: (i) to create and maintain the Protein Sequence Database as a comprehensive, non-redundant, well verified collection, organized according to biological principles, including structural, functional and evolutionary relationships; (ii) to provide a research tool that supports the study of protein sequences, their structural and functional properties, and their biological origins; (iii) to freely distribute the database to the public by the most accessible means including the PIR Web site (see Table 1) and CD-ROM; and (iv) to collaborate with other databases in organizing and coordinating the presentation of biomolecular structural information (4).

FEATURES OF THE PROTEIN SEQUENCE DATABASE

The PIR-International Protein Sequence Database has the following major features.

Non-redundancy

The database is non-redundant; identical and highly similar sequences from the same species are merged into a single entry. In merged entries, each separately reported sequence is represented in a manner that clearly shows any differences with the canonical sequence shown in the entry and that allows the reported sequence to be reconstructed on the PIR Web site.

Classification

PIR sequences are classified by sequence similarity into superfamilies, families and homology domains. Alignments of these

*To whom correspondence should be addressed. Tel. +1 202 687 2121; Fax: +1 202 687 1662; Email: pirmail@nbrf.georgetown.edu

Table 1. PIR Web site URLs

For Information on:	Go To URL:
PIR Home Page	http://www-nbrf.georgetown.edu/pir/
MIPS Home Page	http://www.mips.biochem.mpg.de/
Text Search	http://www-nbrf.georgetown.edu/pir/find.html
Sequence Scan	http://www-nbrf.georgetown.edu/nbrf/scan.html
Sequence Search	http://www-nbrf.georgetown.edu/nbrf/search.html
Complete Genomes	http://www-nbrf.georgetown.edu/pir/genome.html
PIR Alignment Search	http://www-nbrf.georgetown.edu/nbrf/getaln.html
MIPS Protfam Project	http://speedy.mips.biochem.mpg.de/mips/programs/classification.html
MIPS FastA Database	http://speedy.mips.biochem.mpg.de/mips/programs/fasta.html
Atlas CD-ROM	http://www-nbrf.georgetown.edu/pir/atcd.html
PIR Documents (word lists, annotation)	http://www-nbrf.georgetown.edu/pir/doc/index.html
Editorial Board	http://www-nbrf.georgetown.edu/pir/eb/edbd.html

families are available. Full-scale family classification assists database organization, improves database integrity and supports database searches by gene families.

Standardized annotation

The PIR Database is a value-added database in which entries are annotated to include important features not found in the original submission. Full citations are given, including article titles. Genetic information is provided, including map position, intron positions and start codon (if different from AUG). Feature annotations and other terminology have been standardized and restricted vocabularies are enforced to provide greater accuracy and consistency.

Cross references

To optimize information retrieval, PIR entries are cross-referenced to major molecular and reference databases, including Medline, GenBank, EMBL, DDBJ, Protein Data Bank, Human Genome Database and others. Hypertext-links to the cross-referenced database entry are available on the PIR Web site.

Comprehensiveness

The Protein Sequence Database, supplemented with other PIR-maintained databases, comprises the most comprehensive collection of non-redundant protein sequences available.

Public domain with regular releases

The database is freely available to the public and has been updated and released four times per year since 1988. Weekly interim updates of the database are available for searching and browsing on the PIR Web site. All sequence data are available to the public as soon as they are available to the PIR staff.

Information retrieval

The database serves as a major information resource to support biological research. Retrieval and knowledge discovery are facilitated by a variety of search options including various database fields (such as superfamily, authors, features and keywords) and direct database sequence similarity searches. Family classification and multiple sequence alignments, coupled with extensive hypertext-links, make it possible to rapidly find and retrieve information on related sequences in PIR and other molecular databases.

DATABASE ORGANIZATION

The Protein Sequence Database is non-redundant. Therefore, finding sequence reports containing identical or highly similar sequences from the same species and performing the necessary merges is a priority at PIR-International. When sequences are merged the differences between sequences are documented in the new entry and conflicting data are reconciled to give the best canonical sequence for the protein. Bibliographic information and cross-references for individual sequence reports are maintained, and the original submitted sequences can be reconstructed automatically on the PIR Web site. Although our policy is to merge reports of the same sequence into one annotated entry, we do not withhold entries until they are fully merged and annotated.

The database is partitioned into four sections, PIR1, PIR2, PIR3 and PIR4. Currently PIR1 and PIR2 account for ~99% of all entries. Entries in PIR1 are fully classified by superfamily assignment, fully merged with respect to other entries in PIR1, and extensively annotated. Many entries in PIR2 are merged, classified, and annotated as fully as typical PIR1 entries. Entries in PIR3 are not classified, merged or annotated. PIR3 comprises less than 1% of the total database and serves as a temporary holding tank for new entries. PIR4 was created to include sequences identified as not naturally occurring or expressed, such as known pseudogenes, unexpressed ORFs, synthetic sequences, and non-naturally occurring fusion, crossover or frameshift mutations.

Database accuracy and integrity is a major concern of PIR-International. Sequence cross-references are checked for concurrency and accuracy. Checks for syntax, standardization of terms, and data integrity are run on all entries before they enter the database and on the entire database when standardization rules are revised. Discrepancies between published sequences or translations in the nucleotide sequence databases and PIR translations are noted in the entries.

Standardization rules and controlled vocabularies are applied to protein names, organism names, keywords, features, genetic information and other fields. Vocabularies used are derived from international nomenclature commissions or other authoritative sources whenever possible.

Lists of the current keywords, species names, superfamily names, etc. and a Guide to Features Annotations can be found on the PIR Web site.

DATA FROM GENOME PROJECTS

The complete genomes of organisms are being determined at an accelerating pace. The sequences from complete genome determinations have usually entered the Protein Sequence Database in an interim weekly update soon after publication. During the import of genome information, we have been able to improve a number of sequence reports by: (i) correctly representing and annotating non-AUG start codons; (ii) identifying the correct initiation site by comparison to previous entries containing direct amino acid sequence information; (iii) annotating translational frameshift as in peptide chain release factor 2; and (iv) annotating translation exceptions such as selenocysteine in formate dehydrogenase alpha chains. ORFs encoding sequences that require translational frameshifting or readthrough of termination codons are sometimes not translated in the GenBank and EMBL databases. The PIR database provides translations and annotates these sequences from genome projects. A listing of the complete genomes available in the PIR database and associated information is available at the PIR Web site.

SUPERFAMILY AND DOMAIN CLASSIFICATION

PIR is the most extensively classified protein sequence database. Classification of protein sequences into superfamilies and families aids scientists in searching against gene families and in determining the functional and evolutionary relationships among family members. Classification also allows for annotation to be extended to related entries.

The concept of protein superfamilies was originally proposed by Margaret Dayhoff (5,6) and was later refined and developed into a formal model by PIR-International (7,8). In the refined model, two classes of superfamilies are defined: homeomorphic and domain superfamilies. For classification into homeomorphic superfamilies, entries must have global sequence similarity from the amino to the carboxyl end. Family groups are defined as members of a homeomorphic superfamily that share 50% or more sequence identity. Members of a domain superfamily share local sequence similarity to a homology domain, which usually constitutes a functional or structural unit. A complete protein can be a member of only one homeomorphic superfamily, which permits the database to be partitioned into non-overlapping sets. In the Protein Sequence Database, superfamily membership is indicated by name in the 'Superfamily' record of an entry. Homology domains are also annotated as sequence features.

Dr Friedhelm Pfeiffer at MIPS has clustered 93% of the sequences in the PIR database into family groups. The remaining entries in the database were not classified, usually because they were too short or fragmentary. Over 11 000 alignments of families containing two or more sequences are available through the MIPS or PIR Web sites. Every family classified in this way has been assigned a permanent identification number. In addition, more than half of the sequences in the PIR database have been further clustered into homeomorphic superfamilies and assigned permanent identifiers. Sequence alignments of superfamilies are maintained in the PIR-ALN database and are available through the PIR Web site.

SUPPLEMENTARY DATABASES

As part of its effort to produce a protein sequence database that is comprehensive, accurate, and consistent, PIR-International produces a number of supplementary sequence and annotation databases. RESID is the PIR database of modified amino acid residues annotated as features in the Protein Sequence Database. PIR-ALN is a database of alignments of protein sequences produced and curated by PIR staff. The RESID and PIR-ALN databases are described separately in this issue.

The NRL_3D Database (9) is produced by PIR from sequence and annotation information extracted from the Protein Data Bank (PDB) of three-dimensional structures (10). This database makes the sequence information of PDB available for searches and retrieval, and provides cross-reference information for use with the PIR-International Protein Sequence Database.

The FASTADB Database is produced by MIPS from weekly updates of the PIR-International Protein Sequence Database. It contains FastA (11) scores for each database entry with all other sequences in the database. PIR-International uses the FASTADB for identifying sequence relationships for classification and annotation. Entries in FASTADB can be viewed at the MIPS Web site; individual alignments of an entry with related sequences in the FASTADB list can also be displayed.

The PATCHX (12) 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. To reduce redundancy, certain sequences are excluded from PATCHX: questionable sequences, such as those in PIR4; sequences already processed by PIR-International; sequences with very high homology to a PIR database entry; and very short sequences. When PATCHX is used together with the PIR-International database, they provide the most comprehensive collection of protein sequence data currently available in the public domain. PATCHX is available by FTP and on CD-ROM.

LINKS TO OTHER DATABASES

PIR-International maintains concurrent cross-references and links to many other databases, including major nucleotide sequence, protein structure, reference and protein family databases. Table 2 shows a list of database identifiers presently used in the Protein Sequence Database. The identifiers in Table 2 appear on the PIR Web site as hypertext-links to the appropriate database.

The GenBank/EMBL/DDBJ databases, as well as data from various genome sequencing centers, are the sources for most primary data. Entries derived from these databases are presented in the Protein Sequence Database with a citation to the publication, including the Medline unique identifier (MUID), if available. The cross-references include the GenBank/EMBL/DDBJ accession numbers, nucleic acid identifiers (NID), and protein identifiers (PID). Protein identifiers are included in cross-references only if the translation in PIR matches the translation in the cross-referenced database. Discrepancies can occur when entries derived from publications do not agree with translations submitted to the database or when entries to a database are modified. These situations are annotated in the entry.

PIR incorporates and maintains cross-references to genetic information, including standardized gene names, symbols and map positions, in collaboration with genome databases. PIR uses

Table 2. Cross-referenced database identifiers in the PIR databases

External Database object	Database Identifier
MedLine Unique Identifier	MUID
Protein Data Bank Code	PDB
DDBJ Accession Number	DDBJ
EMBL Accession Number	EMBL
GenBank Accession Number	GB
GenBank/EMBL/DDBJ Nucleic Acid Sequence Id	NID
GenBank/EMBL/DDBJ Protein Sequence Id	PID
The Institute for Genome Research CDS Id	TIGR
University of Wisconsin Genome Project	UWGP
Drosophila Database Gene Identification Number	FlyBase
Human Genome Database Gene Symbol and Accession Number	GDB
Mouse Genome Database Gene Symbol and Accession Number	MGI
On-Line Mendelian Inheritance in Man Accession Number	OMIM
MIPS Genome project	MIPS
Yeast Genome Database	SGD
PIR Self Reference	PIR

Cross-references are hypertext-linked to the appropriate database if a facility is available.

information in the Protein Data Bank (10) predominantly for annotation. However, sequences of natural origin that appear with coordinate data in the PDB and that are not published elsewhere are included with cross-references to the PDB code. Links from NRL_3D and PIR to PDB provide a cross-reference between protein sequences and their structures. The ProClass Protein Family Database (13) is a non-redundant database that organizes PIR and Swiss-Prot sequence entries according to global (superfamily) and local (domain) family relationships as defined collectively by PIR superfamilies and PROSITE patterns. Links from the Protein Sequence Database to ProClass provide cross-references to other protein family, family alignment and structural-class databases.

In addition to cross-references to other databases, PIR also maintains cross-references to related entries within the PIR database. Linked entries include alternate splice forms, important strain-specific variations, and entries related by structural or functional annotations.

THE PIR EDITORIAL BOARD

We have created an Editorial Board to channel the expertise of the scientific community into the organization, integration and annotation of the protein sequence data. Members review selected protein groups, advise on information to supplement the sequence data, and contribute their own authored material, such as descriptions and alignments. We provide Web pages with information useful for the editors. We are enlisting additional members and welcome those who would like to participate to contact the PIR.

DATABASE ACCESS AND DISTRIBUTION

The World Wide Web provides the primary means to access the PIR-International Protein Sequence Database. The PIR and MIPS home pages can be found at http://www-nbrf.georgetown.edu/pir and http://www.mips.biochem.mpg.de/ respectively. Table 1 lists the URLs for a number of useful Web pages as entrance points to various resources and search options.

The major search options include: retrieval from the different PIR-maintained databases (Protein Sequence Database, NRL_3D, PIR-ALN and RESID) using a variety of fields (such as superfamily, author, citation, gene name, keyword and protein feature), and sequence similarity searching against the Protein Sequence Database and NRL_3D using BLAST and other programs. When viewing an entry on the PIR Web site, lists of other entries in the same superfamily or sharing the same keywords can be obtained, and alignments of the superfamily or homology domains in PIR-ALN can be displayed through hypertext-links. Sequences provided by users for similarity searching retrieve both sequence entries and family records containing links to alignments and related protein structures. In addition, the entry sequence, published sequence, and any tagged feature or homology domain can automatically be viewed and submitted for a BLAST search of protein databases at the National Center for Biotechnology Information (NCBI). Links to the MIPS Web site allow one to: browse lists of superfamily names and homology domain names; identify all homology domains annotated for a given superfamily and all superfamilies that contain a given homology domain; browse superfamily, family and homology domain alignments; align sequences against profiles for families and homology domains; and inspect FastA results for all Protein Sequence Database entries.

Weekly updates of the Protein Sequence Database are available on the PIR Web site. The quarterly releases of the database are freely available via anonymous FTP at ftp://nbrf.georgetown.edu. The PIR FTP files are in directory [ANONYMOUS.PIR]. Other Web sites and data depositories do not always have the latest quarterly release of the PIR database. One should check the PIR or MIPS Web sites for the weekly interim update and for quarterly release information.

The quarterly release of the database is also available on the ‘Atlas of Protein and Genomic Sequences’ CD-ROM. The ATLAS program provided on the CD-ROM allows simultaneous access to and retrieval from multiple molecular or text databases. Entries can be retrieved from any selected set of databases and any combination of fields including bibliographic information, biological annotations, protein names, superfamily names, organism names, gene names, keywords, features and cross-references to other databases. The ATLAS program also enables selected sets of sequences to be searched directly for user-defined sequences and sequence patterns. The User’s Guide for the ATLAS program is included on the CD-ROM and is available on the PIR Web site. The ATLAS program written in C currently runs on PC-DOS, VAX/VMS, OpenVMS, DEC UNIX, SunOS, SGI/IRIX and Macintosh systems.

ACKNOWLEDGEMENTS

The professional staff of the NBRF acknowledges the assistance and support of Ms Kathryn E. Sidman, PIR Technical Services Coordinator, and Ms Desiree Goins, Project Support Specialist. This publication was supported in part by grant number P41 LM05798 from the National Library of Medicine to NBRF. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Library of Medicine. Work by MIPS was supported by grants from the Bundesministerium für Bildung, Forschung und Technologie (BMBF, FKZ 0311670, 01KW9703/7) and the European Commission (BIOCT-CT-96110). PIR is a registered mark of the NBRF.

REFERENCES

1 George,D.G., Barker,W.C. and Hunt,L.T. (1986) *Nucleic Acids Res*., **14**, 11–15.
2 Dayhoff,M.O., Eck,R.V., Chang,M.A. and Sochard,M.R. (1965) *Atlas of Protein Sequence and Structure* Vol. 1. National Biomedical Research Foundation, Silver Spring, MD.
3 Dayhoff,M.O. (1979) *Atlas of Protein Sequence and Structure* Vol., 5, Supplement 3. National Biomedical Research Foundation, Washington, DC.
4 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., Arminski,L.M., Ledley,R.S., Tsugita,A. and Barker,W.C. (1997) *Nucleic Acids Res*., **25**, 24–27.
5 Dayhoff,M.O. (1976) *Fed. Proc*., **35**, 2132–2138.
6 Dayhoff,M.O., McLaughlin,P.J., Barker,W.C. and Hunt,L.T. (1975) *Naturwissenschaften*, **62**, 154–161.
7 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, 473–481.
8 Barker,W.C., Pfeiffer,F. and George,D.G. (1996) *Methods Enzymol*., **366**, 59–71.
9 Pattabiraman,N., Namboodiri,K., Lowrey,A. and Gaber,B.P. (1990) *Protein Seq. Data Anal*., **3**, 387–405.
10 Abola,E.E., Manning,N.O., Prilusky,J., Stampf,D.R. and Sussman,J.L. (1996) *Res. Natl Stand. Technol*., **101**, 231–241.
11 Pearson,W.R. and Lipman,D.J. (1985) *Science*, **227**, 1435–1441.
12 Barker,W.C., George,D.G., Mewes,H.-W., Pfeiffer,F. and Tsugita,A. (1993) *Nucleic Acids Res*., **21**, 3089–3092.
13 Wu,C., Shivakumar,S. and Huang,H. (1999) *Nucleic Acids Res*., **27**, 272–274.