 Nucleic Acids Research, 2000, Vol. 28, No. 1 41–44

The Protein Information Resource (PIR)

Winona C. Barker*, John S. Garavelli, Hongzhan Huang, Peter B. McGarvey, Bruce C. Orcutt, Geetha Y. Srinivasarao, Chunlin Xiao, Lai-Su L. Yeh, Robert S. Ledley, Joseph F. Janda, Friedhelm Pfeiffer[1], Hans-Werner Mewes[1], Akira Tsugita[2] and Cathy Wu

Protein Information Resource, National Biomedical Research Foundation, 3900 Reservoir Road, NW, Washington, DC 20007, USA, [1]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 and [2]Japan International Protein Information Database, Amakubo 1-16-1, Tsukuba 305-0005, Japan

Received October 1, 1999; Accepted October 4, 1999

ABSTRACT

The Protein Information Resource (PIR) produces the largest, most comprehensive, annotated protein sequence database in the public domain, the PIR-International Protein Sequence Database, in collaboration with the Munich Information Center for Protein Sequences (MIPS) and the Japan International Protein Sequence Database (JIPID). The expanded PIR WWW site allows sequence similarity and text searching of the Protein Sequence Database and auxiliary databases. Several new web-based search engines combine searches of sequence similarity and database annotation to facilitate the analysis and functional identification of proteins. New capabilities for searching the PIR sequence databases include annotation-sorted search, domain search, combined global and domain search, and interactive text searches. The PIR-International databases and search tools are accessible on the PIR WWW site at http://pir.georgetown.edu and at the MIPS WWW site at http://www.mips.biochem.mpg.de . The PIR-International Protein Sequence Database and other files are also available by FTP.

INTRODUCTION

The accelerating pace of genome sequencing projects has greatly increased the volume and complexity of available molecular data. To realize the fullest possible value from the data and to gain a better understanding of the genome, databases and the computational tools for analyzing them are required to allow biologically relevant features in the sequences to be identified and to provide insight on their structure and function. For over 30 years, the Protein Information Resource (PIR) has been providing the scientific community with databases and tools for the organization and analysis of protein sequence data (1,2). Together with MIPS and JIPID, we have undertaken a major restructuring to meet the challenges presented by the rapid growth of largely uncharacterized sequence data and the opportunities provided by the nearly universal access of scientists to the resources available on the WWW. Among the key developments are complete protein family organization for the PIR-International Protein Sequence Database (PSD) and integrated WWW interfaces for user-friendly sequence analysis, database searching and information retrieval.

THE PIR-INTERNATIONAL PROTEIN DATABASES

PIR, MIPS and JIPID constitute the PIR-International consortium that maintains the PIR-International Protein Sequence Database (PSD), the largest publicly distributed and freely available protein sequence database. The database has the following distinguishing features.

- It is a comprehensive, annotated, and non-redundant protein sequence database, containing over 142 000 sequences as of September 1999. Included are sequences from the completely sequenced genomes of 16 prokaryotes, six archaebacteria, 17 viruses and phages, >100 eukaryote organelles and *Saccharomyces cerevisiae*.
- The collection is well organized with >99% of entries classified by protein family and >57% classified by protein superfamily.
- PSD annotation includes concurrent cross-references to other sequence, structure, genomic and citation databases, including the public nucleic acid sequence databases ENTREZ, MEDLINE, PDB, GDB, OMIM, FlyBase, MIPS/Yeast, SGD/Yeast, MIPS/Arabidopsis and TIGR. Where these databases are publicly and freely accessible and provide suitable WWW access, the cross-references presented on the PIR WWW site are hot-linked so that searchers can consult the most current data.
- The PIR is the only sequence database to provide context cross-references between its own database entries. These cross-references assist searchers in exploring relationships such as subunit associations in molecular complexes, enzyme–substrate interactions, activation and regulation cascades, as well as in browsing entries with shared features and annotations.
- Interim updates are made publicly available on a weekly basis, and full releases have been published quarterly since 1984.

In addition to the PSD, PIR-International distributes or provides WWW access to other sequence and auxiliary databases

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

Table 1. PIR-International sequence and auxiliary databases

Database	Description	Information
PSD	Annotated and classified protein sequences	http://pir.georgetown.edu/pirwww/dbinfo/textpsd.html
PATCHX	Sequences not yet in the PIR-International PSD	http://pir.georgetown.edu/pirwww/dbinfo/patchx.html
ARCHIVE	Sequences as originally reported in a publication or submission	http://pir.georgetown.edu/pirwww/dbinfo/archive.html
NRL_3D	Sequences from three-dimensional structure database PDB	http://pir.georgetown.edu/pirwww/dbinfo/nrl3d.html
FAMBASE	Representative sequences from each protein family	http://pir.georgetown.edu/pirwww/dbinfo/fambase.html
PIR-ALN	Sequence alignments of superfamilies, families and homology domains	http://pir.georgetown.edu/pirwww/dbinfo/piraln.html
RESID	Post-translational modifications with PSD feature information	http://pir.georgetown.edu/pirwww/dbinfo/resid.html
ProClass	Non-redundant sequences organized according to superfamilies and motifs	http://pir.georgetown.edu/gfserver/proclass.html
ProtFam	Sequence alignments of superfamilies	http://www.mips.biochem.mpg.de/proj/protfam/protfam

(Table 1), briefly described below, and maintains several internal data collections used for sequence annotation and integrity checks.

- PATCHX (3) is a non-redundant database assembled by MIPS of publicly available protein sequences not yet in the PIR-International PSD. PIR+PATCHX, a combination of the PSD and PATCHX containing ~300 000 sequences available for similarity searches, is the most complete non-redundant collection of protein sequences available in the public domain.
- ARCHIVE is a database of protein sequences as originally reported in a publication or submission, the only such collection of 'as published' unmerged sequences.
- NRL_3D (4) sequence-structure database is produced from sequence and annotation in the Protein Data Bank (PDB) of three-dimensional structures (5).
- FAMBASE is a collection of representative sequences from each protein family that can be used in a similarity search to reduce search time and improve sensitivity for identifying distant families.
- PIR-ALN (6) is a curated database of sequence alignments of superfamilies, families and homology domains, with annotation information derived from PSD and consensus patterns calculated from the alignments.
- RESID (7) is a database of post-translational modifications with descriptive, chemical, structural and bibliographic information based on feature information in the PSD.
- ProClass (8) is a protein family database that organizes non-redundant PIR-International PSD and SWISS-PROT sequences according to PIR superfamilies and PROSITE patterns.
- ProtFam (9) is a curated database of homology clusters with automatically generated multiple sequence alignments for families, superfamilies and homology domains.

To support both data management and data mining and assist knowledge discovery, the PIR databases are being migrated to an object-relational database management system. A three-tier network computing, architecture provides a framework for distributed object computing and Java-based WWW interfaces connect with the database server for database query and update tasks.

Superfamily and family classification

The pioneering work of Margaret Dayhoff on protein classification based on the superfamily concept (10,11) was refined by PIR-International (12) to assist database organization and molecular evolution studies. Central to the organization and annotation of the PIR-International sequence and auxiliary databases are the protein family relationships which are structured at three levels: (i) superfamilies and families (for full-length sequence similarity), (ii) homology domains (for local functional or structural units), and (iii) motifs (for functional or structural sites). PIR-International has maintained the highest classification rate and provided the most comprehensive classification and alignments of proteins among all major public domain databases.

To deal efficiently with the many new sequences from genome sequencing projects, procedures for family and superfamily classification have been automated. Over 99% of sequences are routinely classified shortly after entry into the database into protein families of sequences that are at least 45% identical. Subsequently, entries are further clustered into regular superfamilies of sequences that share end-to-end homology (but may be rather distantly related) and also into domain superfamilies of proteins sharing at least one common homology domain. There are currently >76 000 sequences in >8900 superfamilies, and 30 000 entries with 370 recognized homology domains in the PSD. Corresponding to the classification are 1500 superfamily, 2100 family and 400 domain alignments in the PIR-ALN database, and 15 000 family and 4500 superfamily alignments in the MIPS ProtFam database. Also available from PIR is the ProClass protein family database containing 92 000 classified entries as well as 1300 motif alignments of ~44 000 PIR-International PSD entries.

THE PIR SEARCH AND ANALYSIS SYSTEM

The PIR search and analysis system provides search engines of three types (Table 2): (i) interactive text-based search engines, which allow Boolean queries of text fields; (ii) standard sequence similarity search engines, including Peptide Match, Pattern Match, BLAST, FASTA, Pairwise Alignment and Multiple Alignment; and (iii) advanced search engines that combine sequence similarity and annotation searches or evaluate gene family relationships, including Annotation-Sorted Similarity Search, Domain Search, Global and Domain

Table 2. PIR Search and Analysis systems

Search engine	Description	Location
Text/Entry	Interactive searching of text fields, refine with multiple queries	http://pir.georgetown.edu/pirwww/search/textpsd.html
BLAST	Sequence similarity search and analysis, graphical output interface	http://pir.georgetown.edu/pirwww/search/similarity.html
FASTA	Sequence similarity search and analysis, graphical output interface	http://pir.georgetown.edu/pirwww/search/fasta.html
Pattern/Peptide	Variety of pattern or peptide matching tools	http://pir.georgetown.edu/pirwww/search/patmatch.html
Pairwise alignments	Alignments of PIR or user-supplied sequences using SSEARCH	http://pir.georgetown.edu/pirwww/search/pairwise.html
Multiple alignments	Alignments of PIR or user-supplied sequences using CLUSTALW	http://pir.georgetown.edu/pirwww/search/multaln.html
PIR Annotation-Sorted search	Displays of BLAST or FASTA matches sorted by user-selected annotation	http://pir.georgetown.edu/pirwww/search/pass.html
Domain search	Domain sequence search using FASTA, graphical output interface	http://pir.georgetown.edu/pirwww/search/domains.html
Global and Domain search	BLAST and FASTA searching of PSD for global and domain similarity with graphical display	http://pir.georgetown.edu/pirwww/search/dmsim.html
Integrated Environment for Sequence Analysis	Convenient interface for entry retrieval and sequence and annotation searching	http://pir.georgetown.edu/pirwww/search/piriesa.html
GeneFIND	Protein family classification by combining several search and alignment tools and the ProClass database	http://pir.georgetown.edu/gfserver/genefind.html

Similarity Search and GeneFIND. Sequence searching can be performed against the PSD, NRL_3D, PATCHX, FAMBASE, ProClass and the combined PSD+PATCHX collections. Text and entry searching is provided for the PSD, NRL_3D, PIR-ALN and RESID databases.

Sequence search and alignment

BLAST (13) and FASTA (14) searches for sequence similarity are available for all sequence databases. The output of these search engines employs a graphical interface showing location of hits within the query sequence and full-length alignments generated by SSEARCH (15). Multiple or pairwise alignments of PSD or user-supplied sequences can be done using CLUSTALW (16) or SSEARCH. PIR pattern or peptide matching programs can (i) match a query sequence against a database of regular expressions (i.e., patterns); (ii) search a user-specified regular expression against a sequence database; or (iii) find an exact match for a user-specified peptide sequence in one of the sequence databases, including the ARCHIVE database of 'as published' sequences.

PIR Similarity Search system

Combining sequence and annotation search, the Annotation-Sorted Similarity Search facility displays BLAST or FASTA matches along with the user-selected annotation (superfamily, family, species, taxonomic group, keyword or all five) in the annotation-sorted order. The matched entries can be selected for multiple alignments against the query sequence using CLUSTALW and displayed using MView (17).

The Domain Similarity Search engine uses FASTA to search against domain sequences compiled from the PIR-International PSD, and displays the PSD entry and domain annotation with a graphical representation of the matched region with links to domain alignments in PIR-ALN. The Global and Domain Similarity Search uses BLAST to search the PSD for global similarity and FASTA to search the domain sequence collection for local similarity. The results are ranked by the global score and show the extent of matches at both the global and domain levels. Any combination of complete sequences and domains can be selected and viewed in a multiple alignment.

The PIR Integrated Environment for Sequence Analysis provides an integrated environment for all above protein analysis tasks, including sequence similarity search, pattern and peptide match, multiple sequence alignment and advanced PIR similarity searches, as well as for entry retrieval by unique superfamily, family, title, species, taxonomic group, domains or keywords.

GeneFIND (18) provides protein family classification and information retrieval by combining several search/alignment tools and the ProClass database in a multi-level filter system, including the MOTIFIND neural networks, BLAST search, SSEARCH sequence alignment, motif pattern matching, hidden Markov motif modeling and CLUSTALW multiple motif alignment.

AVAILABILITY

PIR provides free public access to value-added protein information through its WWW site at http://pir.georgetown.edu and direct file transfer at ftp://nbrfa.georgetown.edu/pir . In addition to the databases (Table 1) and search tools (Table 2), the PIR WWW site also provides associated metadata, including technical bulletins and documentation that serves as metadata dictionaries for the PIR-International PSD. Accessible from the PIR anonymous FTP site are PIR-International databases and many other documents, files and software tools, including the weekly interim updates of the PSD (in NBRF format) and the corresponding sequence file (in FASTA format). The PIR-International PSD quarterly releases (in both NBRF and CODATA formats) are also available at the NCBI FTP server. Other sites and data depositories do not always have the most recent quarterly release of the PSD.

ACKNOWLEDGEMENTS

PIR is a registered mark of NBRF. The work at NBRF is supported by grant number P41 LM05978 from the National Library of Medicine and by gifts from COMPAQ, Pfizer and Dupont. The work at MIPS is 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. 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.
2. Dayhoff,M.O. (1979) *Atlas of Protein Sequence and Structure*, Vol. 5, Supplement 3. National Biomedical Research Foundation, Washington, DC.
3. Barker,W.C., George,D.G., Mewes,H.-W., Pfeiffer,F. and Tsugita,A. (1993) *Nucleic Acids Res.*, **21**, 3089–3092.
4. Pattabiraman,N., Namboodiri,K., Lowrey,A. and Gaber,B.P. (1990) *Protein Seq. Data Anal.*, **3**, 387–405.
5. Abola,E.E., Manning,N.O., Prilusky,J., Stampf,D.R. and Sussman,J.L. (1996) *Res. Natl Stand. Technol.*, **101**, 231–241.
6. Srinivasarao,G.Y., Yeh,L.-S., Marzec,C.R., Orcutt,B.C. and Barker,W.C. (1999) *Bioinformatics*, **15**, 382–390.
7. Garavelli,J.S. (2000) *Nucleic Acids Res.*, **28**, 209–211 (this issue).
8. Wu,C., Xiao,C. and Huang,H. (2000) *Nucleic Acids Res.*, **28**, 273–276 (this issue).
9. Mewes,H.W., Frishman,D., Haase,D., Kaps,A., Lemcke,K., Mannhaupt,G., Pfeiffer,F., Schüller C., Stocker,S. and Weil,B. (2000) *Nucleic Acids Res.*, **28**, 37–40 (this issue).
10. Dayhoff,M.O. (1976) *Fed. Proc.*, **35**, 2132–2138.
11. Dayhoff,M.O., McLaughlin,P.J., Barker,W.C. and Hunt,L.T. (1975) *Naturwissenschaften*, **62**, 154–161.
12. Barker,W.C., Pfeiffer,F. and George,D. (1996) *Methods Enzymol.*, **266**, 59–71.
13. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.
14. Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
15. Smith,T.F. and Waterman,M.S. (1981) *Adv. Appl. Math.*, **2**, 482–489.
16. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) *Nucleic Acids Res.*, **22**, 4673–4680.
17. Brown,N.P., Leroy,C. and Sander,C. (1998) *Bioinformatics*, **14**, 380–381.
18. Wu,C.H., Huang,H. and Shivakumar,S. (1999) *Int. J. Artificial Intelligence Tools*, in press.