 Nucleic Acids Research, 2002, Vol. 30, No. 1 35–37

The Protein Information Resource: an integrated public resource of functional annotation of proteins

Cathy H. Wu*, Hongzhan Huang, Leslie Arminski, Jorge Castro-Alvear, Yongxing Chen, Zhang-Zhi Hu, Robert S. Ledley, Kali C. Lewis, Hans-Werner Mewes[1], Bruce C. Orcutt, Baris E. Suzek, Akira Tsugita[2], C. R. Vinayaka, Lai-Su L. Yeh, Jian Zhang and Winona C. Barker

National Biomedical Research Foundation, Georgetown University Medical Center, 3900 Reservoir Road, NW, Washington, DC 20007, USA, [1]GSF-Forschungszentrum f. 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, Science University of Tokyo, Noda, Japan

Received September 17, 2001; Revised and Accepted October 10, 2001

ABSTRACT

The Protein Information Resource (PIR) serves as an integrated public resource of functional annotation of protein data to support genomic/proteomic research and scientific discovery. The PIR, in collaboration with the Munich Information Center for Protein Sequences (MIPS) and the Japan International Protein Information Database (JIPID), produces the PIR-International Protein Sequence Database (PSD), the major annotated protein sequence database in the public domain, containing about 250 000 proteins. To improve protein annotation and the coverage of experimentally validated data, a bibliography submission system is developed for scientists to submit, categorize and retrieve literature information. Comprehensive protein information is available from *i*ProClass, which includes family classification at the superfamily, domain and motif levels, structural and functional features of proteins, as well as cross-references to over 40 biological databases. To provide timely and comprehensive protein data with source attribution, we have introduced a non-redundant reference protein database, PIR-NREF. The database consists of about 800 000 proteins collected from PIR-PSD, SWISS-PROT, TrEMBL, GenPept, RefSeq and PDB, with composite protein names and literature data. To promote database interoperability, we provide XML data distribution and open database schema, and adopt common ontologies. The PIR web site (http://pir.georgetown.edu/) features data mining and sequence analysis tools for information retrieval and functional identification of proteins based on both sequence and annotation information. The PIR databases and other files are also available by FTP (ftp://nbrfa.georgetown.edu/pir_databases).

INTRODUCTION

The Protein Information Resource (PIR) has been providing the scientific community with annotated protein databases and analysis tools for over three decades. To better support research in functional genomics and proteomics and facilitate knowledge discovery, we have made several new advances in the last year, in addition to further enhancing the PIR-International Protein Sequence Database. Some key developments include: launch of a new submission mechanism for literature data, distribution of a new non-redundant reference protein database, enhancement of the integrated classification database, and redesign of the web site for easy navigation, information retrieval and sequence analysis.

PIR-INTERNATIONAL PROTEIN SEQUENCE DATABASE

The PIR, along with the Munich Information Center for Protein Sequences (MIPS) and the Japan International Protein Information Database (JIPID), continues to enhance and distribute the PIR-International Protein Sequence Database (PSD), a non-redundant, expertly annotated, fully classified and extensively cross-referenced protein sequence database in the public domain. It contains about 250 000 protein sequences with comprehensive coverage across the entire taxonomic range, including sequences from all the publicly available complete genomes.

Superfamily classification

A unique characteristic of the PIR-PSD is the superfamily/family classification (1) that provides complete and non-overlapping clustering of proteins based on global (end-to-end) sequence similarity. Sequences in the same superfamily share common domain architecture (i.e. have the same number, order and types of domains) and do not differ excessively in overall length unless they are fragments or result from alternate splicing or initiators. The automated classification system places new members into existing superfamilies and defines new superfamily clusters using parameters including the

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

percentage of sequence identity, overlap length ratio, distance to neighboring superfamily clusters, and overall domain arrangement. Currently, >99% of sequences are classified into families of closely related sequences (at least 45% identical), and over two-thirds of sequences are classified into over 33 000 superfamilies. The automated classification is being augmented by manual curation of superfamilies, starting with those containing at least one definable domain, to provide superfamily names, brief descriptions, bibliography, list of representative and seed members, as well as domain and motif architecture characteristic of the superfamily.

Bibliography submission and literature mapping

Linking protein data to literature data that describes or characterizes the proteins is crucial for us to increase the amount of experimentally verified data and to improve the quality of protein annotation. Attribution of protein annotations to validated experimental sources provides effective means to avoid propagation of errors that may have resulted from large-scale genome annotation. We have developed a bibliography submission system for the scientific community to submit, categorize and retrieve literature information for PSD protein entries. The submission interface guides users through steps in mapping the paper citation to given protein entries, entering the literature data, and summarizing the literature data using categories such as genetics, tissue/cellular localization, molecular complex or interaction, function, regulation and disease. Also included is a literature information page that provides literature data mining and displays both references cited in PIR and submitted by users.

INTEGRATED PROTEIN CLASSIFICATION DATABASE

The *i*ProClass (integrated Protein Classification) database (2) is designed to provide comprehensive descriptions of all proteins and to serve as a framework for data integration in a distributed networking environment. The database describes family relationships at both global (whole protein) and local (domain, motif, site) levels, as well as structural and functional classifications and features of proteins. The current version (Release 1.0, August 2001) consists of more than 270 000 non-redundant PIR-PSD and SWISS-PROT proteins organized with more than 33 000 PIR superfamilies, 100 000 families, 3400 PIR homology and Pfam domains (3), 1300 ProClass/ProSite motifs (4,5), 280 PIR post-translational modification sites, and links to over 40 databases of protein families, structures, functions, genes, genomes, literature and taxonomy. Protein sequence and superfamily summary reports provide rich annotations such as membership information with length, taxonomy and keyword statistics, extensive cross-references and graphical display of domain and motif regions. Directly linked to the *i*ProClass sequence report are two additional PIR databases, ASDB and RESID (6). PIR-Annotation and Similarity Database (ASDB) lists pre-computed, biweekly updated FASTA neighbors of all PSD sequences with annotation information and graphical displays of sequence similarity matches. PIR-RESID documents over 280 post-translational modifications and links to PSD entries containing either experimentally determined or computationally predicted modifications with evidence tags. Future versions of *i*ProClass and ASDB will be based on the new PIR Non-redundant Reference Protein database (NREF).

PIR-NREF

As a major resource of protein information, one of our primary aims is to provide a timely and comprehensive collection of all protein sequence data that keeps pace with the genome sequencing projects and contains source attribution and minimal redundancy. The PIR-NREF protein database includes sequences from PIR, SWISS-PROT (7), TrEMBL (7), RefSeq (8), GenPept, PDB (9) and other protein databases. The NREF entries, each representing an identical amino acid sequence from the same source organism redundantly presented in one or more underlying protein databases, can serve as the basic unit for protein annotation. The NCBI taxonomy (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/) is used as the ontology for matching source organism names at the species or strain (if known) levels. The NREF report provides source attribution (containing protein IDs, accession numbers and protein names from underlying databases), in addition to taxonomy, amino acid sequence and composite literature data. The composite protein names, including synonyms, alternate names and even misspellings, can be used to assist the ontology development on protein names and the identification of mis-annotated proteins. Related sequences, including identical sequences from different organisms and closely related sequences within the same organism, are also listed. The database presently consists of about 800 000 entries and is updated biweekly.

AVAILABILITY

PIR web site

The PIR web site (http://pir.georgetown.edu) (10) connects data mining and sequence analysis tools to underlying databases for exploration of protein information and discovery of new knowledge. The site has been redesigned to include a user-friendly navigation system and more graphical interfaces and analysis tools. The PIR-PSD and *i*ProClass pages represent primary entry points in the PIR web site. A list of the major PIR pages is shown in Table 1.

The PIR-PSD interface provides entry retrieval, batch retrieval, basic or advanced text searches, and various sequence searches. The *i*ProClass interface also includes both sequence and text searches. The BLAST search (11) returns best-matched proteins and superfamilies, while peptide match allows protein identification based on peptide sequences. Text search involves direct search of the underlying Oracle tables using unique identifiers or combinations of text strings. The NREF database is searchable by BLAST search, peptide match and direct report retrieval based on the NREF ID or the entry identifiers of the source databases. Other sequence searches supported on the PIR web site include FASTA (12), pattern matching, hidden Markov model (HMM) (13) domain and motif search, Smith–Waterman (14) pair-wise alignment, CLUSTALW (15) multiple alignment and GeneFIND (16) family identification.

PIR FTP site

The PIR anonymous FTP site (ftp://nbrfa.georgetown.edu/pir_databases) provides direct file transfer. Files distributed include the PIR-PSD (quarterly release and interim updates), PIR-NREF, other auxiliary databases, other documents, files

Table 1. Major PIR web pages for data mining and sequence analysis

Description	URL
PIR Home	http://pir.georgetown.edu
PIR-PSD	http://pir.georgetown.edu/pirwww/search/textpsd.shtml
iProClass	http://pir.georgetown.edu/iproclass
PIR-NREF	http://pir.georgetown.edu/pirwww/search/pirnref.shtml
PIR-ASDB	http://pir.georgetown.edu/cgi-bin/asdblist.pl?id=CCHU
Bibliography submission	http://pir.georgetown.edu/pirwww/literature.html
List of PIR databases	http://pir.georgetown.edu/pirwww/dbinfo/dbinfo.html
List of PIR search tools	http://pir.georgetown.edu/pirwww/search/searchseq.html
List of completed genomes	http://pir.georgetown.edu/pirwww/search/genome.html
FTP site	ftp://nbrfa.georgetown.edu/pir_databases

and software programs. The PIR-PSD is distributed as flat files in NBRF and CODATA formats, with corresponding sequences in FASTA format. Both PIR-PSD and PIR-NREF are also distributed in XML format with the associated document type definition (DTD) file.

The PIR-PSD, *i*ProClass and PIR-NREF databases have been implemented in Oracle 8i object-relational database system on our Unix server. To enable open source distribution, the databases are being mapped to MySQL and ported to Linux system. To establish reciprocal links to PIR databases, to host a PIR mirror web site or to request PIR database schema, please contact pirmail@nbrf.georgetown.edu.

ACKNOWLEDGEMENTS

PIR is a registered mark of National Biomedical Research Foundation (NBRF). The PIR is supported by grant P41 LM05978 from the National Library of Medicine, National Institutes of Health. The *i*ProClass and RESID databases are supported by DBI-9974855 and DBI-9808414 from the National Science Foundation.

REFERENCES

1. Barker,W.C., Pfeiffer,F. and George,D.G. (1996) Superfamily classification in PIR-International Protein Sequence Database. *Methods Enzymol.*, **266**, 59–71.
2. Wu,C.H., Xiao,C., Hou,Z., Huang,H. and Barker,W.C. (2001) *i*ProClass: an integrated, comprehensive, and annotated protein classification database. *Nucleic Acids Res.*, **29**, 52–54.
3. Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 276–280.
4. Huang,H., Xiao,C. and Wu,C.H. (2000) ProClass protein family database. *Nucleic Acids Res.*, **28**, 273–276.
5. Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 235–238.
6. Garavelli,J.S., Hou,Z., Pattabiraman,N. and Stephens,R.M. (2001) The RESID database of protein structure modifications and the NRL-3D sequence-structure database. *Nucleic Acids Res.*, **29**, 199–201.
7. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
8. Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.
9. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 245–248.
10. McGarvey,P., Huang,H., Barker,W.C., Orcutt,B.C. and Wu,C.H. (2000) The PIR web site: new resource for bioinformatics. *Bioinformatics*, **16**, 290–291.
11. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
12. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparision. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
13. Eddy,S.R., Mitchison,G. and Durbin,R. (1995) Maximum Discrimination hidden Markov models of sequence consensus. *J. Comp. Biol.*, **2**, 9–23.
14. Smith,T.F. and Waterman,M.S. (1981) Comparison of bio-sequences. *Adv. Appl. Math.*, **2**, 482–489.
15. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
16. Wu,C.H., Huang,H. and McLarty,J. (1999) Gene family identification network design for protein sequence analysis. *Int. J. Artif. Intell. Tools*, **8**, 419–432.