Nucleic Acids Research, 2003, Vol. 31, No. 1 345–347
DOI: 10.1093/nar/gkg040

The Protein Information Resource

Cathy H. Wu*, Lai-Su L. Yeh[1], Hongzhan Huang, Leslie Arminski[1], Jorge Castro-Alvear[1], Yongxing Chen[1], Zhangzhi Hu[1], Panagiotis Kourtesis[1], Robert S. Ledley[1], Baris E. Suzek[1], C.R. Vinayaka[1], Jian Zhang[1] and Winona C. Barker[1]

Department of Biochemistry and Molecular Biology and [1]National Biomedical Research Foundation, Georgetown University Medical Center, 3900 Reservoir Road, NW, Box 571414, Washington, DC 20057-1414, USA

Received September 15, 2002; Accepted September 27, 2002

ABSTRACT

The Protein Information Resource (PIR) is an integrated public resource of protein informatics that supports genomic and proteomic research and scientific discovery. PIR maintains the Protein Sequence Database (PSD), an annotated protein database containing over 283 000 sequences covering the entire taxonomic range. Family classification is used for sensitive identification, consistent annotation, and detection of annotation errors. The superfamily curation defines signature domain architecture and categorizes memberships to improve automated classification. To increase the amount of experimental annotation, the PIR has developed a bibliography system for literature searching, mapping, and user submission, and has conducted retrospective attribution of citations for experimental features. PIR also maintains NREF, a non-redundant reference database, and iProClass, an integrated database of protein family, function, and structure information. PIR-NREF provides a timely and comprehensive collection of protein sequences, currently consisting of more than 1 000 000 entries from PIR-PSD, SWISS-PROT, TrEMBL, RefSeq, GenPept, and PDB. The PIR web site (http://pir.georgetown.edu) connects data analysis tools to underlying databases for information retrieval and knowledge discovery, with functionalities for interactive queries, combinations of sequence and text searches, and sorting and visual exploration of search results. The FTP site provides free download for PSD and NREF biweekly releases and auxiliary databases and files.

INTRODUCTION

In order to provide integrated and value-added protein information to the scientific community, the Protein Information Resource (PIR) continues to enhance its three major databases, the Protein Sequence Database (PSD), the Non-redundant REFerence (NREF) sequence database, and the integrated Protein Classification (iProClass) database (1). The sections below describe key developments in the past year.

PIR-PSD

The PIR-PSD is public domain protein sequence database, which currently contains over 283 000 annotated and classified entries, covering the entire taxonomic range. Recent development and annotation efforts have focused on superfamily classification and curation and bibliography mapping and attribution.

Superfamily classification and curation. A unique characteristic of the PIR-PSD is the superfamily classification (2) that provides comprehensive, non-overlapping, and hierarchical clustering of sequences to reflect their evolutionary relationships. To further improve the quality of automated classification, we have conducted systematic superfamily curation that: (i) defines the signature domain architecture (number, order, and types of domains) characteristic of the superfamily, (ii) categorizes *regular* and *associate* members to distinguish sequence entries sharing the signature features from outliers (such as fragments), and (iii) designates *representative* and *seed* members amongst regular members. Several thousand superfamilies have been manually curated. The seed members provide a basis for automatic placement of new sequences into existing superfamilies and for automatic generation of multiple sequence alignments and phylogenetic trees. Currently, over 99% of PSD sequences are classified into families of closely related sequences (at least 45% identical), and over two-thirds of sequences are classified into >36 000 superfamilies.

Bibliography mapping and attribution. To improve the quality of protein annotation by increasing the amount of experimentally verified data with source attribution, the PIR has developed a bibliography information system and conducted retrospective attribution of literature data. The bibliography system allows browsing and searching of extensive literature collected for all protein entries from PubMed and other curated molecular databases, together with an interface for scientists to

*To whom correspondence should be addressed. Tel: +1 2026872121; Fax: +1 2026871662; Email: pirmail@georgetown.edu

categorize and submit literature information for mapped proteins. In PIR-PSD, protein features such as binding sites, structural motifs, and post-translational modifications are tagged with '*experimental*' status for experimentally determined features to distinguish from those that are computationally predicted; however, they had not been associated with literature citations. A systematic manual attribution of experimental features is being carried out with computer-assisted mapping to existing protein bibliographic information. So far, a few thousand experimental features have been associated with publications.

PIR-NREF DATABASE

The PIR-NREF provides a timely and comprehensive collection of protein sequence data, keeping pace with the genome sequencing projects and containing source attribution and minimal redundancy. The database contains all sequences in PIR-PSD, SWISS-PROT (3), TrEMBL (3), RefSeq (4), GenPept, and PDB (5), totaling more than 1 000 000 entries currently. Identical sequences from the same source organism (species) reported in different databases are presented as a single NREF entry with protein IDs, accession numbers, and protein names from each underlying database, as well as amino acid sequence, taxonomy, and composite bibliographic data. Also listed are related sequences identified by all-against-all FASTA search (6), including identical sequences from different organisms, identical subsequences, and highly similar sequences ($\geq$ 95% identity). NREF can be used for sequence searching and protein identification against the entire sequence collection or a subset of one or more genomes. The collective protein names, including synonyms, and the bibliographic information can be used to develop a protein name ontology. The different protein names assigned by different databases may help detect annotation errors, especially those resulting from large-scale genomic annotation.

AVAILABILITY

PIR web site. The PIR web site connects data mining and sequence analysis tools to underlying databases for information retrieval and knowledge discovery, with functionalities for interactive queries, combinations of sequence and annotation text searches, and sorting and visual exploration of search results. The three major databases (PSD, NREF and iProClass) represent primary entry points in the PIR web site, all of which provide text search for entry and list retrieval as well as BLAST search and peptide match. Direct entry report retrieval is based on sequence unique identifiers of all underlying databases, such as PIR, SWISS-PROT, or RefSeq. Basic and advanced text searches return protein entries listed in summary lines with information on protein IDs, matched fields, protein name, taxonomy, superfamily, domain, and motif, with hypertext links to the full entry report and to cross-referenced databases. More than 50 fields are searchable, including about 30 database unique identifiers (e.g., PDB ID, EC number, PubMed ID, and KEGG pathway number) and a wide range of annotation texts (e.g., protein name, organism name, sequence feature, and paper title). The BLAST search and peptide search likewise return lists of matched entries with summary lines that also contain search statistics and matched sequence region. Protein entries returned from text and sequence searches can be selected for further analysis, including BLAST (7) and FASTA search, pattern match, hidden Markov model (HMMER) (8) domain search, ClustalW (9) multiple sequence alignments and Phylip (10) phylogentic tree generation, and graphical display of superfamily, domain and motif relationships. Species-based browsing and searching are supported for about 100 organisms, including over 70 complete genomes. The related sequences in FASTA clusters are retrievable based on sequence unique identifiers where neighbors are listed with annotation information and graphical display of matched sequence region. A list of the major PIR pages is shown in Table 1.

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

Description	Web page URL
PIR Home	http://pir.georgetown.edu
PIR-PSD	~/pirwww/pirpsd.shtml
PIR-NREF	~/pirwww/pirnref.shtml
*i*ProClass	~/iproclass
Related Sequences	~/cgi-bin/relatedseq.pl?id=CCHU
Genome Searching	~/cgi-bin/nfspecies.pl?taxon=9606
Bibliography submission	~/pirwww/biblisubmit.html
List of PIR databases	~/pirwww/dbinfo/dbinfo.html
List of PIR search tools	~/pirwww/search/searchseq.html
FTP site	ftp://ftp.pir.georgetown.edu/pir_databases/

~ = http://pir.georgetown.edu

PIR FTP site. The three PIR databases, PSD, NREF and iProClass, are updated biweekly in the same release schedule and made immediately available from the PIR web site for searching and browsing, as well as from the FTP site for free downloading. PIR-PSD is distributed as flat files in NBRF, CODATA, and XML formats, and in the open source relational database, MySQL, format. The MySQL distribution file contains data files (in relational tables), SQL scripts for creating the database and a user's guide with the database schema. PIR-NREF is distributed as XML files. Both PSD and NREF XML distributions have an associated DTD (Document Type Definition) file. The sequence files of both databases are distributed in FASTA format.

ACKNOWLEDGEMENT

The PIR is supported by grant P41 LM05978 from the National Library of Medicine, National Institutes of Health.

REFERENCES

1. Wu,C.H., Xiao,C., Hou,Z., Huang,H. and Barker,W.C. (2001) iProClass: an integrated and comprehensive protein classification database. *Nucleic Acids Res.*, **29**, 52–54.
2. Barker,W.C., Pfeiffer,F. and George,D.G. (1996) Superfamily classification in PIR-International Protein Sequence Database. *Methods Enzymol.*, **266**, 59–71.

3. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
4. Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.
5. Westbrook,J., Feng,Z., Jain,S., Bhat,T.N., Thanki,N., Ravichandran,V., Gilliland,G.L., Bluhm,W., Weissig,H., Greer,D.S., Bourne,P.E. and Berman,H.E. (2002) The Protein Data Bank: unifying the archive. *Nucleic Acids Res.*, **30**, 245–248.
6. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
7. 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.
8. Eddy,S.R., Mitchison,G. and Durbin,R. (1995) Maximum Discrimination hidden Markov models of sequence consensus. *J. Comp. Biol.*, **2**, 9–23.
9. 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.
10. Felsenstein,J. (1989) PHYLIP—phylogeny inference package (Version 3.2). *Cladistics*, **5**, 164–166.