Protein Information Resource: a community resource for expert annotation of protein data

Winona C. Barker*, John S. Garavelli, Zhenglin Hou, Hongzhan Huang, Robert S. Ledley, Peter B. McGarvey, Hans-Werner Mewes[1], Bruce C. Orcutt, Friedhelm Pfeiffer[1], Akira Tsugita[2], C. R. Vinayaka, Chunlin Xiao, Lai-Su L. Yeh and Cathy Wu

National Biomedical Research Foundation, 3900 Reservoir Road, LR-3, 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, Proteomics Research Laboratory, Tsukuba, Japan

Received October 2, 2000; Accepted October 4, 2000

ABSTRACT

The Protein Information Resource, in collaboration with the Munich Information Center for Protein Sequences (MIPS) and the Japan International Protein Information Database (JIPID), produces the most comprehensive and expertly annotated protein sequence database in the public domain, the PIR-International Protein Sequence Database. To provide timely and high quality annotation and promote database interoperability, the PIR-International employs rule-based and classification-driven procedures based on controlled vocabulary and standard nomenclature and includes status tags to distinguish experimentally determined from predicted protein features. The database contains about 200 000 non-redundant protein sequences, which are classified into families and superfamilies and their domains and motifs identified. Entries are extensively cross-referenced to other sequence, classification, genome, structure and activity databases. The PIR web site features search engines that use sequence similarity and database annotation to facilitate the analysis and functional identification of proteins. The PIR-International databases and search tools are accessible on the PIR web site at http://pir.georgetown.edu/ and at the MIPS web site at http://www.mips.biochem.mpg.de. The PIR-International Protein Sequence Database and other files are also available by FTP.

The Protein Information Resource (PIR) for over three decades has been a community resource that provides protein databases and analysis tools to support research on molecular evolution, functional genomics and computational biology. The PIR, along with the Munich Information Center for Protein Sequences (MIPS) and the Japan International Protein Information Database (JIPID), maintains and distributes the PIR-International Protein Sequence Database, the most comprehensive, well-annotated and non-redundant protein sequence database in the public domain. To further support genomic and proteomic research, we have greatly improved our bioinformatics infrastructure in the last 2 years, which allows us: (i) to continue to provide high quality protein sequence data and annotation, while keeping pace with the large influx of data being generated by genome sequencing projects; (ii) to develop an integrated system of protein databases and analytical tools for expert annotation and knowledge discovery; and (iii) to improve accessibility of our resource and interoperability of our databases. Some key developments include: highly-automated protein sequence classification and annotation, enhanced web site with many new search engines and functionality for protein data mining and analysis, a new integrated classification database that provides comprehensive descriptions of family relationships and functional/structural annotations, database migration into Oracle 8i object-relational database system and database distribution in XML format.

PIR-INTERNATIONAL PROTEIN SEQUENCE DATABASE

The PIR-International Protein Sequence Database (PIR-PSD) is the largest public domain protein sequence database in which entries are annotated and classified into family groups. It contains about 200 000 protein sequences with comprehensive coverage across the entire taxonomic range, including sequences from all the publicly available complete genomes.

Expert annotation

Controlled vocabulary and standard nomenclature. The PIR-PSD is recognized for the high quality of its expert annotation. To promote annotation quality and database interoperability, PIR-International employs controlled vocabulary for most annotations and adopts standard nomenclature whenever applicable. PIR-International staff members add experimentally determined and confidently predicted annotations using original literature, MEDLINE abstracts and expertly curated databases as sources for information and accepted nomenclature. Among the databases used are the PIR-RESID database of post-translational modifications (1), the Enzyme Nomenclature produced by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (http://www.chem.qmw.ac.uk/iubmb/enzyme/) and The Genome Database (http://gdbwww.gdb.org/).

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

Experimental/predicted status. Feature annotations in PSD entries are tagged with 'experimental', 'predicted', 'absent' or 'atypical' status. All experimentally determined sequence modifications are marked 'experimental', while those that are predicted carry a 'predicted' status. To help combat 'transitive identification catastrophe' arising from incorrect preliminary identifications by genome centers, PIR-International has introduced status information in entry titles. These are designated 'validated', 'similarity' or 'imported' to distinguish protein names based on experimental evidence from those that are inferred or imported. The 'validated' and 'similarity' statuses are also used on the function and molecular complex records within entries; those records also include a hypertext-linked MEDLINE unique identifier for the article in which the experimental determination is reported.

Rule-based and classification-driven annotation. PIR uses scripts to propagate information-rich annotations among similar sequences, and to perform integrity checks based on PIR controlled vocabulary and thesaurus of synonyms or alternate names. These scripts use the superfamily/family classification system and sequence patterns and profiles to produce highly specific annotations. False positives are avoided by applying automated annotations only to classified members of the families and superfamilies for which the annotation has been validated.

Expert scientific council. The PIR Expert Scientific Council (ESC) was established last year to provide another source of expert annotation. We have enlisted over 40 scientists who agreed to assist PIR in their areas of expertise. We have also developed a separate web site for communicating with ESC members and made available to them all annotation interfaces and tools used by PIR staff members, including those still under development and testing.

Family classification

Superfamily/family. A unique characteristic of the PIR-PSD is its non-overlapping classification of full-length protein sequences based on the superfamily concept pioneered by Margaret Dayhoff *et al.* (2) and refined by PIR-International (3). To deal efficiently with the many new sequences from genome and other sequencing projects, procedures for family and superfamily classification have been automated. Shortly after entry into the database, >99% of sequences are classified at MIPS into families of closely related sequences (at least 45% identical). Sequences are further clustered at PIR into superfamilies of sequences that share end-to-end homology (but may be rather distantly related) and the same domain arrangement. In the past year superfamily classification increased from >76 000 sequences in >8900 superfamilies to >132 000 sequences in >29 000 superfamilies. In this classification procedure, very stringent (highly statistically significant) conditions are applied for placement of new entries into existing superfamilies and for definition of new superfamilies to achieve both high sensitivity and specificity of classification. This leaves ~30% unclassified entries falling in the 'twilight zone'. We have developed a procedure to classify these entries as 'associate members' (in contrast to 'full members' that are unambiguously assigned) and anticipate close to a 100% superfamily classification rate in the next quarterly release.

Domain/motif and family alignment. Sequences in PIR-PSD are also classified with homology domains that are shared by more than one superfamily and, thus, may constitute evolutionary building blocks. The MIPS-ProtFam database provides automatically generated, comprehensive multiple sequence alignments for superfamilies, families and homology domains, while the PIR-ALN database contains a smaller number of curated alignments (4). In addition to the PIR homology domains, entries are now also classified with Pfam domains (5), both based on hidden Markov models (6). The latter predictions are not directly incorporated into the PSD distribution files, but accessible from the PIR web site using IESA (see below). PROSITE (7) motifs in PSD sequences are identified using the GeneFIND program (8) and maintained in ProClass with multiple motif alignments (9).

OTHER PIR RESOURCE DEVELOPMENT

Scientists depend upon freely available resources to identify protein sequences and form testable hypotheses regarding their properties and functions but often do not know how to make best use of the resources. Many are unaware of the limitations of analysis methods and data sources and fail to evaluate the results of database searches or prediction programs so as to distinguish the probable from the improbable. The naming of translated open reading frames after the most similar sequence found in a database search has led to many misidentifications, which in turn lead to further misidentifications ('transitive identification catastrophe'). The PIR web site (10) connects many useful databases and tools to facilitate the expert annotation of protein sequences and the acquisition, extension and review of knowledge in protein science.

Integrated Environment for Sequence Analysis (IESA)

The IESA is a new web interface (http://pir.georgetown.edu/pirwww/search/piriesa.html) that combines the functions of browsing, searching and sequence similarity analysis for all PIR-PSD sequences, and links to several external databases. With IESA users may: (i) search and retrieve entries by PIR ID, superfamily/family, homology domain, species, taxonomic group, title and keyword; (ii) perform BLAST (11) or FASTA (12) searching, motif pattern matching, ClustalW (13) multiple sequence alignment, or advanced analysis such as PIR Global and Domain Search, and PIR Annotation-Sorted Search; (iii) view statistics of key PIR annotations; and (iv) select specialized sequence groups such as those in human, mouse, yeast and *Escherichia coli* genomes, or cross-referenced to PDB (14), COG (15) or Enzyme Commission (EC) numbers.

Annotation and Similarity Database (ASDB)

The PIR-ASDB is a precomputed, biweekly-updated similarity database that allows rapid retrieval of similarity and annotation information. It is based on an all-against-all FASTA search on the entire PIR-PSD. The web interface (http://pir.georgetown.edu/pirwww/dbinfo/asdb.html) displays a table of PSD entries similar to the query sequence, including similarity measures such as FASTA scores, sequence length, overlap length and percent identity, as well as annotations such as protein name, superfamily/family assignment, species and taxonomic group. Also included is a graphical display to indicate matched regions, with color code reflecting the magnitude of the

FASTA Z-score and a link to the pair-wise alignment. Data can be retrieved based on PIR ID with optional residue range, and sorted by any annotation field. The significance of the observed match can be evaluated by comparing with statistics obtained from randomly shuffled sequences. One can also generate a multiple alignment of any sequences in the table, with a display of color-coded domain regions and links to domain summaries.

Sequence Profile Search

The PIR Sequence Profile Search (http://pir.georgetown.edu/pirwww/search/pirhmm.html) allows users to search a query sequence against protein profiles (hidden Markov models) derived from PIR homology domains, ProClass/PROSITE motifs or Pfam families. A profile can also be built from user-specified sequences, and then used to search against the PIR-PSD. To speed up on-line profile search, the searches are limited to those sequences matched in ASDB for each of the sequences in the profile.

*i*ProClass

The *i*ProClass database (16) is an integrated resource that provides comprehensive 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 *i*ProClass currently consists of more than 210 000 non-redundant PIR and SWISS-PROT proteins organized with more than 29 000 PIR superfamilies, 100 000 MIPS families, 2600 PIR homology and Pfam domains, 1300 ProClass/PROSITE motifs, 280 PIR post-translational modification sites and links to over 30 databases of protein families, structures, functions, genes, genomes, literature and taxonomy.

Object-relational system

To facilitate data management and data mining, we are migrating the PIR-PSD and other auxiliary databases to Oracle 8i object-relational database management system. All new databases (such as *i*ProClass) are directly implemented in Oracle. We use both relational and object models for database design based on ER (entity-relationship) and UML modeling, and adopt a three-tier network computing architecture for database implementation to provide a framework for distributed object computing. A Java-based user-friendly web interface has been developed for querying the database and for supporting database update in both record and batch modes. With this new object-relational database system, we have greatly improved the data organization, data consistency and integrity, information retrieval, database scalability, maintainability and interoperability of our databases.

AVAILABILITY

PIR web and FTP sites

PIR provides free public access to value-added protein information through its web site at http://pir.georgetown.edu and direct file transfer at ftp://nbrfa.georgetown.edu/pir. In addition to the databases and search tools, the PIR web site provides associated metadata, including technical bulletins and documentation, which serve as dictionaries for the PIR-PSD. Accessible from the PIR anonymous FTP site are PIR-International databases and many other documents, files and software tools, including interim updates of the PSD and the corresponding sequence file. The PIR-PSD quarterly updates are also available at the NCBI FTP server. Other sites and data depositories do not always have the most recent release of the PIR-PSD. To establish reciprocal links to PIR databases or to host a PIR mirror web site, please contact pirmail@nbrf.georgetown.edu

File formats

The PIR-International has distributed the PIR-PSD as flat files in NBRF and CODATA formats, and PSD sequence file in FASTA format. Effective with Release 66.0 (September 30, 2000), the PIR-PSD is also distributed in XML (eXtensible Markup Language) format with the associated DTD (Document Type Definition) file.

ACKNOWLEDGEMENTS

PIR is a registered mark of National Biomedical Research Foundation (NBRF). The work at NBRF is supported by grant P41 LM05978 from the National Library of Medicine, and DBI-9974855 from the National Science Foundation. 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. 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.
2. Dayhoff,M.O., McLaughlin,P.J., Barker,W.C. and Hunt,L.T. (1975) Evolution of sequences within protein superfamilies. *Naturwissenschaften*, **62**, 154–161.
3. Barker,W.C., Pfeiffer,F. and George,D.G. (1996) Superfamily classification in PIR-International Protein Sequence Database. *Methods Enzymol.*, **266**, 59–71.
4. Srinivasarao,G.Y., Yeh,L.-S., Marzec,C.R., Orcutt,B.C. and Barker,W.C. (1999) Database of Protein Sequence Alignments: PIR-ALN. *Bioinformatics*, **15**, 382–390.
5. 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.
6. Eddy,S.R., Mitchison,G. and Durbin,R. (1995) Maximum Discrimination hidden Markov models of sequence consensus. *J. Comp. Biol.*, **2**, 9–23.
7. Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.
8. 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.
9. Huang,H., Xiao,C. and Wu,C.H. (2000) ProClass protein family database. *Nucleic Acids Res.*, **28**, 273–276.
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. 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.
14. 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*. (2001), **29**, 214–218.

15. Tatusov,R.L., Galperin,M.Y., Natale,D.A. and Koonin,E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, **28**, 33–36. Updated article in this issue: *Nucleic Acids Res*. (2001), **29**, 22–28.
16. 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.