© 1997 Oxford University Press

The Protein Information Resource (PIR) and the PIR-International Protein Sequence Database

David G. George*, Robert J. Dodson, John S. Garavelli, Daniel H. Haft, Lois T. Hunt, Christopher R. Marzec, Bruce C. Orcutt, Kathryn E. Sidman, Geetha Y. Srinivasarao, Lai-Su L. Yeh, Leslie M. Arminski, Robert S. Ledley, Akira Tsugita[1] and Winona C. Barker

National Biomedical Research Foundation (NBRF), Georgetown University Medical Center, 3900 Reservoir Road, NW, Washington, DC 20007, USA and [1]Japan International Protein Information Database, Science University of Tokyo, 2669 Yamazaki, Noda 278, Japan

Received September 30, 1996; Accepted October 8, 1996

ABSTRACT

From its origin, the PIR has aspired to support research in computational biology and genomics through the compilation of a comprehensive, quality controlled and well-organized protein sequence information resource. The resource originated with the pioneering work of the late Margaret O. Dayhoff in the early 1960s. Since 1988, the Protein Sequence Database has been maintained collaboratively by PIR-International, an association of macromolecular sequence data collection centers dedicated to fostering international cooperation as an essential element in the development of scientific databases. The work of the resource is widely distributed and is available on the World Wide Web, via FTP, E-mail server, CD-ROM and magnetic media. It is widely redistributed and incorporated into many other protein sequence data compilations including SWISS-PROT and the *Entrez* system of the NCBI.

INTRODUCTION

The Protein Information Resource (PIR) has evolved from the *Atlas of Protein Sequence and Structure* established in the early 1960s by the late Margaret O. Dayhoff (1–3). From its origin as a data collection designed to support research on the interrelationships and evolution of proteins, the resource was designed as a tool to support research rather than as a static library. Hence, the information is analyzed, reviewed, and reformulated to represent a dynamic view of biological knowledge. Although PIR preserves submitted and published data in its original form, PIR's mission is not to provide an archive of deposited data; rather, PIR draws on such archives to produce a corrected and current resource of sequence data in a form representative of biological concepts.

The Protein Sequence Database has been maintained by PIR-International (4–9) since 1988. The participating centers include: the Protein Information Resource (PIR) at the National Biomedical Research Foundation (NBRF) in the USA; the Martinsried Institute for Protein Sequences (MIPS) at the Max Planck Institute for Biochemistry in Germany; and the Japan International Protein Information Database (JIPID) at the Science University of Tokyo, Japan. PIR-International is unique in successfully overcoming the political, social and economic hurdles of organizing a global effort to maintain a single integrated scientific database in full collaboration.

The database contains information concerning all naturally occurring, wild-type proteins whose primary structure (the sequence) is known. A major goal of the database project is to provide comprehensive, nonredundant data uniquely organized by homology and taxonomy. In addition to sequence data, the database contains annotation information concerning: (i) the name and classification of the protein and the organism in which it naturally occurs; (ii) references to the primary literature, including information concerning the sequence determination; (iii) the function and general characteristics of the protein, including gene expression, post-translational processing and activation; and (iv) sites and regions of biological interest within the sequence. The database is also unique in maintaining consistency of annotation, with restricted vocabularies employed for features and keywords. Data are accumulated from the published literature, by submissions to PIR-International and by translation of nucleic acid sequences submitted to GenBank (10), the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database (11) and the DNA Data Base of Japan (DDBJ) (12). Entries in the database are cross-referenced to these source databases and others as described in the following sections.

The PIR-International Protein Sequence Database is widely redistributed; it is integrated into other public data sets including those assembled by the National Center for Biotechnology Information (NCBI) (13) and by SWISS-PROT (11), the protein sequence database distributed by the European Biotechnology Institute (EBI) of the EMBL. The database is also distributed by many vendors in conjunction with software packages. Although users may find these software data packages convenient, they

*To whom correspondence should be addressed. Tel: +1 202 687 2121; Fax: +1 202 687 1662; Email: pirmail@nbrf.georgetown.edu should be aware that the database supplied may not be the latest release and may not include all of the information available in the original. (The nodes of PIR-International reserve all rights on their intellectual properties and are not responsible for the versions of the database supplied by any secondary sources.)

ARCHITECTURE OF A DATABASE ENTRY

The primary unit of information in the database is an entry (see 9 for a more complete description of an entry). All relevant information concerning the same protein molecule is combined into a single entry. When the information is available, the precursor form (that corresponding to the sequence translated from the mRNA) of the molecule is represented in the database. This requires that the originally reported sequence data be transformed into the precursor form and be combined (or merged) into a single canonical form that represents the molecule as found in nature. An entry in the database is uniquely identified by the entry identification code. These codes remain unchanged from release to release provided that new information added to the database does not change the nature of the entry fundamentally. When referring to the entry as a whole, PIR entries should be cited by entry identification code.

The REFERENCE records contain information specific to each experimental report. Each reference is assigned a PIR Reference number that uniquely identifies that citation. When the reference corresponds to a report of sequence data, the REFERENCE record contains one or more PIR Accession number subrecords. The reference-specific accession numbers found in these subrecords are assigned to sequences as reported in the literature (or as generated from submitted data). We designate the reported versions of the sequence as *source sequences*. The canonical sequence represents a composite form that can be decomposed into its various constituent source sequences. The residues-specification associated with each accession subrecord constitutes an instruction for performing this decomposition. Hence, the process of merging source sequences results in reduced redundancy in the representation of sequence data without loss of the original information. Source sequences derived from nucleic acids correspond most closely to GenBank/EMBL/DDBJ coding region features (CDSs); the reference-specific accession numbers should be used to cite source sequences.

All other information in the entry refers to the canonical form of the molecule. The information is structured in records to clearly separate distinct classes of information. The GENETICS record, for example, provides information concerning the genetic origin of the protein molecule. The Features records contain information concerning transformed versions of the sequence (PRODUCT records) or regions and sites within the sequences. These data are represented in a standardized form and, when possible, are maintained consistent with information found in other databases, such as PDB, GDB, Flybase, etc.

DATABASE INTEROPERABILITY

There has been much discussion concerning the need to integrate biomolecular and genomic databases seamlessly (14–15). Effort has been devoted to addressing the technical issues of database interoperability, specifically the integration of distributed, heterogeneous resources operating under differing data models and on differing computer architectures. The World Wide Web (WWW) provides a mechanism to associate information in heterogeneous, distributed databases via hypertext links. Following this strategy, the primary task of the data center with respect to database interoperability is to establish the linkages between the information prepared at the center and that available at other data centers.

Entries in genomic and macromolecular sequence databases are complex; the attributes (data fields) are hierarchically grouped into nested entities, such as REFERENCE records, features, etc., with each subentity representing a designated class of information. When linking entries (with different internal structure) from different databases, such as GenBank and PIR, we can distinguish two types of linkages: loose-linkages and tight-linkages. Loose-links are links with undefined attributes. They relate an entry in one database to an entry in another but provide little information concerning the nature of the linkage, how it was established, or which informational entities the link associates. Links can be compiled in a variety of different ways, each approach yielding significantly different sets of links. Without knowledge of how the link was formed its significance is unknown and, hence, its utility is limited. Typically, generalized approaches are used that tend to produce over-full linkage tables (links are not missed but the links may associate unrelated information), which exhibit unexpected behaviors.

For example, while the simplest way to link GenBank and PIR entries would be to link all entries in each database that share a common citation to the literature, such links do not directly relate the sequence data. Sequences in GenBank may represent data compiled from several overlapping published sources and/or data submissions. PIR entries also are constructed from multiple data sources (generally not the same combination of sources as the corresponding GenBank entry). Linking entries by literature would result in a many-to-many relationship. In addition, publications often present more than one sequence, each represented in a different GenBank entry; likewise, several PIR entries will be associated to the same reference resulting in a combinatorial problem when relating the PIR and GenBank sequences via common literature citations. Moreover, many GenBank entries contain multiple protein coding regions each specifying a different protein represented as a distinct PIR entry.

In order to resolve these ambiguities, PIR aims to compile tight-links, i.e., links with defined characteristics that associate well-defined entities via their identifiers. Table 1 provides a list of links currently supported. It lists the identifier for the PIR entity and those of the corresponding external database entitites. Entity types in external databases are assigned unique database tags. Links appear in PIR entries as cross-references with two parts: the database tag and the entry identifier separated by a colon, e.g., TIGR:HI0664. The Database Tag specifies an entity type in an external database; the identifier specifies a particular instance of that type.

Links to Medline and the nucleic acid sequence databases are accumulated during the course of data processing. REFERENCE records corresponding to journal citations are cross-referenced directly to Medline by Medline MUID when the journal citation is found in Medline. For these citations there is a one-to-one correspondence between PIR reference Numbers and Medline MUIDs. Entries in the Protein Data Bank (PDB) (16) contain annotation information. PDB entries are treated as published references and are cited in PIR REFERENCE records. In these cases, the REFERENCE record is cross-referenced to PDB entry by PDB code.

Table 1. Correspondence between PIR Entity Identifier Types and external Database Tags

OIR Identifier Type	Database tag
PIR EntryID (entry identification code)	
PIR Reference Number	
MedLine	MUID
PDB Code	PDB
PIR Accession Number	
TIGR CDS id	TIGR
GenBank/EMBL/DDBJ CDS id	CDS_PID[a]
GenBank/EMBL/DDBJ nucleic acid sequence id	NID[a]
GenBank Accession Number	GB
EMBL Accession Number	EMBL
DDBJ Accession Number	DDBJ
PDB Code	PDB
PIR GENETICS Record	
GDB Accession Number	GDB
FlyBase Accession Number	FLYBASE[a]
ListA Accession Number	LISTA[a]

Note that PIR does not in general provide identifiers for Genetics records; this entity is listed because it corresponds most closely with the indicated external identifiers.

[a]These links are not currently activated on the PIR WWW-server, but the cross-references are available in the database.

Although many entries in GenBank, EMBL and DDBJ are composite entries (composed from overlapping reported nucleic acid sequences), new entries added to these databases generally originate from a single source. These data are submitted to the database and often a publication describing the sequence is contained in the entry. Unless it is determined otherwise, we treat new GenBank/EMBL/DDBJ entries as constituting single reports and represent the data within a single reference record (attributed to the publication, when a citation to a publication is given). GenBank/EMBL/DDBJ coding region features (CDSs) correspond to unique PIR source sequences. PIR translates CDS features and compares the computer translation with the sequence presented under the CDS/translation qualifier and, when possible, with the sequence shown in the publication. Discrepancies are noted and resolved when possible; the PIR source sequence corresponds to the corrected form. The accession-specific cross-references include the CDS_PID (taken from the /db_xREF qualifier) and the primary GenBank/EMBL/DDBJ accession number and NID of the entry in which the CDS appears. In earlier versions of GenBank, NCBI *gi* identifiers were associated with CDSs and GenBank nucleic acid sequences; some of these appear as cross-references (NCBIP and NCBIN, respectively) but will be replaced by the new GenBank identifiers. The CDS identifiers assigned by The Institute for Genomic Research (TIGR) for the *Mycoplasma genitalium* (17), *Haemophilus influenzae* (18) and *Methanococcus jannaschii* (19) genomic sequences are also included as accession-specific cross-references. There are a few instances in which sequences appearing in coordinate-data deposited in the Protein Data Bank (PDB) (16) are not available elsewhere. In these cases cross-references to the PDB entry (by PDB code) are included as accession-specific cross-references.

Gene mapping databases, such as the Genome Data Base (GDB) (20), the yeast gene name LISTA database (21) and the *Drosophila* genome database (FlyBase) (22), compile genetic information, including standardized gene names and symbols and map positions, independent of the macromolecular sequence databases. The GENETICS record of the PIR entry is structured to incorporate this information and maintain it consistently with that found in the gene mapping databases. These data are maintained in collaboration with the corresponding data center.

THE PIR WORLD WIDE WEB (WWW) SITE

The PIR WWW site provides full access to the information generated by the PIR project including: (i) a description of the project, (ii) search and retrieval access facilities to the databases, (iii) anonymous FTP, (iv) facilities for submitting data, (v) general announcements and (vi) staff contact information.

The search and retrieval facilities provide access to weekly updates of the PIR-International Protein Sequence Database and other PIR databases, such as NRL_3D (8) and the sequence alignment database (8). These facilities are grouped into four services: (i) the entry request service, (ii) the text search service, (iii) the selection list service, and (iv) the sequence search request service. The entry request service provides a facility for direct lookup of an entry by any of the entity identifiers listed in Table 1. For example, all PIR entries that include a specified PIR Reference number (or MUID) can be selected or the entry corresponding to a particular *Haemophilus influenzae* coding region can be selected by TIGR unique ID. This service also serves as a hypertext anchor point; external servers can link to specific PIR entries via any of these identifiers (please contact PIR staff for more specific details). The text retrieval service allows for entry retrieval based on Boolean AND operations over substring queries of fields such as Title, Species, Author, Publication, Keywords, Superfamily and Gene Name. The selection list service provides an alternative search mechanism by selecting terms presented on standardized lists from the Species, Keywords, Superfamily or Gene Name fields. The sequence search request service automatically packages the sequences from selected database entries as BLAST sequence requests and submits the search to the NCBI BLAST server (23).

Upon selection, HTML-formatted entries are displayed including full hypertext links to external WWW-servers. Database tags (Table 1) are treated as logical names that identify specific WWW-servers. Cross-referencing is enabled by a generic mechanism whereby new classes of links can be added and target WWW-servers added or redefined by changes in the logical name table. The sequence BLAST search request service can be invoked directly from the entry display.

ANONYMOUS FTP

Anonymous FTP services are available from nbrf.georgetown.edu. The host machine recognizes VAX/VMS directory structure. PIR anonymous FTP files are in directory DISK$BIG1:[ANONYMOUS.PIR]. The 000README. file in this directory describes the contents of the directory and the data available from the site.

THE ATLAS OF PROTEIN AND GENOMIC SEQUENCES CD-ROM

The Atlas of Protein and Genomic Sequences CD-ROM contains the ATLAS Information Retrieval System, the FASTA program (24) for similarity searches, the PIR-International Protein Sequence Database, the NRL_3D database, the PIR-ALN Protein Alignment Database, RESID Database of Residue Modifications, the PATCHX database, and the *Escherichia coli* K12 Genomic Database (7).

The ATLAS program is a fully integrated multidatabase access program that allows simultaneous access to multiple databases. Although designed primarily to handle macromolecular sequence databases, it can operate on textual databases. The program employs a multidatabase, multifield index structure. This design provides a framework that allows simultaneous retrieval from any selected set of databases and any combination of fields within those databases. ATLAS provides a user-friendly environment where entries from selected databases can be linked dynamically for simultaneous retrieval on biological annotations and bibliographic information, such as protein names, superfamily names, homology domains, organism names, gene names, keywords, feature descriptions, authors' names, etc. The ATLAS program also enables selected sets of sequences to be searched directly for exact subsequences or for patterns.

The 'Atlas of Protein and Genomic Sequences User's Guide' is updated with each release and included on the CD-ROM in both PostScript and plain text versions; it can also be obtained separately in printed form.

The Complex Carbohydrate Structure Database (CCSD) (25) and its associated CarbBank software are available on the Atlas CD-ROM. Scott Doubet of CarbBank provides the CCSD, software, and documentation to PIR-International for distribution. CarbBank is the computer management system for the CCSD database files, has a menu-driven user interface, and currently runs on PC- or MS-DOS IBM-compatible microcomputers. A Windows™ version is expected shortly. An Installation Manual and Tutorial for CarbBank can be printed from the CD-ROM.

The CD-ROM is formatted in accordance with the ISO 9660 standard and can be read from any computer system supporting this standard. The ATLAS program currently runs on PC-DOS, VAX/VMS, OpenVMS Alpha AXP, DEC UNIX Alpha AXP, DEC ULTRIX (RISC), SunOS, SGI/IRIX and Macintosh systems. The program is written in the C computer language and complies with the ANSI standard.

DATA DISTRIBUTION ON MAGNETIC TAPES

The PIR-International Protein Sequence Database with associated information is also available on a variety of magnetic media.

HOW TO OBTAIN PIR-INTERNATIONAL DATABASES AND SOFTWARE

For information on currently available database releases or other services, contact the PIR Technical Services Coordinator, National Biomedical Research Foundation, 3900 Reservoir Road, NW, Washington, DC 20007, USA. Tel: +1 202 687 2121; Fax: +1 202 687 1662; electronic mail: PIRMAIL@nbrf.georgetown.edu. In Europe, contact MIPS: Martinsried Institute for Protein Sequences, Max Planck Institute for Biochemistry, D-82152 Martinsried, Germany. Tel: +49 89 8578 2657; Fax: +49 89 8578 2655; electronic mail: mewes@mips.embnet.org. In Asia or Australia, please contact JIPID: Japan International Protein Information Database, Science University of Tokyo, 2669 Yamazaki, Noda 278, Japan. Tel: +81 471 239778; Fax: +81 471 221544; electronic mail: TSUGITA@JPNSUT31.BITNET.

ACKNOWLEDGEMENTS

This publication was supported in part by grant number P41 LM05798 from the National Library of Medicine. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Library of Medicine.

REFERENCES

1 Dayhoff,M.O., Eck,R.V., Chang,M.A. and Sochard,M.R. (1965) *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Silver Spring, MD.

2 Dayhoff,M.O. (1972) *Atlas of Protein Sequence and Structure 1972* vol. 5. National Biomedical Research Foundation, Washington, DC.

3 Dayhoff,M.O. (1979) *Atlas of Protein Sequence and Structure* vol. 5, Supplement 3. National Biomedical Research Foundation, Washington, DC.

4 Keil,B. (1989) In Colwell,R.R. (ed.), *Biomolecular Data: A Resource in Transition.* Oxford University Press, New York, pp. 27–32.

5 Mewes,H.W., George,D.G., Barker,W.C. and Tsugita,A. (1989) In Wittmann-Liebold,B. (ed.), *Methods in Protein Sequence Analysis.* Springer-Verlag, Berlin, pp. 357–360.

6 Barker,W.C., George,D.G., Mewes,H.-W., Pfeiffer,F. and Tsugita,A. (1993) *Nucleic Acids Res.* **21**, 3089–3092.

7 George,D.G., Barker,W.C., Mewes,H.-W., Pfeiffer,F. and Tsugita,A. (1994) *Nucleic Acids Res.* **22**, 3569–3573.

8 George,D.G., Barker,W.C., Mewes,H.-W., Pfeiffer,F. and Tsugita,A. (1996) *Nucleic Acids Res.* **24**, 17–20.

9 George,D.G., Hunt,L.T. and Barker,W.C. (1996) *Methods Enzymol.* **266**, 41–59.

10 Benson,D., Boguski,M., Lipman,D.J. and Ostell,J. (1996) *Nucleic Acids Res.* **24**, 1–5.

11 Rodriguez-Tomé,P., Stoehr,P.J., Cameron,G.N. and Flores,T.P. (1996) *Nucleic Acids Res.* **24**, 6–12.

12 Tateno,Y., Ugawa,Y., Yamazaki,Y., Hayashida,H., Saitou,N. and Gojobori,T. (1991) *CODATA Bull.* **23**(4), 74–75.

13 Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) *Methods Enzymol.* **266**, 141–162.

14 Weissenbach,J. and Bentolia,D. (1996) Integrating maps requires integrated data, *Nature Biotechnol.* **14**, 678.

15 Kingsburg,D.T. (1996) Consensus, common entry, and community curation, *Nature Biotechnol.* **14**, 678.

16 Abola,E.E., Bernstein,F.C., Bryant,S.H., Koetzle,T.F. and Weng,J. (1987) In Allen,F.H., Bergerhoff,G. and Sievers,R. (eds.), *Crystallographic Databases—Information Content, Software Systems, Scientific Applications.* Data Commission of the International Union of Crystallography, Cambridge, pp. 107–132.

17 Fleischman,R.D., Adams,M.D., White.O., Clayton,R.A., Kirkness,E.F., Kerlavage.A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M., *et al.* (1995) *Science* **269**, 496–512.

18 Fraser,C.M., Gocayne,J.D., White,O., Adams,M.D., Clayton,R.A., Fleischmann,R.D., Bult,C.J., Kerlavage,A.R., Sutoon,G., Kelley,J.M., *et al.* (1995) *Science* **270**, 397–403.

19 Bult,C.J., White,O., Olsen,G.J., Zhou,L., Fleischmann,R.D., Sutton,G.G., Blake,J.A., Fitzgerald,L.M., Clayton,R.A., Gocayne,J.D., *et al.* (1996) *Science* **273**, 1058–1072.

20 Keen,G., Burton,J., Crowley,D., Dickinson,E., Espinosa-Lujan,A., Franks,E., Harger,C., Manning,M., March,S., McLeod,M., *et al.* (1996) *Nucleic Acids Res.* **24**, 13–16.

21 Dölz,R., Mossé,M.-O., Slonimski,P.O., Bairoch,A. and Linder,P. (1996) *Nucleic Acids Res.* **24**, 50–52.

22 The Flybase Consortium (1996) *Nucleic Acids Res.* **24**, 53–56.

23 Madden,T.L., Tatusov,R.L. and Zhang,J. (1996) *Methods Enzymol.* **266**, 131–141.

24 Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 2444–2448.

25 Doubet,S. (1991) *CODATA Bull.* **23**(4), 56–58.