Nucleic Acids Research, 1994, Vol. 22, No. 17 3569–3573

The PIR-International Protein Sequence Database

David G.George, Winona C.Barker*, Hans-Werner Mewes[1], Friedhelm Pfeiffer[1] and Akira Tsugita[2]

Protein Information Resource, National Biomedical Research Foundation, Washington, DC, USA, [1]Martinsried Institute for Protein Sequences, Max Planck Institute for Biochemistry, Martinsried, Germany and [2]Japan International Protein Information Database, Science University of Tokyo, Noda, Japan

ABSTRACT

PIR-International is an association of macromolecular sequence data collection centers dedicated to fostering international cooperation as an essential element in the development of scientific databases. A major objective of PIR-International is to continue the development of the Protein Sequence Database as an essential public resource for protein sequence information. This paper briefly describes the architecture of the Protein Sequence Database and how it and associated data sets are distributed and can be accessed electronically.

PIR-INTERNATIONAL AND THE PROTEIN SEQUENCE DATABASE

PIR-International emerged from the concept that biological data are neither generated nor used within any single nation; hence, scientists world-wide have a legitimate stake in the development of biomolecular databases. The interests of American, European, and Japanese scientists can be served best by the participation, as equal partners, of international centers representing these concerns [1–3]. For the past five years, the contributing groups within PIR-International have worked to compile a single, fully integrated Protein Sequence Database. The close collaboration between the groups preserved the central organization of the database while allowing for decentralization of data collection, annotation, and distribution activities. PIR-International provides a model for international scientific database collaboration in that it has successfully overcome problems of data selection, quality control, implementation of standardization, and other issues that arise in the management of an international collaborative project. The participating centers include: the Protein Information Resource (PIR) at the National Biomedical Research Foundation (NBRF); the Martinsried Institute for Protein Sequences (MIPS) at the Max Planck Institute for Biochemistry; and the Japan International Protein Information Database (JIPID) at the Science University of Tokyo.

From its origin [4–6], the goal of the Protein Sequence Database project has been to maintain dynamically a comprehensive set of protein sequences that reflect current biological understanding of their forms and properties and to organize these data by homology and other biological concepts. It is the aim of the project to develop a scientific database as a tool for the investigation of questions in biological research. Achieving these goals requires analysis, evaluation, and correction of the data as part of the data-processing strategy.

The PIR-International Protein Sequence Database uses homology and comparative analysis to provide consistency and to allow inferences to be made. Comparative analysis is one of the foremost tools in biology, underlying many computerized methods of sequence analysis. Even those analyses that do not directly compare sequences and are largely empirical utilize patterns, motifs, or other modes of analysis based on comparative studies. Most of the properties attributed to proteins in the published literature have been assigned by homology.

The database contains information concerning all naturally occurring, wild-type proteins whose primary structure (the sequence) is known. In addition to sequence data, the database contains information (called annotation) concerning: (1) the name and classification of the protein and the organism in which it naturally occurs; (2) references to the primary literature, including information concerning the sequence determination; (3) the function and general characteristics of the protein, including gene expression, post-translational processing, and activation; and (4) sites and regions of biological interest within the sequence. Entries in the database are cross-referenced to other related databases, including GenBank [7], the EMBL Nucleotide Sequence Database [8], the DNA Data Base of Japan (DDBJ) [9], the human Genome Data Base (GDB) [10], the yeast chromosome map LISTA database [11], and MEDLINE. Work is currently underway to cross-reference to the *Drosophila* genome database (FlyBase) [12], the Brookhaven PDB [13], and the Complex Carbohydrate Structure Database (CCSD) of the international CarbBank project [14].

The database is widely redistributed; it is integrated into other public data sets including those assembled by the National Center for Biotechnology Information (NCBI) [15–16] and by SWISS-PROT [17] (the protein sequence database distributed by the European Molecular Biology Laboratory (EMBL) Data Library) and is incorporated into commercial products such as the Genetics Computer Group (GCG) package.

*To whom correspondence should be addressed

Data processing strategy

Experimental methods of sequence determination are error-prone. Most protein sequences are not determined directly but are inferred from the sequences of the corresponding nucleic acid coding regions, introducing additional uncertainty. As the mechanisms of gene expression are not always clearly understood, such inferences may result in serious errors in the sequence data. Annotation information (e.g., features) is obtained by a wide variety of experimental techniques, each with its own intrinsic limitations.

Moreover, biological knowledge concerning protein sequences is dynamic. As more is learned, through additional experimentation and comparative analysis with previously unknown homologous sequences, the information associated with the protein may change fundamentally. In contrast, the published literature is static. It reflects only what was known at the time of publication. Individual reports may be incomplete, inconsistent with other reports of related information, and employ nonstandardized terminology. Ubiquitin, for example, occurs in a variety of tissues in combination with a number of different proteins; its sequence has been reported separately under at least 37 different names. Unless data submitted electronically are routinely updated by the authors to reflect new understandings, submitted data will remain as static as those extracted from the published literature.

PIR-International hierarchically addresses the issues of data verification, data redundancy, and dynamic data maintenance. We consider a *source document* to be any tangible medium that conveys sequence or sequence-related data (annotation), including electronic data depositions residing at one or more data collection centers, traditional published works (journals, monographs, etc.), or other nonelectronic, unpublished bodies of work. We organize the sequence data at four levels: (1) *source sequence* data are extracted from independent source documents; (2) source sequences concerning the same protein are compared and merged to generate a *canonical sequence* constructed from the various reported forms; (3) the canonical sequences are clustered into groups of evolutionarily closely related families; (4) families of closely related sequences are further clustered into superfamilies.

The primary rationale for merging overlapping sequence data is to increase the reliability of the data. The sequence entries become representations of biological entities rather than individual experimental sequence reports. We note discrepancies among reported versions of the same sequence in order to warn users of variations or uncertainties at specific sequence positions. The information is formulated to allow the original source data to be recovered directly. Thus, redundancy in the database is reduced without the elimination of data. In addition, incongruities among closely related sequences often reveal errors such as peptide transpositions, reading-frame shifts, or the use of inappropriate genetic codes. We correct the inferred sequence data accordingly.

Annotation of sequence entries is the most time-intensive task in protein sequence data processing. By processing entries in classes rather than individually, critical processing-time factors are related to the number of superfamilies (and families) and the amount of information known about these superfamilies rather than to the number of individual source sequences. The numbers of families and superfamilies are expected to remain small relative to the number of sequences and to exhibit a slow stable growth as opposed to the exponential growth exhibited by the sequence data.

An important consideration is superfamily coverage, i.e., what fraction of the sequences are expected to fall within families or superfamilies with more than one member and therefore can be processed more efficiently according to this strategy. We estimate that 80–90% of source sequences in the current database will cluster into families or superfamilies containing at least two members. This assessment, however, reflects a history of directed sequencing efforts (targeting genomic regions known to express proteins) and may be biased accordingly. Irrespectively, nearly all available annotation information is associated with sequences that fall within homology classes. Experimental efforts generally are not directed toward understanding the properties of the products of potential coding regions until additional homologs have been discovered.

Experimental determination and verification of the characteristics and behaviors of proteins and protein sequences remains a time-intensive task. The rate at which annotation information (other than that which can be inferred by homology) is elucidated is unlikely to increase significantly; certainly it will not keep pace with the rate of nucleic acid sequence determination. The development of a methodology based on processing homologous classes of sequences will make the annotation problem tractable, thus allowing the database to be maintained by a staff that is stable and limited in size.

Conceptual database design

A formal definition of the database is essential (1) to ensure reliability of the information, (2) to enable the scientific community to contribute to the development of the database and to use the database effectively, and (3) to enable exchange of information between related databases in order to promote database interoperability. PIR-International has undertaken the task of completely and formally defining the information contained in the Protein Sequence Database. This work has been summarized in two documents: (1) PIR-International Protein Sequence Database Definition Document: The Protein Sequence Component (this document is in its final stages of preparation and will be made available by anonymous FTP; contact PIRMAIL@nbrf.georgetown.edu for further information) and (2) An Object-Oriented Sequence Database Definition Language (SDDL) [18]. The most direct effect on the user community of these developments has been some refinements in the formats of the distribution forms of the database. These changes and all subsequent ones are described in the *PIR Technical Development Bulletin*. Contact POSTMASTER@nbrf.georgetown.edu for additional information.

The conceptual model encompasses the data processing strategy and is realized through the component model of the database. Much of the ancillary information concerning the proteins is stored in separate database components. These components serve as repositories for the authoritative version of information that may be shared by a number of sequence entries, e.g., citations or formal names of organisms. The component model was adopted to promote a separation of concerns, i.e., partitioning allows database activities to be focused on specific aspects of the data at hand. In addition, it reduces redundancy in data storage.

The database components fall into six general categories: (1) source data components; (2) canonical sequence component; (3) classification component; (4) nomenclature standardization components; (5) sequence and feature verification components; and (6) database interoperability components.

The first three categories of database components reflect the data processing strategy directly. The Citation Component contains citations to all source documents of interest to the project; the Source Sequence Component serves as a repository for all available source sequence data. The Protein Sequence Component corresponds to the data currently distributed. The Class Component contains the family and superfamily classification scheme. The superfamily classification, the placement numbers, and the Alignment Database [3] provide the material to develop this component. Most other components are largely in place, which has profoundly contributed to our ability to standardize and verify the information contained in the database.

Database growth

The emergence of PIR-International and the concomitant changes in the data processing strategy have allowed the database to achieve a stable, sustained, exponential growth rate over the past three years. The Protein Sequence Database exhibits a doubling time (for number of entries) of 2.4 years. This contrasts with a doubling time of 3.5 years for SWISS-PROT (36,000 nonredundant entries as of February 1994) and for SEQDB (50,428 unmerged entries as of March 1994) from the Protein Research Foundation (Osaka). We estimate about a 30% reduction in the number of sequences when all overlapping reports of the same sequence are merged; this factor does not contribute to the calculated doubling times. Thus the PIR-International Protein Sequence Database is both the largest and the fastest growing of the curated protein sequence databases. Projections based on the growth rate since 1992 suggest that by 1999 there will be nearly 280,000 entries; we expect that an estimate of 200,000 after merging of overlapping reports is realistic.

Standardization and data verification

During the past five years, we have standardized journal abbreviations, organism names, keywords, superfamily names, enzyme names, and features [3]. We developed procedures for enforcing conformance to the standards and for controlling the introduction of new terms. The following fields are now checked by software: protein name, Enzyme Commission number for enzymes, species name (formal name), organism name (common name), special genetic code, reference number, journal abbreviation, superfamily name, and keywords.

In 1992, we began the standardization of feature annotations. Fields recording active sites, binding sites, modified sites, inhibitory sites, cleavage sites, cross-links, and disulfide bonds have now been corrected and standardized. With the cooperation of scientists at Chemical Abstracts, we have compiled a data set of types of amino acid residue modifications. This *Residues* component assists annotators in producing correct feature annotations for covalent binding sites, modified sites, and cross-links. This dataset is publicly available. To aid scientists in finding sequence patterns useful for predicting features, we have assigned to each of these features a status indicating whether the feature has been experimentally determined or predicted, or whether it has been found experimentally to be absent. Some experimentally confirmed features that are unexpected by accepted prediction methods have been marked as atypical.

Database interoperability

Recently, there has been much discussion concerning database interoperability and the need to develop effective mechanisms for interdatabase communication [19]. Effectively cross-mapping the semantics (meaning of data) implied by different database representations is now understood to be of critical importance. The most straightforward approach is to employ a common nomenclature to represent common data elements. The database interoperability components contain sets of information common to the Protein Sequence Database and other nonsequence and nonbibliographic databases. This information is maintained collaboratively with other database centers, allowing the information to be cross-checked for mutual consistency. Such collaborative arrangements have been established with the human GDB project at the Johns Hopkins University [10], the Flybase project [12], the Brookhaven PDB [13], and CarbBank [14]. The NRL_3D data set [3] provides an effective link between the Protein Sequence Database and PDB.

The role of the scientific community in database development

A major new focus of PIR-International is to involve the general scientific community more directly in Protein Sequence Database activities. The participation of the scientific community in database development and management cannot be coordinated effectively, on a large scale, without a directed effort toward the establishment of an appropriate infrastructure. Our approach is for the database centers of PIR-International to serve as focal points for these interactions. The staffs at the centers take on the role of *curators* of the database. Their essential task is to ensure that the information generated by the research community is effectively gathered, correlated, and presented in the best possible light. This approach is novel because we realize that a high level of scientific support by dedicated experts at the database centers is crucial to the success of such an endeavor. We are working toward the development of an infrastructure to support effective interaction with the scientific community by (1) precisely and formally defining the scope and content of the database, including the establishment of clear standards and policies for data submission, (2) continuing to foster interdatabase collaboration, (3) enlisting the aid of the scientific societies in organizing the interactions with the general scientific community, particularly in the area of nomenclature and data exchange standardization, and (4) establishing direct communication links with the Nomenclature Commissions of the International Council of Scientific Unions (ICSU).

Network access to PIR-International and electronic data distribution

PIR-International data are accessible through electronic data networks on a number of file-servers and FTP sites. NBRF and MIPS operate full-function network fileservers that handle database queries, sequence searches, and sequence submissions, in addition to fileserver requests.

The NBRF network request server responds to over 25 database query and general service commands. Most of the database query commands are implemented with calls to the ATLAS program. These commands provide simultaneous access to the PIR1, PIR2, PIR3, NRL_3D, PATCHX [3], Alignment, GenBank, and GBNEW databases. The sequence searching command, SEARCH, is implemented through a version of the FASTA program [20] with output routines modified for network transmission. This command also performs another, unique function: nucleotide sequences are translated in six reading frames and each polypeptide translation is submitted to a FASTA search

against the protein sequence databases. The PIR taxonomic database can be searched with the TAXONOMY command. Complete instructions for the NBRF fileserver can be obtained by sending an E-mail message containing the command HELP (in the body of the message, not on the Subject line) to FILESERV@nbrf.georgetown.edu (or FILESERV@GUNB-RF.BITNET).

The Protein Sequence Database and all other major sequence databases are also accessible at MIPS by electronic mail query and on-line connection. MIPS forwards database transactions to a propagating network of a number of nodes in Europe to provide most up-to-date information. File servers for sequence searching and for database retrieval are also available (for further information contact MIPS@ehpmic.mips.biochem.mpg.de).

Through the effort of Dan Davison at the University of Houston, each PIR-International release and the accompanying NRL_3D release are available for anonymous FTP, WAIS, and Gopher access from the University of Houston Gene-Server. All files are stored as UNIX 16-bit compressed files. An ASCII directory contains the CODATA format files and a VMS directory contains the database files and indexes in VMS format. Utilities to uncompress the data are available at that site for non-UNIX systems. The databases are also available from the NCBI FTP server.

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 for similarity searches, the PIR-International Protein Sequence Database, the NRL_3D database, the Alignment Database, the PATCHX database, and the *Escherichia coli* K12 Genomic Database [21].

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 single 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. With each release of the CD-ROM, the 'Atlas of Protein and Genomic Sequences Installation and User's Guide' is updated. Although included on the CD-ROM, it can also be obtained separately in printed form.

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, OSF/1 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 is distributed four times yearly on magnetic media in VAX/VMS and ASCII formats. The data and programs are available on 9-track tape, on TK50 and TK70 streaming tape cartridges, and on DAT 4mm cartridges. VMS users have the option of VAX/VMS copy or backup formats. The tapes currently include four protein sequence datasets (PIR1, PIR2, PIR3, and NRL_3D); files of associated information such as taxonomic classification of organisms, special genetic codes, list of superfamilies; update information; extensive indexes; and documentation. In addition, the VAX/VMS tape includes the PSQ and NAQ retrieval systems, software to create databases that can be accessed by PSQ or NAQ from user-supplied sequence entries or from GenBank or EMBL databases, and software to create the ASCII flat-file version of the database from the VAX/VMS version. ASCII card image format tapes do not include retrieval software; however, files are supplied containing indexes to authors, accession numbers, reference numbers, species, superfamily names, citations, keywords, and features.

The PIR-International Protein Sequence Database is distributed by many other vendors in conjunction with software packages. The nodes of PIR-International reserve any rights on their intellectual properties and are not responsible for the versions of the database supplied by any secondary sources. Although users may find these software data packages convenient, they 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.

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, D.C. 20007, USA; telephone +1 202 687-2121; FAX +1 202 687-1662; electronic mail PIRMAIL@nbrf.georgetown.edu. In Europe, contact MIPS: Martinsried Institut fur Proteinsequenzen, Max-Planck-Institut fur Biochemie, D-82152 Martinsried bei Munchen, Germany; telephone +49 89 8578 2657; FAX +49 89 8578 2655; electronic mail MIPS@ehpmic.mips.biochem.mpg.de. In Asia or Australia, please contact JIPID: Japan International Protein Information Database, Science University of Tokyo, 2669 Yamazaki, Noda 278, Japan; telephone +81 471 239778; FAX +81 471 221544; electronic mail TSUGITA@JPNSUT31.BITNET and EX5292@JPNSUT30.BITNET.

ACKNOWLEDGEMENTS

We thank Thomas Schneider of the Frederick Cancer Center for assistance in analyzing the database growth figures. PIR-International staff members with principal responsibility for distributed data sets include Friedhelm Pfeiffer (PATCHX), John S. Garavelli (NRL_3D and Residues), Geetha Srinivasarao and Lai-Su Yeh (Alignment Database), and T. Kunisawa (Escherichia coli K12 Genomic Database). The Protein Identification Resource is partially supported by National Institutes of Health Grant LM05206. Development of the ATLAS CD-ROM was partially supported by a grant from Digital Equipment Corporation. MIPS is supported by the Max-Planck-Gesellschaft, the Forschungszentrum f. Umwelt und Gesundheit (GSF), and the European Economic Community BRIDGE Programme Grants BIOT-CT-0167 and 0172.

REFERENCES

1. Keil, B. (1989) in Biomolecular Data: A Resource in Transition, Colwell, R.R. (ed.) Oxford University Press, New York, pp. 27–32.
2. Mewes, H.W., George, D.G., Barker, W.C., and Tsugita, A. (1989) in Methods in Protein Sequence Analysis, Wittmann-Liebold, B. (ed.) Springer-Verlag, Berlin, pp. 357–360.
3. Barker, W.C., George, D.G., Mewes, H.-W., Pfeiffer, F., and Tsugita, A. (1993) Nucl. Acids Res. 21, 3089–3092.
4. 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.
5. Dayhoff, M.O. (1972) Atlas of Protein Sequence and Structure vol. 5. National Biomedical Research Foundation, Washington, DC.
6. Dayhoff, M.O. (1979) Atlas of Protein Sequence and Structure vol. 5, Supplement 3. National Biomedical Research Foundation, Washington, DC.
7. Benson, D., Lipman, D.J., and Ostell, J. (1993) Nucl. Acids Res. 21, 2963–2965.
8. Rice, C.M., Fuchs, R., Higgins, D.G., Stoehr, P.J., and Cameron, G.N. (1993) Nucl. Acids Res. 21, 2967–2971.
9. Tateno, Y., Ugawa, Y., Yamazaki, Y., Hayashida, H., Saitou, N., and Gojobori, T. (1991) CODATA Bull. 23(4), 74–75.
10. Cuticchia, A.J., Fasman, K.H., Kingsbury, D.T., Robbins, R.J., and Pearson, P.L. (1993) Nucl. Acids Res. 21, 3003–3006.
11. Linder, P., Doelz, Mosse, M.-O., Lazowska, J., and Slonimski, P.P. (1993) Nucl. Acids Res. 21, 3001–3002.
12. Merriam, J., Ashburner, M., Hartl, D.L., and Kafatos, F. (1991) Science 254, 221–225.
13. Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F., and Weng, J. (1987) In Crystallographic Databases — Information Content, Software Systems, Scientific Applications, Allen, F.H., Bergerhoff, G., and Sievers, R. (eds.) Data Commission of the International Union of Crystallography, Cambridge, pp. 107–132.
14. Doubet, S. (1991) CODATA Bull. 23(4), 56–58.
15. Benson, D., Boguski, M., Lipman, D.J., and Ostell, J. (1990) Genomics 6, 389–391.
16. Benson, D. (1991) CODATA Bull. 23(4), 76–78.
17. Bairoch, A., and Boeckman, B. (1993) Nucl. Acids Res. 21, 3093–3096.
18. George, D.G., Orcutt, B.C., Mewes, H.-W., and Tsugita, A. (1993) Protein Seq. Data Anal. 5, 357–399.
19. Snoddy, J., Robbins, R., and Adamson, A. (1993) Human Genome News 5(3), 1–4.
20. Pearson, W.R., and Lipman, D.J. (1988) Proc. Nat. Acad. Sci. USA 85, 2444–2448.
21. Kunisawa, T., Nakamura, M., Watanabe, H., Otsuka, J., Tsugita, A., Yeh, L.-S.L., George, D.G., and Barker, W.C. (1990) Protein Seq. Data Anal. 3, 157–162.