Nucleic Acids Research, Vol. 20, Supplement 2023–2026

The PIR-International Protein Sequence Database

Winona C.Barker, David G.George, Hans-Werner Mewes[1] and Akira Tsugita[2]
Protein Identification Resource, National Biomedical Research Foundation, Washington, DC, USA, [1]Martinsried Institute for Protein Sequences, Max Planck Institute of Biochemistry, Martinsried, Germany and [2]International Protein Information Database of Japan, Science University of Tokyo, Noda, Japan

PIR-INTERNATIONAL

PIR-International is a tripartite association of protein sequence data centers: PIR-America; PIR-Europe; and PIR-Asia. PIR-America is part of the National Biomedical Research Foundation (NBRF) at Georgetown University, Washington DC, USA, and has evolved from the Protein Identification Resource project; it is now directed by Dr. Winona C.Barker. PIR-Europe is part of the Martinsried Institute for Protein Sequences (MIPS) at the Max Planck Institute of Biochemistry, Martinsried, Germany, and is directed by Dr. Hans-Werner Mewes. PIR-Asia is part of the International Protein Information Database of Japan (JIPID) at the Science University of Tokyo and is directed by Professor Akira Tsugita. The unique goals of PIR-International are (1) to dynamically maintain a complete and comprehensive set of experimentally determined protein sequences and related information in accordance with current biological understanding and (2) to organize these data by similarity and evolutionary relationship.

The success of the PIR-International collaboration is seen both in the growth of the database and in the increase in the contributions of MIPS and JIPID. From December 1988, when MIPS and JIPID began contributing substantial amounts of data, until December 1991, the number of entries in the Protein Sequence Database increased 3.4-fold. The comparable increase in the GenBank® /EMBL/DDBJ nucleotide sequence database was 2.8-fold. As of Release 31.0 (December 1991), about 40% of the data in the 36,150 database entries had been contributed by MIPS and JIPID.

THE PROTEIN SEQUENCE DATABASE

The Protein Sequence Database was initiated in the early 1960's by Margaret O.Dayhoff [1–3] for the express purpose of providing a dataset for the support of research on the interrelationships and evolution of proteins. Supporting such research requires that an up-to-date representation of the scientific understanding of the information be maintained. The central goal of the database project has been to maintain this perspective by merging all sequence data corresponding to a particular protein into one representative structure that is consistently annotated with information describing its genetic origin and expression and its biological function.

The Protein Sequence Database is currently divided into three datasets that are different in their organization, information content, and degree of verification [4]. Entries in Section 1 are highly verified, nonredundant, annotated with additional information, and organized by function and sequence similarity. Data from multiple determinations of the sequence of the same molecule are combined into a single entry. In contrast, entries in Section 3 each correspond to a single report of a sequence and only minimal information is given: entry title, accession number (which also serves as the entry identification code), the citation for the publication or submission and a unique 'reference number' assigned to it, and the sequence. Optionally, the formal nomenclature for the species and a cross-reference to the nucleotide sequence databases may be given. These entries are marked as unverified because they have not been reviewed. After such review, entries are moved to Section 2. Entries in Section 2 may be merged, annotated, or classified into protein superfamilies but some of the additional information may not have been subjected to critical review.

As the data in PIR3 are processed and moved to PIR1 and PIR2, a copy of the originally reported sequence is permanently stored in an ARCHIVE. If errors are noted in later processing stages, they are corrected so that an accurate permanent record can be maintained. Thus PIR1, PIR2, and PIR3 contain all the sequence information in its most processed form. PIR3 and the ARCHIVE constitute the set of originally reported sequences. Newly entered data are checked for duplication against the information in PIR3 and in the ARCHIVE, which allows detection and elimination of duplicate data at the point of entry to the dataset.

THE SUPERFAMILY CONCEPT AND PLACEMENT IN THE DATABASE

A superfamily is a group of proteins that share sequence similarity due to common ancestry [5, 6]. From the mid-1970's until recently, the entries in the Protein Sequence Database were organized into superfamilies, families, subfamilies, entries, and subentries by assigning to each entry a set of numbers that uniquely specified both its order (or placement) in the database and its relationship to other entries [4]. This hierarchical organization has proved to be very fruitful and has been employed in a variety of investigations [7–10]. However, this structure was designed before it was recognized that a given protein can belong to more than one superfamily and that similarity among proteins within the same superfamily may be restricted to regions or domains.

In its original formulation, the superfamily classification was based on the assumption that related proteins are usually similar along their entire lengths; this provided no satisfactory mechanism for expressing the complex relationships among proteins that are mosaics of domains that may be found also in otherwise unrelated proteins. In order to group proteins into a superfamily it was necessary to ignore portions of the molecules that were dissimilar. Classification was specified by a scientist after examining the results of various computer methods for detecting and measuring sequence relatedness; this time-consuming procedure was done only for the fully annotated set of entries (PIR1).

We have addressed these problems by decoupling the concept of superfamily classification from relative placement. As currently employed, the term superfamily refers to protein domains rather than to the entire protein (although, in most cases, the domain extends the entire length of the protein). Superfamily relationships are now indicated in a special 'Superfamily': data item, which has been added to all entries in PIR1 (except hypothetical proteins).

This data item may contain one or more superfamily names. Because it is much easier to recognize that two or more domains are related than to express these relationships in a hierarchical classification, the decoupling of superfamily assignment from the hierarchical classification system has markedly improved the speed and efficiency with which superfamily assignments can be made.

The system of assigning numbers to indicate placement of a sequence within a group of similar sequences has been retained as a mechanism to order the entries. The ordering of the sequence entries by number provides a natural structure for obtaining an overview of the types of proteins present in the database and for assessing the depth of data available concerning specific groups of proteins. Sequences are assigned to the same placement group when they can easily be demonstrated to be homologous over the majority of their lengths. These new criteria will allow existing methodology to be adapted for automatic placement assignment.

STANDARDIZATION WITHIN AND AMONG DATABASES

PIR-International has embarked upon a multifaceted effort toward greater standardization within the Protein Sequence Database. Crucial to this effort was the fuller characterization of the information by partitioning it into well-defined fields. This allows the terminology standardization project to be broken into a series of manageable tasks that can be distributed among the PIR-International centers and outside collaborators. We have been developing controlled vocabularies for use in many of these fields. This terminology reflects the usage recommended by the international nomenclature commissions and other authoritative sources when applicable. These standardization efforts are being coordinated with those of other related database projects and with the CODATA Commission on Standard Terminology for Access to Biological Data Banks [12].

At the instigation of the PIR-International, the major protein and nucleic acid database centers adopted a common set of journal abbreviations, the list developed in conformance with certain international standards by the National Library of Medicine. Authors and article titles have also been standardized so that all citations to a given article are identical. Each article is identified with a unique reference number [4], which can be used to retrieve all data from a given article in the PIR-International datasets; these numbers are being correlated with the Medline abstract identifiers.

The standardization of species names is difficult because there is no single authority to consult, different authorities may disagree, and accepted names are generally not available in computer-readable form. Furthermore, accepted names are regularly revised to reflect new ideas about classification of organisms. A staff taxonomist, Dr. Andrzej Elzanowski, maintains the list of accepted species names and the taxonomic classification scheme used in the Protein Sequence Database. The PIR-International taxonomic listing of species is distributed with the database and shared with a number of other database centers, including GenBank [13], EMBL [14], BioMagResBank [15], the Brookhaven Protein Data Bank [16], and NCBI [17].

We have begun the standardization of protein names with the enzymes because they are classified and named by an international commission. We created computer files containing the EC numbers, recommended names, and alternate names in the 1984 edition of Enzyme Nomenclature [18], as modified by corrections and additions that have since been published [19–21]. (As with species names, a computer-readable version of this reference work was not available at the time.) We also drafted a policy on punctuation, word order, and representation of Greek letters and of super- and subscripts, in the protein names. The names of all enzymes with assigned EC numbers were standardized in the database. For most entries it is possible to use the recommended name, followed by the EC number, followed by any needed modifiers. Other commonly used names are placed in the 'Alternate names:' field.

The standardization of keywords is well under way. Over 1700 keywords that appeared in this previously uncontrolled field were examined to locate synonyms, variations of the same term, and unsuitable terms. (We decided, for example, that keywords will not give taxonomic information or homology information, both of which are given in other defined fields.) Keywords convey information in certain categories including protein function and activity, location at which the protein is active or synthesized, higher order structure of the protein, including multiple chains and cofactors. However, the meaning of keywords can be ambiguous. For example, 'mitochondrion' may refer to the compartment where the protein functions, the compartment where the protein is synthesized, or the genome in which the protein is encoded. To more clearly convey such information, as well as that now appearing in free text comments, the information will be partitioned into distinct fields over the next few years. Standardization of terminology for features is currently underway.

In certain fields, the Protein Sequence Database incorporates terminology standards used or set by other databases. In particular, the gene symbols, gene names, and map positions from the appropriate genomic databases are being added. Collaborations have been initiated with the Genome Data Base for human genes [22] at The Johns Hopkins University and with Michael Ashburner in Cambridge, England, who maintains the Drosophila Genome Data Base (Flybase) [23]. In each case, a set of common data (gene symbol, map position, gene name, protein name) will be collaboratively maintained. These data will provide the basis for cross-linking information in the databases. Human and *Drosophila* genetic information in PIR entries will be standardized first, followed by other species as similar collaborations can be established.

WORK IN PROGRESS

Comprehensive annotation of the protein sequence data entered into the Protein Sequence Database in the future will be possible only by the help of suitable software to support the systematic evaluation of data associated with the primary sequence information. To achieve this, a well-defined data structure is an absolute requirement. We have developed a Sequence Database Definition Language (SDDL) for this purpose [24]. Within the context of this language a new format for data distribution and exchange is being developed; this format is an extension of the CODATA format [25]. This approach will allow us to develop a much more flexible representation scheme, while minimizing the effect on the user community. In particular, the implementation of the format will allow us to add additional information to the database, such as processing status, and will result in eliminating the separation of the dataset into three sections. The implementation of the new format and the software being developed will allow us to develop components of associated information (e.g., taxonomy, genetic information). These components will improve quality and consistency.

COMPLEMENTARY ROLE OF THE PROTEIN SEQUENCE DATABASE AND GENINFO®

The National Center for Biotechnology Information (NCBI) of the National Library of Medicine (NLM) is developing GenInfo, a three-level database system. The foundation level will be a stable repository that reflects the published literature as collected by the NCBI, supplemented with directly submitted data as collected by the GenBank staff at Los Alamos National Laboratory (GenBank will be administered by the NCBI beginning in the fall of 1992). At this level, protein and nucleic acid sequence data will be recorded exactly as published and as submitted by the authors without interpretation, refinement, or standardization of nomenclature. Linkages will be maintained to the publication and to its Medline abstract and between coding regions and their conceptual translations. No additional organization will be imposed on the data. Additional information concerning proteins will be included in the foundation level only when it is directly submitted by the research investigators (David Lipman, personal communication). The second level will be the value-added information from the PIR-International database and other comprehensive sources; the third level will comprise sequence-related information generated by other specialized database projects. The NCBI data collection is designed as a stable foundation upon which other databases are built and to which they remain linked; thus, the success of the 'backbone' is largely dependent upon the continued development of the Protein Sequence Database and other value-added datasets. In this respect, PIR-International will directly complement and greatly enhance the activities of the NCBI.

The NCBI GenInfo initiative will supplant some of the activities currently conducted by the PIR-International, including locating articles with sequence data (although journals and books must still be scanned for annotation information) and direct entry of the sequence data. PIR is working with NCBI to develop software to import the NCBI backbone data (from its ASN.1 representation) into the PIR ARCHIVE. The duplication detection mechanisms within the PIR system will allow information previously entered or entered from other sources to be correlated and cross-referenced to the GenInfo backbone (including data submitted to the Los Alamos National Laboratory). Except for the efforts required to analyze and resolve discrepancies, the system will be automatic.

DATA DISTRIBUTION ON MAGNETIC TAPES AND CD-ROM

The databases and programs of PIR-International are distributed on 9-track magnetic tape and TK50 and TK70 cartridges, in VAX/VMS format and in ASCII card image format; the databases are distributed quarterly. VAX/VMS format tapes of the PIR Protein Sequence Database include the Protein Sequence Query (PSQ) database retrieval program [26]. Also included on this tape are programs for creating databases that can be used by PSQ from user-supplied files of entries. We also distribute the NRL 3D database [27] of sequence information extracted from the Brookhaven Protein Data Bank and formatted for use with the PIR sequence analysis software and PSQ program. Output from PSQ can be used with standard molecular modeling programs in conjunction with the Brookhaven Protein Data Bank to display the 3-D structure of identified sequences. 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. Tapes in both formats contain documentation as well as additional files containing the species names (ordered in a taxonomic hierarchy), journal abbreviations, and special genetic codes used in the database. In 1992, we will begin distribution of a CD-ROM containing the PIR Protein Sequence Database, GenBank, and a powerful retrieval program that allows simultaneous text searching of these databases.

The PIR-International Protein Sequence Database has been incorporated into or used as the primary source for other protein sequence databases and is also distributed by many other vendors in conjunction with software packages. The PIR-International is not responsible for the versions of the database supplied by these 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.

ON-LINE ACCESS AND E-MAIL SERVERS

Up-to-date databases and sequence analysis software are accessible on-line from PIR and MIPS, and similar services will be offered by JIPID in mid-1992. In addition to PIR-International, on-line users have access to recent releases of several other databases such as the EMBL Nucleotide Sequence Data Library, the GenBank Genetic Sequence Databank, and a merged protein database (MIPSX). The sequence databases are accessed by a multidatabase retrieval program that not only combines the capabilities of its predecessors, PSQ and NAQ, but can also simultaneously access any or all databases available on the system, thereby eliminating the need to repeat the same query in each database separately. Daily updates of new protein sequence data are made available in a separate retrieval system.

The PIR-International Protein Sequence Database is also accessible by electronic mail query to network fileservers at PIR and MIPS. JIPID is scheduled to install an E-mail server in 1992. Complete instructions for the PIR 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@GUNBRF.BITNET. MIPS operates separate file servers for sequence searching, SEARCH@MIPS.mpg.dbp.de, and for database retrieval, RETRIEVE@MIPS.mpg.dbp.de.

HOW TO OBTAIN PIR-INTERNATIONAL DATABASES, SOFTWARE, AND NEWSLETTERS

For information on currently available database releases, or other services, or for a copy of the PIR Newsletter, contact the PIR Technical Services Coordinator, National Biomedical Research Foundation, 3900 Reservoir Road NW, Washington, D.C. 20007; telephone +1 202 687−2121; FAX +1 202 687−1662; electronic mail PIRMAIL@GUNBRF.BITNET. In Europe, contact MIPS: Martinsrieder Institüt für Proteinsequenzen, Max-Planck-Institüt für Biochemie, D-8033 Martinsried bei München, FRG; telephone +49 89 8578 2656; FAX +49 89 8578 2655; electronic mail MEWES@vax1.mips.mpg.dbp.de. In Asia or Australia, please contact JIPID: International Protein Information Database in Japan, Science University of Tokyo, 2641 Yamazaki, Noda 278, Japan; telephone +81 471 239778; FAX +81 471 221544; electronic mail TSUGITA@JPNSUT31.BITNET and EX5292@JPNSUT30.BITNET.

ACKNOWLEDGMENTS

The Protein Identification Resource is supported by National Institutes of Health Grant LM05206 and National Science Foundation Grant DIR9107540. Improvements to our on-line system were made possible by a grant from Digital Equipment Corporation. MIPS is supported by Bundesministerium für Forschung und Technologie Grant 0319149A and European Economic Community BRIDGE Programme Grants BIOT-CT-0167 and 0172.

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* 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. Barker,W.C., George,D.G., Hunt,L.T. and Garavelli,J.S. (1991) *Nucleic Acids Res.* **19**, 2231−2236.
5. Dayhoff,M.O., McLaughlin,P.J., Barker,W.C. and Hunt,L.T. (1975) *Naturwissenschaften* **62**, 154−161.
6. Dayhoff,M.O. (1976) *Fed. Proc.* **35**, 2132−2138.
7. Pearson,W.R. (1991) *Genomics* **11**, 635−650.
8. Seto,Y., Ikeuchi,Y. and Kanehisa,M. (1990) *Proteins* **8**, 341−351.
9. Lipman,D.J. and Pearson,W.R. (1985) *Science* **227**, 1435−1441.
10. Guigo,R., Johansson,A. and Smith,T. (1991) *CABIOS* **7**, 309−315.
11. Barker,W.C., George,D.G. and Hunt,L.T. (1990) *Meth. Enzymol.* **183**, 31−49.
12. Blaine,L. (1991) *CODATA Bull.* **23**(4), 38−46.
13. Burks,C., Cassidy,M., Cinkosky,M.J., Cumella,K.E., Gilna,P., Hayden,J.E.D., Keen,G.M., Kelley,T.A., Kelly,M., Kristofferson,D. and Ryals,J. (1991) *Nucleic Acids Res.* **19**, 2221−2225.
14. Kahn,P. and Cameron,G. (1990) *Meth. Enzymol.* **183**, 23−31.
15. Markley,J.L., Seavey,B.R. and Farr,E. (1991) *CODATA Bull.* **23**(4), 85−88.
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. Benson,D., Boguski,M., Lipman,D.J. and Ostell,J. (1990) *Genomics* **6**, 389−391.
18. Nomenclature Committee of the International Union of Biochemistry (1984) *Enzyme Nomenclature*. Academic Press, Orlando.
19. Nomenclature Committee of the International Union of Biochemistry (1986) *Eur. J. Biochem.* **157**, 1−26.
20. Nomenclature Committee of the International Union of Biochemistry (1989) *Eur. J. Biochem.* **179**, 489−533.
21. Nomenclature Committee of the International Union of Biochemistry (1990) *Eur. J. Biochem.* **187**, 263−281.
22. Pearson,P.L. (1991) *Nucleic Acids Res.* **19**, 2237−2239.
23. Merriam,J., Ashburner,M., Hartl,D.L. and Kafatos,F. (1991) *Science* **254**, 221−225.
24. George,D.G., Orcutt,B.C., Mewes,H.-W. and Tsugita,A., manuscript in preparation.
25. George,D.G., Mewes,H.W. and Kihara,H. (1987) *Protein Seq. Data Anal.* **1**, 27−39.
26. Orcutt,B.C., George,D.G. and Dayhoff,M.O. (1983) *Annu. Rev. Biophys. Bioeng.* **12**, 419−441.
27. Pattabiraman,N., Namboodiri,K., Lowrey,A. and Gaber,B.P. (1990) *Protein Seq. Data Anal.* **3**, 387—405.