 Nucleic Acids Research, 1998, Vol. 26, No. 1 27–32

The PIR-International Protein Sequence Database

Winona C. Barker*, John S. Garavelli, Daniel H. Haft, Lois T. Hunt, Christopher R. Marzec, Bruce C. Orcutt, Geetha Y. Srinivasarao, Lai-Su L. Yeh, Robert S. Ledley, Hans-Werner Mewes[1], Friedhelm Pfeiffer[1] and Akira Tsugita[2]

Protein Information Resource, National Biomedical Research Foundation, 3900 Reservoir Road, NW, Washington, DC 20007, USA, [1]Munich Information Center for Protein Sequences, Max Planck Institute for Biochemistry, Martinsried, Germany and [2]Japan International Protein Information Database, Science University of Tokyo, Noda, Japan

Received October 2, 1997; Accepted October 8, 1997

ABSTRACT

From its origin the Protein Information Resource (http://www-nbrf.georgetown.edu/pir/) has supported research on evolution and computational biology by designing and compiling a comprehensive, quality controlled, and well-organized protein sequence database. The database has been produced and updated on a regular schedule since 1984. Since 1988 it has been maintained collaboratively by the PIR-International, an association of data collection centers engaged in international cooperation for the development of this research resource during a period of explosive acquisition of new data. As of June 1997, essentially all sequence entries have been classified into families, allowing the efficient application of methods to propagate and standardize annotation among related sequences. The databases are available through the Internet by the World-Wide Web and FTP, or on CD-ROM and magnetic media.

THE PROTEIN INFORMATION RESOURCE

The Protein Information Resource (PIR) was established by the National Biomedical Research Foundation (NBRF) in 1984 as a resource to assist in the identification and interpretation of protein sequence information (1). The PIR evolved from the original NBRF Protein Sequence Database, developed over a 20-year period by the late Margaret O. Dayhoff and published as the 'Atlas of Protein Sequence and Structure' (2,3). PIR-International is a collaboration established in 1988 between the NBRF, the Munich Information Center for Protein Sequences (MIPS), and the Japan International Protein Information Database (JIPID) to collect and publish what is now the oldest database of biomolecular sequence, source, bibliographic and feature information, the PIR-International Protein Sequence Database. The mission of the PIR-International remains: (i) to create and maintain the Protein Sequence Database as a comprehensive, well verified collection, organized according to biological principles, including structural, functional and evolutionary relationships; (ii) to distribute the database by the most accessible means including the Internet, CD and tape media; (iii) to provide a research tool that supports the study of protein sequences, their structural and functional properties, and their biological origins; and (iv) to collaborate with other databases in organizing and coordinating the presentation of biomolecular structural information (4).

UNIQUE FEATURES OF THE PIR-INTERNATIONAL PROTEIN SEQUENCE DATABASE

The PIR-International Protein Sequence Database is unique among comprehensive public domain protein sequence databases in the following respects.

- As of June 1997, essentially all sequence entries are classified into families as discussed below.
- Full citations, including the titles of articles cited, are given; these are hypertext linked to Medline abstracts and associated information in the Entrez system of the National Center for Biotechnology Information (NCBI).
- Each separately reported sequence is represented in a manner that clearly shows any differences from the sequence shown in the entry (the canonical sequence) and that allows the reported sequence to be reconstructed automatically and, if desired, submitted for searches.
- Cross-references to the nucleotide sequence databases are directly associated with the cited sequence to which they refer. The nucleotide database entry accession, NID, and PID are given and serve as hypertext links to the corresponding GenBank or EMBL database entries. These data are checked, and updated as necessary, after each release of GenBank.
- The most complete and current genetic information is provided, including map position, intron positions, and start codon (if different from AUG), along with hypertext links to genome databases.
- Feature annotations are represented with greater accuracy and consistency because standardized format and restricted vocabulary are enforced. The Guide for Features Annotations is available on the PIR Web site.
- Terminology in most retrievable fields has been standardized and restricted vocabularies are enforced. Lists of the current vocabularies can be viewed or retrieved through the PIR Web site.
- The PIR-International Protein Sequence Database contains more citations and more up-to-date data.

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

- It consistently adheres to its announced update schedule. The database has been updated and publicly released 4 times per year for the last 13 years.
- Although our policy is to merge all independent reports of the same sequence into one annotated entry, we do not withhold entries from public view until they are fully merged and annotated. All sequence data that we obtain are available to the public as soon as they are available to the PIR staff.
- Through our Web site and online system, public access is provided to the interim updates normally prepared on a weekly basis.

DATABASE ORGANIZATION

The Protein Sequence Database is partitioned into several sections, currently PIR1, PIR2, PIR3 and PIR4. The section in which an entry appears can be adjusted during an update; however, in the quarterly releases of the database an entry code is unique across all sections. Therefore, we recommend that Protein Sequence Database entries be identified by the tag 'PIR:'. While the appearance of an entry in PIR3 is indicative of its processing status, in all other database sections the processing status of a particular sequence report is indicated by a 'Status' record for the accession.

Entries in PIR1 are fully classified by superfamily assignment and placement numbers, as discussed below. They are fully merged with respect to other entries in PIR1 according to our current understanding, and they are generally annotated; >90% of PIR1 entries have keywords and 70% have features. There is no clear distinction between entries in PIR1 and PIR2; many entries in PIR2 are merged, classified and annotated as fully as typical PIR1 entries. For all entries in PIR1 and PIR2, sequence translations have been verified and bibliographic information has been checked. All entries in PIR1 are fully classified. Entries in PIR2 are classified by family, and they may also be classified by superfamily and either partial or complete placement numbers. Entries with complete placement numbers are fully merged with respect to all other entries in PIR1 and PIR2 with full placement numbers.

Entries in PIR3 have not been subjected to verification of sequence and bibliographic information and they are not classified, merged, or annotated. During the past 2 years, the backlog of entries in PIR3 has been reduced from nearly 23 000 entries to an average of <2000 in any interim update, <2% of the total database. Thus, a far greater proportion of the database is now fully merged, annotated, and classified. We estimate that redundancy, measured as the percentage of entries that could disappear by merging into other entries, has been reduced from >30% 2 years ago to ~10% at the present time. Most of the current redundancy arises either from newly introduced entries or from fragmentary sequences that are difficult to detect as merge candidates with the existing automatic methods.

Finding duplicate sequence reports and performing the necessary merges is a major priority at PIR-International. In merged entries, the differences between reports are documented, conflicting data are reconciled to give the best canonical sequence for the protein, representations for reconstructing the reported sequences are provided, and the individual sources of information about a single protein are recorded. Incomplete and inaccurate entries are replaced in favor of more complete and accurate merged entries.

Several years ago the PIR4 section was introduced for the sake of comprehensiveness, to include sequences identified as not naturally occurring or naturally expressed. Entries in PIR4 have been carefully reviewed and are annotated. This section includes sequences known to be conceptual translations of pseudogenes or otherwise unexpressed potential ORFs that may have mistakenly been assigned identifiers as coding regions. It also includes engineered or synthetic sequences, sequences resulting from fusion, cross-over or frameshift mutations, and sequences of natural polypeptides that are not ribosomally synthesized. No active effort is being made to collect these data. However, as they are acquired during routine data processing and identified, they will be stored and made available in PIR4.

Changes in the structure of the Protein Sequence Database and other PIR databases and in the formats of their distributed flat files are announced in the *PIR Technical Development Bulletin*. This bulletin is distributed by Email ~1 month before the quarterly release of the database in which the changes are scheduled. For information about this bulletin and other PIR publications, contact the PIR Technical Services Coordinator.

FAMILY AND SUPERFAMILY CLASSIFICATION

The concept of protein superfamilies was originally proposed by Margaret Dayhoff (5,6) and was later refined and developed into a formal model by PIR-International (7,8). In the refined model, two classes of superfamilies are defined: homeomorphic superfamilies and domain superfamilies. For classification into homeomorphic superfamilies, sequences are compared from the N- to the C-end. Members of a homeomorphic superfamily are similar from end to end. Members of a domain superfamily share a given type of homology domain, such as 'protein kinase homology'. A completely sequenced protein can be a member of only one homeomorphic superfamily, which permits the database to be partitioned into non-overlapping sets.

Superfamily membership is indicated by name in the 'Superfamily' record of an entry. Even when an entry sequence is fragmentary, the names of all the homology domains found in a full-length member of the superfamily are listed in the 'Superfamily' record. (Homology domains are also annotated as sequence features.) For convenience, we have transiently grouped sequences containing certain homology domains into 'superfamilies' with names such as 'unassigned homeobox proteins' and 'unassigned ATP-binding cassette proteins'. Some of these sequences are not currently classifiable because they are incomplete. When viewing an entry on the PIR Web site, the list of all other entries in the same superfamily can easily be obtained and displayed under 'Associated information'. It is also possible to extract the sequences of defined homology domain features and submit them for searches.

In another approach to classification, many entries have been assigned a 'placement number' that is used to order entries by functional and structural relatedness. Placement numbers are not permanent. Between major releases, placement numbers are assigned that position a new family or superfamily among existing ones; at each major release, the entire set of placement numbers in the database is reassigned.

Dr Friedhelm Pfeiffer at MIPS has clustered 93% of the sequences in the PIR database into families whose members have ~50% or more sequence identity. Less than 5% of entries in the database were not classifiable, usually because they were too

short or fragmentary, and only ~2% of entries were not fully analyzed. Over 10 000 alignments of families containing at least two sequences are available at the MIPS Web Site (http://www.mips.biochem.mpg.de/). Every family classified in this way has been assigned a permanent identification number. About half of the sequences have been further clustered into superfamilies that have also been assigned permanent identifiers.

The completion of the family classification has permitted an acceleration in the program to reduce database redundancy. With all sequences classified, it is usually necessary to look only among members of the same family for candidates to be merged. Moreover, it also allows the efficient application of methods to propagate and standardize annotation among related sequences. The NBRF has developed and is evaluating semi-automated methods to propagate annotation based on three criteria: (i) sequence alignment (used mainly for feature annotations), (ii) family or superfamily membership and (iii) functional classification.

LINKS TO OTHER DATABASES

An entry in the Protein Sequence Database presents a sequence that may have been constructed from several independent reports. In an entry, each published or submitted sequence appears in a way that allows it to be reconstructed easily and used for searching. To avoid ambiguities when cross-referencing information appearing in or obtained from another sequence database, the reported sequence is directly linked to its identifier or accession number in that database.

Non-sequence data in other databases may be cross-referenced but will not have the direct linkage to particular sequence reports that can be maintained with sequences, their accession numbers and associated identifiers. With this in mind, we exercise care in constructing and maintaining cross-references for non-sequence data. Typically when 'name only' approaches are used to link objects in different databases, the result is over-full linkage tables in which few genuine links are missed but many links associate unrelated information. Subsequently, users of these databases encounter erroneous associations or other unexpected behavior. Whenever possible, cross-references to other databases are established through published literature citations, the reports of standardization committees, or consultation with outside experts.

All cross-references to other databases are regularly checked for concurrency. In particular, sequence cross-references are checked not only for concurrency of the identifier but accuracy of the sequence presentation. Cross-references to identifiers and accession numbers in other databases appear with two parts, a short database tag followed by a colon and then the identifier or accession number. When a user accesses PIR databases through our Web site, these cross-references appear as hypertext links if the corresponding database maintains a search engine that responds to the selected identifier or accession number (Table 1).

Links to MedLine and to the nucleic acid sequence databases are produced during initial data processing, during annotator review and entry merging, and during automatic concurrency checking procedures.

PIR-International uses the GenBank/EMBL/DDBJ databases as a source of primary data. Generally GenBank/EMBL/DDBJ entries are processed as single reports and presented in the Protein Sequence Database with a single citation to the publication, if one is provided. The nucleotide CDS features are re-translated and compared with the sequence in the CDS/translation feature. Discrepancies are noted and resolved, when possible by comparison with the sequence shown in the publication, the corrected form being shown in the 'Residues' record. The cross-references include the PID protein identifier from '/db_xref=PID' qualified features, the NID nucleic acid identifier of the entry, and the GenBank/EMBL/DDBJ primary accession number. Secondary accession numbers may be included, particularly if they are assigned at submission and appear in the publication. In the NCBI Backbone Database, 'gi' identifiers were formerly associated with CDSs and other sequences; these identifiers tagged with 'NCBIP:' and 'NCBIN:' have now been removed as cross-references but do appear as notes.

The PIR uses information in the Protein Data Bank (PDB) (9) predominantly for annotation. PDB entries are processed as submissions to the PDB with cross-references to the PDB code. Sequences of natural origin that appear with coordinate data in the PDB and that are not published elsewhere are included as accessions with cross-references to the PDB code.

DATA FROM GENOME PROJECTS

The complete genomes of organisms are now being determined at an accelerating pace. Table 2 lists complete genomes from which sequences have entered PIR-International within the last 2 years.

Also during this period sequences derived from the 16 *Saccharomyces cerevisiae* chromosomes have entered at various times (17). Between the time this manuscript was submitted (September 1997) and the time when it is published, complete genomes of additional organisms may be published, including *Borrelia burgdorferi* (Lyme disease spirochete), *Bacillus subtilis*, *Methanobacterium thermoautotrophicum* and *Archaeoglobus fulgidus*. In all but one case, the sequences from the complete genome determinations listed above entered the Protein Sequence Database in the interim weekly update soon after, in some cases the same week as, publication. We have also been able to merge about half of the pre-existing sequence entries with the corresponding new sequence determinations within several weeks of their appearance in the database. In some cases we have maintained the pre-existing entries and the new genome sequence determinations unmerged in separate entries where the differences might be of significance to the sequencing community. During the import of information from model organisms, we have been able to improve a number of sequence reports by correctly representing and annotating non-AUG start codons where necessary, translational frameshifting as in peptide chain release factor 2, and translation exceptions such as selenocysteine in formate dehydrogenase alpha chains.

Genome databases, such as the the Human Genome Data Base (GDB) (18), compile genetic information, including standardized gene names and symbols and map positions. The PIR incorporates this information and maintains its concurrency in collaboration with the genome databases. In this ongoing collaboration, a special Web facility has been established to help genome databases rapidly verify PIR accession numbers used as concurrent cross-references to the PIR-International. Other interested genome data centers should contact the NBRF for instructions on using this special facility.

Table 1. Database tags appearing in the current PIR-International databases

PIR data object	External Database data object	Database Tag
Reference Number	MedLine unique identifier	MUID
	Brookhaven Protein Data Bank code	PDB
PIR Accession Number	DDBJ accession number	DDBJ
	EMBL accession number	EMBL
	GenBank accession number	GB
	GenBank/EMBL/DDBJ nucleic acid sequence ID	NID
	Brookhaven Protein Data Bank code	PDB
	GenBank/EMBL/DDBJ protein sequence ID	PID
	The Institute for Genome Research CDS ID	TIGR
	University of Wisconsin Genome Project	UWGP
PIR Genetics Record	Drosophila Database gene identification number	FlyBase
	Human Genome Database gene symbol and accession number	GDB
	Mouse Genome Database gene symbol and accession number	MGI
	On-Line Mendelian Inheritance in Man accession number	OMIM

Table 2. Complete genomes entering the PIR-International in the last 2 years

Organism	PIR entry date	No. entries	Reference
Haemophilus influenzae	18-Aug-1995	1675	10
Mycoplasma genitalium	17-Nov-1995	467	11
Methanococcus jannaschii	13-Sep-1996	1736	12
Mycoplasma pneumoniae (ATCC 29342)	27-Feb-1997	676	13
Synechocystis sp. (PCC 6803)	25-Apr-1997	3136	14
Helicobacter pylori	09-Aug-1997	1559	15
Escherichia coli K-12	12-Sep-1997	4257	16

SUPPLEMENTARY DATABASES

As part of its effort to produce a protein sequence database that is comprehensive, accurate, precise and consistent, the PIR-International produces a number of supplementary sequence or annotation databases.

RESID is the PIR database of modified amino acid residues annotated as features in the Protein Sequence Database (19). Due to the large and steadily increasing number of protein structure modifications that require standardized annotation in the PIR-International Protein Sequence Database, the RESID database was introduced in 1995 to assist users and annotators in interpreting features annotations for covalent binding sites, modified sites and cross-links. It is the only publicly available, human- and computer-readable database comprehensively documenting protein structural modifications. The RESID database describes features annotated in the Protein Sequence Database and is accessible through the ATLAS program. For a feature in the Protein Sequence Database, a corresponding RESID database entry provides a systematic chemical name, frequently observed alternate names, Chemical Abstracts Service registry numbers, atomic formulas and weights, appropriate keywords to accompany the feature, and identifies the amino acids that give rise to the modification. In addition to providing chemical information on modifications in more detail than is possible in the Protein Sequence Database, the RESID Database also provides a means for predicting both chemical-average and monoisotopic molecular weights for modified peptides and fragments from the Protein Sequence Database in order to facilitate their identification by mass-spectroscopy (20,21).

The RESID Database was publicly released in 1995 with 181 entries. Within a year it was recognized as a resource useful in identifying post-translationally modified peptides (22). The Database is distributed quarterly on the ATLAS CD-ROM with the PIR-International Protein Sequence Database. Release 11.00 (September 1997) contained 238 entries.

The NRL_3D Database (23) is produced by PIR from sequence and annotation information extracted from the Brookhaven Protein Data Bank of three-dimensional structures. This database makes the sequence information of PDB available for similarity searches and retrieval, and provides cross-reference information for use with the PIR-International Protein Sequence Database. Release 22.0 of the NRL_3D Database (September 1997) contained 11 496 entries.

PIR-ALN is a database of alignments of protein sequences produced and curated by PIR staff. These alignments contain a limited number of selected sequences in order to allow reasonable interactive display. Superfamily alignments support the superfamily classification for the Protein Sequence Database. They contain a selection of the more distantly related sequences in the superfamily, aligned to illustrate important sequence features and the essential

end-to-end homology of the sequences. These alignments typically contain regions where some sequences have sizable insertions or deletions and other regions where there is so little sequence similarity that the alignment is somewhat arbitrary. Homology domain alignments document the working definition of every type of homology domain that appears in 'Superfamily' records and feature annotations so that they can be reliably located and consistently represented. They contain representative domains selected from sequences in different homeomorphic superfamilies. The PIR-ALN database also includes curated alignments representing families of closely related sequences, typically <55% difference between any pair. The NBRF does not maintain a comprehensive collection of curated family alignments in PIR-ALN, but will continue to provide some that are of particular interest. A complete set of automatically generated family and superfamily alignments is available through the MIPS Web site. Hypertext links to both PIR curated alignments and MIPS automatically generated alignments are provided in the 'Associated information' section of each entry through the PIR Web site. Release 17.0 of the PIR-ALN Database (September 1997) contained >3000 entries.

The PATCHX (24) Database is assembled by MIPS from a collection of other public domain sequence databases and includes protein sequences not identical with or contained within sequences in the PIR-International Protein Sequence Database. To reduce non-informative redundancy, certain sequences are excluded from PATCHX: (i) invalid sequences, such as those present in the PIR4 section, (ii) conflicting sequences processed by the PIR-International, detected through PID cross-references, (iii) sequences with very high homology (a FastA score >95% of the self-score) to a PIR-International Protein Sequence Database entry, and (iv) very short sequences (length <20). When PATCHX is used together with the PIR-International databases, they provide the most comprehensive collection of protein sequence data currently available in the public domain.

THE PIR EDITORIAL BOARD

We have created an Editorial Board to channel the expertise of the scientific community into the organization, integration and annotation of the protein sequence data. Members will review selected protein groups, advise on information to supplement the sequence data, and contribute their own authored material, such as descriptions and alignments. We have developed a system for mutual electronic communication and provide a Web site at http://www-nbrf.georgetown.edu/pir/edbd.html with useful information for the editors. We will be enlisting more members and welcome those who would like to participate to contact the NBRF.

DATABASE ACCESS AND DISTRIBUTION

The fastest and easiest way to access the PIR-International Protein Sequence Database is through the PIR and MIPS Web sites. Starting from the PIR homepage http://www-nbrf.georgetown.edu/pir/ you can locate information on:

- descriptions of PIR products and services
- searching and retrieving from PIR databases
- getting PIR databases by anonymous FTP at nbrf.georgetown.edu and other locations
- PIR database current release information
- procedures for linking to the PIR databases on the Web

The following search and retrieval facilities are provided to access the latest weekly interim update of the sequence and alignment databases. For the PIR-International Protein Sequence Database:

- an entry request service retrieves entries based on entry code, PIR reference, PIR accession or cross-reference identifiers for other databases,
- a text request service provides both basic and advanced search forms to retrieve entries using Boolean operations for substrings in title, species, author, publication, keyword, superfamily, feature and genetics fields,
- a selection list service retrieves entries through alphabetized lists of species, keywords, superfamilies, genes and journal abbreviations, and
- a sequence search service scans for exact sequence matches.

For the NRL_3D Sequence Database:

- an entry request service retrieves entries by PDB code,
- a text request service provides a search form to retrieve entries by substrings in title, species, author, publication, keyword, and feature fields, and
- a sequence search service scans for exact sequence matches.

For the PIR-ALN Alignment Database,

- an entry request service retrieves alignments by either entry code or title search.

Through the MIPS Web sites at http://www.mips.bichem.mpg.de you have access to:

- the latest interim update of PIR-International through the ATLAS interface,
- the PROT-FAM collection of automatically generated multiple sequence alignments, containing more than 10 000 family, 2000 superfamily and 300 homology domain alignments used in the classification of entries in the PIR-International Protein Sequence Database.

It is important to note that other Web sites and data depositories do not always have the latest quarterly release of the PIR-International Protein Sequence Databases that are available. Users should check the NBRF or MIPS Web sites for the weekly interim update or quarterly release information.

ANONYMOUS FTP

Anonymous FTP services are available at ftp://nbrf.georgetown.edu . The host machine recognizes VAX/VMS directory structure. PIR anonymous FTP files are in directory [ANONYMOUS.PIR]. The 000README. file in this directory describes the contents of the directory and the data available from the site. The latest quarterly release is available through FTP.

THE ATLAS OF PROTEIN AND GENOMIC SEQUENCES CD-ROM

The ATLAS program allows simultaneous access to and retrieval from multiple macromolecular or textual databases. Retrieval can be from any selected set of the databases and any combination of fields within those databases including: biological annotations and bibliographic information, such as protein names, superfamily names, homology domains, organism names, gene names, keywords, feature descriptions, authors' names, cross-references to other databases, etc. The ATLAS program also enables selected sets of sequences to be searched directly for exact subsequences or for patterns. The User's Guide for the ATLAS program is included

on the CD-ROM in both PostScript and plain text versions; it can also be obtained separately in printed form (Table 3).

Table 3. Contents of the ATLAS CD-ROM

ATLAS Information Retrieval System
PIR-International Protein Sequence Database
NRL_3D Sequence-Structure Database
PIR-ALN Protein Alignment Database
RESID Database of Residue Modifications
PATCHX additional non-redundant entries from other protein sequence databases
A version of the FASTA (25) package
Complex Carbohydrate Structure Database (CCSD) (26) and CarbBank for Windows95™

The CD-ROM is formatted in accordance with the ISO 9660 standard. The ATLAS program is written in C and currently runs on PC-DOS, VAX/VMS, OpenVMS Alpha AXP, DEC UNIX Alpha AXP, DEC ULTRIX (RISC), SunOS, SGI/IRIX and Macintosh systems.

HOW TO CONTACT PIR-INTERNATIONAL

For information on currently available database releases or other services, contact the appropriate node of PIR-International.

In the Americas, contact PIR: PIR Technical Services Co-ordinator, National Biomedical Research Foundation, 3900 Reservoir Road, NW, Washington, DC 20007, USA. Tel: +1 202 687 2121; Fax: +1 202 687 1662; Email: pirmail@nbrf.georgetown.edu

In Europe, contact MIPS: Munich Information Center for Protein Sequences, Max Planck Institute for Biochemistry, D-82152 Martinsried, Germany. Tel: +49 89 8578 2657; Fax: +49 89 8578 2655; Email: mewes@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. Tel: +81 471 239778; Fax: +81 471 221544; Email tsugita@jipdalph.rb.noda.sut.ac.jp

ACKNOWLEDGEMENTS

The professional staff of the NBRF acknowledge the assistance and support of Ms Kathryn E. Sidman, PIR Technical Services Coordinator, and Ms Desiree Goins, Project Support Specialist. The professional staff are especially appreciative of the years of dedicated service by Ms Sidman. 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. PIR is a registered mark of the NBRF.

REFERENCES

1 George,D.G., Barker,W.C. and Hunt,L.T. (1986) *Nucleic Acids Res.*, **14**, 11–15.
2 Dayhoff,M.O., Eck,R.V., Chang,M.A. and Sochard,M.R. (1965) *Atlas of Protein Sequence and Structure* Vol.1. National Biomedical Research Foundation, Silver Spring, MD.
3 Dayhoff,M.O. (1979) *Atlas of Protein Sequence and Structure* Vol.5, Supplement 3. National Biomedical Research Foundation, Washington, DC.
4 George,D.G., Dodson,R.J.,Garavelli,J.S., Haft,D.H., Hunt,L.T., Marzec,C.R., Orcutt,B.C., Sidman,K.E., Srinivasarao,G.Y., Arminski,L.M., *et al.* (1997) *Nucleic Acids Res.*, **25**, 24–27.
5 Dayhoff,M.O. (1976) *Fed. Proc.*, **35**, 2132–2138.
6 Dayhoff,M.O., McLaughlin,P.J., Barker,W.C. and Hunt,L.T. (1975) *Naturwissenschaften*, **62**, 154–161.
7 Barker,W.C., Pfeiffer,F. and George,D.G. (1995) In Atassi,M.Z. and Appella,E. (eds), *Methods in Protein Structure Analysis*, Plenum Publishing, New York, 473–481.
8 Barker,W.C., Pfeiffer,F. and George,D.G. (1996) *Methods Enzymol.*, **366**, 59–71.
9 Abola,E.E., Manning,N.O., Prilusky,J., Stampf,D.R. and Sussman,J.L. (1996) *Res. Natl. Stand. Technol.*, **101**, 231–241.
10 Fleischmann,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.
11 Fraser,C.M., Gocayne,J.D., White,O., Adams,M.D., Clayton,R.A., Fleischmann,R.D., Bult,C.J., Kerlavage,A.R., Sutton,G., Kelley,J.M. *et al.* (1995) *Science*, **270**, 397–403.
12 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**, 1058–1073.
13 Himmelreich,R., Hilbert,H., Plagens,H., Pirkl,E., Li,B.C. and Herrmann,R. (1996) *Nucleic Acids Res.*, **24**, 4420–4449.
14 Kaneko,T., Sato,S., Kotani,H., Tanaka,A., Asamizu,E., Nakamura,Y., Miyajima,N., Hirosawa,M., Sugiura,M., Sasamoto,S. *et al.* (1996) *DNA Res.*, **3**, 109–136.
15 Tomb,J.F., White,O., Kerlavage,A.R., Clayton,R.A., Sutton,G.G., Fleischmann,R.D., Ketchum,K.A., Klenk,H.P., Gill,S., Dougherty,B.A. *et al.* (1997) *Nature*, **388**, 539–547.
16 Blattner,F.R., Plunkett,G.,III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al.* (1997) *Science*, **277**, 1453–1462.
17 Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M. *et al.* (1996) *Science*, **274**, 546,563–567.
18 Fasman,K.H., Letovsky,S.I., Li,P., Cottingham,R.W. and Kingsbury,D.T. (1997) *Nucleic Acids Res.*, **25**, 72–80. [See also this issue *Nucleic Acids Res.* (1998) **26**, 94–97]..
19 Garavelli,J.S. (1993) *Protein Sci.*, **2** (Suppl. 1), abstract 450.
20 Biemann,K. and Scoble,H.A. (1987) *Science*, **237**, 992–998.
21 Takao,T., Yoshino,K., Suzuki,N. and Shimonishi,Y. (1990) *Biomed. Environ. Mass Spectrom.*, **19**, 705–712.
22 Yates,J.R. (1996) *Methods Enzymol.*, **271**, 351–377.
23 Pattabiraman,N., Namboodiri,K., Lowrey,A. and Gaber,B.P. (1990) *Protein Seq. Data Anal.*, **3**, 387–405.
24 Barker,W.C., George,D.G., Mewes,H.-W., Pfeiffer,F. and Tsugita,A. (1993) *Nucleic Acids Res.*, **21**, 3089–3092.
25 Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 2444–2448.
26 Doubet,S. (1991) *CODATA Bull.*, **23**, 56–58.