The PIR protein sequence database

Winona C.Barker*, David G.George, Lois T.Hunt and John S.Garavelli
National Biomedical Research Foundation, Georgetown University Medical Center, Washington, DC 20007, USA

INTRODUCTION

The Protein Sequence Database was initiated at the National Biomedical Research Foundation (NBRF) in the early 1960's by the late Margaret O. Dayhoff as a collection of sequences for the study of evolutionary relationships between proteins. The database is now an international project of an association of protein sequence data collection centers including NBRF, the Martinsried Institute for Protein Sequences (MIPS), and the International Protein Information Database in Japan (JIPID). The three centers cooperate to produce and distribute a single database of 'wild-type' protein sequences. Currently the NBRF effort is supported as the Protein Identification Resource (PIR) project funded by the National Library of Medicine as part of the Biomedical Research Technology Program. This resource aids researchers in identifying and interpreting protein sequence information. The PIR provides an integrated system composed of protein and nucleic acid sequence databases and software designed for the identification and analysis of protein sequences. Although there are many computer centers serving molecular biologists and other sources of software for sequence analysis, the unique contributions of the PIR are maintenance of a comprehensive, carefully edited, and widely disseminated database of protein sequences and associated information [1], and the development of software specifically tailored to the requirements of these data [2,3]. The resource is operated by scientists with extensive experience in the application and interpretation of sequence comparison methods and provides not only the facilities for protein identification but also practical guidance in the interpretation of the results of these methods.

The Protein Sequence Database is currently divided into three sections that are different in their organization, information content, and degree of verification and standardization. This situation arose because budgetary constraints made it impossible to analyze, evaluate, and incorporate new information into a well-organized data set at the same pace as the data accumulated. The three divisions reflect gradations in the level of verification and organization of the data.

From December 1989 to December 1990, the Protein Sequence Database grew from 14,372 to 26,798 entries, an 86% increase, and from 3,977,903 to 7,620,688 residues, a 92% increase. Much of this increase was due to the introduction of 'Section 3. Unverified entries,' which contains 6,444 entries and 1,816,684 residues. 'Section 2. Preliminary entries' increased by 4,785 entries to a total of 12,607 entries and 3,417,043 residues. 'Section 1. Annotated and Classified Entries' contains 7,747 entries and 2,386,941 residues. There are altogether 30,275 citations to 18,257 sources. Direct submissions account for 1,653 of the citations.

Section 1: Annotated and Classified Entries

The Protein Sequence Database of the PIR has been traditionally maintained as a reference data compendium. Such a compendium includes one 'value' for each item of data (or perhaps a range of values) rather than reproducing the results of individual research reports. Our policy has been to minimize redundant information by combining, into a single entry, data from multiple determinations of the sequence of the same molecule. The resulting highly verified, nonredundant data collection that forms Section 1 of the database supports statistical analyses that require a representative collection, is efficient to search, and saves users the considerable time needed to verify, compare, and combine related data from individual reports.

The text portion of fully annotated entries of the Protein Sequence Database provides information useful for evaluating the results of database searches. In addition to literature citations and information about the experimental determination of the sequence, the text may include alternative nomenclature, domain structure, functional sites in the sequence, genetic information, and general information about the source, function, tertiary and quaternary structures, and physicochemical properties of the molecule. This information may be used to locate particular types of proteins in the database or to compare other features when two proteins are found to have similar sequences.

The origin of the database as a set for the study of protein evolution is reflected in the organization of the data, which is based on the concept of a protein superfamily: a group of proteins whose amino acid sequences can be shown to be evolutionarily related [4]. Inclusion of a protein in a specific superfamily implies that the protein is homologous (evolved from a common ancestral sequence) with the other proteins in that superfamily. Each sequence in Section 1 of the Protein Sequence Database is assigned a set of five numbers, the first of which represents the superfamily. The family, subfamily and entry numbers subdivide a superfamily into groups of proteins that are more that 50%, 20% and 5% different, respectively. The superfamily organization is useful in interpreting database searches because a new sequence that is truly homologous with one already in the database should show sequence similarity to the corresponding

* To whom correspondence should be addressed

regions of other members of the superfamily as well. Furthermore, it is possible to choose a representative subset of the database for searching by removing all sequences that are, for example, less that 20% different from another sequence. This superfamily classification was designed before it became apparent that many large proteins are composed of domains and that these distinct regions may have different evolutionary origins and histories. Not only have genes duplicated to give rise to new proteins, but they have fused, gained and lost exons, changed reading frames, and incorporated DNA from other parts of the genome, from other genomes within the organism, or even from other organisms. Sequences that are mostly unrelated are assigned to separate superfamilies even when they contain related domains.

Figure 1 shows a data entry from Section 1 as it appears on the ASCII card-image tape distributed by PIR. This format, which adheres closely to the CODATA standardized format

```
\\\
ENTRY            GDPG          #Type Protein
TITLE            Glutaredoxin - Pig

ALTERNATE-NAME   thioltransferase
DATE             31-Dec-1990 #Sequence 31-Dec-1990 #Text 31-Dec-1990
PLACEMENT          23.0    0.2    1.0    1.0    1.0
SOURCE           Sus scrofa domestica #Common-name domestic pig
ACCESSION        JQ0117\ A29322

REFERENCE        (Liver)
   #Authors      Yang Y., Gan Z.R., Wells W.W.
   #Journal      Gene (1989) 83:339-346
   #Title        Cloning and sequencing the cDNA encoding pig liver
                   thioltransferase.
   #Reference-number JQ0117
   #Accession    JQ0117
   #Molecule-type mRNA
   #Residues     1-106 <YAN>

REFERENCE        (Liver)
   #Authors      Gan Z.R., Wells W.W.
   #Journal      J. Biol. Chem. (1987) 262:6699-6703
   #Title        The primary structure of pig liver thioltransferase.
   #Reference-number A29322
   #Accession    A29322
   #Molecule-type protein
   #Residues     3-4,'A',5-106 <GAN>

COMMENT          This enzyme catalyzes the reduction of a variety of
                   disulfides, including protein disulfides, in the
                   presence of reduced glutathione.
SUPERFAMILY      #Name glutaredoxin
KEYWORDS         electron transport\ deoxyribonucleotide synthesis\
                   redox-active disulfide

FEATURE
   2-106                    #Protein glutaredoxin <MAT>\
   2                        #Modified-site acetylation\
   23-26                    #Disulfide-bonds redox-active
                              (predicted)\
   79-83                    #Disulfide-bonds

SUMMARY        #Molecular-weight 11828  #Length 106  #Checksum  7594
SEQUENCE
                  5         10        15        20        25        30
      1 M A Q A F V N S K I Q P G K V V V F I K P T C P F C R K T Q
     31 E L L S Q L P F K E G L L E F V D I T A T S D T N E I Q D Y
     61 L Q Q L T G A R T V P R V F I G K E C I G G C T D L E S M H
     91 K R G E L L T R L Q Q I G A L K
///
\\\
```

Fig. 1. An entry from the PIR Protein Sequence Database (CODATA format).

recommended for the communication of protein sequence data [5], has several advantages: (1) All data items are labeled with specific identifiers or subidentifiers, so it is easy to write computer programs to access and manipulate the information. (2) It is also 'human' readable. (3) It is extensible by the addition of new identifiers and subidentifiers. (Subidentifiers are preceded by the # symbol.) (4) It explicitly includes some information previously available only in auxiliary files. (5) It is readily convertible to other formats for use by programs and retrieval systems based on these formats; for example, the display in Fig. 1 and that shown in Fig. 2 are automatically generated from the same internal format. In Fig. 2, some of the identifiers are replaced by synonymous labels and some data items are not explicitly labeled.

Section 2: Preliminary Entries

In 1985, the Protein Sequence Database added a file of sequence entries under preparation for the main database. These entries have been compared with other sequences in the database and with the translations of the corresponding nucleotide sequences

```
GDPG
Glutaredoxin - Pig

Alternate names: thioltransferase

Species: Sus scrofa domestica (domestic pig)

Accession: JQ0117; A29322

Yang, Y., Gan, Z.R., and Wells, W.W., Gene 83, 339-346, 1989
    (Liver)
    Title: Cloning and sequencing the cDNA encoding pig liver
    thioltransferase.
    Reference number: JQ0117
    Accession: JQ0117
    Molecule type: mRNA
    Residues: 1-106 <YAN>

Gan, Z.R., and Wells, W.W., J. Biol. Chem. 262, 6699-6703,
    1987 (Liver)
    Title: The primary structure of pig liver
    thioltransferase.
    Reference number: A29322
    Accession: A29322
    Molecule type: protein
    Residues: 3-4,'A',5-106 <GAN>

This enzyme catalyzes the reduction of a variety of
    disulfides, including protein disulfides, in the presence
    of reduced glutathione.

Superfamily: glutaredoxin

Keywords: electron transport; deoxyribonucleotide synthesis;
    redox-active disulfide

Residues            Feature
2-106               Protein: glutaredoxin <MAT>
2                   Modified site: acetylation
23-26               Disulfide bonds: redox-active (predicted)
79-83               Disulfide bonds:

Mol. wt. unmod. chain = 11,828      Number of residues = 106

          5         10        15        20        25        30
 1 M A Q A F V N S K I Q P G K V V V F I K P T C P F C R K T Q
31 E L L S Q L P F K E G L L E F V D I T A T S D T N E I Q D Y
61 L Q Q L T G A R T V P R V F I G K E C I G G C T D L E S M H
91 K R G E L L T R L Q Q I G A L K
```

Fig. 2. Alternative display of the entry in Fig. 1, using the PSQ program.

(when these are available) and have been examined by computer to detect certain well-defined errors and inconsistencies; they are often partially annotated, but not all of the additional information has been subjected to critical staff review. The scientific name of the organism is supplied from a standardized list to allow for sorting entries according to taxonomic classification. The entries for a given species are sorted alphabetically according to protein name.

Section 3. Unverified Entries

In order to make data available more quickly, we have restricted the information initially entered to that which is minimally necessary to make the sequence data useful. An automated, transaction protocol-based system is being developed that will allow this dataset to be remotely updated by E-mail from each of the collaborating centers. Each entry (see Fig. 3) in this dataset corresponds to a single sequence (as published in a single paper or as submitted to a database) and contains the following information only: an entry identification code, an entry title, an optional species line, the reference citation (which optionally includes the title of the publication), a reference number, an accession number (which is identical with the entry identification code), an optional cross-reference line, and the sequence. These data are marked as unverified because they have not been reviewed by scientific staff; when the information is reviewed and further processed, it is moved to Section 1 or 2 of the database. Although initial data entry is very accurate (perhaps 2 or 3 uncorrected errors per thousand characters), about 5% of all sequence reports present conflicting data for the same sequence, e.g., the amino acid translation is inconsistent with the corresponding nucleotide sequence. Moreover, some unverified sequence entries are grossly incorrect, mislabeled, or incorrectly attributed due to misinterpretations occurring in the initial data selection process. While we strongly recommend that researchers always consult the original literature from which database entries are constructed, this is especially important when using data from entries that have not been thoroughly evaluated by the database staff.

Transaction protocols also provide an interface to the database for data submitted by authors. The DEPOSIT command of the PIR network server includes filters that transform the submitted data into a transaction that can be incorporated into Section 3 and made available soon after submission. It will be possible to develop similar filters for the direct incorporation of information processed at other database centers such as GenBank or the proposed 'backbone' database of the National Center for Biotechnology Information.

Modification of Accession Number Policy

The two primary keys to entries in Section 3 are the 'accession number' and the 'reference number.' The accession number is a unique identifier that identifies a single sequence as represented in a single report. The reference number is a unique identifier that identifies a citation to the scientific literature or to a submitted set of sequence data. These numbers are also used as the primary tracking keys once the data have been entered into the computer. They provide a mechanism for preventing redundancy across the three sections of the database.

In Sections 1 and 2 of the database, which contain merged entries, the accession numbers are directly associated with a

```
\\\
ENTRY              S09010        #Type Protein
TITLE              *Alzheimer's disease amyloid beta protein - Human

DATE               06-Sep-1990 #Sequence 06-Sep-1990 #Text 06-Sep-1990
PLACEMENT             0.0     0.0     0.0     0.0     0.0
COMMENT            *This entry is not verified.
SOURCE             Homo sapiens #Common-name man

REFERENCE
   #Authors        Kitaguchi N., Takahashi Y., Oishi K., Shiojiri S.,
                     Tokushima Y., Utsunomiya T., Ito H.
   #Journal        Biochim. Biophys. Acta (1990) 1038:105-113
   #Title          Enzyme specificity of proteinase inhibitor region in
                     amyloid precursor protein of Alzheimer's disease:
                     different properties compared with protease nexin
                     I.
   #Reference-number S09010
   #Accession      S09010

SUMMARY            #Molecular-weight 8114  #Length 73  #Checksum  3785
SEQUENCE
                   5         10        15        20        25        30
      1 E V C S E Q A E T G P C R A M I S R W Y F D V T E G K C A P
     31 F F Y G G C G G N R N N F D T E E Y C M A V C G S A M S Q S
     61 L L K T T Q E P L A R R L
///
\\\
```

Fig. 3. An entry from Section 3 of the Protein Sequence Database.

'residues' line. This line contains an unambiguous syntax that provides the necessary information for regenerating the sequence as originally reported in the publication. Hence both the merged sequence and all of its constituent parts are represented in a single entry rather than being redundantly stored separately.

These developments signify a modification of the accession number policy of the past. Originally an accession number was a permanent label attached to a sequence entry in the database to allow it to be located in future releases of the database, regardless of changes to the entry. Unfortunately, this policy had a fundamental flaw: it is not possible to assign a permanent identity to something that is not permanent. At some stage the entry may be changed to such an extent that it retains little or no resemblance to the original. This has resulted also in the undesirable accumulation of accession numbers as the information in the entries has been modified, combined, and separated.

Our new policy is a refinement of, rather than a departure from, the previous policy. Although the accession numbers are now associated with individual reported sequences rather than with the merged sequence entry, they are still contained within the entry and can be used to locate the information in future releases of the database.

Magnetic Tapes and CD-ROM

The databases and programs of the PIR 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 updated and distributed quarterly. In 1991, 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. Please contact PIR, MIPS, or JIPID as instructed below to obtain further information about CD-ROM releases.

VAX/VMS format magnetic tapes of the PIR Protein Sequence Database include the Protein Sequence Query (PSQ) database retrieval program. With this software, entries can be retrieved by sequence, by species, by author, by citation, by superfamily classification, by taxonomic classification, and by keywords and features. Sequences can be back-translated and possible restriction enzyme cut points located. Amino acid composition tables can be compiled for single sequences or accumulated for a list of sequences. Sequences can be searched for particular strings of residues with or without mismatches. Entries can be copied from the database to a user-owned file for modification and these user-created files can be accessed by some commands of the PSQ program. Also included on this tape are programs for creating databases that can be used by PSQ from user-supplied files of entries. The tape also contains the NRL-3D database of sequence information extracted from the Brookhaven Protein Data Bank and formatted for use with the PIR sequence analysis software and PSQ program. The output from the Match command of 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, species, superfamily names, citations, keywords, and features. Tapes in both formats contain documentation and additional files containing the species names (ordered in a taxonomic hierarchy), journal abbreviations, and special genetic codes used in the database.

On-line Access

The databases and programs of the PIR are accessible via direct dial-up. There is an annual fee for qualified researchers at nonprofit institutions. Others pay per hour connected. In addition to its own protein and nucleic acid sequence databases, the PIR makes recent releases of several other databases available to its on-line users. Among these are the EMBL Nucleotide Sequence Data Library, the GenBank Genetic Sequence Databank, and a merged protein database (MIPSX). The sequence databases are accessed on our computer 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 our system, thereby eliminating the need to repeat the same query in each database separately. The on-line system also includes many other programs for sequence searching, comparison, and analysis.

The PIR E-mail Server

The Protein Sequence Database is accessible by electronic mail query to the PIR network fileserver and database query system. Complete instructions 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 on BITNET. If you are using a network other than BITNET or INTERNET, your return address may not work. In this case, contact Dr. John S. Garavelli at POSTMASTER@GUNBRF before attempting to use the server.

The PIR E-mail server handles database queries, sequence searches, and sequence submissions. To retrieve a particular entry, the user must know its entry code. The ACCESSION, AUTHOR, JOURNAL, KEYWORD, SPECIES, and TITLE commands allow users to search text fields of the databases to obtain codes for relevant entries. These commands work not only on the PIR databases but also on the latest releases of the GenBank and EMBL databases. The SEARCH command enables the location of PIR entries whose sequences are similar to a protein sequence provided by the user, or to the protein sequences translated from six reading frames, according to any one of several different genetic codes, from a nucleotide sequence provided by the user. Sequences are compared with all sections of the Protein Sequence Database using the FASTA algorithm. Currently, the data are updated weekly. In the near future, we expect to provide access to daily updates via the E-mail server.

How to Obtain PIR 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 Institut fr Proteinsequenzen, Max-Planck-Institut fr Biochemie, D-8033 Martinsried bei Mnchen, FRG; telephone +49 89 8578 2656; FAX +49 89 8578 2655; electronic mail MEWES@MIPS.BITNET. 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 241501; FAX +81 471 221544; electronic mail TSUGITA@JPNSUT31.BITNET or, on DIALCOM, 42:CDT0079. The NBRF PIR 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 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.

ACKNOWLEDGEMENT

The Protein Identification Resource is supported by National Institutes of Health Grant LM05206.

REFERENCES

1. Barker, W.C., George, D.G., and Hunt, L.T. (1990) Meth. Enzymol. 183, 31–49.
2. Orcutt, B.C., George, D.G., and Dayhoff, M.O. (1983) Annu. Rev. Biophys. Bioeng. 12, 419–443.
3. George, D.G., Hunt, L.T., and Barker, W.C. (1988) In Lesk, A.M. (ed.), Computational Molecular Biology: Sources and Methods for Sequence Analysis. Oxford University Press, Oxford, pp. 100–115.
4. Dayhoff, M.O., McLaughlin, P.J., Barker, W.C., and Hunt, L.T. (1975) Naturwissenschaften 62, 154–161.
5. George, D.G., Mewes, H.W., and Kihara, H. (1987) Protein Seq. Data Anal. 1, 27–39.