ARGONNE NATIONAL LABORATORY
9700 South Cass Avenue
Argonne, IL 60439

ANL/MCS-TM-200

Overview of Selected Molecular Biological Databases

by

Karen D. Rayl and Terry Gaasterland*

Mathematics and Computer Science Division

Technical Memorandum No. 200

November 1994

This work was supported by the Office of Scientific Computing, U.S. Department of Energy, under Contract W-31-109-Eng-39.

*Participant in the Fall 1993 Science and Engineering Research Semester. This program is coordinated by the Division of Educational Programs. Home institution is Austin College, Sherman, TX 75091.

Contents

Overview of Selected Molecular Biological Databases

by

Karen D. Rayl and Terry Gaasterland

Abstract

This paper presents an overview of the purpose, content, and design of a subset of the currently available biological databases, with an emphasis on protein databases. Databases included in this summary are 3D_ALI, Berlin RNA databank, Blocks, DSSP, EMBL Nucleotide Database, EMP, ENZYME, FSSP, GDB, GenBank, HSSP, LiMB, PDB, PIR, PKCDD, ProSite, and SWISS-PROT. The goal is to provide a starting point for researchers who wish to take advantage of the myriad available databases. Rather than providing a complete explanation of each database, we present its content and form by explaining the details of typical entries. Pointers to more complete "user guides" are included, along with general information on where to search for a new database.

Presentation

In a variety of institutions, efforts are under way to capture and maintain large quantities of molecular biological data in electronic repositories. The data repositories are numerous and vary in form and content. The data that is becoming available covers a continuum of biological knowledge on DNA, RNA, protein molecule structure, and enzyme function within metabolism, to evolutionary phylogenetic relationships of organisms, and many other areas.

To aid people who want to know more about the available databases, we selected sample entries from available data repositories and present an explanation of what is contained in each entry. The explanations summarize and distill information obtained from "user guides," "help files," and "tutorials" that are associated with each data repository. Information about how to access each repository appears at the end of each databases section. The selected data repositories represent our bias toward databases that contain information relevant to protein sequence analysis rather than a complete summary of all the data repositories available in the world.

Information on the following databases[1] is contained in this paper. It is arguable whether the databases should be presented by groups according to topic (e.g., DNA sequence databases or secondary structure databases) or according to their use (e.g., for protein sequence analysis or for RNA studies). However, some databases cover multiple topics. Furthermore, the uses that we perceive

[1] Although the term "databases" is often used to refer to these collections of data, the term "data repository" is actually more accurate: many protein "databases" are really sets of formatted files. The creators and users of these data repositories, however, use the terms database and databank; we follow the standard terminology.

for a database may not be the only ones. Thus, we chose to present the databases in alphabetical order.

1. 3D_ALI Databank (3D_ALIgn)
2. Berlin RNA Databank
3. Blocks Database
4. DSSP (Dictionary of Secondary Structure of Proteins)
5. EMBL Nucleotide Sequence Database
6. EMP (Enzymes and Metabolic Pathways)
7. ENZYME Database
8. FSSP (Families of Structurally Similar Proteins)
9. GDB (Genome Database)
10. GenBank (Genetic Sequence Databank)
11. HSSP (Homology-derived Secondary Structures of Proteins)
12. LiMB (Listing of Molecular Biological Databases)
13. PDB (Protein Databank)
14. PIR (Protein Information Resource)
15. PKCDD (Protein Kinase Catalytic Domain Database)
16. ProSite (Dictionary of Protein Sites and Patterns)
17. SWISS-PROT (Swiss Protein Database)

Each section includes information on the following topics.

- General background: the topic, purpose, intent, and origin of the database.
- A sample entry together with an explanation of each field
- Information on how to obtain the database.

References, and how information to obtain available on-line references, are summarized at the end of each section as well as at the end of this report.

Please be warned: The only way to really understand what information is available in a particular database is to study many of its actual records. These data repositories are evolving creatures in their format,[2] content, and value. Tutorials, user guides, and help files are often dated from a time when the contents or scope of the data differed; hence, they should be used as guides, not authorities. Thus, this overview claims only to be a tool, not an authority. Likewise, neither FTP sites nor e-mail addresses are permanent locations. The information presented here was collected from June 1993 to January 1994.

1 3D_ALI Databank (3D_ALIgn)

3D_ALI, created by Stefano Pascarella and Patrick Argos, is a databank that merges protein structural and sequence information and is highly dependent on databases that will be introduced later. Structural superpositions of proteins with similar main chain conformation were performed to supplement those collected from the literature. Superpositions of sets of proteins were extracted from pairwise superpositions. For each family of structurally aligned proteins, homologous SWISS-PROT sequences were added if they met the following conditions:

1. the sequences were at least 50%[3] homologous in residue identity to one of the structural sequences, and
2. at least 50% of the structural sequence residues were alignable.

These alignments were taken from HSSP, a program that aligns SWISS-PROT sequences with all proteins of known structure. When a sequence could be aligned in under 3D_ALI rules to more than one structure in a superposition group, the best alignment was used. The structural redundancies of HSSP, maintained from the PDB,[4] were eliminated from 3D_ALI. These restrictions allow for confidence that primary sequences share similar tertiary conformation, except perhaps in loop regions, which are less conserved.

A total of 83 topological groups was created, of which 38 groups contained two or more PDB structures while 45 groups contain only one tertiary structure. A total of 2290 SWISS-PROT sequences are aligned to these groups. The sample entry included here is based on an entry in the "two or more PDB structures" category. Files based on only one PDB structure have the same format.

[2] Many of these databases have formats resembling Fortran code, or punch card code. This provides a simple format that can be accessed in many ways and is more "human-readable" than some possible formats. The Fortran language is more commonly known than other computer languages among biologists and has become a self-propagating format for many new data repositories. However, discussions about a better format are leading to an agreement to start using a new format in the near future. EMBL has begun a massive effort to provide access to and a common format for existing databases. The EMBL format now, to large extent, dictates the format for new databases; therefore, one can readily notice similarities in the format of diverse databases associated with EMBL.

[3] Or, alternatively, 35%. Separate files were created for 50% homology and 35% homology requirements.

[4] Many records within PDB are almost identical and thus represent "no new" information. See information later in this paper on PDB for why these redundancies exist.

1.1 Explanation of a 3D_ALI Record

The first section of a 3D_ALI file is simply a header.[5] A title bar line (1)[6] is followed by the FT line (2) which gives the file name and by the DT line (3) which gives the date of file creation. The NS line (4) indicates the number of PDB structures included in the file, with the FN lines (5) giving relevant information from the PDB file. The NA line (6) tells the number of SWISS-PROT sequences included in the file, and the NL line (7) gives the length, in lines, of the structural alignment. The version of SWISS-PROT used is given in the DB line (8). Information regarding the homology threshold used is stated in the TH lines (9). The database holds entries from both the 50/50 and the 35/35 thresholds, with the default being the 50/50 threshold. Records using the alternative 35% threshold can be distinguished in the TH lines, as well as by their assigned names. Author information is given by AU lines (10) NT lines (11), containing notational information, are of great assistance to understanding a 3D_ALI record. The NT lines are found in every record, so the notation information is always readily available.

The next section of an entry introduces the PDB structures and those SWISS-PROT sequences aligned to them. Line (13) provides the header for the section and uses abbreviations similar to HSSP files. As explained in the NT lines, the notations have the following meanings.

- seq_id gives the sequence identifier from the PDB[7] or SWISS-PROT as is appropriate
- %ide tells the percentage of identical residues
- ifir indicates the first residue of the PDB structure in the alignment
- ilast indicates the last residue of the PDB structure in the alignment
- jfir indicates the first residue of the SWISS-PROT sequence in the alignment
- jlast indicates the last residue of the SWISS-PROT sequence in the alignment
- lali specifies the total length of the alignment excluding insertions or deletions
- ngap is number of gaps in the SWISS-PROT sequence relative to the PDB structural sequence
- lgap is the total length of all gaps counted in ngap
- lseq is the length of the SWISS-PROT sequence

Lines (14) and (15) present PDB entries. Notice that there were no SWISS-PROT entries that aligned to 1fcb in line (14). The sequences that aligned to 3b5c, (15), follow beginning at line (16). These lines represent SWISS-PROT entries. They contain the information embedded in notations and end with the SWISS-PROT accession number and entry name.

[5] Please note that lines in many of the sample entries for various databases were clipped on the right to fit on the page.

[6] See line 1 of the sample entry.

[7] PDB records are inset with "->" markers.

The third section of the 3D_ALI record is separated by the "####" of line (17) and presents the structural alignments of each SWISS-PROT sequence with the PDB protein to which it is homologous as stated in line (18). The number of blocks is given in the NB line (19), followed by the number of structures per block in the NC line (20). When more than three PDB records are contained in a file, this section is presented in successive blocks of up to three structures each. Line (21) serves as a header telling which PDB record each section represents. The column abbreviations, as explained in the NT files, for line (22) have the following meanings.

- HSP gives the residue numbering in the HSSP file
- PDB represents the residue numbering in the PDB file
- A specifies the amino acid sequence in the standard one-letter code
- STRUCT delineates the secondary structure designated in the DSSP file
- PHI and PSI define the main-chain dihedral angles
- ACC describes the solvent accessible surface area taken from the DSSP file

Line (23) begins the information for this section with the first amino acid of the SWISS-PROT sequence, as aligned in the HSSP file.

1.2 Sample 3D_ALI Entry

```
 1 ⟹ ***************** FILE 3D_ALI - sequence alignment based on 3D superposition ****************
 2 ⟹ FT   file: cytb.3D_ALI
 3 ⟹ DT    created at EMBL Thu Jun  4 19:09:30 1992
 4 ⟹ NS    2
 5 ⟹ FN    1fcb  chain: A; Range(PDB):    1 -   99         resolution: 2.4 angstroms.
      FN    3b5c  chain: -; Range(PDB):    3 -   87         resolution: 1.5 angstroms.
 6 ⟹ NA    9
 7 ⟹ NL    107
 8 ⟹ DB    RELEASE 21.0 OF EMBL/SWISS-PROT WITH  23742 SEQUENCES
 9 ⟹ TH    threshold_1 - % identity       0.50
      TH    threshold_2 - % alignment      0.50
10 ⟹ AU    This file was generated by the program CHARON
      AU    Stefano Pascarella and Patrick Argos, EMBL, Postfach 10 22 09, 6900 Heidelberg, Germany
      AU
11 ⟹ NT    The sequence alignments and part of the notation included in this file has been taken from the HSSP
      NT    Sander C. and Schneider R. : Database of homology-derived protein structures. Protins, in press.
      NT    Notation:    seq_id = sequence identifier either from PDB or from SWISSPROT.
      NT                 %ide = percentage of identical residues.
      NT                 ifir, ilast = first and last residue of the alignment in the test sequence.
      NT jfir, jlast = first and last residue of the alignment in the aligned protein.
      NT                 lali = length of the alignment excluding insertions and deletions.
      NT                 ngap, lgap = number and total length of all insertions and deletions.
      NT                 lseq2 = length of the entire sequence of the aligned protein.
      NT                 HSP = numeration in the HSSP file.
      NT                 PDB = numeration in the PDB file.
      NT                 A = aminoacid sequence.
      NT                 STRUCT = secondary structure assignment as found in DSSP file.
      NT                 PHI, PSI, CHIx = dihedral angles phi, psi and chi1, chi2....
      NT                 ACC = accessibility as found in DSSPfiles.
      NT     Symbols:    '-'  indicates a gap in the structural alignments.
      NT                 '?'  indicates a structural gap: the sequence is known but the structure is unresolved.
      NT                 ' ' indictes a gap in the sequence alignments.
      NT                 pairs of lower case letters in the alignment bracket an insertion in the sequence.
12 ⟹ EN

          ...................................................................................................
13 ⟹            seq_id          %ide  ifir  ilas  jfir  jlas  lali  ngap  lgap  lseq
14 ⟹ ->    1. 1fcb
15 ⟹ ->    2. 3b5c
16 ⟹    3. CYB5$PIG             0.99     1    85     7    91    85     0     0   133 P00172     CYTOCHROME B5.
         4. CYB5$RABIT           0.95     2    85     8    91    84     0     0   133 P00169     CYTOCHROME B5.
         5. CYB5$HORSE           0.94     1    85     7    91    85     0     0   133 P00170     CYTOCHROME B5.
         6. CYB5$HUMAN           0.94     1    85     7    91    85     0     0   133 P00167     CYTOCHROME B5.
         7. CYB5$ALOSE           0.94     1    84     4    87    84     0     0    87 P00168     CYTOCHROME B5
         8. CYB5$RAT             0.94     2    85     8    91    84     0     0   133 P00173     CYTOCHROME B5.
         9. CYB5$CHICK           0.82     3    85    15    97    83     0     0   138 P00174     CYTOCHROME B5.
        10. NIA$NEUCR            0.77     1    44     1    44    44     0     0    46 P08619     NITRATE REDUCT
        11. CYM5$RAT             0.64     1    81     8    88    81     0     0    92 P04166     CYTOCHROME B5,

          ...................................................................................................
```

```
17 ⟹  ####
18 ⟹  CC    Structure superposition
19 ⟹  NB    1  blocks
20 ⟹  NC    3 structures per block
21 ⟹                                    1fcb                                 3b5c

22 ⟹      HSP   PDB A  STRUCT     PHI    PSI   ACC |HSP   PDB A  STRUCT     PHI    PSI   ACC |HSP   PDB A  STRU
         |-----------------------------------------|-----------------------------------------|-----------------
23 ⟹    1.   1    1 E            360.0  -20.2   15 |   0    0   --------- 000000 000000 0000 |
        2.   2    2 P       -    -60.0  -81.1   33 |   0    0   --------- 000000 000000 0000 |
        3.   3    3 K       -    170.3  -82.3  100 |   0    0   --------- 000000 000000 0000 |
        4.   4    4 L       -     92.0  -25.4  110 |   0    0   --------- 000000 000000 0000 |
        5.   5    5 D       -     43.0  -31.8   79 |   0    0   --------- 000000 000000 0000 |
        6.   6    6 M       -     73.3  -92.7   85 |   0    0   --------- 000000 000000 0000 |
        7.   7    7 N S    S+    169.8  -23.0  136 |   1    3 A            360.0  139.2  148 |
        8.   8    8 K       +   -142.8  -72.7  136 |   2    4 V       -     -69.0  147.6   83 |
        9.   9    9 Q       -     32.2 -113.7  105 |   3    5 K       -   -118.4  126.7  105 |
       10.  10   10 K       -    124.7   72.1  119 |   4    6 Y E     -a  -111.1  137.4  127 |
       11.  11   11 I B     -a   -83.9  169.2    7 |   5    7 Y E     -a  -119.7  142.4   43 |
       12.  12   12 S    >  -   -136.1  144.4   33 |   6    8 T    >  -     -85.0  158.1   51 |
       13.  13   13 P T  4 S+    -50.9   -7.5   30 |   7    9 L H  > S+     -61.0  -36.2   23 |
       14.  14   14 A T  > S+   -116.6  -29.0   72 |   8   10 E H  4 S+     -53.2  -47.0  143 |
       15.  15   15 E T  4 S+    -61.8  -44.9   78 |   9   11 E H >4 S+     -69.8  -45.1   63 |
```

[lines 16-107 deleted]

```
         |-----------------------------------------|-----------------------------------------|-----------------
      ####


24 ⟹  $$$$
25 ⟹  CC   Sequence alignments  -  from 1
26 ⟹  NB   1 blocks
27 ⟹  NC   120 sequences per block

           .........|.........|.........|.........|.........|.........|.........|.........|.........|.........|
        1. E
        2. P
        3. K
        4. L
        5. D
        6. M
28 ⟹    7. NAA AAA  AA
        8. KVVVVVVV EV
        9. QKKKKKKKRMT
       10. KYYYYYYYYDY
       11. IYYYYYYYYLY
       12. STTTTTTTRER
       13. PLLLLLLLLYL
       14. AEEEEEEEEEE
       15. EEEEEEEEEIE
```

[lines 16-107 deleted]

```
           .........|.........|.........|.........|.........|.........|.........|.........|.........|.........|
29 ⟹  %%%%
30 ⟹  CC    Chi angles
31 ⟹  NB     1  blocks
32 ⟹  NC     3  structures per block
33 ⟹                                   1fcb                                    3b5c

34 ⟹      HSP   PDB A    CHI1   CHI2   CHI3   CHI4 |HSP   PDB A    CHI1   CHI2   CHI3   CHI4 |HSP   PDB A    CH
          |----------------------------------------|----------------------------------------|-----------------
35 ⟹    1.   1     1 E   16.91-145.41 -17.97-999.00 |   0    0   -999.99-999.99-999.99-999.99 |
         2.   2     2 P -999.00-999.00-999.00-999.00 |   0    0   -999.99-999.99-999.99-999.99 |
         3.   3     3 K   24.43-140.36  52.43  90.22 |   0    0   -999.99-999.99-999.99-999.99 |
         4.   4     4 L -104.42 -96.28-999.00-999.00 |   0    0   -999.99-999.99-999.99-999.99 |
         5.   5     5 D  -95.49  37.65-999.00-999.00 |   0    0   -999.99-999.99-999.99-999.99 |
         6.   6     6 M  137.41-179.44 -24.92-999.00 |   0    0   -999.99-999.99-999.99-999.99 |
         7.   7     7 N  176.61  69.92-999.00-999.00 |   1    3 A -999.00-999.00-999.00-999.00 |
         8.   8     8 K   22.63  94.40 136.46 157.84 |   2    4 V -179.29-999.00-999.00-999.00 |
         9.   9     9 Q -168.27  64.90   5.56-999.00 |   3    5 K  -69.68 150.84  94.06 -72.79 |
        10.  10    10 K   80.67  60.95 103.84-133.51 |   4    6 Y  -70.79 -71.48-999.00-999.00 |
        11.  11    11 I  -10.25 -91.17-999.00-999.00 |   5    7 Y  -62.03  83.16-999.00-999.00 |
        12.  12    12 S  -51.74-999.00-999.00-999.00 |   6    8 T   63.77-999.00-999.00-999.00 |
        13.  13    13 P -999.00-999.00-999.00-999.00 |   7    9 L -178.06  59.67-999.00-999.00 |
        14.  14    14 A -999.00-999.00-999.00-999.00 |   8   10 E  174.58 178.59   7.44-999.00 |
        15.  15    15 E  171.35  40.85-142.17-999.00 |   9   11 E -179.75  64.95  29.91-999.00 |

                                    [lines 16-107 deleted]

          |----------------------------------------|----------------------------------------|-----------------
36 ⟹  END.
```

Section four of an entry is marked by “$$$$” as shown in (24). This section presents the sequence alignments, as the CC line (25) introduces. In the sample entry there is one block, shown in the NB line (26), with 120 sequences, shown in the NC line (27). Successive columns of the section (28) contain sequences written from left to right according to increasing numerical assignment of lines analogous to (16). Row numbers are an internal system in 3D_ALI for referencing the structural superpositions. As stated in the NT lines, in the sequence columns

- “-” indicates a gap in the structural alignment,
- “?” designates a structural gap where the sequence is known but the structures are unresolved, and
- “ ” presents a gap in the sequence alignments.

The final section of a 3D_ALI record, marked by “%%%%” (29), lists the torsion angles or chi angles (30). Again, the number of blocks is specified by NB (31), the number of structures per block by NC (32), and the PDB record in each section by line (33). Line (34) uses the following notational abbreviations as column specifiers, as explain in the NT lines.

- HSP gives the enumeration from the HSSP file
- PDB gives the enumeration from the PDB file
- A gives the residues in the one letter amino acid code
- CHI1, CHI2, CHI3, and CHI4 give the respective dihedral angles

Line (35) represents the first line of information for the sequence alignment section. The record is ended by “END” as shown in (36).

1.3 Obtaining 3D_ALI

3D_ALI is available from EMBL by anonymous FTP:

ftp ftp.embl-heidelberg.de [or ftp 192.54.41.33]
cd /pub/databases/3d_ali

Information can also be obtained by e-mail:

To: NETSERV@EMBL-Heidelberg.DE
help 3d_ali (as message body)

Questions can be sent

To: NET-HELP@EMBL-Heidelberg.DE
To: ARGOS@EMBL-HEIDELBERG.DE
To: PASCARELLA@VAXRMA.INFN.IT

The primary reference for 3D_ALI is

> Pascarella, S., and P. Argos. 1992. A data bank merging related protein structures and sequences. *Protein Engineering*. 5: 121–137.

1.4 References

Document: 3D_ALI.DOC. Obtained from e-mail to *NETSERV@EMBL-Heidelberg.DE* with message *GET 3D_ALI:3D_ALI.DOC*.

Document: etu.3D_ALI. June 1992. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/3d_ali*.

Document: README. June 1992. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/3d_ali*.

Document: repressor.3D_ALI. June 1992. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/3d_ali*.

Document: wrp.3D_ALI. June 1992. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/3d_ali*.

Pascarella, S. and P. Argos. 1992. A data bank merging related protein structures and sequences. *Protein Engineering*. 5: 121–137.

2 Berlin RNA Databank

The Berlin RNA databank contains 5S rRNA sequences and 5S rRNA gene sequences. It is divided into three files based on the phylogenetic groups of archaebacteria (*ARCHAE.DAT*), eubacteria (*EUBAC.DAT*), and eukaryotes (*EUKAR.DAT*). Entries within each file are also sorted by phylogenetic group, with the phylogenetic group names given in brackets (1)[8] and further subgroups indented and bracketed (2). Sequence entries within each group are sorted alphabetically. If different publications give different sequences for the same species or strain, a capital suffix letter is added to the ID.

2.1 Explanation of a Berlin Record

Berlin has a basic EMBL format[9] that appears in many databases. The first line of each record (3) is the ID line, which gives entry identification in the form

> GENUS.SPECIES.MOLECULE; type of entry; number of BasePairs.

[8] See line 1 of the sample entry.

[9] A modified form of the EMBL format was used in 3D_ALI.

The DaTe line (4) indicates when the entry was created. The DEscription line (5) states the type of molecule in this entry, while the source of the molecule is given in the OrganismSpecies (6) and OrganismClassification (7) lines. Subspecies (#SUBSP), strain (#STRAIN) (8), variety (#VAR), locus (#LOCUS), cell type (#CELLTYPE), or clone (#CLONE) information may be given under the OS leader. References are set off by lines such as (9), which contains the ReferenceNumber. ReferenceAuthor (10), ReferenceTitle (11), and ReferenceLocation (12) lines are also included. The first line of the SeQuence (13) information contains the number of base pairs followed by the number of adenines, cytosines, thymines, and guanines. The sequence is then given without a new leader specifier created. The literature file number is given in the LT line (14). The SP line (15) designates the species for use in alignments and phylogenetic trees, including genus and species specifications followed by "(phylum, class)." SA lines (16) contain the universal sequence alignment. Alignment gaps are indicated by "-"; helix and bulge boundaries are marked by "[ ]" (or "!" where the end of a helix is the start of the next). Odd base pairs are parenthesized by "<" ">". The SI field (17) answers the question, Is secondary structure information included in the SA field? And finally, the record is ended with "//", as in (18).

2.2 Obtaining the Berlin Databank

Berlin is available from EMBL by anonymous FTP:

```
ftp ftp.embl-heidelberg.de          [or ftp 192.54.41.33]
cd /pub/databases/berlin
```

Information can also be obtained by e-mail:

```
To: NETSERV@EMBL-Heidelberg.DE
help Berlin                          (as message body)
```

Questions can be sent

```
To: NET-HELP@EMBL-Heidelberg.DE
```

The primary references for Berlin are

> Specht T., J. Wolters, and V. Erdmann. 1990. Compilation of 5S rRNA and 5S rRNA gene sequences. *Nucleic Acids Research* (Suppl.). 18: 2215-2230.
>
> Specht T., J. Wolters, and V. Erdmann. 1991. *Nucleic Acids Research* (Suppl.). 19: 2189-2191.

2.3 Sample Berlin Entries

```
 1 ⟹ [ARCHAEBACTERIA DIVISION II (EOCYTA)]
 2 ⟹   [SULFOLOBALES/THERMOPROTEALES]
 3 ⟹ ID   DESULFUROCOCCUS.MOBILIS.5SRRNA; DNA; XXX BP.
 4 ⟹ DT   30-AUG-1988
 5 ⟹ DE   5S RRNA
 6 ⟹ OS   DESULFUROCOCCUS MOBILIS
 7 ⟹ OC   PROKARYOTA; ARCHAEBACTERIA; EOCYTA; SULFOLOBALES;
      OC   DESULFUROCOCCACEAE.
 9 ⟹ RN   [1]
10 ⟹ RA   KJEMS J., GARRETT R.A.;
11 ⟹ RT   NOVEL EXPRESSION OF THE RIBOSOMAL RNA GENES IN THE EXTREME
      RT   THERMOPHILE AND ARCHAEBACTERIUM DESULFUROCOCCUS MOBILIS;
10 ⟹ RL   EMBO J. 6:3521-3530(1987).
13 ⟹ SQ   SEQUENCE  133 BP;  24 A;  45 C;  19 T;  45 G.
           ACGGTGCCCG ACCCGGCCAT AGTGGCCGGG CAACACCCGG TCTCGTTTCG AACCCGGAAG
           TTAAGCCGGC CACGTCAGAA CGGCCGTGAG GTCCGAGAGG CCTCGCAGCC GTTCTGAGCT
           GGGATCGGGC ACC
14 ⟹ LT   50521
15 ⟹ SP   DESULFUROCOCCUS MOBILIS       (TH.PROTEALES, DESULFUROCOCC.)
16 ⟹ SA    A C[G G T G C C C G A - - C C C G G C]C - A T - A[G T - G G
      SA    - C C G]G G C - A A C[A C - C C G G]T C - - T C G T T T C G
      SA    A A C[C C G G[A - A - - - -]G T]T A A G C[C G G C C - - A C
      SA   ]G -[T C A G A A C - - G G C]- C[G T G A G G[T]C C]- G A G A
      SA    -[G G C C T C G C]A -[G C C G T T - - - - C - - - T - G A!G
      SA    C T G G G[A]T C G G G C A C C]-
17 ⟹ SI   YES
18 ⟹ //
      ID   PYRODICTIUM.OCCULTUM.5SRRNA; RNA; 130 BP.
      DT   08-NOV-1989
      DE   5S RRNA GENE
 8 ⟹ OS   PYRODICTIUM OCCULTUM #STRAIN PL-19 (=DSM 2709)
      OC   PROKARYOTA; ARCHAEBACTERIA; EOCYTA; SULFOLOBALES; PYRODICTIACEAE.
      RN   [1]
      RA   KAINE B.P., SCHURKE C.M.; STETTER K.O.;
      RT   GENES FOR THE 16S AND 5S RIBOSOMAL RNAS AND THE 7S RNA OF
      RT   PYRODICTIUM OCCULTUM;
      RL   SYSTEM. APPL. MICROBIOL. 12:8-14(1989).
      SQ   SEQUENCE  130 BP;  20 A;  47 C;  14 U;  49 G.
           UGGCCCGACC CGGCCAUAGC GGCCGGGCAA CACCCGGACU CAUGUCGAAC CCGGAAGUUA
           AGCCGGCCGC GUUGGGGGAU GCUGUGGGGU CCGCGAGGCC CCGCAGCGCC CCCAAGCCGG
           GAUCGGGCCG
      LT   50525
      SP   PYRODICTIUM OCCULTUM          (SULFOLOBALES, PYRODICTIACEAE)
      SA    - - -[U G G C C C G A - - C C C G G C]C - A U - A[G C - G G
      SA    - C C G]G G C - A A C[A C - C C G G]A C - - U C A U G U C G
      SA    A A C[C C G G[A - A - - - -]G U]U A A G C[C G G C C - - G C
      SA   ]G -[U U G G G G G[A]- U G C]- U[G U G G G G[U]C C]- G C G A
      SA    -[G G C C C C G C]A -[G C G C C C - - - - C - - - C - A A!G
      SA    C C G G G[A]U C G G G C C G]- -
      SI   YES
      //
```

2.4 References

Document: archae.dat. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/berlin*.

Document: intro.dat. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/berlin*.

Document: HELP BERLIN. May 1992. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/help*.

Document: read.me. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/berlin*.

3 Blocks

The Blocks database seeks to record local sequence homology and is similar in ideology to ProSite. The most highly conserved regions of proteins can be represented as "blocks" of locally aligned sequence segments. Each block in Blocks represents a local multiple alignment of ungapped segments from a group of related proteins. Blocks contains alignments that range from 4 residues to 249 residues. There are currently 2,302 blocks representing 619 groups, collections of proteins that share sequence similarity, documented in ProSite 10.2. (Blocks is highly linked to ProSite and hence SWISS-PROT and other EMBL databases.)

3.1 Explanation of a Blocks Record

An ID line (1)[10] begins a Blocks entry and is derived from the ProSite **ID**. The ACcession line assigns the block number and is derived from *prosite.dat* PS#, while DEscription is taken from **DE** of *prosite.dat*. The BL line (4) summarizes PROTOMAT information and includes a motif designation, the width of the block, the number of sequences in the block, and information relating to the strength of the homology. The alignment follows the BL line (5). The SWISS-PROT ID for the sequence is followed by the number of the first residue in the alignment and then the sequence segment. The record is ended by "//", line (6).

[10] See line 1 of the sample entry.

3.2 Sample Blocks Entries

```
1 ⟹ ID   FIBRONECTIN_2; BLOCK
2 ⟹ AC   BL00023; distance from previous block=(64,1927)
3 ⟹ DE   Type II fibronectin collagen-binding domain proteins.
4 ⟹ BL   WDC motif; width=24; seqs=11; 99.5%=551; strength=2364
5 ⟹ COG2_HUMAN (   367) GRSDGKMWCATTANYDDDRKWGFC

     COG9_HUMAN (   248) GRSDGLPWCSTTANYDTDDRFGFC

     FA12_HUMAN (    65) GRPGPQPWCATTPNFDQDQRWGYC

     MANR_HUMAN (   186) GRSDGWLWCGTTTDYDTDKLFGYC

     SFP1_BOVIN (   111) IGSMWMSWCSLSPNYDKDRAWKYC

     SFP3_BOVIN (    92) GSTFMNYWCSLSSNYDEDGVWKYC

     FINC_BOVIN (   407) GRRDNMKWCGTTQNYDADQKFGFC
     FINC_HUMAN (   438) GRRDNMKWCGTTQNYDADQKFGFC
       FINC_RAT (   438) GRRDNMKWCGTTQNYDADQKFGFC

     MPRI_BOVIN (  1928) VESRARLWCATTANYDRDHEWGFC
     MPRI_HUMAN (  1919) IESRAKLWCSTTADYDRDHEWGFC
6 ⟹ //
     ID   PAIRED_BOX; BLOCK
     AC   BL00034A; distance from previous block=(7,37)
     DE   'Paired box' domain proteins.
     BL   RRP motif; width=60; seqs=8; 99.5%=1137; strength=3781
     PAX1_MOUSE (    41) DISRQLRVSHGCVSKILARYNETGSILPGAIGGSKPRVTTPNVVKHIRDYKQGDPGIFAW

     PAX3_MOUSE (    71) VISRQLRVSHGCVSKILCRYQETGSIRPGAIGGSKPKQVTTPDVEKKIEEYKRENPGMFS

     PAX8_MOUSE (    46) DISRQLRVSHGCVSKILGRYYETGSIRPGVIGGSKPKVATPKVVEKIGDYKRQNPTMFAW

     HMGD_DROME (    56) VISRQLRVSHGCVSKILNRFQETGSIRPGVIGGSKPRVATPDIESRIEELKQSQPGIFSW
     HMGP_DROME (    57) VISRQLRVSHGCVSKILNRYQETGSIRPGVIGGSKPKVTSPEIETRIDELRKENPSIFSW
     HMPR_DROME (    64) VISRQLRVSHGCVSKILNRYQETGSIRPGVIGGSKPRIATPEIENRIEEYKRSSPGMFSW

     PAX6_BRARE (    60) DISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVGKIAQYKRECPSIFAW
     PAX6_HUMAN (    41) DISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAW
     //
```

3.3 Obtaining the Blocks Database

Blocks databases and associated programs are available from NCBI and EMBL by anonymous FTP:

ftp ncbi.nlm.nih.gov
cd repository/Blocks

ftp ftp.embl-heidelberg.de [or ftp 192.54.41.33]
cd /pub/databases/Blocks

Information can be obtained by e-mail:

To: Blocks@howard.fhcrc.org
help (as message body)

To: NETSERV@EMBL-Heidelberg.DE
help Blocks (as message body)

Questions can be sent

To: henikoff@sparky.fhcrc.org
To: henikoff@howard.fhcrc.org

To: NET-HELP@EMBL-Heidelberg.DE

The primary reference for Blocks is

> Wallace, J., and S. Henikoff. 1992. PATMAT: A searching and extraction program for sequence, pattern and block queries and databases. *CABIOS*. 8: 249–254.

3.4 References

Blocks Database (BLOCKS.DAT_6.2). Release 6.2, August 1993. Obtained from anonymous FTP to *ncbi.nlm.nih.gov* in */repository/blocks*.

Document: Announce.blocks_6.2. August 1993. Obtained from anonymous FTP to *ncbi.nlm.nih.gov* in */repository/blocks*. (Equivalent to: Document: blocks.doc. August 1993. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/blocks*.)

Document: Blocks E-Mail Searcher. September 1993. Obtained from e-mail to *blocks@howard.fhcrc.org* with message *help*.

Document: HELP BLOCKS. January 1993. Obtained from e-mail to *NETSERV@EMBL-Heidelberg.DE* with message *help blocks*.

Henikoff, S., and J. Henikoff. 1993. Protein family classification based on searching a database of blocks (Document: blockman.ps). Obtained from anonymous FTP to *sparky.fhcrc.org* in */blocks*.

4 DSSP (Dictionary of Secondary Structure of Proteins)

Secondary structures assigned in the PDB are often subjective and sometimes incomplete. DSSP is a program whose goal is to approximate the intuitive notion of secondary structure with an objective algorithm. The DSSP database represents the output of running the DSSP (Define Secondary Structure of Proteins) program against the PDB. Wolfgang Kabsch and Chris Sander created DSSP. Currently, there are no README, help files, tutorials, or user guides for DSSP, so the following description is not complete.

The DSSP algorithm assigns secondary structure on the basis of a pattern recognition of features extracted from X-ray coordinates. Larger secondary structures are recognized as repeats of the elementary H-bonding patterns of "turn" and "bridge." Whereas the presence or absence of a pattern is based on a continuum of parameters, the existence of an H–bond is determined by a single decision parameter, bond energy. Using H–bonding patterns to find patterns thus greatly simplifies the determination of secondary structural features. "Helices" are defined as repeating turns, "ladders" are formed of repeating bridges, and connected ladders form "sheets". "Bends" are defined by areas displaying locally high curvature, while chirality is determined by the torsional handedness of consecutive alpha carbons. DSSP also assigns solvent exposure for the protein or a section of the protein. DSSP thus defines, consistently, secondary features of PDB proteins used in the creation of other data sets such as HSSP, FSSP, and 3D_ALI.

4.1 Explanation of a DSSP Record

A DSSP file as shown in the sample entry begins with a title bar (1)[11] followed by the primary literature citation for DSSP (2). The HEADER (3), COMPND (4), SOURCE (5), and AUTHOR (6) lines are obtained directly from the PDB.[12] Line (7) lists, as indicated after the numeric values, the total number of residues, the number of chains, the total number of disulfide bridges, the number of intrachain disulfide bridges, and the number of interchain disulfide bridges. Line (8) gives the accessible surface area of the protein. Lines between (8) and (9) likewise give an explanation of the numeric values, following the values with information of the general form

total number of hydrogen bonds of type XXX
total number of hydrogen bonds of type XXX per 100 residues

Line (9) provides the basis for a histogram where

- line (10) gives the number of residues per alpha helix,
- line (11) gives the number of parallel bridges per ladder,
- line (12) gives the number of antiparallel bridges per ladder, and
- line (13) gives the number of ladders per sheet.

[11] See line 1 of the sample entry.

[12] The complete PDB lines were not included, only information contained therein.

The bulk of the entry begins after line (14), which specifies

- # - DSSP specifier, sequential residue number
- RESIDUE - the residue number as taken from the PDB file[13]
- AA - the residue name as taken from the PDB file
- STRUCTURE - the structure assigned by DSSP (see below)
- BP1 and BP2 - bridge partners' residue numbers (BP2 suffixed with beta sheet label)
- ACC - number of water molecules residue contacts
- N-H–>O - hydrogen bond information (There are two subcolumns of information; denote these X,Y.)
  - X denotes the respective atom of this hydrogen bond. For example, if the given residue is regarded as residue I, the the N-H bond of I is bound to the C=O of the residue in the I+X position.
  - Y contains the electrostatic hydrogen bond energy of this bond.
- O–>H-N, N-H–>O, and O–>H-N - hydrogen bond information, in structure paralleling above
- TCO - the cosine of the angle between the C=O atoms of this residue and the C=O atoms of the preceding residue
- KAPPA - the virtual bond angles of α carbons of this residue (I) and the I-2 and I+2 residues[14]
- ALPHA - the virtual torsional angle of the α carbons of residues I-1, I, I+1, I+2 [15]
- PHI and PSI - dihedral angles
- X-CA, Y-CA, and Z-CA - X,Y,Z coordinates of the α carbon, respectively.

The STRUCTURE column is divided into many subcolumns, with the sequence and structural assignments listed vertically. The first subcolumn under STRUCTURE provides the amino acid sequence. The second subcolumn provides a summary of the structure based on the subsequent columns up to the BP columns (these are absolute columns 19–38). The summary structure, of subcolumn 2 under STRUCTURE, uses the following abbreviations.

- H indicates a 4-helix (α-helix)
- B indicates an isolated β-bridge
- E (16) indicates an extended strand that participates in a β-ladder[16]

[13] This is sometimes followed by a capital letter specifying the chain.

[14] Used when defining bend.

[15] Used when defining chirality.

[16] A ladder is defined as a set of one or more consecutive bridges of identical parallel type. A sheet is defined as a set of ladders that are connected by shared residues.

- G indicates a 3-helix (3_{10}-helix)
- I indicates a 5-helix (π-helix)
- T (17) indicates an hydrogen bonded turn
- S indicates a bend

The succeeding columns provide secondary structural details.

- subcolumn 3 – 3-turns/helix (16)(17) [17]
- subcolumn 4 – 4-turns/helix
- subcolumn 5 – 5-turns/helix
- subcolumn 6 – geometrical bend
- subcolumn 7 – chirality, which is "+" (16) or "-" (18)
- subcolumn 8 and 9 – beta sheet information [18]

[17] In all subcolumns regarding helix formation:

- ">" marks the residue whose CO makes the hydrogen bond (15)
- "<" denotes the residue whose NH participates in the hydrogen bond (18)
- "X" is used for residues where both the CO and the NH form hydrogen bonds.

[18] Capital letters indicate antiparallel, lower-case letters parallel participation of the residue. Capital letters are also used to indicate sheets, line (19). "!" is used to indicate a chain break. " " implies a residue of low curvature that does not participate in a hydrogen bond. Disulfide bonds are treated as primary structure and are indicated by lower-case letters in the residue name.

4.2 Sample DSSP Entry

```
 1 ⟹ **** SECONDARY STRUCTURE DEFINITION BY THE PROGRAM DSSP, VERSION OCT. 1988 **** DATE= 6-APR-1992
 2 ⟹ REFERENCE W. KABSCH AND C.SANDER, BIOPOLYMERS 22 (1983) 2577-2637
 3 ⟹ HEADER    HYDROLASE(CARBOXYLIC ESTERASE)          08-OCT-91   1ACE
 4 ⟹ COMPND    ACETYLCHOLINESTERASE (E.C.3.1.1.7)
 5 ⟹ SOURCE    ELECTRIC RAY (TORPEDO $CALIFORNICA)
 6 ⟹ AUTHOR    J.L.SUSSMAN,M.HAREL,I.SILMAN
 7 ⟹   526  2  3  2  1 TOTAL NUMBER OF RESIDUES, NUMBER OF CHAINS, NUMBER OF SS-BRIDGES(TOTAL,INTRACHAIN,INT
 8 ⟹  19205.0   ACCESSIBLE SURFACE OF PROTEIN (ANGSTROM**2)
       350 66.5   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(J)  , SAME NUMBER PER 100 RESIDUES
        46  8.7   TOTAL NUMBER OF HYDROGEN BONDS IN     PARALLEL BRIDGES, SAME NUMBER PER 100 RESIDUES
        28  5.3   TOTAL NUMBER OF HYDROGEN BONDS IN ANTIPARALLEL BRIDGES, SAME NUMBER PER 100 RESIDUES
         0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-5), SAME NUMBER PER 100 RESIDUES
         0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-4), SAME NUMBER PER 100 RESIDUES
         3  0.6   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-3), SAME NUMBER PER 100 RESIDUES
         0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-2), SAME NUMBER PER 100 RESIDUES
         1  0.2   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-1), SAME NUMBER PER 100 RESIDUES
         0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+0), SAME NUMBER PER 100 RESIDUES
         0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+1), SAME NUMBER PER 100 RESIDUES
        40  7.6   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+2), SAME NUMBER PER 100 RESIDUES
        67 12.7   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+3), SAME NUMBER PER 100 RESIDUES
       142 27.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+4), SAME NUMBER PER 100 RESIDUES
        10  1.9   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+5), SAME NUMBER PER 100 RESIDUES
 9 ⟹   1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30     *** HIST
10 ⟹   0  0  0  2  1  1  4  1  1  0  4  0  0  1  0  2  0  0  0  1  0  0  0  0  0  0  0  0  0  0    RESIDUES
11 ⟹   2  1  2  2  1  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    PARALLEL
12 ⟹   1  0  1  3  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    ANTIPARAL
13 ⟹   1  1  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    LADDERS P
14 ⟹   #  RESIDUE AA STRUCTURE BP1 BP2  ACC   N-H-->O  O-->H-N  N-H-->O  O-->H-N    TCO  KAPPA ALPHA  PHI
15 ⟹     1    4   S    >         0   0  126    0, 0.0   2,-3.2   0, 0.0   3,-0.7   0.000 360.0 360.0 360.0
16 ⟹     2    5   E  T 3   +     0   0   73    1,-0.3  23,-0.1   2,-0.1 100, 0.0  -0.287 360.0  30.2  55.2 -
17 ⟹     3    6   L  T 3  S+     0   0   42   -2,-3.2  11,-3.0  11,-0.1   2,-0.7   0.454 103.0  79.2 -96.1 -
18 ⟹     4    7   L  E <   -A   13   0A  33   -3,-0.7   2,-0.4   9,-0.2   9,-0.2  -0.922  60.9-179.6-104.3 1
         5    8   V  E     -A   12   0A   3    7,-3.3   7,-3.6  -2,-0.7   2,-0.5  -0.936  22.7-146.5-115.1 1
         6    9   N  E     +A   11   0A  94   -2,-0.4   2,-0.3   5,-0.3   5,-0.2  -0.895  24.6 175.9 -99.4 1
         7   10   T  E >   -A   10   0A   7    3,-2.2   3,-1.2  -2,-0.5 173,-0.1  -0.804  49.7 -99.3-119.3 1
         8   11   K  T 3  S+     0   0  124  171,-0.3   3,-0.1  -2,-0.3 172,-0.1   0.532 126.3  54.4 -61.0
         9   12   S  T 3  S-     0   0   15    1,-0.4  42,-3.2  41,-0.1   2,-0.3   0.731 125.1 -80.3 -91.0 -
        10   13   G  E <   -A    7   0A   9   -3,-1.2  -3,-2.2  40,-0.2  -1,-0.4  -0.907  61.9 -38.5 155.0-1
19 ⟹    11   14   K  E     -Ab   6  54A  78   42,-0.8  44,-3.4  -2,-0.3   2,-0.4  -0.597  48.4-166.5 -80.4 1
        12   15   V  E     -Ab   5  55A   0   -7,-3.6  -7,-3.3  42,-0.3   2,-0.5  -0.996   8.4-156.7-129.6 1
        13   16   M  E     -Ab   4  56A  70   42,-2.8  44,-3.9  -2,-0.4  -9,-0.2  -0.940  18.3-167.0-114.9 1
        14   17   G        -     0   0   6  -11,-3.0   2,-0.3  -2,-0.5  13,-0.2   0.113  14.7-115.7 -81.1-1
        15   18   T  E     -C   26   0B  44   11,-2.3  11,-2.7  42,-0.3   2,-0.8  -0.958  18.4-119.2-143.0 1
        16   19   R  E     -C   25   0B 118   -2,-0.3   9,-0.2   9,-0.2  42, 0.0  -0.825  32.8-168.9-104.3 1
        17   20   V  E     -C   24   0B  22    7,-1.4   7,-2.7  -2,-0.8   2,-0.1  -0.729  21.2-119.2 -95.9 1
        18   21   P  E     -C   23   0B  98    0, 0.0   2,-0.3   0, 0.0   5,-0.3  -0.503  31.8-176.9 -77.1 1
        19   22   V        -     0   0   2    3,-1.7  -2, 0.0   1,-0.3 112, 0.0  -0.917  49.1 -64.4-154.9 1
        20   23   L  S    S-     0   0  64   -2,-0.3  -1,-0.3   1,-0.2 111, 0.0  -0.081 120.9 -11.0  36.6-1
```

[residues 21-527 deleted]

4.3 Obtaining DSSP

DSSP is available from EMBL by anonymous FTP:

```
ftp ftp.embl-heidelberg.de            [or ftp 192.54.41.33]
cd pub/databases/protein_extras/dssp
```

Questions can be sent

To: NET-HELP@EMBL-Heidelberg.DE
To: sander@embl-heidelberg.de

The primary reference for DSSP is

> Kabash, W., and C. Sander. 1983. Dictionary of protein secondary structures: Pattern recognition of Hydrogen-bonded and geometrical features. *Biopolymers*. 22: 2577–2637.

4.4 References

Document: 1ace.dssp. April 1992. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/protein_extras/dssp*.

Document: 1ak3.dssp. October 1992. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/protein_extras/dssp*.

Document: 9xim.hssp. October 1993. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/protein_extras/hssp*.

Document: cpk2.dssp. October 1992. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/protein_extras/dssp*.

Document: dssp_help.doc. Obtained from Chris Sander at *Chris.Sander@EMBL-Heidelberg.DE*.

Document: zif1.dssp. May 1992. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/protein_extras/dssp*.

Kabash, W., and C. Sander. 1983. Dictionary of protein secondary structures: Pattern recognition of Hydrogen-bonded and geometrical features. *Biopolymers*. 22: 2577–2637.

5 EMBL Nucleotide Sequence Database

The EMBL nucleotide database (EMBL-DB)[19] is a primary repository for genetic sequences. EMBL-DB contains information scanned from the literature and submitted directly. Many journals now require sequences to have been submitted to the EMBL-DB before publication. Sequences are always listed 5' to 3', with the sequence numbering beginning at one with the most 5' base.

[19] We are using the notation EMBL-DB for the EMBL nucleotide database to prevent confusion when referring to other "EMBL databases," databases that are housed at EMBL, and the "physical entity" of EMBL.

5.1 Explanation of an EMBL Record

EMBL-DB entries begin with an ID line (1)[20] which includes the following information.

- entry name, generally the first character of both the genus and species followed by a descriptive three characters
- data class reflecting the level of data verification
    - standard - complete and verified records
    - unreviewed - complete but not verified manually
    - preliminary - only sequence and citation have been verified
    - unannotated - just entry name, citation, and sequence
    - backbone - entries from NCBI backbone database
- molecule type, either DNA or RNA[21]
- division[22]
- sequence length

[20] See line 1 of the sample entry.

[21] Sequences are given in *in vivo* state. Hence, cDNA is given as an RNA sequence.

[22] The EMBL-DB is grouped into mutually exclusive subsets on the basis of taxonomy. Divisons are

- EST - Expressed Sequence Tags or Transcribed Sequence Fragments,
- PHG - Bacteriophage,
- FUN - Fungi,
- INV - Invertebrates,
- ORG - Organelles,
- PRI - Primates,
- ROD - Rodents,
- MAM - Other Mammals,
- VRT - Other Vertebrates,
- PLN - Plants,
- PRO - Prokaryotes,
- SYN - Synthetic,
- UNC - Unclassified, and
- VRL - Viruses.

Each section of an EMBL-DB record is separated with an "XX" line (2) simply for ease of human reading. The ACcession number (3) is important in that it provides stable access to records that may have been merged or been given a new entry name between releases. Each AC number is terminated with a semicolon.

Two DaTe lines are given, the date of record creation (4) and the date of the last record update (5). The DEscription line (6) gives a free-format description of the molecule. KeyWord lines (7) provide another, more formatted, text-based handle on the record. The OrganismSpecies line (8) generally includes the Latin genus and species followed by the more common English name(s) in parentheses. OrganismClassification lines (9) delineate the full taxonomic classification.

Citation information is given next in the EMBL-DB record. Each reference is assigned a ReferenceNumber (10) to clearly separate references and may be followed by ReferenceComment lines. The ReferencePosition (11) is used in cases in which one or more continuous base sequences can be attributed to the reference, thus removing ambiguity. ReferenceAuthors (11), ReferenceTitle (12), and ReferenceLocation (13) follow. In this case, there is no RT; this record might represent a sequence that was submitted only to the database. Therefore, only a ";" is located on the RT line. For publications, the RL line includes the journal citation. The function of the RL lines is currently being extended to include contact information, such as Internet addresses.

DR lines may follow the reference information providing cross references to other databases, in particular SWISS-PROT entries obtained by translating the EMBL-DB entry. The following are example DR abbreviations.

- SWISS-PROT for the SWISS-PROT database
- TFD for the Transcription Factor Database
- EPD for the Eukaryotic Promoter Database
- FLYBASE for the *Drosophila* Genetic Database
- CPGISLE for the CpG Islands Database

An example DR line is

```
DR   SWISS-PROT; P03593; V90K_AMV
```

where the first section indicates the database (SWISS-PROT), followed by the accession number or primary access identifier (P03593), and concluding with the entry name or secondary identifier (V90K_AMV).

5.2 Sample EMBL Nucleotide Sequence Database Entry

```
 1 ⟹  ID   ATTS1692   standard; RNA; EST; 341 BP.
 2 ⟹  XX
 3 ⟹  AC   Z26952;
       XX
 4 ⟹  DT   16-OCT-1993 (Rel. 37, Created)
 5 ⟹  DT   16-OCT-1993 (Rel. 37, Last updated, Version 1)
       XX
 6 ⟹  DE   A. thaliana transcribed sequence; clone YBY001; 5' end; similar to
       DE   dTDP-D-GLUCOSE 4,6-DEHYDRATASE - Shigella flexneri.
       XX
 7 ⟹  KW   expressed sequence tag; partial cDNA sequence.
       XX
 8 ⟹  OS   Arabidopsis thaliana
 9 ⟹  OC   Eukaryota; Plantae; Embryobionta; Magnoliophyta; Magnoliopsida;
       OC   Dilleniidae; Capparales; Brassicaceae.
       XX
10 ⟹  RN   [1]
11 ⟹  RP   1-341
12 ⟹  RA   Morris P.C., Guerrier D., Barbet JC., Giraudat J.;
13 ⟹  RT   ;
14 ⟹  RL   Submitted (23-SEP-1993) to the EMBL Data Library by:
       RL   CNRS, GDR-1003 ACS, INRA, laboratoire de Biologie Moleculaire, BP
       RL   27, 31326 Castanet-Tolosan cedex, France.
       RL   E-mail:gdr-svp@toulouse.inra.fr. On behalf of: Genetique
       RL   Moleculaire d'Arabidopsis, ISV - UPR40, CNRS, Avenue de la
       RL   Terrasse, 91198 Gif-sur-Yvette Cedex, France.
       RL   E-mail:Giraudat@cnrs-gif.fr.
       XX
```

[reference 2 deleted]

```
       XX
15 ⟹  CC   Cloning vector: Lambda ZAPII oriented ;
       XX
       CC   full automatic.
       XX
       CC   similarity detected by BLASTX against GenPept entry YEPRFBFGHJ_6.
       XX
16 ⟹  FH   Key             Location/Qualifiers
       FH
17 ⟹  FT   source          1..341
18 ⟹  FT                   /organism="Arabidopsis thaliana"
19 ⟹  FT                   /clone="YBY001"
20 ⟹  FT                   /tissue_type="Flowering tips of ecotype Landsberg erecta"
21 ⟹  FT                   /clone_lib="Gif-SiliqueB"
       XX
22 ⟹  SQ   Sequence 341 BP; 89 A; 74 C; 72 G; 102 T; 4 other;
23 ⟹       tgattactgt tctaatctga agaagcttaa tccttctaaa tcctctccca acttcaagtt        60
            tgtgaaagga gatatcgcca gtgcctgatc tcgtcaacta ccttctcatc actgaagaaa       120
            tcgacaccat tatgcacttt gctgctcaaa cccatgttga caattctttc ggtaatagct       180
            ttgagtttac caagaacaat atctatggta cccatgtcct tttggaagct tgtaaagtca       240
            ctggccagat caggaggttc atccatgtga gtactgatga ggtctatggg agagactgga       300
            tgnggatnnc ttcagtgggg tnattcacgg agggcttctc a                           341
24 ⟹  //
```

Comments are free text indicated by "CC" (15). Feature tables, on the other hand, are formatted. FH lines (16) make the record more easily read by humans by providing column headers for the feature table. The feature table (FT) provides a framework for formatted annotations. Source annotations (17) include "/organism" (18), "/clone" (19), "tissue" (20), and "clone_lib" (21).

The actual sequence follows an SQ line (22) which provides sequence information such as length and the number of adenines, cytosines, guanines, thymines, and other modified nucleotides. The sequence is given in the standard one-letter amino acid code. Each sequence line (23) is concluded with a count of the base pairs on that line. The record terminates with "//" (24).

5.3 Obtaining EMBL

EMBL nucleotide sequence databases is available from EMBL by anonymous FTP:

ftp ftp.embl-heidelberg.de [or ftp 192.54.41.33]
cd /pub/databases/embl/release

Information can also be obtained by e-mail:

To: NETSERV@EMBL-Heidelberg.DE
help Nuc (as message body)

Questions can be sent

To: NET-HELP@EMBL-Heidelberg.DE

5.4 References

Document: 931031.dat. October 1993. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/embl/new*.

Document: featuretable.doc. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/embl/doc*.

Document: HELP NUC. October 1993. Obtained from e-mail to *NETSERV@EMBL-Heidelberg.DE* with message *HELP NUC*.

EMBL Data Library: Nucleotide Sequence Database: Release Notes. 1993. Release 36, September 1993. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/embl/doc*.

EMBL Data Library: Nucleotide Sequence Database: User Manual. 1993. Release 36, September 1993. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/embl/doc*.

6 EMP (Enzymes and Metabolic Pathways)

EMP includes data relating to two subjects — enzymology and metabolism. The enzymological section includes data on over 70enzymes classified by EC numbers. In addition, it includes records on 600 new, unclassified enzymes (i.e., enzymes that have not yet been assigned EC numbers). The database now includes data on approximately 3000 distinct enzymes. Records in the database are structured distillations of the factual content of research articles. An attempt has been made to capture all relevant facts included in each article (there are over 300 distinct fields used to encode data from articles). Encoded articles are selected from the leading international journals on biochemistry. A journal article is encoded into one or more EMP records (an average of about two records is made for each article). The database now contains over 14,500 records. Enzyme mechanisms are represented in both textual and graphical forms (the graphical representation is based on an extended form of the Cleland notation).

Besides records on enzymatic data, the database also contains over 1000 records describing metabolic pathways. These include tabular, graphic, and equational representations of pathways and their regulatory mechanisms. This collection is being used to generate summaries for a number of organisms that are being actively sequenced; these include *Mycoplasma capricolum*, *Escherichia coli*, *Salmonella typhimurium*, *Bacillus subtilus*, *Pseudomonas aeruginosa*, and *Saccharomyces cerevisiae.* The database is being used to generate reconstructions of the metabolic networks for these organisms, often from less than 50% of the complete genome.

These records contain data on over 1500 organisms. As the following shows, attempts are being made to generate relatively complete collections for the most important model organisms.

- 1300 records on humans
- 1300 records on *Escherichia coli*
- 1000 records on *Saccharomyces cerevisiae*
- 1300 records on plants

6.1 Explanation of a EMP Record

Because EMP is designed to allow for a wide variety of information, a multitude of defined fields can be used. For this reason, we have included a sample list of fields as well as sample entry at the end of this section. However, the sample list of fields represents less than half of the field designations EMP offers.

6.2 Obtaining EMP

EMP is in preparation for a CD-ROM release for public use. Information about EMP can be obtained from http://www.mcs.anl.gov/home/compbio/ emp.html or e-mail to gaasterland@mcs.anl.gov.

6.3 Sample EMP Field Headers

**	Staff Footnote
***	Annotator's Footnote
AA	Amino Acids
AAB	Annotator's Abstract
AAS	Amino Acid Sequence or Crossreference
AB	Abstract
ABC	Absorption Coefficient
ABD	Antibody
AC	Field Name
ACC	Apparent Catalytic Constant
CPE	Cell Cycle Period
CPH	Cell Cycle Phase
CRR	Cross-Reactivity
CS	Cell Species or Extracellular Fluid
CSH	CD Spectrum Shoulder
CSO	Carbon Source
CSZ	Colony Size
CT	Cell Type
CUC	Cultivation Conditions
CUT	Cultivation Temperature
EC	Enzyme Code
ECO	Enzyme Concentration
EF	Enzyme Form
EFF	Treatment Effect
EN	Recommended Enzyme Name
ENC	Enzyme Composition
EPR	EPR Signals
ESC	Sedimentation Constant
ESE	Elementary Step Equation
ETN	Electron Transfer Number
GT	Generation Time
GY	Growth Yield
HCC	Host Cell Culture Category
HCL	Host Cell Line
HCN	Host Common Name
HCS	Host Cell Species
HCT	Host Cell Type
HDS	Host Developmental Stage
HMT	Host Metabolic Type
HOM	Primary Structure Homology
IDE	Induction Degree
IN	Inhibitor
IND	Inducer
INT	Inhibition Type
IP	Isoelectric Point (pI)
IS	Ionic Strength
IT	Index Terms
IZN	Isozyme Number
KA	Activation Constant
KM	Michaelis Constant
KT	Type of Kinetics or Functional Dependence
LA	Language
LC	Light Conditions
LI	Light Intensity
LIG	Ligand
LOC	Intramolecular Location
LS	Limiting Substrate
MA	Molecular Activity
MAC	Molar Absorption Coefficient
MF	Multifunctional Enzyme Name
MFC	Multifunctional Enzyme Code
MFF	Multifunctional Enzyme Form
MID	Maximal Inhibition Degree
MM	Molecular Mass
MME	Modification Mechanism
MOD	Enzyme Modifier
MOE	Modification Effect
MP	Enzyme Modification Process
MPW	Metabolic Pathway Name
NSO	Nitrogen Source
NSU	Non-Substrate
OC	Optimal Conditions
OCN	Organism Common Name
ON	Other Enzyme Name
OPN	Other Protein Name
OR	Organism Systematic Name
OS	Organizational Source
OTR	Organism Treatment
OVR	Overall Reaction
PA	Preparation Specific Activity
PHO	Optimal pH
PHR	pH Range
PHS	Physiological State
PKV	pK Values
PN	Protein Name
PR	Product
PRF	Protein Form
PRG	Prosthetic Group
PRI	Primer
PS	Purification Steps
PVM	Preparation Maximal Velocity
TI	Title
TO	Optimal Temperature
TPC	Total Protein Content
TR	Enzyme Treatment
VAL	Value of Parameter or Variable
VEL	Overall Reaction Velocity
VL	Variable Ligand

6.4 Sample EMP Entry

```
AN  SEL93335-02/ESM52-4
**  BioBank:SEL01.12.93-2/01.12.93
*** 21 1.58; S 12-01-93 06:19pm/12-01-93 07:54pm
RN  SEL93333-01
CC  Metabolic_pathways
ST  metabolic_map
OR  ‘Escherichia_coli‘
CCC Bacteria
OCN enterobacterium
CTY Gram-negative
MET facultatively_anaerobic; respiratory and heterofermentative;
    chemoorganotrophic; enterobacterial_common_antigen-containing;
    menaquinone-containing; heme-nonrequiring; NAD-nonrequiring;
    organic_nitrogen_sources-nonrequiring; indole-producing;
    methyl_red-positive; Voges-Proskauer-negative;
hydrogen_sulfide-producing;
    arginine-dihydrolase-negative; catalase-positive;
deoxyribonuclease-
    positive; ’beta’-galactosidase-positive;
’gamma’-glutamyltransferase-
    positive; lipase-negative; lysine_decarboxylase-positive;
    nitrate_reductase_A-positive; ornithine_decarboxylase-positive;
    oxidase-negative; phenylalanine_deaminase-negative; tetrathionate_
    reductase-negative; urease-negative; growth_in_KCN-positive;
    nitrate-reducing; gas-producing (D-glucose); acid-producing
(L-arabinose,
    dulcitol, esculin, D-glucose, lactose, maltose, D-mannitol,
D-mannose,
    melibiose, mucate, raffinose, L-rhamnose, salicin, D-sorbitol,
sucrose,
    trehalose, D-xylose); acid-nonproducing (D-adonitol, D-arabitol,
    cellobiose, ‘myo‘-inositol, 2-ketogluconate,
’alpha’-methyl-D-glucoside);
    utilizing: acetate, lactose; nonutilizing: citrate, malonate;
gelatin-
    liquefying; yellow_pigment-nonproducing
TG  Enterobacteriaceae
APN serine_degradation_(cytosol,_plasma_membrane);
    serine_catabolism_(cytosol,_plasma_membrane);
MPW serine--NH(,3)_catabolism_(cytosol,_plasma_membrane)
SPN
L-serine--NH(,3)_catabolism_(tetrahydrofolate,_lipoylprotein)_(cytosol,_plasma_membrane)
OVR L-serine + 2 tetrahydrofolate + lipoylprotein = CO(,2) + NH(,3) +
H(,2)O
          + 2 5,10-methylenetetrahydrofolate + dihydrolipoylprotein
MPW G SERNH3CYTPLM.CAT !!
Serine--NH(,3)_Catabolism_(cytosol,_plasma_membrane)
```

```
MPW E RI |  EC       |  EN                                     |  MFC
      ---------------------------------------------------------------------------
      R1 |  2.1.2.1  |  glycine_hydroxymethyltransferase       |  cytosol

      R2 |  1.4.4.2  |  glycine_dehydrogenase_(decarboxylating) |  1.4.4.2/2.1.2.10
      R3 |  2.1.2.10 |  aminomethyltransferase                 |  1.4.4.2/2.1.2.10

      RI |  MF                     |  SL
      ---------------------------------------------------------------------------
      R1 |  (-)                    |  (-)
      R2 |  glycine_cleavage_system |  plasma_membrane
      R3 |  glycine_cleavage_system |  plasma_membrane


      RI |  RE
                 |  RD |  REV
      ---------------------------------------------------------------------------------------------
      R1 |  5,10-methylenetetrahydrofolate + glycine + H(,2)O = tetrahydrofolate + L-serine  |  B  |  R
      R2 |  glycine + lipoylprotein = `S`-aminomethyldihydrolipoylprotein + CO(,2)           |  F  |  IR
      R3 |  `S`-aminomethyldihydrolipoylprotein + tetrahydrofolate =
                dihydrolipoylprotein + 5,10-methylenetetrahydrofolate + NH(,3) [**]        |  F  |  R


COM synonym_substitutions:
    tetrahydrofolate = (6`S`)-tetrahydrofolate;
    5,10-methylenetetrahydrofolate = (6`R`)-5,10-methylenetetrahydrofolate

GN    RI |  EC       |  GN     |  AGN |  GEN
      ------------------------------------------------------------------
      R1 |  2.1.2.1  |  `glyA` |  (-) |  glycine_A
      R2 |  1.4.4.2  |  `gcvP` |  (-) |  glycine_cleavage_system_P-protein
      R2 |  (-)      |  `gcvH` |  (-) |  glycine_cleavage_system_H-protein
      R3 |  2.1.2.10 |  `gcvT` |  (-) |  glycine_cleavage_system_T-protein

      RI |  GNP                     |  GMP(min) |  AAS
      ------------------------------------------------------------------
      R1 |  EC_2.1.2.1              |  55       |  SWISS-PROT_P00477; PIR_XYECS; PROSITE_PS00096; EC-2D-GEL_G043.8
      R2 |  EC_1.4.4.2              |  (-)      |  (-)
      R2 |  H-protein, lipoylprotein |  (-)      |  SWISS-PROT_P23884; PROSITE_PS00189(lipoamide_binding)
      R3 |  EC_2.1.2.1              |  (-)      |  SWISS-PROT_P27248

      RI |  NS
      ------------------------------------------------------------------
      R1 |  GenBank_J01620; GenBank_J01621; EMBL_V00283; ECOGENE_EG10408
      R2 |  GenBank_M57690; GenBank_M97263; ECOGENE_EG10371
      R3 |  GenBank_M97263; ECOGENE_EG11442

MPW M RI |  SER |  GLY |  CO(,2) |  NH(,3) |  H(,2)O |  THF |  5,10-methylene-THF |  LP |  AMDHLP |  DHLP
      R1 |  -1  |  1   |  0      |  0      |  1      |  -1  |  1                  |  0  |  0      |  0
      R2 |  0   |  -1  |  1      |  0      |  0      |  0   |  0                  |  -1 |  1      |  0
      R3 |  0   |  0   |  0      |  1      |  0      |  -1  |  1                  |  0  |  -1     |  1

DES   5,10-methylene-THF = 5,10-methylenetetrahydrofolate;
      AMDHLP = `S`-aminomethyldihydrolipoylprotein;
      DHLP = dihydrolipoylprotein;
      GLY = glycine;
      LP = lipoylprotein;
      SER = L-serine;
      THF = tetrahydrofolate

REP T RI |  EC       |  RD |  REP                               |  TYP
      R1 |  2.1.2.1  |  B  |  glycine                           |  substrate_repression
      R1 |  2.1.2.1  |  B  |  L-serine                          |  product_repression
      R1 |  2.1.2.1  |  B  |  L-methionine                      |  end_product_repression
      R1 |  2.1.2.1  |  B  |  5-amino_4-imidazole_carboxamide_  |  end_product_repression
                           |        riboside_5'-phosphate       |
      R1 |  2.1.2.1  |  B  |  thymine                           |  end_product_repression
      R2 |  1.4.4.2  |  F  |  inosine                           |  end_product_repression
      R3 |  2.1.2.10 |  F  |  inosine                           |  end_product_repression

IND T RI |  EC       |  IND     |  TYP
      R2 |  1.4.4.2  |  glycine |  feed-forward_induction
      R3 |  2.1.2.10 |  glycine |  feed-forward_induction

SYM 5,10-methylenetetrahydrofolate = (6‘R‘)-5,10-methylenetetrahydrofolate;
    aminomethyltransferase = T-protein = protein_T;
    glycine_dehydrogenase_(decarboxylating) = P-protein = protein_P;
    lipoylprotein = H-protein = protein_H;
    L-serine = serine; tetrahydrofolate = (6‘S‘)-tetrahydrofolate
```

6.5 References

Rayl, K., T. Gaasterland, and R. Overbeek. March 1994. Automating the determination of 3D protein structure. Mathematics and Computer Science Division, Argonne National Laboratory, Preprint MCS-P417-0294.

7 ENZYME

The ENZYME databank contains the EC (Enzyme Code) numbers assigned to known enzymes. Although the ENZYME databank contains a limited amount information, that information is valuable and highly useful as crosslinks between other databases. EC numbers are not assigned or determined in any way by the maintainers of the ENZYME databank but instead are issued by the NC-IUBMB (Nomenclature Committee of the International Union of Biochemistry and Molecular Biology). EC numbers are assigned on the basis of function. Therefore, one physical object of an enzyme may have more than one EC number. EC numbers are a four-place series of the form –.–.–.– where each place further delineates the function in question. The specifications for current EC numbers can be obtained with the ENZYME data bank in the file *enzclass.txt*.[23]

7.1 Explanation of an ENZYME Record

The ENZYME databank was originated by Amos Bairoch. Since EC numbers themselves are flexible and subject to change, the ENZYME databank must cope with unstable records. Each ENZYME record contains first unique ID number (1).[24] This number provides stable access to the database. If an enzyme is assigned a new EC number upon further determination of its function, or if an EC number is deleted upon discovering that an enzyme does not in fact have a specified function, this information is maintained in ENZYME as shown in (2) and (3), respectively. Thus, given an outdated EC number, ENZYME allows a researcher to find the current EC number. The DE line (4) contains either transfer or deletion information, or the recommended enzyme name. Alternative names may be found on the AN line (5). The CA line (6) indicates reactions catalyzed by the enzyme. The CF line (7) indicates known necessary cofactors. Comments lines, CC (8), are free form with each new comment delineated by "-!-" and can provide much additional information. ENZYME is also conveniently cross-referenced to other major databases. DR lines (9) provide the cross-links to SWISS-PROT by giving the SWISS-PROT accession number(s) and entry name(s) for related and pertinent records. SWISS-PROT (described later) is one of the more powerful databases containing many cross-references to other databases. DI lines (10) provide cross-links to MIM (Mendelian Inheritance in Man: Catalogs of autosomal dominant, austosomal recessive, and X-linked phenotypes), which gives information on diseases associated with the enzyme function. Each record is divided by "//" as shown in (11).

[23] See sample enzyme numbers included in this section.

[24] See line 1 of the sample entry.

7.2 Sample EC Numbers

```
1. -. -.-  OXIDOREDUCTASES.
1. 1. -.-   ACTING ON THE CH-OH GROUP OF DONORS.
1. 1. 1.-    WITH NAD(+) OR NADP(+) AS ACCEPTOR.
1. 1. 2.-    WITH A CYTOCHROME AS ACCEPTOR.
1. 1. 3.-    WITH OXYGEN AS ACCEPTOR.
1. 1. 4.-    WITH A DISULFIDE AS ACCEPTOR.
1. 1. 5.-    WITH A QUINONE OR SIMILAR COMPOUND AS ACCEPTOR.
1. 1.99.-    WITH OTHER ACCEPTORS.
1. 2. -.-   ACTING ON THE ALDEHYDE OR OXO GROUP OF DONORS.
1. 2. 1.-    WITH NAD(+) OR NADP(+) AS ACCEPTOR.
1. 2. 2.-    WITH A CYTOCHROME AS ACCEPTOR.
1. 2. 3.-    WITH OXYGEN AS ACCEPTOR.
1. 2. 4.-    WITH A DISULFIDE AS ACCEPTOR.
1. 2. 7.-    WITH AN IRON-SULFUR PROTEIN AS ACCEPTOR.
1. 2.99.-    WITH OTHER ACCEPTORS.
```

[1.3.-.- to 1.18.-.- deleted]

```
2. -. -.-  TRANSFERASES.
3. -. -.-  HYDROLASES.
4. -. -.-  LYASES.
5. -. -.-  ISOMERASES.
6. -. -.-  LIGASES.
```

7.3 Sample ENZYME Entries

```
 1 ⟹ ID   1.1.1.68
 2 ⟹ DE   TRANSFERRED ENTRY: 1.7.99.5.
      //
      ID   1.1.1.70
 3 ⟹ DE   DELETED ENTRY.
      //
      ID   1.1.1.2
 4 ⟹ DE   ALCOHOL DEHYDROGENASE (NADP+).
 5 ⟹ AN   ALDEHYDE REDUCTASE (NADPH).
 6 ⟹ CA   AN ALCOHOL + NADP(+) = AN ALDEHYDE + NADPH.
 7 ⟹ CF   ZINC.
 8 ⟹ CC   -!- SOME MEMBERS OF THIS GROUP OXIDIZE ONLY PRIMARY ALCOHOLS; OTHERS ACT
      CC       ALSO ON SECONDARY ALCOHOLS.
      CC   -!- MAY BE IDENTICAL WITH EC 1.1.1.19, EC 1.1.1.33 AND EC 1.1.1.55.
      CC   -!- A-SPECIFIC WITH RESPECT TO NADPH.
 9 ⟹ DR   P14550, ALDX_HUMAN;  P27800, ALDX_SPOSA;
      //
      ID   1.1.1.29
      DE   GLYCERATE DEHYDROGENASE.
      CA   (R)-GLYCERATE + NAD(+) = HYDROXYPYRUVATE + NADH.
10 ⟹ DI   OXALOSIS II (GLYCERICACIDURIA); MIM:260000.
      DR   P13443, DHGY_CUCSA;
11 ⟹ //
```

7.4 Obtaining ENZYME

ENZYME is available from EMBL by anonymous FTP:

```
ftp ftp.embl-heidelberg.de          [or ftp 192.54.41.33]
cd /pub/databases/enzyme
```

To obtain ENZYME, see also the document *ENZYME.GET*. Information can be obtained by e-mail:

To: NETSERV@EMBL-Heidelberg.DE
help ENZYME (as message body)

Questions can be sent

To: NET-HELP@EMBL-Heidelberg.DE
To: bairoch@cmu.unige.ch

The primary citation for ENZYME data bank is

> Bairoch, A. 1993. The ENZYME data bank. *Nucleic Acids Research*. 21: 3155–3156.

7.5 References

Bairoch, A. 1993. Document name: ENZCLASS.TXT. Release 13.0, July 1993. Obtained from e-mail to *NETSERV@EMBL-Heidelberg.DE* with message *GET ENZYME:ENZCLASS.TXT*.

Bairoch, A. 1993. Document name: ENZYME.GET. Release 13.0, July 1993. Obtained from e-mail to *NETSERV@EMBL-Heidelberg.DE* with message *GET ENZYME:ENZYME.GET*.

Bairoch, A. 1993. ENZYME data bank. Release 13.0, July 1993. Obtained from e-mail to *NETSERV@EMBL-Heidelberg.DE* with message *GET ENZYME:ENZYME.DAT*.

Bairoch, A. 1993. The ENZYME data bank. *Nucleic Acids Research*. 21: 3155–3156.

Bairoch, A. 1993. The ENZYME data bank user manual. Release 13.0, July 1993. Obtained from e-mail to *NETSERV@EMBL-Heidelberg.DE* with message *GET ENZYME:ENZUSER.TXT*.

Document: README. August 1993. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/enzyme*.

8 FSSP (Families of Structurally Similar Proteins)

The FSSP database of Liisa Holm et al. is similar in ideology to the 3D-ALI database. 3D_ALI consists of superpositions that are, in part, obtained from the literature or manipulated by hand. However, FSSP is generated fully automatically and represents the output of running three alignment programs (Suppos, Comp3D, and Dali) against the PDB. For each protein of known structure, searches are performed to find proteins that contain similar three-dimensional substructures. Each

FSSP file contains the search structure, structural alignments with each relatives that has 70-30% sequence homology, and other proteins of the data set that contain substructures significantly similar to the search structure.[25] In other words, each file contains a superpositioned group of structures using one particular structure as the starting point. FSSP groups are nonexclusive: one protein may contain substructures of more than one family. The structural alignments in a file include proteins in the "twilight zone"[26] of sequence homology.

The three different search algorithms each have advantages and disadvantages. (For strong global homologies, the three algorithms give similar alignments.) Suppos is a quick search algorithm based on clustering fragment pairs with similar backbone conformation into a set of globally aligned 3D matches. Suppos finds the largest common 3D substructure within cutoff parameters. Comp3D employs a dynamic programming algorithm to determine the optimal alignment. Insertions, deletions, one-sequence gaps, and gaps in both sequences are all allowed by Comp3D. Comp3D can also detect interdomain motion. Dali uses Monte Carlo optimization to compare distance matrices for pairs of structures. Dali detects topological reconnections and geometrical distortions. Although Dali is the slowest of the three algorithms, it is the most general method and is sensitive and accurate. File names in FSSP include the method used and follow the general format of PDB_entry_name.method.FSSP.[27] A Dali record has been selected for the example here, but records employing other methods use the same format.

8.1 Explanation of a FSSP Record

Each FSSP record has a leading section initiated with the version of the FSSP database (1)[28] and the PDB entry name for the record under PDBID (2). The DATE of the database creation is given on line (3). Information on the database (4), method (5), and constraints (6) (7) follows. The primary reference for FSSP (8) as well as contact (9) and availability information (10) is included. The HEADER (11), COMPND (12), SOURCE (13), and AUTHOR (14) lines of the PDB record are given, in an edited form. The sequence length (15), number of aligned sequences (16), number of distinct chains (17), and chain being used (18) precede NOTATION lines (19). Notation lines are provide valuable information for reading a FSSP file, because they delineate some of the notation used. Since NOTATION lines are found in every record, notation information is always readily available to the user.

[25] Only the best alignment between a pair is included.

[26] The "twilight zone" refers to proteins that display sequence homology that may or may not imply structural similarity. This range is defined differently by different researchers but can be considered to include 0–30% for sequence-based alignments.

[27] For example, the FSSP entry for PDB record 1bmv generated by the Dali algorithm would be 1bmv.dali.FSSP. The corresponding Suppos record would be 1bmv.suppos.FSSP, and so on.

[28] See line 1 of the sample entry.

8.2 Sample FSSP Entry

```
 1 ⟹ FSSP      FAMILIES OF STRUCTURALLY SIMILAR PROTEINS, VERSION 0.2 1993
 2 ⟹ PDBID     1bmv-2
 3 ⟹ DATE      file generated on  6-Oct-93
 4 ⟹ DATABASE  204 chains with 30 % sequence identity cutoff, based on
      DATABASE  PDB-select by Hobohm & al., Protein Science 1, 409-417.
 5 ⟹ METHOD    Dali version 1.0 (1993)
      METHOD    Holm, L., Sander, C. (1993) J.Mol.Biol. in press
 6 ⟹ PARAMETER elastic alignment with similarity threshold 0.20
 7 ⟹ THRESHOLD This file has been filtered to contain only hits that have
      THRESHOLD similarity scores > three standard deviations above the average.
 8 ⟹ REFERENCE Holm, L., Ouzounis, C., Tuparev, G., Vriend, G., Sander, C. (1992)
      REFERENCE A database of protein structure families with common folding
      REFERENCE motifs.  Protein Science 1, 1691-1698
 9 ⟹ CONTACT   e-mail (Internet)    Holm@EMBL-Heidelberg.DE or
      CONTACT   Sander@EMBL-Heidelberg.DE / phone +49-6221-387361 /
      CONTACT   fax +49-6221-387306
10 ⟹ AVAILABLE Free academic use. Commercial users must apply for licence.
      AVAILABLE No incorporation into other databases.
11 ⟹ HEADER    VIRUS                                   09-OCT-89   1BMV
12 ⟹ COMPND    BEAN POD MOTTLE VIRUS (MIDDLE COMPONENT)
13 ⟹ SOURCE    BOUNTIFUL BEAN
14 ⟹ AUTHOR    J.E.JOHNSON
15 ⟹ SEQLENGTH   374
16 ⟹ NALIGN       37
17 ⟹ NCHAIN        2 chain(s) in data set /data/dssp/1bmv.dssp
18 ⟹ CHAIN       2
19 ⟹ NOTATION: STRID1/STRID2: PDB identifiers of search structure and aligned
      NOTATION:      structure with chain identifier
      NOTATION: RMSD: positional root mean square deviation of superimposed CA atoms
      NOTATION:      in Angstroms
      NOTATION: LALI: total length of the aligned fragments.  The list of alignments
      NOTATION:      is sorted by length.
      NOTATION: LSEQ2: length of the entire chain of the aligned structure.
      NOTATION: %IDE: percentage of residue identity in the alignment
      NOTATION: REVERS: number of fragments matching in reversed chain direction
      NOTATION: PERMUT: number of topological permutations
      NOTATION: NFRAG: total number of aligned fragments
      NOTATION: TOPO: 'S' sequential connectivity of aligned fragments;
      NOTATION:       'N' non-sequential alignment
```

[10 NOTATION lines deleted]

```
20 ⟹ ## PROTEINS : PDB/chain identifiers and structural alignment statistics
21 ⟹   NR. STRID1 STRID2 RMSD LALI LSEQ2 %IDE REVERS PERMUT NFRAG TOPO  PROTEIN
        1: 1bmv-2 2mev-3  3.8  166   231    6      1      1    11 N     MENGO ENCEPHALOMYOCARDITIS VIRUS COAT PR
        2: 1bmv-2 1r09-2  3.2  154   255    6      1      1    11 N     RHINOVIRUS 14 (/HRV$14) COMPLEX WITH ANT
22 ⟹    3: 1bmv-2 1r09-3  3.0  151   236    9      0      0     8 S     RHINOVIRUS 14 (/HRV$14) COMPLEX WITH ANT
23 ⟹    4: 1bmv-2 2mev-1  4.1  147   268   14      0      1    16 N     MENGO ENCEPHALOMYOCARDITIS VIRUS COAT PR
        5: 1bmv-2 2tbv-B  3.8  144   284   11      1      1    17 N     TOMATO BUSHY STUNT VIRUS
```

[lines 6-37 deleted]

```
24 ⟹  ## ALIGNMENTS     1 -   25
25 ⟹   SeqNo PDBNo AA STRUCTURE BP1 BP2  ACC NOCC  . . . . : . . . . : . . . . : . . . . : .
```

[lines 1-25 deleted]

```
         25 3025 2 I  E     -    0   0    1    7  . F_ D_ . . . . Ce Ke . . He . . . . . . . . .
26 ⟹     26 3026 2 S  E    -P  170   OG   0    9  . Ye Te . . . . Ee Ae . . Ve . . . . . Ds . . .
         27 3027 2 Q  E    -P  169   OG  69   11  . Te Le . . . . Le Ve . . Ve . . . . . T_ Se Ks .
         28 3028 2 S  E    -P  168   OG   0   13  . L_ I_ . . . . Se Ce Ne . A_ . . . . . Ve Ve S_ .
         29 3029 2 K  E    -P  167   OG 125   13  . D_ Le . . . . Te Ve Ve . N_ . . . . . Ee Se A_ .
         30 3030 2 I  E    -P  166   OG   3   11  . S_ Ie . . . . Ee Le Te . . . . . . . Le Fe A_ .
```

[lines 31-374 deleted]

```
27 ⟹  ## ALIGNMENTS    26 -   37
       SeqNo PDBNo AA STRUCTURE BP1 BP2  ACC NOCC  . . . . : . . . . .: . . . . . : . . . . . : . . . . . :
```

[lines 1-24 deleted]

```
         25 3025 2 I  E     -    0   0    1    7  . . . . He . . De .. . .
         26 3026 2 S  E    -P  170   OG   0    9  . . . . Ke . . Re .. . Kb
         27 3027 2 Q  E    -P  169   OG  69   11  . . . . Ae . . Ne .. . A_
```

[lines 28-374 deleted]

```
28 ⟹  ## FRAGMENTS: ranges of superimposed residues
29 ⟹    NR. STRID1 STRID2                          RANGE1 <--> RANGE2
30 ⟹     1: 1bmv-2 2mev-3 103 ASP (3103) - 106 THR (3106) <-->  61 ALA (  61) -  58 VAL (  58)            (REVERS
          1: 1bmv-2 2mev-3 186 ASN (2004) - 197 MET (2015) <-->  43 ASP (  43) -  54 ILE (  54) (PERMUTED)
          1: 1bmv-2 2mev-3 198 GLY (2016) - 208 ARG (2026) <-->  63 PRO (  63) -  73 LYS (  73)
          1: 1bmv-2 2mev-3 210 VAL (2028) - 221 ALA (2039) <-->  77 LEU (  77) -  88 CYS (  88)
          1: 1bmv-2 2mev-3 229 MET (2047) - 232 PRO (2050) <-->  90 ALA (  90) -  93 PHE (  93)
          1: 1bmv-2 2mev-3 233 ASN (2051) - 271 LEU (2089) <-->  95 ALA (  95) - 133 ALA ( 133)
          1: 1bmv-2 2mev-3 276 ILE (2094) - 289 GLY (2107) <--> 140 ASP ( 140) - 153 GLY ( 153)
          1: 1bmv-2 2mev-3 291 ILE (2109) - 302 GLU (2120) <--> 154 LEU ( 154) - 165 ILE ( 165)
          1: 1bmv-2 2mev-3 305 LEU (2123) - 318 LEU (2136) <--> 168 THR ( 168) - 181 THR ( 181)
          1: 1bmv-2 2mev-3 321 ASP (2139) - 337 THR (2155) <--> 182 ASN ( 182) - 198 PRO ( 198)
          1: 1bmv-2 2mev-3 338 ILE (2156) - 364 SER (2182) <--> 201 CYS ( 201) - 227 PRO ( 227)
```

[179 lines deleted]

```
         17: 1bmv-2 3blm   210 VAL (2028) - 200 LEU (2018) <-->   6 LEU (  36) -  16 VAL (  46)            (REVERS
         17: 1bmv-2 3blm   198 GLY (2016) - 193 LEU (2011) <-->  17 TYR (  47) -  22 LYS (  52)            (REVERS
         17: 1bmv-2 3blm    37 VAL (3037) -  34 LYS (3034) <-->  96 TYR ( 129) -  99 ASN ( 132)            (REVERS
         17: 1bmv-2 3blm   104 ILE (3104) - 111 ASP (3111) <--> 157 LEU ( 190) - 164 GLY ( 197) (PERMUTED)
31 ⟹    17: 1bmv-2 3blm   301 GLN (2119) - 295 VAL (2113) <--> 194 LYS ( 227) - 200 ASP ( 233) (PERMUTED) (REVERS
         17: 1bmv-2 3blm    87 ALA (3087) -  82 GLU (3082) <--> 202 SER ( 235) - 207 THR ( 240)            (REVERS
         17: 1bmv-2 3blm   257 TYR (2075) - 245 ASP (2063) <--> 208 TYR ( 241) - 220 LYS ( 253) (PERMUTED) (REVERS
         17: 1bmv-2 3blm   353 ASN (2171) - 338 ILE (2156) <--> 221 GLY ( 254) - 236 ASN ( 269) (PERMUTED) (REVERS
         17: 1bmv-2 3blm   212 ARG (2030) - 217 ILE (2035) <--> 252 SER ( 285) - 257 PHE ( 290) (PERMUTED)
```

[190 lines deleted]

The section of the FSSP record marked by "## PROTEINS" (20) provides summary information on each protein used in the file. The headers in line (21) translate, as explained in the NOTATION lines.

- NR - protein number for cross-referencing sections with the FSSP file
- STRID1 - PDB ID, followed by chain specifiers, for the base protein with which others are to be compared
- STRID2 - PDB ID, followed by chain specifiers, for the protein being aligned to the base protein
- RMSD - positional root mean square deviation of superimposed α C atoms
- LALI - total length of the aligned fragments[29]
- LSEQ2 - total length of entire chain of aligned structure,
- %IDE - percentage of residue identity
- REVERS - number of fragments that match in the reverse direction of the chain
- PERMUT - the number of topological permutations
- NFRAG - total number of aligned fragments
- TOPO - sequential, S (22) , or nonsequential, N (23), connectivity of aligned fragments
- PROTEIN - the COMPND file of the PDB file for the structure being aligned to the base protein

Multiple alignments of the proteins follow "## ALIGNMENTS" as shown in line (24). All sequences are listed in vertical columns. Line (25) contains the headers.

- SeqNo refers to the sequential number of the residue in the base protein
- PDBNo gives the number of the residue as in the PDB file of the base protein
- AA gives the name of the base protein's amino acid as in the PDB file
- STRUCTURE gives the secondary structure as indicated in the corresponding DSSP file
- BP1 and BP2 indicate bridge partners as given by the corresponding DSSP file
- ACC represents the solvent accessibility as given by the DSSP file
- NOCC presents the number of aligned structures spanning this position

[29] The alignments are listed by length.

"....:...." represents a ruler marking columns. Sequences are presented according to the numbers assigned to them under the NR field of the "## PROTEIN" section of the FSSP file. Each residue is given as an upper-case letter, with the structure assigned to it, by DSSP, following as a lowercase letter. For example, on line (26), 2mev-3, in column 1, does not align at this position while 4sbv-A, in column 8, aligns to the amino acid cytosine of secondary structure "e".[30] When more sequences are aligned to the base protein than will fit across the file, a new "## ALIGNMENTS" section is created, as in (27).

The "## FRAGMENTS" section lists the matching fragment regions. Line (28) contains the following headers.

- NR - protein number for cross-referencing sections with the FSSP file
- STRID1 - PDB ID, followed by chain specifiers, for the base protein with which others are to be compared
- STRID2 - PDB ID, followed by chain specifiers, for the protein being aligned to the base protein
- RANGE1 and RANGE2 - the respective beginning and ending points for the matching fragment region including, for the base protein followed by the aligned protein.
  - the sequential residue numbers
  - the three-letter abbreviation for the amino acid
  - the PDB residue number

Therefore, line (29) reads as

> The first aligning fragment between 1bmv-2, the base protein, and 2mev-3 begins at the 1bmv-2 residue Asp with sequential number 103 and PDB number 3103 and the 2mev-3 residue Thr with sequential number 106 and PDB number 3106. This aligned fragment continues until 1bmv-2's sequentially numbered 61 Ala, PDB number 61 and 2mev-3's Val residue at sequential position 58, PDB position 58. The 2mev-3 fragment aligns in the "reverse" order respective to 1bmv-2, as indicated in the farthest right column.

Fragments are also allowed to align in a "permuted" alignment (30) or both a "permuted" and "reversed" alignment (31).

8.3 Obtaining FSSP

FSSP is available from EMBL by anonymous FTP:

[30] See explanation given in overview of DSSP for secondary structure abbreviations.

```
ftp ftp.embl-heidelberg.de          [or ftp 192.54.41.33]
cd /pub/databases/protein_extras/fssp
```

Questions can be sent

To: NET-HELP@EMBL-Heidelberg.DE
To: holm@embl-heidelberg.de
To: sander@embl-heidelberg.de
To: vriend@embl-heidelberg.de

The primary reference for FSSP is

> Holm, L., C. Ouzounis, C. Sander, G. Tuparev, and G. Vriend. 1992. A database of protein structure families with common folding motifs. *Protein Science*. 1: 1691–1698.

8.4 References

Document: 1bmv2_comp3D.fssp. September 1993. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/protein_extras/fssp*.

Document: 1bmv2_dali.fssp. October 1993. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/protein_extras/fssp*.

Document: 1bmv2_suppos.fssp. September 1993. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/protein_extras/fssp*.

Holm, L., C. Ouzounis, C. Sander, G. Tuparev, and G. Vriend. 1992. A database of protein structure families with common folding motifs. Prerelease of paper submitted to *Protein Science*.

Holm, L., C. Ouzounis, C. Sander, G. Tuparev, and G. Vriend. 1992. A database of protein structure families with common folding motifs. *Protein Science*. 1: 1691–1698.

9 GDB (Genome Database)

GDB is a relational database designed to support the efforts of groups mapping and sequencing the human genome. Its goal is to provide an environment for data editing and scanning for the Human Genome Initiative. EMBL-DB and GenBank were organized along very different lines, serving instead solely as data repositories for nucleotide sequences from all organisms. GDB contains the following four categories of data.

- map objects, primarily
  - genes
  - DNA segments
  - fragile sites

- map location
- genetic disease[31] and locus information
- citations

Data structures had to be created to store map information independent of the map type. Other types of data, such as polymorphism, mutational, homology, and probe information, are appended to these categories. The database is being extended to handle such objects as chromosomal breakpoints, restriction sites, meitotic crossover sites, and partial genetic maps. However, physical and genetic maps are not distinguished by storage or presentation. A GDB entry cannot be approved unless it has a source reference; thus, it is easy to trace information to its source. GDB was begun at the Laboratory for Applied Research in Academic Information of the Welch Medical Library, with most of its data derived by the editorial staff from the literature.

9.1 Explanation of a GDB Record

Because of the relational and interactive nature of GDB, a sample entry is not included here.

9.2 Obtaining GDB

Information on GDB is available by anonymous FTP:

```
ftp mendel.welch.jhu.edu          [or 128.220.59.42]
cd /gbd-5.0/data-dicts.ps
```

Questions can be sent

To: gdbhelp@welch.jhu.edu
To: davidk@welchdev.welch.jhu.edu

9.3 References

DATA DICTIONARY (Document: data_dict.ps.Z.) Version 5.0.1. Obtained from anonymous FTP to *mendel.welch.jhu.edu* in */gbd-5.0*.

Pearson, P. 1991. The genome data base (GDB) — a human gene mapping repository. *Nucleic Acids Research*. 19: 2237–2239.

[31] Upon entering the GDB environment, users have the choice of accessing map or disease information. After this initial choice, a user can travel freely between the two types of information. GDB is linked to MIM (Mendelian Inheritance of Man).

10 GenBank (Genetic Sequence Databank)

GenBank is similar in principle to the EMBL Nucleotide Sequence Database and serves as a repository for known genetic sequences. GenBank has close connections to the EMBL database and all information available in GenBank is also contained in the EMBL database. The EMBL database is becoming the preferred nucleotide database because its format is linked to other major databases at EMBL.[32]

10.1 Explanation of a GenBank Record

The first line of a GenBank entry is designated LOCUS (1)[33]. This line contains

- the locus name or entry name[34]
- the number of base pairs in the entry
- the division to which the record belongs[35]
- the record creation date or date of last major modification to the record

The DEFINITION field (2) provides a descriptive summary about the sequence. The ACCESSION line (3) includes first the primary accession number, here J01636, followed by secondary (often old) accession numbers that have been subsumed. The accession numbers provide stable access to the database and are separated by spaces. The KEYWORDS (4) provide another means of scanning

[32] This feature makes cross-referencing between EMBL database especially practical and allows similar software to be used to scan and extract information from diverse databases.

[33] See line 1 of the sample entry.

[34] The first three characters are usually taken from the organism name, with the last two letters drawn from other descriptive designations.

[35] There are 14 divisions within GenBank, based on taxonomy. These are

- PRI - primate sequences,
- ROD - rodent sequences,
- MAM - other mammalian sequences,
- VRT - other vertebrate sequences,
- INV - invertebrate sequences,
- PLN - plant, fungal, and algal sequences,
- BCT - bacterial sequences,
- RNA - structural RNA sequences,
- VRL - viral sequences,
- PHG - bacteriophage sequences,
- SYN - synthetic sequences,
- UNA - unannotated sequences,
- EST - expressed sequence tags (aka transcribed sequences fragments), and
- PAR - patent sequences.

and collecting various records. Each key word is separated with a ";" and the line terminates at the ".". The SOURCE field consists of an initial subfield, line (5), containing free-format information including an abbreviated form of the organism name and the molecule type. The second subfield, marked "ORGANISM" (6), gives the formal scientific name and the full taxonomic classification.

The REFERENCE section consists of five subsections. Each reference is first assigned a number (7). The range of bases (7) or other type of information (12) reported in this citation is given in "( )". The AUTHORS (8), TITLE (9), and JOURNAL (10) for the reference follow. The STANDARD line declares the degree of annotation and review:

- unannotated records - include citation and sequence only
- simple records - include organism name, coding regions, citations, and sequence
- full records - have all information
- automatic records - are subjected to automated checks only
- staff_entry records - have passed both automatic and annotator checks
- staff_review records - have in addition passed review by senior annotators or outside experts

The first COMMENT (13) in this record delineates the "sites" referred to in the sequence. Each site is associated with its reference source. For example, (14) specifies the type of site information that was available in reference 3. Free-format, paragraph-style COMMENTs follow (15). COMMENTs include cross-references to other sequence entries, documentation of changes to the LOCUS field, and various other free-form remarks. The FEATURES section forms a table that holds formated annotations. A finite number of specifiers are allowed in the FEATURES table, such as mutation (17), mRNA (18), misc(ellaneous)_signal (19), variation (20), and source (21). Under each of these are a finite number of subspecifiers such as "/organism" following source (21). The CDS (22) regions provide protein sequences coded for by the nucleic acid sequence.

The BASE COUNT line (23) provides the number of adenines, cytosines, guanines, and thymines in the sequence. The ORIGIN line (24) is used to give a pointer to the sequence start site. The sequence follows (25) in the 5' to 3' direction with the number of the first nucleotide of that line given in the far left column. The record is terminated with "//", (26).

10.2 Sample GenBank Entry

```
 1 ⟹ LOCUS       ECOLAC       7477 bp ds-DNA             BCT       05-MAY-1993
 2 ⟹ DEFINITION  E.coli lactose operon with lacI, lacZ, lacY and lacA genes.
 3 ⟹ ACCESSION   J01636 J01637 K01483 K01793
 4 ⟹ KEYWORDS    acetyltransferase; beta-D-galactosidase; galactosidase; lac operon;
                 lac repressor protein; lacA gene; lacI gene; lacY gene; lacZ gene;
                 lactose permease; mutagenesis; palindrome; promoter region;
                 thiogalactoside acetyltransferase.
 5 ⟹ SOURCE      Escherichia coli DNA; mRNA; clone lambda-h80dlac DNA; clone puk217;
                 pgm8 (see comment).
 6 ⟹   ORGANISM  Escherichia coli
                 Prokaryotae; Gram-negative facultatively anaerobic rods;
                 Enterobacteriaceae.
 7 ⟹ REFERENCE   1  (bases 1243 to 1266)
 8 ⟹   AUTHORS   Gilbert,W. and Maxam,A.
 9 ⟹   TITLE     The nucleotide sequence of the lac operator
10 ⟹   JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 70, 3581-3584 (1973)
11 ⟹   STANDARD  full automatic
     REFERENCE   2  (bases 1246 to 1308)
       AUTHORS   Maizels,N.M.
       TITLE     The nucleotide sequence of the lactose messenger ribonucleic acid
                 transcribed from the UV5 promoter mutant of Escherichia coli
       JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 70, 3585-3589 (1973)
       STANDARD  full automatic
12 ⟹ REFERENCE   3  (sites)
       AUTHORS   Gilbert,W., Maizels,N. and Maxam,A.
       TITLE     Sequences of controlling regions of the lactose operon
       JOURNAL   Cold Spring Harb. Symp. Quant. Biol. 38, 845-855 (1974)
       STANDARD  full automatic
```

[REFERENCE 4-37 deleted]

```
13 ⟹ COMMENT
14 ⟹             [3]  sites; UV5 mRNA transcripts and operator mutants.
                 [(in) Sund,H. and Blauer,G. (eds.);Protein-Ligand Interactions:
                 193-207;Walter de]  sites; operator mutational analysis.
                 [7]  sites; S1 and mung bean nuclease action on operator DNA.
                 [9]  sites; class I, II and III promoter mutant analysis.
                 [13]  sites; lacI mutant analysis.
```

[20 "sites;" comments deleted]

```
15 ⟹             [1] first reports a 27 bp operator(sites 1240-1266) with two-fold
                 symmetries; the operator has also been defined to be bases
                 1246-1266 or bases 1239-1273 [8]. [(in) Kjeldgaard,N.C. and Maaloe,
                 O.(eds);Control of ribosome synthesis: 138-143;A] explores the
                 ability of lac repressor protein to affect methylation of
                 operator DNA.  [8] argues that DNA on both sides of the 21
                 bp operator (bases 1246-1266) affects repressor binding
                 but that the sequences of this DNA are probably not
                 critical. [5] gives a larger sequence known as the
                 promoter-operator region for the wild-type, whereas [2] and
                 [26] give portions of this region for the mutant strain UV5.
```

[25 lines of comments deleted]

```
16 ⟹ FEATURES             Location/Qualifiers
17 ⟹      mutation        16
                          /note="c in wild-type; t in 'up' promoter mutant I-Q [11]"
18 ⟹      mRNA            51..1230
                          /note="lacI (repressor) mRNA; preferred in vivo 3' end
                          [12],[29]"
19 ⟹      misc_signal     1162..1199
                          /note="cap protein binding site"
20 ⟹      variation       1183..1186
```

[9 FEATURE entries deleted]

```
21 ⟹      source          1..7477
                          /organism="Escherichia coli"
22 ⟹      CDS             79..1161
                          /gene="lacI"
                          /note="lac repressor protein;  (gtg start codon)"
                          /codon_start=1
                          /translation="MKPVTLYDVAEYAGVSYQTVSRVVNQASHVSAKTREKVEAAMAE
                          LNYIPNRVAQQLAGKQSLLIGVATSSLALHAPSQIVAAIKSRADQLGASVVVSMVERS
                          GVEACKAAVHNLLAQRVSGLIINYPLDDQDAIAVEAACTNVPALFLDVSDQTPINSII
                          FSHEDGTRLGVEHLVALGHQQIALLAGPLSSVSARLRLAGWHKYLTRNQIQPIAEREG
                          DWSAMSGFQQTMQMLNEGIVPTAMLVANDQMALGAMRAITESGLRVGADISVVGYDDT
                          EDSSCYIPPSTTIKQDFRLLGQTSVDRLLQLSQGQAVKGNQLLPVSLVKRKTTLAPNT
                          QTASPRALADSLMQLARQVSRLESGQ"
```

[2 CDS FEATURE entries deleted - 34 lines]

```
          CDS             5727..6338
                          /gene="lacA"
                          /note="thiogalactoside acetyltransferase;  (ttg start
                          codon)"
                          /codon_start=1
                          /translation="LNMPMTERIRAGKLFTDMCEGLPEKRLRGKTLMYEFNHSHPSEV
                          EKRESLIKEMFATVGENAWVEPPVYFSYGSNIHIGRNFYANFNLTIVDDYTVTIGDNV
                          LIAPNVTLSVTGHPVHHELRKNGEMYSFPITIGNNVWIGSHVVINPGVTIGDNSVIGA
                          GSIVTKDIPPNVVAAGVPCRVIREINDRDKHYYFKDYKVESSV"
23 ⟹ BASE COUNT     1739 a   1991 c   2004 g   1743 t
24 ⟹ ORIGIN      HindII site [Nature 274, 762-765 (1978)].
25 ⟹        1 gacaccatcg aatggcgcaa aacctttcgc ggtatggcat gatagcgccc ggaagagagt
           61 caattcaggg tggtgaatgt gaaaccagta acgttatacg atgtcgcaga gtatgccggt
          121 gtctcttatc agaccgtttc ccgcgtggtg aaccaggcca gccacgtttc tgcgaaaacg
          181 cgggaaaaag tggaagcggc gatggcggag ctgaattaca ttcccaaccg cgtggcacaa
          241 caactggcgg gcaaacagtc gttgctgatt ggcgttgcca cctccagtct ggccctgcac
          301 gcgccgtcgc aaattgtcgc ggcgattaaa tctcgcgccg atcaactggg tgccagcgtg
          361 gtggtgtcga tggtagaacg aagcggcgtc gaagcctgta aagcggcggt gcacaatctt
          421 ctcgcgcaac gcgtcagtgg gctgatcatt aactatccgc tggatgacca ggatgccatt
          481 gctgtggaag ctgcctgcac taatgttccg gcgttatttc ttgatgtctc tgaccagaca
          541 cccatcaaca gtattatttt ctcccatgaa gacggtacgc gactgggcgt ggagcatctg
          601 gtcgcattgg gtcaccagca aatcgcgctg ttagcgggcc cattaagttc tgtctcggcg
```

[sequence deleted to base pair 7477]

```
26 ⟹ //
```

10.3 Obtaining GenBank

On September 30, 1992, the repository of GenBank was moved from Los Alamos to NCBI (National Center for Biotechnology Information). GenBank is available from NCBI by anonymous FTP:

ftp ncbi.nlm.nih.gov

Records may also be obtained via NETSERVER (e-mail)

To: retrieve@ncbi.nlm.nih.gov
DATALIB genbank
BEGIN (as message body)
the_accession_number_of_record [ACC]

Information can also be obtained by e-mail:

To: retrieve@ncbi.nlm.nih.gov
help (as message body)

or FTP:

ftp ncbi.nlm.nih.gov

Questions can be sent

To: retrieve-help@ncbi.nlm.nih.gov
To: gcr@t10.Lanl.GOV

10.4 References

Burks, C., et al. 1991. GenBank. *Nucleic Acids Research*. 19: 2221–2225.

Document: ECOLAC [J01636]. May 1993. Obtained from e-mail to *retrieve@ncbi.nlm.nih.gov* with message *DATALIB genbank BEGIN J01636 [ACC].*

Document: GBREL.TXT. Release 79.0, October 1993. Obtained from anonymous FTP to *ncbi.nlm.nih.gov* in */genbank*.

Document: LAMBDA [V00636]. December 1992. Obtained from e-mail to *retrieve@ncbi.nlm.nih.gov* with message *DATALIB genbank BEGIN J02459 [ACC].*

Document: README. Release 79.0, October 1993. Obtained from anonymous FTP to *ncbi.nlm.nih.gov* in */genbank*.

Document: RETRIEVE E-Mail Server. October, 1993. Obtained from e-mail to *retrieve@ncbi.nlm.nih.gov* with message *help*.

11 HSSP (Homology-derived Secondary Structures of Proteins)

For each PDB file, and thus for each protein of known 3D structure, the HSSP database contains a file with all SWISS-PROT sequence homologues properly aligned to the PDB protein.[36] The database thus serves as a repository of aligned primary sequence families and of implied secondary and tertiary structures. Each HSSP file merges 1D, 2D, and 3D structural data. The HSSP database represents the output of running the program HSSP against all PDB entries.

Secondary structures are carried over directly from the PDB entry, and tertiary structures can be modeled by using special 3D modeling software to fit the sequences of unknown structure to the structural template of the PDB protein. Thus, for each PDB file, xxxx.ent, there exists an HSSP file, xxxx.hssp, which contains the PDB protein (with repeating identical chains removed), derived 2D structure, solvent accessibility, all SWISS-PROT sequences defined homologous (based on a homology threshold curve), and a measure of the sequence variability at each position. HSSP was created by Chris Sander and Reinhard Schneider.

11.1 Explanation of an HSSP Record

HSSP files are divided into four sections: HEADERS, PROTEINS, ALIGNMENTS, and SEQUENCE PROFILE. The headers section begins with the specification of the version of HSSP used to generate the file (1).[37] The PDB protein on which the file is based is identified by PDBID (2) with the date of creation following under DATE (3). SEQBASE (4) names the sequence database from which the aligned sequences were obtained, commonly either EMBL/SWISS-PROT or PIR/NBRF. PARAMETER lines (5) delineate parameters used in the HSSP program relating to homology determination and alignment.[38] The THRESHOLD line (6) describes the homology threshold curve used. Information on the primary reference for HSSP (7), sources for information and answer about HSSP (8), and availability of HSSP (9) is also included. The header section continues with information about the PDB entry, including the PDB HEADER (10), COMPND (11), SOURCE (12), and AUTHOR (13) lines.[39] Information derived from HSSP is also included, such as the sequence length (14), number of distinct chains (15), and the number of aligned sequences (16). Notation lines provide

[36] The average number of SWISS-PROT sequences aligned to each PDB structure is 130.

[37] See line 1 of the sample entry.

[38]
- smin = lowest similarity between amino acids,
- smax = highest similarity,
- gap-open = penalty for opening a gap,
- gap-elongation = penalty for extending a gap,
- maxdel = maximum deletion length, and
- other information.

[39] The complete PDB lines were not included, only information contained therein.

valuable information to reading an HSSP file because they explain some of the notation used. The NOTATION lines are found in every record.

The second section of an HSSP file, PROTEINS, gives the pairwise alignment data for each protein to the structurally homologous PDB protein. It is marked by "## PROTEINS" (18). The headers (19), as explained in the NOTATION lines, identify the following:

- NR - line identifier
- ID - the SWISS-PROT entry name for the aligned, homologous protein
- STRID - the PDB identifier, for proteins of known 3D structure
- %IDE - the percentage of the alignment's residues identity
- %WSIM - a weighted measure of the alignment's similarity
- IFIR - the first residue of the PDB protein in the alignment
- ILAS - the last residue of the PDB protein in the alignment
- JFIR - the first residue of the sequence aligned to the PDB protein
- JLAS - the last residue of the sequence aligned to the PDB protein
- LAI - the length of the alignment not counting insertions or deletions
- NGAP - the number of insertions or deletions
- LGAP - the total length of the insertions and deletions excluded under the LAI
- LSEQ2 - the total length of the protein sequence aligned to the PDB protein
- ACCNUM - the SWISS-PROT accession number
- PROTEIN - the SWISS-PROT DE (description) line

Line (20) begins the data, and line (21) aligns a sequence of known 3D structure, as indicated by the STRID entry.

"## ALIGNMENTS" (22) initiates the third section of an HSSP entry. This section provides family alignment details residue by residue. Sequences are listed vertically, with the leftmost entry at the top. The notation used in the column specifiers of line (23) is explained in the NOTATION lines.

- SeqNo indicates the sequential residue number of the PDB protein as obtained from the corresponding DSSP file
- PDBNo indicates the residue's number followed by the name as obtained from the PDB file
- AA specifies the amino acid in the SWISS-PROT entry being aligned to the PDB protein, in the one-letter code

- STRUCTURE summarizes the secondary structure[40] as in the DSSP file
- BP1 and BP2 dictate β bridge partners as obtained from the DSSP file
- ACC gives the surface area of the residue that is solvated, acquired from the DSSP file
- NOCC assigns the number of aligned sequences at this position
- VAR indicates the sequence variability derived from the NALIGN alignments

"....:....1....:" provides a "ruler". The ruler can be used to identify the columns to the proteins being aligned to the PDB entry according to the numbers assigned to them under the NR lines (19). In these columns, lower case letters indicate an insertion at this point in the aligned sequence, while "...." indicates deletions. Lower-case letters in the PDB sequence, under the header "PDBNo", indicate disulfide bridges. Line (24) begins the text. The "!" in line (25) indicate a chain termination, with the next chain following immediately.

The sequence profile, marked "## SEQUENCE PROFILE" (26), provides a count of each amino acid in each position obtained by counting the residues in that position for all aligned sequences, including the PDB sequence. (Numbers are scaled such that 100 means only that type of amino acid is always found in this position.) Line (27) specifies the headers, with text following on line (28). The headers are explained in the NOTATION lines.

- SeqNo - residue number as specified in the DSSP file for the PDB protein
- PDBNo - residue number and identity as given in the PDB file
- the amino acid possibilities under the one letter codes[41]
- NOCC - number of sequences with amino acids in this position
- NDEL - number of sequences containing deletions, relative to the PDB protein, at this position
- NINS - number of sequences containing insertions, relative to the PDB protein, at this position
- ENTROPY - measure of sequence variability at this position
- RELENT - relative entropy scaled from 0 to 100
- WEIGHT - conserved weight derived from clustering and used in the alignment program

The final section of an HSSP file begins with "## INSERTION" (29) and identifies by lower-case letters the insertions that were indicated in the ## ALIGNMENTS section. This is a new addition to the HSSP database and is not currently explained in the NOTATION lines. Line (30) provides the following column headers:

[40] For example: hydrogen bonding patterns of turns and helices, geometrical bend, chirality, and β ladder and sheet formation.

[41] **Please note that many of these columns were deleted from the sample entry because of space limitations.**

- AliNo for the number of the alignment assigned in the ## PROTEINS section under NR
- IPOS for the last residue position preceding the insertion in the PDB protein
- JPOS for the last residue position preceding the insertion in the sequence aligned to the PDB protein
- Len for the total length of the insertion
- Sequence for the actual amino acid sequence of the insertion, contained in the lower-case brackets given in section ## ALIGNMENTS

For this record, the first insertion is at NR 9, as shown in line (31). The HSSP record ends with "//", line (32).

11.2 Obtaining HSSP

HSSP is available from EMBL by anonymous FTP:

```
ftp ftp.embl-heidelberg.de          [or ftp 192.54.41.33]
cd /pub/databases/hssp
```

Questions can be sent

To: NET-HELP@EMBL-Heidelberg.DE

The primary reference for HSSP is

> Sander, C., and R. Schneider. 1991. Database of homology-derived protein structures. *Proteins*. 9: 56–68.

11.3 References

Document: 9xim.hssp. October 1993. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/hssp*.

Document: INDEX. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/hssp*.

Document: README. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/hssp*.

Sander, C., and R. Schneider. 1991. Databases of homology-derived protein structures and the structural meaning of sequence alignment. *PROTEINS: Structure, Function, and Genetics*. 9: 56–68.

Schneider, R. and Sander, C. HSSP: A database of structure-sequence alignments. HSSP release 1.0. Obtained from e-mail to *NETSERV@EMBL-Heidelberg.DE* with message *GET PROTEIN-DATA:HSSP.DOC*.

11.4 Sample HSSP Entry

```
 1 ⟹ HSSP       HOMOLOGY DERIVED SECONDARY STRUCTURE OF PROTEINS , VERSION 1.0 1991
 2 ⟹ PDBID      9xim
 3 ⟹ DATE       file generated on  6-Oct-93
 4 ⟹ SEQBASE    RELEASE 26.0 OF EMBL/SWISS-PROT WITH  31808 SEQUENCES
 5 ⟹ PARAMETER  SMIN: -0.5  SMAX:  1.0
      PARAMETER  gap-open:  3.0 gap-elongation:  0.1
      PARAMETER  conservation weights
      PARAMETER  no insertions/deletions in secondary structure allowed
      PARAMETER  alignments sorted according to:DISTANCE
 6 ⟹ THRESHOLD  according to t(L)=(290.15 * L ** -0.562) +  5
 7 ⟹ REFERENCE  Sander C., Schneider R. : Database of homology-derived protein structures. Proteins, Proteins,
 8 ⟹ CONTACT    e-mail (INTERNET) Schneider@EMBL-Heidelberg.DE or Sander@EMBL-Heidelberg.DE / phone +49-6221-3
 9 ⟹ AVAILABLE  Free academic use. Commercial users must apply for license.
      AVAILABLE  No inclusion in other databanks without permission.
10 ⟹ HEADER     ISOMERASE(INTRAMOLECULAR OXIDOREDUCTSE)
11 ⟹ COMPND     D-*XYLOSE ISOMERASE (E.C.5.3.1.5) MUTANT WITH GLU 186
12 ⟹ SOURCE     (ACTINOPLANES $MISSOURIENSIS) /E186Q$ MUTANT GENE EXPRESSED
13 ⟹ AUTHOR     J.JANIN
14 ⟹ SEQLENGTH  1567
15 ⟹ NCHAIN        4 chain(s) in 9xim data set
16 ⟹ NALIGN       40
17 ⟹ NOTATION : ID: EMBL/SWISSPROT identifier of the aligned (homologous) protein
      NOTATION : STRID: if the 3-D structure of the aligned protein is known, then STRID is the Protein Data Ba
      NOTATION : from the database reference or DR-line of the EMBL/SWISSPROT entry
      NOTATION : %IDE: percentage of residue identity of the alignment
      NOTATION : %SIM (%WSIM):  (weighted) similarity of the alignment
      NOTATION : IFIR/ILAS: first and last residue of the alignment in the test sequence
      NOTATION : JFIR/JLAS: first and last residue of the alignment in the alignend protein
      NOTATION : LALI: length of the alignment excluding insertions and deletions
      NOTATION : NGAP: number of insertions and deletions in the alignment
      NOTATION : LGAP: total length of all insertions and deletions
      NOTATION : LSEQ2: length of the entire sequence of the aligned protein
      NOTATION : ACCNUM: SwissProt accession number
      NOTATION : PROTEIN: one-line description of aligned protein
      NOTATION : SeqNo,PDBNo,AA,STRUCTURE,BP1,BP2,ACC: sequential and PDB residue numbers, amino acid (lower ca
      NOTATION : structure, bridge partners, solvent exposure as in DSSP (Kabsch and Sander, Biopolymers 22, 25
      NOTATION : VAR: sequence variability on a scale of 0-100 as derived from the NALIGN alignments
```

[10 NOTATION lines deleted]

```
18 ⟹ ## PROTEINS : EMBL/SWISSPROT identifier and alignment statistics
19 ⟹   NR.    ID         STRID   %IDE %WSIM IFIR ILAS JFIR JLAS LALI NGAP LGAP LSEQ2 ACCNUM     PROTEIN
20 ⟹     1 : xyla_actmi          1.00  1.00 1179 1570    3  394  392    0    0  394  P12851     XYLOSE ISOMERASE (EC 5.3.1.5).
          2 : xyla_actmi          1.00  1.00  394  785    3  394  392    0    0  394  P12851     XYLOSE ISOMERA
          3 : xyla_actmi          1.00  1.00    1  392    3  394  392    0    0  394  P12851     XYLOSE ISOMERA
          4 : xyla_actmi          1.00  1.00  787 1177    4  394  391    0    0  394  P12851     XYLOSE ISOMERA
```

[lines 5-16 deleted]

```
21 ⟹    17 : xyla_strru 1XIS    0.68  0.80 1180 1570    3  386  384    2    7  387  P24300     XYLOSE ISOMERA
         18 : xyla_strru 1XIS    0.68  0.80  395  785    3  386  384    2    7  387  P24300     XYLOSE ISOMERA
         19 : xyla_stral 6XIA    0.68  0.80 1180 1570    3  386  384    2    7  390  P24299     XYLOSE ISOMERA
         20 : xyla_stral 6XIA    0.68  0.81  787 1177    3  386  384    2    7  390  P24299     XYLOSE ISOMERA
```

[lines 21-40 deleted]

```
22 ⟹ ## ALIGNMENTS    1 -   40
23 ⟹  SeqNo  PDBNo AA STRUCTURE BP1 BP2  ACC NOCC  VAR  ....:....1....:....2....:....3....:....4....:....5....
24 ⟹       1   3 A V              0   0  167    3   11    V   L              V
           2   4 A Q        -     0   0   73    9    9    Q   Q   QQ   Q     Q  Q    Q    E
           3   5 A A        -     0   0   25    9   21    A   A   PP   P     P  P    P    P
           4   6 A T    >   -     0   0   78    9   15    T   T   TT   T     T  T    T    K
           5   7 A R  G >  S+     0   0  161    9   17    R   P   PP   P     P  P    P    P
           6   8 A E  G 3  S+     0   0   80    9   14    E   D   EE   E     A  E    E    E
           7   9 A D  G <  S-     0   0   24    9   12    D   D   DD   D     D  D    D    H
           8  10 A K    <   +     0   0   59    9   23    K   K   KR   R     H  R    K    R
           9  11 A F  E     +a  287   0A    1    9    0    F   F   FF   F     F  F    F    F
          10  12 A S  E     -ab 288  48A    0    9   25    S   S   TT   T     T  T    T    T
```

[SeqNo from 11-392 deleted]

```
25 ⟹     393      !  !           0   0    0    0    0
         394   3 B V              0   0  118    3   11   V     L              V
         395   4 B Q        -     0   0   73    9    9   Q     Q  Q   Q  Q    Q  Q  Q      E
         396   5 B A        -     0   0   27    9   18   A     A  P   P  P    P  P  P      P
```

[SeqNo from 397-1570 deleted]

```
26 ⟹ ## SEQUENCE PROFILE AND ENTROPY
27 ⟹  SeqNo PDBNo   V   L   I   [columns  C   H   R   K   Q   E   N   D  NOCC NDEL NINS ENTROPY RELENT WEIGHT
28 ⟹     1    3 A  67  33   0     M-T    0   0   0   0   0   0   0   0    3    0    0  0.637     58  1.02
         2    4 A   0   0   0   deleted]  0   0   0   0  89  11   0   0    9    0    0  0.349     16  1.09
         3    5 A   0   0   0             0   0   0   0   0   0   0   0    9    0    0  0.530     24  0.93
```

[SeqNo 4-392 deleted]

```
       393          0   0   0             0   0   0   0   0   0   0   0    0    0    0  0.000      0  1.00
       394    3 B  67  33   0             0   0   0   0   0   0   0   0    3    0    0  0.637     58  1.02
       395    4 B   0   0   0             0   0   0   0  89  11   0   0    9    0    0  0.349     16  1.09
       396    5 B   0   0   0             0   0   0   0   0   0   0   0    9    0    0  0.530     24  0.93
```

[SeqNo 397-1570 deleted]

```
29 ⟹ ## INSERTION LIST
30 ⟹  AliNo  IPOS  JPOS   Len Sequence
31 ⟹      9  1551   368     1 rAa
         10  1159   369     1 aAw
         11   766   368     1 rAa
         12   373   368     1 rAa
         21  1061   278     1 gFp
         22   277   279     1 fPn
         23   670   279     1 fPn
         24  1455   279     1 fPn
         25   373   368     1 rAa
         26   766   368     1 rAa
         27  1551   368     1 rAa
32 ⟹ //
```

12 LiMB (Listing of Molecular Biological Databases)

LiMB is a database of available molecular biological databases that is produced at Los Alamos National Laboratory. It is formatted as a simple directory, but its role will be extended to serve as a centralized information resource.

12.1 Explanation of a LiMB Record

LiMB comes in two formats: *limbshort.txt* and *limb.txt*.

12.1.1 LiMBshort

LiMBshort is ideal for quick lookup of a limited amount of information on a databases. The first line (1)[42] of a LiMBshort entry gives the entry name, while the line (2) contains the name of the actual database. A short summary of the charter or intent of the database follows (3). Line (4), dat.pri, lists the primary data items available in the databases, with secondary data items listed under data.sec (5). The record is terminated by "//" (6).

12.1.2 LiMB

Full LiMB entries provide much more information about each database. The initial line (1) is again the entry name, but the next line provides a stable accession number that will always be associated with that database. The history of the LiMB entry, such as creation date and revision dates, composes the next section (3). The status line (4) indicates how the information was obtained (most LiMB information was obtained through questionnaires), followed by the name (5), address (6), telephone number (7), fax number (8), and Internet address (9) for the respondent who provided the information about the database. Information about where to direct general inquiries, including name (10), address (11), telephone number (12), fax number (13), and Internet address (14), follows. Information about sources of contributing data is listed in a similar format (15) followed by the appropriate source for acquiring data (16).

The "name.now line" (17) indicates the official name of the database and is followed by "name.alt" (18) for alternative, official or unofficial, names or "name.obs" (19) for obsolete or incorrect names for the database that are commonly encountered. The source line (20) explains the source of information contained in the database; funding for the database is cited under funding (21). Information on formal collaborators for collection, "collab.in", and distribution, "collab.ou", of the data may follow the source line preceding the funding line. Recent publications describing the database are found under citations (22), and a short description of the intent is located under charter (23).

Other databases that are cross referenced by the database are listed under crossref (24). The primary

[42] See line 1 of the sample entry.

data items (25) and the secondary data items (26) are followed by information on the type of hardware (27) necessary for maintaining the database. Other information about the computer aspects of the database includes operating system for the hardware (28), software system used to maintain the data (29), programming language used for the software system (30), software distributed with the database (31), format used for distributing flat text files (32), and limitations on access to the data (33).

Information on the frequency of database updates (34) is followed by information about contributions.

- Can contributions be made to database on-line? (35)
- Can contributions be made to database on magnetic tape? (36)
- Can contributions be made to database on floppy disk? (37)
- Can contributions be made to database by electronic mail? (38)
- Can contributions be made to database on hardcopy? (39)

Next, modes of distribution are delineated.

- Is database distributed on-line? (40)
- Is database distributed on magnetic tape? (41)
- Is database distributed on floppy disk? (42)
- Is database distributed by electronic mail? (43)
- Is database distributed on hardcopy? (44)

The final information contained in the record tells the number of bytes contained in the database (45), the number of bytes contained by primary data items (46), the number of entries in the database (47), and any other information or comments (48). Records are terminated by "///" (49).

12.2 Sample LiMB Entries

12.2.1 Sample LiMBshort Entries

```
1 ⟹ entry      LIMB
2 ⟹ name.now   LiMB
3 ⟹ charter    The goal of LiMB is to provide the scientific community with a
               comprehensive overview of databases relevant to molecular biology
               and related data sets.
4 ⟹ data.pri   databases [molecular biology]
5 ⟹ data.sec   database access information; database contribution information;
               database characteristics; literature citations [molecular
               biology]; database maintenance [hardware]; database maintenance
               [software]
6 ⟹ ///
    entry      PDB
    name.now   Protein Data Bank
    charter    PDB seeks comprehensive coverage of bibliographic, atomic
               coordinate and crystallographic structure factor data for
               biological macromolecules.
    data.pri   atomic coordinates [biomacromolecule]
    data.sec   functional features [biomacromolecule]; literature citations
               [biomacromolecule]; molecular properties [biomacromolecule]
    ///
```

12.2.2 Sample LiMB Entry

```
 1 ⟹ entry       LIMB
 2 ⟹ number      10012
 3 ⟹ history     fm 04/07/87 initial entry
                 jl 12/10/87 update
                 jl 02/08/88 update using memo from cb
                 jl 12/06/89 updated for release 1.2
                 gk 06/21/90 updated entry from returned questionnaire
 4 ⟹ status      questionnaire
 5 ⟹ res.nam     Dr. Christian Burks
 6 ⟹ res.add     T-10, MS K710
                 Los Alamos National Laboratory
                 Los Alamos, NM 87545
                 U.S.A.
 7 ⟹ res.tel     -
 8 ⟹ res.fax     (505) 665-3493
 9 ⟹ res.net     cb@intron.lanl.gov
10 ⟹ gen.nam     LiMB Database
11 ⟹ gen.add     T-10, MS K710
                 Los Alamos National Laboratory
                 Los Alamos, NM 87545
                 U.S.A.
12 ⟹ gen.tel     (505) 667-9455
13 ⟹ gen.fax     -
14 ⟹ gen.net     limb@life.lanl.gov
15 ⟹ con.nam     LiMB Database
     con.add     T-10, MS K710
                 Los Alamos National Laboratory
                 Los Alamos, NM 87545
                 U.S.A.
     con.tel     (505) 667-9455
     con.fax     -
     con.net     limb@life.lanl.gov
16 ⟹ acc.nam     LiMB Database
     acc.add     T-10, MS K710
                 Los Alamos National Laboratory
                 Los Alamos, NM 87545
                 U.S.A.
     acc.tel     (505) 667-9455
     acc.fax     -
     acc.net     limb@life.lanl.gov
17 ⟹ name.now    LiMB
18 ⟹ nam.alt     Listing of Molecular Biology Databases
19 ⟹ nam.obs     -
20 ⟹ source      standard questionnaires completed by the staffs of other
                 databases
21 ⟹ funding     LANL
22 ⟹ citation    [1] Burks, C., Lawton, J., Bell, G.  (1988) The LiMB database.
                 <Science> 241, 888.
                 [2] Lawton, J., Burks, C., Martinez, F.  (1989) Overview of the
                 LiMB database.  <Nucleic Acids Research> 17, 5885-5899.
```

```
23 ⟹ charter     The goal of LiMB is to provide the scientific community with a
                 comprehensive overview of databases relevant to molecular biology
                 and related data sets.
24 ⟹ crossref    -
25 ⟹ data.pri    databases [molecular biology]
26 ⟹ data.sec    database access information; database contribution information;
                 database characteristics; literature citations [molecular
                 biology]; database maintenance [hardware]; database maintenance
                 [software]
27 ⟹ hardware    Sun 4/60
28 ⟹ op.sys      UNIX
29 ⟹ dbms        SYBASE
30 ⟹ language    -
31 ⟹ software    -
32 ⟹ format      flat text file: line type record
33 ⟹ access      no limitations
34 ⟹ updates     1 per year
35 ⟹ con.onl     yes
36 ⟹ con.mag     no
37 ⟹ con.flp     yes
38 ⟹ con.elm     yes
39 ⟹ con.pap     yes
40 ⟹ acc.onl     no
41 ⟹ acc.mag     no
42 ⟹ acc.flp     yes
43 ⟹ acc.elm     yes
44 ⟹ acc.pap     yes
45 ⟹ byt.all     0.23 Mb
46 ⟹ byt.pri     0.00 Mb
47 ⟹ ent.pri     120
48 ⟹ comment     numbers given are for release 2.0
49 ⟹ ///
```

12.3 Obtaining LiMB

LiMB is available from EMBL by anonymous FTP:

```
ftp ftp.embl-heidelberg.de          [or ftp 192.54.41.33]
cd /pub/databases/limb
```

Information can also be obtained by e-mail:

```
To: NETSERV@EMBL-Heidelberg.DE
help LIMB                            (as message body)
```

Questions can be sent to

```
To: NET-HELP@EMBL-Heidelberg.DE
To: gwr@life.lanl.gov
```

The primary reference for LiMB is

Keen, G. et al. 1992. *Math. Comput. Modelling.* 16: 93–101.

12.4 References

Burks, C. Document: limb.doc. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/limb.*

Document: limb.help. July 1992. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/help.*

Lawton, J., F. Martinez, and C. Burks. 1989. Overview of the LiMB database. *Nucleic Acid Research.* 17: 5885-5899.

LiMB Database. (Document: limb.txt.) Release 3.0. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in
/pub/databases/limb.

LiMBshort Database. (Document: limbshort.txt.) Release 3.0. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/limb.*

13 PDB (Protein Databank)

Probably the most prominent protein database is the PDB of Brookhaven National Laboratory. The PDB is the largest repository for 3D protein structures determined by X-ray crystallography or nuclear magnetic resonance (NMR) and contains examples of all known unique protein families. The April 1993 release of the PDB contains 1110 fully annotated atomic coordinates entries. The number of entries in the PDB increases dramatically with each new release.

13.1 Explanation of a PDB Record

PDB records have the now-familiar Fortran format. The HEADER line (1)[43] includes the functional classification of the macromolecule, the date the record was entered, and the ID code (of one number and three-letters) for the record. (The last two columns are contained on all lines; their functionality will be highlighted shortly.) If the record has been subsumed or made obsolete by a newer entry, an OBSLTE line follows the HEADER pointing the user to the new entry. Line (2) is the COMPouND line giving the name of the molecule and identifying information, including the EC number. The SOURCE (e.g., species, organ, tissue, mutant) is highlighted in (3), while the AUTHORs are given in line (4).

Whenever a PDB entry is revised or changed, the changes are documented in the REVisionDATe (5) lines. The first number, after the REVDAT identifier, numbers the revision; note that the most recent revision is listed first. The revision date is listed next, followed by a specifier and a number that indicates the type of revision.

- 3 for modifications affecting coordinates or transformations
- 2 for modifications to the CONECT records
- 1 for all other types of modifications, such as typographical corrections
- 0 for the initial entry

Then a short list of the type of lines modified is provided. For example, the revisions of (5) included revisions to SEQRES, TURN, and ATOM entries.

The specifier given in the REVDAT following the date of the revision is the record's ID followed by a letter of the alphabet. Notice that this identifier replaces the "standard" identifier of the record ID in the next-to-the-last column of each line of the record on lines that were modified on this revision data. Thus, the new specifier for "REVDAT 9" (5) is 8LYZH. The lines of this record modified under "REVDAT 9" can be traced by noting which lines have 8LYZH as the next to last column entry. The last column of each line "counts" the line; thus the order in which lines were entered can be traced. Using the last two columns together with the REVDAT information, one can tell which record lines are primary or revisions, and for the revisions, the researcher can determine, for example, "this was the fifth line revised on 14-JUL-86."

If the record under study had superseded other older records, the REVDAT would have been followed by a SPRSDE line indicating which old PDB records had been subsumed. Typically, after the REVDAT lines, the primary citation for the entry is given on the JouRNL lines. Line (6) gives the AUTHors. Next, the TITLe line (7) of the article is given, which may occupy more than one line, as shown on line (8) with the TITL 2 specifier is given. Further lines give the name of the journal (9) in which the article was published, specified as REFerence, and the REFerenceNumber (10) of the journal.

[43] See line 1 of the sample entry.

REMARK lines are next. Remarks come in types including citations, free-format comments, and correction data. The remarks are each numbered following the REMARK specifier. Line (11) initiates the first remark, REMARK 1. The remarks of type 1 are, here, the references the PDB entry authors used to develop their structure. Line (12) clearly states this, since REMARK of the *1* st class to be presented is a REFERENCE and this begins reference 1. The first line (13) of this REFERENCE 1 is the AUTH line for the citation, followed by the TITL (14), REF (15), and REFN (16) parallel in format to the lines used to specify the journal citation for the PDB record itself (6–10). Lines (17), (18), (19), and (20) specify remarks giving the resolution, refinement, free-format comments, and correction information, respectively. (The deleted REMARK lines held correction information.) Each of these types of remarks receives its own remark number.

The next section of the PDB entry contains information about the actual macromolecule, generally a protein. Primary structure, secondary structure, and tertiary structure information are each given in turn. Line (21) begins the SEQuenceRESidues establishing the primary structure. Each SEQRES line is numbered in the second column followed by the number of residues in the third column. The amino acid sequence, using the three letter abbreviations, is given, followed by the standard last two columns. The secondary structures of $\alpha-$helices, β sheets, hairpin turns, and disulfide bonds are next described.

HELIX lines, for example (22), contain the following in respective columns.

- the helix number
- the chain identifier
- the first amino acid of the helix
- the number of the first amino acid
- the last amino acid of the helix
- the number of the last amino acid
- the class of the helix

Ten helix classes are recongized.

- 1 - Right-handed α (default)
- 2 - Right-handed ω
- 3 - Right-handed π
- 4 - Right-handed γ
- 5 - Right-handed 3_{10}
- 6 - Left-handed α
- 7 - Left-handed ω

- 8 - Left-handed γ
- 9 - 2_7 ribbon/helix
- 10 - Polyproline

β sheets are designated in SHEET lines such as (23). The SHEET lines give the following information.

- strand number
- sheet identifier
- number of strands in the sheet
- first amino acid in the sheet
- number of the first amino acid
- last amino acid in the sheet
- number of the last amino acid
- sense of this strand with respect to the previous strand[44]
- atom of the amino acid of strand N that bonds to strand N-1
- atom of the amino acid of strand N-1 that is bonded to by strand N

Therefore, starting at line (23), information on the first sheet can be paraphrased

> The first strand of the first sheet, which has 2 strands, begins with the first amino acid, Lys, and continues to the third amino acid, Phe. The second-strand of this sheet starts with the amino acid 38, Phe, and ends with amino acid 40, Thr. It is oriented antiparallel to the first strand. The nitrogen of the Thr in position 40 of the second strand hydrogen bonds to the oxygen in the Lys, position 1, of the first strand.

Note that with a β barrel, a closed sheet, would be indicated by the last strand and the first strand of the sheet being identical.

[44] Sense indicated by

- 0 sense of strand 1,
- 1 strand N parallel to strand N-1, or
- -1 strand N antiparallel to strand N-1.

.

13.2 Sample PDB Entry

```
 1 ⟹ HEADER    HYDROLASE (O-GLYCOSYL)                  16-SEP-77   8LYZ      8LYZ   3
 2 ⟹ COMPND    LYSOZYME (E.C.3.2.1.17) IODINE-INACTIVATED                    8LYZ   4
 3 ⟹ SOURCE    HEN (GALLUS GALLUS) EGG WHITE                                 8LYZ   5
 4 ⟹ AUTHOR    C.R.BEDDELL,C.C.F.BLAKE,S.J.OATLEY                            8LYZ   6
 5 ⟹ REVDAT   9   14-JUL-86 8LYZH   3       SEQRES TURN   ATOM               8LYZH  1
     REVDAT   8   22-OCT-84 8LYZG   1       SHEET                            8LYZG  1
     REVDAT   7   27-JAN-84 8LYZF   1       REMARK                           8LYZF  1
     REVDAT   6   30-SEP-83 8LYZE   1       REVDAT                           8LYZE  1
     REVDAT   5   01-MAR-82 8LYZD   1       REMARK                           8LYZE  2
     REVDAT   4   21-MAY-81 8LYZC   3       ATOM                             8LYZE  3
     REVDAT   3   25-MAY-78 8LYZB   1       SEQRES                           8LYZE  4
     REVDAT   2   01-NOV-77 8LYZA   1       SSBOND                           8LYZE  5
     REVDAT   1   24-OCT-77 8LYZ    0                                        8LYZE  6
 6 ⟹ JRNL        AUTH   C.R.BEDDELL,C.C.F.BLAKE,S.J.OATLEY                   8LYZ   7
 7 ⟹ JRNL        TITL   AN X-RAY STUDY OF THE STRUCTURE AND BINDING          8LYZ   8
 8 ⟹ JRNL        TITL 2 PROPERTIES OF IODINE-INACTIVATED LYSOZYME            8LYZ   9
 9 ⟹ JRNL        REF    J.MOL.BIOL.                   V.  97   643 1975      8LYZ  10
10 ⟹ JRNL        REFN   ASTM JMOBAK  UK ISSN 0022-2836                 070   8LYZ  11
11 ⟹ REMARK   1                                                              8LYZ  12
12 ⟹ REMARK   1 REFERENCE 1                                                  8LYZ  13
13 ⟹ REMARK   1  AUTH   R.DIAMOND                                            8LYZ  14
14 ⟹ REMARK   1  TITL   REAL-SPACE REFINEMENT OF THE STRUCTURE OF HEN        8LYZ  15
     REMARK   1  TITL 2 EGG-WHITE LYSOZYME                                   8LYZ  16
15 ⟹ REMARK   1  REF    J.MOL.BIOL.                   V.  82   371 1974      8LYZ  17
16 ⟹ REMARK   1  REFN   ASTM JMOBAK  UK ISSN 0022-2836                 070   8LYZ  18
     REMARK   1 REFERENCE 2                                                  8LYZ  19
     REMARK   1  AUTH   D.C.PHILLIPS                                         8LYZ  20
     REMARK   1  TITL   CRYSTALLOGRAPHIC STUDIES OF LYSOZYME AND ITS         8LYZ  21
     REMARK   1  TITL 2 INTERACTIONS WITH INHIBITORS AND SUBSTRATES          8LYZ  22
     REMARK   1  EDIT   E.F.OSSERMAN,R.F.CANFIELD,S.BEYCHOK                  8LYZ  23
     REMARK   1  REF    LYSOZYME                                  9 1974      8LYZ  24
     REMARK   1  PUBL   ACADEMIC PRESS,NEW YORK                              8LYZ  25
     REMARK   1  REFN                   ISBN 0-12-528950-2             977   8LYZD  1
```

[REF 3-12 deleted]

```
     REMARK   2                                                              8LYZ  95
17 ⟹ REMARK   2 RESOLUTION. 2.5 ANGSTROMS.                                   8LYZ  96
     REMARK   3                                                              8LYZ  97
18 ⟹ REMARK   3 REFINEMENT. BY THE MODEL-BUILDING AND REAL-SPACE             8LYZ  98
     REMARK   3  REFINEMENT PROCEDURES OF R. DIAMOND. REFER TO REFERENCE 1   8LYZ  99
     REMARK   3  ABOVE AND REMARK 4 BELOW.                                   8LYZ 100
     REMARK   4                                                              8LYZ 101
19 ⟹ REMARK   4 THE ONLY SIGNIFICANT FEATURES ON THE DIFFERENCE MAP ARE IN   8LYZ 102
     REMARK   4 THE REGION OF GLU 35 AND TRP 108 SIDE CHAINS - THE OE2 ATOM 8LYZ 103
     REMARK   4 OF GLU 35 FORMS A COVALENT BOND WITH THE CD1 ATOM OF TRP     8LYZ 104
     REMARK   4 108.  AN INTERACTIVE COMPUTER GRAPHICS SYSTEM WAS USED TO    8LYZ 105
     REMARK   4 MANIPULATE THESE SIDE CHAINS IN THE RS5D COORDINATE SET OF   8LYZ 106
     REMARK   4 R. DIAMOND (1974), ENTRY 2LYZ IN THE PROTEIN DATA BANK, SO   8LYZ 107
     REMARK   4 THAT A FIT TO THE ELECTRON DENSITY MAP WAS OBTAINED.         8LYZ 108
     REMARK   4 THESE COORDINATES, THEREFORE, ARE IDENTICAL TO THE RS5D      8LYZ 109
     REMARK   4 ENTRY APART FROM PORTIONS OF THESE TWO SIDE CHAINS.          8LYZ 110
     REMARK   5                                                              8LYZA  1
20 ⟹ REMARK   5 CORRECTION.                                                  8LYZA  2
     REMARK   5  ADD SSBOND RECORDS.                                         8LYZA  3
     REMARK   5  01-NOV-77.                                                  8LYZA  4
```

[REMARKS 6-12 deleted]

```
21 ⟹  SEQRES   1    129  LYS VAL PHE GLY ARG CYS GLU LEU ALA ALA ALA MET LYS  8LYZ 111
       SEQRES   2    129  ARG HIS GLY LEU ASP ASN TYR ARG GLY TYR SER LEU GLY  8LYZ 112
       SEQRES   3    129  ASN TRP VAL CYS ALA ALA LYS PHE GLU SER ASN PHE ASN  8LYZ 113
       SEQRES   4    129  THR GLN ALA THR ASN ARG ASN THR ASP GLY SER THR ASP  8LYZB  3
       SEQRES   5    129  TYR GLY ILE LEU GLN ILE ASN SER ARG TRP TRP CYS ASN  8LYZB  4
       SEQRES   6    129  ASP GLY ARG THR PRO GLY SER ARG ASN LEU CYS ASN ILE  8LYZB  5
       SEQRES   7    129  PRO CYS SER ALA LEU LEU SER SER ASP ILE THR ALA SER  8LYZ 117
       SEQRES   8    129  VAL ASN CYS ALA LYS LYS ILE VAL SER ASP GLY ASN GLY  8LYZH  7
       SEQRES   9    129  MET ASN ALA TRP VAL ALA TRP ARG ASN ARG CYS LYS GLY  8LYZ 119
       SEQRES  10    129  THR ASP VAL GLN ALA TRP ILE ARG GLY CYS ARG LEU      8LYZ 120
22 ⟹  HELIX    1   A ARG      5  HIS     15  1                                 8LYZ 121
       HELIX    2   B LEU     25  GLU     35  1                                 8LYZ 122
       HELIX    3   C CYS     80  LEU     84  5                                 8LYZ 123
       HELIX    4   D THR     89  LYS     96  1                                 8LYZ 124
23 ⟹  SHEET    1  S1 2 LYS     1  PHE     3  0                                 8LYZ 125
       SHEET    2  S1 2 PHE    38  THR    40 -1  N  THR    40   O  LYS     1   8LYZG  5
       SHEET    1  S2 3 ALA    42  ASN    46  0                                 8LYZ 127
       SHEET    2  S2 3 SER    50  GLY    54 -1  N  ASN    46   O  SER    50   8LYZ 128
       SHEET    3  S2 3 GLN    57  SER    60 -1  N  TYR    53   O  ILE    58   8LYZ 129
24 ⟹  TURN     1  T1 LYS    13  GLY    16     TYPE I.                         8LYZ 130
       TURN     2  T2 LEU    17  TYR    20     NEARLY TYPE II CONFORMATION.    8LYZ 131
       TURN     3  T3 ASN    19  GLY    22     NEARLY TYPE II CONFORMATION.    8LYZ 132
       TURN     4  T4 TYR    20  TYR    23     NEARLY TYPE II CONFORMATION.    8LYZ 133
       TURN     5  T5 GLY    54  GLN    57     TYPE I,BETW STRNDS 2,3 SHT S2.  8LYZ 134
       TURN     6  T6 ASN    59  TRP    62     NEARLY TYPE I CONFORMATION.     8LYZ 135
       TURN     7  T7 THR    69  SER    72     NEARLY TYPE I CONFORMATION.     8LYZ 136
       TURN     8  T8 ASN    74  ASN    77     TYPE I.                         8LYZ 137
       TURN     9  T9 ASN   103  ASN   106     TYPE I.                         8LYZH  8
       TURN    10 T10 CYS   115  THR   118     TYPE II (IMPERFECT).            8LYZ 139
       TURN    11 T11 ILE   124  CYS   127     TYPE II (IMPERFECT).            8LYZ 140
25 ⟹  SSBOND   1 CYS      6    CYS    127                                      8LYZA  5
       SSBOND   2 CYS     30    CYS    115                                      8LYZA  6
       SSBOND   3 CYS     64    CYS     80                                      8LYZA  7
       SSBOND   4 CYS     76    CYS     94                                      8LYZA  8
26 ⟹  CRYST1   79.100   79.100   37.900  90.00  90.00  90.00 P 43 21 2     8  8LYZ 141
27 ⟹  ORIGX1      1.000000 0.000000  0.000000       0.000000                   8LYZ 142
       ORIGX2      0.000000 1.000000  0.000000       0.000000                   8LYZ 143
       ORIGX3      0.000000 0.000000  1.000000       0.000000                   8LYZ 144
28 ⟹  SCALE1       .012642 0.000000  0.000000       0.000000                   8LYZ 145
       SCALE2      0.000000  .012642  0.000000       0.000000                   8LYZ 146
       SCALE3      0.000000 0.000000   .026385       0.000000                   8LYZ 147
29 ⟹  ATOM      1  N   LYS     1       3.240  10.040  10.380  1.00  0.00      8LYZ 148
       ATOM      2  CA  LYS     1       2.390  10.410   9.250  1.00  0.00      8LYZ 149
       ATOM      3  C   LYS     1       2.460  11.920   9.100  1.00  0.00      8LYZ 150
       ATOM      4  O   LYS     1       2.580  12.670  10.100  1.00  0.00      8LYZ 151
       ATOM      5  CB  LYS     1        .950   9.960   9.490  1.00  0.00      8LYZ 152
       ATOM      6  CG  LYS     1       -.050  10.450   8.450  1.00  0.00      8LYZ 153
       ATOM      7  CD  LYS     1      -1.470  10.060   8.820  1.00  0.00      8LYZ 154
       ATOM      8  CE  LYS     1      -2.350   9.920   7.590  1.00  0.00      8LYZ 155
       ATOM      9  NZ  LYS     1      -3.680   9.380   7.960  1.00  0.00      8LYZ 156
       ATOM     10  N   VAL     2       2.390  12.350   7.850  1.00  0.00      8LYZ 157
```

[ATOM 11-998 deleted]

```
        ATOM    999  CD1 LEU   129      -12.970  22.550   8.090  1.00  0.00      8LYZ1146
        ATOM   1000  CD2 LEU   129      -13.000  20.080   8.010  1.00  0.00      8LYZ1147
30 ⟹   TER    1002      LEU   129                                                8LYZ1148
31 ⟹   CONECT   48   47  981                                                     8LYZ1149
        CONECT  238  237  889                                                     8LYZ1150
        CONECT  277  275  820                                                     8LYZ1151
        CONECT  513  512  630                                                     8LYZ1152
        CONECT  601  600  724                                                     8LYZ1153
        CONECT  630  513  629                                                     8LYZ1154
        CONECT  724  601  723                                                     8LYZ1155
        CONECT  820  277  819  822                                                8LYZ1156
        CONECT  889  238  888                                                     8LYZ1157
        CONECT  981   48  980                                                     8LYZ1158
32 ⟹   MASTER      124    0    0    4    5   11    0    6 1000    1   10   10    8LYZH 17
33 ⟹   END                                                                       8LYZ1160
```

Hairpin turns, also known as β and γ bends, which occur in the structure outside of helices, are identified in the TURN lines. TURN lines, such as (24), contain the following information.

- turn number
- turn identifier
- first residue of the turn
- number of the first amino acid
- last residue of the turn
- number of the last amino acid
- a comment on the type of the turn

Disulfide bonds are specified similarly under the SSBOND lines (25).

- disulfide bond number
- first amino acid of the bond
- number of the first amino acid
- last amino acid of the bond
- number of the last amino acid

If there were further sites, SITE lines could be included at this point; they would include the groups creating the site. The nature of the site would be fully explained in the REMARK section.

Information regarding coordinate transformations follows the secondary structure information. CRYST1 (26) delineates the unit cell parameters in the order a, b, c, α, β, and γ followed by the "space group symbol" and Z. There are three ORIGX lines: ORIGX1 (27), ORIGX2, and ORIGX3. The ORIGX lines form a transformation matrix between the originally submitted coordinates and the orthogonal coordinates contained in the file. The SCALE lines, of SCALE1 (28) , SCALE2, and SCALE3, also form a transformation matrix from the orthogonal coordinates to the fractional crystallographic coordinates.

The coordinates for each atom in each amino acid are established in the ATOM and HETATM lines. The format is logically parallel to that used earlier in specifying secondary structure. For example, line (29) tells that the first atom, a nitrogen, of the first amino acid, Lys, is located at the coordinates X=3.240, Y=10.040, Z=10.380, Occupancy=1, Temperature factor=0, and no footnote citation is given. ATOM lines are used for "standard" groups, generally one of the twenty unmodified amino acids, while the HETATM delineator is given to "nonstandard" groups such as ATP, Coenzyme A, or Heme. Brookhaven's PDB organizers have tried to develop a uniform nomenclature to deal with nonstandard prosthetic groups and cofactors. ATOM, or HETATM, lines are given for every atom in the macromolecule. TER (30) indicates chain terminators. Whereas all previously mentioned line specifiers (e.g., HEADER) must come in a specific order, ATOM, HETATM, and TER lines, within this section of the PDB record, come in their appropriate order. Thus TER lines occur where chains are terminated and may be followed by the ATOM specifiers for the next chain, and ATOM/HETATM headers may occur in whatever order is appropriate.

CONECT records (31) give the connectivity information of covalent bonds, hydrogen bonds, and salt bridges that occur within the molecule. The MASTER (32) line provides checksums of the number of records in the file. It provides the total number different record line types.

- REMARK
- FTNOTE
- HET
- HELIX
- SHEET
- TURN
- SITE
- transformation records, ORIGX+SCALE+MATRIX
- atomic coordinate records, ATOM+HETATM
- TER
- CONECT
- SEQRES records

The record ends with the END line (33).

Please note: Most, but not all, possible line specifiers for a PDB record were explained here.

13.3 Obtaining PDB

PDB is available from Brookhaven National Laboratory by anonymous FTP:

```
ftp pdb.pdb.bnl.gov
cd /fullrelease/uncompressed_files         [or /fullrelease/compressed_files]
```

Information can also be obtained by e-mail:

To: fileserv@pdb.pdb.bnl.gov
send info your_e-mail_address (as message body)

To: NETSERV@EMBL-Heidelberg.DE
GET PROTEINDATA:PDBFORMAT.PS (as message body)

or by gopher:

oeder@bnl.gov

Questions can be sent

To: pdb@bnl.gov

13.4 References

Protein Data Bank. July 1993 release. Obtained from anonymous FTP to *pdb.pdb.bnl.gov* to */full-release/compressed_files*.

Protein Data Bank Atomic Coordinate and Bibliographic Entry Format Description. Obtained from e-mail to *NETSERV@EMBL-Heidelberg.DE* with message *PROTEINDATA:PDBFORMAT.PS*.

Protein Data Bank Quarterly Newsletter. 1992. Number 61, July 1992. Obtained from e-mail to *NETSERV@EMBL-Heidelberg.DE* with message *GET PROTEINDATA:NEWSLETTER.DOC*

Protein Data Bank Quarterly Newsletter. 1993. Number 64, April 1993. Obtained from e-mail to *fileserv@pdb.pdb.bnl.gov* with message *send info _your_e-mail_address_*.

14 PIR (Protein Information Resource)

PIR is a collection of sequences originally designed to study the evolutionary relationships between proteins. PIR is therefore organized on the idea of protein superfamilies. Superfamilies consist of homologous proteins that appear evolutionarily related on the basis of amino acid sequence. The superfamily design was initiated, however, before it became apparent that large proteins are often composed of domains from different evolutionary origins, obtained through fusion, gain and loss of exons, shifted reading frames, incorporation of foreign DNA, or rearrangement of native DNA. To accommodate such "unevolutionary" homologies, PIR assigns sequences that are mostly unrelated evolutionarily to separate superfamilies, even though they may contain related domains. Margaret Dayhoff initiated the PIR database.

14.1 Explanation of a PIR Record

PIR records begin in line (1)[45] with a ‘‘>’’ followed by the sequence type: P1 for proteins or F1 for fragments, and then the entry code. The second line (2) provides a title, including the name of the protein and the source separated by "-". The sequence, in the standard one-letter code, immediately begins (3) without any specifiers. All of these lines are required.

Line (4) presents alternative names for the protein, and line (5) specifies functions and activities of the protein, including all EC numbers. Comment lines begin with "C;", for example (6) and (7), and a subject delineator precedes the free format text comment. "Species" (6), "Accession" (7), "Superfamily", "Keywords", and "Comment" are examples of acceptable subject delineators.

The reference information is introduced by "R;" as in line (8) which lists the names of the authors followed by the citation without an initiator (9). Further comments about the reference, such as title (10), reference number (11), information about the content of the reference (12) (13) (14), and cross-reference information for the reference (15), are introduced with "A;".

Further comments of the form "C;" or "A;" about the record itself may then follow, such as (16–22). The primary difference between "C;" and "A;" comments is their format. "C;" comments have a subject specifier before their free-format text. "A;" comments are not required to have one of the finite tags but are entirely free format.

The feature line begins with "F;" (23) and describes functional regions of the sequence. Following the "F;" the field contains the residue numbers of the beginning and end of the site. Single-residue sites are distinguished by having only the single residue listed. Further examples of "F;" lines are

- F;24/Active site: Asp,
- F;35,97,82/Binding site: lipid,
- F;15-25,34-97/Protein: basic protease inhibitor, and
- F;254-289/Region: transmembrane.

[45] See line 1 of the sample entry.

14.2 Sample PIR Entry

In "NBRF" Physical File Format

Internal VAX/VMS Distribution Format

Format of the Flat Files **.REF* and **.SEQ*

```
 1 ⟹ >P1;DEHUHS
 2 ⟹ 3beta-hydroxy-Delta5-steroid dehydrogenase multifunctional protein I - human
 3 ⟹  M T G W S C L V T G A G G F L G Q R I I R L L V K E K E L K
       E I R V L D K A F G P E L R E E F S K L Q N K T K L T V L E
       G D I L D E P F L K R A C Q D V S V I I H T A C I I D V F G
       V T H R E S I M N V N V K G T Q L L L E A C V Q A S V P V F
       I Y T S S I E V A G P N S Y K E I I Q N G H E E E P L E N T
       W P A P Y P H S K K L A E K A V L A A N G W N L K N G G T L
       Y T C A L R P M Y I Y G E G S R F L S A S I N E A L N N N G
       I L S S V G K F S T V N P V Y V G N V A W A H I L A L R A L
       Q D P K K A P S I R G Q F Y Y I S D D T P H Q S Y D N L N Y
       T L S K E F G L R L D S R W S F P L S L M Y W I G F L L E I
       V S F L L R P I Y T Y R P P F N R H I V T L S N S V F T F S
       Y K K A Q R D L A Y K P L Y S W E E A K Q K T V E W V G S L
       V D R H K E T L K S K T Q *
 4 ⟹ N;Alternate names: progesterone reductase
 5 ⟹ N;Contains: 3beta-hydroxy-Delta5-steroid dehydrogenase (EC 1.1.1.145); steroid Delta-isomerase (EC 5.3.3.1)
 6 ⟹ C;Species: Homo sapiens (man)
 7 ⟹ C;Accession: A36551; A32746; A23657
 8 ⟹ R;Lorence, M.C., Corbin, C.J., Kamimura, N., Mahendroo, M.S., and Mason, J.I.
 9 ⟹ Mol. Endocrinol. 4, 1850-1855, 1990 (Liver)
10 ⟹ A;Title: Structural analysis of the gene encoding human 3beta-hydroxysteroid dehydrogenase/Delta(5->4)-isom
11 ⟹ A;Reference number: A36551; MUID:91186993
12 ⟹ A;Accession: A36551
13 ⟹ A;Molecule type: DNA
14 ⟹ A;Residues: 1-373 <LOR>
15 ⟹ A;Cross-reference: GB:M63395
     R;Luu The, V., Lachance, Y., Labrie, C., Leblanc, G., Thomas, J.L., Strickler, R.C., and Labrie, F.
     Mol. Endocrinol. 3, 1310-1312, 1989
     A;Title: Full length cDNA structure and deduced amino acid sequence of human 3beta-hydroxy-5-ene steroid de
     A;Reference number: A32746; MUID:89384668
     A;Accession: A32746
     A;Molecule type: mRNA
     A;Residues: 1-373 <LUU>
     A;Cross-reference: GB:M27137
```

[Ref 3 deleted]

```
16 ⟹ C;Comment: This complex enzyme catalyzes the conversion of Delta5-ene-3beta-hydroxy steroid precursors into
     Delta4-3-ketosteroids; its function is crucial for the biosynthesis of all classes of hormonal steroids.
     Deficiency in this enzyme system causes severe reduction of steroid formation by the adrenals and gonads
     and is usually lethal in early life.
17 ⟹ C;Comment: Two isozymes, I and II, are found in human. This isozyme I is almost exclusively in the placenta
     and skin and also predominately in mammary gland tissue.
18 ⟹ C;Map position: 1p13.1
19 ⟹ C;Gene name: GDB:HSD3B1
20 ⟹ C;Introns: 49/1, 104/1
21 ⟹ C;Superfamily: 3beta-hydroxy-Delta5-steroid dehydrogenase
22 ⟹ C;Keywords: intramolecular oxidoreductase; isomerase; NAD; oxidoreductase; steroid biosynthesis
23 ⟹ F;2-373/Protein: 3beta-hydroxy-Delta5-steroid dehydrogenase multifunctional protein (experimental) <MAT>
```

14.3 Obtaining PIR

Information about PIR can be obtained by e-mail:

To: PIRSERVER@NBRF.Georgetown.Edu
HELP (as message body)

Questions can be sent

To: PIRMAIL@GUNBRF.BITNET
To: MARZEC@NBRF.Georgetown.Edu

14.4 References

Barker, W., D. George, L. Hunt, and J. Garavelli. 1991. The PIR protein sequence database. *Nucleic Acids Research*. 19: 2231–2236.

Document: e-mail message from C. Marzec. October 1993. Obtained from e-mail to *MARZEC@NBRF.Georgetown.Edu*.

Document: PIR Request. NBRF-PIR MAILSERVER version 1.49. Obtained from e-mail to *PIRSERVER@NBRF.Georgetown.Edu* with message *HELP*.

George, D. and W. Barker. 1989. User's guide for the protein sequence query program of the Protein Identification Resource (PIR)*. Document PSQM-0589. May 1989. Obtained from e-mail to *MARZEC@NBRF.Georgetown.Edu*.

Protein Sequence Database of PIR-International. PIR Document PRDBFS-1292: Databases File Structure and Format Specification. December 1992. Obtained from e-mail to *MARZEC@NBRF.Georgetown.Edu*.

Protein Sequence Database: Sequence File Format. PIR Document PRFILE-1292. December 1992. Obtained from e-mail to *MARZEC@NBRF.Georgetown.Edu*.

15 PKCDD (Protein Kinase Catalytic Domain Database)

Many biological databases are created by one individual for personal use. These generally contain specific, often limited, information. A good example of a limited database is PKCDD, which contains the amino acid sequences of the catalytic domains of all members of the protein kinase family whose sequence was available as of the last release date. PKCDD was produced by S. Hanks, A. Quinn, and T. Hunter.

PKCDD is contained in two files. All reference information is contained in *PKCDD.REF*, while *PKCDD.FASTA* contains the actual sequences of the catalytic domain in FastA/Pearson format. If equivalent sequences were reported from more than one vertebrate species, only one is given in PKCDD. The format of each file is simple; but since the information is separate, it is more cumbersome to manipulate.

15.1 Explanation of a PKCDD Record

The first line (1)[46] of the *PKCDD.FASTA* contains a ">" marker, an identifying name for the sequence, the number of bases in the sequence, and a checksum identifier. Line (2) begins the sequence of the catalytic domain. By then looking at *PKCDD.REF*, we discover that sequence "cAPK-a" is a human cAMP-dependent protein kinase, while ECAPK is a *C. elegans* cAMP-dependent protein kinase. The files *PKCDD.FASTA* and *PKCDD.REF* are not intrinsically linked, but they are linked by the identifying name. An entry in *PKCDD.REF* (3) also gives the primary reference for the sequence and the GenBank accession number.

15.2 Obtaining PKCDD

PKCDD is available from EMBL by anonymous FTP:

```
ftp ftp.embl-heidelberg.de        [or ftp 192.54.41.33]
cd /pub/databases/pkcdd
```

Information can also be obtained by e-mail:

```
To: NETSERV@EMBL-Heidelberg.DE
help PKCDD                          (as message body)
```

Questions can be sent

```
To: NET-HELP@EMBL-Heidelberg.DE
To: quinn@salk-sc2.sdsc.edu
```

15.3 References

Document: README. July 1993. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/pkcdd*.

Document: PKCDD.FASTA. April 1993. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/pkcdd*.

Hanks, S., and A. Quinn. 1993. Protein Kinase Catalytic Domain Database references. Release April 4, 1993. Obtained from e-mail to *NETSERV@EMBL-Heidelberg.DE* with message *GET PKCDD:PKCDD.REF*.

[46] See line 1 of the sample entry.

15.4 Sample PKCDD Entries

15.4.1 Excerpt of *PKCDD.FASTA*

```
1 ⟹ >CAPK-ALPHA, 255 bases, 39C8DED5 checksum.
2 ⟹ FERIKTLGTGSFGRVMLVKHKETGNHYAMKILDKQKVVKLKQIEHTLNEK
    RILQAVNFPFLVKLEFSFKDNSNLYMVMEYVPGGEMFSHLRRIGRFSEPH
    ARFYAAQIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVK
    GRTWTLCGTPEYLAPEIILSKGYNKAVDWWALGVLIYEMAAGYPPFFADQ
    PIQIYEKIVSGKVRFPSHFSSDLKDLLRNLLQVDLTKRFGNLKNGVNDIK
    NHKWF
    >ECAPK, 308 bases, B8B83D5A checksum.
    FDRIKTLGTGSFGRVMLVKHKQSGNYYAMKILDKQKVVKLKQVEHTLNEK
    RILQAIDFPFLVNMTFSLKDNSNLYMVLEFISGGEMFSHLRRIGRFSEPH
    SRFYAAQIVLAFEYLHSLDLIYRDLKPENLLIDSTGYLKVTDFGFAKRVK
    GRTWTLCGTPEYLAPEIILSKGYNKAVDWWALGVLIYEMAAGYPPFFADQ
    PIQIYEKIVSGKVKFPSHFSNELKDLLKNLLQVDLTKRYGNLKNGVADIK
    NHKWFGSTDWIAIYQKKIEAPFLPKCRGPGDASNFDDYEEEPLRISGTEK
    CAKEFAEF
```

15.4.2 Excerpt of *PKCDD.REF*

```
3 ⟹ cAPK-a:   Human cAMP-dependent protein kinase catalytic subunit, alpha-form
                    (Maldonado and Hanks (1988) Nucl. Acids Res. 16, 8189-8190)
                    Genbank Accession P17612

    ECAPK:    C.elegans cAMP-dependent protein kinase catalytic subunit C
                    (Gross et al (1990) J Biol Chem 265:6896-6907)
                    Genbank Accession   M37114 J05289 M35424
```

16 ProSite (Dictionary of Protein Sites and Patterns)

The ProSite database contains biologically significant patterns and sites that can be easily used to classify an unknown protein into a known family of proteins. Instead of relying on overall sequence alignment to identify structure, ProSite allows for local sequence alignment matches. The patterns are short, allowing for specificity. The goal is to obtain a core pattern that will detect all the proteins in a certain family, or subfamily, without matching proteins outside of this group. ProSite, which is composed of two files *prosite.dat* and *prosite.doc*, was originally created by Amos Bairoch.

16.1 Explanation of a ProSite Record

Prosite.dat begins with an ID line (1)[47] which includes the entry name and the entry type (pattern, rule, or matrix). The ACcession number (2) provides stable access to the data, since entry names can be changed. The DaTe line (3) includes three parts of information: the date the record was created, the date of the last modification to *prosite.dat*, and the date of the last modification to *prosite.doc*. The DEscription line (4) is a free-format short description of the subject of the record.

Line PAttern, (5) and (6), defines the ProSite pattern using the standard IUPAC one-letter amino acid codes. Further, information is also included. holds:

- x is any amino acid
- "[ ]" enclose acceptable ambiguities[48]
- "{ }" enclose unacceptable ambiguities[49]
- < indicates pattern is restricted to the N-terminus
- > indicates pattern is restricted to the C-terminus
- a period ends the pattern

If the record was a rule or a matrix, RU or MA lines would occur, respectively.

[47] See line 1 of the sample entry.

[48] For example, [ALT] = Ala or Leu or Thr is acceptable in this position.

[49] For example, {AM} = any amino acid in this position except Ala or Met.

16.2 Sample ProSite Entries

16.2.1 Excerpt of *prosite.dat*

```
 1 ⟹ ID   LEUCINE_ZIPPER; PATTERN.
 2 ⟹ AC   PS00029;
 3 ⟹ DT   APR-1990 (CREATED); APR-1990 (DATA UPDATE); APR-1990 (INFO UPDATE).
 4 ⟹ DE   Leucine zipper pattern.
 5 ⟹ PA   L-x(6)-L-x(6)-L-x(6)-L.
      CC   /TAXO-RANGE=??E?V;
      CC   /SKIP-FLAG=TRUE;
      DO   PDOC00029;
      //
      ID   HOMEOBOX; PATTERN.
      AC   PS00027;
      DT   APR-1990 (CREATED); JUN-1992 (DATA UPDATE); DEC-1992 (INFO UPDATE).
      DE   'Homeobox' domain signature.
 6 ⟹ PA   [LIVMFY]-x(5)-[LIVM]-x(4)-[IV]-[RKQ]-x-W-x(8)-[RK].
 7 ⟹ NR   /RELEASE=24,28154;
      NR   /TOTAL=187(187); /POSITIVE=175(175); /UNKNOWN=0(0); /FALSE_POS=12(12);
      NR   /FALSE_NEG=9(9);
 8 ⟹ CC   /TAXO-RANGE=??E??; /MAX-REPEAT=1;
 9 ⟹ DR   P02833, HMAN_DROME, T; P07548, HMDF_DROME, T; P20009, HMDL_DROME, T;
      DR   P18488, HMES_DROME, T; P10035, HMH2_DROME, T; P28468, HOX1_HALRO, T;
      DR   P17208, BRN3_MOUSE, P; P20266, BRN3_RAT  , P; P20912, HM16_XENLA, P;
      DR   P20823, HNFA_HUMAN, N; P22361, HNFA_MOUSE, N; P15257, HNFA_RAT  , N;
      DR   P22197, ALF_ARATH , F; P08704, CDGT_KLEPN, F; P80064, HPPD_PSESP, F;
```

[70+ DR lines deleted (most were T)]

```
10 ⟹ 3D   1HDD;
11 ⟹ DO   PDOC00027;
12 ⟹ //
```

16.2.2 Excerpt of *prosite.doc*

```
13 ⟹ {PDOC00027}
14 ⟹ {PS00027; HOMEOBOX}
15 ⟹ {BEGIN}
      *******************************
16 ⟹ * 'Homeobox' domain signature *
      *******************************

17 ⟹ The 'homeobox' is a protein domain of 60 amino acids [1 to 5] which
      was first identified in a number of Drosophila homeotic and
      segmentation proteins.  It has since been found to be extremely well
      conserved in many other animals, including vertebrates.  This domain
      binds DNA through a helix-turn-helix type of structure.  Proteins
      which contain an homeobox domain are likely to play an important role
      in development.  Most of these proteins are known to be sequence
      specific DNA-binding transcription factors.  The homeobox domain has
      also been found to be very similar to a region of the yeast mating
      type proteins, which are sequence specific DNA-binding proteins that
      act as master switches in yeast differentiation by controlling gene
      expression in a cell type-specific fashion.
```

A schematic representation of the homeobox domain is shown below. The helix-turn-helix region is shown by the symbols 'H '(for helix), and 't' (for turn).

```
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxHHHHHHHHtttHHHHHHHHHxxxxxxxxxx
|        |         |         |         |         |         |
1       10        20        30        40        50        60
```

The pattern we developed to detect homeobox sequences is 24 residues long and spans positions 34 to 57 of the homeobox domain.

18 ⟹ -Consensus pattern: [LIVMFY]-x(5)-[LIVM]-x(4)-[IV]-[RKQ]-x-W-x(8)-[RK]
-Sequences known to belong to this class detected by the pattern: ALL, except for Drosophila cut and om(1d), for maize knotted-1, and for liver specific transcription factor LF-B1 (HNF1-alpha and -beta) which has a very atypical homeobox domain.
-Other sequence(s) detected in SWISS-PROT: 12 other proteins.
-Note: a majority of the proteins which contain an homeobox domain can be classified, on the basis of their sequence characteristics, in 3 subfamilies: antennapedia, engrailed, and paired. We have developed specific patterns that characterize these subfamilies of proteins. We also have a specific pattern for the homeobox proteins that contain a "POU" domain.
-Last update: December 1992 / Text revised.

19 ⟹ [1] Gehring W.J.
Trends Biochem. Sci. 17:277-280(1992).
[2] Gehring W.J.
Science 236:1245-1252(1987).
[3] Scott M.P., Tamkun J.W., Hartzell G.W. III
Biochim. Biophys. Acta 989:25-48(1989).
[4] Gehring W.J., Hiromi Y.
Annu. Rev. Genet. 20:147-173(1986).
[5] Schofield P.N.
Trends Neurosci. 10:3-6(1987).

20 ⟹ {END}

The NumericalResults section contains information relevant to the results of a scan of SWISS-PROT with this ProSite pattern. In line (7), the following holds:

- /RELEASE specifies the SWISS-PROT release used
- /TOTAL specifies the number of hits on SWISS-PROT[50]
- /POSITIVE specifies the number of hits on proteins known to belong to the class the ProSite patterns indicates[51]
- /UNKNOWN indicates the number of hits to proteins that could belong to the class
- /FALSE_POS indicates the number of hits to proteins that are known to be false hits
- /FALSE_NEG indicates the number of known missed hits

Each of these subfields, except "/RELEASE", gives numbers in the form x(y), where x is the number of hits and y is the number of sequences.[52]

Comment lines, CC, such as (8), provide annotations on various defined types, for example "taxonomy" (/TAXO-RANGE).[53] DR lines (9) provide the cross-references to SWISS-PROT entries that are hit with the pattern, or missed. The SWISS-PROT accession number is followed by the ID, and finally the type of hit is defined by the following:

- T indicates positive hit
- N indicates a false negative
- P indicates a potential hit[54]
- ? indicates unknown
- F indicates a false positive

3D lines (10) provide pointers to PDB entries, and DO lines (11) point to the proper entry in *prosite.doc*. The record is terminated with "//" (12).

The first line of the *prosite.doc* file (13) is the accession number for the documentation of a ProSite pattern, which is pointed to by the DO line of the *prosite.dat* as mentioned in line (11). Line (14) gives the accession number corresponding to the data file for the pattern, from *prosite.dat*, and the

[50] The number of SWISS-PROT sequences to which this ProSite pattern matches.

[51] "Known" means there is existing literature indicating the sequence belongs to the family.

[52] For example, in a protein with multiple Zn-fingers, $x > y$.

[53]
- a = archaebacteria
- b = bacteriophage
- e = eukaryotes
- p = prokaryotes
- v = eukaryotic viruses

[54] Although this sequence belongs to the set under consideration, it was not picked up because the region that contains the pattern is not yet available in the databank.

entry name. If this file provides documentation for more than one data file, all files are listed.[55] The "BEGIN" (15) begins the actual text. The type or name of the pattern is set off as shown in (16), and the text (17) follows in free format much like a short abstract. Other formatted lines are also included in the text (18). References used to develop the documentation file are listed in a simple format such as in (19). The documentation file is ended with "END", as shown in (20).

16.3 Obtaining ProSite

ProSite is available from EMBL by anonymous FTP:

```
ftp ftp.embl-heidelberg.de          [or ftp 192.54.41.33]
cd /pub/databases/prosite
```

Information can also be obtained by e-mail:

```
To: NETSERV@EMBL-Heidelberg.DE
help ProSite                            (as message body)
```

Questions can be sent

```
To: NET-HELP@EMBL-Heidelberg.DE
To: bairoch@cmu.unig
```

16.4 References

Bairoch, A. 1991. PROSITE: A dictionary of sites and patterns in proteins. *Nucleic Acids Research*. 19: 2241–2245.

Bairoch, A. 1993. Prosite: A dictionary of protein sites and patterns: User manual. Release 10.2. July 1993. Obtained from e-mail to *NETSERV@EMBL-Heidelberg.DE* with message *GET PROSITE:PROSUSER.TXT*.

Document: HELP PROSITE. August 1993. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/help*.

PROSITE.DAT. Release 10.2. July 1993. Obtained from e-mail to *NETSERV@EMBL-Heidelberg.DE* with message *GET PROSITE:PROSITE.DAT*.

PROSITE.DOC. Release 10.2. July 1993. Obtained from e-mail to *NETSERV@EMBL-Heidelberg.DE* with message *GET PROSITE:PROSITE.DOC*.

17 SWISS-PROT (Swiss Protein Database)

SWISS-PROT houses amino acid translations of DNA sequences contained in the EMBL nucleotide sequence database and all PIR annotated sequence data, as well as original data. Sequence data

[55] For example, in the patterns specifying the trypsin family of serine proteases, one data file detects the serine active-site residues and another data file detects the histidine active-site residues.

corresponds to protein form before posttranslation and is closely linked and cross-referenced to other EMBL databases, such as PDB and ProSite. A strong effort was made to include annotations about the extent of different sequence domains present in each entry. All information relating to sequence features is stored in computer-readable tables. The ultimate goal of SWISS-PROT "is to provide a complete annotated protein sequence data bank where all the data is easily retrievable by computer programs and is stored in a format similar to that of the EMBL Nucleotide Sequence Database." SWISS-PROT is the work of Amos Bairoch.

17.1 Explanation of a SWISS-PROT Record

The ID record (1)[56] is divided into four parts. The entry name is of the form X_Y, where X is a four-character abbreviation of the protein name and Y a five-character abbreviation of the source. The data class, either standard or preliminary, tells the state of the verification of the record. The third element is the molecule type, generally PRoTein, followed by the sequence length. The ACcession number (2) or numbers provide stable access to the entry, since the entry names may change from release to release. There are always three date lines representing (3) the date the record was created, (4) the last date the sequence was updated, and (5) the last annotation update. DEscription (6) provides a free-format description of the protein, or molecule, while GeneName (7) give the name of the gene from which it is derived. Note that GN may include the logical connectives "and" and "or" where appropriate. The source taxonomy is delineated in the OrganismSpeices (8) lines, which give both the Latin and English names (for viruses, only the English name is given) and the OrganismClassification lines (9). If the protein is found in more than one species, the OS lines will list all species.

The next set of information represents the literature citations. Each reference is first separated by a ReferenceNumber (10). The ReferencePostion line (11) describes the extent of the work carried out by the authors (for example, sequence, review, characterization, mutagenesis, or structure). ReferenceMedline (12) number allows for the reference to be easily accessed. ReferenceAuthors (13) and ReferenceLocation (14) complete the reference section.

[56] See line 1 of the sample entry.

17.2 Sample SWISS-PROT Entry

```
 1 ⟹ ID   TNFA_HUMAN     STANDARD;      PRT;   233 AA.
 2 ⟹ AC   P01375;
 3 ⟹ DT   21-JUL-1986 (REL. 01, CREATED)
 4 ⟹ DT   21-JUL-1986 (REL. 01, LAST SEQUENCE UPDATE)
 5 ⟹ DT   01-DEC-1992 (REL. 24, LAST ANNOTATION UPDATE)
 6 ⟹ DE   TUMOR NECROSIS FACTOR PRECURSOR (TNF-ALPHA) (CACHECTIN).
 7 ⟹ GN   TNFA.
 8 ⟹ OS   HOMO SAPIENS (HUMAN).
 9 ⟹ OC   EUKARYOTA; METAZOA; CHORDATA; VERTEBRATA; TETRAPODA; MAMMALIA;
      OC   EUTHERIA; PRIMATES.
10 ⟹ RN   [1]
11 ⟹ RP   SEQUENCE FROM N.A.
12 ⟹ RM   87217060
13 ⟹ RA   NEDOSPASOV S.A., SHAKHOV A.N., TURETSKAYA R.L., METT V.A.,
      RA   AZIZOV M.M., GEORGIEV G.P., KOROBKO V.G., DOBRYNIN V.N.,
      RA   FILIPPOV S.A., BYSTROV N.S., BOLDYREVA E.F., CHUVPILO S.A.,
      RA   CHUMAKOV A.M., SHINGAROVA L.N., OVCHINNIKOV Y.A.;
14 ⟹ RL   COLD SPRING HARB. SYMP. QUANT. BIOL. 51:611-624(1986).
      RN   [2]
      RP   SEQUENCE FROM N.A.
      RM   85086244
      RA   PENNICA D., NEDWIN G.E., HAYFLICK J.S., SEEBURG P.H., DERYNCK R.,
      RA   PALLADINO M.A., KOHR W.J., AGGARWAL B.B., GOEDDEL D.V.;
      RL   NATURE 312:724-729(1984).
```

[REFs 3-8 deleted]

```
15 ⟹ CC   -!- FUNCTION: CYTOKINE WITH A WIDE VARIETY OF FUNCTIONS: IT CAN
      CC       CAUSE CYTOLYSIS OF CERTAIN TUMOR CELL LINES, IT IS IMPLICATED
      CC       IN THE INDUCTION OF CACHEXIA, IT IS A POTENT PYROGEN CAUSING
      CC       FEVER BY DIRECT ACTION OR BY STIMULATION OF IL-1 SECRETION, IT
      CC       CAN STIMULATE CELL PROLIFERATION & INDUCE CELL DIFFERENTIATION
      CC       UNDER CERTAIN CONDITIONS.
      CC   -!- SUBUNIT: HOMOTRIMER.
      CC   -!- SUBCELLULAR LOCATION: SYNTHESIZED AS A TYPE II MEMBRANE
      CC       PROTEIN, THEN UNDERGOES POST-TRANSLATIONAL CLEAVAGE LIBERATING
      CC       THE EXTRACELLULAR DOMAIN.
      CC   -!- SIMILARITY: BELONGS TO THE TUMOR NECROSIS FACTOR FAMILY.
16 ⟹ DR   EMBL; X02910; HSTNFA.
      DR   EMBL; M16441; HSTNFAB.
      DR   EMBL; X01394; HSTNFR.
      DR   EMBL; M10988; HSTNFAA.
17 ⟹ DR   PIR; B23784; QWHUN.
18 ⟹ DR   PDB; 1TNF; 15-JAN-91.
19 ⟹ DR   MIM; 191160; NINTH EDITION.
20 ⟹ DR   PROSITE; PS00251; TNF.
21 ⟹ KW   CYTOKINE; CYTOTOXIN; TRANSMEMBRANE; GLYCOPROTEIN; SIGNAL-ANCHOR;
      KW   MYRISTYLATION; 3D-STRUCTURE.
22 ⟹ FT   PROPEP        1     76
      FT   CHAIN        77    233       TUMOR NECROSIS FACTOR.
      FT   TRANSMEM     36     56       SIGNAL-ANCHOR (TYPE-II PROTEIN).
      FT   LIPID        19     19       MYRISTATE.
      FT   LIPID        20     20       MYRISTATE.
      FT   DISULFID    145    177
```

```
        FT   MUTAGEN     108    108       R->W: BIOLOGICALLY INACTIVE.
        FT   MUTAGEN     112    112       L->F: BIOLOGICALLY INACTIVE.
        FT   MUTAGEN     162    162       S->F: BIOLOGICALLY INACTIVE.
        FT   MUTAGEN     167    167       V->A,D: BIOLOGICALLY INACTIVE.
        FT   MUTAGEN     222    222       E->K: BIOLOGICALLY INACTIVE.
        FT   CONFLICT     63     63       F -> S (IN REF. 5).
23 ⟹   SQ   SEQUENCE   233 AA;  25644 MW;  279986 CN;
             MSTESMIRDV ELAEEALPKK TGGPQGSRRC LFLSLFSFLI VAGATTLFCL LHFGVIGPQR
             EEFPRDLSLI SPLAQAVRSS SRTPSDKPVA HVVANPQAEG QLQWLNRRAN ALLANGVELR
             DNQLVVPSEG LYLIYSQVLF KGQGCPSTHV LLTHTISRIA VSYQTKVNLL SAIKSPCQRE
             TPEGAEAKPW YEPIYLGGVF QLEKGDRLSA EINRPDYLDF AESGQVYFGI IAL
24 ⟹   //
```

Comments or notes are found on the CC lines (15), with each new comment initiated by "-!-" and a topic specifier, of which there are a finite number then free-form text. DR lines provide cross-references to other databases such as EMBL Nucleotide databases (16), PIR (17), PDB (18), MIM (19), and ProSite (20). SWISS-PROT also includes cross references to the following.

- REBASE (Restriction enzyme database)
- TFD (Transcription factors database)
- EcoGene section of the EcoSeq/EcoMap integrated *E. coli* databases of NCBI
- ECO2DBASE (gene-protein database of *E. coli* 2D-gel spots)
- Human keratinocyte 2D gel protein database of Aarhus and Ghent
- SWISS-2DPAGE (human 2D gel protein database of University of Geneva)
- FlyBase (database of *Drosophila* genetic maps)
- HIV Sequence Database (including nucleic and amino acid sequences for human retroviruses)

KeyWords (21) are also given, making it easier to link SWISS-PROT to other databases and search for a general topic.

FT lines (22) form a "feature table" that provides a precise but simple method of annotating sequence data. A finite list of identifiers can be used to being each line followed by the "from" and "to" amino acid positions. A small comment or explanation completes the line. The actual sequence information is initiated with a line (23) including the number of amino acids, the molecular weight, and the "checking number." The sequence begins on the next line without a new first column identifier. The record is ended by "//" as shown in line (24).

17.3 Obtaining SWISS-PROT

SWISS-PROT is available from EMBL by anonymous FTP:

```
ftp ftp.embl-heidelberg.de          [or ftp 192.54.41.33]
cd /pub/databases/swissprot
```

Information can also be obtained by e-mail:

```
To: NETSERV@EMBL-Heidelberg.DE
help prot                          (as message body)
```

Questions can be sent

```
To: NET-HELP@EMBL-Heidelberg.DE
```

17.4 References

Bairoch, A. 1993. The SWISS-PROT protein sequence data bank: User manual. Release 25, April 1993. Obtained from Email to *NETSERV@EMBL-Heidelberg.DE* with message *GET PROT:USERMAN.TXT*.

Bairoch, A., and B. Boeckmann. 1991. The SWISS-PROT protein sequence data bank. *Nucleic Acids Research*. 19: 2247–2249.

General Database Locations

EMBL is currently collecting many of the cited databases, with the objective of providing convenient one-site access. Questions about unknown databases can be sent to the NETSERVE at EMBL, *NETSERV@EMBL-Heidelberg.DE*, or to the human support group, *NET-HELP@EMBL-Heidelberg.DE*. NCBI is also collecting many databases; questions to NCBI can be sent to *repository@ncbi.nlm.nih.gov*.

Conclusion

This paper has reviewed 18 different molecular biological databases that are more or less relevant to the tasks of sequence and structure analysis. As we mentioned in the introduction, our interest was to present the highlights of each database and a pointer to access it. Since biological databases are constantly changing in size, focus, and number, we appreciate readers' pointers to new databases or corrections to our presentation (send comments to gaasterland@mcs.anl.gov or rayl@mcs.anl.gov).

This paper is available in WWW form at

http://www.mcs.anl.gov/home/compbio/databases/databases.html.

References

[1] Bairoch, A., and B. Boeckmann. 1991. The SWISS-PROT protein sequence data bank. *Nucleic Acids Research*. 19: 2247–2249.

[2] Bairoch, A. 1993. Document name: ENZCLASS.TXT. Release 13.0, July 1993. obtained from e-mail to *NETSERV@EMBL-Heidelberg.DE* with message *GET ENZYME:ENZCLASS.TXT*.

[3] Bairoch, A. 1993. Document name: ENZYME.GET. Release 13.0, July 1993. Obtained from e-mail to *NETSERV@EMBL-Heidelberg.DE* with message *GET ENZYME:ENZYME.GET*.

[4] Bairoch, A. 1993. ENZYME Data Bank. Release 13.0, July 1993. Obtained from e-mail to *NETSERV@EMBL-Heidelberg.DE* with message *GET ENZYME:ENZYME.DAT*.

[5] Bairoch, A. 1993. Prosite: A dictionary of protein sites and patterns: User manual. Release 10.2. July 1993. Obtained from e-mail to *NETSERV@EMBL-Heidelberg.DE* with message *GET PROSITE:PROSUSER.TXT*.

[6] Bairoch, A. 1993. The ENZYME data bank. *Nucleic Acids Research*. 21: 3155–3156.

[7] Bairoch, A. 1993. The SWISS-PROT protein sequence data bank: User Manual. Release 25, April 1993. Obtained from e-mail to *NETSERV@EMBL-Heidelberg.DE* with message *GET PROT:USERMAN.TXT*.

[8] Bairoch, A. 1991. PROSITE: A dictionary of sites and patterns in proteins. *Nucleic Acids Research*. 19: 2241–2245.

[9] Bairoch, A. 1993. The ENZYME data bank user manual. Release 13.0, July 1993. Obtained from e-mail to *NETSERV@EMBL-Heidelberg.DE* with message *GET ENZYME:ENZUSER.TXT*.

[10] Barker, W., D. George, L. Hunt, and J. Garavelli. 1991. The PIR protein sequence database. *Nucleic Acids Research*. 19: 2231–2236.

[11] Blocks Database (BLOCKS.DAT_6.2). Release 6.2, August 1993. obtained from anonymous FTP to *ncbi.nlm.nih.gov* in */repository/blocks*.

[12] Burks, C. Document: limb.doc. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/limb*.

[13] Burks, C., et al. 1991. GenBank. *Nucleic Acids Research*. 19: 2221–2225.

[14] ATA DICTIONARY (Document: data_dict.ps.Z.) Version 5.0.1. obtained from anonymous FTP to *mendel.welch.jhu.edu* in */gbd-5.0*.

[15] Document: 1ace.dssp. April 1992. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/protein_extras/dssp*.

[16] Document: 1ak3.dssp. October 1992. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/protein_extras/dssp*.

[17] Document: 9xim.hssp. October 1993. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/protein_extras/hssp*.

[18] Document: RETRIEVE E-Mail Server. October, 1993. Obtained from e-mail to *retrieve@ncbi.nlm.nih.gov* with message *help*.

[19] Document: cpk2.dssp. October 1992. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/protein_extras/dssp*.

[20] Document: dssp_help.doc. Obtained from Chris Sander at *Chris.Sander@EMBL-Heidelberg.DE*.

[21] Document: zif1.dssp. May 1992. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/protein_extras/dssp*.

[22] Document: 1bmv2_comp3D.fssp. September 1993. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/protein_extras/fssp*.

[23] Document: 1bmv2_dali.fssp. October 1993. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/protein_extras/fssp*.

[24] Document: 1bmv2_suppos.fssp. September 1993. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/protein_extras/fssp*.

[25] Document: 3D_ALI.DOC. Obtained from e-mail to *NETSERV@EMBL-Heidelberg.DE* with message *GET 3D_ALI:3D_ALI.DOC*.

[26] Document: 931031.dat. October 1993. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/embl/new*.

[27] Document: 9xim.hssp. October 1993. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/hssp*.

[28] Document: Announce.blocks_6.2. August 1993. Obtained from anonymous FTP to *ncbi.nlm.nih.gov* in */repository/blocks*. [equivalent to: Document: blocks.doc. August 1993. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/blocks*.]

[29] Document: Blocks E-Mail Searcher. September 1993. Obtained from e-mail to *blocks@howard.fhcrc.org* with message *help*.

[30] Document: ECOLAC [J01636]. May 1993. Obtained from e-mail to *retrieve@ncbi.nlm.nih.gov* with message *DATALIB genbank BEGIN J01636 [ACC]*.

[31] Document: e-mail message from C. Marzec. October 1993. Obtained from e-mail to *MARZEC@NBRF.Georgetown.Edu*.

[32] Document: GBREL.TXT. Release 79.0, October 1993. Obtained from anonymous FTP to *ncbi.nlm.nih.gov* in */genbank*.

[33] Document: HELP BERLIN. May 1992. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/help*.

[34] Document: HELP BLOCKS. January 1993. Obtained from e-mail to *NETSERV@EMBL-Heidelberg.DE* with message *help*.

[35] Document: HELP NUC. October 1993. Obtained from e-mail to *NETSERV@EMBL-Heidelberg.DE* with message *HELP NUC*.

[36] Document: HELP PROSITE. August 1993. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/help*.

[37] Document: INDEX. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/hssp*.

[38] Document: LAMBDA [V00636]. December 1992. Obtained from e-mail to *retrieve@ncbi.nlm.nih.gov* with message *DATALIB genbank BEGIN J02459 [ACC]*.

[39] Document: PIR Request. NBRF-PIR MAILSERVER version 1.49. Obtained from e-mail to *PIRSERVER@NBRF.Georgetown.Edu* with message *HELP*.

[40] Document: PKCDD.FASTA. April 1993. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/pkcdd*.

[41] Document: README. August 1993. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/enzyme*.

[42] Document: README. July 1993. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/pkcdd*.

[43] Document: README. June 1992. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/3d_ali*.

[44] Document: README. Release 79.0, October 1993. Obtained from anonymous FTP to *ncbi.nlm.nih.gov* in */genbank*.

[45] Document: README. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/hssp*.

[46] Document: archae.dat. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/berlin*.

[47] Document: etu.3D_ALI. June 1992. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/3d_ali*.

[48] Document: featuretable.doc. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/embl/doc*.

[49] Document: intro.dat. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/berlin*.

[50] Document: limb.help. July 1992 Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/help*.

[51] Document: read.me. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/berlin*.

[52] Document: repressor.3D_ALI. June 1992. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/3d_ali*.

[53] Document: wrp.3D_ALI. June 1992. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/3d_ali*.

[54] EMBL Data Library: Nucleotide Sequence Database: Release Notes. 1993. Release 36, September 1993. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/embl/doc*.

[55] EMBL Data Library: Nucleotide Sequence Database: User Manual. 1993. Release 36, September 1993. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/embl/doc*.

[56] George, D. and W. Barker. 1989. User's guide for the protein sequence query program of the Protein Identification Resource (PIR)*. Document PSQM-0589. May 1989. Obtained from e-mail to *MARZEC@NBRF.Georgetown.Edu*.

[57] Hanks, S., and A. Quinn. 1993. Protein Kinase Catalytic Domain Database References. Release April 4, 1993. Obtained from e-mail to *NETSERV@EMBL-Heidelberg.DE* with message *GET PKCDD:PKCDD.REF*.

[58] Henikoff, S., and J. Henikoff. 1993. Protein family classification based on searching a database of blocks (Document: blockman.ps). Obtained from anonymous FTP to *sparky.fhcrc.org* in */blocks*.

[59] Holm, L., C. Ouzounis, C. Sander, G. Tuparev, and G. Vriend. 1992. A database of protein structure families with common folding motifs. Pre-release of paper submitted to *Protein Science*.

[60] Holm, L., C. Ouzounis, C. Sander, G. Tuparev, and G. Vriend. 1992. A database of protein structure families with common folding motifs. *Protein Science*. 1: 1691–1698.

[61] Kabash, W., and C. Sander. 1983. Dictionary of protein secondary structures: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 22: 2577–2637.

[62] Lawton, J., F. Martinez, and C. Burks. 1989. Overview of the LiMB database. *Nucleic Acid Research*. 17: 5885–5899.

[63] LiMB Database. (Document: limb.txt.) Release 3.0. Obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/limb*.

[64] LiMBshort Database. (Document: limbshort.txt.) Release 3.0. obtained from anonymous FTP to *ftp.embl-heidelberg.de* in */pub/databases/limb*.

[65] PROSITE.DAT. Release 10.2. July 1993. Obtained from e-mail to *NETSERV@EMBL-Heidelberg.DE* with message *GET PROSITE:PROSITE.DAT*.

[66] PROSITE.DOC. Release 10.2. July 1993. Obtained from e-mail to *NETSERV@EMBL-Heidelberg.DE* with message *GET PROSITE:PROSITE.DOC*.

[67] Pascarella, S., and P. Argos. 1992. A data bank merging related protein structures and sequences. *Protein Engineering*. 5: 121–137.

[68] Pearson, P. 1991. The genome data base (GDB) — a human gene mapping repository. *Nucleic Acids Research*. 19: 2237–2239.

[69] Protein Data Bank Atomic Coordinate and Bibliographic Entry Format Description. Obtained from e-mail to *NETSERV@EMBL-Heidelberg.DE* with message *PROTEINDATA:PDBFORMAT.PS*.

[70] Protein Data Bank Quarterly Newsletter. 1992. Number 61, July 1992. Obtained from e-mail to *NETSERV@EMBL-Heidelberg.DE* with message *GET PROTEINDATA:NEWSLETTER.DOC*.

[71] Protein Data Bank Quarterly Newsletter. 1993. Number 64, April 1993. Obtained from e-mail to *fileserv@pdb.pdb.bnl.gov* with message *send info _your_e-mail_address_*.

[72] Protein Data Bank. July 1993 release. Obtained from anonymous FTP to *pdb.pdb.bnl.gov* to */fullrelease/compressed_files*.

[73] Protein Sequence Database of PIR-International. PIR Document PRDBFS-1292: Databases File Structure and Format Specification. December 1992. Obtained from e-mail to *MARZEC@NBRF.Georgetown.Edu*.

[74] Protein Sequence Database: Sequence File Format. PIR Document PRFILE-1292. December 1992. Obtained from e-mail to *MARZEC@NBRF.Georgetown.Edu*.

[75] Rayl, K., T. Gaasterland, and R. Overbeek. March 1994. Automating the determination of 3D protein structure. Mathematics and Computer Science Division, Argonne National Laboratory, Preprint MCS-P417-0294.

[76] Sander, C. and R. Schneider. 1991. Databases of homology-derived protein structures and the structural meaning of sequence alignment. *PROTEINS: Structure, Function, and Genetics*. 9: 56–68.

[77] Schneider, R., and Sander, C. HSSP: A databases of structure-sequence alignments. HSSP release 1.0. Obtained from e-mail to *NETSERV@EMBL-Heidelberg.DE* with message *GET PROTEINDATA:HSSP.DOC*.