Open access • Journal Article • DOI:10.1093/NAR/20.SUPPL.2019

The SWISS-PROT protein sequence data bank — Source link

Amos Marc Bairoch, Brigitte Boeckmann

Institutions: University of Geneva

Published on: 25 Apr 1991 - Nucleic Acids Research (Oxford University Press)

Topics: Sequence database, UniProt, Peptide sequence and Data bank

Article

The SWISS-PROT protein sequence data bank

BAIROCH, Amos Marc, BOECKMANN, Brigitte

Nucleic Acids Research, Vol. 20, Supplement 2019–2022

The SWISS-PROT protein sequence data bank

Amos Bairoch and Brigitte Boeckmann[1]
Department of Medical Biochemistry, University of Geneva, 1 rue Michel Servet, 1211 Geneva 4, Switzerland and [1]European Molecular Biology Laboratory, Heidelberg, Germany

INTRODUCTION

SWISS-PROT is an annotated protein sequence database established in 1986 and maintained collaboratively, since 1988, by the Department of Medical Biochemistry of the University of Geneva and the EMBL Data Library [1].

SOURCES OF THE SEQUENCE DATA

Sequence data in SWISS-PROT originates from three different sources:

- From the Protein Sequence Database of the Protein Identification Resource (PIR) [2]
- From translation of entries from the EMBL Nucleotide Sequence Database [1]
- From the literature

FORMAT

The SWISS-PROT protein sequence data bank is composed of sequence entries. Each sequence entry is composed of lines. Different types of lines, each with their own format, are used to record the various data which make up the entry. For standardization purposes the format of SWISS-PROT follows as closely as possible that of the EMBL Nucleotide Sequence Database [3]. A sample SWISS-PROT entry is shown in Figure 1.

WHAT DISTINGUISHES SWISS-PROT FROM OTHER PROTEIN SEQUENCE DATABASES?

Annotation

To be useful to the majority of users a protein sequence database should contain as much data as possible on each of the proteins that it describes. In SWISS-PROT, as in most other sequence databases, two classes of data can be distinguished: the core data and the annotation. For each sequence entry the core data consist of the following items:

- Sequence data
- Citation information (bibliographical references)
- Taxonomic data (description of biological source of the protein)

The annotation consists of the description of the following items:

- Function(s) of the protein
- Post-translational modification(s). For example carbohydrates, phosphorylation, acetylation, GPI-anchor, etc.
- Domains and sites. For example calcium binding regions, ATP-binding sites, zinc fingers, homeobox, kringle, etc.
- Secondary structure (using information from the DSSP database) [4]
- Quaternary structure
- Similarities to other proteins
- Disease(s) associated with deficiencie(s) in the protein
- Sequence conflicts, variants, etc.

We try to include as much annotation information as possible in SWISS-PROT. To obtain this information we use, in addition to the publications that reports new sequence data, review articles to periodically update the annotations of families or groups of proteins. We also make use of external experts who send us their comments and updates concerning specific groups of proteins about which they are knowledgeable. We believe that our having systematic recourse both to publications other than those reporting the core data and to subject referees represents a unique and beneficial feature of SWISS-PROT.

In SWISS-PROT, annotation is mainly found in the comment lines (CC), in the feature table (FT) and in the keyword lines (KW). Most comments are classified by 'topics'; this approach permits the easy retrieval of specific categories of data from the database.

Minimal redundancy

We try as much as possible to minimize the redundancy of the database. Many sequence databases contain, for a given protein sequence, separate entries which correspond to different literature reports. Typically one finds a report that corresponds to a fragment of the protein sequenced at the level of the polypeptide, one or more reports reflecting the results of laboratories that have sequenced that protein at the cDNA level, and finally reports from data provided by genomic sequencing. This state of affairs has many drawbacks; for example it is not easy to obtain an overall view of the current state of the knowledge about a given protein, and similarity search programs will pick up the same protein many times during searches. In SWISS-PROT we try as much as possible to merge all these data and, if conflicts exist between various sequencing reports, we indicate them in the feature table.

Integration with other databases

It is important to provide the users of biomolecular databases with a degree of integration between the three types of sequence-related databases (nucleic acid sequences, protein sequences and protein tertiary structures) as well as with specialized data collections. So as to provide tools that will allow software developers to implement such an integrated approach we have cross-referenced SWISS-PROT with many other databases. Currently cross-references are provided for the following databases:

- EMBL Nucleotide Sequence Database [1]
- PDB, the Brookhaven Protein Data Bank [5] which stores crystallographic coordinates of proteins

```
ID   CAH2_HUMAN     STANDARD;      PRT;   259 AA.
AC   P009T8;
DT   21-JUL-1986 (REL. 01, CREATED)
DT   21-JUL-1986 (REL. 01, LAST SEQUENCE UPDATE)
DT   01-MAR-1992 (REL. 21, LAST ANNOTATION UPDATE)
DE   CARBONIC ANHYDRASE II (EC 4.2.1.1) (CARBONATE DEHYDRATASE II).
GN   CA2.
OS   HOMO SAPIENS (HUMAN).
OC   EUKARYOTA; METAZOA; CHORDATA; VERTEBRATA; TETRAPODA; MAMMALIA;
OC   EUTHERIA; PRIMATES.
RN   [1]
RP   SEQUENCE FROM N.A.
RM   87231043
RA   MONTGOMERY J.C., VENTA P.J., TASHIAN R.E., HEWETT-EMMETT D.;
RL   NUCLEIC ACIDS RES. 15:4687-4687(1987).
RN   [2]
RP   SEQUENCE FROM N.A.
RM   88085190
RA   MURAKAMI H., MARELICH G.P., GRUBB J.H., KYLE J.W., SLY W.S.;
RL   GENOMICS 1:159-166(1987).
RN   [3]
RP   SEQUENCE.
RM   77006079
RA   HENDERSON L.E., HENRIKSSON D., NYMAN P.O.;
RL   J. BIOL. CHEM. 251:5457-5463(1976).
RN   [4]
RP   SEQUENCE.
RM   74143468
RA   LIN K.-T.D., DEUTSCH H.F.;
RL   J. BIOL. CHEM. 249:2329-2337(1974).
RN   [5]
RP   SEQUENCE OF 1-76 FROM N.A.
RM   86077780
RA   VENTA P.J., MONTGOMERY J.C., HEWETT-EMMETT D., TASHIAN R.E.;
RL   BIOCHIM. BIOPHYS. ACTA 826:195-201(1985).
RN   [6]
RP   X-RAY CRYSTALLOGRAPHY, 2.0 ANGSTROMS.
RM   72111787
RA   LILJAS A., KANNAN K.K., BERGSTEN P.-C., WAARA I., FRIDBORG K.,
RA   STRANDBERG B., CARLBOM U., JARUP L., LOVGREN S., PETEF M.;
RL   NATURE NEW BIOL. 235:131-137(1972).
RN   [7]
RP   X-RAY CRYSTALLOGRAPHY, 2.0 ANGSTROMS.
RM   89315726
RA   ERIKSSON A.E., JONES T.A., LILJAS A.;
RL   PROTEINS 4:274-282(1988).
RN   [8]
RP   X-RAY CRYSTALLOGRAPHY, 2.0 ANGSTROMS.
RM   89315727
RA   ERIKSSON A.E., KYLSTEN P.M., JONES T.A., LILJAS A.;
RL   PROTEINS 4:283-293(1988).
RN   [9]
RP   VARIANT JOGJAKARTA.
RM   83100296
RA   JONES G.L., SOFRO A.S.M., SHAW D.C.;
RL   BIOCHEM. GENET. 20:979-1000(1982).
RN   [10]
RP   VARIANT MELBOURNE.
RM   83236368
RA   JONES G.L., SHAW D.C.;
RL   HUM. GENET. 63:392-399(1983).
CC   -!- FUNCTION: REVERSIBLE HYDRATATION OF CARBON DIOXIDE.
CC   -!- CATALYTIC ACTIVITY: H(2)CO(3) = CO(2) + H(2)O.
CC   -!- THERE ARE AT LEAST 6 ENZYMATIC FORMS OF CARBONIC ANHYDRASE: CA-I
CC       (OR B), CA-II (OR C), CA-III (OR M), CA-IV, CA-V AND CA-VI.
CC   -!- DISEASE: DEFECTS IN CA2 ARE THE CAUSE OF OSTEOPETROSIS WITH RENAL
CC       TUBULAR ACIDOSIS (MARBLE BRAIN DISEASE).
DR   EMBL; Y00339; HSCA2.
DR   EMBL; X03251; HSCAII.
DR   EMBL; J03037; HSCAIIA.
DR   PIR; A01141; CRHU2.
DR   PIR; A23202; A23202.
DR   PIR; A27175; A27175.
DR   PDB; 1CA2; 15-JAN-90.
DR   PDB; 2CA2; 15-APR-90.
DR   PDB; 3CA2; 15-APR-90.
DR   MIM; 259730; NINTH EDITION.
DR   PROSITE; PS00162; CARBONIC_ANHYDRASE.
KW   LYASE; ACETYLATION; ZINC; 3D-STRUCTURE.
FT   INIT_MET      0      0
FT   MOD_RES       1      1       ACETYLATION.
FT   ACT_SITE     63     63
FT   ACT_SITE     66     66
FT   METAL        93     93       ZINC (CATALYTIC).
FT   METAL        95     95       ZINC (CATALYTIC).
FT   METAL       118    118       ZINC (CATALYTIC).
FT   ACT_SITE    126    126
FT   ACT_SITE    196    198
FT   VARIANT      17     17       K -> E (JOGJAKARTA).
FT   VARIANT     235    235       P -> H (MELBOURNE).
```

```
FT   VARIANT    251    251       N -> D.
FT   TURN         8     10
FT   HELIX       12     18
FT   HELIX       20     23
FT   TURN        25     26
FT   STRAND      31     32
FT   TURN        34     36
FT   STRAND      38     39
FT   TURN        41     42
FT   STRAND      46     49
FT   TURN        51     52
FT   STRAND      55     60
FT   STRAND      65     69
FT   STRAND      77     80
FT   TURN        81     82
FT   STRAND      87     96
FT   TURN       100    101
FT   STRAND     107    108
FT   TURN       109    110
FT   STRAND     111    111
FT   STRAND     115    123
FT   HELIX      124    126
FT   HELIX      129    132
FT   TURN       133    134
FT   TURN       136    137
FT   STRAND     139    148
FT   HELIX      153    165
FT   TURN       168    169
FT   STRAND     171    173
FT   HELIX      179    182
FT   STRAND     189    194
FT   TURN       199    200
FT   STRAND     205    210
FT   STRAND     214    216
FT   HELIX      218    224
FT   TURN       225    226
FT   STRAND     228    228
FT   TURN       232    233
FT   STRAND     238    238
FT   TURN       250    251
FT   STRAND     255    256
SQ   SEQUENCE   259 AA;  29115 MW;  365693 CN;
     SHHWGYGKHN GPEHWHKDFP IAKGERQSPV DIDTHTAKYD PSLKPLSVSY DQATSLRILN
     NGHAFNVEFD DSQDKAVLKG GPLDGTYRLI QFHFHWGSLD GQGSEHTVDK KKYAAELHLV
     HWNTKYGDFG KAVQQPDGLA VLGIFLKVGS AKPGLQKVVD VLDSIKTKGK SADFTNFDPR
     GLLPESLDYW TYPGSLTTPP LLECVTWIVL KEPISVSSEQ VLKFRKLNFN GEGEPEELMV
     DNWRPAQPLK NRQIKASFK
//
```

Figure 1. A sample entry from SWISS-PROT

PIR, the protein sequence database of the Protein Identification Resource [2]

EcoGene, from the EcoSeq/EcoMap integrated Escherichia coli database [6].

FlyBase, the Drosophila Genetic database prepared by Michael Ashburner at the Department of Genetics, University of Cambridge.

Gene-protein database of Escherichia coli K-12 (2D-gel spots) [7].

HIV, the human retroviruses and AIDS Database [8]

OMIM, the on-line version of the book 'Mendelian Inheritance in Man' [9]

PROSITE, the Dictionary of Protein Sites and Patterns [10]

REBASE, the restriction enzymes database [11]

TFD, the transcription factors data bank [12]

Cross-references are provided in the form of pointers to information related to SWISS-PROT entries and found in data collections other than SWISS-PROT. They are implemented using a specific type of line, the 'DR' (for *D*ata bank *R*eference) line. For example the sample sequence shown in Figure 1 contains DR lines that point to EMBL, PIR, PDB, OMIM, and PROSITE. In that particular example it is therefore possible to retrieve the nucleic acid sequence(s) that encodes for that protein (EMBL), the X-ray crystallographic atomic coordinates (PDB), or the description of genetic disease(s) associated with that protein (OMIM).

CONTENT OF THE CURRENT RELEASE

Release 21.0 of SWISS-PROT (March 1992) contains 23742 sequence entries, comprising 7,866,596 amino acids abstracted from 23919 references. The data file (sequences and annotations) requires 40 Mb of disk storage space. The database is distributed with 15 documentation and index files (user's manual, release notes, list of organisms, citation index, keyword index, etc.) that require about 11 Mb of disk space.

DISTRIBUTION

SWISS-PROT is distributed on magnetic tape and on CD-ROM by the EMBL Data Library. The CD-ROM contains both SWISS-PROT and the EMBL Nucleotide Sequence Database as well as other data collections and some database query and retrieval software for MS-DOS PC compatible computers. For all enquiries regarding the subscription and distribution of SWISS-PROT one should contact:

EMBL Data Library
European Molecular Biology Laboratory
Postfach 10.2209, Meyerhofstrasse 1
6900 Heidelberg, Germany

Telephone: (+49 6221) 387 258
Telefax : (+49 6221) 387 519 or 387 306
Electronic network address: datalib@EMBL-heidelberg.de

Individual sequence entries can be obtained from the EMBL File Server [13]. Detailed instructions on how to make best use of this service, and in particular on how to obtain protein sequences, can be obtained by sending to the network address netserv@EMBL-heidelberg.de the following message:

HELP
HELP PROT

If you have access to a computer system linked to the Internet you can obtain SWISS-PROT using FTP (File Transfer Protocol), from the following file servers:

GenBank On-line Service [14]
Internet address: genbank.bio.net (134.172.1.160)

NCBI (National Library of Medecine, NIH, Washington D.C., U.S.A.)
Internet address: ncbi.nlm.nih.gov (130.14.20.1)

ExPASy (Expert Protein Analysis System server, University of Geneva, Switzerland)
Internet address: expasy.hcuge.ch (129.195.254.61)

The present distribution frequency is four releases per year. No restrictions are placed on use or redistribution of the data.

REFERENCES

1. Stoehr P.J., Cameron G.N. Nucleic Acids Res. 20:xxx-xxx(1992).
2. Barker W.C., Hunt L.T., George D.G., Garavelli J.S. Nucleic Acids Res. 19:2231–2236(1991).
3. EMBL Data Library Nucleotide Sequence Database User Manual, Release 29 of December 1991.
4. Kabsch W., Sander C. Biopolymers 22:2577–2637(1983).
5. Abola E.E., Bernstein F.C., Koetzle T.F. Computational molecular biology. Sources and methods for sequence analysis, Lesk A.M., Editor, pp69–81, Oxford University Press, Oxford, (1988).
6. Rudd K.E., Miller W., Werner C., Ostell J., Tolstoshev C., Satterfield S.G. Nucleic Acids Res. 19:637–647(1991).
7. VanBogelen R.A., Neidhardt F.C. Electrophoresis 12:955–994(1991).
8. Human Retroviruses and AIDS 1990. A compilation and analysis of nucleic acid and amino acid sequences. Myers G., Rabson A.B. , Josephs S.F. , Smith T.F. , Berzofsky J.A., Wong-Staal F., Editors. Theoretical Biology and Biophysics Group T-10, Los Alamos National Laboratory (1990).
9. McKusick V.A. Mendelian Inheritance in Man. Catalogs of autosomal dominant, autosomal recessive, and X-linked phenotypes; Ninth edition; Johns Hopkins University Press, Baltimore, (1990).
10. Bairoch A. Nucleic Acids Res. 20:2013–2018(1992).
11. Roberts R., Macelis D. Nucleic Acids Res. 20:2167–2180(1992).
12. Ghosh D. Trends Biochem. Sci. 16:445–447(1991).
13. Stoehr P.J., Omond R.A. Nucleic Acids Res. 17:6763–6764(1989).
14. Benton D. Nucleic Acids Res. 18:1517–1520(1990).