UNIVERSITÉ DE GENÈVE

Archive ouverte UNIGE
https://archive-ouverte.unige.ch

Article scientifique | Article | 1993

Published version | Open Access

This is the published version of the publication, made available in accordance with the publisher's policy.

The SWISS-PROT protein sequence data bank, recent developments

Bairoch, Amos Marc; Boeckmann, Brigitte

How to cite

BAIROCH, Amos Marc, BOECKMANN, Brigitte. The SWISS-PROT protein sequence data bank, recent developments. In: Nucleic acids research, 1993, vol. 21, n° 13, p. 3093–3096. doi: 10.1093/nar/21.13.3093

This publication URL: https://archive-ouverte.unige.ch/unige:36852
Publication DOI: 10.1093/nar/21.13.3093

© This document is protected by copyright. Please refer to copyright holder(s) for terms of use.

© 1993 Oxford University Press

Nucleic Acids Research, 1993, Vol. 21, No. 13 3093–3096

The SWISS-PROT protein sequence data bank, recent developments

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 [1] 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 [2]. The SWISS-PROT protein sequence data bank consist of sequence entries. Sequence entries are composed of different lines types, each with their own format. For standardization purposes the format of SWISS-PROT [3] follows as closely as possible that of the EMBL Nucleotide Sequence Database. A sample SWISS-PROT entry is shown in Figure 1.

The SWISS-PROT database distinguishes itself from other protein sequence databases by three distinct criteria:

Annotation

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 consists of the sequence data; the citation information (bibliographical references) and the taxonomic data (description of the biological source of the protein) while the annotation consists of the description of the following items:

Function(s) of the protein
Post-translational modification(s)
Domains and sites
Secondary structure
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 report new sequence data, review articles to periodically update the annotations of families or groups of proteins. We also make use of external experts, who have been recruited to send us their comments and updates concerning specific groups of proteins.

Minimal redundancy

Many sequence databases contain, for a given protein sequence, separate entries which correspond to different literature reports. In SWISS-PROT we try as much as possible to merge all these data so as to minimize the redundancy of the database. If conflicts exist between various sequencing reports, they are indicated in the feature table of the corresponding entry.

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. SWISS-PROT is currently cross-referenced with twelve different databases. 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. For example the sample sequence shown in Figure 1 contains DR (*D*ata bank *R*eference) lines that point to EMBL, PIR, PDB, OMIM, and PROSITE. In this 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), the description of genetic disease(s) associated with that protein (OMIM), or the pattern specific for that family of proteins (PROSITE).

RECENT DEVELOPMENTS

Integration of information from 2D gel databases

Enormous progress has been made in two-dimensional (2D) gel techniques in the last few years. One of the consequences of this evolution has been the development of databases that contain master gels from a variety of mammalian tissues or from bacterial sources. These databases will play an increasingly important role in the analysis of genomes and of molecular diseases. 2D gel databases generally contain one or more master images of the gels that correspond to the tissue or organism studied; spots on these images are attributed an identification code and a variable percentage of these spots are linked to known proteins. The identification of a protein on a 2D gel is generally carried out using antibodies or by microsequencing. Microsequencing of 2D gel spots also produces partial sequences and physico-chemical data for a number of yet uncharacterized proteins.

SWISS-PROT has committed itself to work in close collaboration with a number of groups developing 2D gel databases. Since last year, cross-references to the gene-protein database of Escherichia coli K-12 (ECO2DBASE) [4] have been available and symmetrically that database now contains cross-references to SWISS-PROT. As a second step we have expanded our links to 2D gel databases by integrating data from the following sources:

The Human 2D gel protein database of the Faculty of Medicine of the University of Geneva (known as SWISS-2DPAGE).

```
ID   TNFA_HUMAN     STANDARD;      PRT;   233 AA.
AC   P01375;
DT   21-JUL-1986 (REL. 01, CREATED)
DT   21-JUL-1986 (REL. 01, LAST SEQUENCE UPDATE)
DT   01-DEC-1992 (REL. 24, LAST ANNOTATION UPDATE)
DE   TUMOR NECROSIS FACTOR PRECURSOR (TNF-ALPHA) (CACHECTIN).
GN   TNFA.
OS   HOMO SAPIENS (HUMAN).
OC   EUKARYOTA; METAZOA; CHORDATA; VERTEBRATA; TETRAPODA; MAMMALIA;
OC   EUTHERIA; PRIMATES.
RN   [1]
RP   SEQUENCE FROM N.A.
RM   87217060
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.;
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).
RN   [3]
RP   SEQUENCE FROM N.A.
RM   85137898
RA   SHIRAI T., YAMAGUCHI H., ITO H., TODD C.W., WALLACE R.B.;
RL   NATURE 313:803-806(1985).
RN   [4]
RP   SEQUENCE FROM N.A.
RM   86016093
RA   NEDWIN G.E., NAYLOR S.L., SAKAGUCHI A.Y., SMITH D.H.,
RA   JARRETT-NEDWIN J., PENNICA D., GOEDDEL D.V., GRAY P.W.;
RL   NUCLEIC ACIDS RES. 13:6361-6373(1985).
RN   [5]
RP   SEQUENCE FROM N.A.
RM   85142190
RA   WANG A.M., CREASEY A.A., LADNER M.B., LIN L.S., STRICKLER J.,
RA   VAN ARSDELL J.N., YAMAMOTO R., MARK D.F.;
RL   SCIENCE 228:149-154(1985).
RN   [6]
RP   X-RAY CRYSTALLOGRAPHY (2.6 ANGSTROMS).
RM   90008932
RA   ECK M.J., SPRANG S.R.;
RL   J. BIOL. CHEM. 264:17595-17605(1989).
RN   [7]
RP   MUTAGENESIS.
RM   91184128
RA   OSTADE X.V., TAVERNIER J., PRANGE T., FIERS W.;
RL   EMBO J. 10:827-836(1991).
RN   [8]
RP   MYRISTOYLATION.
RA   STEVENSON F.T., BURSTEN S.L., LOCKSLEY R.M., LOVETT D.H.;
RL   J. EXP. MED. 176:1053-1062(1992).
CC   -!- FUNCTION: TNF IS MAINLY SECRETED BY MACROPHAGES, IT IS A CYTOKINE
CC       WITH A WIDE VARIETY OF FUNCTIONS: IT CAN CAUSE CYTOLYSIS OF
CC       CERTAIN TUMOR CELL LINES, IT IS IMPLICATED IN THE INDUCTION OF
CC       CACHEXIA, IT IS A POTENT PYROGEN CAUSING FEVER BY DIRECT ACTION
CC       OR BY STIMULATION OF INTERLEUKIN 1 SECRETION, IT CAN STIMULATE
CC       CELL PROLIFERATION AND INDUCE CELL DIFFERENTIATION UNDER CERTAIN
CC       CONDITIONS.
CC   -!- SUBUNIT: HOMOTRIMER.
CC   -!- SUBCELLULAR LOCATION: SYNTHESIZED AS A TYPE II MEMBRANE PROTEIN,
CC       THEN UNDERGOES POST-TRANSLATIONAL CLEAVAGE LIBERATING THE
CC       EXTRACELLULAR DOMAIN.
CC   -!- DISEASE: CACHEXIA ACCOMPANIES A VARIETY OF DISEASES, INCLUDING
CC       CANCER AND INFECTION, AND IS CHARACTERIZED BY GENERAL ILL HEALTH
CC       AND MALNUTRITION.
CC   -!- SIMILARITY: BELONGS TO THE TUMOR NECROSIS FACTOR FAMILY.
DR   EMBL; X02910; HSTNFA.
DR   EMBL; M16441; HSTNFAB.
DR   EMBL; X01394; HSTNFR.
DR   EMBL; M10988; HSTNFAA.
DR   PIR; B23784; GMHU.
DR   PDB; 1TNF; 15-JAN-91.
DR   MIM; 191160; TENTH EDITION.
DR   PROSITE; PS00251; TNF.
KW   CYTOKINE; CYTOTOXIN; TRANSMEMBRANE; GLYCOPROTEIN; SIGNAL-ANCHOR;
KW   MYRISTYLATION; 3D-STRUCTURE.
FT   PROPEP       1     76
FT   CHAIN       77    233       TUMOR NECROSIS FACTOR.
FT   TRANSMEM    36     56       SIGNAL-ANCHOR (TYPE-II MEMBRANE PROTEIN).
FT   LIPID       19     19       MYRISTATE.
FT   LIPID       20     20       MYRISTATE.
FT   DISULFID   145    177
FT   MUTAGEN    105    105       L->S: LOW ACTIVITY.
FT   MUTAGEN    108    108       R->W: BIOLOGICALLY INACTIVE.
FT   MUTAGEN    112    112       L->F: BIOLOGICALLY INACTIVE.
FT   MUTAGEN    160    160       A->V: 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).
FT   STRAND      89     93
FT   TURN        99    100
FT   TURN       109    110
FT   STRAND     112    113
FT   TURN       115    116
FT   STRAND     118    119
FT   STRAND     124    125
FT   STRAND     130    143
FT   STRAND     152    159
FT   STRAND     166    170
FT   STRAND     173    174
FT   TURN       183    184
FT   STRAND     189    202
FT   TURN       204    205
FT   STRAND     207    212
FT   HELIX      215    217
FT   STRAND     218    218
FT   STRAND     227    232
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
//
```

Figure 1. A sample entry from SWISS-PROT.

SWISS-2DPAGE currently contains data concerning plasma [5] and liver [6] proteins, but will soon include additional tissues.

The Human keratinocyte 2D gel protein database from the universities of Aarhus and Ghent [7] (known as AARHUS/GHENT-2DPAGE).

For both of the above databases we provide:

a) Cross-references to the identificators for the spots corresponding to known or unknown microsequenced proteins.

b) We have created new entries for microsequences that correspond to novel, yet unidentified, proteins.

c) In some cases we have entered the extent of the microsequences for already known proteins. This was done for proteins which are not yet well characterized. The availability of such microsequences allows, for example, to confirm the position of a signal sequence cleavage site or to confirm the correctness of a translated genomic sequence.

In the near future the collaboration with the group of Denis Hochstrasser which produces the SWISS-2DPAGE database will be expanded in the following directions:

a) The MELANIE software package [8] which is a complete system for the analysis of 2D gels and which is developed by the group of Hochstrasser will allow its users to navigate back and forth between SWISS-2DPAGE and SWISS-PROT.

b) A file server will be set up that will allow anyone with a network connection to obtain annotated graphic files containing the region of the gels that correspond to a selected SWISS-PROT entry linked to SWISS-2DPAGE.

Integration of secondary and tertiary structure data

Thanks to recent advances in experimental techniques there has been a significant increase in the number of protein sequences that have been characterized at the level of their tertiary structure either by X-ray crystallography or by NMR-based methods. A particular effort has been made to provide access to this category of information from inside SWISS-PROT. This effort is conceptualized by the following attributes:

a) Thanks to a collaboration with the group of Chris Sander at EMBL, the feature table of sequence entries of proteins whose tertiary structure is known experimentally contains the secondary structure information corresponding to that protein. The secondary structure assignment is made according to the Dictionary of Secondary Structure of Proteins (DSSP) [9] and the information is extracted from the coordinate data sets of the Protein Data Bank (PDB) [10]. In the feature table three types of secondary structure are specified: helices (key 'HELIX'), beta-strand (key 'STRAND') and turns (key 'TURN'). Residues not specified in one of these classes are in a 'loop' or 'random-coil' structure.

b) Cross-references are available to entries in both sections of the PDB database (annotated and preliminary). In addition the protein sequence entries that are linked to PDB contain the keyword '3D-STRUCTURE'.

c) We try to include, in SWISS-PROT as many bibliographical references as possible to papers dealing with structural data

that originate from X-ray crystallography or NMR studies. These references are prefixed by RP lines such as those shown below:

```
RP X-RAY CRYSTALLOGRAPHY (n.n ANGSTROMS).
RP STRUCTURE BY NMR.
RP 3D-STRUCTURE MODELLING.
```

Human genetic diseases

An increasing number of human genetic diseases are being characterized at the molecular level. We have integrated information concerning these diseases in SWISS-PROT. In particular we provide:

a) Cross-references to OMIM, the on-line version of the book 'Mendelian Inheritance in Man' [11]. This database provides a wealth of data on mapped and sequenced human genes including a full description of the phenotype of known Mendelian disorders as well as information relative to known allelic variants. Currently there are more than 1700 human protein sequence entries in SWISS-PROT which are cross-referenced to OMIM. A document file (MIMTOSP.TXT) is distributed with SWISS-PROT that lists these entries and their corresponding OMIM number(s).

b) When a human protein is known to be involved in a genetic disorder a brief description of that disease is available in the comments section (CC lines) of that entry. As shown in the example below the 'DISEASE' topic is used for such a purpose:

```
CC -!- DISEASE: DEFECTS IN SOD1 ARE THE CAUSE OF
CC     FAMILIAL AMYOTROPHIC
CC     LATERAL SCLEROSIS (FALS), A DEGENERATIVE
CC     DISORDER OF MOTOR
CC     NEURONS IN THE CORTEX, BRAINSTEM AND SPINAL
CC     CORD.
```

c) Point mutations that affect a single amino acid and which are linked with the occurrence of a disease are indicated in the feature table (FT lines) of the relevant entry. As shown in the example below the 'VARIANT' key is used for such a purpose:

```
FT  VARIANT    93    93   D → G (ALABAMA; MODERATE
                          HEMOPHILIA).
FT  VARIANT    96    96   Q → P (NEW LONDON; SEVERE
                          HEMOPHILIA).
FT  ARIANT    102   102   C → R (BASEL; SEVERE
                          HEMOPHILIA).
FT  VARIANT   110   110   D → N (OXFORD-D1; SEVERE
                          HEMOPHILIA).
```

Escherichia coli as a model organism

Thanks to a very fruitful collaboration with Ken Rudd of the National Center for Biotechnology Information (NCBI) protein sequences that originate from the chromosome of Escherichia coli K12 are considered to be a paradigm for what we want to achieve in term of the completeness and quality of the data in SWISS-PROT. The hallmarks of this undertaking are listed below.

a) These entries are cross-referenced to the EcoGene section of the EcoSeq/EcoMap integrated Escherichia coli database [12] and also, as described in subsection 2a above, to the gene-protein 2D gel database of Escherichia coli K-12 (ECO2DBASE) [4].

b) New *Escherichia coli* sequences are entered and annotated on a weekly basis and are immediately made available to the scientific community.

c) Existing *Escherichia coli* sequence entries are constantly updated to add data concerning their functions, to resolve sequence conflicts, to add references and comments, to update gene designations, etc.

d) We have implemented the EcoGene gene name nomenclature for unnamed Escherichia coli hypothetical proteins and proteins of unknown function. They are assigned gene names based upon their position on the genomic physical map. They all begin with the letter 'Y'. The next two letters designate which 1/100th of the map (starting at the thr locus) contain the ORF in the order Yaa, Yab,..Yaj, Yba, Ybb,..Ybj,..., Yja,..Yjj. ORF's within any one of these 100 intervals are given a fourth letter (a-z) that serves to distinguish them but is not meant to convey position information.

e) We provide a document file (ECOLI.TXT) that specifically lists all the E.coli K12 chromosomal sequence entries in SWISS-PROT along with their primary and synonymous gene designations.

PRACTICAL INFORTMATION

Content of the current release

Release 25.0 of SWISS-PROT (April 1993) contains 29,955 sequence entries, comprising 10,214,020 amino acids abstracted from 29,176 references. The data file (sequences and annotations) requires 52 Mb of disk storage space. The database is distributed with 17 documentation and index files (user's manual, release notes, list of organisms, citation index, keyword index, etc.) that require about 14 Mb of disk space.

How to obtain SWISS-PROT

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 and Apple MacIntosh 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 the 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:

EMBL anonymous FTP server
Internet address: *ftp.EMBL-heidelberg.de* (or 192.54.41.33)

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

Basel Biozentrum Biocomputing server (EMBnet SWISS node)
Internet address: *bioftp.unibas.ch* (or 131.152.8.1)

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

National Institute of Genetics (Japan) FTP server
Internet address: *ftp.nig.ac.jp* (133.39.16.66)

You can also obtain SWISS-PROT entries using various Internet Gopher servers that specialize in biosciences (biogophers) [14]. Gopher is a distributed document delivery service that allows a neophyte user to access various types of data residing on multiple hosts in a seamless fashion.

No restrictions are placed on use or redistribution of the data.

Release frequency

The present distribution frequency is four releases per year. Weekly updates are also available; these updates are available by anonymous FTP. Three files are updated every week:

new_seq.dat	Contains all the new entries since the last full release.
upd_seq.dat	Contains the entries for which the sequence data has been updated since the last release.
upd_ann.dat	Contains the entries for which one or more annotation fields have been updated since the last release.

These files are available on the EMBL, NCBI, EMBnet Swiss node and Expasy servers, whose Internet addresses are listed above.

REFERENCES

1. Bairoch A., Boeckmann B. Nucleic Acids Res. 20:2019–2022(1992).
2. Higgins D.G., Fuchs R., Stoehr P.J., Cameron G.N. Nucleic Acids Res. 20:2071–2074(1992).
3. Bairoch A. SWISS-PROT protein sequence data bank user manual, Release 25 of April 1993.
4. VanBogelen R.A., Sankar P., Clark R.L., Bogan J.A., Neidhardt F.C. Electrophoresis 13:1014–1054(1992).
5. Hughes G.J., Frutiger S., Paquet N., Ravier F., Pasquali C., Sanchez J.-C., James R., Tissot J.-D., Bjellqvist B., Hochstrasser D.F. Electrophoresis 13:707–714(1992).
6. Hochstrasser D.F., Frutiger S., Paquet N., Bairoch A., Ravier F., Pasquali C., Sanchez J.-C., Tissot J.-D., Bjellqvist B., Vargas R., Appel R.D., Hughes G.J. Electrophoresis 13:992–1001(1992).
7. Celis J.E., Rasmussen H.H., Madsen P., Leffers H., Honore B., Dejgaard K., Gesser B., Olsen E., Gromov P., Hoffmann H.J., Nielsen M., Celis A., Basse B., Lauridsen J.B., Ratz G.P., Nielsen H., Andersen A.H., Walbum E., Kjaergaard I., Puype M., Van Damme J., Vandekerckhove J. Electrophoresis 13:893–959(1992).
8. Appel R., Hochstrasser D.F., Funk M., Vargas J.R., Pellegrini C., Muller A.F., Scherrer J.-R. Electrophoresis 12:722–735(1991).
9. Kabsch W., Sander C. Biopolymers 22:2577–2637(1983).
10. Koetzle T. CODATA Bulletin 23:83–84(1991).
11. McKusick V.A. Mendelian Inheritance in Man. Catalogs of autosomal dominant, autosomal recessive, and X-linked phenotypes; Tenth edition; Johns Hopkins University Press, Baltimore, (1991).
12. Rudd K.E., Miller W., Werner C., Ostell J., Tolstoshev C., Satterfield S.G. Nucleic Acids Res. 19:637–647(1991).
13. Stoehr P.J., Omond R.A. Nucleic Acids Res. 17:6763–6764(1989).
14. Gilbert D. Trends Biochem. Sci. 18:107–108(1993).