The SWISS-PROT protein sequence data bank and its supplement TrEMBL

Amos Bairoch* and Rolf Apweiler[1]

Department of Medical Biochemistry, University of Geneva, 1 rue Michel Servet, 1211 Geneva 4, Switzerland and [1]The EMBL Outstation – The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received October 17, 1996; Accepted October 21, 1996

ABSTRACT

SWISS-PROT is a curated protein sequence database which strives to provide a high level of annotations (such as the description of the function of a protein, structure of its domains, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases. Recent developments of the database include: an increase in the number and scope of model organisms; cross-references to two additional databases; a variety of new documentation files and the creation of TrEMBL, a computer annotated supplement to SWISS-PROT. This supplement consists of entries in SWISS-PROT-like format derived from the translation of all coding sequences (CDS) in the EMBL nucleotide sequence database, except the CDS already included in SWISS-PROT.

INTRODUCTION

SWISS-PROT (1) is an annotated protein sequence database established in 1986 and maintained collaboratively, since 1987, by the Department of Medical Biochemistry of the University of Geneva and the EMBL Data Library [now the EMBL Outstation–The European Bioinformatics Institute (EBI) (2)]. The SWISS-PROT protein sequence data bank consists 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 (see below).

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). 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

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.

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.

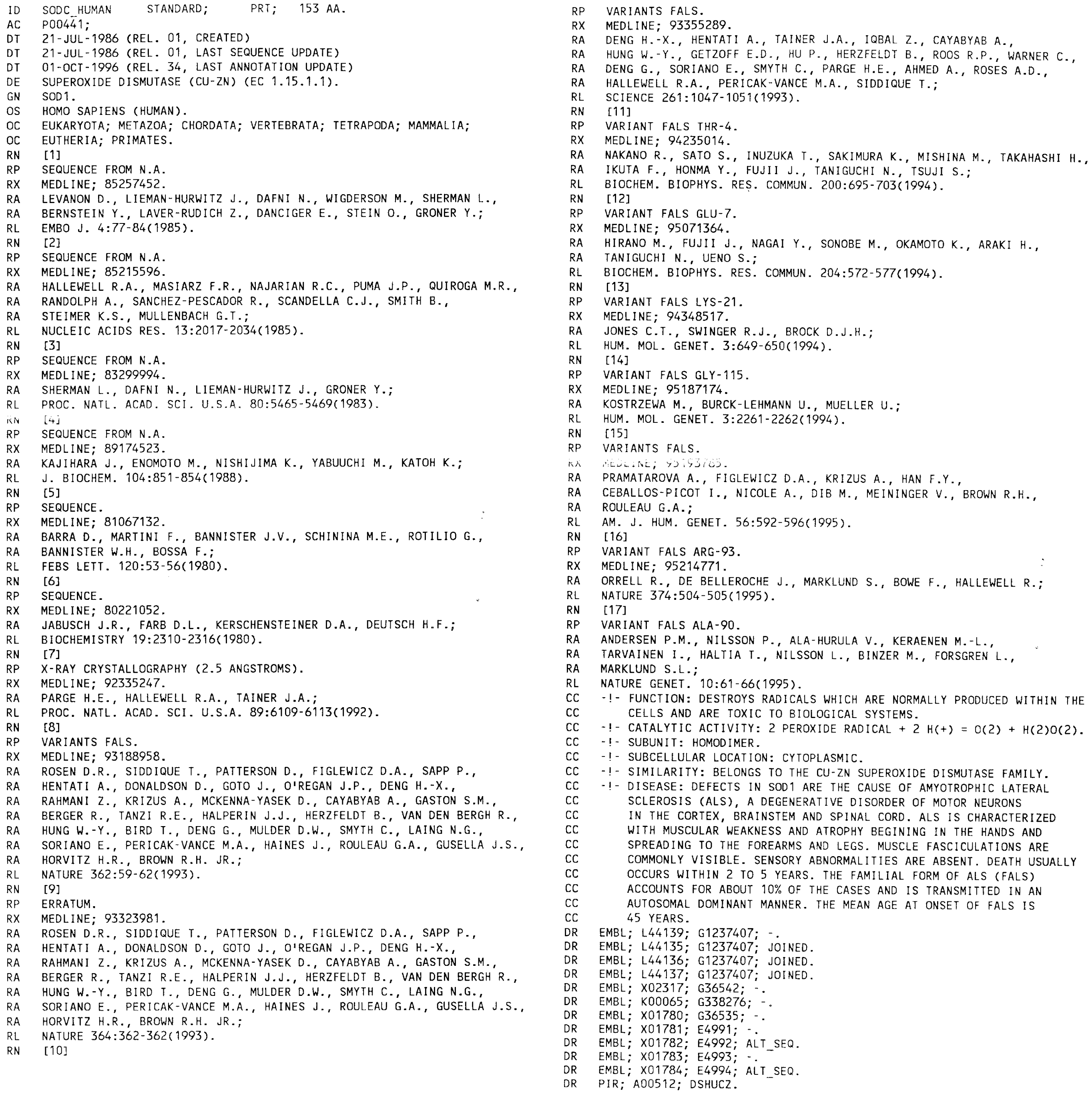
Minimal redundancy

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-

* To whom correspondence should be addressed. Tel: +41 22 784 4082; Fax: +41 22 702 5502; Email: bairoch@cmu.unige.ch



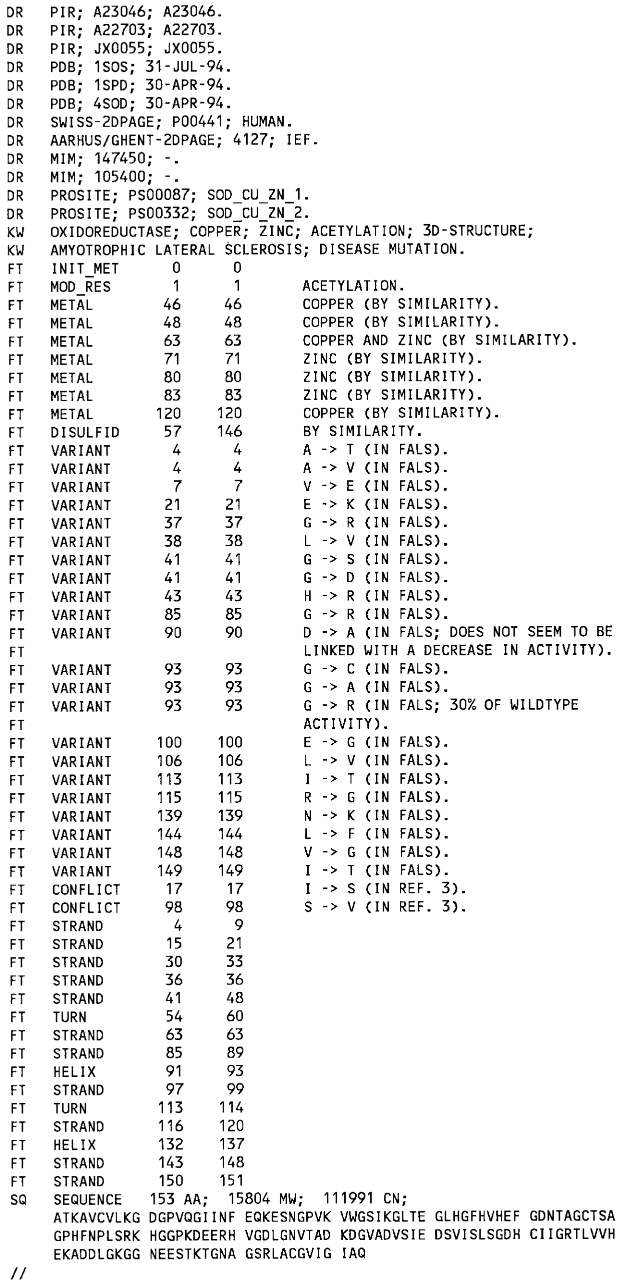
```
ID   SODC_HUMAN     STANDARD;      PRT;   153 AA.
AC   P00441;
DT   21-JUL-1986 (REL. 01, CREATED)
DT   21-JUL-1986 (REL. 01, LAST SEQUENCE UPDATE)
DT   01-OCT-1996 (REL. 34, LAST ANNOTATION UPDATE)
DE   SUPEROXIDE DISMUTASE (CU-ZN) (EC 1.15.1.1).
GN   SOD1.
OS   HOMO SAPIENS (HUMAN).
OC   EUKARYOTA; METAZOA; CHORDATA; VERTEBRATA; TETRAPODA; MAMMALIA;
OC   EUTHERIA; PRIMATES.
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE; 85257452.
RA   LEVANON D., LIEMAN-HURWITZ J., DAFNI N., WIGDERSON M., SHERMAN L.,
RA   BERNSTEIN Y., LAVER-RUDICH Z., DANCIGER E., STEIN O., GRONER Y.;
RL   EMBO J. 4:77-84(1985).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE; 85215596.
RA   HALLEWELL R.A., MASIARZ F.R., NAJARIAN R.C., PUMA J.P., QUIROGA M.R.,
RA   RANDOLPH A., SANCHEZ-PESCADOR R., SCANDELLA C.J., SMITH B.,
RA   STEIMER K.S., MULLENBACH G.T.;
RL   NUCLEIC ACIDS RES. 13:2017-2034(1985).
RN   [3]
RP   SEQUENCE FROM N.A.
RX   MEDLINE; 83299994.
RA   SHERMAN L., DAFNI N., LIEMAN-HURWITZ J., GRONER Y.;
RL   PROC. NATL. ACAD. SCI. U.S.A. 80:5465-5469(1983).
RN   [4]
RP   SEQUENCE FROM N.A.
RX   MEDLINE; 89174523.
RA   KAJIHARA J., ENOMOTO M., NISHIJIMA K., YABUUCHI M., KATOH K.;
RL   J. BIOCHEM. 104:851-854(1988).
RN   [5]
RP   SEQUENCE.
RX   MEDLINE; 81067132.
RA   BARRA D., MARTINI F., BANNISTER J.V., SCHININA M.E., ROTILIO G.,
RA   BANNISTER W.H., BOSSA F.;
RL   FEBS LETT. 120:53-56(1980).
RN   [6]
RP   SEQUENCE.
RX   MEDLINE; 80221052.
RA   JABUSCH J.R., FARB D.L., KERSCHENSTEINER D.A., DEUTSCH H.F.;
RL   BIOCHEMISTRY 19:2310-2316(1980).
RN   [7]
RP   X-RAY CRYSTALLOGRAPHY (2.5 ANGSTROMS).
RX   MEDLINE; 92335247.
RA   PARGE H.E., HALLEWELL R.A., TAINER J.A.;
RL   PROC. NATL. ACAD. SCI. U.S.A. 89:6109-6113(1992).
RN   [8]
RP   VARIANTS FALS.
RX   MEDLINE; 93188958.
RA   ROSEN D.R., SIDDIQUE T., PATTERSON D., FIGLEWICZ D.A., SAPP P.,
RA   HENTATI A., DONALDSON D., GOTO J., O'REGAN J.P., DENG H.-X.,
RA   RAHMANI Z., KRIZUS A., MCKENNA-YASEK D., CAYABYAB A., GASTON S.M.,
RA   BERGER R., TANZI R.E., HALPERIN J.J., HERZFELDT B., VAN DEN BERGH R.,
RA   HUNG W.-Y., BIRD T., DENG G., MULDER D.W., SMYTH C., LAING N.G.,
RA   SORIANO E., PERICAK-VANCE M.A., HAINES J., ROULEAU G.A., GUSELLA J.S.,
RA   HORVITZ H.R., BROWN R.H. JR.;
RL   NATURE 362:59-62(1993).
RN   [9]
RP   ERRATUM.
RX   MEDLINE; 93323981.
RA   ROSEN D.R., SIDDIQUE T., PATTERSON D., FIGLEWICZ D.A., SAPP P.,
RA   HENTATI A., DONALDSON D., GOTO J., O'REGAN J.P., DENG H.-X.,
RA   RAHMANI Z., KRIZUS A., MCKENNA-YASEK D., CAYABYAB A., GASTON S.M.,
RA   BERGER R., TANZI R.E., HALPERIN J.J., HERZFELDT B., VAN DEN BERGH R.,
RA   HUNG W.-Y., BIRD T., DENG G., MULDER D.W., SMYTH C., LAING N.G.,
RA   SORIANO E., PERICAK-VANCE M.A., HAINES J., ROULEAU G.A., GUSELLA J.S.,
RA   HORVITZ H.R., BROWN R.H. JR.;
RL   NATURE 364:362-362(1993).
RN   [10]
RP   VARIANTS FALS.
RX   MEDLINE; 93355289.
RA   DENG H.-X., HENTATI A., TAINER J.A., IQBAL Z., CAYABYAB A.,
RA   HUNG W.-Y., GETZOFF E.D., HU P., HERZFELDT B., ROOS R.P., WARNER C.,
RA   DENG G., SORIANO E., SMYTH C., PARGE H.E., AHMED A., ROSES A.D.,
RA   HALLEWELL R.A., PERICAK-VANCE M.A., SIDDIQUE T.;
RL   SCIENCE 261:1047-1051(1993).
RN   [11]
RP   VARIANT FALS THR-4.
RX   MEDLINE; 94235014.
RA   NAKANO R., SATO S., INUZUKA T., SAKIMURA K., MISHINA M., TAKAHASHI H.,
RA   IKUTA F., HONMA Y., FUJII J., TANIGUCHI N., TSUJI S.;
RL   BIOCHEM. BIOPHYS. RES. COMMUN. 200:695-703(1994).
RN   [12]
RP   VARIANT FALS GLU-7.
RX   MEDLINE; 95071364.
RA   HIRANO M., FUJII J., NAGAI Y., SONOBE M., OKAMOTO K., ARAKI H.,
RA   TANIGUCHI N., UENO S.;
RL   BIOCHEM. BIOPHYS. RES. COMMUN. 204:572-577(1994).
RN   [13]
RP   VARIANT FALS LYS-21.
RX   MEDLINE; 94348517.
RA   JONES C.T., SWINGER R.J., BROCK D.J.H.;
RL   HUM. MOL. GENET. 3:649-650(1994).
RN   [14]
RP   VARIANT FALS GLY-115.
RX   MEDLINE; 95187174.
RA   KOSTRZEWA M., BURCK-LEHMANN U., MUELLER U.;
RL   HUM. MOL. GENET. 3:2261-2262(1994).
RN   [15]
RP   VARIANTS FALS.
RX   MEDLINE; 95193785.
RA   PRAMATAROVA A., FIGLEWICZ D.A., KRIZUS A., HAN F.Y.,
RA   CEBALLOS-PICOT I., NICOLE A., DIB M., MEININGER V., BROWN R.H.,
RA   ROULEAU G.A.;
RL   AM. J. HUM. GENET. 56:592-596(1995).
RN   [16]
RP   VARIANT FALS ARG-93.
RX   MEDLINE; 95214771.
RA   ORRELL R., DE BELLEROCHE J., MARKLUND S., BOWE F., HALLEWELL R.;
RL   NATURE 374:504-505(1995).
RN   [17]
RP   VARIANT FALS ALA-90.
RA   ANDERSEN P.M., NILSSON P., ALA-HURULA V., KERAENEN M.-L.,
RA   TARVAINEN I., HALTIA T., NILSSON L., BINZER M., FORSGREN L.,
RA   MARKLUND S.L.;
RL   NATURE GENET. 10:61-66(1995).
CC   -!- FUNCTION: DESTROYS RADICALS WHICH ARE NORMALLY PRODUCED WITHIN THE
CC       CELLS AND ARE TOXIC TO BIOLOGICAL SYSTEMS.
CC   -!- CATALYTIC ACTIVITY: 2 PEROXIDE RADICAL + 2 H(+) = O(2) + H(2)O(2).
CC   -!- SUBUNIT: HOMODIMER.
CC   -!- SUBCELLULAR LOCATION: CYTOPLASMIC.
CC   -!- SIMILARITY: BELONGS TO THE CU-ZN SUPEROXIDE DISMUTASE FAMILY.
CC   -!- DISEASE: DEFECTS IN SOD1 ARE THE CAUSE OF AMYOTROPHIC LATERAL
CC       SCLEROSIS (ALS), A DEGENERATIVE DISORDER OF MOTOR NEURONS
CC       IN THE CORTEX, BRAINSTEM AND SPINAL CORD. ALS IS CHARACTERIZED
CC       WITH MUSCULAR WEAKNESS AND ATROPHY BEGINING IN THE HANDS AND
CC       SPREADING TO THE FOREARMS AND LEGS. MUSCLE FASCICULATIONS ARE
CC       COMMONLY VISIBLE. SENSORY ABNORMALITIES ARE ABSENT. DEATH USUALLY
CC       OCCURS WITHIN 2 TO 5 YEARS. THE FAMILIAL FORM OF ALS (FALS)
CC       ACCOUNTS FOR ABOUT 10% OF THE CASES AND IS TRANSMITTED IN AN
CC       AUTOSOMAL DOMINANT MANNER. THE MEAN AGE AT ONSET OF FALS IS
CC       45 YEARS.
DR   EMBL; L44139; G1237407; -.
DR   EMBL; L44135; G1237407; JOINED.
DR   EMBL; L44136; G1237407; JOINED.
DR   EMBL; L44137; G1237407; JOINED.
DR   EMBL; X02317; G36542; -.
DR   EMBL; K00065; G338276; -.
DR   EMBL; X01780; G36535; -.
DR   EMBL; X01781; E4991; -.
DR   EMBL; X01782; E4992; ALT_SEQ.
DR   EMBL; X01783; E4993; -.
DR   EMBL; X01784; E4994; ALT_SEQ.
DR   PIR; A00512; DSHUCZ.
```

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 26 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, among others, DR (Data bank Reference) lines that point to EMBL, 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 description of genetic disease(s) associated with that protein (OMIM), the 3D structure (PDB) or the pattern specific for that family of proteins (PROSITE).

RECENT DEVELOPMENTS

Model organisms

We have selected a number of organisms that are the target of genome sequencing and/or mapping projects and for which we intend to: (i) Be as complete as possible. All sequences available at a given time should be immediately included in SWISS-PROT.



```
DR   PIR; A23046; A23046.
DR   PIR; A22703; A22703.
DR   PIR; JX0055; JX0055.
DR   PDB; 1SOS; 31-JUL-94.
DR   PDB; 1SPD; 30-APR-94.
DR   PDB; 4SOD; 30-APR-94.
DR   SWISS-2DPAGE; P00441; HUMAN.
DR   AARHUS/GHENT-2DPAGE; 4127; IEF.
DR   MIM; 147450; -.
DR   MIM; 105400; -.
DR   PROSITE; PS00087; SOD_CU_ZN_1.
DR   PROSITE; PS00332; SOD_CU_ZN_2.
KW   OXIDOREDUCTASE; COPPER; ZINC; ACETYLATION; 3D-STRUCTURE;
KW   AMYOTROPHIC LATERAL SCLEROSIS; DISEASE MUTATION.
FT   INIT_MET      0      0
FT   MOD_RES       1      1       ACETYLATION.
FT   METAL        46     46       COPPER (BY SIMILARITY).
FT   METAL        48     48       COPPER (BY SIMILARITY).
FT   METAL        63     63       COPPER AND ZINC (BY SIMILARITY).
FT   METAL        71     71       ZINC (BY SIMILARITY).
FT   METAL        80     80       ZINC (BY SIMILARITY).
FT   METAL        83     83       ZINC (BY SIMILARITY).
FT   METAL       120    120       COPPER (BY SIMILARITY).
FT   DISULFID     57    146       BY SIMILARITY.
FT   VARIANT       4      4       A -> T (IN FALS).
FT   VARIANT       4      4       A -> V (IN FALS).
FT   VARIANT       7      7       V -> E (IN FALS).
FT   VARIANT      21     21       E -> K (IN FALS).
FT   VARIANT      37     37       G -> R (IN FALS).
FT   VARIANT      38     38       L -> V (IN FALS).
FT   VARIANT      41     41       G -> S (IN FALS).
FT   VARIANT      41     41       G -> D (IN FALS).
FT   VARIANT      43     43       H -> R (IN FALS).
FT   VARIANT      85     85       G -> R (IN FALS).
FT   VARIANT      90     90       D -> A (IN FALS; DOES NOT SEEM TO BE
FT                                LINKED WITH A DECREASE IN ACTIVITY).
FT   VARIANT      93     93       G -> C (IN FALS).
FT   VARIANT      93     93       G -> A (IN FALS).
FT   VARIANT      93     93       G -> R (IN FALS; 30% OF WILDTYPE
FT                                ACTIVITY).
FT   VARIANT     100    100       E -> G (IN FALS).
FT   VARIANT     106    106       L -> V (IN FALS).
FT   VARIANT     113    113       I -> T (IN FALS).
FT   VARIANT     115    115       R -> G (IN FALS).
FT   VARIANT     139    139       N -> K (IN FALS).
FT   VARIANT     144    144       L -> F (IN FALS).
FT   VARIANT     148    148       V -> G (IN FALS).
FT   VARIANT     149    149       I -> T (IN FALS).
FT   CONFLICT     17     17       I -> S (IN REF. 3).
FT   CONFLICT     98     98       S -> V (IN REF. 3).
FT   STRAND        4      9
FT   STRAND       15     21
FT   STRAND       30     33
FT   STRAND       36     36
FT   STRAND       41     48
FT   TURN         54     60
FT   STRAND       63     63
FT   STRAND       85     89
FT   HELIX        91     93
FT   STRAND       97     99
FT   TURN        113    114
FT   STRAND      116    120
FT   HELIX       132    137
FT   STRAND      143    148
FT   STRAND      150    151
SQ   SEQUENCE   153 AA;  15804 MW;  111991 CN;
     ATKAVCVLKG DGPVQGIINF EQKESNGPVK VWGSIKGLTE GLHGFHVHEF GDNTAGCTSA
     GPHFNPLSRK HGGPKDEERH VGDLGNVTAD KDGVADVSIE DSVISLSGDH CIIGRTLVVH
     EKADDLGKGG NEESTKTGNA GSRLACGVIG IAQ
//
```

Figure 1. A sample entry from SWISS-PROT.

This also includes sequence corrections and updates. (ii) Provide a higher level of annotations. (iii) Cross-references to specialized database(s) that contain, among other data, some genetic information about the genes that code for these proteins. (iv) Provide specific indices or documents.

The organisms currently selected are: *Arabidopsis thaliana* (mouse-ear cress), *Bacillus subtilis*, *Caenorhabditis elegans* (worm), *Candida albicans*, *Dictyostelium discoideum* (slime mold), *Drosophila melanogaster* (fruit fly), *Escherichia coli*, *Haemophilus influenzae*, *Homo sapiens* (human), *Mycobacterium tuberculosis*, *Mycoplasma genitalium*, *Saccharomyces cerevisiae* (budding yeast), *Salmonella typhimurium*, *Schizosaccharomyces pombe* (fission yeast) and *Sulfolobus solfataricus* (Table 1).

Collectively these organisms represent about a third of the total number of sequence entries in SWISS-PROT. In the last few months we included in SWISS-PROT fully annotated versions of the protein sequence entries encoded on the complete genome of *M.genitalium* as well as entries originating from the full sequence of yeast chromosomes VII, X, XIII and XIV. We plan to finish annotating all of the remaining yeast sequences (mainly from chromosomes IV, XII, XV and XVI) in early 1997.

Documentation files

SWISS-PROT is distributed with a large number of documentation files. Some of these files have been available for a long time (the user manual, release notes, the various indices for authors, citations, keywords, etc.), but many have been created recently and we are continually adding new files. Table 2 list all the documents that are currently available or that will be added in the next few months.

New cross-references

We have recently added cross-references that link SWISS-PROT to the following databases: (i) the Harefield Hospital 2D gel protein databases (4) prepared under the supervisation of Mike Dunn; and (ii) the Maize genome 2D Electrophoresis database (MAIZE-2DPAGE).

Currently, SWISS-PROT is linked to 26 different databases and has consolidated its role as the major focal points of biomolecular databases interconnectivity. In release 34, there were an average of 3.3 cross-references for each sequence entry.

Table 1. Model organisms in SWISS-PROT

Organism	Database	Index file	Number of sequences
A.thaliana	None yet	In preparation	562
B.subtilis	SubtiList	SUBTILIS.TXT	1783
C.albicans	None yet	CALBICAN.TXT	124
C.elegans	WormPep	CELEGANS.TXT	1208
D.discoideum	DictyDB	DICTY.TXT	265
D.melanogaster	FlyBase	In preparation	910
E.coli	EcoGene	ECOLI.TXT	3605
H.influenzae	HiDB	HAEINFLU.TXT	1591
H.sapiens	MIM	MIMTOSP.TXT	4000
M.genitalium	MgDB	In preparation	425
M.tuberculosis	None yet	In preparation	474
S.cerevisiae	LISTA	YEAST.TXT	4340
S.typhimurium	StyGene	SALTY.TXT	616
S.pombe	None yet	POMBE.TXT	955
S.acidocaldarius	None yet	None yet	42

Table 2. List of documents available in SWISS-PROT

File name	Description
userman .txt	User manual
relnotes.txt	Release notes
submit.txt	Submission of sequence data to SWISS-PROT
shortdes.txt	Short description of entries in SWISS-PROT
jourlist.txt	List of abbreviations for journals cited
keywlist.txt	List of keywords in use
speclist.txt	List of organism identification codes
tisslist.txt	List of tissues (in RC line)*
experts.txt	List of on-line experts for PROSITE and SWISS-PROT
acindex.txt	Accession number index
autindex.txt	Author index
citindex.txt	Citation index
keyindex.txt	Keyword index
speindex.txt	Species index
7tmrlist.txt	List of 7-transmembrane G-linked receptors entries
aatrnasy.txt	List of aminoacyl-tRNA synthetases
allergen.txt	Nomenclature and index of allergen sequences
calbican.txt	Index of *C.albicans* entries and their corresponding gene designations*
cdlist.txt	CD nomenclature for surface proteins of human leucocytes
celegans.txt	Index of *C.elegans* entries and corresponding gene designations and WormPep cross-references
dicty.txt	Index of *D.discoideum* entries and corresponding gene designations and DictyDB cross-references
ec2dtosp.txt	Index of *E.coli* Gene-protein database entries referenced in SWISS-PROT
ecoli.txt	Index of *E.coli* K12 chromosomal entries and corresponding EcoGene cross-references
embltosp.txt	Index of EMBL Database entries referenced in SWISS-PROT
extradom.txt	Nomenclature of extracellular domains
glycosid.txt	Index of glycosyl hydrolases classified by families on the basis of sequence similarities
haeinflu.txt	Index of *H.influenzae* RD chromosomal entries
hoxlist.txt	Vertebrate homeobox proteins: nomenclature and index
humchr20.txt	Index of protein sequence entries encoded on human chromosome 20*
humchr21.txt	Index of protein sequence entries encoded on human chromosome 21
humchr22.txt	Index of protein sequence entries encoded on human chromosome 22
humchrx.txt	Index of protein sequence entries encoded on human chromosome X*
humchry.txt	Index of protein sequence entries encoded on human chromosome Y
mimtosp.txt	Index of MIM entries referenced in SWISS-PROT
nomlist.txt	List of nomenclature related references for proteins
pdbtosp.txt	Index of Brookhaven PDB entries referenced in SWISS-PROT
peptidas.txt	Classification of peptidase families and index of peptidase entries
plastid.txt	List of chloroplast and cyanelle encoded proteins
pombe.txt	Index of *S.pombe* entries in SWISS-PROT and corresponding gene designations
restric.txt	List of restriction enzyme and methylase entries
ribosomp.txt	Index of ribosomal proteins classified by families on the basis of sequence similarities*
salty.txt	Index of *S.typhimurium* LT2 chromosomal entries and corresponding StyGene cross-references
subtilis.txt	Index of *B.subtilis* 168 chromosomal entries and corresponding SubtiList cross-references
yeast.txt	Index of *S.cerevisiae* entries and corresponding gene designations
yeast1.txt	Yeast Chromosome I entries
yeast2.txt	Yeast Chromosome II entries
yeast3.txt	Yeast Chromosome III entries
yeast5.txt	Yeast Chromosome V entries
yeast6.txt	Yeast Chromosome VI entries
yeast7.txt	Yeast Chromosome VII entries*
yeast8.txt	Yeast Chromosome VIII entries
yeast9.txt	Yeast Chromosome IX entries
yeast10.txt	Yeast Chromosome X entries*
yeast11.txt	Yeast Chromosome XI entries
yeast14.txt	Yeast Chromosome XIV entries*

*Documents that have been created since last year.

TrEMBL, A COMPUTER ANNOTATED SUPPLEMENT TO SWISS-PROT

Introduction

Ongoing genome sequencing and mapping projects have dramatically increased the number of protein sequences to be incorporated into SWISS-PROT. Since we do not want to dilute the quality standards of SWISS-PROT by incorporating sequences into SWISS-PROT without proper sequence analysis and annotation, we cannot speed up the incorporation of new incoming data indefinitely. However, as we also want to make sequences available as fast as possible, we introduced with SWISS-PROT release 34, TrEMBL (TRanslation of EMBL nucleotide sequence database), a supplement to SWISS-PROT. TrEMBL consists of computer-annotated entries in SWISS-PROT-like format derived from the translation of all coding sequences (CDS) in the EMBL nucleotide sequence database, except for CDS already included in SWISS-PROT.

The production of TrEMBL has emphasized the importance of linking not only to a whole EMBL nucleotide sequence entry but to linking within that entry at the CDS feature level. This linking has now been achieved by using the ‘PID’, the Protein IDentification number found in the ‘/db_xref’ qualifier tagged to every CDS in the EMBL nucleotide sequence database. The DR lines of SWISS-PROT and TrEMBL entries pointing to an EMBL database entry are now citing the EMBL AC number as primary identifier and the PID as secondary identifier. In all cases where a ‘PID’ is already integrated into SWISS-PROT a ‘/db_xref’ qualifier citing the corresponding SWISS-PROT entry is added to the EMBL nucleotide sequence database CDS labelled with this ‘PID’. In the remaining cases a ‘/db_xref’ qualifier is pointing to the corresponding TrEMBL entry. This approach enables us to point precisely from a given SWISS-PROT entry to one of potentially many CDS in the corresponding EMBL entry and vice versa.

This change will allow the development of software tools that automatically retrieve the part of a nucleotide sequence entry that codes for a specific protein. This will be especially useful in the context of World-Wide Web as it will render obsolete the current situation where, for example, one needs to retrieve the complete sequence of a yeast chromosome when one wants the nucleotide sequence coding for a specific protein encoded on that chromosome.

Current status

The translation of all CDS in the EMBL Nucleotide Sequence Database release 48 (September 1996) resulted in the creation of 199 000 TrEMBL preentries. Around 80 000 of these preentries were already as sequence reports in SWISS-PROT and excluded from TrEMBL. Then the remaining ~119 000 sequence entries have been automatically merged whenever possible to reduce redundancy in TrEMBL. This step led to ~110 000 TrEMBL entries. We have split TrEMBL in two main sections; SP-TrEMBL and REM-TrEMBL.

SP-TrEMBL (SWISS-PROT TrEMBL) contains the entries (~85 000) which should be incorporated into SWISS-PROT. SWISS-PROT accession numbers have been assigned to these entries. SP-TrEMBL is partially redundant against SWISS-PROT, since ~40 000 of these entries are only additional sequence reports of proteins already in SWISS-PROT. We will try to merge these sequence reports as fast as possible with the already existing SWISS-PROT entries for these proteins, so as to make SWISS-PROT and TrEMBL completely non-redundant.

For SP-TREMBL to act as a computer-annotated supplement to SWISS-PROT, new procedures have been introduced whereby valuable annotation has been added automatically. First, all TrEMBL entries are scanned for all PROSITE (5) patterns compatible with their taxonomic range. The results are added to the annotator’s section of the TrEMBL entry that is not visible to the public. Among all of the patterns, some of them are known to be very reliable (i.e. no known false positive). These are used to enhance the information content of the DE, CC, DR and KW fields by adding information about the potential function of the protein, metabolic pathways, active sites, cofactors, binding sites, domains, subcellular location and other annotation to the entry whenever appropriate. We also make use of the ENZYME database (6), using the EC number as a reference point, to generate standardized description lines for enzyme entries and to allow information such as catalytic activity, cofactors and relevant keywords to be taken from ENZYME and to be added automatically to SP-TrEMBL entries.

Furthermore we make use of specialized genomic databases like FlyBase (7) to parse information like the correct gene nomenclature and cross-references to these databases into TrEMBL entries.

REM-TrEMBL (REMaining TrEMBL) contains the entries (~20 000) that we do not want to include in SWISS-PROT. This section is organized in five subsections:

(i) Most REM-TrEMBL entries are immunoglobulins and T-cell receptors. We stopped entering immunoglobulins and T-cell receptors into SWISS-PROT, because we only want to keep the germ line gene derived translations of these proteins in SWISS-PROT and not all known somatic recombinated variations of these proteins. At the moment there are >12 000 immunoglobulins and T-cell receptors in TrEMBL. We would like to create a specialized database dealing with these sequences as a further supplement to SWISS-PROT and keep only a representative cross-section of these proteins in SWISS-PROT.

(ii) Another category of data which will not be included in SWISS-PROT are synthetic sequences. Again, we do not want to leave these entries in TrEMBL. Ideally one should build a specialized database for artificial sequences as a further supplement to SWISS-PROT.

(iii) Fragments with less than eight amino acids.

(iv) Coding sequences captured from patent applications. A thorough survey of these entries have shown that apart for a small minority (which have already been integrated in SWISS-PROT), most of these sequence contains either erroneous data or concern artificially generated sequences outside the scope of SWISS-PROT.

(v) The last subsection consists of CDS translations where we have strong evidence to believe that these CDS are not coding for real proteins.

PRACTICAL INFORMATION

Content of the current release

Currently (October 1996) SWISS-PROT contains ~60 000 sequence entries, comprising 21 million amino acids abstracted from ~50 000 references. The data file (sequences and annotations) requires 120 Mb of disk storage space. The documentation and index files require ~40 Mb of disk space. No restrictions are placed on use or redistribution of the data.

How to obtain SWISS-PROT

SWISS-PROT is distributed on CD-ROM by the EMBL Outstation – The European Bioinformatics Institute (EBI) (2). 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 The EMBL Outstation – The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. Telephone: (+44 1223) 494 400; Telefax : (+44 1223) 494 468; Email: datalib@ebi.ac.uk .

Individual sequence entries can be obtained from the EBI network fileserver. 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@ebi.ac.uk 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:

EBI anonymous FTP server
Internet address: ftp.ebi.ac.uk (or 192.54.41.33)

NCBI Repository (National Library of Medicine, NIH, Washington, DC, USA)
Internet address: ncbi.nlm.nih.gov (or 130.14.20.1)

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

National Institute of Genetics (Japan) FTP server
Internet address: ftp2.ddbj.nig.ac.jp (or 133.39.3.6)

How to submit data to SWISS-PROT

To submit data to SWISS-PROT and for all enquiries regarding the submission of SWISS-PROT one should contact SWISS-PROT, The EMBL Outstation – The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. Telephone: (+44 1223) 494 462; Telefax: (+44 1223) 494 468; Email: datasubs@ebi.ac.uk (for submission), junker@ebi.ac.uk (for enquiries).

Interactive access to SWISS-PROT

The most efficient and user-friendly way to browse interactively in SWISS-PROT is to use the World-Wide Web (WWW) molecular biology server ExPASy (8) as well as the one developed by the EBI. WWW is a global information retrieval system merging the power of world-wide networks, hypertext and multimedia. Through hypertext links, it gives access to documents and information available on thousands of servers around the world. Using a WWW browser [such as Mosaic(TM), Netscape Navigator(TM) or Microsoft Internet Explorer(TM)], one has access to all the hypertext documents stored on the ExPASy and EBI servers (as well as many other WWW servers).

The ExPASy server was made available to the public in September 1993. In October 1996 a cumulative total of 8 million connections was attained. It may be accessed through its Uniform Resource Locator (URL - the addressing system defined in WWW), which is: http://expasy.hcuge.ch/ . The EBI server is accessible under: http://www.ebi.ac.uk/

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 EBI, NCBI and ExPASy servers, whose Internet addresses are listed above.

REFERENCES

1 Bairoch,A. and Apweiler,R. (1996) *Nucleic Acids Res*. **24**, 21–25.
2 Rodriguez-Tome,P., Stoehr,P.J., Cameron,G.N. and Flores,T.P. (1996) *Nucleic Acids Res*. **24**, 6–12.
3 Bairoch,A. (1996) SWISS-PROT protein sequence data bank user manual, Release 34 of October 1996.
4 Corbett,J.M., Wheeler,C.H., Baker,C.S., Yacoub,M.H. and Dunn,M.J. (1994) *Electrophoresis* **15**, 1459–1465.
5 Bairoch,A., Bucher,P. and Hofmann,K. (1996) *Nucleic Acids Res*. **24**, 189–196.
6 Bairoch,A. (1996) *Nucleic Acids Res*. **24**, 221–222.
7 Flybase consortium (1996) *Nucleic Acids Res*. **24**, 53–56.
8 Appel,R.D., Bairoch,A. and Hochstrasser,D.F. (1994) *Trends Biochem. Sci*. **19**, 258–260.