Nucleic Acids Research, 1996, Vol. 24, No. 1 21–25

The SWISS-PROT protein sequence data bank and its new 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, Hinxton Hall, Hinxton, Cambridge CB10 1RQ, UK

Received October 3, 1995; Revised and Accepted October 13, 1995

ABSTRACT

SWISS-PROT is a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domain structure, post-translational modifications, variants, etc), a minimal level of redundancy and a 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 seven additional databases; a variety of new documentation files; the creation of TREMBL, an unannotated 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 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; 2). The SWISS-PROT protein sequence data bank consists of sequence entries. Sequence entries are composed of different line 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 a description of the following items: (i) function(s) of the protein; (ii) post-translational modification(s), for example carbohydrates, phosphorylation, acetylation, GPI-anchor, etc.; (iii) domains and sites, for example calcium binding regions, ATP binding sites, zinc fingers, homeobox, kringle, etc.; (iv) secondary structure; (v) quaternary structure; (vi) similarities to other proteins; (vii) disease(s) associated with deficiency of the protein; (viii) 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', an approach which permits 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-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 24 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 data bank reference (DR) lines that point to EMBL, PIR, OMIM and PROSITE. In this particular example it is therefore possible to retrieve the nucleic acid sequence(s) that encodes that protein (EMBL), the description of genetic disease(s) associated with that protein (OMIM) or the pattern specific for that family of proteins (PROSITE).

* To whom correspondence should be addressed

```
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-SEP-1995 (REL. 32, 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., GETTOFF 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; 95071366.
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; 94340517.
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-HUROLA 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; X02317; G36542; -.
DR   EMBL; K00065; G338276; -.
DR   EMBL; X01780; HSSOD1G1.
DR   EMBL; X01781; HSSOD12.
DR   EMBL; X01782; HSSOD13.
DR   EMBL; X01783; HSSOD14.
DR   EMBL; X01784; HSSOD15.
DR   PIR; A00512; DSHUCZ.
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; 11TH EDITION.
DR   MIM; 105400; 11TH EDITION.
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.

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, including sequence corrections and updates); (ii) provide a higher level of annotation; (iii) cross-reference to specialized databases 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); *Dictyostelium discoideum* (slime mold); *Drosophila melanogaster* (fruit fly); *Escherichia coli*; *Haemophilus influenzae*; *Homo sapiens* (human); *Saccharomyces cerevisiae* (budding yeast); *Salmonella typhimurium*; *Schizosaccharomyces pombe* (fission yeast); *Sulfolobus solfataricus*. Details of the database entries for these organisms are given in Table 1.

Table 1. Organisms entered in the data bank

Organism	Database	Index file	Number of sequences
A.thaliana	None yet	In preparation	399
B.subtilis	SubtiList	subtilis.txt	1329
C.elegans	WormPep	celegans.txt	828
D.discoideum	DictyDB	dicty.txt	210
D.melanogaster	FlyBase	In preparation	761
E.coli	EcoGene	ecoli.txt	3423
H.influenzae	HiDB	haeinflu.txt	1499
H.sapiens	MIM	mimtosp.txt	3250
S.cerevisiae	LISTA	yeast.txt	3347
S.typhimurium	StyGene	salty.txt	602
S.pombe	None yet	pombe.txt	404
S.solfataricus	None yet	None yet	61

Collectively these organisms represent 30% of the total number of sequence entries in SWISS-PROT.

In the last few months we have included in SWISS-PROT fully annotated versions of the protein sequence entries encoded on the complete genome of *Haemophilus influenzae*, as well as entries originating from the full sequence of yeast chromosomes I, II, III, V, VI, VIII, IX and XI.

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 continuously 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 LISTA database of yeast (*Saccharomyces cerevisiae*) genes coding for proteins prepared under the supervision of Patrick Linder at the University of Geneva (4);

(ii) the *Saccharomyces* Genome Database (SGD or SacchDB) prepared under the supervision of Mike Cherry at Stanford University;

(iii) the Yeast Electrophoresis Protein Database (YEPD) prepared under the supervision of Jim Garrells from the Quest Protein Database Center of the Cold Spring Harbor Laboratory (5);

(iv) the StyGene section of the StySeq/StyMap integrated *Salmonella typhimurium* LT2 database prepared by Ken Rudd at the National Center for Biotechnology Information (NCBI);

(v) the SubtiList relational database for the *Bacillus subtilis* 168 genome prepared under the supervision of Ivan Moszer at the Pasteur Institute (6);

(vi) the database of Homology-derived Secondary Structure of Proteins (HSSP) prepared under the supervision of Chris Sander at the EMBL (7);

(vii) the transcription factor database (Transfac) developed by Edgar Wingender and Rainer Knueppel from the Gesellschaft fuer Biotechnologische Forschung mbH in Braunschweig (8).

Currently, SWISS-PROT is linked to 24 different databases and has consolidated its role as the major focal point of biomolecular database interconnectivity. In release 32 there were an average of 3.5 cross-references for each sequence entry.

TREMBL, an unannotated supplement to SWISS-PROT

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 without proper sequence analysis and annotation, we cannot speed up the incorporation of new incoming data indefinitely. However, as we also want to make the sequences available as fast as possible we will introduce with SWISS-PROT release 33 an unannotated 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 CDS already included in SWISS-PROT.

We name this supplement TREMBL (TRanslation from EMBL), since the translation tools used to create translations of the CDS are based on the program 'TREMBL' written by Thure Etzold at the EMBL in Heidelberg.

Translation of all CDS in the EMBL nucleotide sequence database release 44 resulted in the creation of 145 000 TREMBL pre-entries. Around 65 000 of these pre-entries were already present as sequence reports in SWISS-PROT and were excluded from TREMBL. The remaining ~80 000 sequence entries have been automatically merged whenever possible, to reduce redundancy in TREMBL. This step led to ~70 000 TREMBL entries, which supplement SWISS-PROT.

Table 2.

File name	Description
userman.txt	User manual
relnotes.txt	Release notes
submit.txt	Submission of sequence data to the SWISS-PROT data bank*
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
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 receptor entries
aatrnasy.txt	List of aminoacyl-tRNA synthetases*
allergen.txt	Nomenclature and index of allergen sequences*
cdlist.txt	CD nomenclature for surface proteins of human leucocytes
celegans.txt	Index of *Caenorhabditis elegans* entries and corresponding gene designations and WormPep cross-references
dicty.txt	Index of *Dictyostelium discoideum* entries and corresponding gene designations and DictyDB cross-references
ec2dtosp.txt	Index of *Escherichia coli* gene–protein database entries referenced in SWISS-PROT
ecoli.txt	Index of *Escherichia 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*
glycosyl.txt	Index of glycosyl hydrolases classified by families on the basis of sequence similarities
haeinflu.txt	Index of *Haemophilus influenzae* RD chromosomal entries*
hoxlist.txt	Vertebrate homeobox proteins: nomenclature and index
humchr21.txt	Index of protein sequence entries encoded on human chromosome 21*
humchr22.txt	Index of protein sequence entries encoded on human chromosome 22*
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 *Schizosaccharomyces pombe* entries in SWISS-PROT and corresponding gene designations*
restric.txt	List of restriction enzymes and methylases entries
ribosomp.txt	Index of ribosomal proteins classified by families on the basis of sequence similarities
salty.txt	Index of *Salmonella typhimurium* LT2 chromosomal entries and corresponding StyGene cross-references*
subtilis.txt	Index of *Bacillus subtilis* 168 chromosomal entries and corresponding SubtiList cross-references*
yeast.txt	Index of *Saccharomyces 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*
yeast8.txt	Yeast chromosome VIII entries*
yeast9.txt	Yeast chromosome IX entries*
yeast11.txt	Yeast chromosome XI entries

Documents created since last year are flagged with an asterisk.

We have split TREMBL into two main sections, SP-TREMBL and REM-TREMBL. SP-TREMBL (SWISS-PROT TREMBL) contains entries (~55 000) 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 ~30 000 of these SP-TREMBL 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. REM-TREMBL (REMaining TREMBL) contains those entries (~15 000) that we do not wish to include in SWISS-PROT. This section is organized into four subsections.

(i) Most REM-TREMBL entries are immunoglobulins and T-cell receptors. We have stopped entering immunoglobulins and T-cell receptors into SWISS-PROT, because we want to keep only germ line gene-derived translations of these proteins in SWISS-PROT and not all known somatic recombinant variations of these proteins. At the moment there are >10 000 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 is 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) A third subsection consists of fragments with less than seven amino acids.

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

The creation of TREMBL as a supplement to SWISS-PROT was not only for the purpose of producing a more complete and up to date protein sequence collection. We used this task to also achieve a deeper integration of the EMBL nucleotide sequence database with SWISS-PROT + TREMBL.

We used the PID, the Protein IDentification number found in the /db_xref qualifier tagged to every CDS in the EMBL nucleotide sequence database, as the ID of the TREMBL entries created from these CDS. In all 65 000 cases where an EMBL nucleotide sequence database CDS was already present as a sequence report in SWISS-PROT the SWISS-PROT DR lines of the corresponding SWISS-PROT entries have been updated by 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.

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 that part of a nucleotide sequence entry that codes for a specific protein. This will be especially useful in the context of the 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.

PRACTICAL INFORMATION

Content of the current release

Release 32.0 of SWISS-PROT (October 1995) contains 48 440 sequence entries, comprising 17 000 000 amino acids abstracted from ~43 000 references. The data file (sequences and annotations) requires 90 Mb disk storage space. The documentation and index files require ~30 Mb 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 subscription to and distribution of SWISS-PROT one should contact The EMBL Outstation–The European Bioinformatics Institute, Hinxton Hall, Hinxton, Cambridge CB10 1RQ, UK (tel +44 1223 494 400; fax +44 1223 494 468; email datalib@ebi.ac.uk).

Individual sequence entries can be obtained from the EBI file server. 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 query to the network address netserv@ebi.ac.uk

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 (ftp.ebi.ac.uk or 192.54.41.33);

NCBI Repository, National Library of Medicine, NIH, Washington, DC (ncbi.nlm.nih.gov or 130.14.20.1);

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

National Institute of Genetics (Japan) ftp server (ftp.nig.ac.jp or 133.39.16.66).

How to submit data to SWISS-PROT

To submit data to SWISS-PROT and for all enquiries regarding submission to SWISS-PROT one should contact SWISS-PROT, The EMBL Outstation–The European Bioinformatics Institute, Hinxton Hall, Hinxton, Cambridge CB10 1RQ, UK [tel. +44 1223 494 462; fax +44 1223 494 468; email datasubs@ebi.ac.uk (for submissions), 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 (9), as well as that developed by the EBI. WWW is a global information retrieval system merging the power of worldwide networks, hypertext and multimedia. Through hypertext links it gives access to documents and information available on thousands of servers around the world. To access a WWW server one needs a WWW browser. Popular browsers available for most computer platforms include Mosaic™, developed at the National Center for Supercomputing Applications (NCSA) of the University of Illinois at Champaign (obtainable by anonymous ftp from ftp.ncsa.uiuc.edu), and Netscape Navigator™, from Netscape Communications Corp. (available from ftp.netscape.com). Using a WWW browser 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. On August 1995 a cumulative total of 2 000 000 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, although weekly updates are also available. These updates are available by anonymous ftp. Three files are updated every week: new_seq.dat, containing all the new entries since the last full release; upd_seq.dat, containing the entries for which the sequence data has been updated since the last release; upd_ann.dat, containing 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 Boeckmann,B. (1994) *Nucleic Acids Res.*, **22**, 3578–3580.
2 Emmert,D.B., Stoehr,P.J., Stoesser,G. and Cameron,G.N. (1994) *Nucleic Acids Res.*, **22**, 3445–3449.
3 Bairoch,A. (1995) *SWISS-PROT Protein Sequence Data Bank User Manual*, Release 32, October.
4 Doelz,R., Mosse,M.-O., Slonimski,P.P., Bairoch,A. and Linder,P. (1994) *Nucleic Acids Res.*, **22**, 3459–3461.
5 Latter,G.I., Boutell,T., Monardo,P.J., Kobayashi,R., Futcher,B., McLaughlin,C.S. and Garrels,J.I. (1995) *Electrophoresis*, **16**, 1170–1174.
6 Moszer,I., Glaser,P. and Danchin,A. (1995) *Microbiology*, **141**, 261–268.
7 Sander,C. and Schneider,R. (1994) *Nucleic Acids Res.*, **22**, 3597–3599.
8 Wingender,E. (1994) *J. Biotechnol.*, **35**, 273–280.
9 Appel,R.D., Bairoch,A. and Hochstrasser,D.F. (1994) *Trends Biochem. Sci.*, **19**, 258–260.