38–42 Nucleic Acids Research, 1998, Vol. 26, No. 1

The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998

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 22, 1997; Accepted October 24, 1997

ABSTRACT

SWISS-PROT (http://www.expasy.ch/) 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, its domains structure, 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 improvements to TrEMBL, a computer annotated supplement to SWISS-PROT. TrEMBL 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 line types, each with their own format. For standardisation 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: (i) annotation, (ii) minimal redundancy and (iii) integration with other databases.

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. For example homodimer, heterotrimer, etc.
- 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 minimise 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 specialised data

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

```
ID   CD4L_HUMAN     STANDARD;      PRT;   261 AA.
AC   P29965;
DT   01-APR-1993 (REL. 25, CREATED)
DT   01-APR-1993 (REL. 25, LAST SEQUENCE UPDATE)
DT   01-FEB-1997 (REL. 35, LAST ANNOTATION UPDATE)
DE   CD40 LIGAND (CD40-L) (TNF-RELATED ACTIVATION PROTEIN) (TRAP) (T CELL
DE   ANTIGEN GP39) (CD154 ANTIGEN).
GN   CD40LG OR CD40L OR TRAP.
OS   HOMO SAPIENS (HUMAN).
OC   EUKARYOTA; METAZOA; CHORDATA; VERTEBRATA; TETRAPODA; MAMMALIA;
OC   EUTHERIA; PRIMATES.
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE; 93076854.
RA   GRAF D., KORTHAEUER U., MAGES H.W., SENGER G., KROCZEK R.A.;
RL   EUR. J. IMMUNOL. 22:3191-3194(1992).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE; 93049181.
RA   HOLLENBAUGH D., GROSMAIRE L., KULLAS C., CHALUPNY J.,
RA   BRAESCH-ANDERSEN S., NOELLE R., STAMENKOVIC I., LEDBETTER J.,
RA   ARUFFO A.;
RL   EMBO J. 11:4313-4321(1992).
RN   [3]
RP   SEQUENCE FROM N.A.
RX   MEDLINE; 93145330.
RA   SPRIGGS M.K., ARMITAGE R.J., STOCKBINE L., CLIFFORD K.N.,
RA   MACDUFF B.M., SATO T.A., MALISZEWSKI C.R., FANSLOW W.C.;
RL   CELL 72:291-300(1993).
RN   [4]
RP   SEQUENCE FROM N.A.
RX   MEDLINE; 93094757.
RA   SPRIGGS M.K., ARMITAGE R.J., STROCKBINE L., CLIFFORD K.N.,
RA   MACDUFF B.M., SATO T.A., MALISZEWSKI C.R., FANSLOW W.C.;
RL   J. EXP. MED. 176:1543-1550(1992).
RN   [5]
RP   SEQUENCE FROM N.A.
RX   MEDLINE; 93138085.
RA   GAUCHAT J.F.M., AUBRY J., MAZZEI G.J., LIFE P., JOMOTTE T., ELSON G.,
RA   BONNEFOY J.Y.;
RL   FEBS LETT. 315:259-266(1993).
RN   [6]
RP   SEQUENCE FROM N.A.
RA   SHIMADZU M., TERASAKI H., NINOMIYA R., SHIMIZU S., NUNOI H.,
RA   MATSUDA I.;
RL   SUBMITTED (FEB-1995) TO EMBL/GENBANK/DDBJ DATA BANKS.
RN   [7]
RP   X-RAY CRYSTALLOGRAPHY (2.0 ANGSTROMS) OF 116-261.
RX   MEDLINE; 96131874.
RA   KARPUSAS M., HSU Y.-M., WANG J.-H., THOMPSON J., LEDERMAN S.,
RA   CHESS L., THOMAS D.;
RL   STRUCTURE 3:1031-1039(1995).
RN   [8]
RP   VARIANTS HIGM1 ARG-36 AND GLY-140.
RX   MEDLINE; 93156839.
RA   KORTHAEUER U., GRAF D., MAGES H.W., BRIERE F., PADAYACHEE M.,
RA   MALCOLM S., UGAZIO A.G., NOTARANGELO L.D., LEVINSKY R.J.,
RA   KROCZEK R.A.;
RL   NATURE 361:539-541(1993).
RN   [9]
RP   VARIANT HIGM1 GLU-123.
RX   MEDLINE; 93156840.
RA   DISANTO J.P., BONNEFOY J.Y., GAUCHAT J.F., FISCHER A.,
RA   DE SAINT BASILE G.;
RL   NATURE 361:541-543(1993).
RN   [10]
RP   VARIANTS HIGM1 ARG-128; GLY-129 AND PRO-235.
RX   MEDLINE; 93145330.
RA   ARUFFO A., FARRINGTON M., HOLLENBAUGH D., LI X., MILATOVICH A.,
RA   NONOYAMA S., BAJORATH J., GROSMAIRE L.S., STENKAMP R., NEUBAUER M.,
RA   ROBERTS R.L., NOELLE R.J., LEDBETTER J.A., FRANCKE U., OCHS H.D.;
RL   CELL 72:291-300(1993).
RN   [11]
RP   VARIANTS HIGM1 PRO-155; ASP-211 AND VAL-227.
RX   MEDLINE; 93174270.
RA   ALLEN R.C., ARMITAGE R.J., CONLEY M.E., ROSENBLATT H., JENKINS N.A.,
RA   COPELAND N.G., BEDELL M.A., EDELHOFF S., DISTECHE C.M.,
RA   SIMONEAUX D.K., FANSLOW W.C., BELMONT J., SPRIGGS M.K.;
RL   SCIENCE 259:990-993(1993).
RN   [12]
RP   VARIANTS HIGM1 ALA-126; ARG-140 AND GLU-144.
RX   MEDLINE; 95233438.
RA   MACCHI P., VILLA A., STRINA D., SACCO M.G., MORALI F., BRUGNONI D.,
RA   GILIANI S., MANTUANO E., FASTH A., ANDERSSON B., ZEGERS B.J.M.,
RA   CAVAGNI G., REZNICK I., LEVY J., ZAN-BAR I., PORAT Y., AIRO P.,
RA   PLEBANI A., VEZZONI P., NOTARANGELO L.D.;
RL   AM. J. HUM. GENET. 56:898-906(1995).
RN   [13]
RP   VARIANTS HIGM1 ARG-36; CYS-140; SER-231; MET-254 AND GLY-227 DEL.
RA   NONOYAMA S., SHIMADZU M., TORU H., SEYAMA K., NUNOI H., NEUBAUER M.,
RA   YATA J.-I., OCH H.D.;
RL   HUM. GENET. 99:624-627(1997).
CC   -!- FUNCTION: MEDIATES B-CELL PROLIFERATION IN THE ABSENCE OF CO-
CC       STIMULUS AS WELL AS IGE PRODUCTION IN THE PRESENCE OF IL-4.
CC       INVOLVED IN IMMUNOGLOBULIN CLASS SWITCHING.
CC   -!- SUBUNIT: HOMOTRIMER.
CC   -!- SUBCELLULAR LOCATION: TYPE II MEMBRANE PROTEIN. ALSO EXISTS AS AN
CC       EXTRACELLULAR SOLUBLE FORM.
CC   -!- TISSUE SPECIFICITY: SPECIFICALLY EXPRESSED ON ACTIVATED CD4+
CC       T-LYMPHOCYTES.
CC   -!- DATABASE: NAME=CD40Lbase; NOTE=European CD40L defect database;
CC       WWW="http://www.expasy.ch/www/cd40lbase.html";
CC       FTP="ftp.expasy.ch/databases/cd40lbase".
CC   -!- DISEASE: DEFECTS IN CD40LG ARE THE CAUSE OF AN X-LINKED
CC       IMMUNODEFICIENCY WITH HYPER-IGM (HIGM1), AN IMMUNOGLOBULIN ISOTYPE
CC       SWITCH DEFECT CHARACTERIZED BY ELEVATED CONCENTRATIONS OF SERUM
CC       IGM AND DECREASED AMOUNTS OF ALL OTHER ISOTYPES. AFFECTED MALES
CC       PRESENT AT AN EARLY AGE (USUALLY WITHIN THE FIRST YEAR OF LIFE)
CC       RECURRENT BACTERIAL AND OPPORTUNISTIC INFECTIONS, INCLUDING
CC       PNEUMOCYSTIS CARINII PNEUMONIA AND INTRACTABLE DIARRHEA DUE TO
CC       CRYPTOSPORIDIUM INFECTION. DESPITE SUBSTITUTION TREATMENT WITH
CC       INTRAVENOUS IMMUNOGLOBULIN, THE OVERALL PROGNOSIS IS RATHER POOR,
CC       WITH A DEATH RATE OF ABOUT 10% BEFORE ADOLESCENCE.
CC   -!- SIMILARITY: BELONGS TO THE TUMOR NECROSIS FACTOR FAMILY.
DR   EMBL; X68550; G37270; -.
DR   EMBL; Z15017; G38484; -.
DR   EMBL; X67878; G38412; -.
DR   EMBL; L07414; G180124; -.
DR   EMBL; D31797; G1518170; -.
DR   EMBL; D31793; G1518170; JOINED.
DR   EMBL; D31794; G1518170; JOINED.
DR   EMBL; D31795; G1518170; JOINED.
DR   EMBL; D31796; G1518170; JOINED.
DR   PIR; S25684; S25684.
DR   PIR; S26694; S26694.
DR   PIR; S28017; S28017.
DR   PIR; S28852; S28852.
DR   PIR; JH0793; JH0793.
DR   HSSP; P27548; 1CDA.
DR   MIM; 308230; -.
DR   PROSITE; PS00251; TNF.
KW   CYTOKINE; TRANSMEMBRANE; GLYCOPROTEIN; SIGNAL-ANCHOR;
KW   DISEASE MUTATION.
FT   DOMAIN        1     22       CYTOPLASMIC (POTENTIAL).
FT   TRANSMEM     23     46       SIGNAL-ANCHOR (TYPE-II MEMBRANE PROTEIN).
FT   DOMAIN       47    261       EXTRACELLULAR (POTENTIAL).
FT   DISULFID    178    218       POTENTIAL.
FT   CARBOHYD    240    240       POTENTIAL.
FT   VARIANT      36     36       M -> R (IN HIGM1).
FT   VARIANT     123    123       A -> E (IN HIGM1).
FT   VARIANT     126    126       V -> A (IN HIGM1).
FT   VARIANT     128    129       SE -> RG (IN HIGM1).
FT   VARIANT     140    140       W -> C (IN HIGM1).
FT   VARIANT     140    140       W -> G (IN HIGM1).
FT   VARIANT     140    140       W -> R (IN HIGM1).
FT   VARIANT     144    144       G -> E (IN HIGM1).
FT   VARIANT     155    155       L -> P (IN HIGM1).
FT   VARIANT     211    211       T -> D (IN HIGM1).
FT   VARIANT     227    227       G -> V (IN HIGM1).
FT   VARIANT     227    227       MISSING (IN HIGM1).
FT   VARIANT     231    231       L -> S (IN HIGM1).
FT   VARIANT     235    235       A -> P (IN HIGM1).
FT   VARIANT     254    254       T -> M (IN HIGM1).
SQ   SEQUENCE   261 AA;  29273 MW;  DC2AD21F CRC32;
     MIRTYNQTSP RSAATGLPIS MKIFMYLLTV FLITQMIGSA LFAVYLHRRL DKIEDERNLH
     EDFVFMKTIQ RCNTGERSLS LLNCEEIKSQ FEGFVKDIML NKEETKKENS FEMQKGDQNP
     QIAAHVISEA SSKTTSVLQW AEKGYYTMSN NLVTLENGKQ LTVKRQGLYY IYAQVTFCSN
     REASSQAPFI ASLCLKSPGR FERILLRAAN THSSAKPCGQ QSIHLGGVFE LQPGASVFVN
     VTDPSQVSHG TGFTSFGLLK L
//
```

Figure 1. A sample entry from SWISS-PROT.

collections. SWISS-PROT is currently cross-referenced with 30 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:

- Be as complete as possible. All sequences available at a given time should be immediately included in SWISS-PROT. This also includes sequence corrections and updates.
- Provide a higher level of annotations.
- Cross-references to specialised database(s) that contain, among other data, some genetic information about the genes that code for these proteins.
- 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*, *Helicobacter pylori*, *Homo sapiens* (human), *Methanococcus jannaschii*, *Mus musculus* (mouse), *Mycobacterium tuberculosis*, *Mycoplasma genitalium*, *Saccharomyces cerevisiae* (budding yeast), *Salmonella typhimurium*, *Schizosaccharomyces pombe* (fission yeast) and *Sulfolobus solfataricus*.

Table 1 lists, for each of the model organisms, the name of the specialised database to which cross-references are available, the name of the SWISS-PROT index file and the number of sequences in SWISS-PROT.

Table 1. Model organisms in SWISS-PROT

Organism	Database	Index file	Number of sequences
A.thaliana	None yet	In preparation	652
B.subtilis	SubtiList	SUBTILIS.TXT	1854
C.elegans	WormPep	CELEGANS.TXT	1725
C.albicans	None yet	CALBICAN.TXT	167
D.discoideum	DictyDB	DICTY.TXT	272
D.melanogaster	FlyBase	FLY.TXT	1000
E.coli	EcoGene	ECOLI.TXT	4038
H.influenzae	HiDB	HAEINFLU.TXT	1669
H.pylori	HpDB	HPYLORI.TXT	250
H.sapiens	MIM	MIMTOSP.TXT	4600
M.jannaschii	MjDB	MJANNASC.TXT	1061
M.musculus	MGD	MGDTOSP.TXT	2952
M.tuberculosis	None yet	In preparation	698
M.genitalium	MgDB	MGENITAL.TXT	470
S.cerevisiae	SGD	YEAST.TXT	4744
S.typhimurium	StyGene	SALTY.TXT	680
S.pombe	None yet	POMBE.TXT	1042

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
tisslist.txt	List of tissues
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
bloodgrp.txt	Blood group antigen proteins [*]
calbican.txt	Index of Candida albicans entries in SWISS-PROT and their corresponding gene designations
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
fly .txt	Index of Drosophila entries and cross-references to FlyBase [*]
glycosid.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 homeotic Hox proteins: nomenclature and index
hpylori .txt	Index of Helicobacter pylori strain 26695 chromosomal entries [*]
humchr19.txt	Index of protein sequences encoded on human chromosome 19 [*]
humchr20.txt	Index of protein sequences encoded on human chromosome 20
humchr21.txt	Index of protein sequences encoded on human chromosome 21
humchr22.txt	Index of protein sequences encoded on human chromosome 22
humchrx .txt	Index of protein sequences encoded on human chromosome X
humchry .txt	Index of protein sequences encoded on human chromosome Y
metallo .txt	Classification of metallothioneins and index of the entries in SWISS-PROT [*]
mgdtosp .txt	Index of MGD entries referenced in SWISS-PROT [*]
mgenital.txt	Index of Mycoplasma genitalium strain G-37 chromosomal entries [*]
mjannasc.txt	Index of Methanococcus jannaschii entries [*]
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 their 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 Salmonella typhimurium LT2 chromosomal entries and corresponding StyGene cross-references
subtilis.txt	Index of Bacillus subtilis 168 chromosomal entries and corresponding SubtiList cross-references
upflist .txt	List and index of Uncharacterized Protein Families
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
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
yeast13 .txt	Yeast Chromosome XIII entries [*]
yeast14 .txt	Yeast Chromosome XIV entries
yeast15 .txt	Yeast Chromosome XV entries [*]

[*] Documents that have been created since last year.

Collectively these organisms represent ~40% of the total number of sequence entries in SWISS-PROT. We are currently attempting to finish the integration into SWISS-PROT of all the putative proteins from *E.coli*, *B.subtilis* and yeast.

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 lists all the documents that are currently available.

New cross-references

We have recently added cross-references that link SWISS-PROT to the Mouse Genome Database (MGD) (4), to the TIGR Microbial genome database (5) and to the Plant Gene Nomenclature Database (MENDEL) (6).

Currently, SWISS-PROT is linked to 30 different databases and has consolidated its role as the major focal points of biomolecular databases interconnectivity. In release 35, there is an average of 3.3 cross-references for each sequence entry.

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 in early 1997, 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 emphasised 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 accession 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.

Current status

In October 1997, TrEMBL release 5 was produced. Release 5 was based on the translation of all 277 000 CDS in the EMBL

Nucleotide Sequence Database release 52. Around 100 000 of these CDS were already sequence reports in SWISS-PROT and thus excluded from TrEMBL. The remaining ~177 000 sequence entries have been automatically merged whenever possible to reduce redundancy in TrEMBL. This step led to ~150 000 TrEMBL entries.

We have split TrEMBL in two main sections; SP-TrEMBL and REM-TrEMBL: SP-TrEMBL (SWISS-PROT TrEMBL) contains the entries (~120 000 in release 5) 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 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 (7) 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 (8), using the EC number as a reference point, to generate standardised 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 specialised genomic databases like MGD (4) and FlyBase (9) to parse information like the correct gene nomenclature and cross-references to these databases into TrEMBL entries.

REM-TrEMBL (REMaining TrEMBL) contains the entries (~30 000 in release 5) that we do not want to include in SWISS-PROT. This section is organised in five subsections:

1. 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 more than 15 000 immunoglobulins and T-cell receptors in TrEMBL. We would like to create a specialised 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.
2. 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 specialised database for artificial sequences as a further supplement to SWISS-PROT.
3. Fragments with less than eight amino acids.
4. 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.
5. 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 SWISS-PROT release

Currently (October 1997), SWISS-PROT contains ~68 500 sequence entries, comprising 24.8 million amino acids abstracted from ~56 000 references. The data file (sequences and annotations) requires 135 Mb of disk storage space. The documentation and index files require ~45 Mb of disk space. No restrictions are placed on use or redistribution of the data.

Interactive access to SWISS-PROT and TrEMBL

The most efficient and user-friendly way to browse interactively in SWISS-PROT or TrEMBL is to use the World-Wide Web (WWW) molecular biology server ExPASy (10,11) as well as the one developed by the EBI. The ExPASy Web server was made available to the public in September 1993. On October 1997 a cumulative total of 17 million connections was attained. It may be accessed through its Uniform Resource Locator (URL - the addressing system defined in WWW), which is: http://www.expasy.ch/

The EBI server is accessible under: http://www.ebi.ac.uk/

On both the ExPASy and the EBI Web servers, you can use the Sequence Retrieval System (SRS) (12) software package to query and retrieve sequence entries.

How to obtain the full SWISS-PROT and/or TrEMBL releases

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

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:

- ExPASy (Expert Protein Analysis System) server, University of Geneva, Switzerland
  Internet address: ftp.expasy.ch
- EBI anonymous FTP server
  Internet address: ftp.ebi.ac.uk
- NCBI Repository (National Library of Medicine, NIH, Washington D.C., USA)
  Internet address: ncbi.nlm.nih.gov
- National Institute of Genetics (Japan) FTP server
  Internet address: ftp2.ddbj.nig.ac.jp

How to submit data or updates/corrections to SWISS-PROT

To submit new sequence 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. Tel: +44 1223 494 462; Fax: +44 1223 494 468; Email: datasubs@ebi.ac.uk (for submission); junker@ebi.ac.uk (for enquiries).

To submit updates and/or corrections to SWISS-PROT you can either use the E-mail address: swiss-prot@expasy.ch or the WWW address: http://www.expasy.ch/sprot/sp_update_ form.html

Release frequency, weekly updates and non-redundant data sets

The present distribution frequency is four releases per year. Weekly updates are also available; these updates are available by anonymous FTP. For SWISS-PROT, 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, ExPASy and NCBI servers, whose Internet addresses are listed above.

Every week we produce a complete non-redundant protein sequence collection by providing three compressed files (these are in the directory '/databases/sp_tr_nrdb' on the ExPASy FTP server and in /pub/databases/sp_tr_nrdb on the EBI server): sprot.dat.Z, trembl.dat.Z and trembl_new.dat.Z.

This set of non-redundant files is especially important for two types of users:

(i) Managers of similarity search services. They can now provide what is currently the most comprehensive and non-redundant data set of protein sequences.

(ii) Anybody wanting to update their full copy of SWISS-PROT and TrEMBL at their own schedule without having to wait for full releases of SWISS-PROT or of TrEMBL.

REFERENCES

1 Bairoch,A. and Apweiler,R. (1997) *Nucleic Acids Res*. **25**, 31–36.
2 Stoesser,G., Sterk,P., Tuli,M.A., Stoehr,P.J. and Cameron,G.N. (1997) *Nucleic Acids Res*. **25**, 7–13. [See also this issue *Nucleic Acids Res.* (1998) **26**, 8–15].
3 Bairoch,A. (1997) SWISS-PROT protein sequence data bank user manual, Release 35 of October 1997.
4 Blake,J.A., Richardson,J.E., Davisson,M.T., Eppig,J.T. and the Mouse Genome Informatics Group (1997) *Nucleic Acids Res*. **25**, 85–91. [See also this issue *Nucleic Acids Res.* (1998) **26**, 130–137].
5 http://www.tigr.org/tdb/mdb/mdb.html
6 http://jiio6.bbsrc.ac.uk/
7 Bairoch,A., Bucher,P. and Hofmann,K. (1997) *Nucleic Acids Res*. **25**, 217–221.
8 Bairoch,A. (1996) *Nucleic Acids Res*. **24**, 221–222.
9 Flybase consortium (1997) *Nucleic Acids Res*. **25**, 63–66. [See also this issue *Nucleic Acids Res.* (1998) **26**, 85–88].
10 Appel,R.D., Bairoch,A. and Hochstrasser,D.F. (1994) *Trends Biochem. Sci*. **19**, 258–260.
11 Bairoch,A., Appel,R.D. and Peitsch,M.C. (1997) *PDB Newsletter* **81**, 5–7.
12 Etzold,T. and Argos,P. (1993) *Comput. Appl. Biosci*., **9**, 49–57.