The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999

Amos Bairoch* and Rolf Apweiler[1]

Swiss Institute of Bioinformatics, Centre Medical Universitaire, 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 7, 1998; Revised and Accepted October 28, 1998

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 high level of integration with other databases. Recent developments of the database include: cross-references to 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. The URLs for SWISS-PROT on the WWW are: http://www.expasy.ch/sprot and http://www.ebi.ac.uk/sprot

INTRODUCTION

SWISS-PROT (1) is an annotated protein sequence database, which was created at the Department of Medical Biochemistry of the University of Geneva and has been a collaborative effort of the Department and the European Molecular Biology Laboratory (EMBL), since 1987. SWISS-PROT is now an equal partnership between the EMBL and the newly created Swiss Institute of Bioinformatics (SIB). The EMBL activities are carried out by its Hinxton 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 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 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:

(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, homeoboxes, SH2 and SH3 domains, etc.

(iv) Secondary structure. For example α-helix, β-sheet, etc.

(v) Quaternary structure. For example homodimer, heterotrimer, etc.

(vi) Similarities to other proteins

(vii) Disease(s) associated with deficiencie(s) in the protein

(viii) Sequence conflicts, variants, etc.

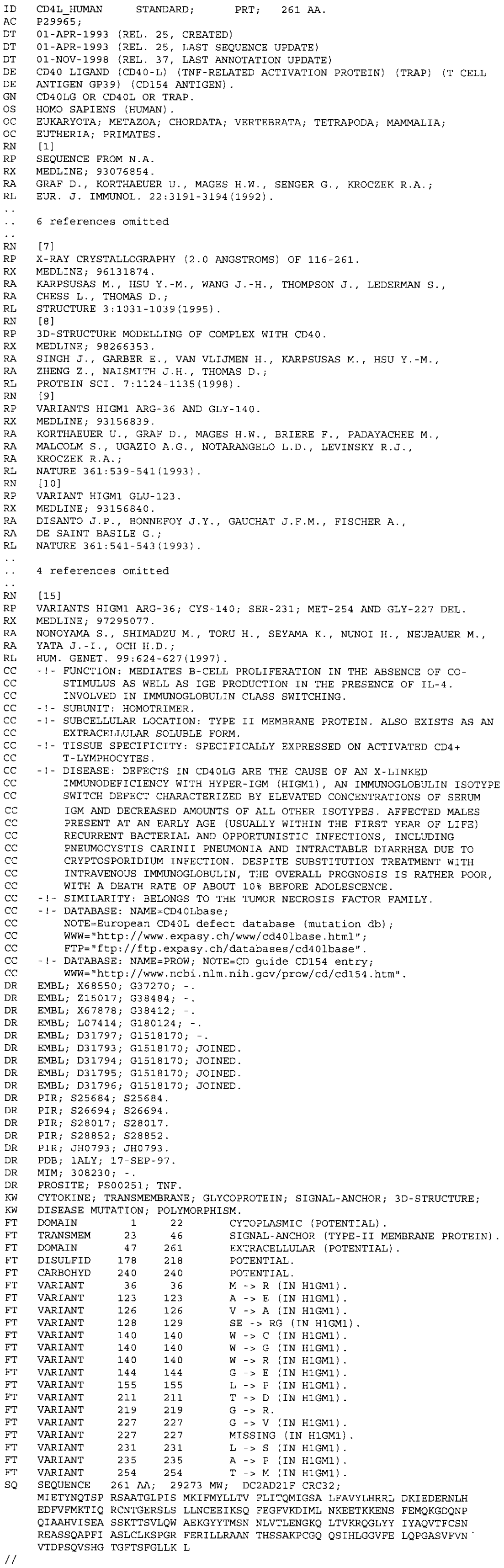
We try to include as much annotation information as possible in SWISS-PROT. To obtain this information we use, in addition to the publications reporting new sequence data, review articles to periodically update the annotations of families or groups of proteins. We also make use of external experts, who 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

*To whom correspondence should be addressed. Tel: +41 22 702 5477; Fax: +41 22 702 5502; Email: amos.bairoch@medecine.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-NOV-1998 (REL. 37, 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).
..
..   6 references omitted
..
RN   [7]
RP   X-RAY CRYSTALLOGRAPHY (2.0 ANGSTROMS) OF 116-261.
RX   MEDLINE; 96131874.
RA   KARPSUSAS 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   3D-STRUCTURE MODELLING OF COMPLEX WITH CD40.
RX   MEDLINE; 98266353.
RA   SINGH J., GARBER E., VAN VLIJMEN H., KARPSUSAS M., HSU Y.-M.,
RA   ZHENG Z., NAISMITH J.H., THOMAS D.;
RL   PROTEIN SCI. 7:1124-1135(1998).
RN   [9]
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   [10]
RP   VARIANT HIGM1 GLU-123.
RX   MEDLINE; 93156840.
RA   DISANTO J.P., BONNEFOY J.Y., GAUCHAT J.F.M., FISCHER A.,
RA   DE SAINT BASILE G.;
RL   NATURE 361:541-543(1993).
..
..   4 references omitted
..
RN   [15]
RP   VARIANTS HIGM1 ARG-36; CYS-140; SER-231; MET-254 AND GLY-227 DEL.
RX   MEDLINE; 97295077.
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   -!- 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.
CC   -!- DATABASE: NAME=CD40Lbase;
CC       NOTE=European CD40L defect database (mutation db);
CC       WWW="http://www.expasy.ch/www/cd40lbase.html";
CC       FTP="ftp://ftp.expasy.ch/databases/cd40lbase".
CC   -!- DATABASE: NAME=PROW; NOTE=CD guide CD154 entry;
CC       WWW="http://www.ncbi.nlm.nih.gov/prow/cd/cd154.htm".
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   PDB; 1ALY; 17-SEP-97.
DR   MIM; 308230; -.
DR   PROSITE; PS00251; TNF.
KW   CYTOKINE; TRANSMEMBRANE; GLYCOPROTEIN; SIGNAL-ANCHOR; 3D-STRUCTURE;
KW   DISEASE MUTATION; POLYMORPHISM.
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 H1GM1).
FT   VARIANT     123    123       A -> E (IN H1GM1).
FT   VARIANT     126    126       V -> A (IN H1GM1).
FT   VARIANT     128    129       SE -> RG (IN H1GM1).
FT   VARIANT     140    140       W -> C (IN H1GM1).
FT   VARIANT     140    140       W -> G (IN H1GM1).
FT   VARIANT     140    140       W -> R (IN H1GM1).
FT   VARIANT     144    144       G -> E (IN H1GM1).
FT   VARIANT     155    155       L -> P (IN H1GM1).
FT   VARIANT     211    211       T -> D (IN H1GM1).
FT   VARIANT     219    219       G -> R.
FT   VARIANT     227    227       G -> V (IN H1GM1).
FT   VARIANT     227    227       MISSING (IN H1GM1).
FT   VARIANT     231    231       L -> S (IN H1GM1).
FT   VARIANT     235    235       A -> P (IN H1GM1).
FT   VARIANT     254    254       T -> M (IN H1GM1).
SQ   SEQUENCE   261 AA;  29273 MW;  DC2AD21F CRC32;
     MIETYNQTSP 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.

exist between various sequencing reports, they are indicated in the feature table of the corresponding SWISS-PROT 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 collections. 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 codes 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. This also includes sequence corrections and updates;

(ii) Provide a higher level of annotation;

(iii) Cross-references to specialised 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*, *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 above 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.

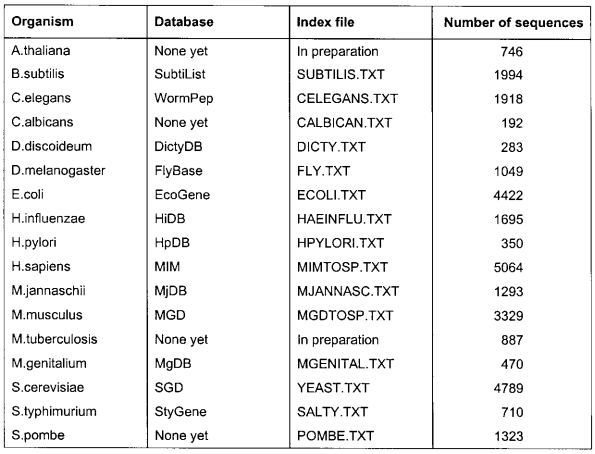
Collectively these organisms represent about 40% of the total number of sequence entries in SWISS-PROT. We are currently attempting to finish the integration into SWISS-PROT of all the putative proteins from *E.coli*, *B.subtilis*, *M.jannaschii* and yeast.

New model organisms will soon be added to the list, these will include at least one additional archebacterial species, a cyanobacteria (probably *Synechocystis* sp. PCC 6803) and a plant (probably maize).

Documentation files

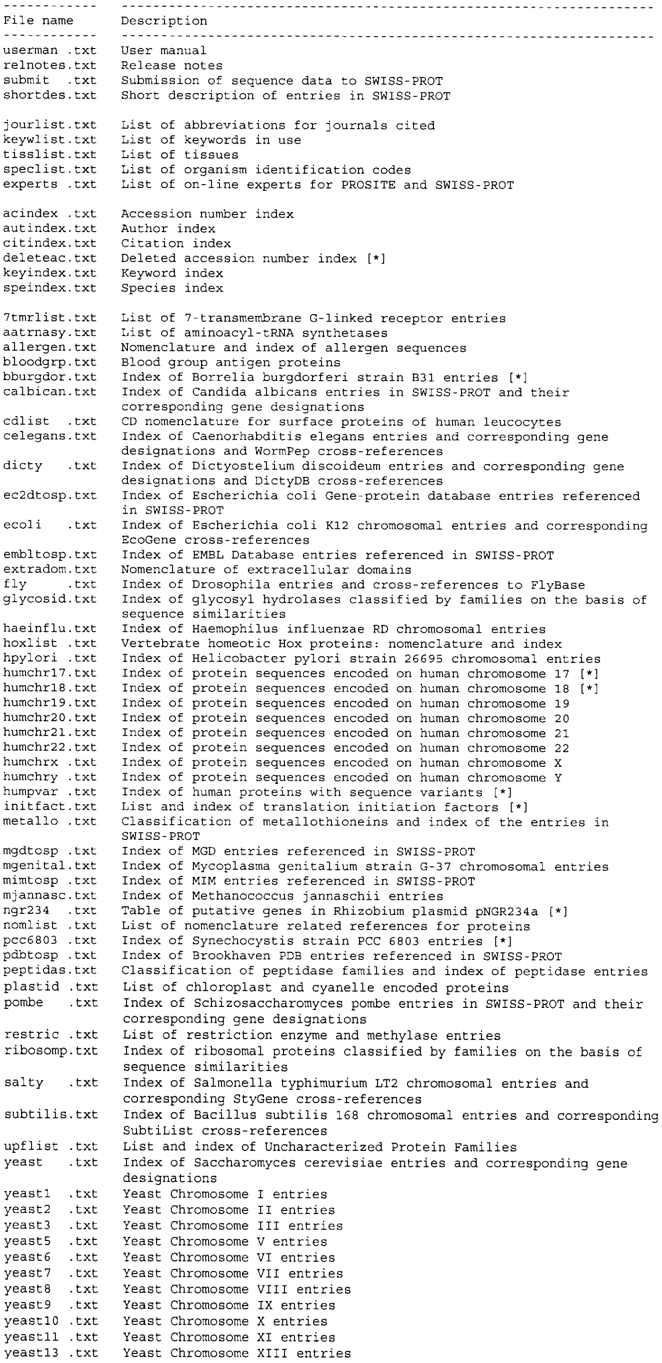
SWISS-PROT is distributed with a large number of documentation files. Some of these files have been available for a long time (the user manual, release notes, the various indices for authors,

Table 1. Model organisms in SWISS-PROT



Organism	Database	Index file	Number of sequences
A.thaliana	None yet	In preparation	746
B.subtilis	SubtiList	SUBTILIS.TXT	1994
C.elegans	WormPep	CELEGANS.TXT	1918
C.albicans	None yet	CALBICAN.TXT	192
D.discoideum	DictyDB	DICTY.TXT	283
D.melanogaster	FlyBase	FLY.TXT	1049
E.coli	EcoGene	ECOLI.TXT	4422
H.influenzae	HiDB	HAEINFLU.TXT	1695
H.pylori	HpDB	HPYLORI.TXT	350
H.sapiens	MIM	MIMTOSP.TXT	5064
M.jannaschii	MjDB	MJANNASC.TXT	1293
M.musculus	MGD	MGDTOSP.TXT	3329
M.tuberculosis	None yet	In preparation	887
M.genitalium	MgDB	MGENITAL.TXT	470
S.cerevisiae	SGD	YEAST.TXT	4789
S.typhimurium	StyGene	SALTY.TXT	710
S.pombe	None yet	POMBE.TXT	1323

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
deleteac.txt	Deleted accession number 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
bburgdor.txt	Index of Borrelia burgdorferi strain B31 entries [*]
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
humchr17.txt	Index of protein sequences encoded on human chromosome 17 [*]
humchr18.txt	Index of protein sequences encoded on human chromosome 18 [*]
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
humpvar .txt	Index of human proteins with sequence variants [*]
initfact.txt	List and index of translation initiation factors [*]
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
mimtosp .txt	Index of MIM entries referenced in SWISS-PROT
mjannasc.txt	Index of Methanococcus jannaschii entries
ngr234 .txt	Table of putative genes in Rhizobium plasmid pNGR234a [*]
nomlist .txt	List of nomenclature related references for proteins
pcc6803 .txt	Index of Synechocystis strain PCC 6803 entries [*]
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

*Documents that have been created since last year.

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 Pfam Protein families' database of alignments and HMMs (3).

Currently, SWISS-PROT is linked to 29 different databases and has consolidated its role as the major focal point of biomolecular databases interconnectivity. In release 36, there is an average of 3.5 cross-references for each sequence entry.

Implicit links

The 'explicit' links stored in the 'DR' lines of the flat file version of SWISS-PROT are supplemented by an additional category of links that we term 'implicit'. Implicit links are only available through the ExPASy WWW version of SWISS-PROT (see the practical information section) and are automatically generated by the server software. They further enhance the interoperability offered by SWISS-PROT by allowing users to navigate through additional and complementary information resources. There are two broad categories of implicit links as outlined below.

(i) There are many databases that have been developed in the last 10 years that are completely based on SWISS-PROT and offer a specific analytical view of the database. For example, the ProDom (4) and DOMO (5) databases describe an automatically derived domain view of each protein in SWISS-PROT; the ProtoMap (6) database is a hierarchical classification of all SWISS-PROT proteins. As these databases use SWISS-PROT primary accession numbers, it is possible to add implicit links from any SWISS-PROT entry to the corresponding entry in such an external database.

(ii) There are specialized databases that share with SWISS-PROT some form of unambiguous 'identifiers'. A typical example is 'GeneCard' (7), a database containing information on human genes. GeneCard can be accessed using the HUGO (Human Genome Organization) approved gene symbol of a relevant gene. Because SWISS-PROT also uses, as the first name listed on the GN line, the HUGO approved symbol, it is possible to automatically generate a link between SWISS-PROT and GeneCard for those human sequences that have been assigned a gene name.

While implicit links are quite useful, one must remember that:

(i) If one prints or saves an entry from the ExPASy server, it will contain lines that do not exist in the distributed version accessible through various software packages or from other Web servers.

(ii) In some cases such automatically generated links can fail. For example, a new SWISS-PROT entry may not yet have a corresponding entry in a derived database. Or, to take the example of GeneCard, it could happen that a gene symbol has not been 'synchronized' (either GeneCard has updated a gene name before SWISS-PROT or the reverse).

TrEMBL—a computer annotated supplement to SWISS-PROT

Introduction. Due to the increased data flow from genome projects to the sequence databases we face a number of challenges to our way of database annotation. Maintaining the high quality of sequence and annotation in SWISS-PROT requires careful sequence analysis and detailed annotation of every entry. This is the rate-limiting step in the production of SWISS-PROT. On one hand we do not wish to relax the high editorial standards of SWISS-PROT and it is clear that there is a limit to how much we can accelerate the annotation procedures. On the other hand, it is also vital that we make new sequences available as quickly as possible. To address this concern, we introduced in 1996 TrEMBL (Translation of EMBL nucleotide sequence database). TrEMBL consists of computer-annotated entries derived from the translation of all coding sequences (CDS) in the EMBL database, except for CDS already included in SWISS-PROT.

Current status. In August 1998, TrEMBL release 7 was produced. Release 7 was based on the translation of all 327 000 CDS in the EMBL Nucleotide Sequence Database release 55. Around 109 000 of these CDS were already as sequence reports in SWISS-PROT and thus excluded from TrEMBL. The remaining 218 000 sequence entries have been automatically merged whenever possible to reduce redundancy in TrEMBL. This step led to 193 860 TrEMBL entries.

We have split TrEMBL into two main sections; SP-TrEMBL and REM-TrEMBL: SP-TrEMBL (SWISS-PROT TrEMBL) contains the entries (165 420 in release 7) 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. For TrEMBL to act as a computer-annotated supplement to SWISS-PROT, new procedures have been introduced to remove redundancy and to automatically add highly reliable annotation.

The first step is the reduction of redundancy. All full-length proteins in SP-TrEMBL with the same sequence are merged into one entry. All fragment proteins with the same sequence from the same organism are merged, provided they do not belong to a highly variable category of proteins like MHC proteins or viral proteins. For all SWISS-PROT entries, the CRC32 checksums of all the different annotated sequence reports are calculated and compared with the checksums of all SP-TrEMBL entries. Identified matches are removed from SP-TrEMBL and integrated into the corresponding SWISS-PROT entries. Merging sub-fragments with full-length sequences and conflicting sequence reports about the same sequence further reduces the redundancy. Although these merging operations are automated, all merged entries are finally checked by biologists to avoid the merging of sequences from two different but highly similar genes into one entry. We use LASSAP (8) to identify sub-fragments to be merged with full-length sequences and to identify conflicting sequence reports about the same sequence. This new set of matches is removed from SP-TrEMBL and integrated into the corresponding SWISS-PROT or SP-TrEMBL entries.

The second post-processing step is the information enhancing process. All SP-TrEMBL entries are scanned for PROSITE patterns (9). If a matching pattern is found, a three-step procedure is used to reduce the number of false positive hits. Firstly, the taxonomic classification of the SP-TrEMBL entry must be within the known taxonomic range of the PROSITE pattern. For instance, a match of an *a priori* prokaryotic pattern against a human protein is regarded as false positive and filtered out. Secondly, the significance of the PROSITE pattern match is checked. This is done by a second check of the SP-TrEMBL sequence with a set of secondary patterns derived from the PROSITE pattern. These secondary patterns are computed with the eMotif algorithm (10). The PROSITE database contains a list of all SWISS-PROT proteins that are true members of the relevant protein family. For each pattern, the true positive sequences are aligned and fed into eMotif, which computes a nearly optimal set of regular expressions, based on statistical rather than biological evidence. We used a stringency of 10^{-9}, so that each eMotif pattern is expected to produce on random a false positive hit in 10^9 matches. Thirdly, in cases where a protein family is characterised by more than one PROSITE signature, all signatures must be found in the entry. For instance, bacterial rhodopsins have a signature for a conserved region in helix C and another signature for the retinal binding lysine. If an SP-TrEMBL entry matches only the helix-C-pattern, but not the retinal-binding pattern, it will not be regarded as a bacterial rhodopsin.

The raw PROSITE hits and all results of the confirmation steps are stored in a hidden section of the SP-TrEMBL entry, but only those hits that satisfy all confirmation conditions are made publicly visible in a DR PROSITE line. Approximately 35% of all SP-TrEMBL entries can be characterised by a PROSITE signature but only around 30% of all SP-TrEMBL entries are true positive matches. The characterisation based only on PROSITE patterns would lead to 10–20% of false positive assignments. The confirmation steps reduce the level of characterisation by nearly a third to 25%. At this stage, we achieve a level of less than 0.07% of false positive assignments.

Whenever an SP-TrEMBL entry is recognised by our procedures as a true member of a certain protein family, annotation about the potential function, active sites, cofactors, binding sites, domains, subcellular locations is added to the entry. The main source of the annotation is compiled by extracting the annotation that is common to all SWISS-PROT entries of the relevant protein family. For every protein family, a 'virtual SWISS-PROT entry' is created computationally, which is based on the specific annotation valid for all SWISS-PROT members of this family. If we are sure that a new SP-TrEMBL protein belongs to a certain family, we can immediately transfer the annotation of the virtual entry for this family. The annotation is flagged as annotation based on comparative analysis ('BY SIMIILARITY').

The 'virtual SWISS-PROT entries' have a far-reaching effect on SP-TrEMBL. For example, the virtual entry for Rubisco affects more than 2000 SP-TrEMBL entries. Therefore we developed a system to decompose these virtual entries into rules, which are stored in a relational database. This rule-based system enables us to express the membership criteria for each protein family in a formal language. Furthermore, subfamilies have been introduced to meet the SWISS-PROT standard more closely. For instance, the ribosomal protein L1 family is found in all known species, but the annotation added to SP-TrEMBL entries of this family obviously depends on the taxonomic kingdom. The description reads ‘50S RIBOSOMAL PROTEIN L1’ for prokaryotes, archaebacteria, chloroplasts and cyanelles, and ‘60S RIBOSOMAL PROTEIN L10A’ for non-chloroplast encoded proteins of eukaryotes.

We also use the ENZYME database (11), 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 use specialised databases like FlyBase (12) and MGD (13) to transfer information such as the correct gene nomenclature and cross-references to these databases into SP-TrEMBL entries. The automatic analysis and annotation of TrEMBL entries is redone and updated at every TrEMBL release.

REM-TrEMBL (REMaining TrEMBL) contains the entries (about 28 440 in release 7) that we do not want to include in SWISS-PROT. This section is organised into 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 more than 18 000 immunoglobulins and T-cell receptors in REM-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.

(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 specialised 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 has shown that apart from a small minority (which has already been integrated in SWISS-PROT), most of these sequences contain either erroneous data or concern artificially generated sequences outside the scope of SWISS-PROT.

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

PRACTICAL INFORMATION

The use of SWISS-PROT is free for academic users. However, we implemented in September 1998 a system of annual subscription fee for commercial users of the database. The SIB and the EMBL/EBI mandated a new company, Geneva Bioinformatics (GeneBio) (see http://www.genebio.com) to act as their representative for the purpose of concluding the necessary license agreements and levying the fees. The funds raised will be used at SIB and the EBI to bring SWISS-PROT up to date, to keep it up to date, and to further enhance its quality. Further information on this new system is available from the WWW addresses: http://www.expasy.ch/announce/ and http://www.ebi.ac.uk/news.html

Content of the current SWISS-PROT release

Currently (November 1998), SWISS-PROT contains ~76 000 sequence entries, comprising 27.2 million amino acids abstracted from ~60 000 references. The data file (sequences and annotations) requires 155 Mb of disk storage space. The documentation and index files require ~55 Mb of disk space.

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 (14) as well as the one developed by the EBI. The ExPASy Web server was made available to the public in September 1993. In October 1998 a cumulative total of 34 million connections was attained. It may be accessed through its URL, 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) (15) software package to query and retrieve sequence entries. The EBI and SIB also offer a range of search services to run Smith–Waterman, FASTA and BLAST sequence similarity searches against SWISS-PROT and TrEMBL.

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) 494 444; 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 anonymous FTP (File Transfer Protocol), from the following file servers: ftp.expasy.ch and ftp.ebi.ac.uk

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); datalib@ebi.ac.uk (for enquiries).

To submit updates and/or corrections to SWISS-PROT you can either use the Email 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 current 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:

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

For TrEMBL, a file containing all the new entries since the last full release (trembl_new.dat) is updated every week.

These files are available on the EBI and ExPASy servers, whose Internet addresses are listed above.

Every week we also 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.

Swiss-Shop

Swiss-Shop is an automated sequence alerting system which allows users to obtain, by Email, new sequence entries relevant to their field(s) of interest. Keyword-based and sequence/pattern-based requests are possible. Every time a weekly SWISS-PROT release is performed, all new database entries matching the user-specified search keywords or patterns and the entries showing sequence similarities to the user-specified sequence will be sent automatically to the user by Email. Swiss-Shop requests can be submitted to: http://www.expasy.ch/swisshop/

REFERENCES

1 Bairoch,A. and Apweiler,R. (1998) *Nucleic Acids Res*., **26**, 38–42.
2 Stoesser,G., Moseley,M.A., Sleep,J., McGowran,M., Garcia-Pastor,M. and Sterk,P. (1998) *Nucleic Acids Res*., **26**, 8–15.
3 Sonnhammer,E.L., Eddy,S.R., Birney,E., Bateman,A. and Durbin,R. (1998) *Nucleic Acids Res*., **26**, 320–322.
4 Corpet,F., Gouzy,J. and Kahn,D. (1998) *Nucleic Acids Res*., **26**, 323–326.
5 Gracy,J. and Argos,P. (1998) *Bioinformatics*, **14**, 164–187.
6 Yona,G., Linial,N., Tishby,N. and Linial,M. (1998) *ISMB*, **6**, 212–221.
7 Rebhan,M. and Prilusky,J. (1997) *Electrophoresis*, **18**, 2774–2780.
8 Glemet,E. and Codani,J.-J. (1997) *Comput. Applic. Biosci*., **13**, 137–143.
9 Bairoch,A., Bucher,P. and Hofmann,K. (1997) *Nucleic Acids Res*., **25**, 217–221.
10 Nevill-Manning,C.G., Wu,T.D. and Brutlag,D.L. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 5865–5871.
11 Bairoch,A. (1996) *Nucleic Acids Res*., **24**, 221–222.
12 Flybase Consortium (1998) *Nucleic Acids Res*., **26**, 85–88.
13 Blake,J.A., Eppig,J.T., Richardson,J.E., Davisson,M.T. and the Mouse Genome Informatics Group (1998) *Nucleic Acids Res*., **26**, 130–137.
14 Appel,R.D., Bairoch,A. and Hochstrasser,D.F. (1994) *Trends Biochem. Sci*., **19**, 258–260.
15 Etzold,T. and Argos,P. (1993) *Comput. Applic. Biosci*., **9**, 49–57.