What you can find in Swiss-Prot ?

Function of the protein; enzyme-specific information (catalytic activity, cofactors, metabolic pathway, regulation mechanisms); biologically relevant domains and sites; post-translational modifications (PTM); molecular weights determined by mass spectrometry; subcellular location(s) of the protein; tissue-specific expression of the protein; development-specific expression of the protein; secondary and quaternary structure information; splice isoforms; polymorphisms; similarities to other proteins; use of the protein as a pharmaceutical drug or in a biotechnological process; diseases associated with deficiencies in the protein; sequence conflicts; standardised nomenclature and controlled vocabularies; non-experimental qualifiers for predicted or propagated data; documentation files (`http://www.expasy.org/sprot/sp-docu.html`), etc.

What you can find through Swiss-Prot ?

Detailed expertise that goes beyond the scope of Swiss-Prot is made available via cross-references to specialised data collections such as the DDBJ/EMBL/GenBank nucleotide sequence databases, 2D and 3D protein structure databases, various protein domain and family characterisation databases, PTM databases, species-specific data collections, variant and disease databases; a list of cross-referenced databases is available at `http://www.expasy.org/cgi-bin/lists?dbxref.txt`.

Cross-references indicated in the DR lines are used to provide 'explicit' links to many databases; additionally, 'implicit' links are created on the fly by the ExPASy server. Unique and stable feature identifiers (FTId) in the FT lines allow referral and links to position-specific annotation.

Interactive access to Swiss-Prot and TrEMBL

The most efficient and user-friendly way to browse interactively in Swiss-Prot and/or TrEMBL is to use the Sequence Retrieval System (SRS) that is available on the ExPASy web server at `http://www.expasy.org/`, and its mirror sites, or on the EBI server at `http://www.ebi.ac.uk/`.

Tools

Links to various sequence analysis tools are provided from the ExPASy web server at `http://www.expasy.org/tools/` and the EBI server at `http://www.ebi.ac.uk/Tools/`.

How to obtain a local copy of Swiss-Prot and TrEMBL ?

Swiss-Prot and TrEMBL can be obtained by anonymous ftp from the ExPASy server `ftp.expasy.org` and EBI server `ftp.ebi.ac.uk/pub/` in the original Swiss-Prot flat file format, fasta format, and XML format. For detailed information see `http://www.expasy.org/sprot/download.html`.

Weekly updated complete non-redundant data sets for Swiss-Prot and TrEMBL are provided for ftp download. Swiss-Prot can also be obtained on CD-ROM from the EBI.

Submission of updates and new data

To submit **updates** and/or corrections to Swiss-Prot and for any enquiries you can either use the e-mail address `swiss-prot@expasy.org` or the WWW address `http://www.expasy.org/sprot/update.html`.

To submit **new sequence data** to Swiss-Prot, see `http://www.ebi.ac.uk/swissprot/Submissions/submissions.html` or contact us by e-mail at `datasubs@ebi.ac.uk`.

How to cite Swiss-Prot ?

If you want to cite Swiss-Prot in a publication, please use the following reference:
Boeckmann B., Bairoch A., Apweiler R., Blatter M.-C., Estreicher A., Gasteiger E., Martin M.J., Michoud K., O'Donovan C., Phan I., Pilbout S. and Schneider M.
The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**:365-370 (2003).

This document was written and designed by Brigitte Boeckmann from the Swiss Insitute of Bioinformatics and distributed by the Publications Committee of EMBnet.

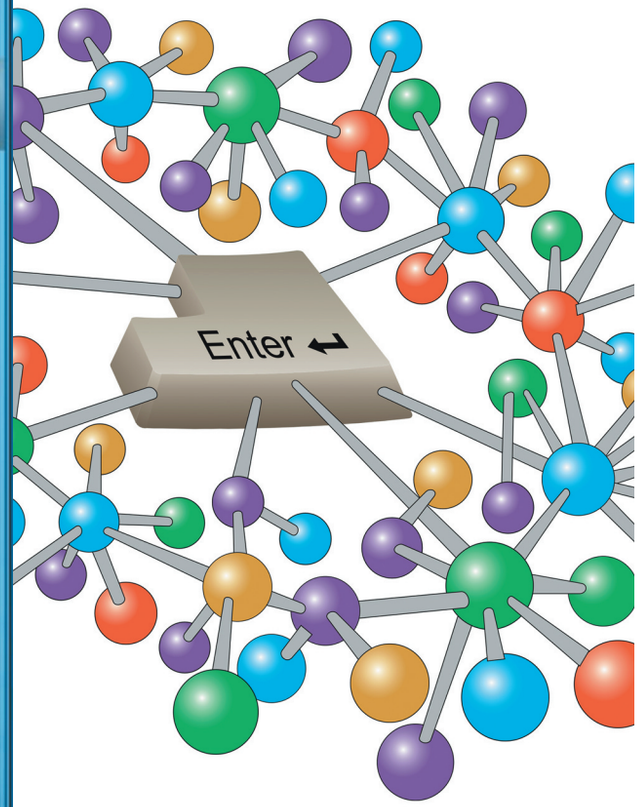
EMBnet - European Molecular Biology network - is a network of bioinformatics support centres situated primarily in Europe. Most countries have a national node which can provide training courses and other forms of help for users of bioinformatics software.

`http://www.embnet.org/`

If you have found this publication useful, please let us know.

A Quick Guide To Swiss-Prot and TrEMBL

EMBnet



A Quick Guide Swiss-Prot & TrEMBL

A Quick Guide To Swiss-Prot & TrEMBL

This is an introduction to the Swiss-Prot protein knowledgebase and its computer-annotated supplement TrEMBL. Swiss-Prot was established by Amos Bairoch in 1986 and is maintained collaboratively by the Swiss Institute of Bioinformatics (SIB) and the European Bioinformatics Institute (EBI). TrEMBL was created in 1996 by the EBI.

The Swiss-Prot knowledgebase is a high quality manually annotated and non-redundant protein sequence database, which brings together experimental results, computed features and scientific conclusions. Swiss-Prot provides annotated entries for all species, but concentrates on the annotation of entries from human and model organisms of distinct taxonomic groups to ensure the presence of high quality annotation for representative members of all protein families. Protein families and groups of proteins are regularly reviewed to keep up with current scientific findings.

TrEMBL is a computer-annotated supplement to Swiss-Prot, which strives to gather all protein sequences that are not yet represented in Swiss-Prot. A perpetually increasing level of automated annotation is incorporated into TrEMBL. The format of TrEMBL entries is Swiss-Prot-like. Manually annotated TrEMBL entries are moved to Swiss-Prot and keep the same accession numbers.

The Swiss-Prot, TrEMBL and PIR protein database activities have united to form the Universal Protein Knowledgebase (**UniProt**) consortium. The central database has two sections, corresponding to Swiss-Prot and TrEMBL.

The Swiss-Prot format

Each entry is composed of lines, beginning with a two-character line code. The format of the distinct line types is described in the user manual at `http://www.expasy.org/sprot/userman.html`.
Note: on some servers, Swiss-Prot entries are shown in the user-friendly view known as **NiceProt**.

Entry information: the entry name is indicated in the ID line, but the stable and unique identifier of an entry is the first (primary) accession number in the AC line.

```
ID   AHA1_HUMAN     STANDARD;      PRT;   338 AA.
AC   O95433; Q96IL6; Q9P060;
DT   16-OCT-2001 (Rel. 40, Created)
DT   16-OCT-2001 (Rel. 40, Last sequence update)
DT   15-SEP-2003 (Rel. 42, Last annotation update)
```

Name and origin of the protein (DE lines). Gene names and locus names are shown in the GN line. If the coding gene is not nuclear, the OG line indicates its origin. The taxonomy database of NCBI is cross-referenced in the OX line.

```
DE   Tetanus toxin precursor (EC 3.4.24.68) (Tentoxylysin)
DE   [Contains: Tetanus toxin light chain
DE   (Tetanus toxin chain L); Tetanus toxin heavy chain
DE   (Tetanus toxin chain H)].
GN   TETX OR CTP60.
OS   Clostridium tetani.
OG   Plasmid pE88, and Plasmid 75 Kbp.
OC   Bacteria; Firmicutes; Clostridia; Clostridiales;
OC   Clostridiaceae; Clostridium.
OX   NCBI_TaxID=1513;
```

References concern sequences, protein structure, function, PTMs, tissue-specific expression, variants, etc. The RP and RC lines store information relevant to the reference cited.

```
RN   [2]
RP   SEQUENCE FROM N.A. (ISOFORM 1), AND TISSUE SPECIFICITY.
RC   STRAIN=C57BL/6; TISSUE=Skeletal muscle;
RX   MEDLINE=21536607; PubMed=11679416;
RA   DeHaven J.E., Robinson K.A., Nelson B.A., Buse M.G.;
RT   ìA novel variant of glutamine:fructose-6-phosphate
RT   amidotransferase-1 (GFAT1) mRNA is selectively
RT   expressed in striated muscle.î;
RL   Diabetes 50:2419-2424(2001).
```

Comment blocks (CC) start with a topic that indicates the type of comment.

```
CC   -!- FUNCTION: Catalyzes the oxidative decarboxylation
CC       of glutaryl-CoA to crotonyl-CoA and CO(2) in the
CC       degradative pathway of L-lysine, L-hydroxylysine,
CC       and L-tryptophan metabolism. It uses electron
CC       transfer  avoprotein as its electron acceptor.
CC   -!- CATALYTIC ACTIVITY: Glutaryl-CoA + acceptor =
CC       crotonoyl-CoA + CO(2) + reduced acceptor.
CC   -!- COFACTOR: FAD.
CC   -!- PATHWAY: Degradative pathway of L-lysine,
CC       L-hydroxylysine, and L-tryptophan metabolism.
CC   -!- SUBUNIT: Homotetramer.
CC   -!- SUBCELLULAR LOCATION: Mitochondrial matrix.
CC   -!- ALTERNATIVE PRODUCTS:
CC       Event=Alternative splicing; Named isoforms=2;
CC       Name=Long;
CC         IsoId=Q92947-1; Sequence=Displayed;
CC       Name=Short;
CC         IsoId=Q92947-2; Sequence=VSP_000145;
CC         Note=Inactive;
CC   -!- TISSUE SPECIFICITY: The 2 isoforms have been found
CC       in  broblasts and liver.
CC   -!- DISEASE: Defects in GCDH are the cause of glutaric
CC       acidemia type I (GA-I) [MIM:231670]. GA-I is an
CC       autosomal recessive metabolic disorder
CC       characterized by progressive dystonia and athetosis
CC       due to gliosis and neuronal loss in the basal
CC       ganglia.
CC   -!- SIMILARITY: Belongs to the acyl-CoA dehydrogenase
CC       family.
```

Cross-references in the DR line allow links to many specialised databases via the database name and a unique identifer.

```
DR   EMBL; D55674; BAA09525.1; -.
DR   EMBL; M94630; AAA35781.1; ALT_SEQ.
DR   PDB; 1HD0; 18-MAY-00.
DR   SWISS-2DPAGE; Q14103; HUMAN.
DR   Genew; HGNC:5036; HNRPD.
DR   GK; Q14103; -.
DR   MIM; 601324; -.
DR   GO; GO:0005634; C:nucleus; TAS.
DR   GO; GO:0003723; F:RNA binding activity; TAS.
DR   GO; GO:0006401; P:RNA catabolism; TAS.
DR   InterPro; IPR000504; RNA_rec_mot.
DR   Pfam; PF00076; rrm; 2.
DR   SMART; SM00360; RRM; 2.
DR   PROSITE; PS50102; RRM; 2.
```

Keywords are controlled vocabulary, which summarise the information of an entry.

```
KW   Transport; Sugar transport; Outer membrane;
KW   Transmembrane; Porin; Signal; Plasmid; 3D-structure.
```

Features: More than 30 feature keys (e.g. SIGNAL, CHAIN) may refer to regions or positions in a sequence. Some feature types are associated with unique feature identifiers (FTId). Non-experimental qualifiers ('Potential', 'Probable' and 'By similarity') indicate the experimental status of a feature and may also be found in the CC lines.

```
FT   SIGNAL        1     27       POTENTIAL.
FT   CHAIN        28    672       INTER-ALPHA-TRYPSIN
FT                                 INHIBITOR HEAVY CHAIN H1.
FT   PROPEP      673    911       POTENTIAL.
FT   PEPTIDE     181    184       PHAGOCYTOSIS UPTAKE SIGNAL
FT                                 (POTENTIAL).
FT   DOMAIN      290    450       VWFA.
FT   DOMAIN      387    911       HYALURONAN BINDING.
FT   DISULFID    244    247       BY SIMILARITY.
FT   DISULFID    268    540       BY SIMILARITY.
FT   CARBOHYD    285    285       N-LINKED (GLCNAC...)
FT                                (COMPLEX).
FT                                 /FTId=CAR_000138.
FT   CARBOHYD    653    653       O-LINKED (GALNAC...).
FT                                 /FTId=CAR_000213.
FT   BINDING     672    672       CHONDROITIN 4-SULFATE,
FT                                 CROSS-LINK SITE.
FT   VARIANT     263    263       S -> T (IN dbSNP:1042777).
FT                                 /FTId=VAR_011873.
FT   VARIANT     595    595       Q -> R (in allele ITIH1*2
FT                                 and allele ITIH1*3).
FT                                 /FTId=VAR_004020.
FT   CONFLICT     51     51       V -> T (IN REF. 5).
```

Sequence information. The molecular weight is calculated from the sequence shown and does not consider any experimental findings. The 64-bit Cyclic Redundancy Check (CRC64) value facilitates the identification of identical sequences. The termination (//) line designates the end of an entry.

```
SQ   SEQUENCE   45 AA;  5140 MW;  3E6B661E0342CA01 CRC64;
     MMSCLILRIF ILIKEGVISM AQDIISTIGD LVKWIIDTVN KFTKK
//
```