

A Quick Guide To UniProtKB Swiss-Prot & TrEMBL

UniProt, the Universal Protein

Resource, is produced by the UniProt Consortium, a



collaboration between the Swiss Institute of Bioinformatics (SIB), the European Bioinformatics Institute (EBI) and the Georgetown University Medical Center's Protein Information Resource (PIR). The mission of UniProt is to support biological research by maintaining a high quality central database that serves as a stable, comprehensive and accurately annotated protein sequence knowledgebase, the **UniProtKB** database. UniProtKB comprises 2 sections: the Swiss-Prot Protein Knowledgebase and the TrEMBL Protein Database, which together give access to all known protein sequences.

UniProtKB/Swiss-Prot 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.

UniProtKB/TrEMBL is a computer-annotated supplement to Swiss-Prot, which strives to gather all protein sequences that are not yet represented in Swiss-Prot. The protein sequences are derived from the translation of coding sequences (CDS) submitted to the nucleic acid databases (i.e. EMBL). 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 then moved to Swiss-Prot and keep the same accession numbers.



The UniProtKB/Swiss-Prot format

Each sequence 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 a 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 AHSAL_HUMAN STANDARD; PRT; 338 AA.
AC O95433; Q96IL6; Q9P060;
DT 21-FEB-2001, integrated into UniProtKB/Swiss-Prot.
DT 01-MAY-1999, sequence version 1.
DT 07-FEB-2006, entry version 45.
```

Name and origin: Protein name, synonyms and abbreviations are indicated in the DE line, gene and locus names are shown in the GN line. Furthermore species and taxonomy information are available from this section.

```
DE Tetanus toxin precursor (EC 3.4.24.68) (Tentoxylysin)
DE [Contains:Tetanus toxin light chain (Tetanus toxin chain
DE L); Tetanus toxin heavy chain (Tetanus toxin chain H)].
GN Name=tetX; OrderedLocusNames=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, post-translational modifications, tissue-specific expression, variants, etc. The RP line specifies the information extracted from the reference.

```
RN [1]
RP NUCLEOTIDE SEQUENCE [MRNA], FUNCTION, TISSUE SPECIFICITY, AND
RP DEVELOPMENTAL STAGE.
RC STRAIN=Bristol N2;
RX PubMed=10662646;
RA Karashima T., Sugimoto A., Yamamoto M.;
RT «Caenorhabditis elegans homologue of the human azoospermia factor DAZ
RT is required for oogenesis but not for spermatogenesis.»;
RL Development 127:1069-1079(2000).
```

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

```
CC -!- FUNCTION: Catalyzes the oxidative decarboxylation of glutaryl-CoA
CC to crotonyl-CoA and CO(2) in the degradative pathway of L-lysine,
CC L-hydroxylysine, and L-tryptophan metabolism. It uses electron
CC transfer flavoprotein as its electron acceptor. The short isoform
CC is inactive.
CC -!- CATALYTIC ACTIVITY: Glutaryl-CoA + acceptor = crotonoyl-CoA +
CC CO(2) + reduced acceptor.
CC -!- COFACTOR: FAD.
CC -!- PATHWAY: Degradative pathway of L-lysine, L-hydroxylysine, and L-
CC 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 -!- TISSUE SPECIFICITY: The 2 isoforms have been found in fibroblasts
CC and liver.
CC -!- DISEASE: Defects in GCDH are the cause of glutaric acidemia type I
CC (GA-I) (MIM:231670). GA-I is an autosomal recessive metabolic
CC disorder characterized by progressive dystonia and athetosis due
CC to gliosis and neuronal loss in the basal ganglia. Macrocephaly is
CC often seen at birth. Patients with the disorder accumulate and
CC excrete glutaric, 3-hydroxyglutaric, and glutaconic acid.
CC -!- SIMILARITY: Belongs to the acyl-CoA dehydrogenase family.
CC -!- DATABASE: NAME=GeneReviews;
CC WWW=http://www.genetests.org/query?gene=GCDH*.
```

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

```
DR EMBL; D55674; BAA09525.1; -, mRNA.
DR EMBL; AF026126; AAC23474.1; -, Genomic_DNA.
DR EMBL; BC002401; AAH02401.1; -, mRNA.
DR EMBL; M94630; AAA35781.1; ALT_SEQ; mRNA.
DR PDB; 1HD0; NMR; A=98-172.
DR PDB; 1HD1; NMR; A=98-172.
DR SWISS-2DPAGE; Q14103; HUMAN.
DR Ensembl; ENSG00000138668; Homo sapiens.
DR HGNC; HGNC:5036; HNRPD.
DR H-InVdb; HIX0004333; -.
DR Reactome; Q14103; -.
DR MIM; 601324; -.
DR GO; GO:0003723; F:RNA binding; TAS.
DR InterPro; IPR012677; a_b_plait_nuc_bd.
```

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

```
KW 3D-structure; Carbohydrate metabolism; Diabetes mellitus;
KW Direct protein sequencing; Hormone; Pharmaceutical; Signal.
```

Features: More than 30 feature keys (e.g. SIGNAL, DISULFID) 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 20
FT PEPTIDE 21 55 Insulin-like 3 B chain.
FT /FTId=PRO_0000016140.
FT PROPEP 58 104 C peptide like (Potential).
FT /FTId=PRO_0000016141.
FT PEPTIDE 106 131 Insulin-like 3 A chain.
FT /FTId=PRO_0000016142.
FT DISULFID 34 116 Interchain (between B and A chains) (By
FT similarity).
FT DISULFID 46 129 Interchain (between B and A chains) (By
FT similarity).
FT DISULFID 115 120 By similarity.
FT VARIANT 24 24 A -> G.
FT /FTId=VAR_013231.
FT VARIANT 43 43 V -> L.
FT /FTId=VAR_013232.
FT VARIANT 49 49 P -> S (could be a rare polymorphism;
FT identified in a male with
FT undermasculinised genitalia and intra-
FT abdominal testes).
FT /FTId=VAR_013233..
```

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 ILIKGVISM AQDIISTIGD LWRWIIDTVN KPTKK
//
```

UniProtKB/Swiss-Prot specificity and utility

Data integrated into Swiss-Prot, including the protein sequence and current knowledge on each protein, are **manually checked and continuously updated**.

In order to have **minimal redundancy** and improve **sequence reliability**, all protein sequences encoded by a same gene are merged into a single entry.

A special emphasis is laid on the **annotation of biological events which generate protein diversity** that cannot be predicted at the genomic level. Alternative products (alternative splicing, RNA editing...) and post-translational modification

(PTMs) are extensively annotated. For additional information, see Boeckmann et al., *C. R. Biol.* **328**:882-899 (2005).

Swiss-Prot is highly cross-referenced and thus a **central hub for biological data**, a platform linking together all the protein resources.

Swiss-Prot is particularly suitable for similarity searches, protein identification (proteomics) and training of prediction software tools.

What you can find in UniProtKB/Swiss-Prot

Function of the protein; enzyme-specific information (catalytic activity, cofactors, metabolic pathway, regulation mechanisms); biologically relevant domains and sites; post-translational modifications (PTMs); molecular weights determined by mass spectrometry; subcellular location(s); tissue-specific expression; development-specific expression; secondary and quaternary structure information; alternative protein products (derived from alternative splicing, alternative promoter usage, alternative initiation, RNA editing...); 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 UniProtKB/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 EMBL/GenBank/DDBJ 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.

Interactive access to UniProtKB

You can access UniProtKB from <http://www.uniprot.org/> through the 'Text Search' tool. The most efficient way to browse interactively in Swiss-Prot and/or TrEMBL is to use the Sequence Retrieval System (SRS) that is available on the ExPASy server at <http://www.expasy.org/>, and its mirror sites, or on the EBI server at <http://www.ebi.ac.uk>.

How to obtain a local copy of UniProtKB

Swiss-Prot and TrEMBL can be obtained by anonymous ftp from the ExPASy <ftp.expasy.org> and EBI servers <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>.

Biweekly updated complete non-redundant data sets for Swiss-Prot and TrEMBL are provided for ftp download.

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 UniProtKB

If you want to cite UniProtKB in a publication, please use the following reference:

The Universal Protein Resource (UniProt)
Bairoch A., Apweiler R., Wu C.H., Barker W.C., Boeckmann B., Ferro S., Gasteiger E., Huang H., Lopez R., Magrane M., Martin M.J., Natale D.A., O'Donovan C., Redaschi N., Yeh L.S.
Nucleic Acids Res. **33**:D154-D159(2005).

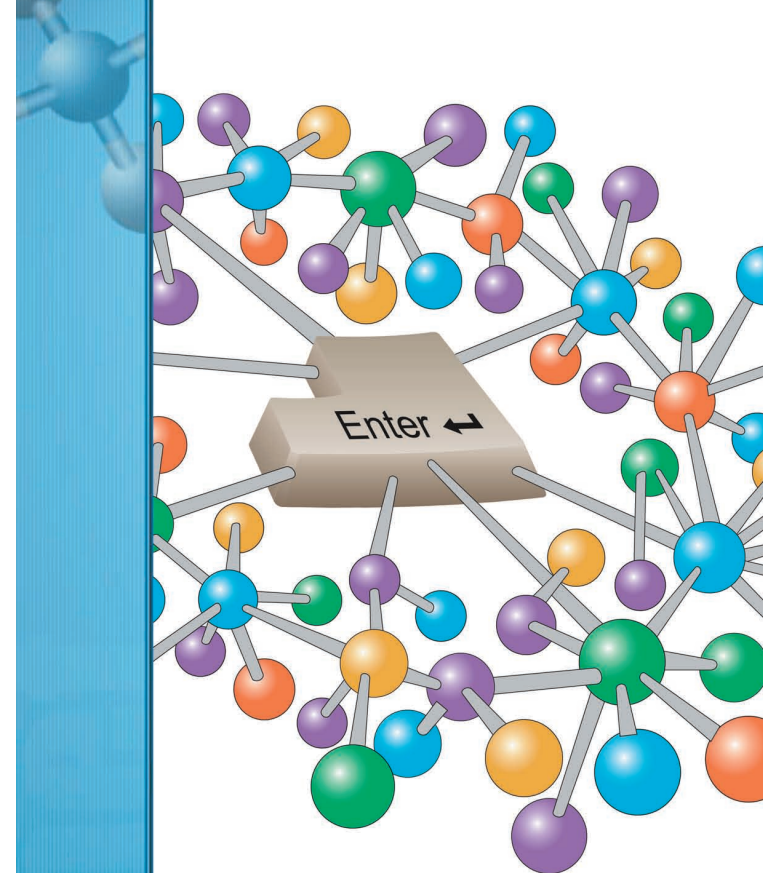


This document was written and designed by Brigitte Boeckmann from the Swiss Institute of Bioinformatics and being distributed by P&PR Publications Committee of EMBnet.

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/>

A Quick Guide To UniProtKB, Swiss-Prot and TrEMBL
Second edition © 2006



EMBnet

A Quick Guide
UniProtKB
Swiss-Prot & TrEMBL