EMBL-EBI

European Bioinformatics Institute
is an Outstation of the
European Molecular Biology Laboratory.

Xenopus and Zebrafish Annotation in the UniProt Knowledgebase (UniProtKB)

Rebecca E. Foulger[1] and UniProt Consortium[1,2,3]

[1]European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, UK
[2]Swiss Institute of Bioinformatics, Geneva, Switzerland
[3]Protein Information Resource, Georgetown University, Washington DC, USA

Introduction to UniProt

- UniProt (Universal Protein Resource: http://www.uniprot.org) provides a central resource of protein sequences with functional annotation. The UniProt Consortium is a collaboration between the Swiss Institute of Bioinformatics (SIB), the European Bioinformatics Institute (EBI) and the Protein Information Resource (PIR).

- UniProt Knowledgebase (UniProtKB) contains the manually annotated UniProtKB/Swiss-Prot section and the automatically annotated UniProtKB/TrEMBL section.

UniProtKB/TrEMBL

Computer-annotated protein sequences

Sequences with 100% identity from the same organism are merged

UniProtKB/Swiss-Prot

Manually annotated protein sequences

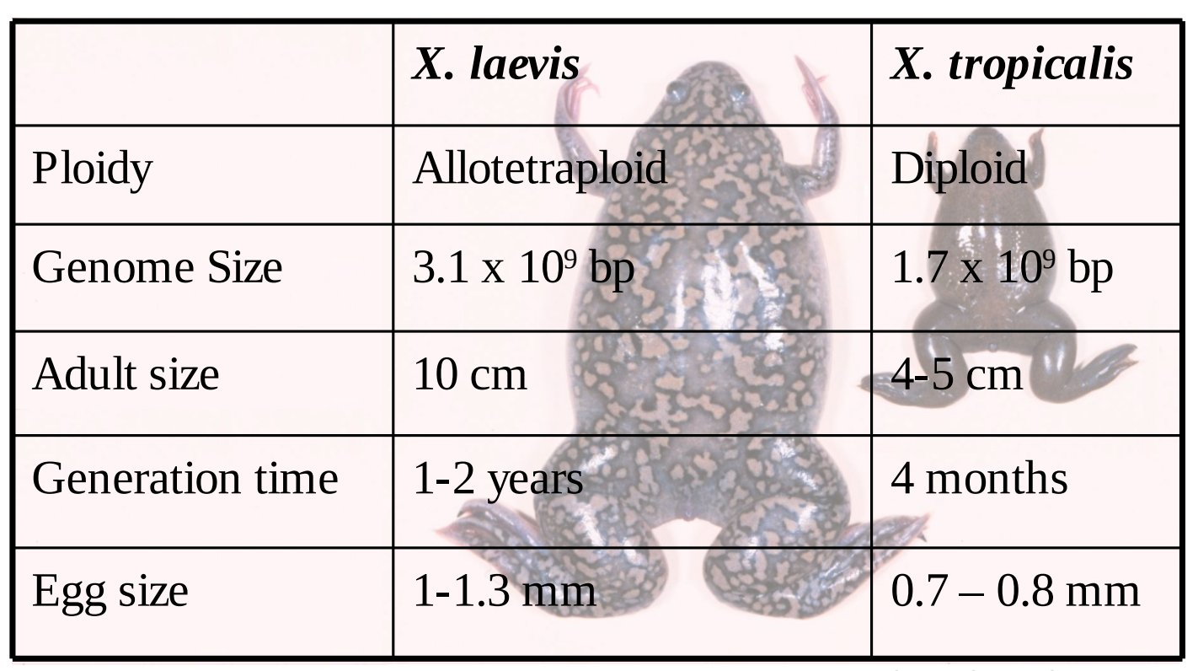
Non-redundant and high level of accuracy

Why frogs and fish?

- *Xenopus laevis* (African clawed frog) and *Danio rerio* (zebrafish) are both model organisms in the laboratory, as they are easy to accommodate and can produce large numbers of externally-developing embryos that can be easily viewed and manipulated.

- With a generation time of 2 to 3 months, *D. rerio* is amenable to genetic analysis, and the Sanger Institute began sequencing the zebrafish genome in 2001.The extensive similarity between the zebrafish and human genomes means many human developmental and disease genes have counterparts in the zebrafish.

- *X. laevis* has a much longer generation time of 1-2 years, and its tetraploid genome creates difficulties for genetic analysis. The related species *Xenopus tropicalis* (Western clawed frog) is diploid with a relatively small genome and a shorter generation time, so is much better suited for genetic studies. Sequencing of the *X. tropicalis* genome has begun and is supported by cDNA and EST sequencing projects, making it an ideal time to focus on Xenopus curation.



	X. laevis	*X. tropicalis*
Ploidy	Allotetraploid	Diploid
Genome Size	3.1 x 10^9 bp	1.7 x 10^9 bp
Adult size	10 cm	4-5 cm
Generation time	1-2 years	4 months
Egg size	1-1.3 mm	0.7 – 0.8 mm

Xenopus photos by Enrique Amaya.

The Challenge of Duplicated Genomes

- Both *Danio rerio* and *Xenopus laevis* have undergone whole genome duplications, resulting in duplicated genes for multiple loci.

- Protein sequences encoded by duplicated *X. laevis* and *D. rerio* genes are annotated as separate UniProtKB/Swiss-Prot entries.

Accession	Entry name	Status	Protein names	Gene names	Organism
Q9W707	FXF1B_XENLA	★	**Forkhead box protein F1-B** (FoxF1-B) (FoxF1b) (Fork head domain-related protein 13') (xFD-13')	**foxf1-B**	Xenopus laevis (African clawed frog)
Q9W706	FXF1A_XENLA	★	**Forkhead box protein F1-A** (FoxF1-A) (FoxF1a) (Fork head domain-related protein 13) (xFD-13)	**foxf1-A**	Xenopus laevis (African clawed frog)

A and B nomenclature is used to distinguish the duplicated sequences

Synonyms from the literature are recorded

Accession	Entry name	Status	Protein names	Gene names	Organism
Q6PC54	SN25B_DANRE	★	**Synaptosomal-associated protein 25-B** (SNAP-B) (Synaptosome-associated protein 25.2) (SNAP-25.2)	**snap25b** (Snap) (snap25.2)	Danio rerio (Zebrafish) (Brachydanio rerio)
Q5TZ66	SN25A_DANRE	★	**Synaptosomal-associated protein 25-A** (SNAP-25A) (Synaptosome-associated protein 25.1) (SNAP-25.1)	**snap25a** (Snap) (snap25.1) (si:dkeyp-8f4.6)	Danio rerio (Zebrafish) (Brachydanio rerio)

- For *D. rerio*, the UniProtKB A/B nomenclature matches ZFIN. For *X. laevis*, the A/B naming is taken from the literature where available, or assigned by a curator when the literature doesn’ t specify.

Curation of a *X. tropicalis* TrEMBL Entry

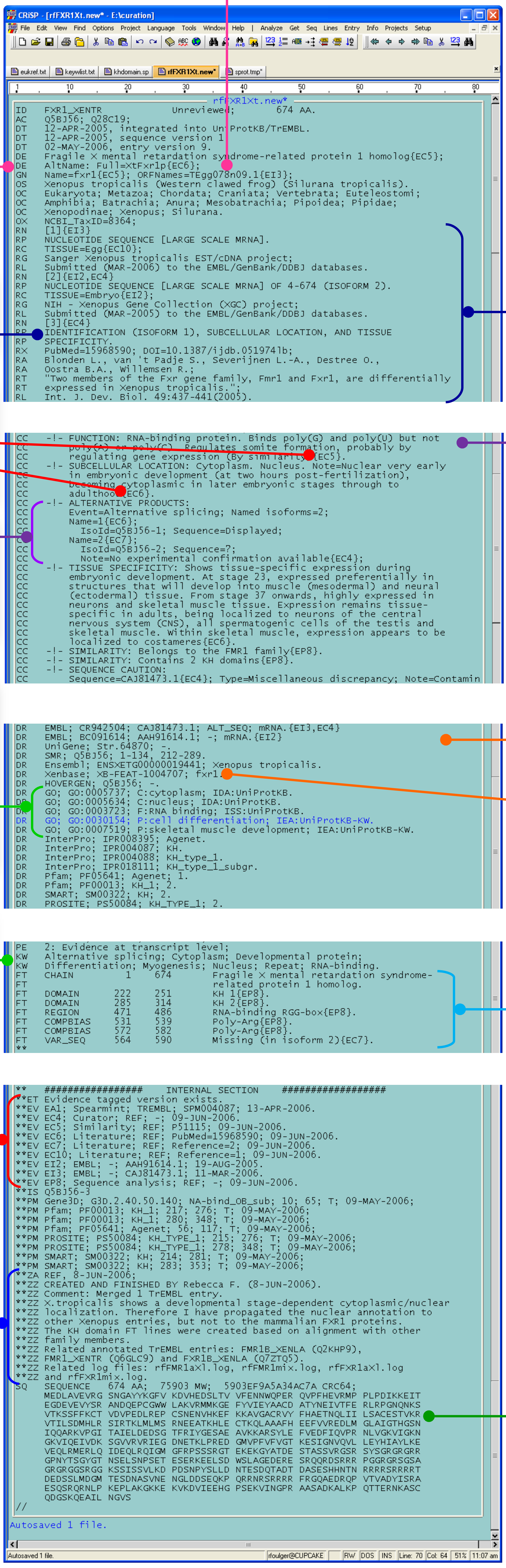
- We use a protein-by-protein approach to curation.

- PubMed searches, requests from users, cross-reference updates and sequence revisions are all used to identify which proteins are of priority to curate.

- Entries are curated using our flat-file editor, CRiSP.

For both Xenopus and zebrafish, protein and gene nomenclature is generally propagated from the human ortholog, with species-specific names added as synonyms

Identifiers from large-scale projects are recorded



```
ID   FXR1_XENTR              Unreviewed;       674 AA.
AC   Q5BJ56; Q28C19;
DT   12-APR-2005, integrated into UniProtKB/TrEMBL.
DT   12-APR-2005, sequence version 1
DT   02-MAY-2006, entry version 9.
DE   Fragile X mental retardation syndrome-related protein 1 homolog{EC5};
DE   AltName: Full=XtFxr1p{EC6};
GN   Name=fxr1{EC5}; ORFNames=TEgg078n09.1{EI3};
OS   Xenopus tropicalis (Western clawed frog) (Silurana tropicalis).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Amphibia; Batrachia; Anura; Mesobatrachia; Pipoidea; Pipidae;
OC   Xenopodinae; Xenopus; Silurana.
OX   NCBI_TaxID=8364;
RN   [1]{EI3}
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE MRNA].
RC   TISSUE=Egg{EC10};
RG   Sanger Xenopus tropicalis EST/cDNA project;
RL   Submitted (MAR-2006) to the EMBL/GenBank/DDBJ databases.
RN   [2]{EI2,EC4}
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE MRNA] OF 4-674 (ISOFORM 2).
RC   TISSUE=Embryo{EI2};
RG   NIH - Xenopus Gene Collection (XGC) project;
RL   Submitted (MAR-2005) to the EMBL/GenBank/DDBJ databases.
RN   [3]{EC4}
RP   IDENTIFICATION (ISOFORM 1), SUBCELLULAR LOCATION, AND TISSUE
RP   SPECIFICITY.
RX   PubMed=15968590; DOI=10.1387/ijdb.051974lb;
RA   Blonden L., van 't Padje S., Severijnen L.-A., Destree O.,
RA   Oostra B.A., Willemsen R.;
RT   "Two members of the Fxr gene family, Fmr1 and Fxr1, are differentially
RT   expressed in Xenopus tropicalis.";
RL   Int. J. Dev. Biol. 49:437-441(2005).

CC   -!- FUNCTION: RNA-binding protein. Binds poly(G) and poly(U) but not
CC       poly(A) or poly(C). Regulates somite formation, probably by
CC       regulating gene expression (By similarity){EC5}.
CC   -!- SUBCELLULAR LOCATION: Cytoplasm. Nucleus. Note=Nuclear very early
CC       in embryonic development (at two hours post-fertilization),
CC       becoming cytoplasmic in later embryonic stages through to
CC       adulthood{EC6}.
CC   -!- ALTERNATIVE PRODUCTS:
CC       Event=Alternative splicing; Named isoforms=2;
CC       Name=1{EC6};
CC         IsoId=Q5BJ56-1; Sequence=Displayed;
CC       Name=2{EC7};
CC         IsoId=Q5BJ56-2; Sequence=?;
CC         Note=No experimental confirmation available{EC4};
CC   -!- TISSUE SPECIFICITY: Shows tissue-specific expression during
CC       embryonic development. At stage 23, expressed preferentially in
CC       structures that will develop into muscle (mesodermal) and neural
CC       (ectodermal) tissue. From stage 37 onwards, highly expressed in
CC       neurons and skeletal muscle tissue. Expression remains tissue-
CC       specific in adults, being localized to neurons of the central
CC       nervous system (CNS), all spermatogenic cells of the testis and
CC       skeletal muscle. Within skeletal muscle, expression appears to be
CC       localized to costameres{EC6}.
CC   -!- SIMILARITY: Belongs to the FMR1 family{EP8}.
CC   -!- SIMILARITY: Contains 2 KH domains{EP8}.
CC   -!- SEQUENCE CAUTION:
CC       Sequence=CAJ81473.1{EC4}; Type=Miscellaneous discrepancy; Note=Contamin

DR   EMBL; CR942504; CAJ81473.1; ALT_SEQ; mRNA.{EI3,EC4}
DR   EMBL; BC091614; AAH91614.1; -; mRNA.{EI2}
DR   UniGene; Str.64870; -.
DR   SMR; Q5BJ56; 1-134, 212-289.
DR   Ensembl; ENSXETG00000019441; Xenopus tropicalis.
DR   Xenbase; XB-FEAT-1004707; fxr1.
DR   HOVERGEN; Q5BJ56; -.
DR   GO; GO:0005737; C:cytoplasm; IDA:UniProtKB.
DR   GO; GO:0005634; C:nucleus; IDA:UniProtKB.
DR   GO; GO:0003723; F:RNA binding; ISS:UniProtKB.
DR   GO; GO:0030154; P:cell differentiation; IEA:UniProtKB-KW.
DR   GO; GO:0007519; P:skeletal muscle development; IEA:UniProtKB-KW.
DR   InterPro; IPR008395; Agenet.
DR   InterPro; IPR004087; KH.
DR   InterPro; IPR004088; KH_type_1.
DR   InterPro; IPR018111; KH_type_1_subgr.
DR   Pfam; PF05641; Agenet; 1.
DR   Pfam; PF00013; KH_1; 2.
DR   SMART; SM00322; KH; 2.
DR   PROSITE; PS50084; KH_TYPE_1; 2.

PE   2: Evidence at transcript level;
KW   Alternative splicing; Cytoplasm; Developmental protein;
KW   Differentiation; Myogenesis; Nucleus; Repeat; RNA-binding.
FT   CHAIN         1    674       Fragile X mental retardation syndrome-
FT                                related protein 1 homolog.
FT   DOMAIN      222    251       KH 1{EP8}.
FT   DOMAIN      285    314       KH 2{EP8}.
FT   REGION      471    486       RNA-binding RGG-box{EP8}.
FT   COMPBIAS    531    539       Poly-Arg{EP8}.
FT   COMPBIAS    572    582       Poly-Arg{EP8}.
FT   VAR_SEQ     564    590       Missing (in isoform 2){EC7}.
**

**   #################    INTERNAL SECTION    ##################
**ET Evidence tagged version exists.
**EV EA1; Spearmint; TREMBL; SPM004087; 13-APR-2006.
**EV EC4; Curator; REF; -; 09-JUN-2006.
**EV EC5; Similarity; REF; P51115; 09-JUN-2006.
**EV EC6; Literature; REF; PubMed=15968590; 09-JUN-2006.
**EV EC7; Literature; REF; Reference=2; 09-JUN-2006.
**EV EC10; Literature; REF; Reference=1; 09-JUN-2006.
**EV EI2; EMBL; -; AAH91614.1; 19-AUG-2005.
**EV EI3; EMBL; -; CAJ81473.1; 11-MAR-2006.
**EV EP8; Sequence analysis; REF; -; 09-JUN-2006.
**IS Q5BJ56-3
**PM Gene3D; G3D.2.40.50.140; NA-bind_OB_sub; 10; 65; T; 09-MAY-2006;
**PM Pfam; PF00013; KH_1; 217; 276; T; 09-MAY-2006;
**PM Pfam; PF00013; KH_1; 280; 348; T; 09-MAY-2006;
**PM Pfam; PF05641; Agenet; 56; 117; T; 09-MAY-2006;
**PM PROSITE; PS50084; KH_TYPE_1; 215; 276; T; 09-MAY-2006;
**PM PROSITE; PS50084; KH_TYPE_1; 278; 348; T; 09-MAY-2006;
**PM SMART; SM00322; KH; 214; 281; T; 09-MAY-2006;
**PM SMART; SM00322; KH; 283; 353; T; 09-MAY-2006;
**ZA REF, 8-JUN-2006;
**ZZ CREATED AND FINISHED BY Rebecca F. (8-JUN-2006).
**ZZ Comment: Merged 1 TrEMBL entry.
**ZZ X.tropicalis shows a developmental stage-dependent cytoplasmic/nuclear
**ZZ localization. Therefore I have propagated the nuclear annotation to
**ZZ other Xenopus entries, but not to the mammalian FXR1 proteins.
**ZZ The KH domain FT lines were created based on alignment with other
**ZZ family members.
**ZZ Related annotated TrEMBL entries: FMR1B_XENLA (Q2KHP9),
**ZZ FMR1_XENTR (Q6GLC9) and FXR1B_XENLA (Q7ZTQ5).
**ZZ Related log files: rfFMR1aXl.log, rfFMR1mix.log, rfFXR1aXl.log
**ZZ and rfFXR1mix.log.
SQ   SEQUENCE   674 AA;  75903 MW;  5903EF9A5A34AC7A CRC64;
     MEDLAVEVRG SNGAYYKGFV KDVHEDSLTV VFENNWQPER QVPFHEVRMP PLPDIKKEIT
     EGDEVEVYSR ANDQEPCGWW LAKVRMMKGE FYVIEYAACD ATYNEIVTFE RLRPGNQNKS
     VTKSSFFKCT VDVPEDLREP CSNENVHKEF KKAVGACRVY FHAETNQLII LSACESTVKR
     VTILSDMHLR SIRTKLMLMS RNEEATKHLE CTKQLAAAFH EEFVVREDLM GLAIGTHGSN
     IQQARKVPGI TAIELDEDSG TFRIYGESAE AVKKARSYLE FVEDFIQVPR NLVGKVIGKN
     GKVIQEIVDK SGVVRVRIEG DNETKLPRED GMVPFVFVGT KESIGNVQVL LEYHIAYLKE
     VEQLRMERLQ IDEQLRQIGM GFRPSSSRGT EKEKGYATDE STASSVRGSR SYSGRGRGRR
     GPNYTSGYGT NSELSNPSET ESERKEELSD WSLAGEDERE SRQQRDSRRR PGGRGRSGSA
     GRGRGGSRGG KSSISSVLKD PDSNPYSLLD NTESDQTADT DASESHHNTN RRRRSRRRRT
     DEDSSLMDGM TESDNASVNE NGLDDSEQKP QRRNRSRRRR FRGQAEDRQP VTVADYISRA
     ESQSRQRNLP KEPLAKGKKE KVKDVIEEHG PSEKVINGPR AASADKALKP QTTERNKASC
     QDGSKQEAIL NGVS
//
Autosaved 1 file.
```

Functional annotation is attributed to individual papers

References include both large scale projects (including XGC/ZGC and Sanger cDNA projects), and sequences and papers from individual labs. A PubMed search is performed to find additional characterization papers. All available papers are manually curated and added to a TrEMBL entry

Evidence tags are used throughout the entry to show the source of the annotation

A wide range of functional data is taken from papers. Data is also propagated from similar UniProtKB/Swiss-Prot entries, and predicted by sequence analysis tools

Isoforms are annotated based on all available sequences and literature

Most cross-references are added automatically. Some (E.g. PROSITE, EMBL/DDBJ/GenBank, and MOD X-ref lines) can be added or modified by a curator

GO annotation is added manually using the Protein2GO editor. IEA annotations are generated from automatic methods, including KW:GOterm mapping

Reciprocal links to Xenbase gene pages were introduced at the end of 2008

Xenbase

Computer-generated keywords (KWs) are verified by a curator, and further keywords are added manually

Key residues, domains and motifs are annotated based on sequence analysis tools (such as Anabelle), similarity to existing UniProtKB/Swiss-Prot entries, and information in the literature

Evidence tags show the source of annotation

Internal comments are used to explain annotation and point to related annotated entries

The longest isoform is usually displayed. Where conflicts exist, for frogs and fish the consensus sequence is normally displayed

Acknowledgements

UniProt is supported by grants from the National Institutes of Health, European Commission, Swiss Federal Government and PATRIC BRC.

UniProt

European Bioinformatics Institute (EMBL-EBI)

Swiss Institute of Bioinformatics (SIB)

Protein Information Resource (PIR)

Email: help@uniprot.org

URL: www.uniprot.org