One gene, several proteins: sequence annotation in UniProtKB/Swiss-Prot

UniProt
the Universal protein resource

www.uniprot.org

20 Years SIB swissprot EBI

One of the main goals of UniProtKB/Swiss-Prot is to provide the scientific community with the most accurate protein sequence and the description of all its variations.

EMBL/GenBank/DDBJ Coding sequences (CDS)
High redundancy

→

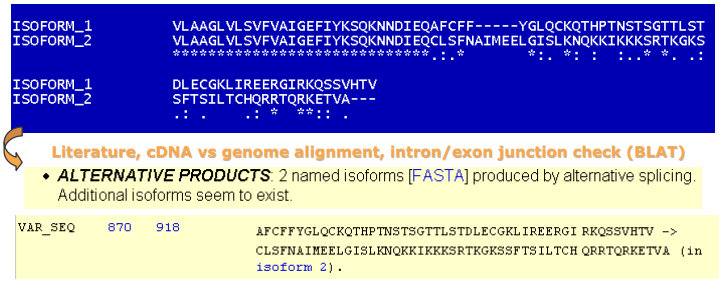
UniProtKB/TrEMBL
100% identical CDS are merged
(within the same species)

→

UniProtKB/Swiss-Prot
- Sequences encoded by one gene are merged (including fragments, alternative splicing isoforms, polymorphisms (SAPs), …);
- The longest sequence is preferentially shown to allow the description of a maximum of features.

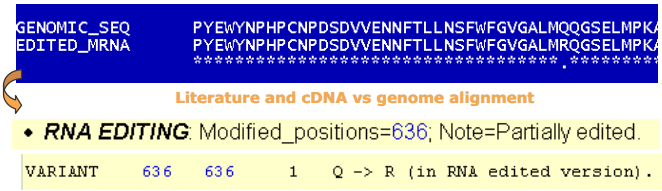
Manual sequence annotation

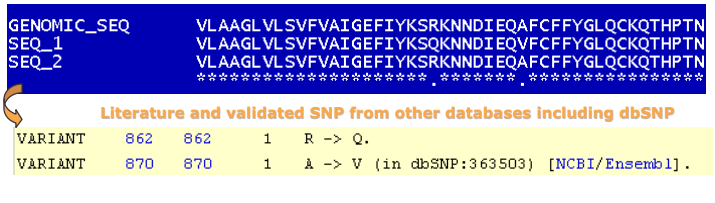
1. In the merging process, **UniProtKB/TrEMBL accession** numbers are kept, as well as **references** (publications or submissions) and **cross-references** to EMBL/GenBank/DDBJ; **Protein sequencing data** is integrated, if available;
2. Discrepancies (SAPs, alternative promoter usage, alternative splicing, RNA editing, alternative translation initiation or unknown reasons) are analyzed and documented at **single amino acid level**;
3. Sequence comparison between **related organisms** (e.g. human-mouse-rat) allows confirmation of sequence length, sometimes alternative splicing isoforms or SAPs;
4. Once the sequence corrected, high-performance bioinformatics tools are run and predicted domains, post-translational modifications (PTMs), etc. are critically reviewed before integration.



Alternative sequences annotated in:
References: *RP* line, e.g. *NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 2).*
Comments: topic *ALTERNATIVE PRODUCTS*; contains information about alternative promoter usage, alternative splicing and alternative translation initiation.
Keywords: *Alternative splicing, Alternative promote usage, Alternative initiation*
Sequence features: topic *FT VAR_SEQ* characterized by a unique identifier (/FTId=VSP_000128).

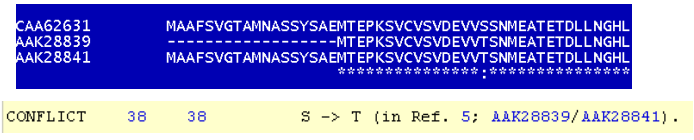
RNA editing annotated in:
References: *RP* line, e.g. *RNA EDITING OF POSITIONS 636.*
Comments: topic *RNA EDITING*.
Keywords: *RNA editing.*
Sequence features: e.g. FT VARIANT 636 636 Q -> R (in RNA edited version). /FTId=VAR_000304
Sequence: the edited sequence is shown in the entry, when fully edited.





Polymorphisms annotated in:
References: *RP* line, e.g. *VARIANT VAL-98.*
Comments: topic *POLYMORPHISM* and/or *DISEASE.*
Keywords: *Polymorphism* and/or *Disease mutation.*
Sequence features: topic *FT VARIANT* characterized by a unique identifier (FTId=VAR_123456). Cross-references to dbSNP.

Conflicts
Discrepancies of unknown origin, ranging from sequencing errors to yet uncharacterized polymorphisms.



Cross-references to nucleic acid sequence databases
Various tags in DR EMBL /GenBank/DDBJ lines allow the description of discrepancies between the submitted CDS and the UniProtKB/Swiss-Prot sequence:
ALT_INIT, **ALT_TERM**: erroneous start or stop;
ALT_FRAME:presence of frameshift(s) in the submitted nucleotide sequence. Linked to comment CAUTION: e.g. *Ref.X (PID) sequence differs from that shown due to a frameshift in position Y;*
ALT_SEQ: erroneous gene/CDS prediction or combination of several problems. Linked to comment CAUTION: e.g. *Ref.x (PID) sequence differs from that shown due to erroneous gene model prediction.*

UniProt
the Universal protein resource

Swiss Institute of Bioinformatics (SIB)
European Bioinformatics Institute (EMBL-EBI)
Protein Information Resource (PIR)

Contact
help@uniprot.org