UniProt Manual Curation SOP

Author: UniProt Consortium
Version: 2.0
Effective Date: November 2014

1. Abstract

The UniProt manual curation process comprises manual review of results from a range of sequence analysis programs and literature curation of experimental data as well as attribution of all information to its original source. Curators also assign GO terms to all manually curated entries.

2. Introduction

This SOP describes the manual curation procedure used by the UniProt Consortium members at the European Bioinformatics Institute (EBI), the SIB Swiss Institute of Bioinformatics and the Protein Information Resource (PIR). The UniProt Knowledgebase (UniProtKB) consists of two sections, UniProtKB/Swiss-Prot and UniProtKB/TrEMBL. TrEMBL records are enriched with automatic classification and annotation while Swiss-Prot records are manually curated by a team of biologists. The manual curation procedure results in a newly curated or updated UniProtKB/Swiss-Prot record.

3. Requirements

3.1 Data requirements

One or more UniProtKB records requiring manual curation

3.2 Software requirements

UniProt curation editor, UniProt sequence analysis platform, Protein2GO curation tool, QuickGO browser

3.3 Compute requirements

Windows PC, network connection

4. Procedure

4.1 Select entry to curate

Entries are selected for manual curation based on defined curation priorities described at http://www.uniprot.org/program/.

4.2 Run sequence similarity searches

BLAST (1) searches are run with the sequence of the selected entry to identify additional sequences from the same gene.

4.3 Identify homologs

Reciprocal BLAST searches as well as phylogenetic resources such as Ensembl Compara (2) are used to identify homologous proteins which can be curated at the same time as the originally selected entry to ensure data consistency across related proteins.

4.4 Lock entries

Entries which are selected for manual curation are “locked” using an internal tool which prevents duplication of curation efforts by ensuring that no one else works on the entry until it is finished and “unlocked”.

4.5 Merge entries from same gene

Entries from the same gene and same species are merged into a single record to minimise redundancy in the database. Differences between sequence reports are identified using sequence alignment programs. Three sequence alignment programs are incorporated into the UniProt curation environment: 1) T-Coffee version 3.67 (3), 2) Muscle version 3.6 (4), 3) ClustalW version 1.83 (5). The underlying causes of any sequence differences such as alternative splicing, natural variations, frameshifts, incorrect initiation sites, incorrect exon boundaries and unidentified conflicts are documented in the merged record.

4.6 Sequence analysis

Sequences are analysed using a range of sequence analysis programs as shown in Table 1. The programs are integrated into a platform which allows all or a subset of selected programs to be launched simultaneously from the UniProt curation editor.

Table 1. Sequence analysis tools used during the UniProtKB manual curation process

Program	Version	Prediction
Topology		
Signal P (6)	3.0	Presence and location of signal peptides
TargetP (6)	1.1	Presence and location of transit peptides
Predotar (7)	1.03	Mitochondrial, plastid or ER targeting sequences
ESKW* (8)	UniProt-modified version 1.0	Transmembrane domains
MEMSAT (9)	UniProt-modified version 1.8a	Transmembrane domains
TMHMM (10)	2.0	Transmembrane domains
Phobius (11)	Unknown	Discriminates transmembrane and signal regions
Post-translational modifications		
GPI-predictor (12)	1.0	GPI lipid anchor sites
NetNGlyc (13)	1.0	N-glycosylation sites
NetOGlyc (14)	3.1	O-glycosylation sites
NMT Predictor (15)	1.0	N-terminal myristoylation sites
Sulfinator (16)	1.0	Tyrosine sulfation sites
Domains		
ps_scan	1.0	Internal PROSITE profile, pattern and rule scanning program
InterPro (17)	Uses latest versions of InterPro and InterProScan	Retrieves non-PROSITE motif matches using InterPro database or InterProScan
Coils (18)	2.2	Coiled-coil regions
polyAA	1.0	Internal program which identifies homopolymeric stretches of amino acids
REPEAT (19)	1.1	Identifies the following repeats: Ankyrin, Armadillo, HAT, HEAT, Kelch, Leucine-rich, PFTA, PFTB, RCC1, TPR, WD40

*ESKW = transmembrane prediction algorithm by Eisenberg, Schwarz, Komaromy and Wall

Automatically selected results are returned in a graphical interface which allows visualisation of the predictions (Figure 1). Selected features are shown in green and unselected features are shown in red. The selected/unselected state of a feature can be toggled by clicking on it.

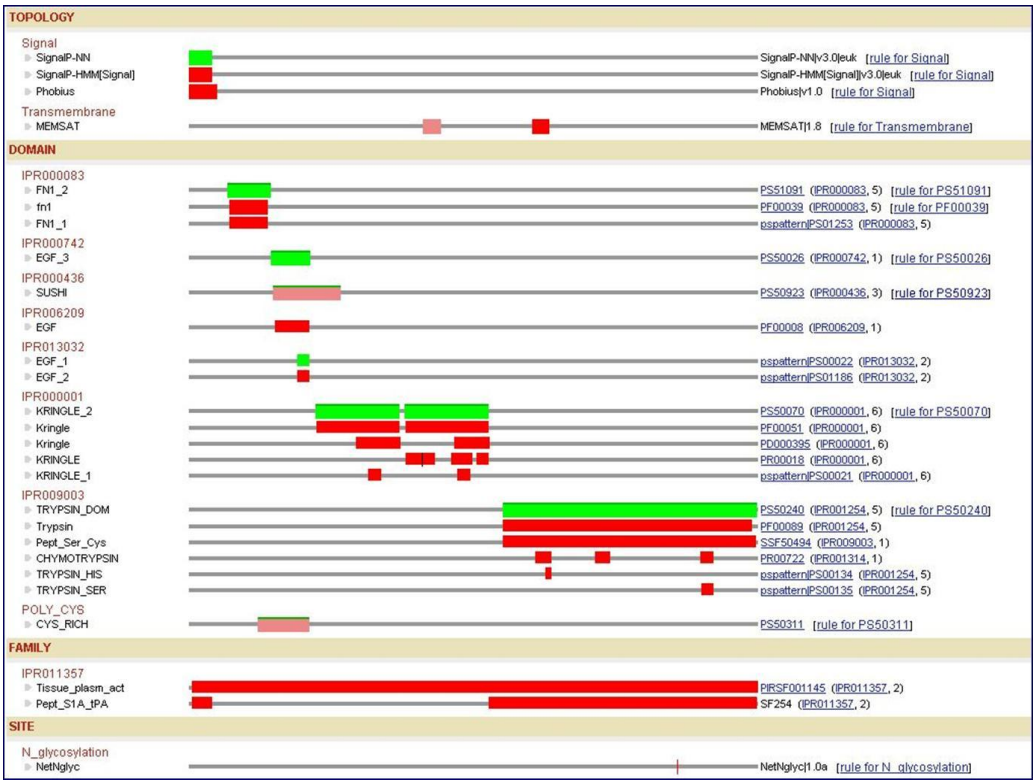
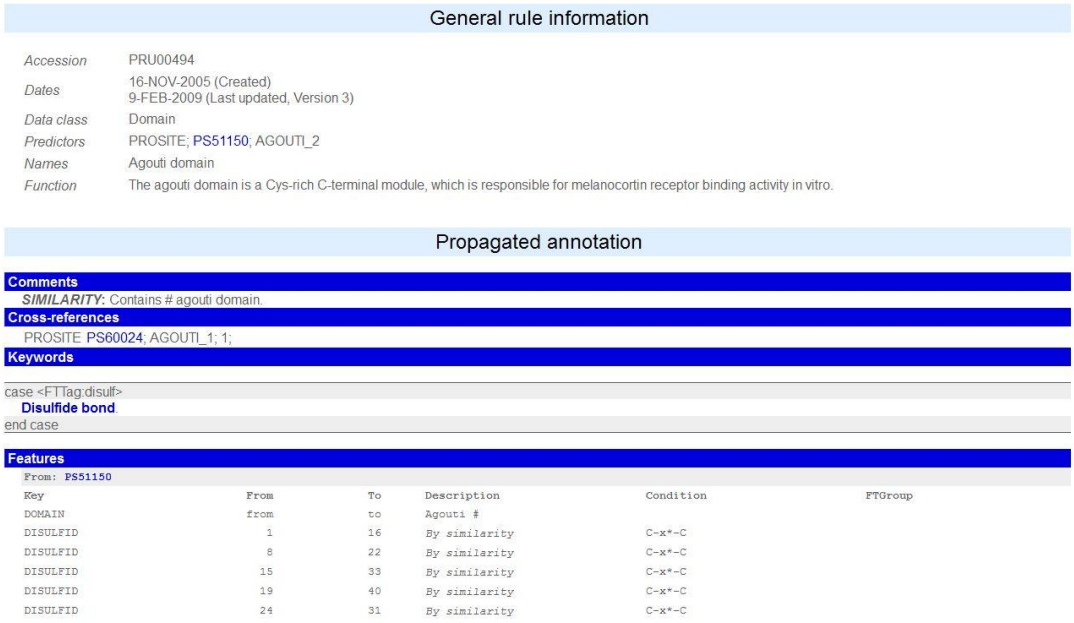


Figure 1. UniProtKB sequence analysis results displayed in graphical interface

All predictions are manually reviewed and relevant results are selected for inclusion in the entry. The sequence analysis platform then transforms the selected features into UniProtKB annotation by applying a set of automatic annotation rules (Figure 2).



General rule information

Accession	PRU00494
Dates	16-NOV-2005 (Created) 9-FEB-2009 (Last updated, Version 3)
Data class	Domain
Predictors	PROSITE; PS51150; AGOUTI_2
Names	Agouti domain
Function	The agouti domain is a Cys-rich C-terminal module, which is responsible for melanocortin receptor binding activity in vitro.

Propagated annotation

Comments

SIMILARITY: Contains # agouti domain.

Cross-references

PROSITE PS60024; AGOUTI_1; 1;

Keywords

case <FTTag:disulf>
Disulfide bond
end case

Features

From: PS51150

Key	From	To	Description	Condition	FTGroup
DOMAIN	from	to	Agouti #		
DISULFID	1	16	*By similarity*	C-x*-C	
DISULFID	8	22	*By similarity*	C-x*-C	
DISULFID	15	33	*By similarity*	C-x*-C	
DISULFID	19	40	*By similarity*	C-x*-C	
DISULFID	24	31	*By similarity*	C-x*-C	

Figure 2. Example of an annotation rule incorporated into the UniProt sequence analysis platform

The resulting annotations are imported into the UniProt curation editor in a format where they can be transferred to the relevant entries.

4.7 Identification of relevant scientific literature

Relevant papers are identified using literature and text-mining resources such as PubMed (20), Europe PubMed Central (21), iHOP (22) and TextPresso (23). Access is also provided to the UniProt Additional Bibliography through the curation editor. The Additional Bibliography includes references which have been imported from a range of external databases to supplement the literature in UniProtKB and aids in identification of relevant papers.

4.8 Literature curation

The full text of each paper is read and information is extracted and added to the entry using the UniProt curation editor. General biological information is added in a defined set of annotations as documented at http://www.uniprot.org/manual/general_annotation and includes a wide range of information related to the role of the protein such as its function, subcellular location, interactions with other proteins and subunit structure. Position-specific annotation is included in a set of sequence features as documented at http://www.uniprot.org/manual/sequence_annotation. These features describe regions or sites of interest in the protein sequence including post-translational modifications, binding sites, enzyme active sites and local secondary structure. In addition, all references which are used during the curation procedure are added to the entry with details of what information has been extracted from each paper.

4.9 Family-based curation

Putative homologs which were identified in step 4.3 are evaluated and curated according to the steps outlined above. Annotation is standardized and propagated across homologous proteins to ensure data consistency.

4.10 Evidence attribution

All information added to a UniProtKB entry during the manual curation process is attributed to its original source so that users can trace the origin of each piece of information and evaluate it. This is

done through the use of a subset of evidence codes from the Evidence Code Ontology (ECO) (24). There are seven ECO evidence codes used in manually curated entries as shown in Table 2.

Table 2. Evidence Code Ontology (ECO) codes used during the UniProt manual curation process

ECO code	Term name	Usage
ECO:0000269	experimental evidence used in manual assertion	Information for which there is published experimental evidence
ECO:0000303	non-traceable author statement used in manual assertion	Information based on author statements in scientific articles for which there is no experimental support
ECO:0000250	sequence similarity evidence used in manual assertion	Information which has been propagated from a related experimentally characterised protein
ECO:0000312	imported information used in manual assertion	Information which has been imported from another database and manually verified
ECO:0000305	curator inference used in manual assertion	Information which has been inferred by a curator based on his/her scientific knowledge or on the scientific content of an article
ECO:0000255	match to sequence model evidence used in manual assertion	Information originating from the UniProt automatic annotation systems or any of the sequence analysis programs used during the manual curation process and which has been manually verified
ECO:0000244	combinatorial evidence used in manual assertion	Information which is manually curated based on a combination of experimental and computational evidence

Full details of the evidences used in UniProtKB are available at http://www.uniprot.org/manual/evidences.

4.11 GO annotation

Gene Ontology (GO) terms are assigned based on experimental data from the literature. Relevant terms are identified using the QuickGO (25) browser and are assigned to entries using the Protein2GO curation tool. This tool has been developed within the UniProt group and is used both by UniProt and by other members of the GO Consortium. GO terms are also propagated to homologous proteins where appropriate. The procedure is described in more detail at http://www.ebi.ac.uk/GOA/ManualAnnotationEfforts.

4.12 Quality control and integration

All finished entries are run through a series of automated checks which verify a large number of biological rules such as the positions and relevance of amino acids cited in the entry. Any reported errors are corrected. Once an entry has passed the automated checks, it undergoes manual review by a senior curator to ensure that all relevant sequences have been merged, that all relevant literature has been added, that the annotation has been added correctly, and that all relevant sequence analysis results have been included. Once an entry has passed the automated and manual quality control checks, it is integrated into the database.

4.13 Unlock finished entries

Integrated entries are unlocked so that they are available for further curation.

5. Implementation
N/A

6. Discussion
N/A

7. Related documents and references

1. Altschul S.F., Madden T.L., Schaffer A.A. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., 25, 3389–3402.

2. Vilella,A.J., Severin,J., Ureta-Vidal,A. et al. (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 19:27–35.

3. Notredame C., Higgins D. and Heringa J. (2000) T_Coffee: a novel method for multiple sequence aligments. J. Mol. Biol. 302:205–217.

4. Edgar R.C. (2004) MUSCLE, multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

5. Thompson J.D., Higgins D.G., Gibson T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

6. Emanuelsson O., Brunak S., von Heijne G., Nielsen H. (2007) Locating proteins in the cell using TargetP, SignalP, and related tools. Nat. Protoc. 2:953-971.

7. Small I., Peeters N., Legeai F., Lurin C. (2004) Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. Proteomics 4:1581-1590.

8. Eisenberg D., Schwarz E., Komaromy M., Wall R. (1984) Analysis of membrane and surface protein sequences with the hydrophobic moment plot. J. Mol. Biol. 179:125-42

9**.** Jones, D.T., Taylor, W.R., Thornton, J.M. (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology. Biochemistry 33:3038-3048.

10. Krogh A., Larsson B., von Heijne G., Sonnhammer E.L.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. 305:567-580.

11. Käll L., Krogh A., Sonnhammer E.L.L. (2004) A combined transmembrane topology and signal peptide prediction method. J. Mol. Biol. 338:1027-1036.

12. Eisenhaber F., Eisenhaber B., Kubina W., Maurer-Stroh S., Neuberger G., Schneider G., Wildpaner M. (2003) Prediction of lipid posttranslational modifications and localization signals from protein sequences: big-Pi, NMT and PTS1. Nucleic Acids Res. 31:3631-3634.

13. http://www.cbs.dtu.dk/services/NetNGlyc/

14. Julenius K., Mølgaard A., Gupta R., Brunak S. (2005) Prediction, conservation analysis and structural characterization of mammalian mucin-type O-glycosylation sites. Glycobiology 15:153-164.

15. Maurer-Stroh S., Eisenhaber B., Eisenhaber F. (2002) N-terminal N-myristoylation of proteins: prediction of substrate proteins from amino acid sequence. J. Mol. Biol. 317:541-557.

16. Monigatti F., Gasteiger E., Bairoch A., Jung E. (2002) The Sulfinator: predicting tyrosine sulfation sites in protein sequences. Bioinformatics 18:769-770.

17. Hunter S., Apweiler R., Attwood T.K., Bairoch A., Bateman A., Binns D., Bork P., Das U., Daugherty L., Duquenne L., Finn R.D., Gough J., Haft D., Hulo N., Kahn D., Kelly E., Laugraud A., Letunic I., Lonsdale D., Lopez R., Madera M., Maslen J., McAnulla C., McDowall C., Mistry J., Mitchell A., Mulder N., Natale D., Orengo C., Quinn A.F., Selengut J.D., Sigrist C.J., Thimma M., Thomas P.D., Valentin F., Wilson D., Wu C.H., Yeats C. (2009) InterPro: the integrative protein signature database. Nucleic Acids Res. 37:D211-D215.

18. Lupas A., Van Dyke M., Stock J. (1991) Predicting coiled coils from protein sequences. Science 252:1162-1164.

19. Andrade M.A., Ponting C., Gibson T., Bork P. (2000) Identification of protein repeats and statistical significance of sequence comparisons. J. Mol. Biol. 298:521-537.

20. http://www.ncbi.nlm.nih.gov/pubmed

21. McEntyre J.R., Ananiadou S., Andrews S., Black W.J., Boulderstone R., Buttery P., Chaplin D., Chevuru S., Cobley N., Coleman L.A., Davey P., Gupta B., Haji-Gholam L., Hawkins C., Horne A., Hubbard S.J., Kim J.H., Lewin I., Lyte V., MacIntyre R., Mansoor S., Mason L., McNaught J., Newbold E., Nobata C., Ong E., Pillai S., Rebholz-Schuhmann D., Rosie H., Rowbotham R., Rupp C.J., Stoehr P., Vaughan P. (2011) UKPMC: a full text article resource for the life sciences. Nucleic Acids Res. 39:D58-D65.

22. Hoffmann, R., Valencia, A. (2004) A gene network for navigating the literature. Nat. Genet. 36:664.

23. Müller H.M., Kenny E.E., Sternberg P.W. (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biol. 2:e309.

24. Chibucos M.C., Mungall C.J., Balakrishnan R., Christie K.R., Huntley R.P., White O., Blake J.A., Lewis S.E., Giglio M. (2014) Standardized description of scientific evidence using the Evidence Ontology (ECO). Database 2014; article ID bau075; doi: 10.1093/database/bau075.

25. Binns D., Dimmer E., Huntley R., Barrell D., O'Donovan C., Apweiler R. (2009) QuickGO: a web-based tool for Gene Ontology searching. Bioinformatics 25:3045-3046.

8. Revision history

Version	Author	Date	Change made
1.0	Michele Magrane	7 June 2011	Established SOP
2.0	Michele Magrane	19 November 2014	Updated section 4.10 (Evidence attribution)