Standard operating procedure (SOP) for HAMAP annotation rule creation

Author: HAMAP
Version: 1.0
Effective Date: June 2014

1. Abstract

HAMAP annotation rules are used for the automatic functional annotation of protein sequences. The rules specify the annotations that can be applied to members of a protein family. The process of HAMAP annotation rule creation involves: i) the identification of all characterized members of a protein family and their manual curation in UniProtKB/Swiss-Prot (these serve as rule templates); ii) the construction of the annotation rule by the extraction of suitable annotation topics from the templates and the definition of (optional) conditions (in the form of control statements) that specify when these annotations apply; iii) the validation of the final rule by applying it to a representative set of UniProtKB entries.

2. Introduction

This SOP describes the procedure used by the curators of the Swiss-Prot group of the SIB Swiss Institute of Bioinformatics to create annotation rules for HAMAP (High-quality Automated and Manual Annotation of Proteins), a system for the classification and annotation of protein sequences (1). Classification is achieved using HAMAP family profiles with the annotation occurring at a subsequent step based on the evaluation of associated annotation rules. These specify the annotations that can be applied to members of a protein family and associated conditions, if any. HAMAP annotation rules and HAMAP family profiles are used in the automatic annotation of UniProtKB (2) and can be used for the analysis of any protein sequence on the HAMAP website (http://hamap.expasy.org/).

3. Requirements

3.1 Data requirements

HAMAP family profile and seed alignment for a protein family, constructed according to the standard operating procedure (SOP) for HAMAP family profiles creation (3).

3.2 Software requirements

UniProt curation editor, urucreate Perl script

3.3 Compute requirements

Windows PC, network connection

4. Procedure

4.1 Selection and annotation of template entries

Characterized members of a protein family (as defined by true-positive matches to the HAMAP family profile) are identified by literature search and subsequently annotated in UniProtKB/Swiss-Prot according to the standard operating procedure (SOP) for UniProt manual curation (4). These curated records will serve as annotation templates for the HAMAP annotation rule.

4.2 Rule construction

The curator runs a Perl script that parses a file that contains the template entries that were selected in 4.1 plus all seed alignment members. The output of this procedure is a 'skeleton' HAMAP annotation rule in UniRule format (5). This 'skeleton' rule contains all annotation blocks collected from the first record of the source file (ideally, the best characterized of the template entries) in the annotation section. The rule contains additional sections that are populated by the script based on the analysis of all entries in the source file (see below). The 'skeleton' rule must be completed by the curator before being validated.

The following sections are completed by the curator:

4.2.1 Header section

The Header section contains technical information about the HAMAP rule. The curator must give a meaningful name to the annotation rule.

Furthermore, the TR (Trigger) line describes which HAMAP family profile triggers the application of the current annotation rule.

If more than one profile triggers an annotation rule, all profile accession numbers must be added to the rule.

If a rule only applies to a taxonomic subset of the match set of an associated profile, and at least one other annotation rule exists for other sequences of the same match set, a taxonomic condition must be added to the TR line in each of these rules to prevent matching sequences from being annotated by two different rules.

4.2.2 Annotation section

All annotations that can be propagated from the template(s) to the target sequences, such as protein names, gene names, functional annotation, GO terms (6), keywords and sequence features are selected and combined to build the final rule (Table 1).

Table 1. Typical annotation topics found in a UniProtKB/Swiss-Prot record and their use in HAMAP annotation rules for annotation propagation

Annotation topics	Propagated	Not propagated
Protein names	Recommended name, EC number, Alternative name(s), Short name(s), Includes:, Contains: Flags: Precursor	Flags: Fragment(s)
Gene names	Names, Synonyms	OrderedLocusNames, ORFnames

Annotation topics	Propagated	Not propagated
Functional annotation	FUNCTION, CATALYTIC ACTIVITY, COFACTOR, ENZYME REGULATION, PATHWAY, SUBUNIT, SUBCELLULAR LOCATION, INDUCTION, DOMAIN, PTM, MISCELLANEOUS, SIMILARITY, CAUTION	BIOPHYSICOCHEMICAL PROPERTIES, INTERACTION, ALTERNATIVE PRODUCTS, TISSUE SPECIFICITY, DEVELOPMENTAL STAGE, MASS SPECTROMETRY, POLYMORPHISM, DISEASE, DISRUPTION PHENOTYPE, TOXIC DOSE, BIOTECHNOLOGY, PHARMACEUTICAL, SEQUENCE CAUTION, WEB RESOURCE
GO terms	All applicable	
Keywords	Categories: Biological process, Cellular component, Developmental stage, Domain, Ligand, Molecular function, Post-translational modification	Categories: Coding sequence diversity, Disease, Technical term
Sequence features	SIGNAL, PROPEP, TRANSIT, CHAIN, PEPTIDE, TOPO_DOM, INTRAMEM, TRANSMEM, DOMAIN, REPEAT, CA_BIND, ZN_FING, DNA_BIND, NP_BIND, REGION, COILED, MOTIF, COMPBIAS, ACT_SITE, METAL, BINDING, SITE, MOD_RES, LIPID, CARBOHYD, DISULFID, CROSSLNK	NON_STD, VAR_SEQ, VARIANT, MUTAGEN, UNSURE, CONFLICT, NON_CONS, NON_TER, HELIX, TURN, STRAND

If several template entries are available and the content of identical annotation topics varies between the different templates (e.g. in different organisms), multiple variants of the annotation topic can be added to the rule, and control statements added to these annotation topics limit their propagation to only a subset of matching sequences that also satisfy these additional conditions. The same is true if particular annotation topics should only be propagated to a subset of the protein family members (e.g. annotations specific to defined taxonomic groups) or if the propagation of an annotation is dependent on the presence of particular sequence features or residues in the target protein sequence. All possible control statements that are available to the curator, their syntax, use, as well as more details and examples can be found in the UniRule format documentation (5).

The conditions are based on relevant biological information collected from the literature or from curator knowledge, and are used to ensure that the annotation is only applied where appropriate, to guarantee the production of annotation of the same quality as that produced by expert curation.

4.2.3 Cross-references

Cross-references to PROSITE (7), Pfam (8), TIGRFAMs (9), PRINTS (10) and/or PIRSF (11) are added to the rule where applicable, and can serve as additional checks for annotation propagation. In the cross-references section, the curator can specify the expected number of occurrences of each available signature. For any target sequence that has a different number of matches to these signatures a warning is generated in the annotation output.

For PROSITE profiles associated with a ProRule (12), the curator can select this ProRule to be triggered for annotation of positive matches to the PROSITE profile. The annotation generated by the

ProRule will be merged with the annotations specified in the HAMAP annotation rule during the annotation propagation step in the HAMAP pipeline.

The cross-reference section is also used to specify the sequence analysis programs (e.g. signal sequence- and transmembrane region-predictors) that should be run on the target sequences. Sequence analysis tools must also be associated with annotation rules that define UniProtKB format annotations based on the analysis results. The syntax and use of the DR lines used to trigger the sequence analysis tools is explained in the UniRule format documentation (5) and a list of the sequence analysis tools used can be found in the SOP for UniProt manual curation (4).

4.2.4 Computing section

The computing section contains additional checks that are performed on the target sequences and that are specified by the curator to ensure proper annotation propagation and prevent propagation to entries not meeting all criteria.

4.2.4a Size

This field specifies the size range within which the size of a target protein must reside to not generate a warning. The field is prefilled with the shortest and the longest size value of the protein sequences from the source file. The size range can be modified by the curator.

4.2.4b Related

This field lists all other HAMAP annotation rules that may produce cross-annotations to members of the family to which the actual rule is applied. This may be the case for HAMAP family profiles for which multiple (usually non-overlapping, taxonomically separated) annotation rules exist, or for HAMAP family profiles that have overlapping match sets (usually proteins classified in a subfamily, which match both a profile for the subfamily as well as the profile for the broader superfamily). In the latter case, the curator must specify if a rule listed in the 'Related' line supersedes the current rule, i.e. matches to the current rule should be disregarded and not annotated if a match to the superseding rule is found. This ensures that each protein sequence is annotated only by a single rule.

4.2.4c Template

This field lists the UniProtKB accession numbers of all entries (templates) that were manually curated and for which there is experimental evidence or structural data that was used for constructing the family and its annotation rule. The field is prefilled with the accession number of the first record of the source file. The curator must complete this field with at least all accession numbers of entries listed as templates in the seed alignment and the feature section of the rule, plus optionally all other entries that provided characterization information for rule construction.

4.2.4d Scope

This field provides the possibility to specify the taxonomic scope of the annotation rule. It is prefilled with the highest hierarchical taxonomic node(s) representing all protein records from the source file. It can be edited by the curator. For any entries outside of the taxonomic scope of specified in this field a warning is generated in the annotation output.

4.2.4e Comments (Optional)

There are 2 sections in the annotation rule that give the curator the opportunity to store comments. The "Comments" field stores comments for users that will be visible publicly on the rule page on hamap.expasy.org. Additionally, an "Internal comments" section allows the curator to store comments and explanations only visible to other curators, which may be of use when revisiting the rule.

4.3 Validation of the rule

After a rule has been built, the curator validates the rule by applying the rule to target entries. The curator will apply the rule to all UniProtKB/Swiss-Prot matches of the associated HAMAP family profile as well as to a representative set of uncharacterized UniProtKB/TrEMBL entries matching the profile. The curator verifies that the annotation output for UniProtKB/Swiss-Prot entries corresponds to what would be expected if the entries were manually curated, and spot-checks the annotated UniProtKB/TrEMBL entries for any inconsistencies or warnings that may prevent propagation of annotation via UniProt's automatic annotation pipeline (3).

4.4 Putting rules in production

Once a HAMAP annotation rule has passed validation, it is integrated together with its associated HAMAP family profile into the HAMAP database.

5. Implementation

N/A

6. Discussion

N/A

7. Related documents and references

1. Pedruzzi I., Rivoire C., Auchincloss A.H., Coudert E., Keller G., de Castro E., Baratin D., Cuche B.A., Bougueleret L., Poux S., Redaschi N., Xenarios I., Bridge A. and the UniProt Consortium (2013) HAMAP in 2013, new developments in the protein family classification and annotation system. Nucleic Acids Res, 41,D584-D589.
2. UniProt's automatic annotation programs (available at http://www.uniprot.org/program/automatic_annotation)
3. Standard operating procedure (SOP) for HAMAP family profiles creation (available at ftp://ftp.expasy.org/databases/hamap/SOP_HAMAP_profile_creation.pdf)
4. Standard operating procedure (SOP) for UniProt manual curation (available at http://www.uniprot.org/docs/sop_manual_curation.pdf)
5. UniRule format documentation (available at http://hamap.expasy.org/unirule/unirule.html)
6. Gene Ontology Consortium. (2013) Gene Ontology annotations and resources. Nucleic Acids Res, 41, D530-535
7. Sigrist C.J., de Castro E., Cerutti L., Cuche B.A., Hulo N., Bridge A., Bougueleret L., Xenarios I. (2013) New and continuing developments at PROSITE. Nucleic Acids Res, 41, D344-347.

8. Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J. et al. (2012) The Pfam protein families database. Nucleic Acids Res, 40, D290-301.
9. Selengut, J.D., Haft, D.H., Davidsen, T., Ganapathy, A., Gwinn-Giglio, M., Nelson, W.C., Richter, A.R. and White, O. (2007) TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. Nucleic Acids Res, 35, D260-264
10. Attwood T.K., Coletta A., Muirhead G., Pavlopoulou A., Philippou P.B., Popov I., Romá-Mateo C., Theodosiou A., Mitchell A.L. (2012) The PRINTS database: a fine-grained protein sequence annotation and analysis resource--its status in 2012. Database (Oxford), 2012:bas019. doi: 10.1093/database/bas019.
11. Nikolskaya AN, Arighi CN, Huang H, Barker WC, Wu CH. (2007) PIRSF family classification system for protein functional and evolutionary analysis. Evol Bioinform Online, 2, 197-209.
12. Sigrist C.J., De Castro E., Langendijk-Genevaux P.S., Le Saux V., Bairoch A., Hulo N. (2005) ProRule: a new database containing functional and structural information on PROSITE profiles. Bioinformatics, 21, 4060-4066.

8. Revision history

Version	Author	Date	Change made
1.0	Ivo Pedruzzi	1 June 2014	Established SOP