Standard operating procedure (SOP) for HAMAP family profiles creation

Author: HAMAP
Version: 1.1
Effective Date: May 2014

1. Abstract

HAMAP family profiles are used to determine family membership of protein sequences. The process of HAMAP family profile creation involves the manual selection of trusted family member sequences (seed), the creation of a multiple sequence alignment (seed alignment), the automatic generation of a profile from this seed alignment, and the validation of the resulting profile by using it to scan against UniProtKB and verifying the score distribution of matching proteins. Curators can alter the composition of the seed alignment, profile generation parameters and the score threshold of the profile in order to enhance the specificity of the family profile, performing iterative profile searches until a satisfactory score distribution is obtained.

2. Introduction

This SOP describes the procedure used by the curators of the Swiss-Prot group of the Swiss Institute of Bioinformatics (SIB) to create family profiles for HAMAP (High-quality Automated and Manual Annotation of Proteins), a system for the classification and annotation of protein sequences (1). HAMAP family profiles are used to determine family membership of protein sequences. The HAMAP profiles are linked to manually curated HAMAP annotation rules, which specify the annotations that can be applied to members of the protein family and which are used in the automatic annotation of UniProtKB.

3. Requirements

3.1 Data requirements

One or more curated UniProtKB/Swiss-Prot records of characterized proteins known to belong to the same protein family

3.2 Software requirements

UniProt curation editor, HAMAP profile builder

3.3 Compute requirements

Windows PC, network connection

4. Procedure

4.1 Selection of seed members

The set of trusted member sequences normally includes all characterized family members from UniProtKB/Swiss-Prot, plus a representative selection of other sequences that provide broad taxonomic coverage of the target family. Sequences are selected using iterative and reciprocal BLAST searches (2), and the resulting sets are compared to those from other resources of protein families

and homologs including HOGENOM (3), OrthoDB (4), TIGRFAMs (5), Pfam (6), and PROSITE (7), and, if available, the corresponding scientific literature on the subject.

4.2 Reduction of redundancy

The resulting set of member sequences, which can be several thousands, is reduced to a manageable number for multiple sequence alignments, usually a number less than 300 but at least 20 sequences. This is done by applying filters like restricting the source of sequences to reference proteomes (http://www.uniprot.org/taxonomy/complete-proteomes), or applying computational methods like CD-HIT (8) to reduce the redundancy in protein sequences based on percentage of sequence similarity. It is important that template entries (characterized UniProtKB/Swiss-Prot entries that are used for sequence-linked annotation (feature) propagation by associated HAMAP rules) are not eliminated in this step.

4.3 Generation of seed alignment and correction of sequences

All protein sequences that are included in the seed alignment are manually checked, and where necessary corrected. This may typically involve rectification of erroneous start sites or erroneous gene model predictions. These corrections are subsequently integrated into UniProtKB/Swiss-Prot, thereby guaranteeing that the corrected sequences remain fixed. Template entries that are needed for feature propagation (see above) are noted in the header of the seed alignment (Figure 1).

4.4 Generation of the profile

The HAMAP profile builder is then used to generate the profile from the seed alignment. The redundancy in the alignment is decreased by removing a sequence from each pair of sequences that share a certain degree of residue identity. The curator can chose between values from 0%-90% sequence identity for sequence removal ('hamapseed_q' value), the standard value being 70%. However, care is taken to preserve characterized sequences in the seed set, and especially the sequence used in the rule as a template for feature propagation. Both termini of the alignment are then trimmed until a column with less than a certain amount of gaps occurs. Again, curators define the threshold for terminal column trimming from values between 0%-90% ('hamapseed_p' value), the standard value being 90%. Curators can then choose between two methods to generate the profile. Either a profile is generated by the standard procedure used in the construction of the PROSITE database by using pfmake of the PFTOOLS package (http://web.expasy.org/pftools/). The curator has the choice of which substitution matrix (BLOSUM30, BLOSUM45 (default), BLOSUM62 or BLOSUM80) (9) to apply over the alignment ('pfmake_matrix' value), in domain-global alignment mode (as described in 10, 11, and 12). The choice of the substitution matrix may influence whether the resulting profile is more specific in detecting closely related sequences or more sensitive for comparing distant related sequences. Alternatively, curators may choose to build profile HMMs with the HMMER 2 package (http://hmmer.janelia.org/software). The profile HMM will nevertheless be converted to PROSITE format profiles with htop from the PFTOOLS package, a program that transforms a HMMER ASCII-formatted HMM into an equivalent PROSITE profile. HAMAP profiles built with pfmake are optimized to be selective while being able to detect remote homologies to achieve maximal sensitivity, but profile HMMs have proven superior when this sensitivity comes at the price of specificity (absence of false positives). The method that has been used and all optional parameters chosen to construct the final profile are noted in the header section of the seed alignment (default is

pfmake if no method is specified, the default values for all other options are mentioned above and only values that differ from the defaults need to be specified) (Figure 1).

```
CLUSTAL W (1.83) multiple sequence alignment template=KCY_METJA profile_method=hmmbuild


A1R0P4_BORT9      ---------------MRIAVSSKSGCGNTTISGMLAKHYGLKLIN--YTFHDIAREKN--
B5RN17_BORDL      ---------------MRIAISSKSGCGNTTVSSMLAKYYGLKLIN--YTFHDIAREKN--
KCY2_BORBU        ---------------MKIALSGKSGCGNTTVSGMIAKHYGLEFIN--YTFHDIAREHN--
E0RT27_SPITD      ---------------MIIAISGKSGCGNSTVSRMVAERLGLRWIN--YTFRNIAQERG--
F5YHG8_TREPZ      -------MKK----DIRIAISGKSGCGNTTISRMVSDTLGLRFIN--FTFRSLAEEKG--
F5YAD1_TREAZ      -------MSNKPISETKIAISGKSGCGNTTVSRLVADALELRFIN--FTFRSLAAEKG--
KCY_TREPA         ---------------MRIAVSGASGCGNTTVSALLAERLGLPLVN--YTFRNIARELG--
KCY_AERPE         ------------------MISGPPGSGKSTYAKRLAEDLGLSYYSTGTIFRSIARERG--
KCY_STAMF         --------------MVVIVISGPPGGGKTTQARRVAEYFSLRYYSAGMIFREIARSRG--
KCY_SULAC         ---------------MKIIISGPPGSGKSSVAKILSSKLSIKYVSAGLIFRDLAKRMN--
KCY_SULTO         ---------------MIIVISGPPGSGKSTVAKILSKNLSLKYISAGHIFRELAEKEG--
KCY_SULSO         ---------------MIIIISGPPGSGKTSVAIKLANELSYKFISAGKIFRDIAQSMG--
KCY_IGNH4         --------------MTVVVISGPPGSGKSTVAKKLASELGLRFVSAGSVFRKLAEEIG--
KCY_PYRAE         --------------MVVIAVSGQPGSGKTTIAREIARVLGLPLVSSGLLFREMAARMG--
KCY_METAC         ---------------MQITVSGLPGSGTTTLSRLLSDYYELELISSGEIFRRMAKERG--
KCY_METBU         ---------------MLLTISGLPGSGTTTVGKLLAEHYSVDIISAGDVFRGLAKERG--
KCY_METB6         ---------------MRITVSGLPGSGTTSLSRYLSERYGFTMISAGEVFRQCAKEHN--
KCY_METPE         ---------------MRITISGPPGSGTTSLSKHLAAEHNLKLISAGEVFRQLAREKG--
KCY_METHJ         ---------------MRITISGLPGSGTTSLTYHLAEMHRLDVISAGEVFRQMARERG--
KCY_UNCMA         ---------------MIITLSGQPGSGKTSVAKELAEKYGFVVISAGEQFRKLAAERG--
KCY_METTP         ---------------MIITISGPPGSGTSTLARGLSEVLGVRWVNSGDLFRRIAAERG--
KCY_METJA         ---------------MIITIGGLPGTGTTTIAKMIAEKYNLRHVCAGFIFREMAKEMG--
```

Figure 1. Example seed alignment with a template entry (KCY_METJA) and the profile generation method used (hmmbuild) specified in the header

The constructed profile is calibrated using the standard PROSITE procedure. The profile is scanned against a database of randomized protein sequences from UniProtKB, and the parameters of an extreme value distribution are estimated from the score distribution obtained. These parameters are subsequently used in the normalization of the raw scores using an affine transformation (as described in 13).

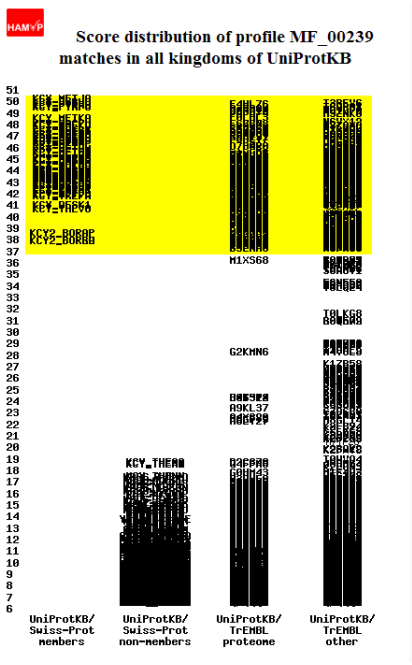
4.5 Validation of the profile

After profile construction and calibration, all matches to the profile are extracted from UniProtKB and the lowest scoring member sequence of the seed alignment is used to define an initial threshold value (or trusted cutoff score) for the normalized scores to each profile (Figure 2). Curators validate the profile by inspecting the match score distribution and cutoff value as well as the species distribution of positive matches to ascertain that the profile only hits probable true members of the protein family and that it is in line with the current knowledge or the eventual literature available for the protein family.

Figure 2. Typical output of a profile builder scan, showing the score distribution of matches, the options chosen for building the profile, and a list of matching proteins

Curators may choose to alter the composition of the original seed alignment in order to enhance the specificity of the family profile, performing iterative profile searches until a satisfactory score distribution is obtained (see also Figure 5).

The cutoffs are validated by inspecting the distribution of non-member scores (true negatives), which are the scores of characterized and annotated UniProtKB/Swiss-Prot records known to belong to a different protein family (Figure 3).



List of UniProtKB/Swiss-Prot true positive matches for the profile MF_00239
(*i.e.* HAMAP MF_00239 UniProtKB/Swiss-Prot members)

ac	id	kingdom	score	score_diff	description	organism
Q58071	KCY_METJA	Archaea	50.213	+12.972	Cytidylate kinase	Methanocaldococcus jannaschii (strain ATCC 43067 / DSM 2661 / JAL-1 / JCM 10045 / NBRC 100440)
O58988	KCY_PYRHO	Archaea	49.932	+12.691	Cytidylate kinase	Pyrococcus horikoshii (strain ATCC 700860 / DSM 12428 / JCM 9974 / NBRC 100139 / OT-3)
Q9UZJ6	KCY_PYRAB	Archaea	49.667	+12.426	Cytidylate kinase	Pyrococcus abyssi (strain GE5 / Orsay)
Q8U2L4	KCY_PYRFU	Archaea	49.514	+12.273	Cytidylate kinase	Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1)
Q8TZB3	KCY_METKA	Archaea	48.542	+11.301	Cytidylate kinase	Methanopyrus kandleri (strain AV19 / DSM 6324 / JCM 9639 / NBRC 100938)
O28379	KCY_ARCFU	Archaea	48.011	+10.77	Cytidylate kinase	Archaeoglobus fulgidus (strain ATCC 49558 / VC-16 / DSM 4304 / JCM 9628 / NBRC 100126)
Q5JJE3	KCY_THEKO	Archaea	47.867	+10.626	Cytidylate kinase	Thermococcus kodakaraensis (strain ATCC BAA-918 / JCM 12380 / KOD1)
B9LPV4	KCY_HALLT	Archaea	47.441	+10.2	Cytidylate kinase	Halorubrum lacusprofundi (strain ATCC 49239 / DSM 5036 / JCM 8891 / ACAM 34)
A7I5R5	KCY_METB6	Archaea	47.433	+10.192	Cytidylate kinase	Methanoregula boonei (strain 6A8)
Q3IMV8	KCY_NATPD	Archaea	47.159	+9.918	Cytidylate kinase	Natronomonas pharaonis (strain ATCC 35678 / DSM 2160)
C6A187	KCY_THESM	Archaea	47.159	+9.918	Cytidylate kinase	Thermococcus sibiricus (strain MM 739 / DSM 12597)

Figure 3. Graphical view of the match score distribution of a HAMAP profile. Matching proteins are sorted in UniProtKB/Swiss-Prot members, UniProtKB/Swiss-Prot non-members, UniProtKB/TrEMBL proteomes as well as others to assess the occurrence of false negative or false positive matches in the UniProtKB database.

Curators can manually adjust this cutoff to include lower scoring member sequences, or raise it to reduce the possibility of false positive matches (Figure 4).

```
ID   Cytidyl_kinase_type2; MATRIX.
AC   MF_00239;
DT   DEC-2013 (DATA UPDATE).
DE   Cytidylate kinase [cmk].
CC   /VERSION=4;
MA   /GENERAL_SPEC: ALPHABET='ACDEFGHIKLMNPQRSTVWY'; LENGTH=191; LOG_BASE=1.071779; P0=0.9972;
MA       P=   7.552363,   1.698108,   5.303439,   6.320015,   4.078187,   6.844419,   2.240667,
MA   /DISJOINT: DEFINITION=PROTECT; N1=6; N2=186;
MA   /NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=1.03045; R2=0.0080365; TEXT='-LogE';
MA   /CUT_OFF: LEVEL=1; SCORE=2514; N_SCORE=37.241; MODE=1; TEXT='!';
MA   /CUT_OFF: LEVEL=0; SCORE=2513; N_SCORE=37.241; MODE=1; TEXT='?';
MA   /CUT_OFF: LEVEL=-1; SCORE=-1061; N_SCORE=8.5; MODE=1; TEXT='??';
MA   /DEFAULT: B0=*; B1=*; E0=*; E1=*; M1=*; I1=*;
MA   /I: B0=-85; B1=-85; BD=-49;
MA   /M: SY='M'; M=-25,-24,-49,-42,8,-41,-30,-3,-10,-4,46,-37,-41,-35,-8,-32,-6,6,-28,-25; M0=-
MA   /I: MM=0; MI=-100; MD=-111; IM=-9; II=-11; DM=-7; DD=-14;
MA   /M: SY='R'; M=-24,-25,-39,-33,-2,-39,-27,20,0,-1,-17,-3,-39,-4,21,3,-24,11,-29,-25; M0=-12;
MA   /I: MM=0; MI=-100; MD=-111; IM=-9; II=-11; DM=-7; DD=-14;
MA   /M: SY='I'; M=-44,-39,-71,-68,-43,-70,-70,39,-68,-13,-30,-66,-66,-66,-69,-64,-44,4,-62,-57;
MA   /I: MM=0; MI=-100; MD=-111; IM=-9; II=-11; DM=-7; DD=-14;
MA   /M: SY='T'; M=10,11,-59,-53,-34,-51,-43,3,-50,-29,5,-49,-52,-47,-49,-43,31,15,-41,-37; M0=-
MA   /I: MM=0; MI=-101; MD=-111; IM=-9; II=-11; DM=-7; DD=-14;
MA   /M: SY='I'; M=-44,-39,-71,-68,-44,-70,-71,35,-68,-4,-31,-67,-67,-67,-70,-64,-44,20,-63,-58;
M0=-44;
```

Figure 4. The normalized threshold scores are stored in the profile and can be adjusted manually

4.6 Putting profiles in production

Once the HAMAP family profile has passed validation, the seed alignment and the profile are integrated into the HAMAP database (Figure 5).

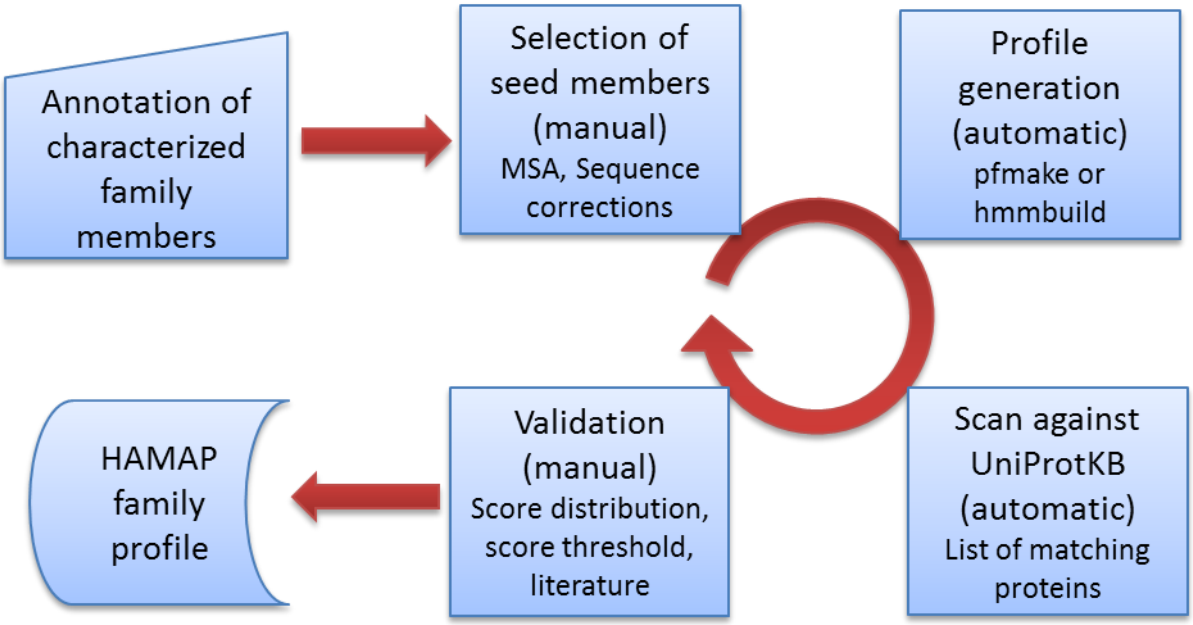


Figure 5. Typical workflow of a HAMAP family profile generation

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. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. J Mol Biol, 215, 403-410.
3. Penel, S., Arigon, A.M., Dufayard, J.F., Sertier, A.S., Daubin, V., Duret, L., Gouy, M. and Perriere, G. (2009) Databases of homologous gene families for comparative genomics. BMC Bioinformatics, 10 Suppl 6, S3.
4. Waterhouse, R.M., Zdobnov, E.M., Tegenfeldt, F., Li, J. and Kriventseva, E.V. (2011) OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011. Nucleic Acids Res, 39, D283-288.
5. 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.

6. 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.
7. Sigrist, C.J., Cerutti, L., de Castro, E., Langendijk-Genevaux, P.S., Bulliard, V., Bairoch, A. and Hulo, N. (2010) PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res, 38, D161-166.
8. Huang, Y., Niu, B., Gao, Y., Fu, L. and Li, W. (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics, 26, 680-682.
9. Henikoff, S., Henikoff, J.G. (1991) Automated assembly of protein blocks for database searching. Nucleic Acids Res, 19, 6565-6572.
10. Lüthy, R., Xenarios, I., Bucher, P. (1994) Improving the sensitivity of the sequence profile method. Protein Sci, 3, 139-146.
11. Bucher, P., Karplus, K., Moeri, N., Hofmann, K. (1996) A flexible motif search technique based on generalized profiles. Computers Chem, 20, 3-23.
12. Sigrist, C.J., et al. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. Brief. Bioinform, 3, 265-274.
13. Pagni, M. and Jongeneel, C.V. (2001) Making sense of score statistics for sequence alignments. Brief Bioinform, 2, 51-67.

8. Revision history

Version	Author	Date	Change made
1.0	Ivo Pedruzzi	1 October 2013	Established SOP
1.1	Ivo Pedruzzi	1 May 2014	Added more options to 4.4, Generation of the profile. Updated Figure 2