Article

HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot

DE OLIVEIRA LIMA, Tania, *et al*.

Abstract

The growth in the number of completely sequenced microbial genomes (bacterial and archaeal) has generated a need for a procedure that provides UniProtKB/Swiss-Prot-quality annotation to as many protein sequences as possible. We have devised a semi-automated system, HAMAP (High-quality Automated and Manual Annotation of microbial Proteomes), that uses manually built annotation templates for protein families to propagate annotation to all members of manually defined protein families, using very strict criteria. The HAMAP system is composed of two databases, the proteome database and the family database, and of an automatic annotation pipeline. The proteome database comprises biological and sequence information for each completely sequenced microbial proteome, and it offers several tools for CDS searches, BLAST options and retrieval of specific sets of proteins. The family database currently comprises more than 1500 manually curated protein families and their annotation templates that are used to annotate proteins that belong to one of the HAMAP families. On the HAMAP website, individual sequences as well as whole genomes [...]

Reference

DE OLIVEIRA LIMA, Tania, *et al*. HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. *Nucleic acids research*, 2009, vol. 37, no. Database issue, p. D471-8

DOI : 10.1093/nar/gkn661
PMID : 18849571

Published online 11 October 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D471–D478***
doi:10.1093/nar/gkn661

HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot

Tania Lima*, Andrea H. Auchincloss, Elisabeth Coudert, Guillaume Keller, Karine Michoud, Catherine Rivoire, Virginie Bulliard, Edouard de Castro, Corinne Lachaize, Delphine Baratin, Isabelle Phan, Lydie Bougueleret and Amos Bairoch

Swiss-Prot Group, Swiss Institute of Bioinformatics, 1 rue Michel-Servet, 1211 Geneva 4, Switzerland

Received August 28, 2008; Accepted September 19, 2008

ABSTRACT

The growth in the number of completely sequenced microbial genomes (bacterial and archaeal) has generated a need for a procedure that provides UniProtKB/Swiss-Prot-quality annotation to as many protein sequences as possible. We have devised a semi-automated system, HAMAP (High-quality Automated and Manual Annotation of microbial Proteomes), that uses manually built annotation templates for protein families to propagate annotation to all members of manually defined protein families, using very strict criteria. The HAMAP system is composed of two databases, the proteome database and the family database, and of an automatic annotation pipeline. The proteome database comprises biological and sequence information for each completely sequenced microbial proteome, and it offers several tools for CDS searches, BLAST options and retrieval of specific sets of proteins. The family database currently comprises more than 1500 manually curated protein families and their annotation templates that are used to annotate proteins that belong to one of the HAMAP families. On the HAMAP website, individual sequences as well as whole genomes can be scanned against all HAMAP families. The system provides warnings for the absence of conserved amino acid residues, unusual sequence length, etc. Thanks to the implementation of HAMAP, more than 200 000 microbial proteins have been fully annotated in UniProtKB/Swiss-Prot (HAMAP website: www.expasy.org/sprot/hamap).

INTRODUCTION

The increasing number of completely sequenced microbial genomes represents an unparalleled opportunity to achieve a better understanding of prokaryotes, including their metabolic pathways, virulence factors, phylogeny, etc. However, the sequences themselves are not enough. It is of fundamental importance that these genomes be annotated with high quality and that the nomenclature be standardized.

Since the publication in 1995 of the complete *Haemophilus influenzae* genome (1), more than 700 bacterial and archaeal genomes have been entirely sequenced; the development of new sequencing techniques, such as parallel pyrosequencing of 454 Life Sciences (2) and Solexa/Illumina Genome Analyzer sequencing-by-synthesis technology (3), has greatly increased the amount of sequenced data that is generated, and they complement the classic Sanger DNA sequencing method (4). Public databases currently hold more than 100Gb of sequence and this amount will continue to increase exponentially as sequencing centres will soon have an annual throughput of several gigabases each.

Most of the proteins coming from these sequencing projects will probably never be characterized, and the annotation at the DNA level is succinct. Sequencing centres have developed automated pipelines from a combination of methods, such as sequence similarity, presence of domains and pathway prediction, among many other sequence analysis methods usually employed (5) to attempt to annotate the proteome of a certain microorganism. Though the prediction of coding sequences (CDSs) is usually very good, the quality of the functional annotation attached to them is very variable.

Many methods have been developed to improve genome functional annotation, including the use of genomic

*To whom correspondence should be addressed. Tel: +41 22 379 5050; Fax: +41 22 379 5858; Email: tania.lima@isb-sib.ch
Present address:
Isabelle Phan, Seattle Biomedical Research Institute (SBRI), 307 Westlake Avenue North, Seattle, WA 98109-2591, USA

context information (6), mapping of pathways in orthologous groups (7), or defining protein function based on protein–protein interactions (8). Genome annotation by the scientific community using Wiki software has lately been the focus of several initiatives (9–11), but one of the major hurdles is the establishment of common standards for the annotation provided by each expert. Since sequencing centres and users in general rely on large protein databases, and especially on UniProtKB/Swiss-Prot (12), to annotate new genomes and identify new proteins, we consider it to be an important mission of UniProtKB to provide as many annotated proteins as possible, with the highest possible quality.

In order to address this need, we have implemented HAMAP (High-quality Automated and Manual Annotation of microbial Proteomes), a semi-automated pipeline system within UniProtKB/Swiss-Prot, dedicated to high-throughput, high-quality annotation of proteins from microbial complete proteomes, that also provides complete proteome sets that are consistent and non-redundant. Its aim is to maximize the complementarity between manual and automated annotation; the HAMAP system is composed of two databases and an automatic annotation pipeline. It targets proteins from bacteria, archaea and plastids, the latter being included due to their bacterial origin.

On the HAMAP website (www.expasy.org/sprot/hamap), two databases are available: one that provides curated information on all bacterial, archaeal and plastid proteomes—only fully sequenced and assembled genomes submitted to the public databases and whose CDSs have been annotated are taken into account—and a family database that contains all manually created protein families and annotation templates (also called 'family rules'). There is also a tool for user-derived complete protein annotation (protein recommended name, gene name, function, subunit, membership to a protein family, sequence features, etc., as specified in the family annotation template) that is provided upon submission of either one protein sequence, if it belongs to one of the HAMAP families, or of a complete genome even before submission to the public DNA databases. Since the system provides not only annotation, but also warnings regarding atypical N-termini, lack of conserved residues and many other features, we believe that this tool can help the scientific community in the annotation of whole microbial genomes or any protein from bacteria, archaea and plastids.

THE PROTEOME DATABASE

The proteome database (www.expasy.org/sprot/hamap/proteomes.html) is developed jointly with the UniProt team at the European Bioinformatics Institute. Its aim is to provide, in a relational database, information on the biology, genome and taxonomy of each completely sequenced proteome that has been submitted to the public DNA databases. Whole-genome-shotgun genomes (WGSs) are not incorporated into the proteome database.

On the 'HAMAP proteomes' homepage, a list of all available proteomes is provided, plus a link to all sequenced microorganisms that are known to interact with other organisms (for example, a list of sequenced strains that are avirulent, animal intracellular parasites, plant symbionts, etc).

A page is provided for each complete proteome added; this page contains three sections: general information, genome(s) sequenced, and tools.

(i) The 'General Information' section contains:

- taxonomic information;
- information on the biology and genomics of the sequenced strain; and
- presence of some morphological characteristics.

(ii) The 'Genome(s) sequenced' section describes all DNA elements (chromosome and plasmids), with links to the DNA database and the reference to the paper, if the genome has been published, plus links to external databases that refer to the genome in question. This database is constantly updated: as papers are published, the references are added to the database and to the UniProtKB entries themselves.

(iii) The 'Tools' section contains:

- the genome viewer, which allows the user to see the CDSs encoded on a particular region of the sequenced genome;
- BLAST searches against all proteins from the proteome;
- a link to download all UniProtKB entries for the proteome, either in UniProtKB format or in FASTA format;
- a link to retrieve all characterized or identified proteins from the proteome; this is based on the 'Protein existence' line present in each UniProtKB entry [for details see (12)]; and
- a link to retrieve all proteins from the proteome for which a 3D structure is available.

This database is extensively curated in several aspects: plasmids (which are not always submitted simultaneously with the chromosome sequences) are attached to the proteome sets to form complete genomes; extensive information on the sequenced strain is presented; cross-references to relevant sites are manually added and maintained, as is information on genome publications. The complete proteome sets presented contain both annotated entries from UniProtKB/Swiss-Prot and from its supplement, UniProtKB/TrEMBL (12).

At the time of writing, the proteome database contained pages for 622 bacterial proteomes, 53 archaeal proteomes and 133 plastid proteomes.

THE FAMILY DATABASE

The HAMAP annotation system was designed (13) to propagate manually generated annotation to all members of a given protein family in an automated, but controlled way. The system is based on protein families and their annotation templates, which are created manually by

curators (see below) and which are used as the annotation template for the propagation of annotation to members of a protein family. Members of HAMAP families are identified using a profile collection (see below).

Three types of protein families are dealt with by the HAMAP annotation system:

(1) Proteins that belong to well-characterized families, a family being a manually compiled collection of orthologs. Their function is known, i.e. has been described for at least one or several members, and has been well studied in one or more species;
(2) UPFs, i.e. uncharacterized protein families, are conserved proteins found in several species but for which no function is known at present; and
(3) proteins belonging to complex families, such as ABC transporters.

The main components of the HAMAP annotation system are the protein families and their annotation templates, the alignments and the profiles that are generated from them, and the annotation pipeline. Each component is explained in the following sections.

HAMAP protein families and annotation templates

The annotation templates are manually created and contain all the annotation that will be propagated to the members of a family. In order to create the annotation template, all characterized proteins that belong to this family are manually annotated according to UniProtKB/Swiss-Prot standards; this means that curators perform a thorough, detailed and in-depth review of the existing literature on a certain protein, including proteins from genomes that are not fully sequenced. The information available on these proteins provides the contents of each annotation template (family rule), and these proteins are listed in the field 'Template' in each family rule (see below). Most available papers are read and used to annotate the characterized proteins. This manual annotation and additional BLAST similarity searches (14) are used by curators to define what information can be safely propagated to other prokaryotes and to manually select the set of member sequences that will be used to build the seed alignment. In other words, curators determine the nature and extent of the annotation that can be propagated to orthologs.

The advantage of the manual intervention by curators who continuously revise the existing literature is that annotation templates and protein families are periodically revised to ensure that the annotation is as up-to-date as possible, and also to ensure that the organisms represented are as divergent as possible. This is important for the generation and maintenance of profiles. Also, if a curator comes across experimental evidence that contradicts the propagated annotation, the entire family is revised and the annotation template is updated taking into account the new available experimental evidence. Manual curation ensures that most available experimental knowledge is represented in the database, even though this is a slow, time-consuming process that usually lags behind the pace at which new evidence becomes available.

At present, more than 1500 protein families and their annotation templates are available on the HAMAP website (www.expasy.org/sprot/hamap/families.html).

Each annotation template for a HAMAP protein family (Figure 1) has a unique identifier of the format MF_xxxxx. They contain several fields, among which (for detailed information on all the fields present in HAMAP annotation templates see www.expasy.org/unirule/unirule_web_view.html#General):

- general information, such as last revision date;
- annotation that can be propagated to all members, such as protein name (which usually includes only the recommended name of a protein, but can also include some alternative, synonymous names if appropriate); gene name when available; general annotation lines such as function, catalytic activity, subunit, subcellular location, PTMs and the name of the family to which the protein belongs, among other information; keywords; relevant sequence features, such as active sites, metal-binding residues, domains, topology, etc.;
- Gene Ontology (GO) terms (15), which are manually selected by the curators after thorough review of the existing literature and of the available terms;
- cross-references to PROSITE (16), Pfam (17), TIGRFAMs (18), PRINTS (19) and/or PIRSF (20);
- UniProtKB accession numbers of all entries (templates) that were manually annotated and for which there is experimental evidence or structural data that was used to build the family and its annotation template; and
- sets of member sequences divided by taxonomic groups.

The use of conditional statements ('cases' and conditions) ensures that the annotation is only applied where appropriate, to guarantee the production of annotation of the same quality as that produced by manual curation (see Figure 2 for some examples).

Cases and conditions are derived from relevant biological information collected from the literature; cases can restrict the propagation of annotation to a specific taxonomic group, for example, or be dependent (in this case a 'condition' statement exists in the annotation template) on the presence of a specific amino acid residue, or group of residues, for the annotation to be propagated. The annotation templates are designed to perform numerous checks on the sequences themselves as well, such as sequence length, aberrant N-termini, absence of expected sequence features, among others.

On the website, the protein families and their annotation templates can be browsed by protein name, gene name, pathway, scope (archaeal, bacterial and/or plastid families), etc.

Alignments and profiles

Once the seed members of a protein family are manually selected, the sequences are aligned using ClustalW (21), MUSCLE (22) or T-Coffee (23). The alignments are manually verified, and sometimes manually edited. The sequences themselves are also manually corrected if

Figure 1. Example of a HAMAP protein family annotation template (family rule), MF_00074 (www.expasy.org/unirule/MF_00074). Annotation templates contain three sections: 'General rule information', 'Propagated annotation' and 'Additional information'. General information comprises: family identification number (MF_xxxxx), dates of creation and revision, 'Data class', i.e. that the whole protein is annotated by the family rule and not only a specific domain, and 'Predictors', which contain the distribution of matches and the alignment that was used to generate the family profile. The 'Propagated annotation' section contains the information that is propagated to all members of a protein family, or to some, if the field is

appropriate, for example if they result from a frameshift or are too long or too short at their N-terminus. The alignments are used both for the automated generation of identification profiles used to generate family matches [for details see (13)], and for the propagation of sequence features by similarity to the template sequence.

The whole collection of HAMAP profiles can be downloaded by ftp at ftp.expasy.org/databases/hamap/.

Detailed explanations about the database, the fields in the annotation templates and the annotation pipeline in general, plus a comprehensive user manual, can be found in the 'Documents' section (www.expasy.org/sprot/hamap/hamap_doc.html).

THE ANNOTATION PIPELINE

The annotation pipeline was set up to optimize the interaction between programs and curators and to ensure that 'problematic' sequences will always be re-directed to manual check and curation. The aim is to propagate annotation as carefully as possible; built-in checks and limitations will prevent a protein sequence from being annotated in case of doubt. The aim is always to achieve quality rather than maximal coverage.

In brief, the system works as follows (Figure 3): after a complete genome is deposited in DDBJ/EMBL/GenBank (24) entries are produced containing the original annotation that was provided by the submitter, plus, in some cases, automatically added additional annotation. These entries are stored in UniProtKB/TrEMBL, the unreviewed section of UniProtKB. All microbial and plastid protein sequences in UniProtKB/TrEMBL are run daily against the HAMAP profile collection and family members are identified. Matches with a score above the cutoff are annotated using the annotation templates and are integrated into UniProtKB/Swiss-Prot; problematic proteins (for example, sequences having unusual length, missing conserved amino acid residues or having aberrant N-termini) generate warnings and are channeled to manual review and annotation.

UniProtKB/Swiss-Prot entries that belong to a HAMAP family, i.e. manually curated templates and entries that are the product of the automated annotation pipeline, can be identified by the cross-reference to HAMAP and the corresponding family number (in the 'Cross references' field, under 'Family and domain databases', MF_xxxxx).

TOOLS

On the website 'Tools' section, several analysis and retrieval tools are available: users can scan one protein sequence or a whole genome against the collection of HAMAP families; specific sets can be retrieved (characterized or identified proteins from specific proteomes, or sequences for which there are 3D structures available).

Submission of sequences or genomes for analysis

On the HAMAP tools page (www.expasy.org/sprot/hamap/index.html#tools), sequences can be submitted and checked whether they belong to any HAMAP family.

Two types of scan can be performed: 'quick scan', for one or a few sequences, and 'advanced scan', for whole microbial genomes.

After submission, results are displayed on the website. If a sequence hits one or more HAMAP families (a distinction is made between a 'true' membership, which is above the trusted cut-off, and a 'weak' match, below the trusted cut-off), the user is directed to the corresponding protein family and its annotation template containing the annotation that is applied to the respective family members.

If a whole genome is submitted, the results are password-protected and can be retrieved on the 'HAMAP Scan results' page, with full annotation and warnings regarding N-termini that are too long or too short, absence of conserved amino acid residues (which can be useful to check potential sequencing errors or frameshifts), absence of expected domains, etc.

Retrieval of sets of characterized/existent proteins or with 3D structures

With this tool, users can retrieve specific sets of proteins for which some characterization is available, i.e. the protein has been found to exist through mass spectrometry, in 2D gels, etc., or for which there is some literature, according to standards defined by the UniProtKB 'Protein Existence' line (12). A typical use would be to retrieve all 'characterized' proteins of a bacterium or archaeon (for example, retrieve all 'characterized' proteins of *H. influenzae*, or for a group of organisms, such as enterobacteria). The same can be done for retrieval of proteins for which there is at least one 3D structure available.

CONCLUSION

The HAMAP database makes available to the scientific community and genome sequencing centres a collection of manually curated microbial protein families and profiles that can be useful for the functional annotation of protein sequences or microbial genomes. The automated pipeline can be used to detect occasional sequence errors by making use of the warnings generated by the system.

preceded by 'cases' or 'conditions'. For MF_00074, the function field will be different depending on the taxonomic origin, but all proteins will have 'Cytoplasm' as subcellular location and all belong to the family 'RNA methyltransferase rsmG'. It also contains cross-references to other protein family databases, such as Pfam and TIGRFAMS, and manually selected GO terms. Additional information includes the size range of members of this family, if there are protein families related to this one, the list of characterized protein(s) that were used to compile information for the creation of the protein family and its annotation template (for MF_00074, literature is found for the proteins of *E. coli*, *Bacillus subtilis*, *Microbacterium tuberculosis* and *Streptomyces coelicolor*), the scope, i.e. the taxonomic groups covered by this family, if in at least one member this protein is fused to another protein either in the N-terminal or C-terminal region, and whether there are duplicates or whether in some species the protein is encoded on a plasmid. In the 'UniProtKB rule member sequences' section, complete sets of member proteins can be retrieved, taxonomic distribution can be browsed, and specific sets of proteins can be retrieved.

```
From MF_00112

case <OC:Archaea>
ID   GGGPS
DE   RecName: Full=Geranylgeranylglyceryl phosphate synthase;
DE            Short=GGGP synthase;
DE            Short=GGGPS;
DE            EC=2.5.1.41;
DE   AltName: Full=(S)-3-O-geranylgeranylglyceryl phosphate synthase;
DE   AltName: Full=Phosphoglycerol geranylgeranyltransferase;
end case
case <OC:Bacillales>
ID   PCRB
DE   RecName: Full=Putative glycerol-1-phosphate prenyltransferase;
DE            EC=2.5.1.-;
GN   Name=pcrB;
end case

From MF_01544

case <Property:Membrane=1>
CC   -!- SUBCELLULAR LOCATION: Cell membrane; Single-pass membrane protein
CC       (Potential).
end case
case <Property:Membrane=2>
CC   -!- SUBCELLULAR LOCATION: Cell inner membrane; Single-pass membrane
CC       protein (Potential).
end case

From MF_01624

FT   From: ARSC_STAAU (P0A006)
FT   ACT_SITE     10     10       Nucleophile; for reductase activity and
FT                                phosphatase activity (By similarity).
FT   Condition: C
FT   ACT_SITE     82     82       Nucleophile; for reductase activity (By
FT                                similarity).
FT   Condition: C
FT   ACT_SITE     89     89       Nucleophile; for reductase activity (By
FT                                similarity).
FT   Condition: C
FT   DISULFID     10     82       Redox-active; alternate (By similarity).
FT   Condition: C-x*-C
FT   DISULFID     82     89       Redox-active; alternate (By similarity).
FT   Condition: C-x*-C

From MF_01339

FT   From: RBL2_RHORU (P04718)
FT   ACT_SITE    166    166       Proton acceptor (By similarity).
FT   Condition: K
FT   ACT_SITE    287    287       Proton acceptor (By similarity).
FT   Condition: H
FT   METAL       191    191       Magnesium; via carbamate group (By
FT                                similarity).
FT   Group: 1; Condition: K
FT   METAL       193    193       Magnesium (By similarity).
FT   Group: 1; Condition: D
FT   METAL       194    194       Magnesium (By similarity).
FT   Group: 1; Condition: E
FT   MOD_RES     191    191       N6-carboxylysine (By similarity).
FT   Condition: K
XX
```

Figure 2. Examples of uses of the conditional statements 'case' and 'conditions' in family annotation templates (family rules). MF_00112 (www.expasy.org/unirule/MF_00112): an example of ID/protein name/gene name propagation depending on taxonomic distinction. In archaea, no gene name has been assigned but enzyme function has been proven in several different species, whereas the gene name pcrB is used only in Bacillales, with a function that has only been suggested for *B. subtilis*. Note that the reaction catalyzed by the archaeal protein has no biological significance in bacteria, since GGGP is a specific precursor of archaeal membrane lipids. MF_01544 (www.expasy.org/unirule/MF_01544): Subcellular location is predicted based on the number of membranes the bacterium possesses. MF_01624 (www.expasy.org/unirule/MF_01624): an example of conditions used for active site and disulfide bond feature propagation. If the indicated amino acid(s) are not present in the appropriate position(s) in the sequence, the feature is not propagated and a warning is generated, necessitating manual intervention. MF_01339 (www.expasy.org/unirule/MF_01339): an example of active site, metal and modified residue feature propagation. In the last two examples, the template entry used to derive the information is also indicated.

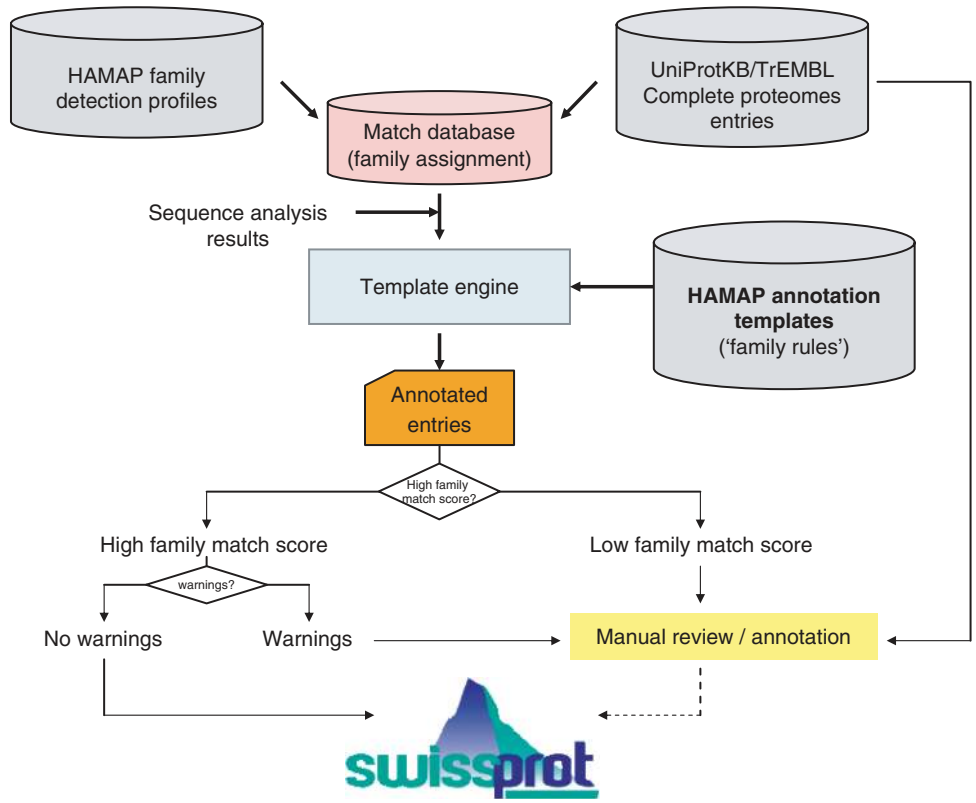

Figure 3. The HAMAP annotation pipeline. UniProtKB/TrEMBL complete proteome entries matching a HAMAP family detection profile (derived from an alignment of manually selected family members; matches to those profiles are stored in a 'Match database', allowing assignment of family membership) are passed through a 'template engine' that applies the annotation found in the corresponding HAMAP annotation template (and resolving its conditional statements) to generate UniProtKB/Swiss-Prot annotation. If the system generates warnings, or if the matching score is low, the entry is channelled to manual annotation; entries without warnings are directly integrated into UniProtKB/Swiss-Prot. UniProtKB entries for which there is available literature are manually annotated.

The HAMAP system as a whole has greatly increased the speed at which microbial protein sequences are annotated in UniProtKB/Swiss-Prot and we believe that this has been achieved without lowering the standards for which UniProtKB/Swiss-Prot is renowned. The coverage of HAMAP families keeps increasing as new families are manually created—at the moment, about 25% of the *Escherichia coli* K-12 proteins belong to a HAMAP family.

We hope that the HAMAP resource can help the annotation of complete genomes, improving the quality of CDS prediction and functional annotation.

The development of the system and its website is an ongoing effort and future plans include the addition of phylogenetic analysis to help establish true orthology, checks of consistency within pathways and taking into account the conservation of gene neighborhoods, improvements in the generation of identification profiles and, especially, the coverage of all housekeeping genes.

ACKNOWLEDGEMENTS

We wish to thank Alexandre Gattiker for the design and implementation of the initial HAMAP pipeline, Sandrine Pilbout for all the taxonomy-related work, Nicole Redaschi, Thomas Kappler and Paul Kersey for database management, and Nicolas Hulo and Christian Sigrist for help with alignments and profiles.

FUNDING

The Swiss-Prot group is part of the Swiss Institute of Bioinformatics (SIB) and of the UniProt Consortium. Swiss-Prot group activities are supported by the Swiss Federal Government through the Federal Office of Education and Science, and by the National Institutes of Health (grant 2 U01 HG02712-04). Additional support comes from the European Commission contract FELICS (021902RII3) and from PATRIC BRC (NIH/NIAID contract HHSN 266200400035C). Funding for open access charges: Swiss Federal Government through the Federal Office of Education and Science.

Conflict of interest statement. None declared.

REFERENCES

1. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.-F., Dougherty,B.A., Merrick,J.M. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.

2. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J., Chen,Z. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
3. Bennett,S. (2004) Solexa Ltd. *Pharmacogenomics*, **5**, 433–438.
4. Sanger,F., Nicklen,S. and Coulson,A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA*, **74**, 5463–5467.
5. Stothard,P. and Wishart,D.S. (2006) Automated bacterial genome analysis and annotation. *Curr. Opin. Microbiol.*, **9**, 505–510.
6. Wolf,Y.I., Rogozin,I.B., Kondrashov,A.S. and Koonin,E.V. (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res.*, **11**, 356–372.
7. Mao,F., Su,Z., Olman,V., Dam,P., Liu,Z. and Xu,Y. (2006) Mapping of orthologous genes in the context of biological pathways: an application of integer programming. *Proc. Natl Acad. Sci. USA*, **103**, 129–134.
8. Marcotte,E.M., Pellegrini,M., Ng,H.L., Rice,D.W., Yeates,T.O. and Eisenberg,D. (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**, 751–753.
9. Salzberg,S.L. (2007) Genome re-annotation: a wiki solution? *Genome Biol.*, **8**, 102–102.
10. Elsik,C.G., Worley,K.C., Zhang,L., Milshina,N.V., Jiang,H., Reese,J.T., Childs,K.L., Venkatraman,A., Dickens,C.M., Weinstock,G.M. *et al.* (2006) Community annotation: procedures, protocols, and supporting tools. *Genome Res.*, **16**, 1329–1333.
11. Mons,B., Ashburner,M., Chichester,C., van Mulligen,E., Weeber,M., den Dunnen,J., van Ommen,G.J., Musen,M., Cockerill,M., Hermjakob,H. *et al.* (2008) Calling on a million minds for community annotation in Wiki proteins. *Genome Biol.*, **9**, R89–R89.
12. UniProt Consortium (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*, **36**, D190–D195.
13. Gattiker,A., Michoud,K., Rivoire,C., Auchincloss,A.H., Coudert,E., Lima,T., Kersey,P., Pagni,M., Sigrist,C.J., Lachaize,C. *et al.* (2003) Automated annotation of microbial proteomes in SWISS-PROT. *Comput. Biol. Chem.*, **27**, 49–58.
14. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
15. Gene Ontology Consortium (2008) The Gene Ontology project in 2008. *Nucleic Acids Res.*, **36**, D440–D444.
16. Hulo,N., Bairoch,A., Bulliard,V., Cerutti,L., Cuche,B.A., de Castro,E., Lachaize,C., Langendijk-Genevaux,P.S. and Sigrist,C.J. (2008) The 20 years of PROSITE. *Nucleic Acids Res.*, **36**, D245–D249.
17. Sammut,S.J., Finn,R.D. and Bateman,A. (2008) Pfam 10 years on: 10,000 families and still growing. *Brief. Bioinform.*, **9**, 210–219.
18. Haft,D.H., Selengut,J.D. and White,O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.*, **31**, 371–373.
19. Attwood,T.K., Bradley,P., Flower,D.R., Gaulton,A., Maudling,N., Mitchell,A.L., Moulton,G., Nordle,A., Paine,K., Taylor,P. *et al.* (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.*, **31**, 400–402.
20. Wu,C.H., Nikolskaya,A., Huang,H., Yeh,L.S., Natale,D.A., Vinayaka,C.R., Hu,Z.Z., Mazumder,R., Kumar,S., Kourtesis,P. *et al.* (2004) PIRSF: family classification system at the Protein Information Resource. *Nucleic Acids Res.*, **32**, D112–D114.
21. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTALW: 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.
22. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113–113.
23. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
24. Kulikova,T., Akhtar,R., Aldebert,P., Althorpe,N., Andersson,M., Baldwin,A., Bates,K., Bhattacharyya,S., Bower,L., Browne,P. *et al.* (2007) EMBL nucleotide sequence database in 2006. *Nucleic Acids Res.*, **35**, D16–D20.