AUTOMATISATION IN UNIPROTKB / SWISS-PROT ANNOTATION: NEW RULES AND TOOLS

[1] The HAMAP team, [1] The PROSITE team, [1] Edouard de Castro, [1] Alan Bridge, and the UniProt Consortium.

[1] Swiss-Prot Group, Swiss Institute of Bioinformatics, CMU, 1 Michel Servet, 1211 Geneva 4.

Introduction

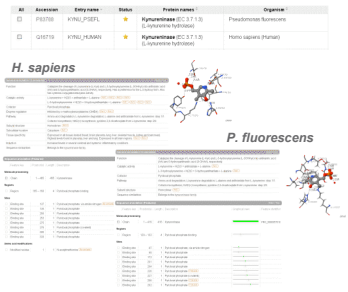
The development of next generation sequencing technologies promises a massive increase in the rate of submission of new protein sequences to sequence databases such as the Universal Protein Resource Knowledge Base, UniProtKB. At UniProtKB/Swiss-Prot we propose to meet this challenge by continuing to expand and develop systems for the automatic propagation of existing annotation to newly submitted protein sequences. These developments will promote the standardization of ortholog annotation both across and within kingdoms and significantly enhance our ability to accurately annotate new protein sequences which are being produced at an ever increasing rate.

The HAMAP pipeline and automatic annotation of eukaryotic proteins

The existing prototype for automatic annotation in Swiss-Prot is the HAMAP project (High-quality Automated and Manual Annotation of microbial Proteomes http://www.expasy.org/sprot/hamap/index.html), which specifically targets microbial proteins. We have recently modified the existing HAMAP annotation pipeline to allow the automatic annotation of eukaryotic proteins. Steps 1 to 4 below illustrate a typical workflow for the construction of an **annotation template** or **family rule**. Such rules serve as the source of annotations for new entries that match an associated sequence profile. Steps 5a-c describe the application of a new tool for the automatic creation of such family rules from an existing set of manually annotated UniProtKB/Swiss-Prot entries.

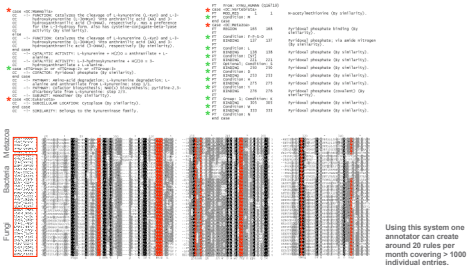
We have identified eukaryotic proteins that are candidates for automatic annotation in two ways. First, during a recent annotation marathon targeting conserved proteins of the slime mold *Dictyostelium discoideum*, we manually annotated over 1000 proteins with orthologs in at least two other major eukaryotic phyla (*Homo sapiens/Mus musculus* and/or *Drosophila melanogaster* and/or *Saccharomyces cerevisiae/Schizosaccharomyces pombe*). *D. discoideum* separated from metazoa and fungi prior to the metazoa:fungi split, and so annotation of *D. discoideum* orthologs provides an ideal opportunity for the identification of common annotations suitable for transfer to other eukaryotic orthologs. **Second**, we analyzed the coverage of eukaryotic proteins by existing HAMAP annotation rules. HAMAP includes 1551 family rules covering 214838 proteins (UniProtKB/Swiss-Prot Release 56.7 of 20-Jan-2009). Of these rules, 167 potentially match vertebrate proteins, while 228 match fungi and 147 match arthropods. The taxonomic scope of these existing HAMAP rules could conceivably be expanded to cover eukaryotes.

1. Manually annotated UniProtKB/Swiss-Prot templates
Experimentally characterized proteins belonging to defined protein families.



2. Manual creation of seed alignment and family rule from existing template(s).

3. Annotation template (family rule) and associated sequence profile
The annotation template contains textual annotations such as protein nomenclature and functional information, including terms from controlled vocabularies, as well as annotations associated with defined positions in the sequence. Both types of annotations may be conditional upon the protein satisfying certain criteria such as membership of a particular taxonomic grouping (*) or the presence of particular residues (*).



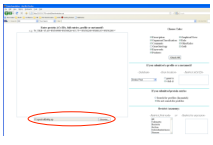
Using this system one annotator can create around 20 rules per month covering > 1000 individual entries.

4. a) Automatic retrieval and annotation of sequences that match the family profile.

4. b) Manual check of warnings – missing features, possible erroneous initiator codon, etc.

5. Automatic creation of family rule.

5. a) Select existing entries.



5. b) Examine variability of existing annotations.

5. c) Automatic extraction of common annotations. Check proposed annotations, specify conditions where appropriate.

6. Automatically annotated UniProtKB/Swiss-Prot entries.

UniProt is mainly supported by the National Institutes of Health (NIH) grant 2 U01 HG02712-04. Additional support for the EBI's involvement in UniProt comes from the European Commission (EC)'s FELICS grant (021902RII3) and from the NIH grant 1R01HGO2273-01. Swiss-Prot activities at the SIB are supported by the Swiss Federal Government through the Federal Office of Education and Science and the European Commission contract FELICS (021902RII3). PIR activities are also supported by the NIH grants and contracts HHSN266200400061C, NCI-caBIG, and 1R01GM080646-01, and the National Science Foundation (NSF) grant IIS-0430743. PROSITE is supported by the Swiss National Science Foundation, grant n. 315230-116864 and the InterPro grant n. 213037 from the European Union.

Contact
help@uniprot.org
www.uniprot.org

See also:
www.isb-sib.ch