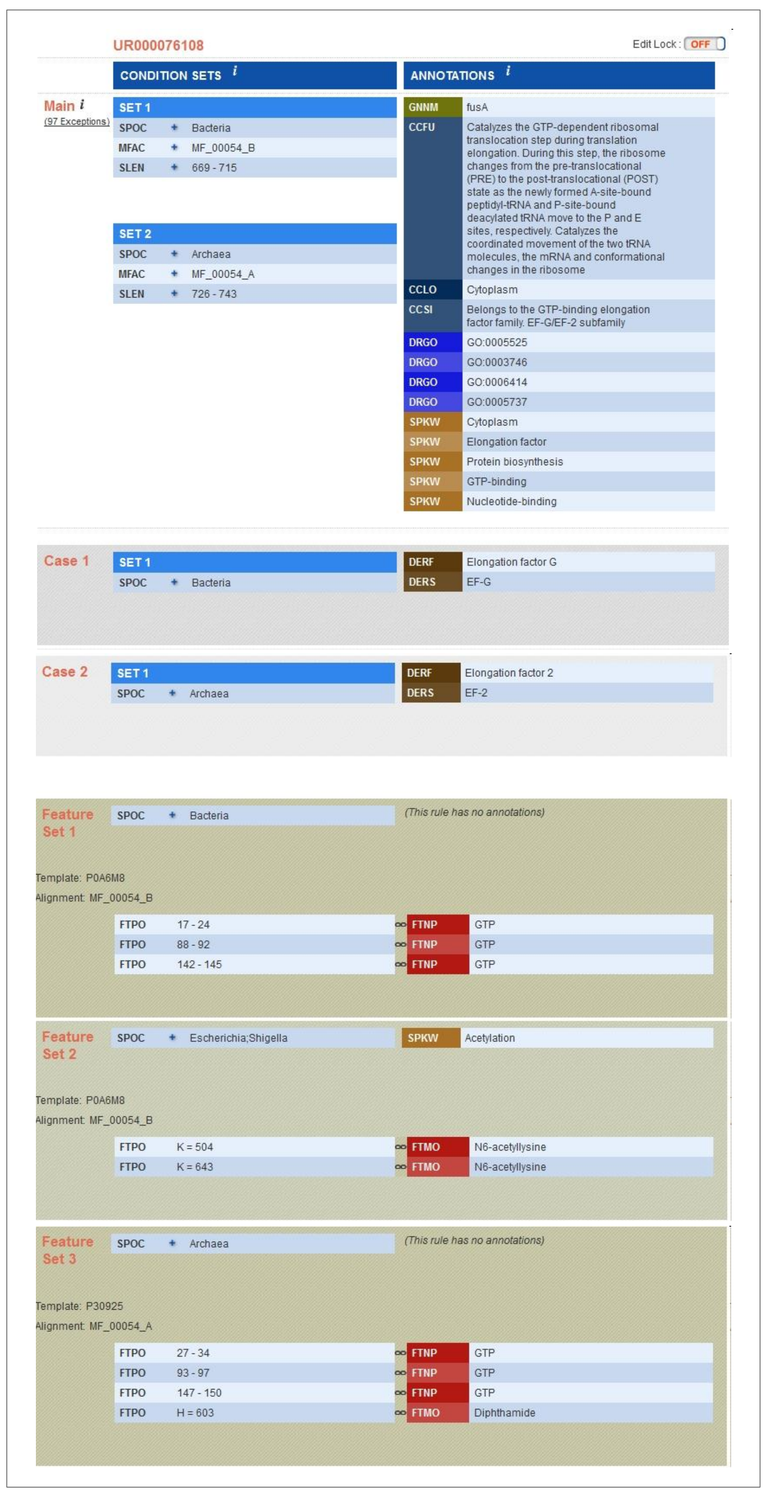
Cecilia Arighi[3], Ivo Pedruzzi[2], Klemens Pichler[1] and the UniProt Consortium[1,2,3]
[1]EMBL-European Bioinformatics Institute, Cambridge, UK
[2]SIB Swiss Institute of Bioinformatics, Geneva, Switzerland
[3]Protein Information Resource, Georgetown University, Washington DC & University of Delaware, USA

The UniRule system in UniProtKB - leveraging manual Swiss-Prot annotation to TrEMBL

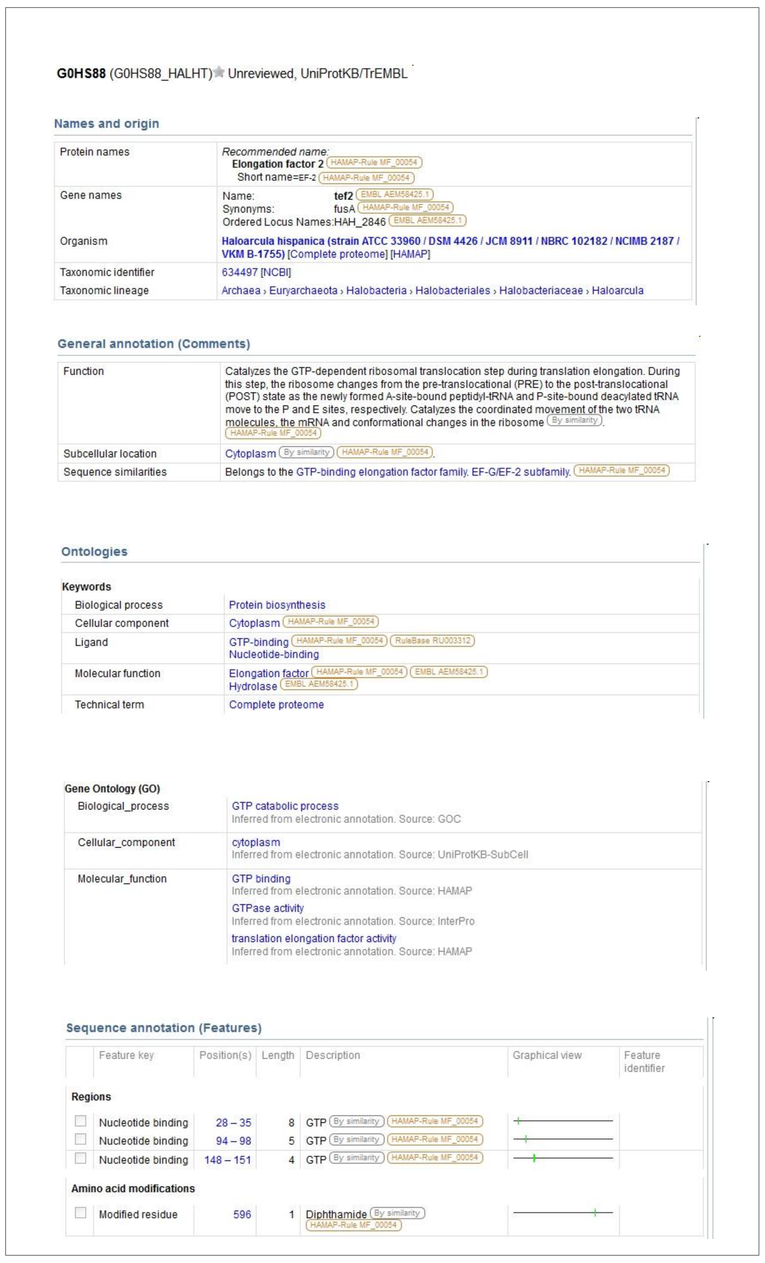
The UniProt Knowledgebase

The UniProt Knowledgebase (UniProtKB) is a freely accessible resource for functional information on proteins with accurate, consistent, rich and comprehensive annotation. In addition to manually reviewed proteins (Swiss-Prot), UniProtKB holds a large number of nucleotide-derived protein sequences, many of which are annotated by automatic annotation procedures (TrEMBL).



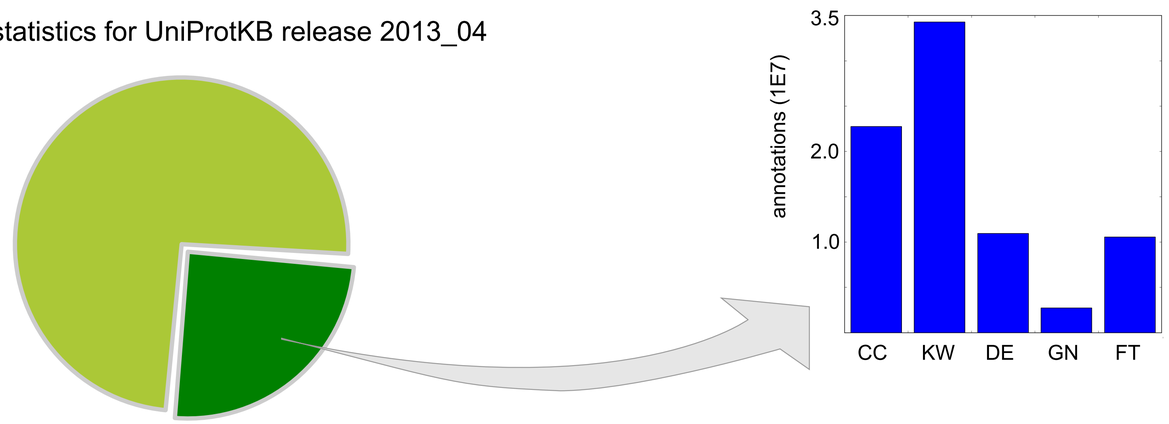
Challenges of data growth...

Currently, reviewed entries constitute only about 2% of UniProtKB; manual curation is time-intensive and a lot of published experimental data focuses on a rather limited range of model organisms. At the same time, an increasing number of new sequences is added to UniProtKB (TrEMBL) at ever increasing speed. However, for a large proportion of those new sequences there is no experimental data available.



Left, screenshot of a rule in the UniRule tool. The rule contains case statements for taxon-specific protein names and sequence features.
Top, screenshot of an entry with the above rule applied to it. Annotations propagated by the annotation rule are clearly tagged

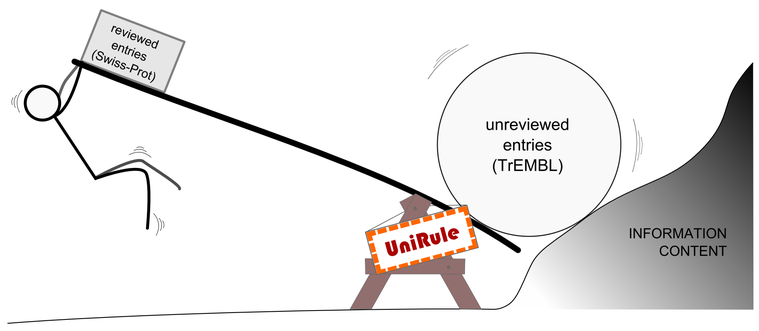
UniRule statistics for UniProtKB release 2013_04



No. applied rules:	3167
No. entries hit:	8,398,405
No. entries in TrEMBL:	32,153,798

CC: functional annotation
KW: keywords
DE: protein names
GN: gene names
FT: sequence features

... and a solution



The UniRule system leverages manual curation for the automatic annotation of unreviewed UniProtKB entries. UniRule integrates rules from historically distinct automatic annotation systems (HAMAP, PIRNR/PIRSR, Rulebase) in a single pipeline. It consists of manually created annotation rules that specify functional annotations and the conditions which must be satisfied for them to be applied, such as taxonomic scope, family membership as defined by InterPro and the presence of specific sequence features.

UniRule tool

The UniRule tool is the interface used by curators. It is used for both the creation and maintenance of rules. It holds all the rules in the UniRule system and allows the editing of rules individually or in batch mode. Statistics on rule performance are calculated on the fly. Rules can be imported and exported in XML format.

Rule creation

Rule creation is part of an integrated workflow that starts with the manual curation of UniProtKB/Swiss-Prot records. Protein names, gene names, functional annotation, GO terms, keywords and sequence features are selected and combined with InterPro signatures and taxonomic restrictions to create a rule. Protein family signatures can be custom-built if necessary. Once created, fully annotated TrEMBL entries hit by the rule can be previewed directly in the UniRule tool.

Rule maintenance

The UniRule tool provides a range of statistics for evaluation of rule performance (confidence, sensitivity, coverage). Rules performing suboptimally are flagged and are then reviewed by curators, thus ensuring the timely correction of any inaccuracies.

Rule application

Rules are applied anew with each UniProtKB release, keeping the propagated annotation up-to-date. On the UniProtKB website, information added by automatic annotation procedures is clearly highlighted as such.

People

R. Antunes[1], J. Argasinska[1], C. Arighi, A.H. Auchincloss[2], D. Baratin[2], E. Barrera[1], A. Bridge[2], E. Coudert[2], B.A. Cuche[2], E. de Castro[2], F. Fazzini[1], E. Hatton-Ellis[1], G. Keller[2], K. Laiho[3], M. Martin[1], A. McDougall[1], D. Natale[3], C. O'Donovan[1], I. Pedruzzi[2], K. Pichler[1], D. Poggioli[1], C. Rivoire[2], C.R. Vinayaka[3], T. Wardell[1], L.-S. Yeh[3]

Funding:
UniProt is funded by the European Molecular Biology Laboratory, National Institutes of Health, European Union, Swiss Federal Government, British Heart Foundation and National Science Foundation

Email: help@uniprot.org
URL: www.uniprot.org