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

EMBL-EBI

UniRule - Increasing Annotation Depth of Unreviewed Protein Entries in UniProtKB

The UniProt Knowledgebase

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information. One of the databases provided by UniProt is the UniProt Knowledgebase (UniProtKB) which contains expertly reviewed (UniProtKB/Swiss-Prot) and unreviewed (UniProtKB/TrEMBL) records.

Challenges of data growth and a solution

Currently, reviewed records constitute only about 1% of UniProtKB; expert curation is time-intensive and most published experimental data focuses on a rather limited range of model organisms. At the same time, the number of unreviewed records in UniProtKB (UniProtKB/TrEMBL) is growing continuously, yet for a large proportion of these records there is no experimental data available.
The UniRule system leverages expert curation for the automatic annotation of unreviewed UniProtKB records. 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.

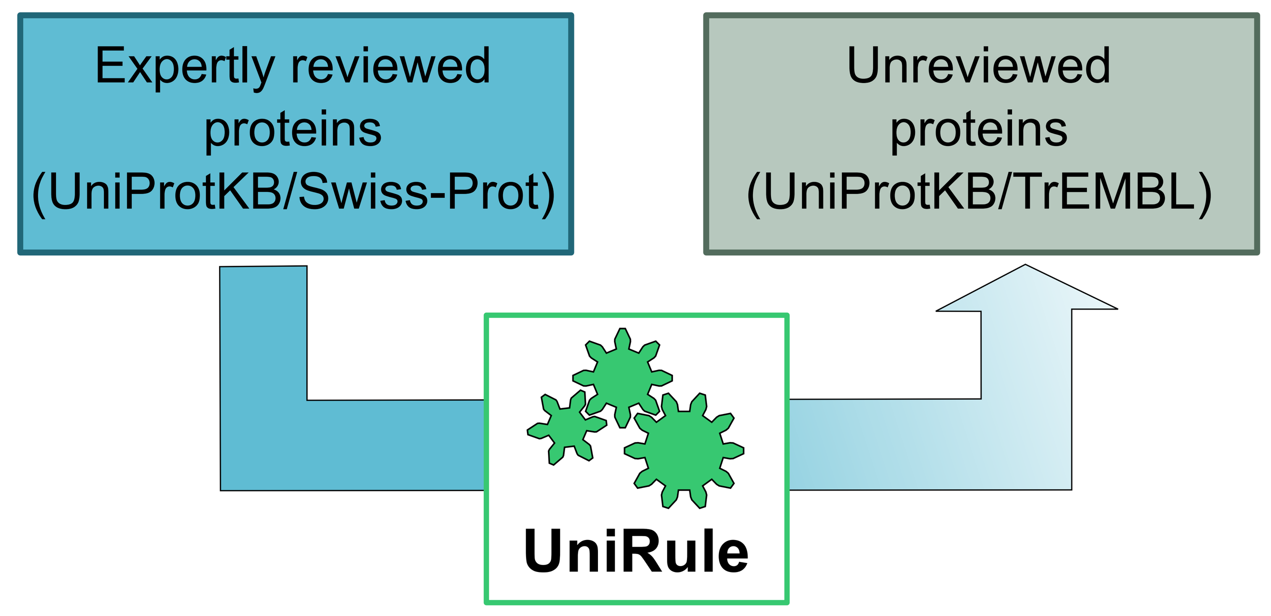


Figure 1: Dataflow via UniRule

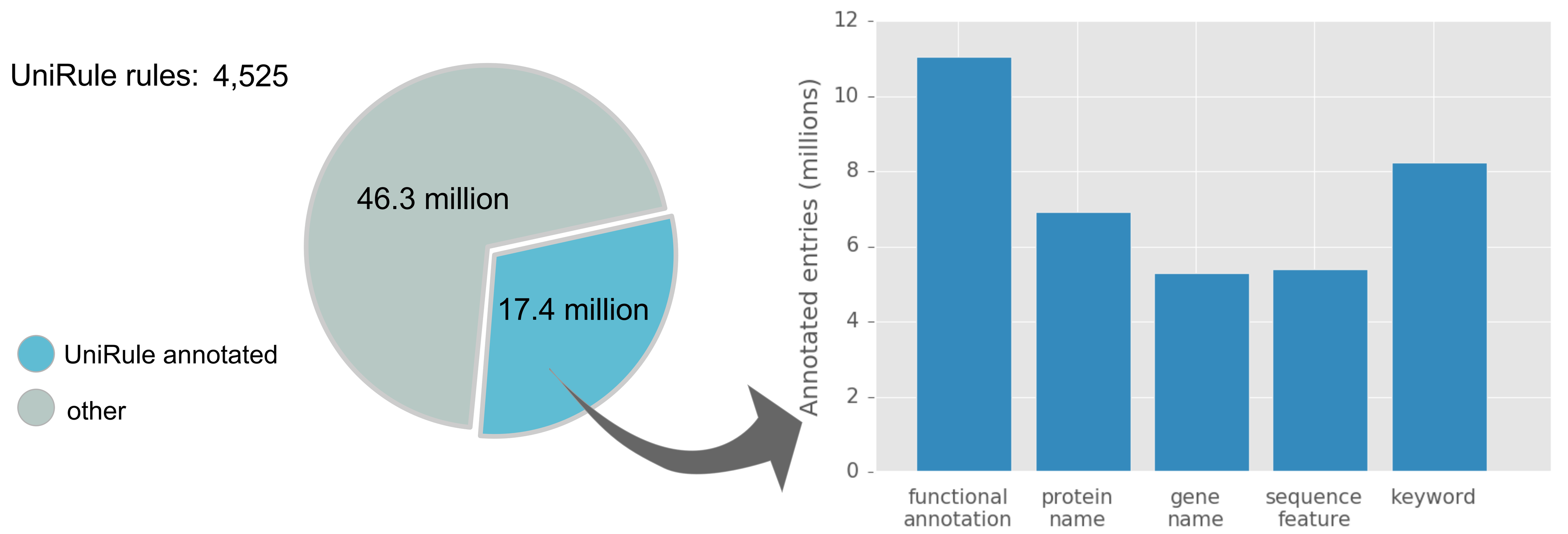


Figure 2: UniRule statistics for UniProtKB release 2016_04

Data types propagated by UniRule

Rule creation

Rule creation is part of an integrated workflow that starts with the expert 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.

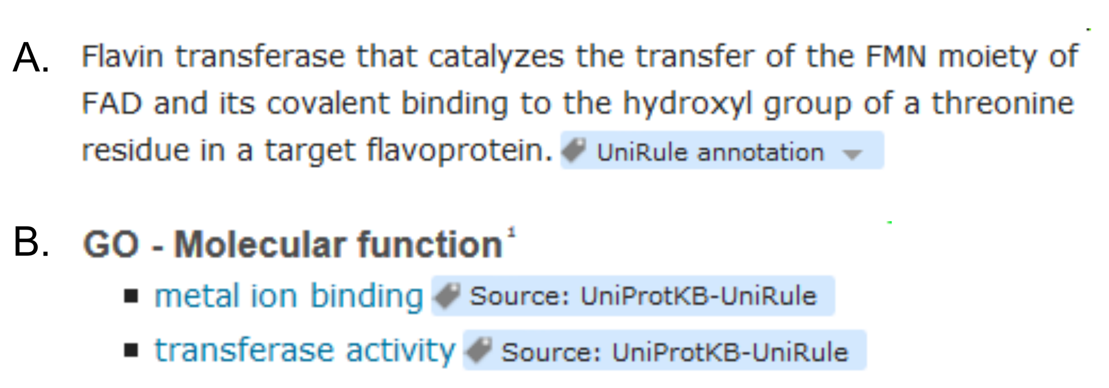


Figure 3: Evidence tags pointing to a UniRule rule. A, Annotation describing protein function. B, Associated Gene ontology terms

UniRule tool

The UniRule tool is a web interface used by curators 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.
Rules can be built either step by step with the tool providing validation and assistance or they can be imported using an XML exchange format. The tool also exports to this format.
Statistics on rule performance are calculated against all reviewed UniProtKB entries on the fly. This way, the effects of individual changes to a rule can be assessed. In case the effects are found to be detrimental to rule performance, a 'history' functionality allows previous versions to be reinstated.

People

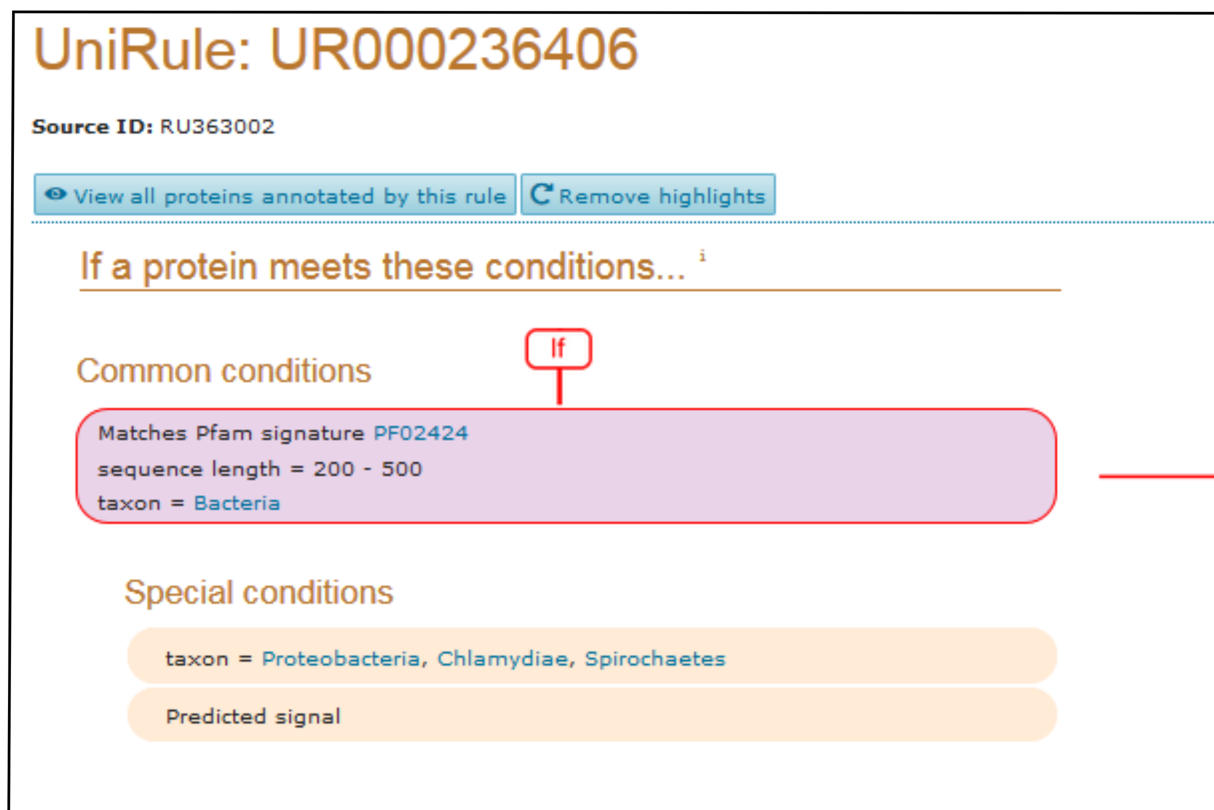
R. Antunes[1], C. Arighi[3], D. Baratin[2], A. Bridge[2], E. Coudert[2], B.A. Cuche[2], E. de Castro[2], J.S. Garavelli[3], E. Hatton-Ellis[1], G. Keller[2], K. Laiho[3], M. Martin[1], A. MacDougall[1], C. O'Donovan[1], I. Pedruzzi[2], K. Pichler[1], D. Poggioli[1], L. Pureza[1], N. Redaschi[2], A. Renaux[1], C. Rivoire[2], C.R. Vinayaka[3], V. Volynkin[1], Q. Wang[3], L.-S. Yeh[3], H. Zellner[1]

Rule maintenance

A range of statistics are used to evaluate the performance of rules (confidence, sensitivity, coverage). Statistics are recalculated after every change made to a rule and with every release. Rules performing suboptimally are flagged and are then reviewed by curators, thus ensuring the timely correction of any inaccuracies.

Rule application

Rules are reapplied 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 using evidence tags (see Figure 3). The tags can also be used as search terms to specifically search for and/or filter out annotation added by a rule.



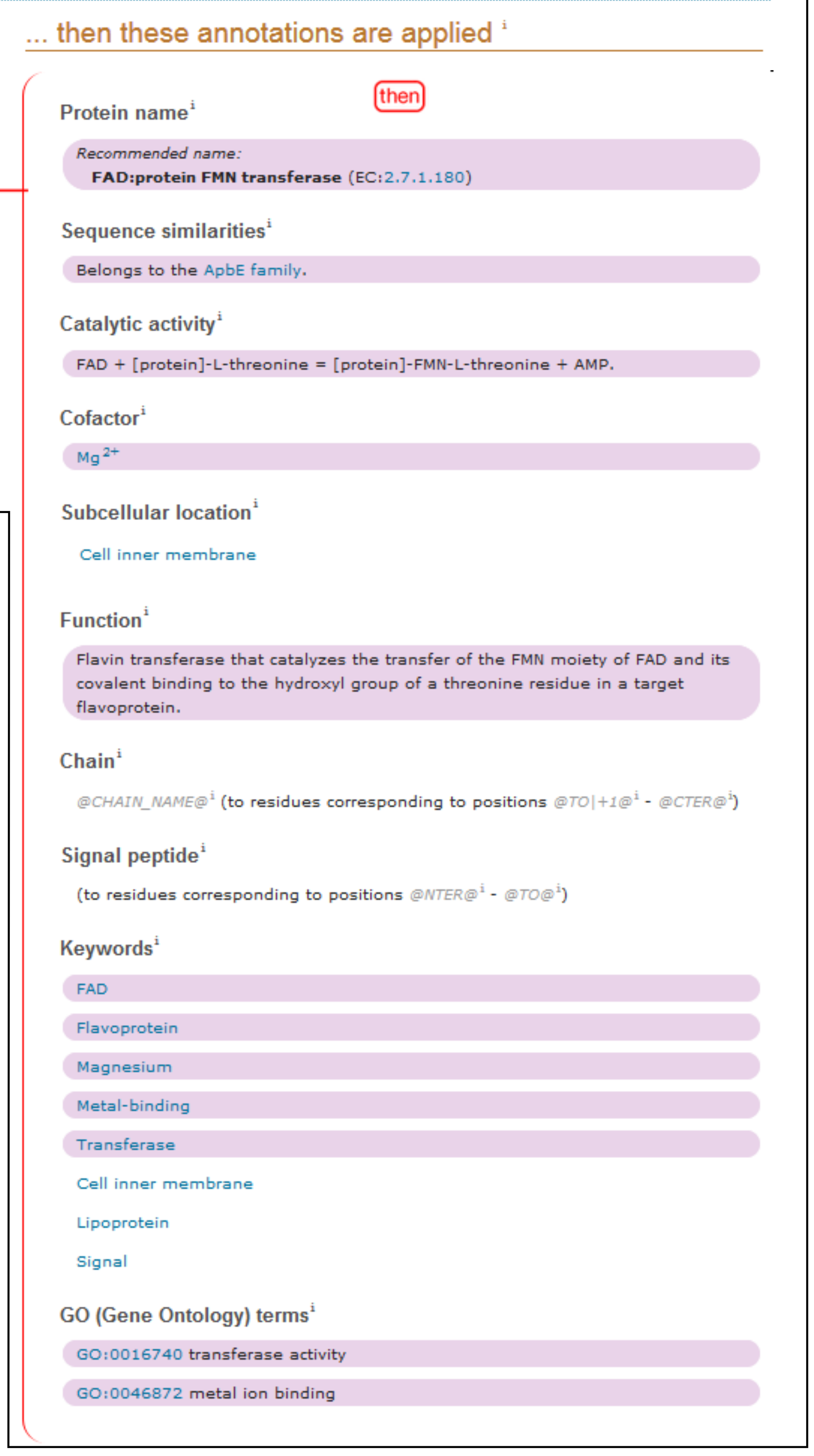


Figure 4: Screenshot of a UniRule rule on uniprot.org

UniRule data in UniProtKB

UniRule-generated annotation in UniProtKB and all rules underlying the data can be explored. Clicking on conditions or annotations will highlight links between the two, thus explaining the rule's logic.
Figure 4 shows an example of a UniRule rule as it is displayed on uniprot.org including rule identifiers for reference and a quick link to a tabular view of all unreviewed entries touched by this rule. The rule defines the overall sequence space it covers using a Pfam signature, a taxonomic restriction to bacteria and constraints on the length of target sequences ('common conditions'). The bulk of the annotations is propagated if the common conditions are met (this is highlighted in the screenshot). Special conditions apply for propagation of the subcellular location comment. Furthermore, the rule calls a predictive algorithm (SignalP) for annotating potential signal peptides.

Funding: UniProt is funded by National Institutes of Health, European Molecular Biology Laboratory, Swiss Federal Government, British Heart Foundation, Parkinson's Disease United Kingdom and National Science Foundation

UniProt EMBL-EBI PIR SIB Swiss Institute of Bioinformatics

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