EMBL-EBI

European Bioinformatics Institute
is an Outstation of the
European Molecular Biology Laboratory.

European Bioinformatics Institute

Rule-based annotation in UniProtKB

Klemens Pichler[1], Diego Poggioli[1], Claire O'Donovan[1] and the UniProt consortium[1,2,3]

[1]EMBL-European Bioinformatics Institute, Cambridge, UK
[2]Swiss Institute of Bioinformatics, Geneva, Switzerland
[3]Protein information Resource, Washington DC, USA

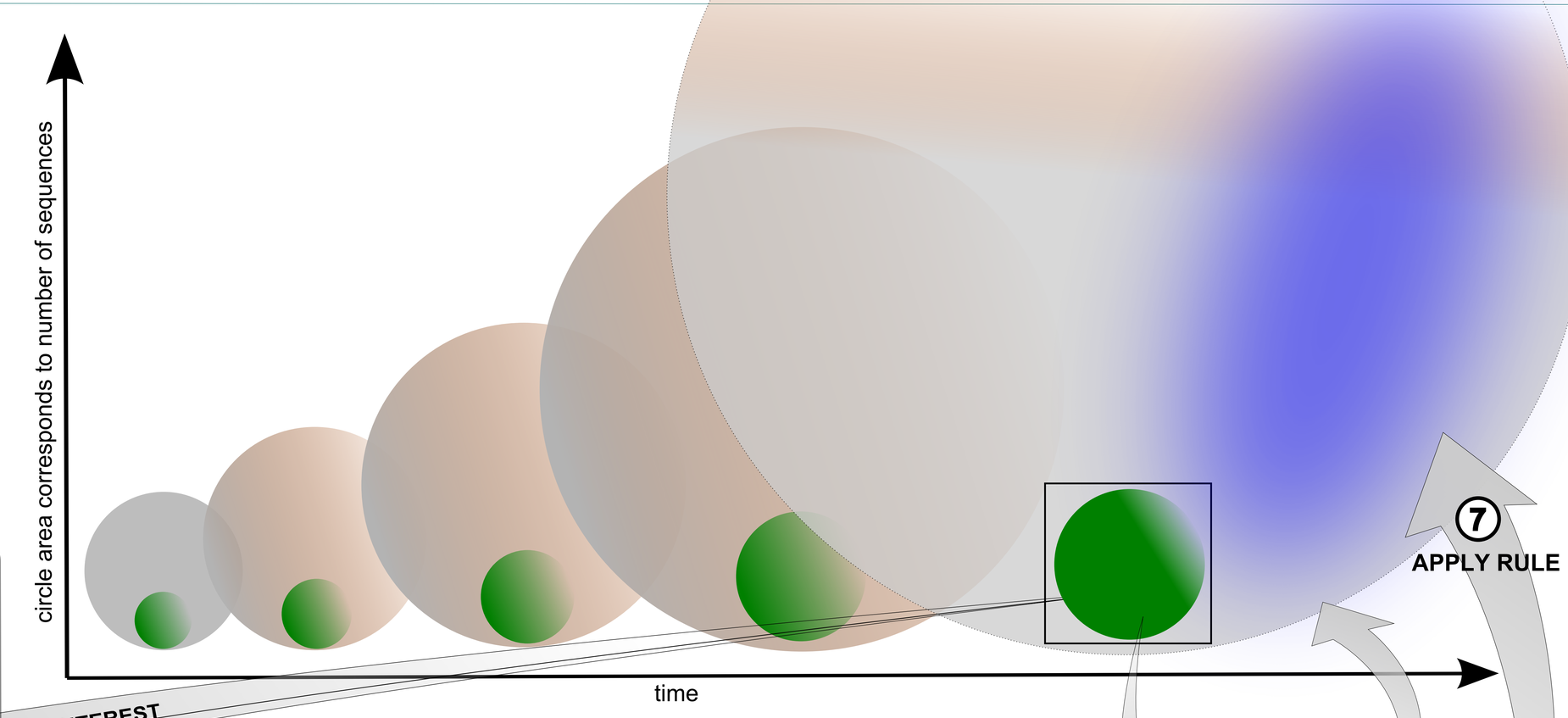
UniProt

THE CHALLENGE

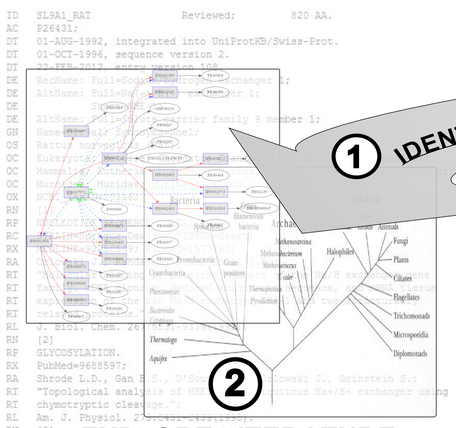
> increasing number of new sequences added at ever increasing speed
> no experimental data for a large proportion of new sequences

VS

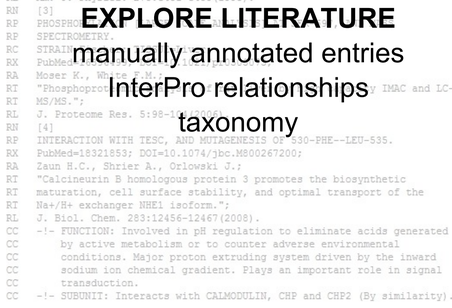
> manual curation is time-intensive
> a lot of published experimental data focuses on limited range of model organisms



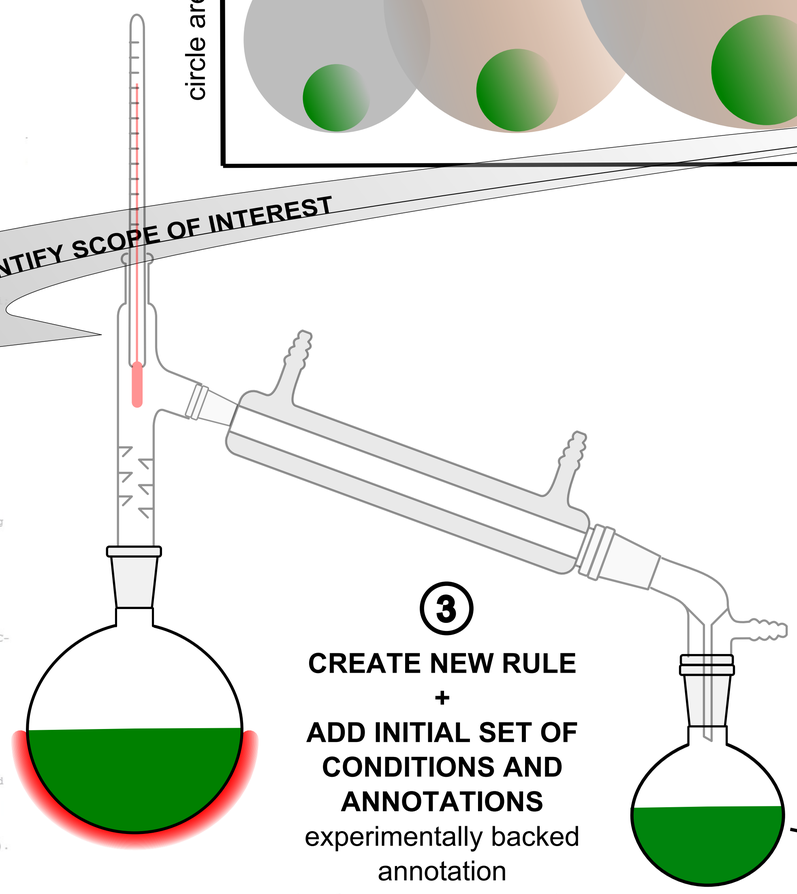
THE SOLUTION



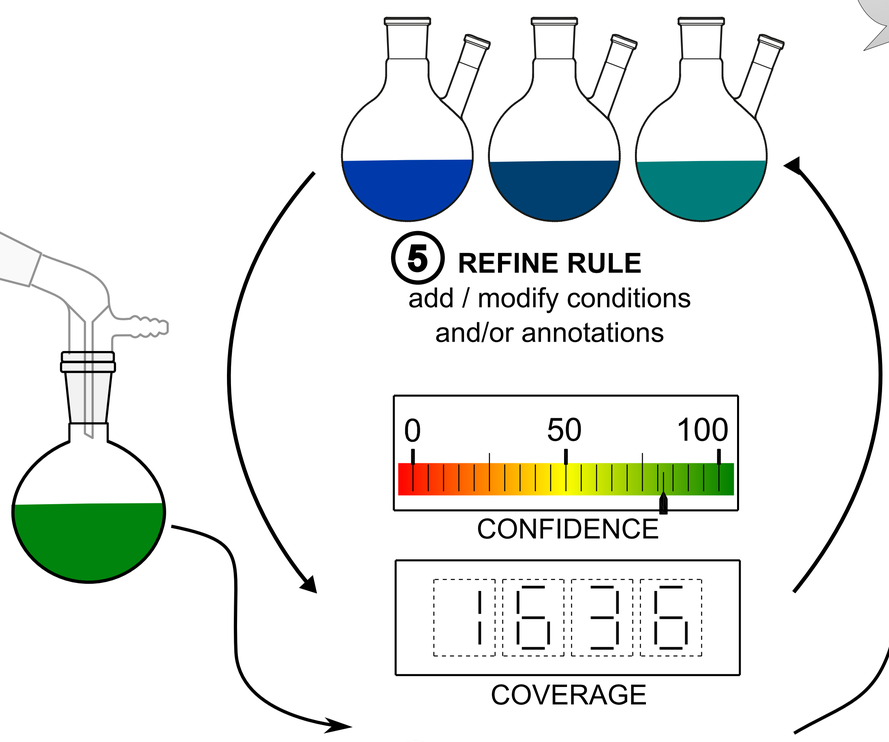
1 IDENTIFY SCOPE OF INTEREST



2
EXPLORE LITERATURE
manually annotated entries
InterPro relationships
taxonomy

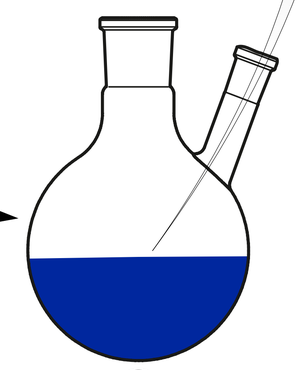


3
CREATE NEW RULE
+
ADD INITIAL SET OF CONDITIONS AND ANNOTATIONS
experimentally backed annotation
InterPro member signatures
taxonomy



4
RUN STATISTICS
check confidence
evaluate false positives (FP) and false negatives (FN)
evaluate coverage

5
REFINE RULE
add / modify conditions and/or annotations



6
CHECK BEHAVIOUR OF APPLIED RULE
+
SAVE RULE

DETAILS OF RULE CREATION

1 **Identify scope for rules** through: manual curation work, a jamboree, user requests, collaborations, output from SAAS, a list of InterPro entries that have not yet been used in a rule.

2 **Filter and evaluate data** (e.g. using scripts parsing relevant annotated entries): consistent protein names, gene names, functional annotation, GO terms, keywords and sequence features.

Can an existing rule be extended to cover more entries or propagate more annotation? Can it be complemented with a more specific rule, e.g. family vs. subfamily?

Annotations and protein family signatures can be custom built, too!

3 **Create and maintain rules in a web-based tool** providing lots of specialized functionality.
Rules have conditions and annotations.
Conditions: InterPro member signatures, sequence and proteome properties, taxonomy
Annotations: all types of data deemed safe to propagate automatically.

4 **Run and evaluate statistics** which are computed based on manually annotated entries and generated on the fly. Only rules with 100% confidence are put into production, i.e. where all the manually annotated entries meet all the conditions and

5 **Balance specificity of annotation and coverage** (number of entries hit by a rule) while maintaining high confidence. Rules failing the statistics are flagged for review.

6 A **preview of a rule** as it is applied to a certain entry is being implemented. Saving a rule generates a history, which allows reverting to previous versions.

7 **Rules are applied anew with each UniProtKB release** keeping propagated annotation up-to-date.

S The Statistical Automated Annotation System (SAAS) uses decision trees to generate simple rules.

NEW **Annotation of sequence features:**
1. using predictors
> secretory signal sequences (SignalP 4.0) *
> transmembrane domains (TMHMM) *
> * Phobius as a discriminator
> coiled-coils (coils)
2. using alignments
active sites, motifs, binding sites for metals/ nucleotides, binding regions etc.

NEW **Case statements** extend a given rule by allowing additional annotation, specific to only a subset of entries hit by that rule, to be propagated.

RULE

CONDITIONS	ANNOTATIONS
PF00999 + PS01077 + eukaryota OR archaea	[Function] Binds to the 23S rRNA [Similarity] Belongs to the ribosomal protein L37e family [Keywords] Metal-binding, Zinc, Ribonucleoprotein, Ribosomal protein, RNA-binding, rRNA-binding case(archaea) [Name] Ribosomal protein L37e case(eukaryota) [Name] Ribosomal protein L37

Acknowledgements

UniProt is funded by the European Molecular Biology Laboratory, the US National Institutes of Health, the European Union and the Swiss Federal Government.

Email: help@uniprot.org

URL: www.uniprot.org

EMBL- EBI
Wellcome Trust Genome
Campus
Hinxton
Cambridge
CB10 1SD
UK

T +44 (0) 1223 494 444
F +44 (0) 1223 494 468
http://www.ebi.ac.uk