Rule-based annotation in UniProtKB

Kiemens Pichler1, Diego Poggioli1, Claire O’Donovan1
and the UniProt consortium1,2,3

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

THE CHALLENGE

- increasing number of new sequences added at ever increasing speed
- no experimental data for a large proportion of new sequences
- 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
3. CREATE NEW RULE
   - add initial set of conditions and annotations
4. RUN STATISTICS
   - check confidence
5. REFINES RULE
   - add / modify conditions and/or annotations
6. CHECK BEHAVIOUR OF APPLIED RULE
7. APPLY 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.

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