UniProt-GOA: A central resource for data integration and GO annotation.

Mélanie Courtot, Aleksandra Shypitsyna, Elena Speretta, Alexander Holmes, Tony Sawford, Tony Wardell, Maria J. Martin, and Claire O'Donovan

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD
United Kingdom
mcourtot@ebi.ac.uk

Abstract. The Gene Ontology (GO) is a well-established, structured vocabulary used in the functional annotation of gene products. GO terms are used to replace the multiple nomenclatures used by scientific databases that can hamper data integration. Currently, GO consists of more than 41,000 terms describing the molecular function, biological process and subcellular location of a gene product in a generic cell.

The UniProt-Gene Ontology Annotation (UniProt-GOA) project [1] provides high-quality manual and electronic GO annotations, historically to proteins within the UniProt Knowledgebase. Recently, support for annotation of RNAs via RNAcentral IDs and to macromolecular complexes, identified by IntAct Complex Portal IDs, was added. For many species, no experimental data is available: electronic annotations are the only source of information for biological investigation, and it is therefore critical that solid pipelines for data integration across multiple resources be implemented. For example, we rely on Ensembl [2] to project GO annotations automatically according to orthology between species, or InterPro [3] to identify proteins with similar signatures to which GO terms describing the conserved function or location can be associated. In September 2015, an additional 1.5 million annotations from the UniProt Unified Rule (UniRule [4]) system were added electronically.

In addition to increasing the number of annotations available, UniProt-GOA also supports, as part of manual curation, the addition of information about the context of a GO term, such as the target gene or the location of a molecular function, via annotations extensions. For example, we can now describe that a gene product is located in a specific compartment of a specific cell type (e.g., gene product that localizes to the nucleus of a keratinocyte [5]). Annotation extensions are amenable to sophisticated queries and reasoning. A typical use case is for researchers studying a protein that is causative of a specific rare cardiac phenotype: they will be more interested in specific cardiomyocytes cell differentiation proteins than all proteins involved in cell differentiation.

Annotation files for various reference proteomes are released monthly, including human, mouse, rat, zebrafish, cow, chicken, dog, pig, Arabidopsis and Dictyostelium, as well as a file for the multiple species within UniProtKB. The UniProt-GOA dataset can be queried through our user-friendly

QuickGO browser6 or downloaded in a parsable format via the EMBL-EBI [7] and GO Consortium FTP [8] sites.
UniProt-GOA is the largest and most comprehensive open-source contributor of annotations to the GO Consortium annotation effort. The UniProt-GOA dataset has increasingly been integrated into tools that aid in the analysis of large datasets resulting from high-throughput experiments thus assisting researchers in biological interpretation of their results.

Keywords: gene ontology, curation, data integration, uniprot

References

1. UniProt-GOA website, http://www.ebi.ac.uk/GOA
2. The Ensembl project, http://www.ensembl.org/index.html
3. InterPro: protein sequence analysis & classification, http://www.ebi.ac.uk/interpro/
4. UniRule, Rule-based evidence for UniProtKB annotation, http://www.uniprot.org/help/unirule
5. Huntley, Rachael P., et al. A method for increasing expressivity of Gene Ontology annotations using a compositional approach. BMC bioinformatics 15.1 (2014): 155.
6. QuckGO, A fast browser for Gene Ontology terms and annotations, http://www.ebi.ac.uk/QuickGO
7. EBI FTP server, ftp://ftp.ebi.ac.uk/pub/databases/GO/goa
8. Gene Ontology FTP, ftp://ftp.geneontology.org/pub/go/gene-associations