Manual and Automatic Annotation of Plants and Animals in the UniProt Knowledgebase and the Gene Ontology

Claire O’Donovan
UniProt
EMBL-EBI

Introduction to UniProt

Background of UniProt

Since 2002 a merger and collaboration of three databases:

- Swiss-Prot & TrEMBL
- PIR-PSD

Funded mainly by NIH (US) to be the highest quality, most thoroughly annotated protein sequence database

Integrating sequence data

Submissions

Literature scanning

UniProtKB - Protein KnowledgeBase

Expert manual annotation

Reviewed sequences

Unreviewed sequences

Submitted and automatic annotation

Cross-linked to ~150 other databases

Expertly curated functional annotation
Curation of a UniProt/SwissProt entry

Annotation comments

- FUNCTION
- SUBCELLULAR LOCATION
- ALTERNATIVE PRODUCTS
- TISSUE SPECIFICITY
- DEVELOPMENTAL STAGE
- INDUCTION
- SIMILARITY
- CATALYTIC ACTIVITY
- COFACTOR
- ENZYME REGULATION
- BIOPHYSICOCHEMICAL PROPERTIES
- PATHWAY
- SUBUNIT
- INTERACTION
- PTM
- RNA EDITING
- MASS SPECTROMETRY
- DOMAIN
- POLYMORPHISM
- DISRUPTION PHENOTYPE
- ALLERGEN
- DISEASE
- TOXIC DOSE
- BIOTECHNOLOGY
- PHARMACEUTICAL
- MISCELLANEOUS
- CAUTION
- SEQUENCE CAUTION
- WEB RESOURCE

Data import

- Gene names
- Additional bibliography
- GO terms
- Interactions
- Data from structures

Bioinformatics Portal

- www.uniprot.org/help/biocuration
- Provides details about our annotation priorities
- Chordata protein annotation program
- Prokaryotic protein annotation program
- Fungal protein annotation program
- Plant protein annotation program
- Drosophila protein annotation program
- Caenorhabditis protein annotation program
- Viral protein annotation program
- Animal toxin protein annotation program
- SOPs for our manual and automatic annotation processes

Evidence attribution

- System which allows linking of all information in an entry to its original source
- Went public in the flatfile in September 2014
- Allows users:
  - to trace origin of all data
  - to differentiate easily between experimental and computational data
  - to assess data reliability
UniProtKB data growth

UniProtKB/Swiss-Prot contains 547,357 manually curated entries
UniProtKB/TrEMBL contains 89,451,166 unreviewed entries

UniProtKB, reference proteome sequence sets

A reference proteome is the proteome of a representative, well-studied model organism or an organism of interest for biomedical research.

2,351 reference proteomes selected to cover major taxonomic nodes

Two complementary approaches

UniRule
Manually curated rules created by curation team

SAAS
Automatic decision tree-based rule-generating system

Rules are reapplied to UniProtKB/TrEMBL as part of each four-weekly release with both automatic and manual QA procedures ensuring they are still valid

Automatic annotation

- Allows annotation of UniProtKB/TrEMBL in an efficient and scalable manner with a high degree of accuracy
- Based on annotation rules which are created, tested and validated against published experimental data in UniProtKB/Swiss-Prot
- Rules are linked to InterPro member database signatures and define annotations to be added and conditions which must be fulfilled
- Common annotation in Swiss-Prot is transferred to related family members in TrEMBL if they fulfill rule conditions

Introduction to GO
The Gene Ontology

- A way to capture biological knowledge for individual gene products in a written and computable form
- A set of concepts and their relationships to each other arranged as a hierarchy

The Concepts in GO

1. Molecular Function
   - An elemental activity or task or job
   - Protein kinase activity
   - Insulin receptor activity

2. Biological Process
   - A commonly recognised series of events
   - Cell division

3. Cellular Component
   - Where a gene product is located
   - Mitochondrion
   - Mitochondrial matrix
   - Mitochondrial inner membrane

Ontology structure

- Hierarchical
  - Terms can have more than one parent
  - Terms can have more than one child
- Terms are linked by relationships:
  - is_a
  - part_of
  - regulates (and +/- regulates)
  - has_part
  - occurs_in

These relationships allow for complex analysis of large datasets

Inconsistency in English language

- Same name for different concepts
  - Cell
  - or
  - ??

- Different names for the same concept
  - Eggplant
  - Brinjal
  - Aubergine
  - Melongene

Same for biological concepts

→ Comparison is difficult – in particular across species or across databases

Just one reason why the Gene Ontology (GO) is needed...
Aims of the GO project

• Compile the ontologies
  - currently over 38,000 terms
  - constantly increasing and improving

• Annotate gene products using ontology terms
  - around 40 groups provide annotations

• Provide a public resource of data and tools
  - regular releases of annotations
  - tools for browsing/querying annotations and editing the ontology

UniProt-Gene Ontology Annotation (UniProt-GOA) project at the EBI

• Largest open-source contributor of annotations to GO
• Provide annotation for more than 441,000 species

A GO annotation is …

…a statement that a gene product;

1. has a particular molecular function
   or is involved in a particular biological process
   or is located within a certain cellular component

2. as determined by a particular method

3. as described in a particular reference

Electronic Annotation

• Quick way of producing large numbers of annotations
• Annotations use less-specific GO terms
• Only source of annotation for ~438,000 non-model organism species

Manual Annotation

• Time-consuming process producing lower numbers of annotations (~2,800 taxa covered)
• Annotations tend to use more specific GO terms

UniProt-GOA incorporates annotations made using two methods

Broad taxonomic coverage

We have annotations for species that may not have a dedicated curation effort;

e.g. for 1,400 Solanaceae species' we have ~360,000 annotations for ~64,000 proteins

We provide annotation files for more common species…

…as well as less common species that have;

• Complete proteome
• >25% GO annotation coverage

Electronic annotation methods

1. Mapping of external concepts to GO terms
  e.g. InterPro2GO, UniProt Keyword2GO, Enzyme Commission2GO

Electronic Annotation

• non membrane spanning protein tyrosine kinase activity
Annotations are high-quality and have an explanation of the method (GO_REF).

**Electronic annotation methods**

2. Automatic transfer of manual annotations to orthologs

- e.g. Human
- Macaque
- Chimpanzee
- Quokka Pig
- Rat
- and more

- Arabidopsis
- Rice
- Ensembl compara
- Poplar
- Brachypodium
- Maize
- Poplar
- Grape
- and more

Annotations are high-quality and have an explanation of the method (GO_REF).

**Manual annotation by UniProt**

High-quality, specific annotations made using:

- Full text peer-reviewed papers
- A range of evidence codes to categorise the types of evidence found in a paper, e.g. Inferred by Direct Assay (IDA), Inferred by Mutant Phenotype (IMP), Inferred from Physical Interaction (IPI)

**Where can you find annotations?**

- UniProtKB
- Ensembl
- Entrez gene

**Annotation Files**

Contain all information for each annotation

**GO Consortium website**

- http://www.geneontology.org/cgi-bin/references.cgi

**UniProt-GOA website**

- http://www.ebi.ac.uk/GOA/downloads.html

**AmiGO**

GO browsers

**Help / Feedback**

Stuck? Just ask – active help and support team

- http://www.ebi.ac.uk/training/online/course/uniprot-exploring-protein-sequence-and-functional

- Feedback – if you find something wrong, outdated, missing etc please tell us.

  help@uniprot.org/goa@ebi.ac.uk
Acknowledgements

• All UniProt Consortium members
• All GO Consortium members
• Database providers, submitters and researchers with whom we interact
• Funders