Introduction to the Gene Ontology and GO annotation resources

Rachael Huntley
UniProt
EBI
PAGXXII
January 2014

What is 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

Anatomy of a GO term

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
Why do we need GO?

Inconsistency in English language
• *Same* name for *different* concepts

Eggplant
Aubergine
Brinjal
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...

Reasons for the Gene Ontology
• Inconsistency in English language
• Increasing amounts of biological data available
• Increasing amounts of biological data to come

Increasing amounts of biological data available
Search on ‘DNA repair’...get almost 65,000 results

Expansion of sequence information
Reasons for the Gene Ontology

- Inconsistency in English language
- Increasing amounts of biological data available
- Increasing amounts of biological data to come
- Large datasets need to be interpreted quickly

GO provides a way of consolidating all the sequence and experimental information into one easily accessible resource

Who is GO?

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

Aims of the GO project

GO Annotation
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

<table>
<thead>
<tr>
<th>Accession</th>
<th>Name</th>
<th>GO ID</th>
<th>GO term name</th>
<th>Reference</th>
<th>Evidence code</th>
</tr>
</thead>
<tbody>
<tr>
<td>P00505</td>
<td>GOT2</td>
<td>GO:0004069</td>
<td>aspartate transaminase activity</td>
<td>PMID:2731362</td>
<td>IDA</td>
</tr>
</tbody>
</table>

Annotation Extensions

Allows curators to combine GO terms with:

• other GO terms
• terms from external ontologies
• database identifiers representing genes or their products

“Pap1 is a transcription factor that regulates transcription of several oxidative stress response genes, including trf, pmf1, obr1, hgl1 and ctt1”

“TMEM115 is located in the nucleus that is part of an epithelial cell that is part of the cervix epithelium”

UniProt-GOA incorporates annotations made using two methods

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 taxons covered)
• Annotations tend to use more specific GO terms
• Manual annotation is essential for creating predictions

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 well-studied species…

… as well as less well-studied 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

GO:0004715; non-membrane spanning protein tyrosine kinase activity
Annotations are high-quality and have an explanation of the method (GO_REF).

2. Automatic transfer of manual annotations to orthologs

- Ensembl
- Arabidopsis
- Rice
- Macaque
- Chimpanzee
- Guinea Pig
- Rat
- Mouse
- Cow
- Dog
- Chicken

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

Electronic annotation methods

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)

Electronic annotations 213,544,213
Manual annotations* 1,448,170

Dec 2013 Statistics

* Includes manual annotations integrated from external model organism and specialist groups

Number of annotations in UniProt-GOA database

Where can you find annotations?

UniProtKB
Ensembl
Entrez gene

Annotation Files
Contain all information for each annotation

GO Consortium website

UniProt-GOA website
Numerous species-specific files
Uses of GO annotation data

• Access gene product functional information
• Analyse high-throughput genomic or proteomic datasets
• Validation of experimental techniques
• Get a broad overview of the functions of a proteome
• Obtain functional information for novel gene products

How scientists use the GO

Some examples...

Term enrichment

• Most popular type of GO analysis
• Determines which GO terms are more often associated with a specified list of genes/proteins compared with a control list or rest of genome
• Many tools available to do this analysis
• User must decide which is best for their analysis

Numerous Third Party Tools

Annotating novel sequences

- BLAST can be used to find similar sequences with GO annotation which give an indication of the functional roles of the new sequence
- Two tools currently available:
  - AmiGO BLAST – searches the GO Consortium database
  - BLAST2GO – searches the NCBI database

* Note BLAST alone is not considered reliable enough to make automated functional predictions

Using the GO to provide a functional overview for a large dataset

- Many GO analysis tools use GO slims to give a broad overview of the dataset
- GO slims are cut-down versions of the GO and contain a subset of the terms in the whole GO
- GO slims usually contain less-specialised GO terms
Slimming the GO using the ‘true path rule’…however annotations can be mapped up to a smaller set of parent GO terms:

GO slims

Custom slims are available for download:

or you can make your own using:
• QuickGO
  http://www.ebi.ac.uk/QuickGO
• AmIGO’s GO slimmer
  http://amigo.geneontology.org/cgi-bin/amigo/slimmer

The UniProt QuickGO browser

Search GO terms or proteins
Find sets of GO annotations
Map-up annotations with GO slims
Questions on how to use QuickGO?
Contact goa@ebi.ac.uk

www.ebi.ac.uk/QuickGO

UniProt-GOA

Curators:
  Rachael Huntley
  Prudence Mutwoto
  Aleksandra Shypitsyna

Software developers:
  Tony Sawford
  Carlos Bonilla

Team leaders:
  Claire O’Donovan
  Maria Martin

Email: goa@ebi.ac.uk

http://www.ebi.ac.uk/GOA

UniProt

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.

Acknowledgements

Members of:
  InterPro
  IntAct
  HAMAP
  Ensembl
  Ensembl Genomes

GO Consortium

Funding

Additional support for the EBI’s involvement in UniProt comes from the NIH grant (HG02277), British Heart Foundation grant FS/13/180/32330 and SNB. SwissProt activities at the EBI are supported by the Swiss Federal Government through the The State Secretariat for Education, Research and Innovation (SERI) and the European Commission (contracts EBM2007-002101 and no.GEO/GENO for the UniProtKB project, and contracts BIOTEC-021676, BIOTEC-025048, BIOTEC-025500, and BIOTEC-025500 for the SwissProt project) and the National Science Foundation (NSF) grant DBI-1052153.