Introduction to the Gene Ontology and GO annotation resources

Rachael Huntley

UniProt-GOA

EBI

PAGXX

January 2012
Why do we need GO?
• Inconsistency in English language
Inconsistency in English language

- **Same** name for **different** concepts

Cell

or

??

[Image of a prison cell on the left and a cell diagram on the right]
• Different names for the same concept

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
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

<table>
<thead>
<tr>
<th>ID</th>
<th>GO:0005634</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>nucleus</td>
</tr>
<tr>
<td>Definition</td>
<td>A membrane-bounded organelle of eukaryotic cells in which chromosomes are housed and replicated. In most cells, the nucleus contains all of the cell’s chromosomes except the organellar chromosomes, and is the site of RNA synthesis and processing. In some species, or in specialized cell types, RNA metabolism or DNA replication may be absent.</td>
</tr>
<tr>
<td>Comment</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type</th>
<th>Synonym</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>exact cell nucleus</td>
</tr>
</tbody>
</table>

**Cross-references associated with this term:**

<table>
<thead>
<tr>
<th>Database</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERPRO</td>
<td>IPR000003</td>
</tr>
<tr>
<td>INTERPRO</td>
<td>IPR000116</td>
</tr>
<tr>
<td>INTERPRO</td>
<td>IPR000135</td>
</tr>
</tbody>
</table>
Ontology structure

• Directed acyclic graph
  Terms can have more than one parent

• 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
Aims of the GO project

• Compile the ontologies
  - currently over 35,000 terms
  - constantly increasing and improving

• Annotate gene products using ontology terms
  - around 30 groups provide annotations

• Provide a public resource of data and tools
  - regular releases of annotations
  - tools for browsing/querying annotations and editing the ontology
GO Annotation
UniProt-Gene Ontology Annotation (UniProt-GOA) database at the EBI

- Largest open-source contributor of annotations to GO
- Provides annotation for more than 397,000 species
- Our priority is to annotate the human proteome
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>
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 many non-model organism species

**Manual Annotation**
- Time-consuming process producing lower numbers of annotations
- Annotations tend to use very specific GO terms
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
2. Automatic transfer of manual annotations to orthologs

Annotations are high-quality and have an explanation of the method (GO_REF)
Manual annotation by GOA

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. IDA, IMP, IPI
Number of annotations in UniProt-GOA database

<table>
<thead>
<tr>
<th>Annotation Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electronic annotations</td>
<td>110,247,289</td>
</tr>
<tr>
<td>Manual annotations*</td>
<td>920,557</td>
</tr>
</tbody>
</table>

Dec 2011 Statistics

* Includes manual annotations integrated from external model organism and specialist groups
How to access and use GO annotation data
Where can you find annotations?

UniProtKB

Ensembl

Entrez gene
Gene Association Files

17 column files containing all information for each annotation

UniProt-GOA website
Numerous species-specific files

http://www.ebi.ac.uk/GOA/downloads.html
**ABL1**

Tyrosine-protein kinase ABL1
protein from *Homo sapiens* (human)

**Term Associations**

<table>
<thead>
<tr>
<th>Accession, Term</th>
<th>Ontology</th>
<th>Qualifier</th>
<th>Evidence</th>
<th>Reference</th>
<th>Assigned by</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:00300036 : actin cytoskeleton organization</td>
<td>biological</td>
<td>process</td>
<td>ISS</td>
<td>GO:0060024</td>
<td>UniProtKB</td>
</tr>
<tr>
<td>GO:0007155 : cell adhesion</td>
<td>biological</td>
<td>process</td>
<td>IFA</td>
<td>GO:0008004</td>
<td>UniProtKB</td>
</tr>
<tr>
<td>GO:00068630 : DNA damage response, signal transduction resulting in induction of apoptosis</td>
<td>biological process</td>
<td></td>
<td>TAP</td>
<td>Probeset Inc.</td>
<td></td>
</tr>
<tr>
<td>GO:0006298 : mismatch repair</td>
<td>biological</td>
<td>process</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The EBI's QuickGO browser

QuickGO is a fast web-based browser for Gene Ontology terms and annotations, which is provided by the UniProt:GOA group at the EBI.

Search GO terms or proteins

Find sets of GO annotations

QuickGO Tips

- QuickGO can help make customized GO slims, which can be used to 'map up' your gene/protein identifiers to broadly categorise their functions or subcellular locations. Click here for more information.
- QuickGO's search facility can be used to find information on GO terms as well as information on annotation to distinct proteins. Just type a partial GO term name/synonym or a gene symbol, protein name or sequence identifier into the Search box.

Example queries

QuickGO can be queried for both GO terms and proteins:
- Search for terms by keyword or ID: *apopha*, GO:0006615
- Search for proteins by name or accession: *bopomys*, P08727
- List all terms in an ontology: Biological Process, Molecular Function, Cellular Component
How scientists use the GO

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

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

Tools for Searching and Browsing GO

The following tools make use of the GO ontologies or the gene associations provided by Consortium members. Being listed on this page does not represent an endorsement by the GO Consortium, nor has the Consortium tested the tool or found that it uses the Consortium information accurately. This page is provided to promote an exchange of information between users and software developers.

Tools for Analysis of Data Sets, e.g. gene expression / microarray data

The following tools make use of the GO ontologies or the gene associations provided by Consortium members. Being listed on this page does not represent an endorsement by the GO Consortium, nor has the Consortium tested the tool or found that it uses the Consortium information accurately. This page is provided to promote an exchange of information between users and software developers.

AmiGO

AmiGO provides an interface to search and browse GO and view the terms with which they are associated, as well as gene product annotations. AmiGO also provides a facility to download GO and gene product annotations.

Avadis

Avadis is a data analysis and visualization tool for gene expression data. Avadis has a built-in Gene Ontology browser to view ontology hierarchies. There are common ontology paths for multiple genes. Genes can be clustered based on ontology terms to identify functional signatures in gene expression clusters.
Analysis of high-throughput genomic datasets

Bregje Wertheim at the Centre for Evolutionary Genomics, Department of Biology, UCL and Eugene Schuster Group, EBI.
Validation of experimental techniques

Rat liver plasma membrane isolation

(Cao et al., Journal of Proteome Research 2006)
Annotating novel sequences

• Can use BLAST queries to find similar sequences with GO annotation which can be transferred to the new sequence

• Two tools currently available;

  **AmiGO BLAST** – searches the GO Consortium database

  **BLAST2GO** – searches the NCBI database
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’

Many gene products are associated with a large number of descriptive, leaf GO nodes:
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-GOA group

Curators: Emily Dimmer
           Rachael Huntley
           Yasmin Alam-Faruque
           Prudence Mutowu

Software developer: Tony Sawford

Team leaders: Rolf Apweiler
              Claire O’ Donovan

Email: goa@ebi.ac.uk

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

Members of:

UniProtKB
InterPro
IntAct
HAMAP
Ensembl
Ensembl Genomes

GO Consortium

Funding

National Human Genome Research Institute (NHGRI)
Kidney Research UK
British Heart Foundation
EMBL