Introduction to Gene Ontology

Training: February 2011
THE ONTOLOGIES... the scope of GO

1. **Biological Process**
   - Divided into 3 main levels:
     - cellular processes (E.g. cell cycle)
     - multicellular organismal processes (E.g. blood circulation)
     - multi-organism processes (E.g. mating, pollination)

2. **Molecular Function** (E.g. protein kinase activity)

3. **Cellular Component**
   - Divided into:
     - component (E.g. cell, mitochondrion, plastid)
     - component part (E.g. cell part, mitochondrial part, plastid part)
Process-Function Links in GO

- GO was originally three completely independent hierarchies, with no relationships between them.

- Biological processes are ordered assemblies of molecular functions.

- As of 2009 we have started making relationships between biological process and molecular function in the live ontology:
  - Functions that regulate processes e.g. transcription regulator regulates transcription.
  - Functions that are part_of processes e.g. transporter part_of transport.
Process-Function Links in GO

Process: protein transport
Function: substrate-specific transporter activity
protein transporter activity

Function: molecular_function
Process: regulation of transcription
transcription regulator activity
Naming GO Terms

All GO term names:

- are lower case
- are singular
  - mitochondrion
- are expanded, not abbreviated (where sensible)
  - epidermal growth factor binding
- are as precise as possible
  - syncytium formation
  - cell growth
Defining GO terms

- All terms have a textual definition, ideally of the form:
  - “An A that B” where A is the is_a parent (genus) and B is what differentiates it from its parent (differentia)
    - phosphorylation: The process of introducing a phosphate group into a molecule, usually with the formation of a phosphoric ester, a phosphoric anhydride or a phosphoric amide.
    - protein phosphorylation: The process of introducing a phosphate group on to a protein.

- have a beginning and end (for processes)

- Many (most!) definitions don’t conform to this yet, however, and it is an ongoing process to improve them
Relations Between GO Terms

- is_a
- part_of
- regulates
- has_part (full GO only)

http://www.geneontology.org/GO.ontology-ext.relations.shtml
is_a

- If A is a B, then A is a subtype of B
  - mitotic cell cycle is a cell cycle
  - lyase activity is a catalytic activity.

- Transitive relationship: can infer up the graph
part_of

- Necessarily part of
- Wherever B exists, it is as part of A. But not all B is part of A.

E.g. All replication forks are part of a chromosome. Not all chromosomes have replication forks.

- Transitive (can infer up the graph)
regulates

- One process directly affects another process or quality
- Necessarily regulates: if both A and B are present, B always regulates A, but A may not always be regulated by B

All cell cycle checkpoints regulate the cell cycle.
The cell cycle is not solely regulated by cell cycle checkpoints
regulates

- regulates
  - positively regulates
  - negatively regulates

- Unlike is_a and part_of, regulates does not indicate granularity

- Not transitive through regulates relationships:

  ![Diagram showing relationships]

  cannot draw any conclusions from these statements about the relationship between regulation of anti-apoptosis and cell death.
regulates

- Can infer up through the is_a and part_of relationships

http://www.geneontology.org/GO.ontology.relations.shtml#regulates
- Only used in the full GO file, at the moment

- Relationships are upside down compared to is_a and part_of (go from parent to child)

- Necessarily has part

All nuclei have chromosomes. Not all chromosomes are part of nuclei.
GO Relations

- For all terms in the ontology, you have to be able to reach the root through a complete path of is_a relationships:
  - we call this being is_a complete
  - important for reasoning over the ontology, and for ontology development (everything is_a something)

Nephron development is part_of kidney development

Nephron development is_a anatomical structure development, which is_a developmental process, which is_a biological process
GO Relations

- is_a and part_of relationships operate between terms such that
  - ‘parent’ term = broader (closer to root)
  - ‘child’ term = more specific (closer to leaf node)

- Child terms inherit the meaning of all their parent terms. Otherwise you get TPVs (True Path Violations)
Species Specificity

- We used to have sensu GO terms: eye development (sensu Mammalia), eye development (sensu Insecta) etc

- Sensu terms were replaced with GO terms that describe the differences between organisms. If a word or phrase refers to different entities or processes depending upon the organism, subclasses are created based on differentiating characteristics.

  - eye development
    
    *The process whose specific outcome is the progression of the eye over time, from its formation to the mature structure. The eye is the organ of sight.*

  - camera-type eye development
    
    *The process whose specific outcome is the progression of the camera-type eye over time, from its formation to the mature structure. The camera-type eye is an organ of sight that receives light through an aperture and focuses it through a lens, projecting it on a photoreceptor field.*

  - compound eye development
    
    *The process whose specific outcome is the progression of the compound eye over time, from its formation to the mature structure.*

- GO provides a file containing species-specific terms and the taxa the terms are/are not appropriate for.

  - id: GO:0009526
  - name: plastid envelope
  - relationship: only_in_taxon NCBI_taxonomy:33090 ! Viridiplantae
Ontology Quality Checks

- Editors run checks within the editor (OBO-Edit) for:
  - spelling, extra spaces, full stops, missing references, name redundancy, is_a completeness, cycles....

- You can run a separate reasoner over the ontology to check for redundant relationships. (Tanya and David also do this periodically).

- Chris Mungall (Berkeley) runs a nightly script on the ontology to check for errors
GO Slims

- GO slims are cut-down versions of the GO ontologies containing a subset of the terms, without any of the fine, granular terms.

- Useful for giving a summary of the results of GO annotation of large-scale analyses, such as microarray, expression experiments etc.

- GO provides a generic (non species-specific) GO slim.
- Various databases/groups have submitted GO slims.
- Users can create GO slims to meet their individual needs.

## GO Slims

### Maintained GO slims for download

<table>
<thead>
<tr>
<th>Organism or Usage</th>
<th>Download</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generic GO slim</td>
<td>OBO format</td>
</tr>
<tr>
<td>Developed by GO Consortium</td>
<td></td>
</tr>
<tr>
<td>UniProtKB-GOA and whole proteome analysis</td>
<td>OBO format</td>
</tr>
<tr>
<td>Developed by N. Mulder, M. Pruess</td>
<td></td>
</tr>
<tr>
<td>Plant slim</td>
<td>OBO format</td>
</tr>
<tr>
<td>Developed by <a href="#">The Arabidopsis Information Resource</a></td>
<td></td>
</tr>
<tr>
<td><em>Candida albicans</em></td>
<td>OBO format</td>
</tr>
<tr>
<td>Developed by <a href="#">Candida Genome Database</a></td>
<td></td>
</tr>
<tr>
<td>Protein Information Resource <em>.slim</em></td>
<td>OBO format</td>
</tr>
<tr>
<td>Developed by Darren Natale of PIR</td>
<td></td>
</tr>
<tr>
<td><em>Schizosaccharomyces pombe</em> slim</td>
<td>OBO format</td>
</tr>
<tr>
<td>Developed by Val Wood</td>
<td></td>
</tr>
<tr>
<td>Yeast slim</td>
<td>OBO format</td>
</tr>
<tr>
<td>Developed by <a href="#">Saccharomyces Genome Database</a></td>
<td></td>
</tr>
</tbody>
</table>
As well as the regular textual definitions, we also provide formal definitions - known as cross-products - for some terms.

Cross-products are of the form [genus] relationship [differentia]

- E.g. the cross-product for nuclear chromosome is:
  chromosome part_of nucleus

Cross-products allow automatic reasoning over ontology
- detects inconsistencies, missing relationships etc
- allow better integration with other ontologies
- more systematic ontology development (automatic term generation, textual definition generation)

In live GO, we currently have cross-products for
- most regulation terms
- cellular component part_of
- .... More will be added in batches
Exercise: Term Definitions

- Write GO definitions for the following terms:
Exercise: Relationships

- Create relationships between the following terms: