Gene Ontology (GO)

ISLS - Lecture of 14 October 2004

Yun Bei

Ontology ?

- In computer science, an **ontology** is the attempt to formulate an exhaustive and rigorous conceptual schema within a given domain, a typically hierarchical data structure containing all the **relevant entities** and their **relationships** and **rules** (theorems, regulations) within that domain. – **Wikipedia**
- Ontologies are '**specifications of a relational vocabulary'**. They are sets of defined terms like the sort that you would find in a dictionary, but the terms are **networked**. The terms in a given vocabulary are likely to be restricted to those used in a particular field, and in the case of GO, the terms are all **biological. - GO**

Why GO?

- Wide variations in terminology
- Collaborate consistent descriptions of gene products[1] in different databases
- Vocabulary and relationships, standard annotations, uniform queries
- Extract biological insight from enormous sets of data

[1] GO uses 'gene products' to refer to any protein or RNA encoded by a gene

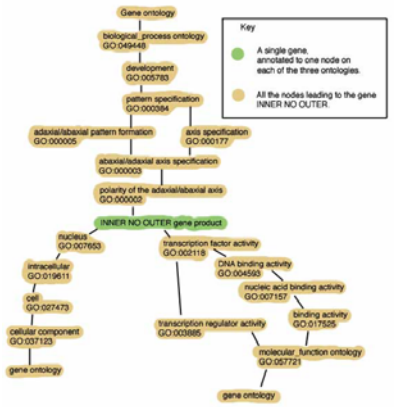
GO - Aim

- Develop and maintain ontologies(vocabularies)
  - in three non-overlapping domains
- Cross-links between ontologies and gene products in biological databases
  - annotations
- Develop software tools for use with GO data
  - Visualization and query
  - i.e. DAG-Edit, AmiGO browser

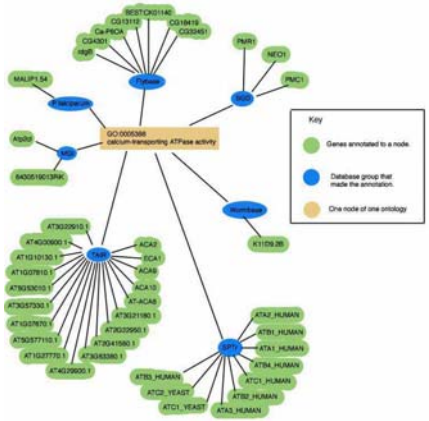
GO - collaborative databases

- In 1998
  - FlyBase (Drosophila)
  - Saccharomyces Genome Database (SGD)
  - Mouse Genome Database (MGD)

A network of defined biological terms



Annotation of genes to the nodes on the ontologies



GO - Ontologies

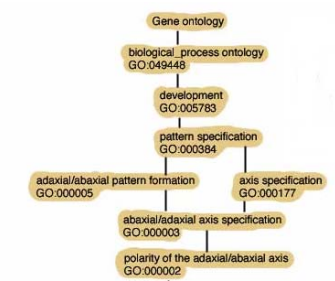
- Molecular Function (MF)
  - Activities
  - i.e. ‘kinase activity’
- Biological Process (BP)
  - Biological goals, one or more ordered assemblies of MFs
  - i.e. ‘cell death’
- Cellular Component (CC)
  - Locations
  - i.e. ‘nuclear inner membrane’
- Sequence Ontology (SO)
  - Sequence features
  - i.e. ‘exon’

GO structure

- GO terms are organized in structures called directed acyclic graphs (DAGs)
  - 'child' can have one or many 'parents'
  - relationship: 'is-a' and 'part-of'
- Flat file format (.obo)

```
[Term]
id: GO:000384
name: pattern specification
namespace: process
def: "The processes that result in the patterns of cell differentiation." [PMID:11566870]
subset: goslim_generic
subset: goslim_goa
subset: goslim_plant
subset: goslim_yeast
synonym: "pattern formation" []
is_a: GO:005783
```

GO - DAG



Annotations

- Links between gene products and GO terms
- Two principles
  - Be attributed to a source: literature source, another database or computational analysis
  - Indicate the evidence.
    - IMP inferred from mutant phenotype
    - IGI inferred from genetic interaction [with <database:gene_symbol[allele_symbol]>]
    - IPI inferred from physical interaction [with <database:protein_name>]
    - ISS inferred from sequence similarity [with <database:sequence_id>]
    - IDA inferred from direct assay
    - IEP inferred from expression pattern
    - IEA inferred from electronic annotation [to <database:id>]
    - TAS traceable author statement
    - NAS non-traceable author statement
    - ND no biological data available
    - IC inferred by curator

Annotation file format

Column	Content	Example
1.	DB	SGD
2.	DB_Object_ID	S0000296
3.	DB_Object_Symbol	PHO3
4.	Qualifier	
5.	GO ID	GO:0015888
6.	DB:Reference(|DB:Reference)	SGD:8788|PMID:2676709
7.	Evidence	IMP
8.	With (or) From	
9.	Aspect	P
10.	DB_Object_Name	acid phosphatase
11.	DB_Object_Synonym(|Synonym)	YBR092C
12.	DB_Object_Type	gene
13.	taxon(|taxon)	taxon:4932
14.	Date	20010118
15.	Assigned_by	SGD

GO slims

- Annotate sets of gene products, gain a high-level view of gene functions
- Subsets of GO
- Customized for specific analysis needs
  - goslim_generic
  - goslim_goa
  - goslim_plant
  - goslim_yeast
- Perl script to generate associations mapped to the slim GO terms
  - map2slim.pl -d go -o OUTPUTFILE SLIMFILE GENE-ASSOC-FILE

GO slims

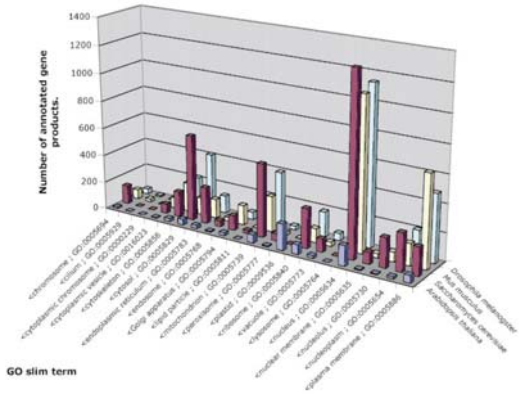


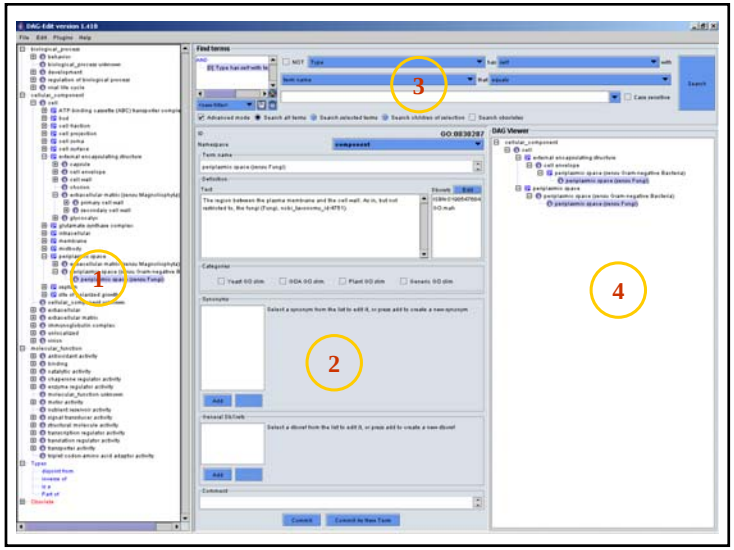
Figure 1. Application of a GO slim set in genome annotation. The number of gene products annotated to each term in each of four model organism genomes is shown for a GO slim set taken from the cellular component ontology (data as of August 1, 2003).

GO Database

- MySQL database
- Help programmers to write tools that use GO data.
- Monthly released database
  - termdb, assocdb, seqdb, etc

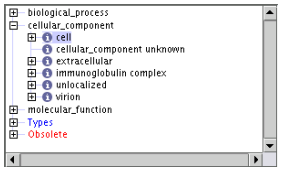
Software/tools

- GO visualization and query tool
  - standalone GO editors DAG-Edit
  - web-based GO browsers
- DAG-Edit
- AmiGO browser
  - web-interface searching and displaying
  - tree-like view of GO structure
  - summary graphical view



DAG-Edit

1. Term editor panel



2. Text editor panel

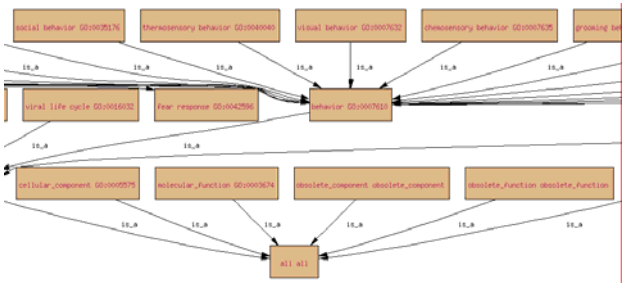
DAG-Edit

3. Search/filter panel

4. DAG viewer

AmiGo

AmiGo



GO Resources – www.geneontology.org

File	Description	Updated
The gene ontologies in OBO flat file format	gene_ontology.obo text file (17886 terms as of October 4, 2004)	Every 30 mins.
The gene ontologies in GO flat file format	**Molecular Function** text file (7422 terms as of October 4, 2004) **Biological Process** text file (8972 terms as of October 4, 2004) **Cellular Component** text file (1472 terms as of October 4, 2004) **Term Definition** text file (Deprecated)	Daily, from the OBO flat file. (Time may vary.)
XML format	Two XML Format Files files are available, one with gene associations and one without. Specific information on the file contents are available from the download page. Older versions are in the archive.	Monthly from the GO flat file.
GO Database/ MySQL format	API documentation, schema diagrams and full descriptions of all tables for the mySQL database developed and maintained by BDGP. The latest MySQL files are available, as are archived versions.	MySQL files updated monthly from the GO flat file.
GO Slims	GO Slims are 'slimmed down' versions of the ontologies that allow you to annotate genomes or sets of gene products to gain a high-level view of gene functions. The currently maintained GO slims are: goslim_generic OBO format old GO format goslim_goa OBO format old GO format goslim_plant OBO format old GO format goslim_yeast OBO format old GO format An archive of GO slims, used in publications etc., is also available. You can the use the map2slim.pl script (follow the 'scripts' link on the bottom left of this GO Database documentation page) to take a GO slim file and a gene association file, and output the associations mapped to the slim terms. To do this, use the command map2slim.pl -d go -o OUTPUTFILE SLIMFILE GENE-ASSOC-FILE	Updated automatically in line with ontology files.
Current Annotations	Download annotations made by each database group.	Database dependent.
Archives of Monthly Releases	An Archive of GO and OBO flat file format ontology files is kept in the FTP site.	Monthly.
FTP Archive	- The publicly accessible FTP archive of the GO project repository contains a direct copy of all files currently in GO cvs. This is the full set of GO files, including all documentation, e-mail and meeting archives, and ontology and annotation files.	Every 30 minutes, from cvs.

What GO is NOT?

- NOT a database of gene sequences
- NOT a way to unify biological databases
- NOT a dictated standard

Status of the GO vocabularies

Totals	July 1, 2000	July 1, 2003
All valid terms[a]	4493	13412
Terms with definitions	250	11105
Terms with synonyms	301	2813
Terms with db cross-references	1042	12317
Associations[b]	30654	7781954
Gene products	13016	1549236
Sequences	0	21916
Paths[c]	30941	314886

[a]Excludes obsolete terms.
[b]Individual associations between any gene product and any GO term.
[c]Parent–child relationships traced from any GO term to the root (molecular function, biological process or cellular component).