Introduction to Gene Ontology

Presenter: Wayne Xu, Ph.D
Computational Genomics Consultant, Supercomputing Institute

Email: wxu@msi.umn.edu
Phone: (612) 624-1447
Help: help@msi.umn.edu

April 13, 2006
Outline

• Introduction
• Gene Ontology and GO Consortium
• GO data descriptive vocabularies
• GO annotation
• GO Databases
• GO Tools
Introduction
Motivation

• Explosively-increasing amount of sequence data leads the creation of many databases for the data management
  – Domain-specific: PIR, PDB, GenBank, TIGR, UniProt, …
  – Organism-specific: AceDB, FlyBase, SGD, MGI,…

• But limitation in data integration:
  – Can list a gene product P53 in all organisms and what it does in these organisms?
  – Can list all “receptor signaling protein tyrosine kinase activity” proteins in all organisms?
  – Can list all “defense response to pathogenic bacteria” proteins in all organisms?
  – Even within the same organism, how do you classify a group of proteins?
Solutions

• The most fundamental questions for the biologists served by these databases revolve around the genes
  – Describe the genes or gene products
  – Genes have relationships to others
  – Gene product has multiple features

• So, the challenge is to develop one common data description schema for all organisms and all databases

• What is a best way?
  – Description
    • Location, function, process
  – Presentation:
    • List
    • Taxonomy
    • Ontology
List

Protein Function process

- No relationships within the same type of concepts
- Very useful for simplest applications
Taxonomy

Protein Function

- Hierarchical relationship among the same type of concept
- But 1:1 relationship between concepts, not the case in genes
Ontology

• Include much richer and more descriptive relationships between concepts
Gene Ontology and GO Consortium
Gene Ontology

• In July 1998, at the Montreal International Conference on Intelligent Systems for Molecular Biology (ISMB) bio-ontologies Workshop
  • Michael Ashburner presented a simple hierarchical controlled vocabulary as Gene Ontology
  • It was agreed by three model databases: FlyBase (Suzanna E Lewis), SGD (Steve Chervitz), and MGI (Judith Blake)
  • The Gene Ontology Consortium was founded
Ontologies

• Ontology is derived from the Greek meaning “a description of what exists”.
• An ontology is used now a description of the concepts and relationships that exist for a community of agents
• Practically write an ontology as a set of definitions of formal vocabulary
• For the purpose of enabling knowledge sharing and reuse
  – Plant ontology (PO): a controlled vocabulary for plant structure (anatomy) and growth stages
  – Trait ontology (TO): a controlled vocabulary to describe each trait as a distinguishable feature, characteristic, quality or phenotypic feature of a developing or mature individual. Examples are glutinous endosperm, disease resistance, plant height, photosensitivity, male sterility, etc.
  – Mammalian Phenotype Ontology
  – Mouse ontology
  – Cell type ontology
  – Sequence Ontology
  – Gene Ontology
  – ...
GO Consortium

• Three major goals:
  – To develop a set of controlled, structured vocabularies – gene ontology (GO) – to describe key domains of molecular biology, gene
  – To apply GO terms in the annotation of genes in biological databases
  – To provide a centralized public resource allowing universal access to the GO, annotation data sets and software tools developed for use with GO data
GO Data Descriptive Vocabularies
GO Vocabularies (Terms)

• Define all gene products by the three organizing GO principles:
  – molecular function
  – biological process
  – cellular component

• Eukaryotes and virus share a same data description schema (controlled vocabularies)
  – problem?
GO Molecular Function

• Describes activities, such as catalytic or binding activities, at the molecular level

• Examples:
  – Broad molecular function terms:
    • catalytic activity,
    • transporter activity,
    • binding;
  – Narrower molecular function terms
    • Adenylate cyclase activity
    • Toll receptor binding
GO Biological Process

• Series of events accomplished by one or more molecular functions

• Examples:
  – Broad biological process terms
    • cellular physiological process
    • signal transduction,
  – Narrower biological process terms:
    • pyrimidine metabolism
    • alpha-glucoside transport.

• Distinguish between a biological process and a molecular function, but the general rule is that a process must have more than one distinct steps

• A biological process is not equivalent to a pathway.
GO Cellular Component

• A component of a cell such as part of some larger object

• Examples:
  – an anatomical structure (e.g. rough endoplasmic reticulum or nucleus)
  – a gene product group (e.g. ribosome, proteasome or a protein dimer)
GO Vocabularies (Terms)

• A gene product has one or more molecular functions and is used in one or more biological processes; it might be associated with one or more cellular components.

• Example, the gene product cytochrome c can be described by
  – the molecular function term oxidoreductase activity,
  – the biological process terms oxidative phosphorylation and induction of cell death,
  – and the cellular component terms mitochondrial matrix and mitochondrial inner membrane.
Define GO Terms

• Controlled Vocabularies,
• Explore into all the three principles and their hierarchical relationships
• must use our extensive domain knowledge of biology
  – GO Consortium
  – Many Curator interest groups
GO Terms

[Term] id: GO:0000002
name: mitochondrial genome maintenance
namespace: biological_process
def: "The maintenance of the structure and integrity of the mitochondrial genome." [GOC:ai]
is_a: GO:0007005 ! mitochondrion organization and biogenesis

[Term] id: GO:0000003
name: reproduction
namespace: biological_process
Alt_id: GO:0019952
def: "The production by an organism of new individuals that contain some portion of their genetic material inherited from that organism." [GOC:go_curators, ISBN:0198506732]
subset: goslim_generic
subset: goslim_plant
subset: gosubset_prok
is_a: GO:0008150 ! biological_process
GO Annotation
GO Gene Annotation

• All GO collaborating databases annotate their gene products (or genes) with GO terms
  – Source
    • Literature
    • another database
    • computational analysis
  – Evidence codes:
    • IMP
    • IGI
    • IPI
    • ISS
    • IDA
    • IEP
    • IEA
    • TAS
    • NAS
    • ND
    • IC
Annotation File Format

• Gene associate file or Mysql gene associate table
  – Link between term and gene or gene product (transcript or protein)

• 15 columns:

  1. DB
  2. DB_Object_ID
  3. DB_Object_Symbol
  4. NOT
  5. GO ID
  6. DB:Reference
  7. Evidence
  8. With (or) from
  9. Aspect
 10. DB_Object_Name
 11. DB_Object_Synonym
 12. DB_Object_Type
 13. Taxon
 14. Date
 15. Assigned_by
GO Database
Gene Ontology Home

The Gene Ontology project provides a controlled vocabulary to describe gene and gene product attributes in any organism. Read more...

Popular Links

Search the Gene Ontology Database

This search uses the browser AmIGO. Browse the Gene Ontology using AmIGO.

GO website

- GO downloads: including ontology files, annotations and the GO database
- Tools for using GO
- Request new terms or ontology changes via the SourceForge tracker system; help with new term submission is available.
- Documentation on all aspects of the GO project and the FAQ
- Gene Ontology mailing lists and contact details

News
Funding and Acknowledgements
Usage Statistics
Gene Ontology Database Downloads

Downloads

The GO monthly releases are available, either as RDF XML or as a MySQL database dump.

Monthly release contains data as of 2006-02-01: [HTTP or FTP].

Weekly release prepared on 2006-02-26: [HTTP or FTP].

Daily release prepared on 2006-03-02: [HTTP or FTP].

View and download all daily, weekly and monthly releases.

Release Cycle

The GO Database is built from the data publicly available as flatfiles from the main GO website. The database is not used for data management, only for querying, either with ArdGO, the go-db-perl modules or with MySQL.

The GO database follows a monthly cycle. Each monthly release takes a few days to build, and requires manual QC. Although the timing of the release is irregular, it always corresponds to the data in the main GO CVS repository and FTP site as of midnight on the first of every month.

Documentation

General Documentation

Consult the GO Software and Databases webpage for software and API details, or the main GO website.

Schema Documentation

GO Database Schema Description

Main documentation on the GO Database, and the table creation SQL

Example queries

Examples of common queries you might want to ask the GO database

Hyperlinked HTML Tables

A web page showing the tables and columns in the GO database. You can traverse the foreign key relationships as hyperlinks. Note: there are no embedded comments on this autogenerated page, also, the tables are not arranged into their modular structure. For full documentation, please refer to the full documentation

Contributed Diagrams

Diagrams showing the structure of the GO database - these are contributed by third parties, and may be incorrect or out of date with respect to the most
Recursive Querying

• Find all *DNA binding* genes

• *term2term* table to iterate through the graph, but this requires multiple SQL calls

• *precompute* the path from every node to all of its ancestors. This goes in the *graph_path* table, which also holds the distance between terms
Query GO Database

• Direct MySQL queries
  – use the mysql command line interface to issue queries

• Query via the perl API
  – need `go-db-perl` for this

• Local copy of AmiGO
  – install AmiGO as a local CGI script, and issue web queries

• Query via your own code
  – write your own code to query the db, using a database driver such as DBI or JDBC

• Query via DBStag
  – use the `stag` module for issuing queries to the GO db and getting back XML. query with arbitrary SQL, or use the `stag templates` provided (see `README`).
## SQL Command Line

Login db1.msi.umn.edu
. /usr/local/mysql/mysql_client
mysql -h 127.0.0.1 -P 9903 -u geneontology -p
Enter password:

```sql
mysql>
select name from db;
+--------------------+
<table>
<thead>
<tr>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>AgBase</td>
</tr>
<tr>
<td>CGD</td>
</tr>
<tr>
<td>DDB</td>
</tr>
<tr>
<td>FB</td>
</tr>
<tr>
<td>GDB</td>
</tr>
<tr>
<td>GeneDB_Lmajor</td>
</tr>
<tr>
<td>GeneDB_Pfalciparum</td>
</tr>
<tr>
<td>GeneDB_Spombe</td>
</tr>
<tr>
<td>GeneDB_Tbrucei</td>
</tr>
<tr>
<td>GOA</td>
</tr>
<tr>
<td>GR</td>
</tr>
<tr>
<td>HGNC</td>
</tr>
<tr>
<td>IntAct</td>
</tr>
<tr>
<td>MGI</td>
</tr>
<tr>
<td>PINC</td>
</tr>
<tr>
<td>Reactome</td>
</tr>
<tr>
<td>RGD</td>
</tr>
<tr>
<td>SANGER</td>
</tr>
<tr>
<td>SGD</td>
</tr>
<tr>
<td>TAIR</td>
</tr>
<tr>
<td>TIGR</td>
</tr>
<tr>
<td>UniProt</td>
</tr>
<tr>
<td>WB</td>
</tr>
<tr>
<td>ZFIN</td>
</tr>
</tbody>
</table>
+--------------------+
24 rows in set (0.04 sec)
```

```sql
mysql> show tables;
+------------------------+
| Tables_in_geneontology |
+------------------------+
| assoc_rel              |
| association            |
| association_qualifier  |
| db                     |
| dbxref                 |
| evidence               |
| evidence_dbxref        |
| gene_product           |
| gene_product_count     |
| gene_product_property  |
| gene_product_seq       |
| gene_product_synonym   |
| graph_path             |
| graph_path2term        |
| instance_data          |
| seq                    |
| seq_dbxref             |
| seq_property           |
| source_audit           |
| species                |
| term                   |
| term2term              |
| term_audit             |
| term_dbxref            |
| term_definition        |
| term_synonym           |
+------------------------+
26 rows in set (0.00 sec)
```
SQL Command Line

Say we want to find the total number of gene products that are BOTH GTP binding (GO:0005525) and immune response (GO:0006955)

```
SELECT count(DISTINCT a1.gene_product_id)
FROM term AS t1
    INNER JOIN graph_path AS p1 ON (t1.id=p1.term1_id)
    INNER JOIN association AS a1 ON (a1.term_id=p1.term2_id)
    INNER JOIN term AS t2 ON (t2.id=p2.term1_id)
    INNER JOIN graph_path AS p2 ON (a2.term_id=p2.term2_id)
    INNER JOIN association AS a2 ON (a2.gene_product_id=a1.gene_product_id)
WHERE t1.acc = 'GO:0005525' AND t2.acc = 'GO:0006955';
```

<table>
<thead>
<tr>
<th>count(DISTINCT a1.gene_product_id)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
</tr>
</tbody>
</table>
#!/usr/local/bin/perl
use GO::AppHandle;

my $dbname = "geneontology";
my $mysqlhost = "127.0.0.1:9903";
my $user = "geneontology";
my $passwd = "gois_here";

$apph = GO::AppHandle-&gt;connect(-dbname=&gt;$dbname, -dbhost=&gt;$mysqlhost, -dbuser=&gt;$user, -dbauth=&gt;$passwd);

$product =$apph-&gt;get_product({symbol=&gt;"Cyp1a1"});
printf "Product; name=\%s  Acc=\%s\n",
    $product-&gt;full_name(),
    $product-&gt;acc();

• -bash-3.00$ ./symbol.pl
• Product; name=cytochrome P450, family 1, subfamily a, polypeptide 1  Acc=MGI:88588
GO Tools
GO Tools


• Consortium Tools:
  • AmiGO
  • DAG-Edit

• Non-Consortium Tools:
  – Search and browse
    • GOFish, QuickGO, ….
  – Annotation
    • Manatee, GeneTools,…
  – Gene expression
    • BiNGO, GeneMerge, GOArray, GO Term Finder, …
  – Others
    • Blast2GO, Generic GO term Mapper, GO SLIM Mapper, …
### Three major goals:

1. **Acp53C14c**, *Acp53C14c*  
   - Gene from *Drosophila melanogaster*, data from FlyBase (FBgn0053530)  
   - **Term**: extracellular region

2. **Acp53Ea**, *Accessory gland-specific peptide 53Ea*  
   - Gene from *Drosophila melanogaster*, data from FlyBase (FBgn0015584)  
   - **Term**: physiological process, post mating behavior, sperm competition, sperm displacement, extracellular region, hormone activity

3. **Acp53Eb**, *Accessory gland protein 53Eb*  
   - Gene from *Drosophila melanogaster*, data from FlyBase (FBgn0024500)  
   - **Term**: sperm displacement

4. **Arp53D**, *Actin-related protein 53D*  
   - Gene from *Drosophila melanogaster*, data from FlyBase (FBgn0011743)  
   - **Term**: cytoskeleton organization and biogenesis
GOFish Tool

Three major goals:

- apoptosis regulator
- death

Selected GO terms:

A. apoptosis regulator
B. death

Search GO terms:

- GO term word(s)
- Start(s) of GO term word(s)
- Start of GO term
- GO term ID(s)

Ranked gene products:

<table>
<thead>
<tr>
<th>QS</th>
<th>Name</th>
<th>A</th>
<th>B</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>Bax</td>
<td>1</td>
<td>1</td>
<td>Bcl2-associ</td>
</tr>
<tr>
<td>21</td>
<td>Bcl2</td>
<td>1</td>
<td>1</td>
<td>B-cell leuq</td>
</tr>
<tr>
<td>31</td>
<td>Apaf1</td>
<td>1</td>
<td>1</td>
<td>apoptotic</td>
</tr>
<tr>
<td>41</td>
<td>Bcl1</td>
<td>1</td>
<td>1</td>
<td>B-cell leuq</td>
</tr>
<tr>
<td>51</td>
<td>Bak1</td>
<td>1</td>
<td>1</td>
<td>BCL2-ant</td>
</tr>
<tr>
<td>61</td>
<td>Bag3</td>
<td>1</td>
<td>1</td>
<td>Bcl2-associ</td>
</tr>
<tr>
<td>71</td>
<td>Bid</td>
<td>1</td>
<td>1</td>
<td>BH3 inter</td>
</tr>
<tr>
<td>81</td>
<td>Fadd</td>
<td>1</td>
<td>1</td>
<td>Fas (TNF)</td>
</tr>
<tr>
<td>91</td>
<td>Bad</td>
<td>1</td>
<td>1</td>
<td>Bcl-associ</td>
</tr>
<tr>
<td>101</td>
<td>Crodd</td>
<td>1</td>
<td>1</td>
<td>CASP2 ar</td>
</tr>
<tr>
<td>111</td>
<td>Cflar</td>
<td>1</td>
<td>1</td>
<td>CASP8 ar</td>
</tr>
<tr>
<td>121</td>
<td>Dffb</td>
<td>1</td>
<td>1</td>
<td>DNA fragt</td>
</tr>
<tr>
<td>131</td>
<td>Dff3</td>
<td>1</td>
<td>1</td>
<td>DNA fragt</td>
</tr>
<tr>
<td>141</td>
<td>Bcl2l</td>
<td>1</td>
<td>1</td>
<td>Bcl2-like</td>
</tr>
<tr>
<td>151</td>
<td>Birc1b</td>
<td>1</td>
<td>1</td>
<td>baculovir</td>
</tr>
<tr>
<td>161</td>
<td>Birc1a</td>
<td>1</td>
<td>1</td>
<td>baculovir</td>
</tr>
<tr>
<td>171</td>
<td>Birc1f</td>
<td>1</td>
<td>1</td>
<td>baculovir</td>
</tr>
<tr>
<td>181</td>
<td>Birc1e</td>
<td>1</td>
<td>1</td>
<td>baculovir</td>
</tr>
</tbody>
</table>

Gene product viewer:

- MGD name: Bcl2-associated X protein
- MGD ID: MGI:99702
- Marker symbol: Bax
- Chromosome: VII
Three major goals:

QuickGO is a fast web-based browser of the Gene Ontology data (see geneontology.org) based at the EBI, as well as the annotation of GO to UniProt and InterPro generated by the GOA project. It integrates into InterPro, providing links between the two data sets that are navigable via the web. Various search facilities also exist.

See the documentation index.

Summary
Onto-Express (OE)

http://vortex.cs.wayne.edu/ontoexpress/servlet/UserInfo

Intelligent Systems and Bioinformatics Laboratory, Wayne State University

- Automatically translate gene lists of differentially regulated genes into functional profiles
- Functional profiles: biochemical function, biological process, cellular role, cellular component, molecular function and chromosome location.
- Statistical significance values are calculated for each category.
Onto-Express (OE)

- Login (c:\temp\go-demo)
- Run Onto-express
- Input:
  - Input file: interested gene list (209) from microarray analysis
  - Organism: (homo sapiens)
  - Input type: (affymetrix probe id)
  - Reference Array: (affymetrix human genome u133a array)
  - Distribution:
  - Correction:
  - Search for:

![Onto-Express Input Form](image)
<table>
<thead>
<tr>
<th>P-Value</th>
<th>Corrected P-Value</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02087</td>
<td>0.002846</td>
<td>1</td>
</tr>
<tr>
<td>0.00277</td>
<td>0.000354</td>
<td>1</td>
</tr>
<tr>
<td>0.03821</td>
<td>0.04231</td>
<td>1</td>
</tr>
<tr>
<td>0.03808</td>
<td>0.04305</td>
<td>1</td>
</tr>
<tr>
<td>0.00131</td>
<td>0.00812</td>
<td>2</td>
</tr>
<tr>
<td>0.01479</td>
<td>0.02428</td>
<td>2</td>
</tr>
<tr>
<td>0.02226</td>
<td>0.0397</td>
<td>2</td>
</tr>
<tr>
<td>1.1E-4</td>
<td>0.00105</td>
<td>7</td>
</tr>
<tr>
<td>0.05585</td>
<td>0.003527</td>
<td>1</td>
</tr>
<tr>
<td>1.0E-5</td>
<td>2.1E-4</td>
<td>4</td>
</tr>
<tr>
<td>0.0</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>6.0E-5</td>
<td>6.3E-4</td>
<td>4</td>
</tr>
<tr>
<td>0.0</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>0.01390</td>
<td>0.08523</td>
<td>2</td>
</tr>
<tr>
<td>8.0E-5</td>
<td>8.3E-4</td>
<td>12</td>
</tr>
</tbody>
</table>

- **Onto-Express (OE)**
- **http://vortex.cs.wayne.edu/ontoexpress/servlet/UserInfo**
Onto-Express (OE)

http://vortex.cs.wayne.edu/ontoexpress/servlet/UserInfo

Functional Category:
Number of Genes:
(Right click on each pie for additional options.)

- apoptosis
- regulation of cell cycle
- intracellular signaling cascade
- proteolysis and peptidolysis
- cell-cell signaling
- antigen processing, exogenous antigen via MHC class II
- antigen presentation, exogenous antigen

Legend:
User Interactions:
- Unselected Function
- Synchronized Function
- More Than Expected
- Less Than Expected
- Same As Expected

Gene Regulation:
- Positive
- Negative
- No Change
MGI Gene Ontology GO_Slim Chart Tool

This GO_Slim Chart Tool bins the genes in your list according to MGI GO Slim definitions to help you discover common.

Step 1: Enter gene names*:
Input the gene names in the text box, as MGI:accID or gene symbol (with carriage returns) or Select a file of gene names, as MGI:accID or gene symbol (with carriage returns).

Step 2: Choose Ontology:
- [ ] Process
- [ ] Function
- [ ] Component

Step 3:
Indicate whether to exclude evidence code IEA (Inferred from Electronic Annotation):
- [ ] Include IEAs
- [ ] Exclude IEAs

Step 4:
- [ ] Search
- [ ] Reset

Uses TermFinder (0.5) implementation of Gavin Sherlock, Stanford Microarray Database.
Three major goals:

Note: overrepresentation of genes in bins for "DNA metabolism" and "stress response" compared with MGI overall.