Systematic Annotation

Mark Voorhies

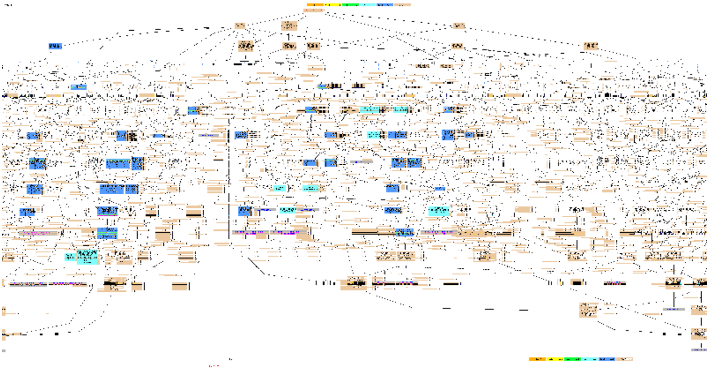
4/5/2011

The Gene Ontology

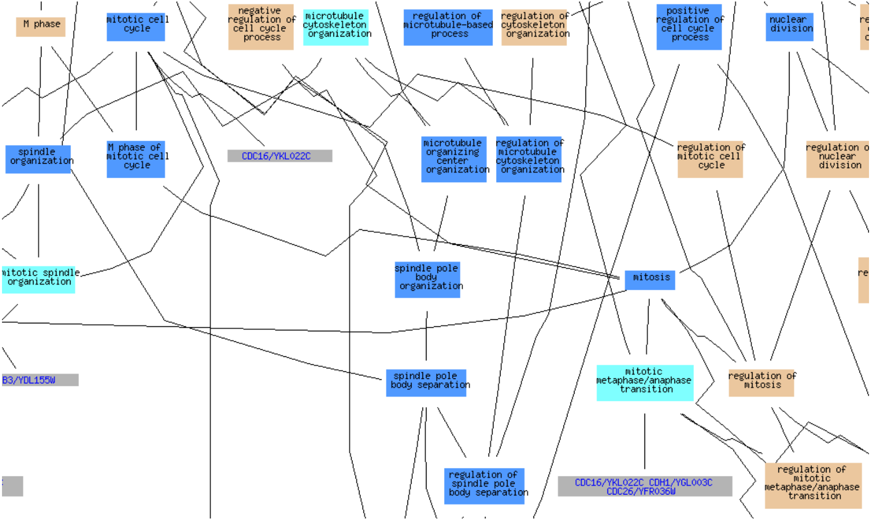
Three directed acyclic graphs (aspects):

- Biological Process
- Molecular Function
- Subcellular Component

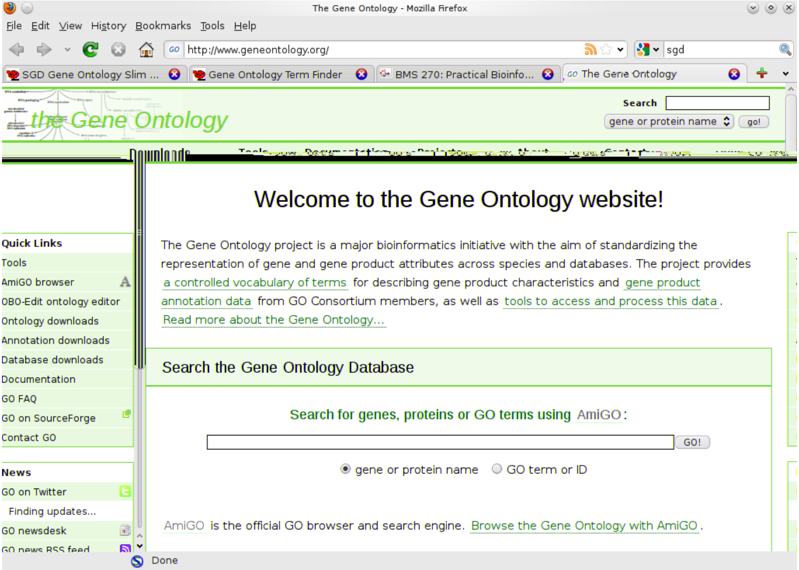
The Gene Ontology



The Gene Ontology


M phase
mitotic cell cycle
negative regulation of cell cycle process
microtubule cytoskeleton organization
regulation of microtubule-based process
regulation of cytoskeleton organization
positive regulation of cell cycle process
nuclear division
re
spindle organization
M phase of mitotic cell cycle
CDC16/YKL022C
microtubule organizing center organization
regulation of microtubule cytoskeleton organization
regulation of mitotic cell cycle
regulation o nuclear division
nitotic spindle organization
spindle pole body organization
mitosis
re
B3/YDL155W
spindle pole body separation
mitotic metaphase/anaphase transition
regulation of mitosis
regulation of spindle pole body separation
CDC16/YKL022C CDH1/YGL003C CDC26/YFR036W
regulation of mitotic metaphase/anaphase transition

The AmiGO browser


The Gene Ontology - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.geneontology.org/
sgd
SGD Gene Ontology Slim ...
Gene Ontology Term Finder
BMS 270: Practical Bioinfo...
The Gene Ontology
the Gene Ontology
Search
gene or protein name
go!
Downloads
Tools
Documentation
Projects
About
Done

Welcome to the Gene Ontology website!

The Gene Ontology project is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases. The project provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data from GO Consortium members, as well as tools to access and process this data. Read more about the Gene Ontology...

Quick Links

- Tools
- AmiGO browser
- OBO-Edit ontology editor
- Ontology downloads
- Annotation downloads
- Database downloads
- Documentation
- GO FAQ
- GO on SourceForge
- Contact GO

News

- GO on Twitter
- Finding updates...
- GO newsdesk
- GO news RSS feed

Search the Gene Ontology Database

Search for genes, proteins or GO terms using AmiGO:

GO!

● gene or protein name ○ GO term or ID

AmiGO is the official GO browser and search engine. Browse the Gene Ontology with AmiGO.

The Gene Ontology

Associating GO terms

How might we annotate genes with GO terms?

Associating GO terms

How might we annotate genes with GO terms?

- By sequence homology (*e.g.*, BLAST)
- By domain homology (*e.g.*, InterProScan)
- Mapping from an annotated relative (*e.g.*, INPARANOID)
- Human curation of the literature (*e.g.*, SGD)

Associating GO terms: Evidence codes

- Experimental
  - EXP: Inferred from Experiment
  - IDA: Inferred from Direct Assay
  - IPI: Inferred from Physical Interaction
  - IMP: Inferred from Mutant Phenotype
  - IGI: Inferred from Genetic Interaction
  - IEP: Inferred from Expression Pattern
- Computational Analysis
  - ISS: Inferred from Sequence or Structural Similarity
  - ISO: Inferred from Sequence Orthology
  - ISA: Inferred from Sequence Alignment
  - ISM: Inferred from Sequence Model
  - IGC: Inferred from Genomic Context
  - RCA: inferred from Reviewed Computational Analysis
- Author Statement
  - TAS: Traceable Author Statement
  - NAS: Non-traceable Author Statement
  - Curator Statement Evidence Codes
  - IC: Inferred by Curator
  - ND: No biological Data available
- Automatically-assigned
  - IEA: Inferred from Electronic Annotation
- Obsolete
  - NR: Not Recorded

The Gene Ontology

- How might we annotate genes with GO terms?
- How do we calculate the significance of the GO terms associated with a particular group of genes?

Sampling with replacement: Mutagenesis

How many transformants do we have to screen in order to “cover” a genome?

Sampling with replacement: Mutagenesis

How many transformants do we have to screen in order to "cover" a genome?
Probability that a transformant has (1) disrupted gene: p_m
Number of genes in organsim: N_g

Sampling with replacement: Mutagenesis

How many transformants do we have to screen in order to “cover” a genome?
Probability that a transformant has (1) disrupted gene: p_m
Number of genes in organsim: N_g
Probability that a specific gene is disrupted in a specific transformant:

$$p_d = p_m \left(\frac{1}{N_g} \right) = \frac{p_m}{N_g} \tag{1}$$

Sampling with replacement: Mutagenesis

How many transformants do we have to screen in order to "cover" a genome?
Probability that a transformant has (1) disrupted gene: p_m
Number of genes in organsim: N_g
Probability that a specific gene is disrupted in a specific transformant:

$$p_d = p_m \left(\frac{1}{N_g} \right) = \frac{p_m}{N_g} \tag{1}$$

Probability of *not* disrupting that gene:

$$p_u = 1 - \frac{p_m}{N_g} \tag{2}$$

Sampling with replacement: Mutagenesis

Probability of *not* disrupting that gene:

$$p_u = 1 - \frac{p_m}{N_g} \tag{3}$$

Sampling with replacement: Mutagenesis

Probability of *not* disrupting that gene:

$$p_u = 1 - \frac{p_m}{N_g} \tag{3}$$

The probability of *not* disrupting that gene n independent times is:

$$p_{u,n} = \left(1 - \frac{p_m}{N_g}\right)^n \tag{4}$$

Sampling with replacement: Mutagenesis

Probability of *not* disrupting that gene:

$$p_u = 1 - \frac{p_m}{N_g} \tag{3}$$

The probability of *not* disrupting that gene n independent times is:

$$p_{u,n} = \left(1 - \frac{p_m}{N_g}\right)^n \tag{4}$$

And the probability *of* disrupting that gene n independent times is:

$$p_{d,n} = 1 - p_{u,n} = 1 - \left(1 - \frac{p_m}{N_g}\right)^n \tag{5}$$

Sampling with replacement: Mutagenesis

Probability of *not* disrupting that gene:

$$p_u = 1 - \frac{p_m}{N_g} \tag{3}$$

The probability of *not* disrupting that gene n independent times is:

$$p_{u,n} = \left(1 - \frac{p_m}{N_g}\right)^n \tag{4}$$

And the probability *of* disrupting that gene n independent times is:

$$p_{d,n} = 1 - p_{u,n} = 1 - \left(1 - \frac{p_m}{N_g}\right)^n \tag{5}$$

This is also the expected genome coverage.

Sampling with replacement: General Cases

Calculating the probability of *zero* events was easy.

$$p_{0,n} = \left(1 - \frac{p_m}{N_g}\right)^n \tag{6}$$

Sampling with replacement: General Cases

Calculating the probability of *zero* events was easy.

$$p_{0,n} = \left(1 - \frac{p_m}{N_g}\right)^n \tag{6}$$

What about exactly k events?

Sampling with replacement: General Cases

Calculating the probability of *zero* events was easy.

$$p_{0,n} = \left(1 - \frac{p_m}{N_g}\right)^n \tag{6}$$

What about exactly k events?
Binomial distribution:

$$p_{k,n} = \binom{n}{k} p_m^k (1 - p_m)^{n-k} \tag{7}$$

Sampling with replacement: General Cases

Calculating the probability of *zero* events was easy.

$$p_{0,n} = \Bigg(1 \qquad\qquad {}_{1}$$

Sampling without replacement: GO Annotation

The binomial distribution assumes that event probabilities are constant:

$$p_{k,n} = \binom{n}{k} p_m^k (1 - p_m)^{n-k} \tag{9}$$

What if there are m virulence factors in our genome, and every time we discover one it is magically removed from our library?

Sampling without replacement: GO Annotation

The binomial distribution assumes that event probabilities are constant:

$$p_{k,n} = \binom{n}{k} p_m^k (1 - p_m)^{n-k} \tag{9}$$

What if there are m virulence factors in our genome, and every time we discover one it is magically removed from our library? Hypergeometric distribution:

$$p_{k,m,n} = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \tag{10}$$

Sampling without replacement: GO Annotation

The binomial distribution assumes that event probabilities are constant:

$$p_{k,n} = \binom{n}{k} p_m^k (1 - p_m)^{n-k} \tag{9}$$

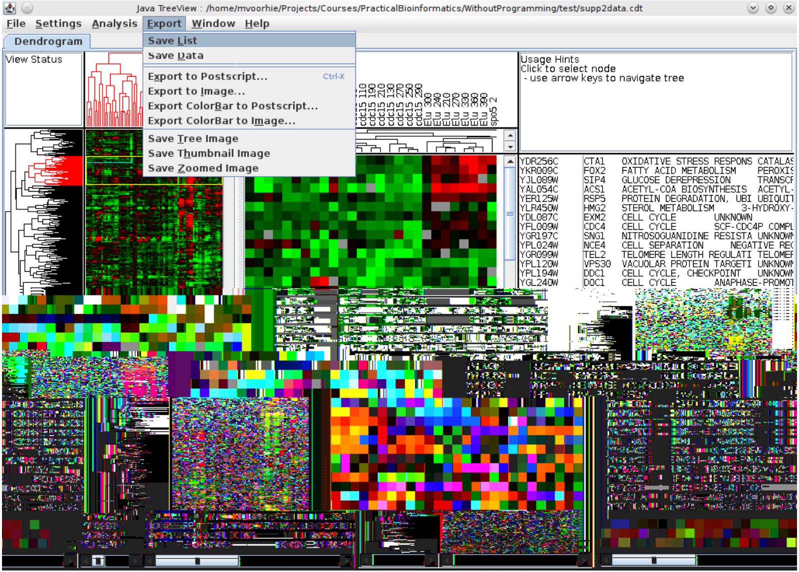
What if there are m virulence factors in our genome, and every time we discover one it is magically removed from our library? Hypergeometric distribution:

$$p_{k,m,n} = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \tag{10}$$

More than one *disjoint* type of label:

$$p_{k_1,k_2,\ldots,m_1,m_2,\ldots,n} = \frac{\prod \binom{m_i}{k_i}}{\binom{N}{n}} \tag{11}$$

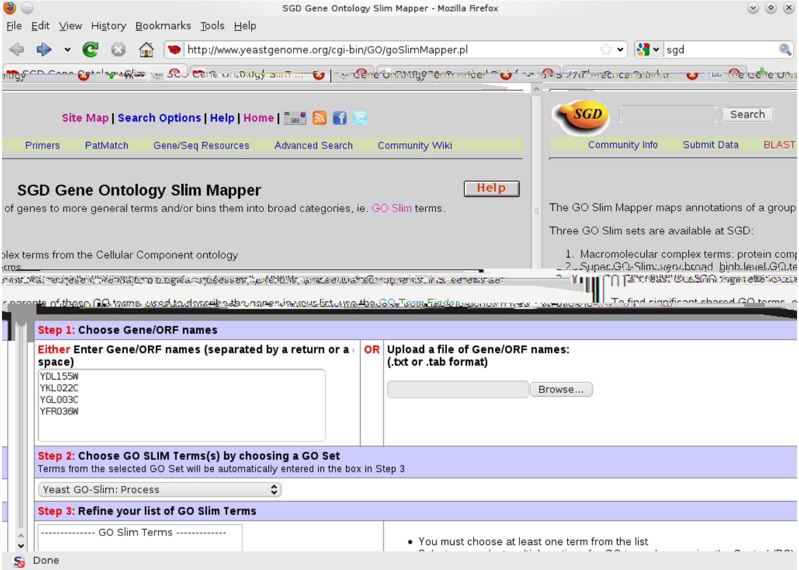
Extracting gene lists from JavaTreeView



Usage Hints
Click to select node
 - use arrow keys to navigate tree

```
YDR256C    CTA1   OXIDATIVE STRESS RESPONS CATALAS
YKR009C    FOX2   FATTY ACID METABOLISM    PEROXIS
YJL089W    SIP4   GLUCOSE DEREPRESSION     TRANSCR
YAL054C    ACS1   ACETYL-COA BIOSYNTHESIS  ACETYL-
YER125W    RSP5   PROTEIN DEGRADATION, UBI UBIQUIT
YLR450W    HMG2   STEROL METABOLISM      3-HYDROXY-
YDL087C    EXM2   CELL CYCLE          UNKNOWN
YFL009W    CDC4   CELL CYCLE          SCF-CDC4P COMPL
YGR197C    SNG1   NITROSOGUANIDINE RESISTA UNKNOWN
YPL024W    NCE4   CELL SEPARATION       NEGATIVE REG
YGR099W    TEL2   TELOMERE LENGTH REGULATI TELOMER
YPL120W    VPS30  VACUOLAR PROTEIN TARGETI UNKNOWN
YPL194W    DDC1   CELL CYCLE, CHECKPOINT   UNKNOWN
YGL240W    DOC1   CELL CYCLE          ANAPHASE-PROMOT
```

The SGD GO Slim Mapper


SGD Gene Ontology Slim Mapper - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.yeastgenome.org/cgi-bin/GO/goSlimMapper.pl
sgd
Site Map | Search Options | Help | Home |
SGD
Search
Primers PatMatch Gene/Seq Resources Advanced Search Community Wiki
Community Info Submit Data BLAST
Help
Browse...
Done

SGD Gene Ontology Slim Mapper

The GO Slim Mapper maps annotations of a group of genes to more general terms and/or bins them into broad categories, ie. GO Slim terms.

Three GO Slim sets are available at SGD:

1. Macromolecular complex terms: protein comp…plex terms from the Cellular Component ontology
2. Super GO-Slim: very broad, high level GO te…rms

…parents of those GO terms used to describe the genes in your list, use the GO Term Finder

To find significant shared GO terms, o…

Step 1: Choose Gene/ORF names

Either Enter Gene/ORF names (separated by a return or a space)

```
YDL155W
YKL022C
YGL003C
YFR036W
```

OR Upload a file of Gene/ORF names: (.txt or .tab format)

Step 2: Choose GO SLIM Terms(s) by choosing a GO Set

Terms from the selected GO Set will be automatically entered in the box in Step 3

Yeast GO-Slim: Process

Step 3: Refine your list of GO Slim Terms

-------------- GO Slim Terms -------------

- You must choose at least one term from the list

Alternatives to Hierarchical Clustering

- GORDER and pre-clustering by SOM

Alternatives to Hierarchical Clustering

- GORDER and pre-clustering by SOM
- Pre-calling number of clusters: k-means and k-medians

Alternatives to Hierarchical Clustering

- GORDER and pre-clustering by SOM
- Pre-calling number of clusters: k-means and k-medians
- Principal Component Analysis (PCA)
    - See also ICA (Independent Component Analysis)

Homework

Read:

- PNAS 98:5116 (SAM)
- BMC Bioinformatics 5:54 (BAGEL)