

# Systematic Annotation

Mark Voorhies

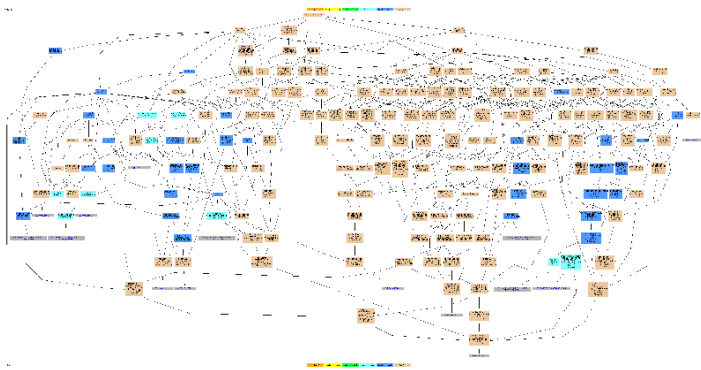
4/5/2012

- RTFM
- PNAS 95:14863

Three directed acyclic graphs (aspects):

- Biological Process
- Molecular Function
- Subcellular Component

# The Gene Ontology





# The AmiGO browser

The screenshot shows a Mozilla Firefox browser window titled "The Gene Ontology - Mozilla Firefox". The address bar contains "http://www.geneontology.org/". The browser's menu bar includes "File", "Edit", "View", "History", "Bookmarks", "Tools", and "Help". The page content features the "the Gene Ontology" logo and a navigation menu with links for "Downloads", "Tools", "Documentation", "Projects", "About", and "Contact". A search bar is located in the top right corner, with a dropdown menu showing "gene or protein name" and a "go!" button. The main heading reads "Welcome to the Gene Ontology website!". Below this, a paragraph describes the project's goal of standardizing gene and gene product attributes. A search box is provided with the text "Search for genes, proteins or GO terms using AmiGO:" and a "GO!" button. Radio buttons allow users to search by "gene or protein name" (selected) or "GO term or ID". A "Quick Links" sidebar on the right lists various tools and resources, including "AmiGO browser", "OBO-Edit ontology editor", and "GO on SourceForge". A "News" sidebar lists "GO on Twitter" and "GO newsdesk". The browser's status bar at the bottom shows "Done".

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...](#)

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](#).

Done

**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

# 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?

How might we annotate genes with 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 organism:  $N_g$

# 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 organism:  $N_g$

Probability that a specific gene is disrupted in a specific transformant:

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

# 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 organism:  $N_g$

Probability that a specific gene is disrupted in a specific transformant:

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

Probability of *not* disrupting that gene:

$$p_u = 1 - \frac{p_m}{N_g} \quad (2)$$

# Sampling with replacement: Mutagenesis

Probability of *not* disrupting that gene:

$$p_u = 1 - \frac{p_m}{N_g} \quad (3)$$



# Sampling with replacement: Mutagenesis

Probability of *not* disrupting that gene:

$$p_u = 1 - \frac{p_m}{N_g} \quad (3)$$

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

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

# Sampling with replacement: Mutagenesis

Probability of *not* disrupting that gene:

$$p_u = 1 - \frac{p_m}{N_g} \quad (3)$$

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

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

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

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

# Sampling with replacement: Mutagenesis

Probability of *not* disrupting that gene:

$$p_u = 1 - \frac{p_m}{N_g} \quad (3)$$

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

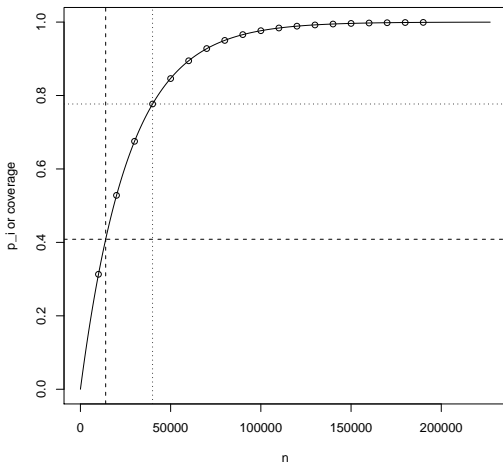
$$p_{u,n} = \left(1 - \frac{p_m}{N_g}\right)^n \quad (4)$$

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

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

This is also the expected genome coverage.

# Sampling with replacement: Mutagenesis



# Sampling with replacement: General Cases

Calculating the probability of *zero* events was easy.

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

# Sampling with replacement: General Cases

Calculating the probability of *zero* events was easy.

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

What about exactly  $k$  events?

# Sampling with replacement: General Cases

Calculating the probability of *zero* events was easy.

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

What about exactly  $k$  events?

Binomial distribution:

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

# Sampling with replacement: General Cases

Calculating the probability of *zero* events was easy.

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

What about exactly  $k$  events?

Binomial distribution:

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

What if there is more than one type of event?



# Sampling with replacement: General Cases

Calculating the probability of *zero* events was easy.

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

What about exactly *k* events?

Binomial distribution:

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

What if there is more than one type of event?

Multinomial distribution:

$$p_{k_1, k_2, \dots, n} = \frac{n!}{\prod k_i!} \prod p_i^{k_i} \quad (8)$$

# Sampling without replacement: GO Annotation

The binomial distribution assumes that event probabilities are constant:

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

# Sampling without replacement: GO Annotation

The binomial distribution assumes that event probabilities are constant:

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

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:

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

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:

$$p_{k,m,n} = \frac{\binom{m}{k} \binom{N-m}{n-k}}{\binom{N}{n}} \quad (10)$$

# Sampling without replacement: GO Annotation

The binomial distribution assumes that event probabilities are constant:

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

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:

$$p_{k,m,n} = \frac{\binom{m}{k} \binom{N-m}{n-k}}{\binom{N}{n}} \quad (10)$$

More than one *disjoint* type of label:

$$p_{k_1, k_2, \dots, m_1, m_2, \dots, n} = \frac{\prod \binom{m_i}{k_i}}{\binom{N}{n}} \quad (11)$$

# Extracting gene lists from JavaTreeView

The screenshot shows the JavaTreeView application interface. The main window displays a dendrogram on the left, a heatmap in the center, and a list of genes with their systematic annotations on the right. A menu is open over the dendrogram, showing options like 'Save List', 'Export to Postscript...', and 'Export ColorBar to Postscript...'. The heatmap shows a color scale from green (low expression) to red (high expression). The gene list on the right includes gene IDs and their corresponding GO terms.

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

YDR256C	CTA1	OXIDATIVE STRESS RESPONSA	CATALAS
YKR009C	FOX2	FATTY ACID METABOLISM	PEROXIS
YJL069W	STP4	GLUCOSE DEPRESSION	TRANSC
YAL054C	ACS1	ACETYL-CoA BIOSYNTHESIS	ACETYL-
YER125W	RSP5	PROTEIN DEGRADATION, UBI	UBIQUITI
YLR450W	HM62	STEROL METABOLISM	3-HYDROXY-
YDL087C	EXM2	CELL CYCLE	UNKNOWN
YFL009W	CDC4	CELL CYCLE	SGF-CDC4P COMPL
YGR197C	SNG1	NITROSOGUANIDINE RESISTA	UNKNOWN
YFL024W	NCE4	CELL SEPARATION	NEGATIVE REG
YGR099W	TEL2	TELOMERE LENGTH REGULATI	TELOMERE
YPL120W	VPS30	VACUOLAR PROTEIN TARGETI	UNKNOWN
YPL194W	DOC1	CELL CYCLE, CHECKPOINT	UNKNOWN
YGL240W	DOC1	CELL CYCLE	ANAPHASE-PROMOT
YML016W	DUR3	TRANSPORT	UREA PERMEASE
YHR154W	ESC4	SILENCING	UNKNOWN
YFL053W	DAK2	CARBOHYDRATE METABOLISM;	DIHYDRO
YBR294W	SUL1	TRANSPORT	SULFATE PERMEAS
YDR439W	LRS4	TRANSCRIPTION/RDNA SILEN	UNKNOWN
YML012W	SPO1	MEIOSIS (SPOR.)	TRANSCRIPTIO
YLR033W	ATP10	ATP SYNTHESIS	F1F0 ATPASE
YGL102W	IME4	MEIOSIS	TRANSCRIPTION FAC
YML091C	RPM2	TRNA PROCESSING, MITOCHO	RNASE P
YMR056C	AAC1	TRANSPORT	MITOCHONDRIAL /
YBR001C	NTH2	TREHALOSE METABOLISM	ALPHA, /
YPL119C	DBP1	mRNA PROCESSING	RNA HELICASE
YGL180W	APG1	AUTOPHAGY	PROTEIN KINASE
YIL171W	HXT12	TRANSPORT	HEXOSE PERMEASE
YIL170W	HXT12	TRANSPORT	HEXOSE PERMEASE
YJL219W	HXT9	TRANSPORT	HEXOSE PERMEASE
YOL159W	HXT11	TRANSPORT	HEXOSE PERMEASE
YLR273C	PTG1	GLUCOSE DEPRESSION (PUTATI	VE)
YGL163C	RAD54	DNA REPAIR	DNA-DEPENDENT /
YPL152W	RFD2	DRUG RESISTANCE	UNKNOWN
YER014W	HEM14	HEME BIOSYNTHESIS	PROTOPORP
YGL203C	KEK1	SECRETION	CARBOXYPEPTIDA
YLR439W	CAR2	ARGININE METABOLISM	ORNITHINE
YPL111W	CAR1	ARGININE METABOLISM	ARGININASE
YML042W	CAT2	FATTY ACID TRANSPORT	CARNITIN
YDR059C	UBC5	PROTEIN DEGRADATION, UBI	E2 UB. .
YGR013W	SNL71	mRNA SPLICING	U1 SNFNP PRC
YJL214W	HXT8	TRANSPORT	HEXOSE PERMEASE

# 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 Gene Ontology Slim ... Gene Ontology Term Finder BMS 270: Practical Bioinfo... GO The Gene Ontology

**SGD** Search

Site Map | Search Options | Help | Home | RSS | Facebook | Twitter

Community Info Submit Data BLAST Primers PatMatch Gene/Seq Resources Advanced Search Community Wiki

## SGD Gene Ontology Slim Mapper Help

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 complex terms from the Cellular Component ontology
2. Super GO-Slim: very broad, high level GO terms
3. Yeast GO-Slim: high level GO terms that represent the major biological processes, functions, and cellular components in *S. cerevisiae*

To find significant shared GO terms, or parents of those GO terms, used to describe the genes in your list, use the [GO Term Finder](#).

### Step 1: Choose Gene/ORF names

**Either** Enter Gene/ORF names (separated by a return or a space) **OR** Upload a file of Gene/ORF names: (.txt or .tab format)

YDL155W  
YKL022C  
YGL003C  
YFR036W

Browse...

### 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

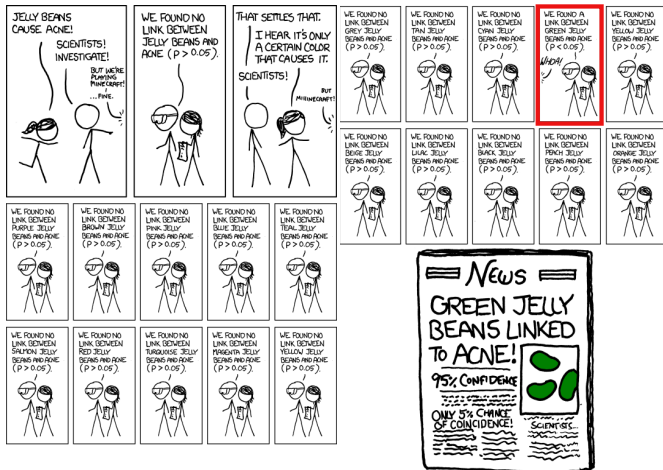
### Step 3: Refine your list of GO Slim Terms

..... GO Slim Terms .....

- You must choose at least one term from the list

Done

# Multiple Hypothesis Testing



<http://xkcd.com/882/>



# 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)

- Download PyMol