Systematic Annotation

Mark Voorhies

4/5/2012

Review

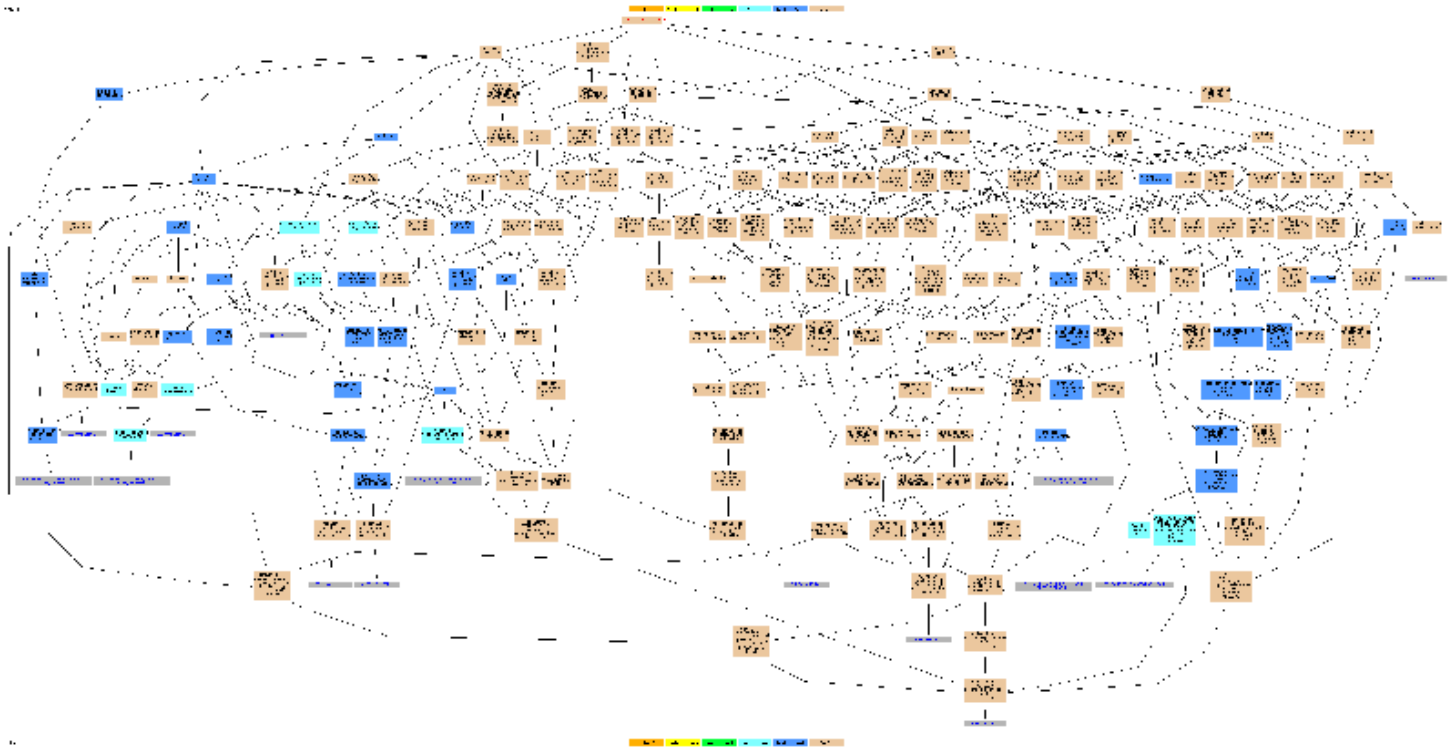
- RTFM
- PNAS 95:14863

The Gene Ontology

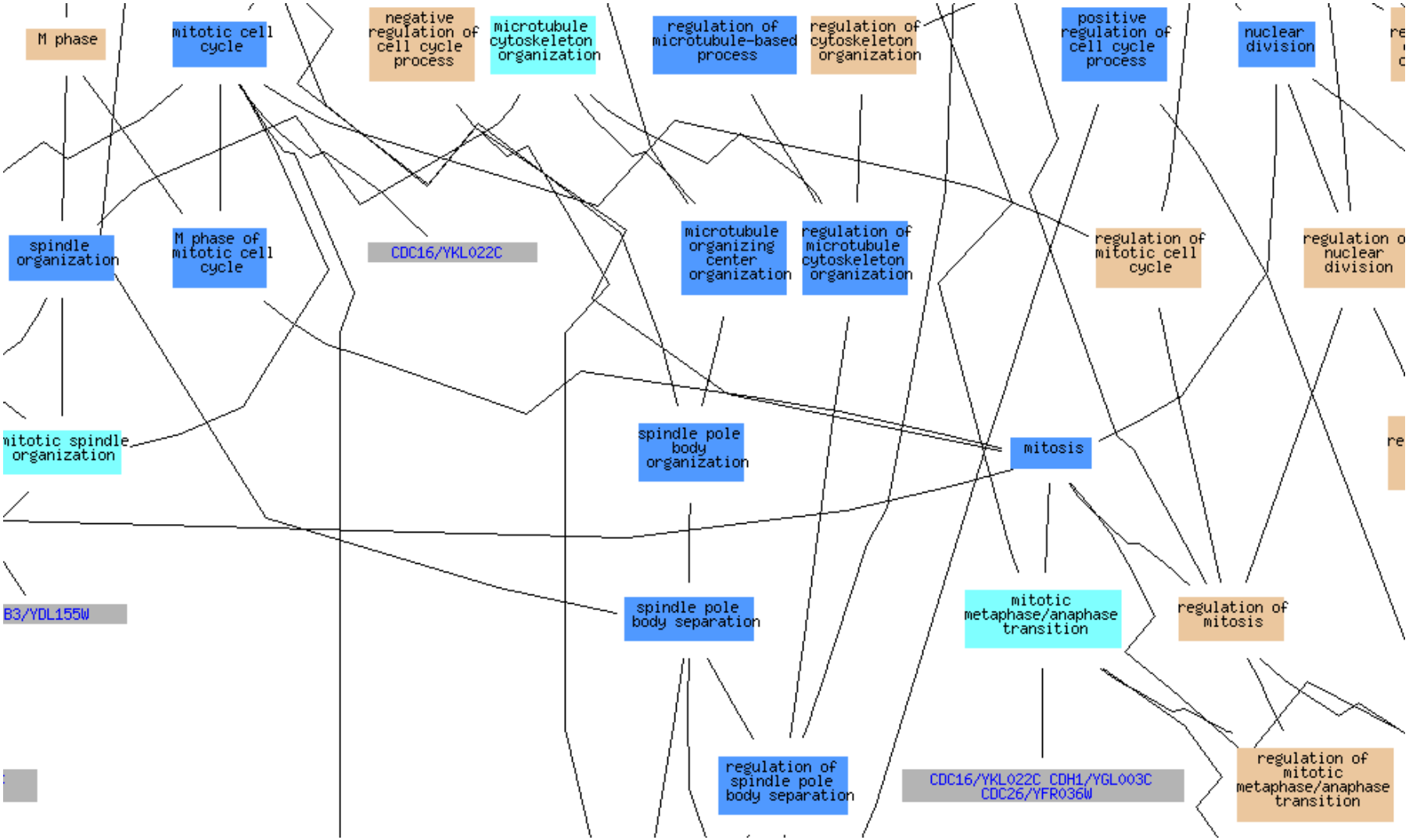
Three directed acyclic graphs (aspects):

- Biological Process
- Molecular Function
- Subcellular Component

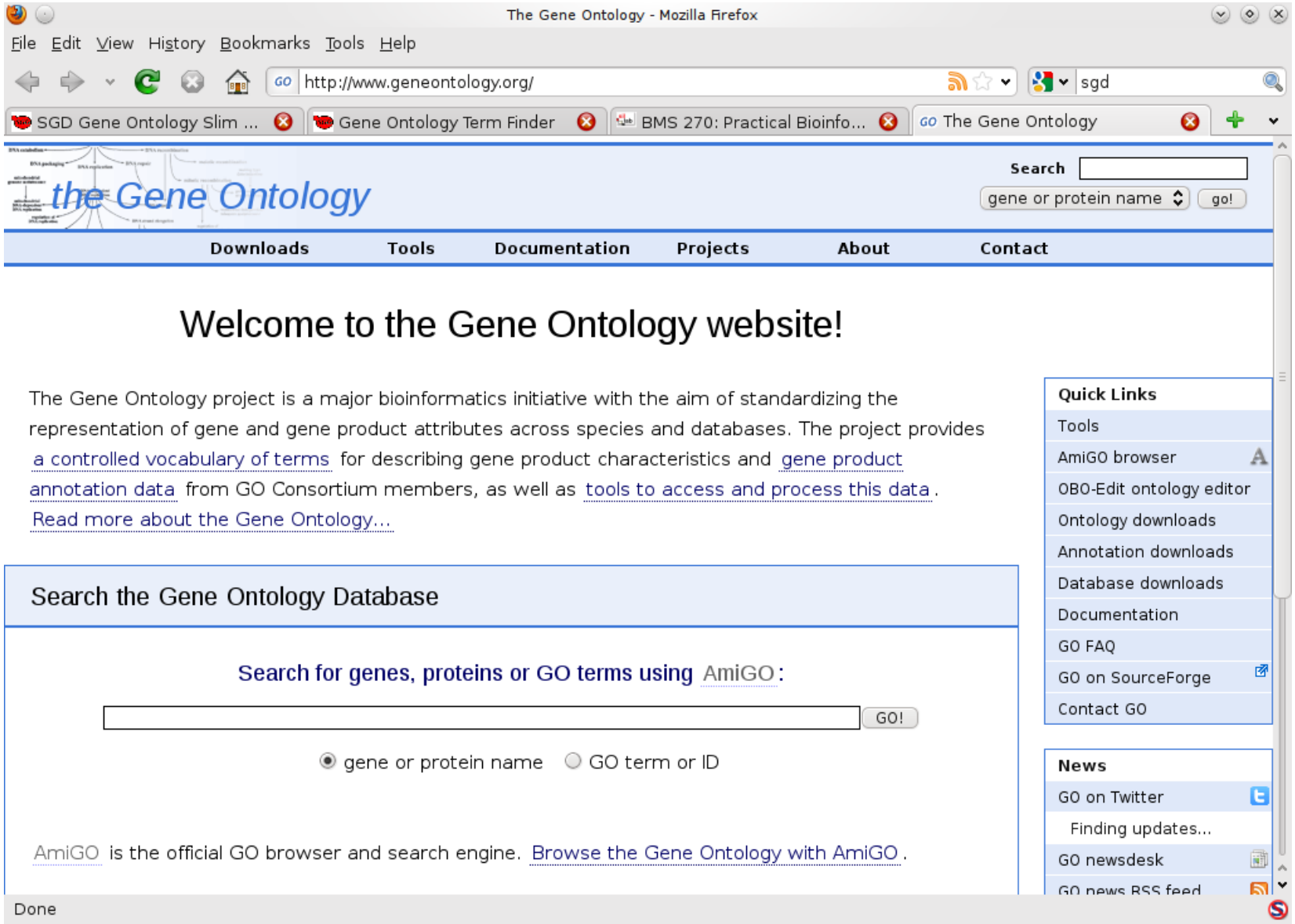
The Gene Ontology



The Gene Ontology



The AmiGO browser



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?

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:

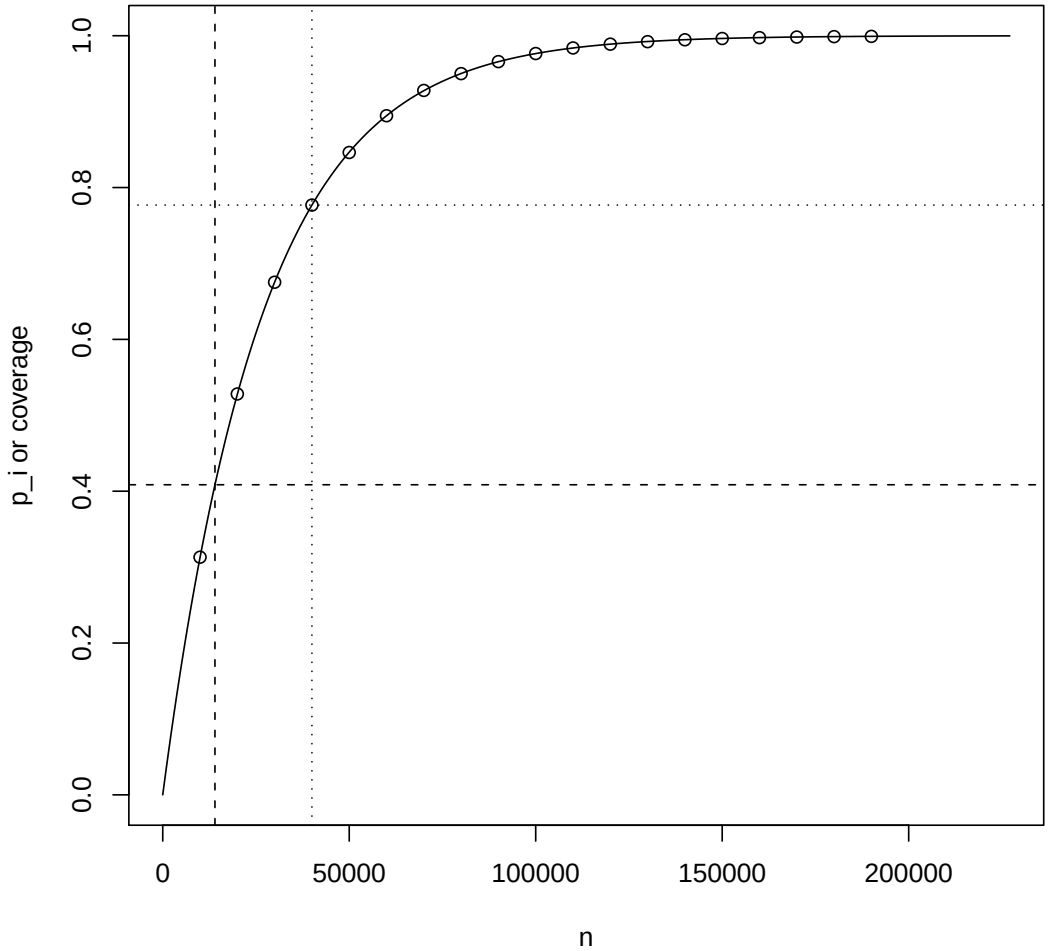
$$p_{u,n} = \left(1 - \frac{p_m}{N_g}\right)^n \tag{4}$$

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: Mutagenesis



Sampling with replacement: General Cases

Calculating the probability of *zero* events was easy.

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

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.

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

What about exactly k events?
Binomial distribution:

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

What if there is more than one type of event?

Sampling with replacement: General Cases

Calculating the probability of *zero* events was easy.

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

What about exactly k events?
Binomial distribution:

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

What if there is more than one type of event?
Multinomial distribution:

$$p_{k_1,k_2,\ldots,n} = \frac{n!}{\prod k_i!} \prod p_i^{k_i} \tag{8}$$

Sampling without replacement: GO Annotation

The binomial distribution assumes that event probabilities are constant:

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

Sampling without replacement: GO Annotation

The binomial distribution assumes that event probabilities are constant:

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

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:

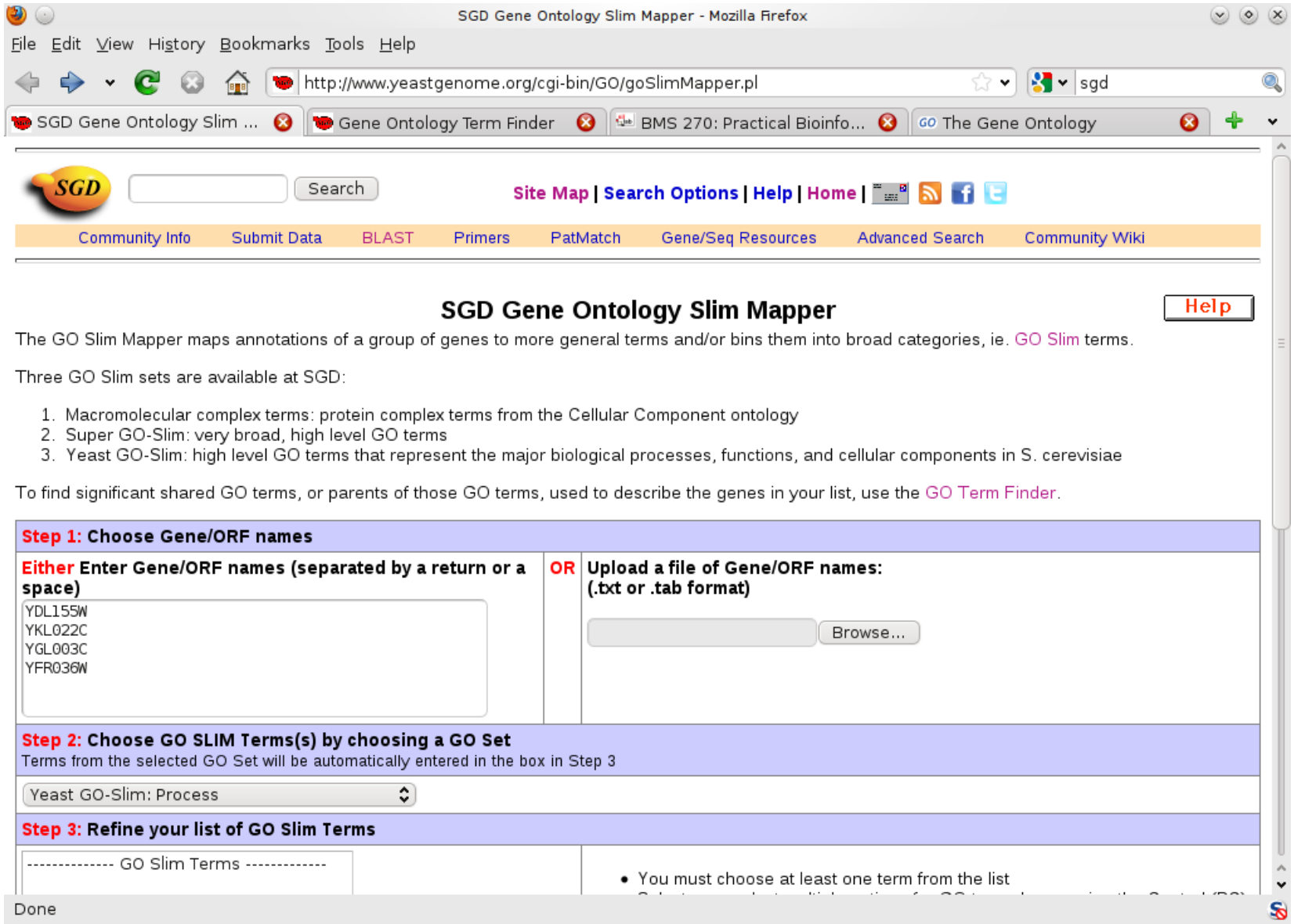
$$p_{k,m,n} = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \tag{10}$$

More than one *disjoint* type of label:

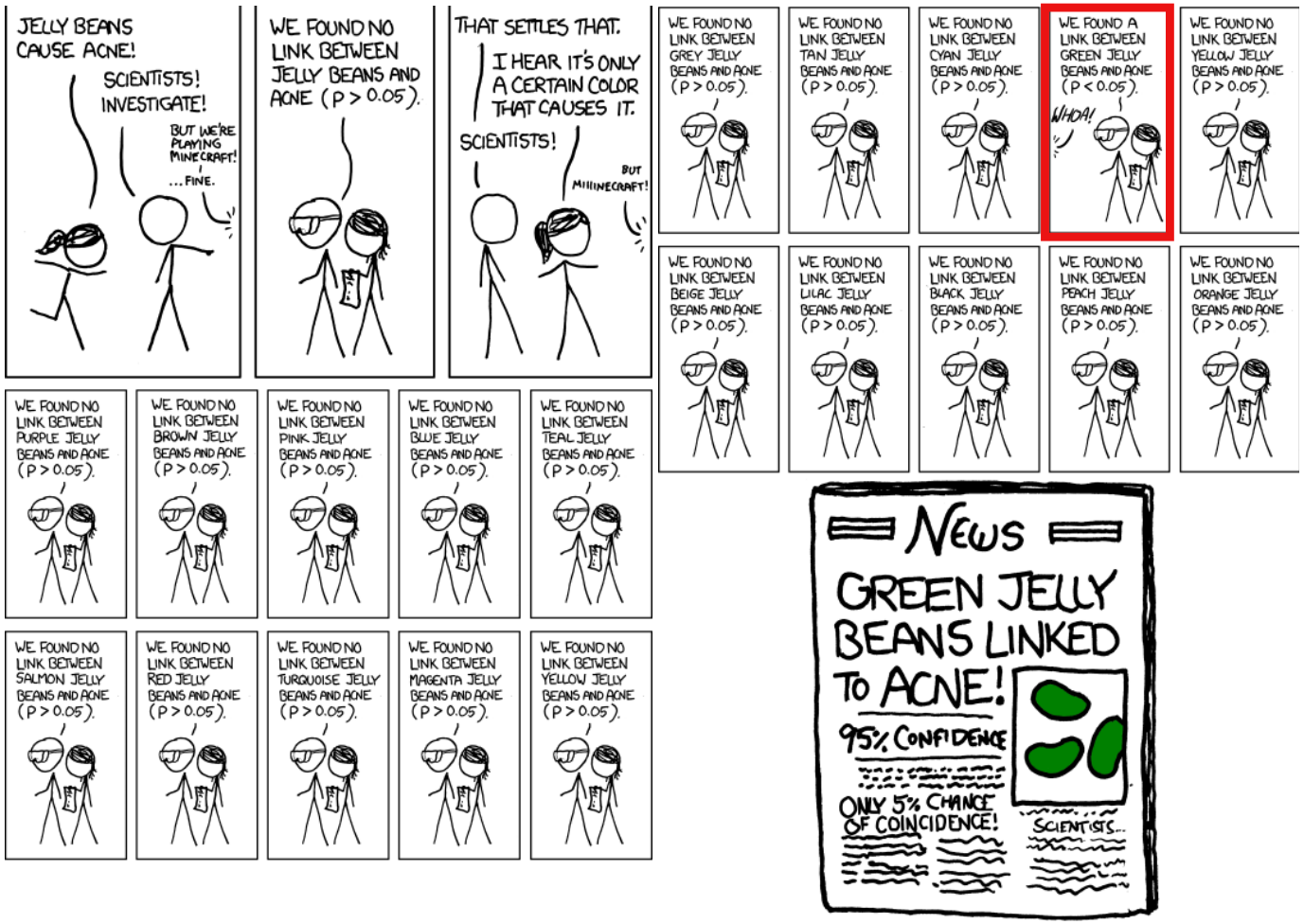
$$p_{k_1,k_2,\ldots,m_1,m_2,\ldots,n} = \frac{\prod \binom{m_i}{k_i}}{\binom{N}{n}} \tag{11}$$

Extracting gene lists from JavaTreeView

The SGD GO Slim Mapper



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)

Homework

- Download PyMol