

Finding and analyzing interesting (sub)groups of genes using the Gene Ontology

CBW Genomics Vancouver
Jochen Brumm
Department of Statistics and
Centre for Molecular Medicine and Therapeutics
University of British Columbia

jochen@stat.ubc.ca

August 19, 2004

Introduction: the Gene Ontology and other classification systems

- ◆ With increasing information available on genes and proteins, there are now a number of initiatives that try to organize this 'functional' information in a computer-accessible manner.
- ◆ The Gene Ontology (GO) is one of these major initiatives, and this afternoon we will introduce what it is and how you could use it
- ◆ We will see that GO is a useful tool in the integrated analysis of genomic data
- ◆ the focus of the afternoon will be on the use of GO to analyze groups of genes (arising from a microarray experiment, say)
- ◆ the lecture will focus on the over- or underrepresentation of a given GO term in a group of genes; the lab will show you how to do this sort of analysis

|← ← → →| ▼ ▸ * X W - 1 -

What is the Gene Ontology ?

The Gene Ontology organizes 'functional' information in a computer-accessible manner. It is comprised of:

- ◆ a controlled vocabulary
- ◆ semantic links between the entries of this vocabulary

So for example, we can express the statement

{'GO':0006950 ('response to stress')} {'is a'} {GO:0050896 ('response to stimulus')}

as a directed graph, which can be digested by a computer unambiguously. In this graph, the nodes are the *GO terms* and they are linked by their relationship.

|← ← → →| ▼ ▽ * X W - 2 -

There is one more requirement on the GO graph: it cannot have cycles. Hence we can describe the whole collection of GO terms and their relationships by a directed acyclic graph (DAG).

|← ← → →| ▼ ▽ * X W - 3 -

Construction of the Gene Ontology

How is GO created/maintained ?

- ◆ it is a community effort
- ◆ you can suggest new GO terms to be added to the GO graph.
- ◆ you can suggest genes to be annotated with GO terms
- ◆ for details, see the FAQ at:

<http://www.geneontology.org/>

|◀ ◀ ▶ ▶| ▼ ▽ * X W - 4 -

- ◆ the information is curated and there are evidence codes that describe on what basis an annotation for a gene is added, for example:
 - TAS traceable author statement
 - IC inferred by curator
 - IPI inferred from physical interaction
 - ISS inferred from sequence or structural similarity
- ◆ so the quality of evidence can vary considerably (TAS probably most stringent)
- ◆ also be aware of potential circularity if you are trying to use GO to validate your results

|◀ ◀ ▶ ▶| ▼ ▽ * X W - 5 -

How to use GO ?

- ◆ disclaimer: these concepts are very abstract, and there are many possible uses.
- ◆ we will focus on the following:
 - say we have a group of genes that are closely related in a set of experiments
 - we try to use GO to identify molecular basis of observed similarity (sort of ...) and identify known members of this mechanism
 - so for example, try to identify a subset of genes in a co-expression cluster that are co-regulated to identify transcription factor binding sites
- ◆ for this, we need to associate the genes in the clusters with GO terms

|← ← → →| ▼ ▸ * X W - 6 -

- ◆ the annotation is done on the most specific level, other GO terms are then implied by the GO association
- ◆ to create counts for GO terms: take each gene down the appropriate paths in the GO graph and add one to the count for each GO term visited
- ◆ this collection of counts is then the basis for further analysis
- ◆ so for example: look at the path 'biological process'-'physiological process'-'metabolism'-'biosynthesis'-'macromolecule biosynthesis'-'protein biosynthesis' - 'charged tRNA modification' which leads in the example of overexpressed genes you'll use later to the counts 660 (all) - 499 - 413 - 177 - 154 - 130 - 2

|← ← → →| ▼ ▸ * X W - 7 -

How to assess the count for a single GO term ?

- ◆ say we found a GO term of interest (say 'protein biosynthesis') and we would like to know if the observed count of this GO term is meaningful
- ◆ that is, how sure are we that 'protein biosynthesis' is really one of the biological processes that led to the observed expression similarity ?
- ◆ judge this by comparing it statistically to 'randomness'; that is we will use the deviation from random association as a measure of relevance

|← ← → →| ▼ ▸ * X W - 8 -

- ◆ we can adopt the following mindset:
 - view the event gene has GO term as a success, the opposite as failure
 - we have two groups: group 1 is the cluster of genes; group 2 is the rest of the genes
 - 'rest': can be all other queried genes on the array or the whole genome
 - note that 'rest' should be either close to a complete genome or a *random* selection of genes
 - now we can ask if the *rate of success in group 1* is different from the *rate of success in group 2*.
- ◆ the next slides will discuss this in some detail

|← ← → →| ▼ ▸ * X W - 9 -

How to compare two rates ?

- ◆ Now we have left all biology behind us and are purely talking about statistics. :-) / :-(
- ◆ Note that the way biological information enters is simply through counts of genes in categories.
- ◆ this shows the strong and weak point of this analysis: it is very generally applicable for any gene-counting problem, but it will not be very powerful
- ◆ How could we compare the rates:
 1. derive an approximate distribution for the *difference of rates*
 2. derive an approximate distribution for the *odds ratio*
 3. derive an exact distribution for a 2x2 table with fixed margins (Fisher's exact test)

|← ← → →| ▼ ▸ * ✕ W - 10 -

- ◆ to fix notation:

y_c = count of GO term in cluster of interest

y_b = count of GO term in background

n_c = count of genes in cluster of interest

n_b = count of genes in background

|← ← → →| ▼ ▸ * ✕ W - 11 -

Difference of rates

- ◆ the estimated rate in the cluster is

$$\hat{p}_c = \frac{y_c}{n_c}$$

- ◆ the estimated rate in the background is

$$\hat{p}_b = \frac{y_b}{n_b}$$

- ◆ the approximation is simply

$$\frac{\hat{p}_c - \hat{p}_b}{\text{se}(\hat{p}_c - \hat{p}_b)} \sim N(0, 1)$$

- ◆ here $N(0, 1)$ is the standard normal distribution

|◀ ◀ ▶ ▶| ▼ ▸ * ✕ W - 12 -

Odds ratio

- ◆ the estimated odds ratio is

$$\hat{\psi} = \frac{\hat{p}_c / (1 - \hat{p}_c)}{\hat{p}_b / (1 - \hat{p}_b)}$$

- ◆ this is the ratio of the relative risk (of being a member of the GO term) in the cluster divided by the relative risk in background

- ◆ we approximate $\log \hat{\psi}$ by a normal distribution; the standard error is

$$\sqrt{1/(n_c - y_c) + 1/y_c + 1/(n_b - y_b) + 1/y_b}$$

|◀ ◀ ▶ ▶| ▼ ▸ * ✕ W - 13 -

The Fisher exact test

We can also represent the data in a 2x2 table:

	not GO	GO	total
cluster	a	b	$a + b$
background	c	d	$c + d$
total	$a + c$	$b + d$	$a + b + c + d$

In the terms of successes and counts, we have $b = y_c$, $d = y_b$, $a + b = n_c$ and $c + d = n_b$. The odds-ratio estimate becomes simply

$$\hat{\psi} = \frac{a \times d}{b \times c}$$

In the Fisher test, we fix all row- and column-totals. With this, we need to only fix one more number in the 2x2 table to get all entries.

|← ← → →| ▼ ▸ * ✕ W - 14 -

Say we take $y_c = b$, the number of successes in the cluster as the random variable. Its probability is given by

$$P(Y = y) = \frac{\binom{n_c}{y} \binom{n_b}{y_c + y_b - y}}{\binom{n_c + n_b}{y_c + y_b}}$$

which reads as: choose y successes out of the cluster independently of (the number of successes minus y) successes out of the background; divide by all possible ways of choosing $(y_c + y_b)$ successes out of the total number of genes. This is a *hypergeometric* distribution (draw red and black balls from an urn without replacement).

To derive a p-value in the one-sided test for over-representation, add up all these probabilities for each y that is as great as or greater than the observed y_c .

|← ← → →| ▼ ▸ * ✕ W - 15 -

How to browse for interesting GO terms ?

- ◆ until now we have only analyzed a single GO term by itself
- ◆ unfortunately, there is no good method (yet ?) that allows to analyze the whole sequence of counts
- ◆ the default procedure to date is to adjust the individual p-values with a multiple testing correction (that may or may not take the correlation of the counts into account)
- ◆ you will evaluate this in the lab section
- ◆ bottom-line: the p-value considering only one GO term is valid, p-values derived from browsing a GO graph have to be considered with caution

|← ← → →| ▼ ▸ * ✕ W - 16 -

- ◆ I would always try to achieve the analysis with as little browsing as possible; you could simply select a few GO terms *a priori* and this ensures appropriate p-values

|← ← → →| ▼ ▸ * ✕ W - 17 -

Outlook; Remarks

- ◆ the creation of ontologies will proliferate, since they allow the exchange of relationships between computers
- ◆ you can also create ontologies for yourself to provide meta-data about your own experiment to share them over the web
- ◆ the induced relationships will play a crucial role in integrated genomics
- ◆ so far we have only considered very simple uses of the GO, but one could anticipate a 'functional distance' based on GO (see for example

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=12835272

|◀ ◀ ▶ ▶| ▼ ▸ * ✕ W - 18 -

- ◆ there are many related initiatives: SNOMED (Systemized nomenclature of medicine); MIPS (provides its own vocabulary); SWISSPROT
- ◆ note also that the manual curation of GO annotations is quite slow
- ◆ text mining would speed this up, but leads to *probabilistic* gene - GO term association
- ◆ a popular tool we won't use is EASE:

<http://david.niaid.nih.gov/david/ease.htm>

- ◆ there is also a bioconductor library

|◀ ◀ ▶ ▶| ▼ ▸ * ✕ W - 19 -

Summary / Take home

- ◆ GO is a hierarchical organization of functional terms
- ◆ the statistical analyses are simple counting techniques that apply to any classification system
- ◆ everything depends on the available annotations

|← ← → →| ▼ ▸ * ✕ W - 20 -