a

Analysis of DNA Chips and Gene Networks **Spring Semester, 2009**

Lecture 14a: January 21, 2010

Lecturer: Ron Shamir *Scribe: Roye Rozov*

Gene Enrichment Analysis

14.1 Introduction

This lecture introduces the notion of enrichment analysis, where one wishes to assign biological meaning to some group of genes. Whereas in the past each gene product was studied individually to assign it functions and roles in biological processes, there now exist tools that allow this process to be automated. By centralizing and disseminating a wealth of prior knowledge about known genes, the Gene Ontology [1] database allows researchers to assign attributes to groups of genes that emerge from their experiments or analyses. The initial group of genes may be some set that was clustered together through expression analysis, bound by the same transcription factor, or chosen based on prior knowledge. To identify larger patterns within this group is to seek enrichment - to assess whether some subset of the group shows significant over-representation of some biological characteristic.

14.2 Gene Ontology (GO)

GO is a set of associations from biological phrases to specific genes that are either chosen by trained curators or generated automatically. GO is designed to rigorously encapsulate the known relationships between biological terms and and all genes that are instances of these terms. The GO associations allow biologists to make inferences about groups of genes instead of investigating each one individually. For example, the early clustering work of Eisen et al. [Figure 14.1] resulted in gene clusters that required manual annotation of each gene in order to interpret what was shared within each cluster. With GO, each gene can be automatically assigned its respective attributes.

14.2.1 Structure of GO

GO terms are organized hierarchically such that higher level terms are more general and thus are assigned to more genes, and more specific decedent terms are related to parents by either

Figure 14.1: A cluster solution manually annotated

"is a" or "part of" relationships. For example, the nucleus is part of a cell, whereas a neuron is a cell. The relationships form a directed acyclic graph (DAG), where each term can have one or more parents and zero or more children. Users may select the level of generality the terms capture and carry out their analysis accordingly [Figure 14.2, 14.3].

Terms are also separated into three categories/ontologies:

- Cellular Component - describes where in the cell a gene acts, what organelle a gene product functions in, or what functional complex an enzyme is part of
- Molecular Function - defines the function carried out by a gene product; one product may carry out many functions; a set of functions together make up a biological process
- Biological Process - some biological phenomena, or "commonly recognized series of events" affecting the state of an organism. Examples include the cell cycle, DNA replication, limb formation, etc.

14.2.2 GO & microarray analysis

GO annotations can be used to complement traditional microarray analysis. Once low level analysis is complete and a group of differentially expressed or significantly affected genes is

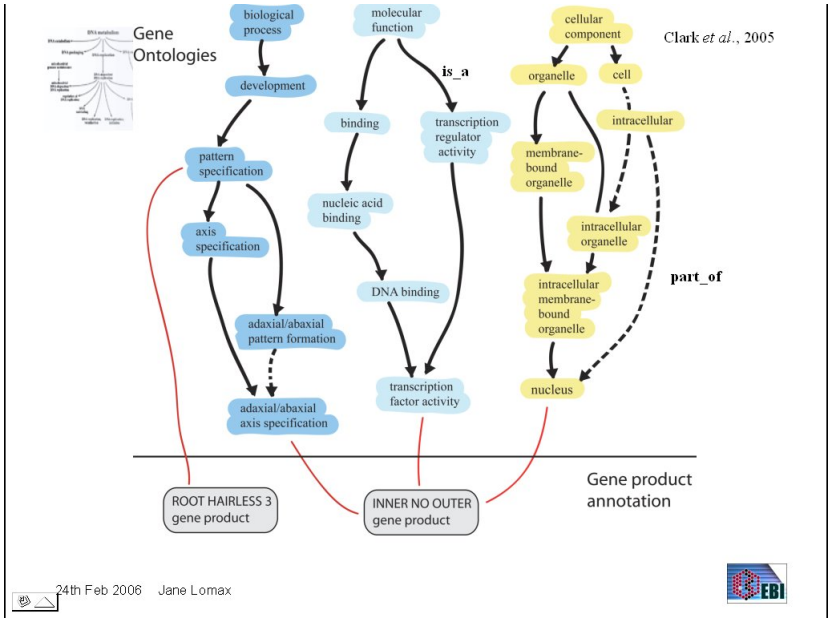

Figure 14.2: complete GO DAG [3]

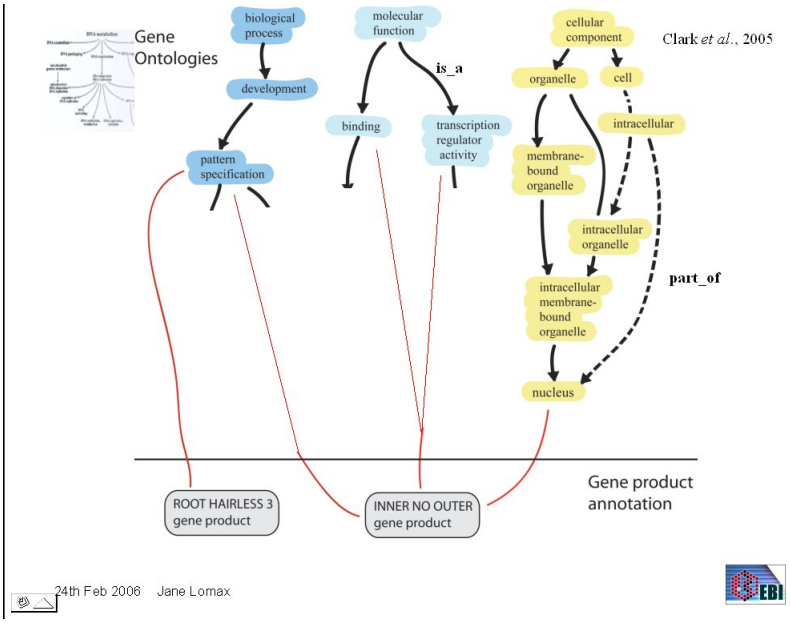

Figure 14.3: collapsed GO DAG [3]

selected, enrichment of GO attributes within the group can be assessed. Many tools exist to address this problem. Given a background gene set (i.e., all genes on the array), and a subset of interesting genes (e.g., all those that are differentially expressed), these programs identify which GO terms are most commonly associated with this subset and test the claim that this association (enrichment) is significantly different from what would be expected by chance, based on the proportions of genes out of the total having each attribute. As an example, consider the table below depicting 100 differentially expressed genes:

Process	Genes on Array	# genes expected in 100 random	# occurred out of 100
mitosis	800/1000	80	80
apoptosis	400/1000	40	40
p. ctrl.cell prol.	100/1000	10	30
glucose transp.	50/1000	5	20

It can clearly be seen that although 80 out of the 100 differentially expressed genes are associated with mitosis, the most interesting group attribute in the 100 might be glucose transport, even though it has seemingly few occurrences. This is because the size of the group that is present is much larger than that expected by chance for this process, meaning that it is over-represented. Examples of tools to determine whether such over-representation is significant in general can be found at `http://www.geneontology.org/GO.tools.microarray.shtml`, in addition to some described below.

14.3 TANGO: Tool for Analysis of GO classes

TANGO [4] tests for significance by assuming genes are sampled from a hypergeometric distribution, an approach which was introduced earlier in the course in the context of promoter analysis (see Lecture 12). For each group and each function in the hierarchy, we have:

Background: n genes, out of which m (the set A) are annotated with a certain function

Target: m' genes (labeled the set T), k of which with the function

Using these parameters, $Pr(|A \bigcap T| = k) = HG(n, m, m', k)$,and the enrichment p-value is $Pr(|A \bigcap T| \geq k) = \sum_{j \geq k} HG(n, m, m', j)$

14.3.1 Corrections

Since the significance test is performed for many groups, a multiple testing correction must be carried out in order to limit false positives. Both the Bonferroni and FDR methods

are too stringent since there exist strong dependencies between groups (since they are often members of the same hierarchy). To get around these limitations, TANGO instead calculates the empirical p value distribution. For a given cluster T_j, TANGO samples many random sets of the same size & computes their p-values vs. each of the annotation sets A_i. Next, it also permutes gene IDs to eliminate dependency between annotation sets and target sets. This correction also applies for testing multiple clusters.

14.3.2 Filtering Redundancies

Because annotation groups can overlap significantly, it is likely that highly related groups will be found significant. To avoid such overlapping results, greedy redundancy filtering is applied. To execute such filtering for a fixed target set T, we compute the approximate p-value for the enrichment of A in T, given the enrichment of another set A'.

This is given by $CondP(T, A|A') = HG(|A'|, |A \bigcap A'|, |T \bigcap A'|, |T \bigcap A \bigcap A'|) * HG(n - |A'|, |A - A'|, |T - A'|, |(T - A') \bigcap A|)$

Following this calculation, annotation sets A_i are sorted by increasing p values, and accepted only if $CondP(T, A_j|A_i) < \beta$ for all i<j. Thus, the parameter β controls how much overlap is allowed.

14.4 GSEA

Gene Set Enrichment Analysis (GSEA) is different from typical enrichment testing in that it takes into account the magnitude of expression differences between conditions for each gene. As such, it addresses the question of whether the expression of the gene set of interest shows significant differences between these conditions. It relies on ~1300 pre-defined gene sets collected from other databases (such as GO or pathway databases) and computational studies that are stored on MSigDB, the database the GSEA calls on. Running GSEA allows the user to restrict the search to specific groups of genes that have attributes that are of interest to the user. These are separated into sets C1-C5, defined as:

C1 positional; including genes on the same chromosome or cytogenic band

C2 curated; taken from pathway databases, publications, expert knowledge

C3 motif; conserved cis-regulatory motifs based on comparative studies

C4 computational; derived from past cancer studies

C5 GO, as above

14.4.1 GSEA Algorithm

GSEA tests for enrichment of some group S among N background genes, similar to TANGO above. GSEA differs in that more information is incorporated into this enrichment calculation. Some expression measure of all the genes is used explicitly in order to assess the correlation of each with a phenotype C assigned to each sample. Genes are ranked based on this correlation to calculate ES(S), as described below.

14.4.1.1 GSEA inputs:

1. Expression data set D with N genes and k samples

2. Ranking procedure to produce Gene list L. Includes a ranking metric (such as correlation) and a phenotype or profile of interest C (e.g., sick vs. healthy in the 2-category case).

3. An exponent p to control the weight of the step

4. Independently derived gene set S of $N_H$genes (e.g., some set taken from MSigDB above)

14.4.1.2 Enrichment score ES(S)

1. Rank order N genes in D to form $L = \{g_1, ..., g_N\}$according to the correlation, $r(g_j) = r_j$, of their expression profiles with C [Figure14.4, A]:.

2. Evaluate the fraction of genes in S ("hits") weighted by their correlation and the fraction of genes not in S ("misses") present up to a given position i in L:

   - $P_{hit}(S,i) = \sum_{g_j \epsilon S, j \leq i} \frac{|r_j|^p}{N_R}$,where $N_R = \sum_{g_j \epsilon S} |r_j|^p$
   - $P_{miss}(S,i) = \sum_{g_j \notin S, j \leq i} \frac{1}{N - N_H}$

   Then, ES(S) is the maximum deviation from zero,$ES(S) = max_i |P_{hit}(S,i) - P_{miss}(S,i)|$, which depends on what both the weight of the correlations and the positions of the genes in S relative to all of the genes in L[Figure14.4, B]. When p=0, ES(S) reduces to the Kolmogorov-Smirnov statistic; when p=1, the score weighs genes in S by their correlation in C normalized by the sum over all correlations in S

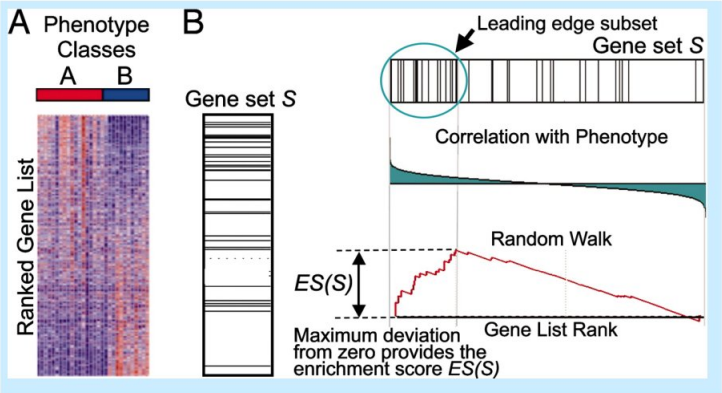

Figure 14.4: GSEA procedure. Genes in expression matrix are sorted based on correlation to phenotype classes (red and blue at the top of D, panel A). The positions of genes in S are noted with black bars to the right of D. ES(S) is calculated based on both the correlations and the positions in L (panel B). [5]

14.4.2 Estimating Significance

Significance is estimated empirically as in TANGO. The observed ES score is compared with the set of scores ES_{NULL}computed by permuting phenotypes. The original phenotype labels are assigned randomly to samples, the genes are sorted based on correlation to these labels, and ES(S) is re-computed. This permutation step is repeated 1000 times to create a histogram of the corresponding enrichment scores ES_{NULL}. Since the positive and negative sides of the distribution behave differently, the nominal P value for S is estimated from $ES_{NULL}$by using the portion of the distribution corresponding to the sign of the observed ES(S).

14.4.3 Multiple Hypothesis Testing

When many gene sets are considered, a correction is performed to account for multiple testing. Sets are normalized for size and significance based on label permutations (as above). Then, an FDR is calculated for each normalized score to estimate the probability of a given score emerging from a false positive finding. The normalized scores past a chosen FDR cutoff correspond to the sets that are reported as enriched. These corrections are carried out as follows:

1. Determine ES(S) for each gene set in the collection.

2. For each set S and each fixed permutation π (out of 1000 performed) of the phenotype labels, reorder the genes in L and determine ES(S,π). This is the same step as that needed to estimate significance.

3. Adjust for variation in the gene set size. Normalize ES(S,π) and the observed ES(S), separately rescaling the positive and negative scores by dividing by the mean of the ES(S,π) to yield the normalized scores NES(S,π) and NES(S). For example, for positive scores:

   - NES(S,π) = $\frac{ES(S,\pi)}{AVE_{ES(S,\pi).\geq 0}[ES(S,\pi)]}$ if ES(S,π) ≥ 0
   - NES(S) = $\frac{ES(S)}{AVE_{ES(S,\pi).\geq 0}[ES(S,\pi)]}$ if ES(S) ≥ 0

4. Compute the FDR. Control the ratio of false positives to the total number of gene sets attaining a fixed level of significance separately for positive (negative) NES(S) and NES(S,π):

   Create a histogram of all NES(S,π) over all S and π. Use this null distribution to compute an FDR q value, for a given NES(S)= $\alpha \geq 0$. The FDR is the ratio of the percentage of all (S,π) with NES(S,π)≥ 0, whose NES(S,π)$\geq \alpha$ divided by the percentage of observed S with NES(S)$\geq \alpha$ and similarly if NES(S)= $\alpha \leq 0$.

- $q = \frac{|\{(S,\pi)|NES(S,\pi)\geq\alpha\}|/|\{(S,\pi)|NES(S,\pi)\geq 0\}|}{|\{S|NES(S)\geq\alpha\}|/|\{S|NES(S)\geq 0\}|}$

14.4.4 Results

Many studies have applied GSEA in diverse settings. One [5]employed GSEA to reanalyze results from two earlier lung cancer studies (called here the Boston and Michigan studies). Each study obtained about 70 expression profiles that were classified either as good or poor outcomes. It was found that there was little overlap (12 genes) between the top 100 genes most correlated to the outcomes in each study, and more strikingly that there were no genes significantly associated with the outcome at a .05 significance level after correcting for multiple testing. This demonstrated the disadvantages of the single gene analysis approach. Using GSEA on the same data, 8 genes in the Boston data and 11 in the Michigan data were found significantly correlated with poor outcome (FDR $\leq$0.25). It was also found that checking the sets correlated with negative outcome from each study against the dataset of the other resulted in significant enrichment [Figure 14.5]. While this result in itself is an improvement over the gene based approach, the bigger advantages were seen in the gene sets that showed significant enrichment. About half the sets were shared between the two studies, and there were several non-identical sets that related to the same processes, such as up-regulation by telomerase, and two different insulin-related sets.

GSEA was also applied in conjunction with motif discovery software (instead of expression) and ChIP- chip measures to predict gene sets targeted by a specific transcription factor[6].

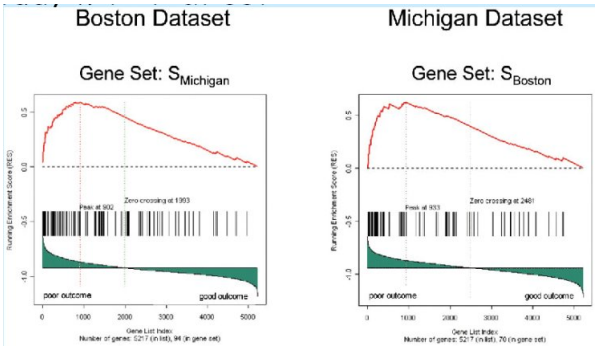

Figure 14.5: The top 100 genes correlated with negative outcome (out of genes present in both studies) from each study showing similar enrichment among all genes (present in both studies) for each individual dataset - i.e., here the Boston signature is compared against the Michigan dataset and vice-versa.

Gene set: BRCA_BRCA1_NEG (edu.mit.broad.msigdb:c2:0818)	
Standard name	BRCA_BRCA1_NEG
LSID	edu.mit.broad.msigdb:c2:0818
Brief description	Genes whose expression is consistently negatively correlated with brca1 germline status in breast cancer - higher expression is associated with BRCA1 tumors
Category	c2: Curated
Sub-category	CLIN: Clinical

Figure 14.6: BRCA1_NEG gene set

The ChIP-chip screened for targets of Nanog, a factor involved in maintaining pluripotency of embryonic stem cells. Hits of the screen were then input into a motif prediction algorithm, which produced new theoretical motifs. These motifs were then compared against promoter sequences of all human sequences and scored based on the matches to the motifs. This score was the input to GSEA, and once run GSEA produced most likely targets sets of Nanog. Among these were several other pluripotency genes, including Nanog itself, which was expected. Surprisingly, some of these genes also belonged to a larger breast cancer gene set characterized as genes upregulated in BRCA1 tumors [Figure 14.6], yielding a result with potential therapeutic implications.

14.5 Conclusions

Enrichment analysis is a means to characterize biological attributes in a given gene set. The GO dataset provides a central collection of such attributes already known and assigned to specific genes. The GO ontologies are split into cellular component, molecular function, and biological process. Using these ontologies one can give meaning to any gene, and when

they are assigned to groups of genes one can define patterns instead of labeling each gene manually. Many tools exist for assessing significance of enrichment within a group. These typically employ hypergeometric (TANGO) testing, but can also be based on a Kolmogorov-Smirnov statistic (GSEA). These tools usually require empirical estimations of p-values and multiple testing corrections.

GSEA is different in character from hypergeometric test based tools, and also offers several advantages. It requires no cutoff to be chosen a priori for gene level significance, and takes into account the effects of all genes - not only a small subgroup to be tested for enrichment. This eliminates bias of the choice, but also allows for the possibility of random results showing up as significant. As a result more corrections need to be made. GSEA also takes into account the strength of each gene's activity, as opposed to only testing for membership in specific groups. It was also shown that the tool is not limited to expression based queries, in that it has also been applied to target identification.

Bibliography

[1] The Gene Ontology Consortium: *Gene Ontology: tool for the unification of biology.* Nature Genetics Volume 25 May 2000

[2] Jane Lomax. *Gene Ontology Tutorial* `www.geneontology.org/teaching_resources/presentations/2006-02_MUGEN_expression-analysis_jlomax.ppt`

[3] Jennifer Deegan.*GO introduction for CS*, EBI (2009)

[4] Tanay, Amos. *Computational Analysis of Transcriptional Programs: Function and Evolution.* PhD Thesis, Tel Aviv University 2005 `http://acgt.cs.tau.ac.il/theses/amos_phd.pdf`

[5] Subramanian et al.: *Gene set enrichment analysis: A knowledge based approach for interpreting genome wide expression profiles.* PNAS Vol. 102 no. 43 October 2005

[6] Dan Scanfeld et al. *Motif Discovery: Algorithm and Application* `web.mit.edu/varun_ag/www/motif.ppt`