

Comparative Analysis of Gene Sets in the Gene Ontology Space under the Multiple Hypothesis Testing Framework

Sheng Zhong¹, Lu Tian¹, Cheng Li^{1,3}, Kai-Florian Storch⁴, Wing H. Wong^{1,2}

¹ Department of Biostatistics, ² Department of Statistics, Harvard University

³ Department of Biostatistical Sciences, Dana Farber Cancer Institute

⁴ Department of Neurobiology, Harvard Medical School

wwong@hsph.harvard.edu

Abstract

The Gene Ontology (GO) resource can be used as a powerful tool to uncover the properties shared among, and specific to, a list of genes produced by high-throughput functional genomics studies, such as microarray studies. In the comparative analysis of several gene lists, researchers maybe interested in knowing which GO terms are enriched in one list of genes but relatively depleted in another. Statistical tests such as Fisher's exact test or Chi-square test can be performed to search for such GO terms. However, because multiple GO terms are tested simultaneously, individual p-values from individual tests do not serve as good indicators for picking GO terms. Furthermore, these multiple tests are highly correlated, usual multiple testing procedures that work under an independence assumption are not applicable. In this paper we introduce a procedure, based on False Discovery Rate (FDR), to treat this correlated multiple testing problem. This procedure calculates a moderately conserved estimator of q-value for every GO term. We identify the GO terms with q-values that satisfy a desired level as the significant GO terms. This procedure has been implemented into the GoSurfer software. GoSurfer is a windows based graphical data mining tool. It is freely available at <http://www.gosurfer.org>

Keywords: Data Mining, Microarray, Gene Ontology, False Discovery Rate, Q-value, Visualization, GoSurfer

1. Introduction

1.1. Background

The Gene Ontology (GO) resource [1] dynamically structures biological knowledge using a controlled vocabulary consisting of GO terms. GO terms are organized in three general categories,

“biological process,” “molecular function,” and “cellular component,” and the terms within each category are linked in defined parent-child relationships that reflect current biological knowledge. On the basis of accumulated information, individual genes from all organisms are systematically associated to GO terms, and these associations continue to grow in complexity and detail as sequence databases and experimental knowledge grow.

GO provides a useful tool to look for the common traits that are shared within a list of genes, which may arise from the analysis of high-throughput genomic data, such as microarray or proteomics data. The common traits are represented by the GO terms that are associated with a large portion of the genes in the gene list. More interestingly, if compared to another gene list, some GO terms are statistically enriched in the original list but relatively depleted in the comparison list, such GO terms may describe some unique features of the original gene list. For example, Genter et al [2] compared several lists of genes with tissue specific expression patterns, and identified the GO terms that are enriched in the annotation of olfactory specific genes. Mazzolini et al [3] gathered a list of genes that are highly expressed in a pancreatic cancer cell line, and identified the GO terms that are enriched in these genes.

Let G_1, G_2, \dots, G_{n_1} and $G_{n_1+1}, G_{n_2+2}, \dots, G_{n_1+n_2}$ denote the genes in list 1 and in list 2, respectively. A particular GO term can be viewed as a function, which maps gene G into $go(G) = 0$ or 1, according to whether gene G is associated with the corresponding GO term. Thus, the null hypothesis of no association between the gene lists and a particular GO term is translated into equal distributions of binary random variables $go(G_1), go(G_2), \dots, go(G_{n_1})$ in list 1 and $go(G_{n_1+1}), go(G_{n_2+2}), \dots, go(G_{n_1+n_2})$ in list 2.

Numerous software tools [4,5,6,7,8,9] have been provided to systematically perform the gene list comparisons in the GO space. The main idea of such comparisons is to use a standard 2×2 table test, to test whether the proportion of genes that are associated with a particular GO term is the same between two lists. For example, there are $Obs_{11} = \sum_{i=1}^{n_1} go(G_i)$ and

$$Obs_{21} = \sum_{i=n_1+1}^{n_1+n_2} go(G_i) \text{ genes associated with a GO}$$

term in List 1 and in List 2, respectively. Similarly, there are $Obs_{12} = n_1 - Obs_{11}$ and $Obs_{22} = n_2 - Obs_{21}$ genes not associated with this GO term, in the two lists. Table 1 illustrates how these numbers are distributed. Under the null hypothesis, one

would expect that $\frac{Obs_{11}}{Obs_{12}} \approx \frac{Obs_{21}}{Obs_{22}}$. And then, for

example, Fisher's exact test and Pearson's chi-square test can be applied to test for such hypotheses. Any GO term with a significant p-value could be highly associated with one list of genes comparing to the other. The usual practice is to set a cutoff on the p-value and identify all the GO terms that satisfy this cutoff. The criterion of $\frac{Obs_{11}}{Obs_{21}} > \frac{Obs_{12}}{Obs_{22}}$ can be

further imposed to choose the GO terms that are specifically enriched in List 1, and vice versa.

Table 1. A 2×2 table for the distribution of associated and not associated genes in two gene lists. Our goal is exploratory in nature, that is, to find as many GO terms that are highly associated with either of lists as possible. Due to the large number of candidate GO terms, the same statistical test would be performed many times — resulting in the well known multiple testing issue in statistics.

	# of genes associated with a GO term	# of genes not associated
List 1	Obs_{11}	Obs_{12}
List 2	Obs_{21}	Obs_{22}

A multiple testing problem is inherited to the procedure described above. Suppose there are totally N GO terms that are associated with at least one gene in either gene list. Therefore totally N hypothesis tests are

performed. If the null hypothesis was true for all GO terms and all the tests were independent to each other, the N p-values would take a uniform (0,1) distribution. Picking GO terms with small p-values becomes statistically problematic because small p-values can happen with larger chances as N grows bigger. In the tests for association between GO terms and gene lists, the problem is more involved because many GO terms are mutually dependent. The dependency comes from two sources: the hierarchical structure of the GO and the usage of multiple GO terms in the annotation of one gene. For example, Cell Proliferation is a parent GO term of Cell Cycle, therefore all the genes that are annotated with the term Cell Cycle must be annotated with the term Cell Proliferation. For another example, human HoxA7 gene has been annotated with 4 GO terms, Development, Nucleus, DNA Dependent Regulation of Transcription, and Transcription Factor Activity. Therefore if add the gene HoxA7 to List 1, the Obs_{11} statistics for all of the 4 GO terms will be simultaneously added by 1.

This multiple testing problem has been recognized by several research groups. The GoMiner group [6] and the GOTree Machine group [7] both raised this problem but decided it is beyond the scopes of their papers. The FuncAssociate group [4] implemented a method [26] to control for FWER (see 1.2), but this method is too insensitive in detecting interesting GO terms (see 1.2 and Dudoit et al [28]). The FatiGO group [9] implemented several previously devised multiple testing methods, but they offered little discussion on the applicability of those methods to the current problem.

We propose to use the False Discovery Rate (FDR) to detect GO terms. We provide a justified procedure to calculate a moderately conserved estimate of q-value [10] (see 1.2) for every GO term, taking into account the dependency among GO terms. For a desired cutoff on q-value, we output the GO terms with q-values smaller than the cutoff. In the GoSurfer software, users can highlight the GO nodes that satisfy the cutoff on the GO tree. There are plenty of other graphical and interactive features in GoSurfer to help users to investigate the significant GO terms.

1.2. Review of multiple testing procedures

For a single statistical test, there are two types of statistical errors linked to it: type I error (false positive) and type II error (false negative). When conducting a statistical test to test the null hypothesis, H_0 , the typical approach is to determine a rejection region R_α such that $\Pr(T \in R_\alpha | H_0) \leq \alpha$, for a selected test statistic, T , and pre-specified type I error, α , first,

and then draw a conclusion based on the observed test statistic $T = t$. For an observed statistic $T = t$, the p-value is defined as $\inf_{t \in R_\alpha} \Pr(T \in R_\alpha | H_0)$ for nested

rejection regions. P-value is viewed as a measure of significance for the observed test statistic.

However, in the presence of multiple testing, the situation becomes much more complicated. It is well known that traditional p-value cutoffs of 0.01 or 0.05 should be made stricter to avoid an abundance of false positive results arising from the fact that a large number of statistical tests are performed at the same time. Various global measures of the risk of false positives in multiple testing have been developed. The commonly used measures can be summarized through quantities in table 2.

Table 2. Quantities in multiple testing, Benjamini and Hochberg [31]

	Not rejected	Rejected	
Null true	U	V	m_0
Alternative true	T	S	m_1
	W	R	m

- Family-wise error rate (FWER). The FWER is the probability of at least one false positive in the m tests, i.e., $P(V \geq 1)$.
- False discovery rate (FDR). The FDR is the expected proportion of false positive among rejected hypotheses, i.e., $E(\frac{V}{R} | R > 0)P(R > 0)$.
- Positive false discovery rate (pFDR). The pFDR is the expected proportion of false positive among rejected hypotheses given at least one hypothesis is rejected, i.e., $E(\frac{V}{R} | R > 0)$.

In general, FWER (FDR) is the most (least) conservative global measure of type I error among them [28] Westfall and Young [26] proposed resampling-based p-value adjustment procedures which offered strong control of FWER with dependent test statistics and improved the classical Bonferroni adjustment. FDR was first proposed by Benjamini and Hochberg [31] and has been extended to account for dependent structure of the test statistics by Benjamini

and Yakutieli [32] and Storey [25]. Many argued that FDR and pFDR are more appropriate measures to use, when the goal of the analysis was to reliably identify certain associations [11]. Various resampling-based test procedures controlling FDR were extensively studied by Van der Laan & Bryan [12], Tusher et al [24], Storey [18], Reiner et al [13], Efron et al [14] and Ge et al [15]. Efron et al [14] established an interesting relationship between FDR and Empirical Bayesian method. Recently, Storey [16] introduced the concept of q-value which is a generalization of p-value. Similar to p-value, for a given rejection region R_α and observed test statistic $T = t$, the q-value is defined as $\inf_{t \in R_\alpha} pFDR(R_\alpha)$. Q-value provides the evidence of significance for each individual test and it automatically accounts for multiple testing by means of pFDR.

Most of the aforementioned methods were motivated and developed for analyzing microarray data. Storey & Tibshirani [11] listed some other genomic studies where the newly developed methods should be used. The data mining in the GO space is another interesting field for such applications. Regardless of many similarities with other genomic studies, there are many unique features to the association analysis in the GO space.

2. Method

Assuming there are N GO terms in consideration: go_1, go_2, \dots, go_N . The data can be summarized in a $N \times (n_1 + n_2)$ matrix with the ij -th entry being $go_i(G_j)$, $i = 1, \dots, N$; $j = 1, \dots, n_1 + n_2$ (Figure 1, upper table). $go_i(G_j)$ takes value 1 if GO term go_i is used in the annotation of gene G_j , and $go_i(G_j)$ takes value 0 if go_i is not used to annotate G_j . Unlike in microarray data, where the dependence across different genes is expected to be weak and restricted to small groups, the dependence among rows in the GO data matrix is strong and may be complicated (see Discussion).

	List 1			List 2		
	G_1	...	G_{n1}	G_{n1+1}	...	G_{n1+n2}
GO_1	$go_1(G_1)$...	$go_1(G_{n1})$	$go_1(G_{n1+1})$...	$go_1(G_{n1+n2})$
GO_2	$go_2(G_1)$...	$go_2(G_{n1})$	$go_2(G_{n1+1})$...	$go_2(G_{n1+n2})$
...						

For every GO term:

	# of genes associated with GO_1	# of genes not associated with GO_1
List 1	$Obs_{11} = \sum_{j=1}^{n1} go(G_j)$	$Obs_{12} = n1 - Obs_{11}$
List 2	$Obs_{21} = \sum_{j=n1+1}^{n1+n2} go(G_j)$	$Obs_{22} = n2 - Obs_{21}$

↓
 X_1^2

Figure 1: Linking genes with GO terms. The $go_i(G_j)$ functions in the upper matrix takes value 1 if GO term i is used to annotate gene G_j , and takes value 0 if not. The upper matrix is referred as the GO data matrix. The lower table shows how to generate the cell counts from the GO data matrix for the 2×2 table of every GO term. Gene-list and GO term association test can be performed based on the cell counts in the lower table.

Various statistical tests can be used for testing association between gene lists and GO terms. In the following, we take a modified Pearson's chi-square test statistic as an example to describe our procedure, which is implemented in **GoSurfer v1.1**. For every individual GO, we calculate a signed Chi-square statistic, X^2 . The statistic is defined as

$$sign(\frac{Obs_{11}}{Obs_{21}} - \frac{Obs_{12}}{Obs_{22}}) \sum_{i=1}^2 \sum_{j=1}^2 \frac{(Obs_{ij} - Exp_{ij})^2}{Exp_{ij}} \quad (1),$$

where Obs_{ij} is defined in Table 1, and Exp_{ij} is the expected number in the ij^{th} cell, that is,

$$Exp_{ij} = \frac{(\sum_{i=1}^2 Obs_{ij}) \times (\sum_{j=1}^2 Obs_{ij})}{\sum_{j=1}^2 Obs_{ij}} \quad (2).$$

The test statistic X^2 provides not only the strength but also the direction of the potential association.

Due to the dependence among rows of the GO data matrix, the test statistics calculated from different rows are dependent. Similar to the SAM procedure [17, 24] (see Discussion), the null distribution can most easily be calculated by permuting the list labels, or one can use bootstrap. Storey & Tibshirani [17] and Westfall & Young [26] provided insights in comparing the two methods. The permutation method has strength in that if the null hypothesis is true, then we are able to calculate the null distribution exactly.

Denote the signed Chi-square statistic for GO term i as X_i^2 . We rank the X_i^2 s from the smallest to the largest. Denote $X_{(j)}^2$ as the j^{th} order statistic of X_i^2 s.

To generate the null distribution of $X_{(j)}^2$, $j = 1, \dots, K$, we permute the list labels in the data matrix, i.e. we randomly reassign the genes into two lists while fixing the total number of genes in each list the same as the original gene list. All the gene to GO mapping functions, $go_i(G_j)$ s, are untouched in the permutation. Following the same procedure as the $X_{(j)}^2$ s are

calculated, in the b^{th} permutation we calculate new $X_{(j)}^2$ s, which are denoted by $X_{(j*)}^{2b}$ s. After a large number of permutations, say B times, we will have an approximation of the distribution of $X_{(j*)}^2$ based on B realizations, $\{X_{(j*)}^{21}, X_{(j*)}^{22}, \dots, X_{(j*)}^{2B}\}$, for every j . Regard this distribution as the background distribution of $X_{(j)}^2$, we can ask the question of how likely the actual $X_{(j)}^2$ can be observed. We compute the following quantity for any GO term k :

$$\hat{q}(k) = \frac{\frac{1}{B} \sum_{b=1}^B \sum_{j=1}^N I\{|X_{(j*)}^{2b}| \geq |X_k^2|\}}{\max\{1, \sum_{j=1}^N I\{|X_{(j)}^2| \geq |X_k^2|\}\}} \hat{\pi}_0(c_0) \quad (3),$$

where $I\{\cdot\}$ is an indicator function and $\hat{\pi}_0(c_0)$ is an estimator for the proportion of true null hypotheses, which is equivalent to Storey's $\hat{\pi}_0(\lambda)$ [18]. We will show in the Appendix section that (3) is a moderately conserved estimate of the q-value for GO term k . In the end, **GoSurfer v1.1** labels each GO term with its corresponding estimated q-value. Users can then identify a list of GOs at a desired q-value level. Compared with SAM, our q-value estimator is adjusted by an additional factor $\hat{\pi}_0(c_0)$, which may improve the conservativeness of FDR estimation in the original

SAM [24] considerably when $\hat{\pi}_0(c_0)$ is not close to 1 (see Discussion for more differences and SAM's improvement).

3. Software

3.1. GoSurfer v1.1

GoSurfer [19] is a windows based graphical interactive data mining tool. We briefly summarize its previous functionalities here. GoSurfer takes one or two list(s) of gene ids as input file(s). The gene ids can be Locuslink ID, Unigene ID, or Affymetrix probe set ID. GoSurfer finds all the GO terms that are associated

with any genes in the input gene list(s), and visualize these GO terms as three hierarchical trees. Each tree corresponds to one of the three general GO categories "biological process," "molecular function," and "cellular component". Users can manipulate the graphic output in various ways. For example, users can trim off the GO terms that are associated with only a small number of the input genes. The Chi-Square test described in the introduction section can also be performed to search for the GO terms that are enriched in the annotation of one input list of genes. Users can click on the GO graph to find the input genes that are associated with the clicked GO term. Figure 2 shows a screen shot of GoSurfer, when it only takes one input list of genes.

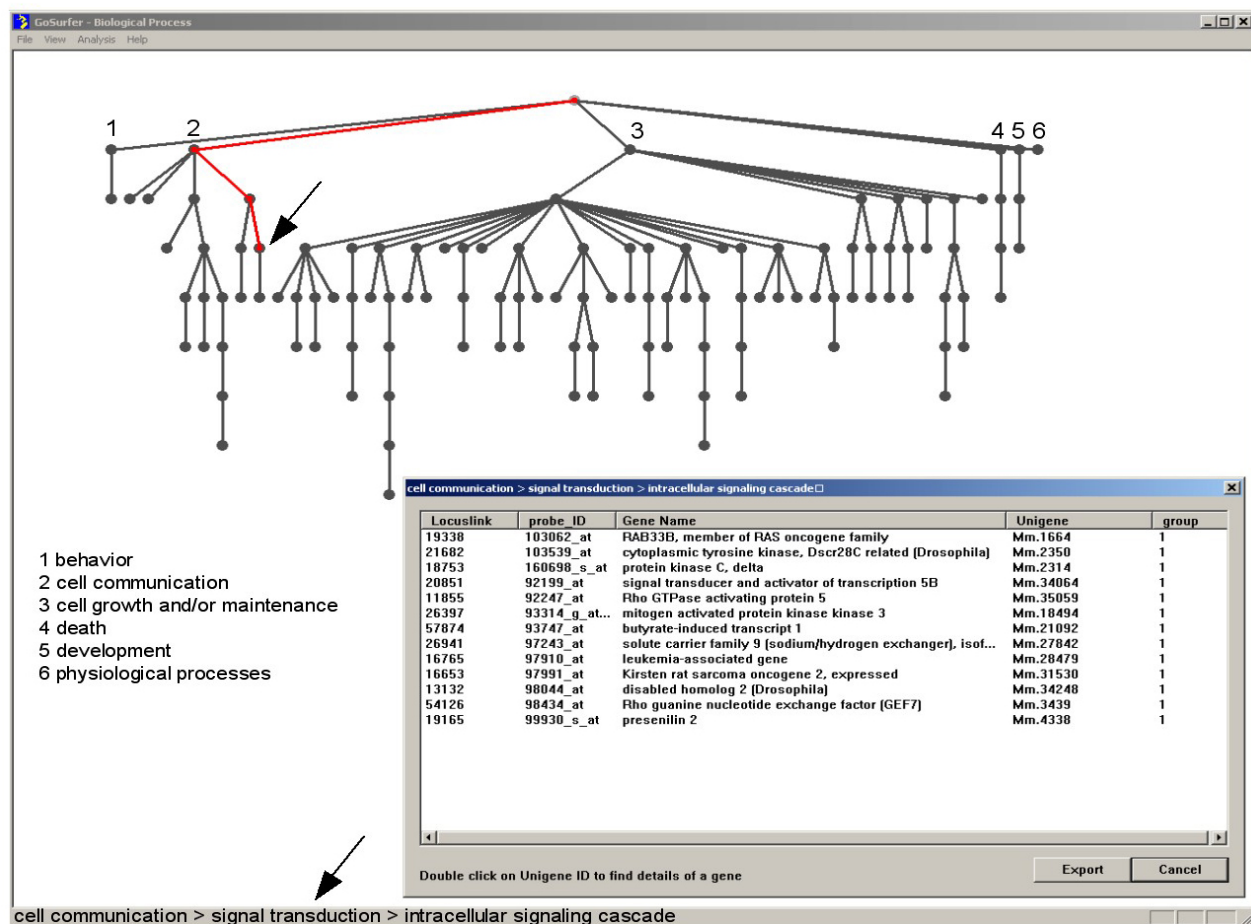


Figure 2: GoSurfer screen shot displaying a GO tree of the biological process category. Each node represents an individual GO term and all GO terms at display are associated with at least one out of 575 liver specific genes (data not shown). For clarity, only terms (nodes) that are associated with at least 4 genes from the data set are shown. The path to the GO term "intracellular signaling cascade" is highlighted in red and corresponds to the terms displayed in the status line (arrows). Inset: pop-up window displaying all genes in the data set that are associated with the GO term "intracellular signaling cascade." Selected nodes are marked with numbers, and the corresponding GO terms are listed underneath the tree structure.

We have implemented our method for controlling for multiple testing (see Method section) into GoSurfer v1.1. Users can use GoSurfer both to calculate the estimated q-value for every GO term and to highlight the GO terms that satisfy a user defined q-value threshold, on the visualized GO trees. To perform such analyses, users first need to input two lists of genes (Affymetrix probe sets) for comparison. Users can click on GoSurfer's menu "Analysis -> FDR" (Figure 3). After the FDR menu being clicked, a popup window will show up, where users can designate the desired location of the output file. This output file records the following information for every GO term: the observed test statistic, the mean of the test statistics under permutation, and the estimated q-value. Users can use menu "File -> Export -> GO info" to obtain more detailed information for every GO term, such as its relative location(s) in the GO tree graph, the number of genes attached to this GO term in every input list, its p-value from a chi-square test, its q-value, etc. This exporting process can take several minutes. A progress indicator will show up at the lower left corner of the software window to help users to monitor the exporting process.



Figure 3: Activating the FDR menu in GoSurfer.

To highlight the GO terms that satisfy a q-value threshold, users can first draw any of the three GO trees using the submenus of the "View" menu. After the GO tree is drawn, users can click on the "Analysis -> Highlight" menu. In the popup window, users can choose to use the multiple testing procedure, and set a threshold on the q-value. The satisfied GO terms will be highlighted accordingly (Figure 4). Users can then use other interactive features to explore these GO terms or manipulate the graph.

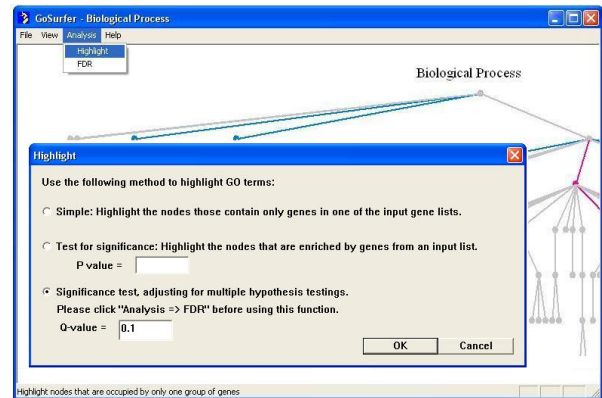
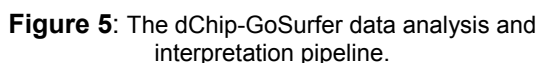


Figure 4: Using the multiple testing procedure to identify interesting GO terms in GoSurfer.

3.2. dChip-GoSurfer interaction

The GoSurfer software can be used interactively with the dChip [20] software, which is specialized in microarray analysis. This interactive feature makes all the described analyses in the GO space more directly accessible to microarray analysis. For example, suppose a researcher compared the microarray data of normal and cancer samples, and found a list of induced genes and a list of suppressed genes in the cancer samples with the dChip software. The researcher can directly call out GoSurfer from dChip. GoSurfer will automatically take the two lists of genes to map onto the GO space. The researcher can then perform subsequent analysis with GoSurfer. Often of the case is that the researcher is not sure about his/her methods in identifying genes, so that he/she needs to resort to GO to make more educated judgments. In such a case the direct interaction between dChip and GoSurfer would greatly facilitate the analytical and discovery process. Some researchers⁶ commented that in the revolutionizing era of biology, three challenging steps have to be taken to make discoveries: experiment, statistical analysis and biological interpretation. The dChip-GoSurfer software suit may greatly facilitate the last two steps. Figure 5 shows the dChip-GoSurfer pipeline of data analysis and visualization.



We used GoSurfer to find the GO terms that are significantly associated with genes showing altered expression in prostate cancer in comparison with normal prostate. We analyzed the data from a microarray study of gene expression in 52 prostate tumor specimens and 50 normal prostates [21].

By a q-value cutoff of 0.1, we have identified 25 GO terms that are enriched in the up-regulated genes in the cancer cells, and 40 GO terms that are enriched in the down-regulated genes in cancer cells (see supplementary Table 1 and Table 2 online). We displayed the GO trees and color-coded the significant GO terms. The example in Figure 7 highlights biological processes that are significantly associated with genes induced (magenta) or repressed (blue) in prostate cancer compared to normal prostate. The preference of blue in the tree display suggests that repression of gene expression could be a major factor contributing to tumor phenotype. Interestingly, a different microarray study on prostate cancer reported that metastatic tumors were distinguished from nonmetastatic by a larger number of down-regulated genes [22]. Biological processes that were significantly associated with genes up-regulated in prostate cancer included “protein metabolism” and “protein biosynthesis” (Figure 7, nodes 6, 7). A few ribosomal complex related GO terms are also heavily associated with cancer induced genes (see Supplementary Table 1 online), perhaps reflecting aberrant proliferative control or energy metabolism. Biological processes that were associated with genes down-regulated in prostate cancer included “regulation of cell proliferation” (node 8), “organogenesis” (node 9), “cell mobility” and “muscle contraction” (nodes 10,11), and a pathway representing surface cell surface receptor signal transduction (nodes 3,12,13). Down-regulation of genes involved in the regulation of cell proliferation points to the expected defect of mitotic control in cancer cells. The down-regulated genes in this node (Supplementary Table 3 online) include several well-

known tumor suppressors, but also many positive regulators of cell proliferation, implying that the regulatory circuits for mitotic control are generally perturbed in prostate tumors. Down-regulation of genes associated with cell mobility and muscle contraction (Supplementary Table 4 online) perhaps reflects the de-differentiated phenotype of tumor cells, since the

normal prostate gland is a contractile organ containing smooth muscle cells. Suppression of genes associated with cell communication (node 12) and signal transduction (node 13) in prostate cancer is in agreement with the reported insensitivity of cancer cells to exogenous anti-growth signals [23].

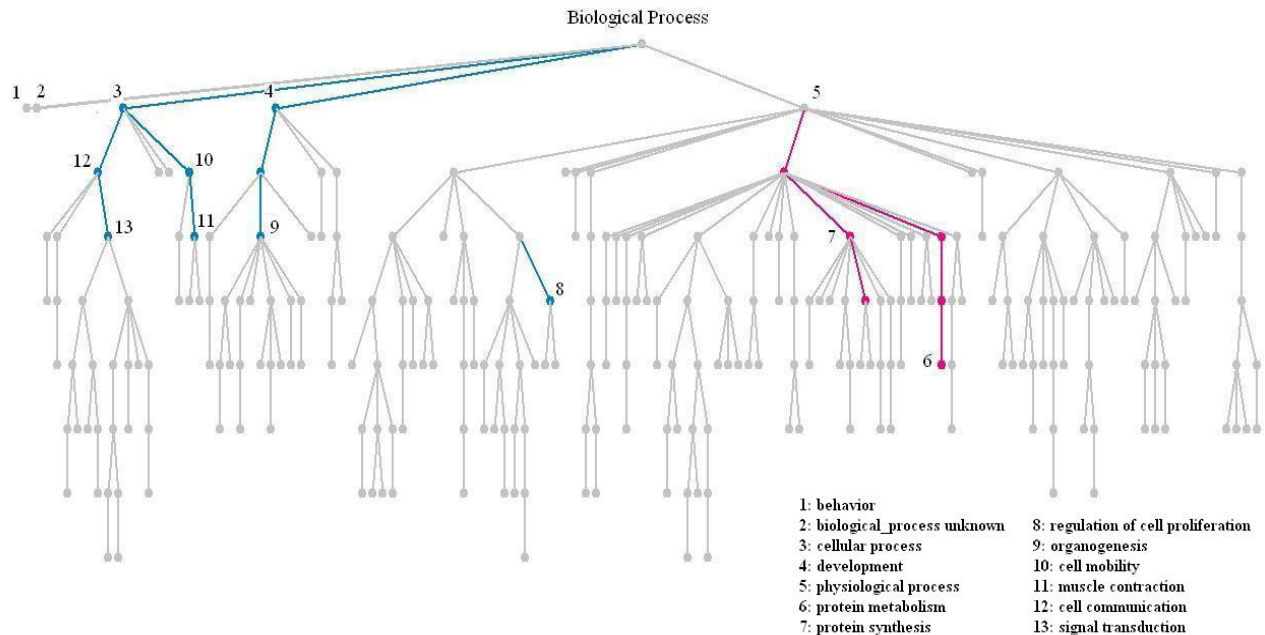


Figure 7: GoSurfer comparison of biological processes significantly (q -value < 0.1) associated with genes up-regulated (magenta) or down-regulated (blue) in prostate cancers compared with normal prostate. For clarity, only terms that are associated with at least 2 genes of the induced or repressed gene sets are shown. Selected nodes are marked with numbers, and the corresponding GO terms are listed underneath the tree structure.

5. Discussion

We have posed the multiple testing problem in the more common than ever practice of searching for enriched GO terms in gene lists. We explored the suitability of using FDR strategies to deal with this problem and proposed a moderately conserved estimator to the q -value for any GO term. We have implemented this method into GoSurfer software, allowing researchers to list and display the GO terms that satisfy any q -value cutoff. This functionality, together with plenty of other graphical and interactive features, has made GoSurfer a useful tool in functional analysis for large gene sets. Especially the dChip-GoSurfer interaction makes data mining in the GO space from microarray data more convenient and efficient.

Although the multiple testing issue in the GO setting has similar flavor to the well explored

microarray setting, the natures of studies, and subsequently, the statistical procedures, make the two settings different. Both the gene to GO mapping and the strong correlation among GO terms have made the multiple testing issue in this particular setting an interesting statistical exercise. We would like to provide a heuristic explanation of the way we handled the correlation. In the permutation, neither the $go_i(G_j)$ transformation nor the GO structure is changed. Suppose under the null, the two lists of genes have no difference, i.e. they come from the same underlining distribution. Here the genes, G_j s, are regarded as random variables. All the gene to GO mappings, $go_i(G_j)$ s, are transformations of G_j s. The joint distribution of $go_i(G_j)$ s is determined by the underlining distribution of all G_j s. As long as the underlining distribution of genes is fixed, the joint distribution of $go_i(G_j)$ s is fixed. Any permutation on the list labels would provide a random sample from underlining distribution of genes, and the consequently

derived new $go_i(G_j)$ s would jointly follow the same pre-fixed distribution. So that every gene permutation would give a random sample of $\{go_i(G_j)\}$ from the same distribution of the actually observed $\{go_i(G_j)\}$. Therefore under the null, the permutation is capable of the generating new $\{go_i(G_j)\}$ with the same correlation structure as the observed $\{go_i(G_j)\}$.

The SAM software [24] used permutation method to calculate FDR estimate in a two-sample comparison setting for microarray data. Our method is different from SAM's in three major ways. 1. SAM asks for a user defined rejection region and estimate the FDR for the designated rejection region, while GoSurfer allows users to directly input a desired q-value cutoff, and automatically finds the corresponding rejection region. 2. The SAM's estimate of FDR does not include the $\hat{\pi}_0(c_0)$ factor, and it appears to be overly conservative, although later a new approach by Storey [25] was added in. 3. An ideological difference: the SAM considers the different rows in the data matrix as different data, while we consider the different rows in the data matrix as different transformations of the same underlining data. If we regarded the different rows in the data matrix as different data, then the correlation across different rows would be too complex to sort out. There are quite a few implicit statistical approximations in the original SAM paper [24], which were later clarified [17]. We have outlined the most important approximations of our procedure (in Appendix). It may help interested readers to have deeper understanding of our procedure, as well as SAM's.

Seldom has any software besides GoSurfer seriously treated the multiple testing issue in the GO space. To our knowledge, only the FuncAssociate [4] team and the FatiGO [9] team have made such efforts. FuncAssociate used Westfall and Young's method [26,27] to control for FWER, which is very conservative [28] and insensitive in detecting interesting GO terms. The FatiGO team implemented 4 methods: Westfall and Young [29,30] (W&Y), Benjamini and Hochberg [31] (B&H), Benjamini and Yekutieli [32] (B&Y) and their own permutation method. The FatiGO authors gave little information on the implementation details and the applicability of these methods to the problem at hand. Among these methods, B&Y seems to be most desirable for two reasons. It works under a very relaxed dependency requirement, and it is not too conservative. The permutation test devised by the FatiGO group actually does not control for multiple testing.

6. Online supplementary materials

The supplementary tables and color figures in this article are available at: http://www.gosurfer.org/Supplementary_materials.htm

7. Acknowledgement

We thank Yofre Cabeza-Arvelaiz for evaluating the software and helpful suggestions. This work is supported by NIH grant 1R01HG02341.

Appendix

To see that (3) is a conserved estimator of the q-value of GO term k , let $R(k) = \{ |X_k^2| \geq X_k^{2*} \}$ denote the rejection region for all the N tests, we have the following approximations:

$$\frac{1}{BN} \sum_{b=1}^B \sum_{j=1}^N I\{|X_{(j^*)}^{2b}| \geq X_k^{2*}\} \approx \Pr(X^2 \in R(k) | H_0^c),$$

$$\frac{1}{N} \max \{1, \sum_{j=1}^N I\{|X_{(j)}^2| \geq X_k^{2*}\}\} \approx \Pr(X^2 \in R(k)),$$

and

$$\begin{aligned} \hat{\pi}_0(c_0) &= \frac{\frac{1}{N} \sum_{j=1}^N I\{|X_{(j)}^2| < c_0\}}{\frac{1}{NB} \sum_{b=1}^B \sum_{j=1}^N I\{|X_{(j^*)}^{2b}| < c_0\}} \approx \frac{\Pr(|X^2| < c_0)}{\Pr(|X^2| < c_0 | H_0)} \\ &\approx \frac{\Pr(|X^2| < c_0 | H_0) \Pr(H_0) + \Pr(|X^2| < c_0 | H_1) \Pr(H_1)}{\Pr(|X^2| < c_0 | H_0)} \approx \Pr(H_0) \end{aligned}$$

where H_0^c denotes the complete null hypothesis: null hypothesis holds true for all the m hypotheses, H_1 is the alternative and c_0 is an appropriately chosen cutoff value such that $\Pr(|X^2| < c_0 | H_1)$ is negligible compared with $\Pr(|X^2| < c_0 | H_0)$ for reasonable alternative H_1 . When the proportion of true null hypothesis, $\Pr(H_0)$, is close to 1, or the "weak dependence"¹⁸ structures of the test statistics under H_0^c and H_0 are similar, it is expected that $\Pr(X^2 \in R(k) | H_0^c) \approx \Pr(X^2 \in R(k) | H_0)$. The two conditions are satisfied in practice by our limited experience. Therefore the (3) approximates

$$\frac{\Pr(X^2 \in R(k) | H_0) \Pr(H_0)}{\Pr(X^2 \in R(k))} = FDR \approx pFDR.$$

Because the rejection region $R(k) = \{ |X^2| \geq |X_k^2| \}$ is so defined that GO term k possesses the smallest $|X^2|$ in the rejection region, the quantity in (3) is an estimate of q-value. It is easy to see that if $\hat{\pi}_0(c_0)$ is a conservative estimate of π_0 (the proportion of true null hypotheses), then (3) is a conserved estimate of q-value. We use a reasonably small c_0 to guarantee the conservativeness.

Reference

- [1] Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-9 (2000).
- [2] Genter, M.B. et al. Microarray-based discovery of highly expressed olfactory mucosal genes: potential roles in the various functions of the olfactory system. *Physiol Genomics* **16**, 67-81 (2003).
- [3] Mazzolini, G. et al. Pancreatic cancer escape variants that evade immunogene therapy through loss of sensitivity to IFNgamma-induced apoptosis. *Gene Ther* **10**, 1067-78 (2003).
- [4] Berriz, G.F., King, O.D., Bryant, B., Sander, C. & Roth, F.P. Characterizing gene sets with FuncAssociate. *Bioinformatics* **19**, 2502-4 (2003).
- [5] Castillo-Davis, C.I. & Hartl, D.L. GeneMerge--post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics* **19**, 891-2 (2003).
- [6] Zeeberg, B.R. et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol* **4**, R28 (2003).
- [7] Zhang, B., Schmoyer, D., Kirov, S. & Snoddy, J. GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics* **5**, 16 (2004).
- [8] Doniger, S.W. et al. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol* **4**, R7 (2003).
- [9] Al-Shahrour, F., Diaz-Uriarte, R. & Dopazo, J. Fatigo: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* **20**, 578-80 (2004).
- [10] Storey J.D. The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics* **31**: 2013-2035 (2003)
- [11] Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**, 9440-5 (2003).
- [12] Van der Laan, M. & Bryan, J. Gene expression analysis with the parametric bootstrap, *Biostatistics*, **2**, 445-461 (2001)
- [13] Reiner, A., Yekutieli, D. & Benjamini, Y. Identifying differentially expressed genes using false discovery rate controlling procedures, *Bioinformatics*, **19**, 368-375 (2003)
- [14] Efron, B., Tibshirani, R., Storey, J.D., & Tusher, V. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, **96**, 1151-1160 (2001)
- [15] Ge, Y., Dudoit, S. & Speed, T. P. Resampling-based multiple testing for microarray data analysis, Sociedad Espanola de Estadística e Investigación Operativa. *Test*, **12**, 1-77 (2003)
- [16] Storey, J. D. A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, **64**, 479-498 (2002)
- [17] Storey J.D., Tibshirani R. SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. *The Analysis of Gene Expression Data: Methods and Software*. Springer, New York (2003)
- [18] Storey, J.D., Taylor, J.E. & Siegmund, D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society B* **66**, 187-198 (2004).
- [19] Zhong S, Storch F, Lipan O, Kao MJ, Weitz C, Wong WH. GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in Gene Ontology space. *Applied Bioinformatics* **3**(3) In press.
- [20] Li, C. & Wong, W.H. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A* **98**, 31-6 (2001).
- [21] Singh, D. et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1**, 203-9 (2002).
- [22] Dhanasekaran, S.M. et al. Delineation of prognostic biomarkers in prostate cancer. *Nature* **412**, 822-6 (2001).
- [23] Hanahan, D. & Weinberg, R.A. The hallmarks of cancer. *Cell* **100**, 57-70 (2000).
- [24] Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**, 5116-21 (2001).
- [25] Storey J.D. A direct approach to false discovery rate. *Journal of Royal Statistical Society B* **64**: 479-498
- [26] Westfall, P.H., Young, S.S. Rsampling-based multiple testing: examples and methods for p-value adjustment. *John Wiley & Sons, Inc., New York* (1993)
- [27] http://llama.med.harvard.edu/FuncAssociate_Methods.html
- [28] Dudoit, S., Shaffer, J.P., Boldrick, J.C. Multiple Hypothesis Testing in Microarray Experiments. *Statistical Science* **18**: 71-103 (2003)
- [29] Westfall, P.H., Young, S.S. Rsampling-based multiple testing: examples and methods for p-value adjustment. *John Wiley & Sons, Inc., New York* (1993)
- [30] http://llama.med.harvard.edu/FuncAssociate_Methods.html
- [31] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, **57**, 289-300 (1995).

[32] Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple hypothesis testing under dependency. 1165-1188 (2001).