Published in final edited form as:
Biometrics. 2014 March ; 70(1): 84–94. doi:10.1111/biom.12122.

A Bayesian Extension of the Hypergeometric Test for Functional Enrichment Analysis

Jing Cao[1,*] and **Song Zhang**[2,**]

[1]Department of Statistical Science, Southern Methodist University, Dallas, TX 75275, U.S.A

[2]Department of Clinical Sciences, UT Southwestern Medical Center, Dallas, TX 75390, U.S.A

Summary

Functional enrichment analysis is conducted on high-throughput data to provide functional interpretation for a list of genes or proteins that share a common property, such as being differentially expressed (DE). The hypergeometric *P*-value has been widely used to investigate whether genes from pre-defined functional terms, e.g., Gene Ontology (GO), are enriched in the DE genes. The hypergeometric *P*-value has three limitations: 1) computed independently for each term, thus neglecting biological dependence; 2) subject to a size constraint that leads to the tendency of selecting less-specific terms; 3) repeated use of information due to overlapping annotations by the true-path rule. We propose a Bayesian approach based on the non-central hypergeometric model. The GO dependence structure is incorporated through a prior on non-centrality parameters. The likelihood function does not include overlapping information. The inference about enrichment is based on posterior probabilities that do not have a size constraint. This method can detect moderate but consistent enrichment signals and identify sets of closely-related and biologically-meaningful functional terms rather than isolated terms. We also describe the basic ideas of assumption and implementation of different methods to provide some theoretical insights, which are demonstrated via a simulation study. A real application is presented.

Keywords

Functional enrichment analysis; Modular enrichment analysis; Hypergeometric *P*- value; Non-central hypergeometric distribution; Gene ontology

1 Introduction

In traditional high-throughput data analysis such as microarray analysis the focus has been on identifying differentially expressed (DE) genes. Researchers have found that the list of identified DE genes are usually difficult to reproduce and bear little unifying biological theme (Subramanian *et al*., 2005). Many methods have been proposed to incorporate biological knowledge accumulated in public databases to detect enriched and pertinent biology (Khatri *et al*., 2002; Hosack *et al*., 2003; Zhang *et al*., 2004; Subramanian *et al*., 2005; Kathri and Draghici, 2005; Efron and Tibshirani, 2007; Newton *et al*., 2007). Many annotation resources have been utilized, including Gene Ontology (GO), protein-protein interactions (e.g., KEGG, Kanehisa and Goto, 2000), protein functional domains, disease associations, bio-pathways, sequence features, homology, gene functional summaries, gene

[*]jcao@smu.edu. [**]song.zhang@utsouthwestern.edu.

7 Supplementary Materials

Web Appendices referenced in Sections 2, 4 and 5.2, as well as the C program and its instruction, are available with this paper at the Biometrics website on Wiley Online Library.

tissue expression, literature, transcription factors, miRNAs, structural motifs, drug targets, etc.

The GO database (Gene Ontology consortium, 2000) is one of the most popular gene description databases used in enrichment analyses, with GO terms (each annotating a group of genes) as the building block. Gene Ontology consists of biological processes, cellular components, and molecular functions. In this paper we generically refer to GO terms as functional terms with the understanding that they are also applicable to the other two components. GO terms are organized in a directed acyclic graph (DAG) of parent-child relationship. A child represents a more specific biological classification. The GO DAG is different from a traditional classification tree in that a GO term is allowed to have more than one parent, a feature known as multiple inheritance. The GO database also follows the true-path rule: genes annotated by a child node are automatically annotated by its parent nodes, and subsequently, by all the ancestral nodes.

Many existing procedures to detect enrichment are based on the hypergeometric test (or its variants including the binomial and Fisher's exact tests). We briefly review the classical hypergeometric test. In a high-throughput experiment, let g be the number of genes annotated to a certain GO term, and let f and d be the total numbers of genes evaluated and DE genes detected, respectively. The number of DE genes annotated to this GO term, denoted by n, indicates the representation of the GO term in the list of DE genes. The null hypothesis is that the functional term is irrelevant to the experiment, which means that a gene being annotated by the GO term and this gene being DE are independent events. Given (g, f, d), we can model n by a hypergeometric distribution under the null hypothesis, and the P-value measuring the significance of enrichment is the tail probability of observing n or more DE genes annotated to the GO term,

$$P-\text{value}=\sum_{k=n}^{\min(g,d)}\frac{\binom{g}{k}\binom{f-g}{d-k}}{\binom{f}{d}}, \quad (1)$$

where $\binom{m}{n}=\frac{m!}{n!(m-n)!}$ is the binomial coefficient.

Due to its computational simplicity and straightforward interpretation, the hypergeometric P-value and its variants have become popular in functional enrichment analysis. However, researchers have noticed some drawbacks of the hypergeometric P-value. For example, the hierarchical structure of the GO DAG can be informative but it is ignored in the hypergeometric test. Suppose we observe an enrichment signal at a particular GO term (a biological function). Examining this term by itself, it might be difficult to determine whether this signal arises from biological truth, experimental error, or random noise. On the other hand, biological functions are interconnected. The activity of a particular function requires supply of inputs and utilization of outputs by other functions. Also, the stimulus activating a function might be effective on other closely related functions. Such biological dependence suggests that if a GO term is truly enriched, we would expect some of the neighboring terms to show similar signals. In short, when we take the inter-relationship of biological functions (represented by the hierarchical structure in the GO DAG) into consideration, moderate but consistent signals from a neighborhood of related GO terms might be more trustworthy than a "strong" signal from an isolated single term.

We use Figure 1 to illustrate the above point. It depicts the hierarchical structure of 19 terms $\{G_j, j = 1, \cdots, 19\}$ from a small region of the GO DAG. We use $F = \{a, b, \cdots, z\}$ to denote the full list of genes, and among them the set of 8 DE genes, denoted by D, are marked by boldface (i.e., $f = 26$ and $d = 8$). Each arrow indicates a parent-child pair. Each rectangle contains the subset of genes annotated by a GO term, where (g_j, n_j) are listed in the first row under each rectangle, together with the hypergeometric P-values based on (1), denoted as $P(\cdot)$. For example, G_2 annotates $g_2 = 11$ genes and $n_2 = 6$ of them are DE. The hypergeometric P-value is 0.034. Figure 1 exhibits some important features of the GO DAG, such as the true-path rule and multiple inheritance (e.g., G_{15} is the child of G_{11} and G_{12} simultaneously).

Enrichment analysis based on the hypergeometric test has some limitations. First, it can not distinguish GO terms with the same (g_j, n_j). In Figure 1, nodes G_8, G_{18}, and G_{19} have the same P-value because they have identical $(g_j, n_j) = (1, 1)$. Examining the whole graph, we might consider G_8 more likely to be enriched because of the stronger evidence of enrichment in its neighborhood (related biological functions). This limitation stems from treating GO terms as isolated entities and ignoring the hierarchical structure. Second, the hypergeometric P-value has a size constraint. For a GO term of size g_j, the smallest possible P-value is attained when all the annotated genes are DE (i.e., $n_j = g_j$). This lower bound is reversely associated with g_j. For example, with ($f = 26$, $d = 8$), the hypergeometric P-value is 0.005 when $(g_j = n_j = 4)$ and 0.046 when $(g_j = n_j = 3)$. If we set the significance level at 0.01, any term with a size less than 4 will be automatically excluded from the inference. From a biologist's point of view, detecting more specific GO terms, which usually have a smaller size (g_j), might be more desirable because they provide more detailed information. The third limitation is the repeated use of information. Due to the true-path rule, the genes annotated by a GO term are involved in the tests for all of its ancestors. For example, the information of gene *a* being DE is used seven times, in the hypergeometric tests for each of (G_2, G_3,$\cdots$, G_8). The information of gene *z* being non-DE, however, is used only once (G_3). It is desirable for a test procedure to appropriately distribute evidence from each gene among GO terms according to the hierarchical structure.

A number of new methods have been proposed trying to address the limitations of the hypergeometric test. Carmona-Saez *et al*. (2006) extracted combinations of annotations that appear in at least a pre-specified number of genes. Then a statistical test is applied to access enrichment. Lewin and Grieve (2006) proposed to group closely related GO nodes and compute a hypergeometric P-value for each group. Alexa *et al.* (2006) evaluated enrichment from leaf to root, downweighting genes annotated by child terms which have already been declared significantly enriched. Grossmann *et al*. (2007) evaluated each GO term conditional on the enrichment at the parent(s). Falcon and Gentleman (2007) developed a R package on a conditional hypergeometric test to utilize parent-child relationship. Goeman and Buhlmann (2007) investigated methodological issues in the analysis of gene expression data in terms of gene sets. Lu *et al.* (2008) developed a probabilistic generative model for GO enrichment analysis. Bauer *et al.* (2010) analyzed functional terms in a Bayesian network which assumes gene responses to be directly associated with the activation of biological functions. Zhang *et al*. (2010) proposed a Bayesian method to model the DE status of individual genes with a hierarchical prior on relevance parameters to account for the GO structure. Stingo *et al*. (2011) developed a Bayesian model which uses information on pathways and gene networks. Wang *et al.* (2011) presented a network-based ontology analysis method. Wei and Pan (2012) investigated integrative modeling of multiple gene networks and diverse genomic data to identify targeted genes of transcription factor. Huang *et al*. (2009a) reviewed existing enrichment tools and classified them into three categories: 1) singular enrichment analysis, which calculates enrichment P-value for each term separately based on a pre-selected gene list; 2) gene set enrichment analysis, which directly

uses gene experimental values (no need to pre-select genes); 3) modular enrichment analysis, which utilizes the inter-relationship among terms to assess enrichment. Some modular enrichment tools allows researchers to simultaneously incorporate multiple annotation resources (e.g., DAVID (Huang *et al*. 2009b), GENECODIS (Nogales-Cadenas *et al*, 2009), and GeneTerm Linker (Fontanillo *et al*, 2011)). For example, DAVID covers over 40 annotation categories.

We propose a Bayesian extension of the hypergeometric test which naturally incorporates the GO structure into enrichment analysis. Like the traditional hypergeometric test, the Bayesian method assumes that if a GO term is not enriched, the number (n_j) of DE genes annotated to the term follows the hypergeometric distribution. If the term is enriched, n_j tends to have a greater value, deviating from the hypergeometric distribution. Such deviation has been described by the non-central hypergeometric distribution (Harkness, 1965). We construct a Bayesian non-central hypergeometric model and the GO structure is incorporated through a hierarchical prior on the non-centrality parameters. According to Huang *et al*. (2009a), the proposed method belongs to the category of modular enrichment analysis. Subramanian *et al*. (2005), on the other hand, would refer to it as enrichment analysis of annotations. We demonstrate that the proposed method can overcome the aforementioned limitations of the hypergeometric *P*-value and produce biologically meaningful results.

2 Method

Researchers have used the example of picking colored balls at random from an urn to explain the idea of the non-central hypergeometric distribution. Suppose we pick d balls without replacement from an urn containing a total of f balls, among which g are red and $(f - g)$ are black. When the sampling is unbiased, each ball has an equal chance of being selected and the number of red balls picked follows the hypergeometric distribution. The non-central hypergeometric distribution arises where the sampling is biased. For example, due to differences in size, weight, or texture, a red ball might have a greater chance of being picked than a black ball. In this case, the number of red balls picked can be modeled by the non-central hypergeometric distribution (Harkness, 1965; Fog, 2008). Denote the number of red/black balls picked as Y_r and Y_b, respectively, and $Y_r \sim Binomial(g, p_r)$ and $Y_b \sim Binomial(f - g, p_b)$. Then the conditional probability of Y_r given $Y_r + Y_b = d$ is

$$
\begin{aligned}
P(Y_r=n|Y_r+Y_b=d) &= \frac{P(Y_r=n)P(Y_b=d-n)}{\sum_i P(Y_r=i)P(Y_b=d-i)} \\
&= \frac{\binom{g}{n} p_r^n (1-p_r)^{g-n} \binom{f-g}{d-n} p_b^{(d-n)} (1-p_b)^{[(f-g)-(d-n)]}}{\sum_i \binom{g}{i} p_r^i (1-pr)^{g-i} \binom{f-g}{d-i} p_b^{(d-i)} (1-pb)^{[(f-g)-(d-i)]}} \\
&= \frac{\binom{g}{n}\binom{f-g}{d-n}\theta^n}{\sum_{i=l}^{s}\binom{g}{i}\binom{f-g}{d-i}\theta^i},
\end{aligned}
\tag{2}
$$

where parameter $\theta \equiv \frac{p_r/(1-p_r)}{p_b/(1-p_b)}$ is called the non-centrality parameter, effectively the odds ratio measuring the sampling bias. The valid summation boundaries in the denominator are $l = max(0, d-(f-g))$ and $s = min(g, d)$. For the special case of $\theta = 1$, or equivalently $p_r = p_b$, it can be shown that the denominator reduces to $\binom{f}{d}$, and (2) becomes a hypergeometric

probability. Equation (2) provides a flexible model to accommodate biased sampling where $\theta \neq 1$. In the context of enrichment analysis, for GO term G_j, we consider the red and black balls in the urn as the groups of genes annotated and not annotated by G_j in the full list, with sizes g_j and $(f - g_j)$, respectively. Hence the enrichment of G_j can be modeled through the non-central hypergeometric distribution for n_j with parameter θ_j. We declare G_j to be enriched if the estimated non-centrality parameter θ_j is significantly greater than 1.

In reality, researchers are usually interested in evaluating the whole DAG simultaneously instead of a single GO term. We propose a Bayesian non-central hypergeometric model to addresses the limitations of the traditional hypergeometric *P*-value: 1) genes annotated by a GO term is repeatedly used in the test of enrichment for all of its ancestors; 2) there is no sharing of information among related GO terms; and 3) it has a size constraint.

We address Limitation 1) by only including the additionally-contributed information by term G_j into the likelihood. Let g_j^* be the number of genes that are most specifically annotated to term G_j, among which, n_j^* are DE genes. In other words, the g_j^* genes are annotated by G_j, but not by any of its child terms. In Figure 1 the most specifically annotated genes by G_5, G_7 and G_8 are (*c*, *s*, *u*), (*b*, *d*, *t*), and (*a*), and we have $(g_5^*=3, n_5^*=1)$, $(g_7^*=3, n_7^*=2)$ and $(g_8^*=1, n_8^*=1)$. Corresponding to (g_j^*, n_j^*), we define $f_j^*=f-(g_j-g_j^*)$ and $d_j^*=d-(n_j-n_j^*)$. Here $(g_j-g_j^*)$ and $(n_j-n_j^*)$ represent information that has been used in the tests of offspring terms of G_j. Then the likelihood from G_j is

$$P(n_j^*|f_j^*, d_j^*, g_j^*, \theta_j)=\frac{\binom{g_j^*}{n_j^*}\binom{f_j^*-g_j^*}{d_j^*-n_j^*}\theta_j^{n_j}}{\sum_{i=l^*}^{s^*}\binom{g_j^*}{i}\binom{f_j^*-g_j^*}{d_j^*-i}\theta_j^i}. \quad (3)$$

Accordingly, the lower and upper bounds for summation are $l^*=max(0, d_j^*-(f_j^*-g_j^*))$ and $s^*=min(g_j^*, d_j^*)$.

We address Limitation 2) by incorporating the GO hierarchical structure through a prior model on the set of non-centrality parameters. First we define $\phi_j \equiv \log(\theta_j)$, hence $-\infty < \phi_j < \infty$. We set $\phi_1 = 0$ for the root node. From top to bottom, we assume the following conditional distribution of ϕ_j ($j = 2, \cdots, J$) given the non-centrality parameters of its parent nodes, denoted by $\boldsymbol{\phi}_{P_j} = \{\phi_k : G_k \in P_j\}$,

$$\varphi_j|\varphi_{P_j}, \sigma^2 \sim \sum_{k:G_k\in P_j}\frac{1}{|P_j|}N(\varphi_k, \sigma^2). \quad (4)$$

Here P_j denotes the set of parent nodes of G_j and $|P_j|$ is the number of GO terms in P_j. Model (4) assumes that ϕ_j arises from a mixture distribution of $|P_j|$ components, each being a normal distribution centered at the non-centrality parameter of one of its parents. With an equal mixing probability $1/|P_j|$ we assume *a priori* that each parent has equal influence on ϕ_j. Parameter σ^2 characterizes the variability among child nodes. The joint prior of $\boldsymbol{\phi} = \{\phi_j : j = 2, \cdots, J\}$ is obtained by the product of (4) over $j = 2, \cdots, J$. This prior provides a mechanism to share information among GO terms based on the DAG structure. It also accommodates multiple inheritance. We assign an inverse gamma prior $IG(0.01, 0.01)$ to σ^2, which has been used extensively in Bayesian models (Gelman *et al*., 2003).

We infer the enrichment of term G_j based on the posterior distribution of ϕ_j, denoted by $[\phi_j \mid (\boldsymbol{f}^*, \boldsymbol{d}^*, \boldsymbol{g}^*, \boldsymbol{n}^*)]$, where $(\boldsymbol{f}^*, \boldsymbol{d}^*, \boldsymbol{g}^*, \boldsymbol{n}^*) = \left\{(f_j^*, d_j^*, g_j^*, n_j^*), j = 2, \cdots, J\right\}$. Specifically, we use $B_j \equiv P(\phi_j > 0 \mid (\boldsymbol{f}^*, \boldsymbol{d}^*, \boldsymbol{g}^*, \boldsymbol{n}^*))$, denoted as the *B*-value, to measure the enrichment of a GO term. It is the posterior probability of G_j being enriched in the DE gene list *D*. By using the *B*-value, we overcome Limitation 3), because the *B*-value has a support of (0,1) and it does not have a size constraint like the hypergeometric *P*-value.

A Markov Chain Monte Carlo (MCMC) sampling algorithm is employed to simulate random samples from the joint posterior distribution. The full conditional distribution of σ^2 is an inverse gamma distribution. A Metropolis-Hasting algorithm is implemented to sample from the full conditional distribution of ϕ_j. See Web Appendix A and B for more details on the MCMC algorithm and the calculation of the *B*-value.

The *B*-value measures the strength of enrichment evidence at each GO term. We can use it as a screening tool to rank GO terms to help researcher select terms for further investigation. To answer questions such as "what should be the cutoff value for the *B*-value", the Bayesian FDR (Newton *et al.*, 2004) has been widely used to account for multiplicity:

$E(FDR|data) = \frac{\sum \delta_j (1 - B_j)}{D}$, where $D = \Sigma_j \delta_j$ is the number of selected GO terms, indicator $\delta_j = 1$ if the *j*th term is identified as enriched (its *B*-value ranks among the top *D* terms), and $\delta_j = 0$ otherwise. $E(FDR|data)$ is the posterior proportion of false discoveries in the identified enriched terms. The Bayesian FDR has been used extensively in Bayesian highthroughput data analysis (Storey, 2002; Do *et al.*, 2005). Examples of other developments in Bayesian paradigm to address statistical significance in multiple comparison include Muller *et al.* (2006) and Scott and Berger (2010). Due to the complexity in the biological data-mining environment (e.g., gene-gene and function-function correlation, annotation redundancy, etc.), the sensitivity and specificity of existing enrichment methods are not yet in the optimal state (Goeman *et al.*, 2007). We agree with the idea that the *B*-value, like other enrichment scores including the *P*-value, should be treated as a scoring system that plays an advisory role such as ranking and suggesting possible relevant annotations, as opposed to an absolute, decision-making role (Huang *et al.*, 2009b). The important guideline to help users in adjusting analytic thresholds is the notion that enriched terms should make sense based on *a priori* biological knowledge of the study (Huang *et al.*, 2009a).

3 Comparison with Existing Methods

Based on Figure 1, we compare the proposed method with a number of existing methods: the hypergeometric *P*-value, the *elim P*-value (Alexa *et al.*, 2006), the parent-child *P*-value (Grossmann *et al.*, 2007), and a Bayesian model-based gene set analysis (Bauer *et al.*, 2010), denoted as MGSA. They have different assumptions and tend to detect different features. With the biological truth unknown, there is no gold standard to compare methods in real studies (Grossmann *et al.*, 2007). We try to demonstrate and understand the distinctive characteristics of each method. In Figure 1 we present (g_j^*, n_j^*) and the *B*-value in the second row under each term. We also use $A(\cdot)$, $G(\cdot)$, and $M(\cdot)$ to present the *elim P*-value, the parent-child *P*-value, and the posterior probability of activation under MGSA, respectively. Stronger enrichment is indicated by larger values of $B(\cdot)$ and $M(\cdot)$, and smaller values of $P(\cdot)$, $A(\cdot)$, and $G(\cdot)$.

Compared with the hypergeometric *P*-value, the *B*-value has several advantages. First, it can distinguish GO terms with the same (g_j, n_j). The three nodes G_8, G_{18}, and G_{19}, all with $(g_j, n_j) = (1, 1)$ and thus equal hypergeometric *P*-values, are differentiated by their *B*-values (B_8

= 0.861, B_{18} = 0.404, and B_{19} = 0.315). Due to intrinsic noises in high-throughput data collection and processing, there can be errors in DE detection, which affects the accuracy of the hypergeometric *P*-values. The proposed method recognizes that neighboring terms on the GO DAG represent related biological functions. Thus enrichment detected in a neighborhood might be more reliable than that detected in an isolated term. The larger *B*-value for G_8 is attributed to the stronger evidence of enrichment in its neighbors. Second, the proposed method mitigates the undue influence of GO term size. The hypergeometric *P*-value identifies G_2 as the most enriched term, although its enrichment level, $n_2/g_2 = 6/11 = 0.55$, is modest. Because it is one of the largest terms ($g_2 = 11$) in Figure 1, it has the smallest *P*-value of 0.034. If we exclude evidence already accounted for in the off-springs, its marginal evidence of enrichment is ($g_2^*=3, n_2^*=2$), the same as G_7. The *B*-value takes into account marginal evidence and enrichment in the neighborhood: for G_7, both of its parent G_5 $(g_5^*=3, n_5^*=1)$ and off-spring G_8 $(g_8^*=1, n_8^*=1)$ contribute supporting evidence, which is not the case for G_2. Thus the *B*-value considers G_7 to have stronger evidence of enrichment.

Both the *elim P*-value (Alexa *et al*., 2006) and the parent-child *P*-value (Grossmann *et al*., 2007) attempt to address the "dependency problem" caused by overlapping annotations between parent-child pairs. The *elim* method tests GO terms from leaf to root, and removes all genes annotated to a significantly enriched term from its ancestors. Thus the *elim P*-value tends to identify highly enriched GO terms that remain significant even after removing enrichment evidence from their offsprings. Considering the small size of Figure 1, we set the *P*-value cutoff at 0.10. Thus G_7 is considered significantly enriched and genes (*a*, *b*, *d*, *t*) are removed from its ancestors (G_2, ⋯, G_6). The *elim P*-value considers G_7 the most significantly enriched. Due to the removal of genes (*a*, *b*, *d*, *t*), none of G_7's ancestors are significant, and the previously top-ranking term G_2 (by both the *B*-value and the hypergeometric *P*-value) now ranks even behind G_{18} and G_{19}. The parent-child *P*-value works from root to leaf, computing the hypergeometric *P*-value for each term in the context of its parent (treating the genes annotated by its parent as the full gene list). This approach identifies GO terms that show stronger enrichment compared to their parents. Effectively it penalizes terms for having highly enriched parents. Take G_8 and G_{19} for example, which have the same (I_A, n_A). Based on the parent-child *P*-value, G_{19} is more enriched than G_8 because G_{19}'s enrichment (1 out of 1) is stronger relative to its parent G_{17} (1 out of 6), than G_8 (1 out of 1) relative to its parent G_7 (3 out of 4). The *B*-value accounts for parent-child relationship through hierarchical prior (4). It puts greater emphasis on consistent enrichment signal from neighborhoods of related GO terms. As a result, G_8 has a larger *B*-value than G_{19} because the enrichment in G_8 is corroborated by its neighbors.

MGSA evaluates all terms at once using a Bayesian network (Bauer *et al*., 2010), assuming that in an experiment, a limited number of terms are activated which activate the genes they annotate (i.e., the true states of these genes become DE). Thus a gene is assumed to be truly DE if and only if at least one of its annotations is activated. The observed status are noisy observations of the true gene states, subject to a false positive rate α and false negative rate β. The inference is based on the posterior probability of each term being activated (enriched). MGSA does not use the GO DAG structure other than the true-path rule. One of its advantages is parsimony because a gene being DE is assumed to be fully explained by the activation of one of its annotations. Declaring additional terms to be active is discouraged because the likelihood function remains the same but a penalty is imposed by the prior. In Figure 1 MGSA identifies G_7, G_{18} and G_{19} as most likely to be enriched, because they efficiently explain the DE status of five genes (*a*, *b*, *d*, *g*, *h*) at the cost of one false negative gene (*t*). Although G_8 shows the same evidence of enrichment as G_{18} and G_{19}, MGSA assigns a smaller score to G_8 because once G_7 is declared active, declaring G_8 active only

leads to penalty by the prior. In contrast, the proposed method assigns a greater *B*-value to G_8 than G_{18} and G_{19} because it does not emphasize parsimony. Instead, it emphasizes the detection of subareas in GO where neighboring terms show consistent enrichment signals. Under this rationale, G_{18} and G_{19} have smaller *B*-values because their enrichment was not corroborated by their neighbors. Both MGSA and *B*-value address overlapping annotations, but via different approaches: MGSA by imposing that overlapping annotations lead to no gain in likelihood but penalty in prior; *B*-value by only including marginal evidence (excluding genes annotated by child terms) into likelihood (3).

In summary, all methods have their strength and weakness. The *elim P*-value and parentchild *P*-value tend to identify highly enriched terms that stand out from their background. These highly enriched terms might bear important biological meaning, but they are vulnerable to random noise or experimental error. MGSA tends to identify a small number of terms among which there is little overlap, and jointly they best account for the DE genes. It is advantageous in parsimony but it might suffer the problem of "not seeing the forest for the tree". That is, MGSA does not detect joint terms which may contain unique biological information that is not held by individual terms (Huang *et al.*, 2009a). The *B*-value tends to identify neighborhoods of related terms where moderate but consistent signals are considered more trustworthy than strong signals from isolated terms. In addition, sharing information among neighbors may mitigate the impact of mis-classified genes on the inference of individual terms. However, it might miss some highly enriched but isolated functions. Incorporating the GO graphical structure via prior implies that mistakes in the GO database might adversely affect its performance.

4 Simulation to Assess Performance

We conduct a simulation to compare the performance of different methods in detecting enriched GO subareas. Based on the DAG in Figure 1, we simulated data (the DE status of genes) following a generative scheme similar to that of Bauer *et al.* (2010). First, we assume the region containing (G_2, ⋯, G_8) to be active (truly enriched). Genes annotated by active terms are set to be non-DE and DE with probabilities β and $1-\beta$, respectively. Genes annotated by inactive terms are set to be non-DE and DE with probabilities $1-\alpha$ and α. We set $\alpha = \beta = 0.3$ and generated 100 datasets. In Figure 2.A we present the precision/recall plot where precision=TP/(TP+FP) and Recall=TP/(TP+FN). Here TP, FP, and FN are the numbers of true positives, false positives, and false negatives, respectively. In Figure 2.B we present the ROC plot. This simulation is not a comparison of overall performance among the enrichment methods. Instead, by setting a subarea as truly enriched, it specifically evaluates the ability of different methods in identifying neighborhoods of related terms. It is no surprise that *B*-value performs best because it is designed to identify related terms. The parent-child *P*-value performs poorly on this task for a simple reason: In order to identify related GO terms, moderate but consistent signal in a neighborhood is considered stronger evidence than extreme signals from isolated terms. The parent-child *P*-value is designed to do exactly the opposite, detecting highly enriched terms that stand out from their parents. Here we assume that related functions tend to be activated together, thus we specified truly enriched terms as a subarea. This assumption implicitly favors the proposed *B*-value. We conduct a second simulation which adopts the assumption of MGSA that DE genes are determined by the activation of a limited number of terms among which there is minimum overlap (see Web Appendix C). Under this simulation truth MGSA has the best performance while the hypergeometric *P*-value has the poorest performance. Thus we caution against over-interpreting Figure 2 as the comparison of overall performance among different methods.

With biological knowledge still evolving, inevitably there are some mistakes and incompletions in existing annotation systems. We conduct a third simulation to assess the performance of enrichment methods under imperfect annotations. We first generate data following the generative scheme described above. At this stage we assume perfect annotations. Then we introduce two types of imperfection: one by switching the annotations of two randomly selected genes, which attempts to mimic potential mistakes in existing annotations; the other by randomly removing a gene from its annotations, which attempts to mimic incompletions in existing annotations. We find that imperfect annotations does lead to poorer performance for all the methods, but their relative performance remains unchanged. It shows that different enrichment methods can produce useful results despite some imperfections in the existing annotation system. Please see Web Appendix D for simulation under the imperfect annotations.

5 Application

5.1 Dataset

We use a microarray data set to demonstrate the proposed Bayesian method. Researchers evaluated the effect of three stimulus on a B cell lymphoma cell line: the B cell antigen receptor (BCR), CD40, and a combination of the two (Basso *et al*., 2005). The expression data was processed using the SAM approach (Tusher *et al*., 2001). The full list contains f = 3952 genes and a cluster of d =196 genes was identified. Genes in this cluster shared a particularly interesting expression pattern: they were all upregulated in response to BCR signaling alone, but this upregulation was suppressed when CD40 signaling was included. These treatment conditions mimic important biological responses of immature B cells (Hsueh and Scheuermann, 2000), which must distinguish between signals delivered by authentic pathogen-derived antigens and by self antigens. In the former case, B cells need to respond by productive proliferation and differentiation into immune effector cells. In the later case, B cell responses need to be suppressed either through the induction of unresponsiveness or apoptotic cell death. It is hypothesized that B cells receiving only one signal, through the BCR, will proliferate and die, but B cells receiving two signals, through the BCR combined with a co-stimulatory receptor like CD40, will proliferate and survive. Thus, genes in the identified cluster are suppressed with the addition of CD40 signaling and could thus be involved either in the cell death response or in the induction of unresponsiveness.

5.2 Result

The full list of 3952 genes are annotated by J =6768 GO terms. We compare the top 100 GO terms selected by the hypergeometric P-value (denoted as the P-list) and by the B-value (denoted as the B-list). First, we examine which method tends to identify sets of related terms instead of isolated ones. We define a GO-set to be a group of GO terms connected through parent-child links. For example, in Figure 1, if we set the threshold of B-value at 0.75, then four terms are above this threshold (G_2,G_5,G_7,G_8). They form two GO-sets, one containing (G_5,G_7,G_8) and the other by G_2 itself. Larger GO-sets might contains more reliable enrichment signals than isolated single terms. Table 1 indicates that the B-value tends to identify larger GO-sets. Specifically, 25 terms in the P-list are isolated single terms, while there are only 7 such terms in the B-list. In addition, there are 15 GO-sets of size 2 (parent-child pairs) in the P-list. Among them, 9 pairs have parent terms with $(g_j^*, n_j^*)=(0,0)$, meaning that those parent terms contribute no additional evidence, and their small P-values are purely due to repeated use of information that has already been used in the test of their off-springs. By comparison, the B-list has far fewer such overlapping cases because the method is based on (g_j^*, n_j^*) instead of (g_j, n_j). Furthermore, the B-value has

identified 3 large GO-sets each with more than 10 GO terms, while the largest GO-set in the *P*-list is of size 8.

Next we compare the biological functional groups identified by GO-sets of size ≥2 in the *P*-list and the *B*-list. A functional group is defined as a region in DAG that is associated with a relatively coherent biological theme and contains at least one such GO-set. Table 2 lists those functional groups identified and the number of GO terms from each list. First, the *P*-list has identified 5 functional groups (groups A – E) and the *B*-list has identified 2 additional functional groups (groups F, G). Among groups A – E, three groups (A, C, and E) are represented by similar numbers of GO terms in both lists. For the other two groups (B and D), the number of GO terms from the *B*-list substantially exceeds that from the *P*-list. Second, the two additional functional groups (F and G) identified by the *B*-list are represented by neighborhoods of 7 and 16 GO terms, indicating strong signals of enrichment that otherwise would have been missed by the hypergeometric *P*-value. For example, Figure 3 shows the regional DAG of functional group G. Among these 16 GO terms from the top 100 *B*-value list, only two disconnected terms (GO:0048167 and GO:0032230) are selected by the *P*-list, providing trivial evidence for this functional group. Group G, which is synaptic transmission, can be biologically associated with the experiment. Specifically, B cell receptor is one of the cell surface receptors that can activate phospholipase C enzymes to govern the opening of calcium ions channel in endoplasmic and reticulum and plasma membrane, and changes in intracellular calcium concentration is an established mechanism to regulate synaptic transmission (Wang *et al*., 2000; Dutting, Brachs, and Mielenz, 2011). In addition, disruption of mitochondrial membrane potential (Group F) has been found to be one of the main causes of BCR-medicated cell cycle arrest and apoptosisi in simulated immature B cells (Katz *et al*., 2001). We also include biological implications of functional groups A–E in Web Appendix E.

Finally we illustrate that *B*-value can distinguish GO terms with the same (g_j, n_j) by incorporating evidence from neighboring terms. There are eight GO terms with $(g_j, n_j) = (2, 2)$ that have the same hypergeometric *P*-value of 0.0024. All of them are in the top-100 *P*-list. The *B*-value suggests that these terms are very different. For example, GO:0007217 (tachykinin signaling pathway) has a *B*-value of 0.948 and it is in the top-100 *B*-list. In contrast, GO:0051319 (G2 phase) has a *B*-value of 0.397 and its rank based on the *B*-value is 5196. To shed light on their difference in *B*-values, we compare their regional DAGs in Figure 4 and 5. The parent of GO:0007217 has stronger enrichment than that of GO: 0051319. In addition, GO:0007217 has 9 siblings, 6 of which have genes represented in the identified cluster ($n_j > 0$). By comparison, GO:0051319 has 6 siblings, none of which has genes represented in the cluster. The support from related GO terms is substantially higher for GO:0007217 than for GO:0051319, and thus GO:0007217 is considered more likely to be associated with the identified list *D*. We conducted a literature search and could not find experimental evidence for the involvement of G2 phase (GO:0051319) in the regulation of B cell function. On the other hand, tachykinin (GO: 0007217) has been found to be secreted during the differentiation of B cell precursors thereby regulating their development (Milne *et al*., 2004). Thus, GO:0007217 does appear to be biologically relevant.

6 Discussion

We have proposed a Bayesian approach to detecting enriched annotations based on a non-central hypergeometric model. To address the issue of repeated use of information, we only include additionally contributed information from each GO term in the likelihood. We encourage sharing of evidence among related biological functions by specifying a hierarchical prior on the non-centrality parameters based on the dependence structure of GO DAG. We also use the *B*-value, which does not have a size constraint, to measure

enrichment. The proposed method is a natural extension of the hypergeometric test, which provides a straightforward connection with the conventional test. More importantly, the mechanism induced by the Bayesian model to share information among related GO terms strengthens the detection of moderate but consistent enrichment signals which helps researchers to identify sets of related terms rather than individual isolated terms.

Currently annotation databases like GO are imperfect and still evolving, which means that mistakes in GO annotations might adversely affect functional enrichment analysis. This is a challenge faced by all enrichment analysis methods. Many studies have demonstrated that although imperfect, incorporating annotation information may help researchers achieve more meaningful results. The proposed method may mitigate the impact of mis-annotations in GO. For example, the hypergeometric test accesses each GO terms separately, thus a mis-annotated gene will greatly impact the *P*-value of a small term. Furthermore, this mis-annotation affects all the ancestors due to the true-path rule. The proposed Bayesian method can reduce the impact of mis-annotation in two ways: First, in likelihood (3), only the additional genes are included. Thus mis-annotation at a GO term does not directly affect its ancestors' likelihood. Second, the hierarchical prior (4) allows borrowing strength among neighboring terms, which considers moderate but consistent signals from a neighborhood stronger evidence than isolated strong signals from individual terms. It is reasonable to believe that a false signal due to mis-annotation at a GO term would not be consistent with the true signals from its neighbors, which would be down-weighted by the Bayesian model.

The proposed Bayesian model is constructed based on the GO DAG structure. In this paper we could not investigate all possible structures of functional terms. However, our paper provides a framework that can be easily extended to account for other structures, where only the prior model (4) needs to be modified accordingly. For example, to make inference on KEGG functional terms which are organized as pathways, a possible choice of the prior is the conditional auto-regressive (CAR) model (Gelfand and Vounatsou, 2003) to incorporate the spatial relationship between KEGG terms. On the other hand, a number of methods have been proposed to measure the similarity between terms from semantic, topological, and functional perspectives (Schlicker *et al.*, 2006; Lerman and Shakhnovich, 2007). Incorporating similarity measures into enrichment analysis might be another interesting topic in modular enrichment methodological research, especially in the Bayesian paradigm.

Some state-of-the-art enrichment tools allow researchers to utilize different sources of annotation simultaneously, including DAVID (Huang *et al.*, 2009b), GENECODIS (Nogales-Cadenas *et al.*, 2009), and GeneTerm Linker (Fontanillo *et al.*, 2011). One type of tools, such as DAVID and GENECODIS, first organize and condense heterogeneous annotation content (such as GO terms, protein pathways, etc.) into term classes based on biological co-occurrences. Then enrichment tests are performed on those term classes. For this type of tools, the proposed Bayesian approach can be implemented at the second stage to account for inter-relationship among newly constructed term classes, if such inter-relationship can be established. Another type of tools, such as GeneTerm Linker, first perform enrichment analysis in different annotation spaces and the outputs are filtered and linked to produce metagroups of coherent biological significance. For this type of tools, the proposed Bayesian method can be implemented at the first stage to improve the accuracy and robustness of individual enrichment analysis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the referees, associate editor, and editor for valuable comments resulting in substantial improvements to the manuscript. This work has been supported in part by NIH 1R15HG006365-01, UT-STAR, and NIH/NCATS UL1TR000451.

References

Alexa A, Rahnenfuhrer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics. 2006; 22:1600–1607. [PubMed: 16606683]

Bauer S, Gagneur J, Robinson PN. Going Bayesian: Model-based gene set analysis of genome-scale data. Nucleic Acids Research. 2010; 38(11):3523–3532. [PubMed: 20172960]

Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. Reverse engineering of regulatory networks in human B cells. Nature Genetics. 2005; 37:382–390. [PubMed: 15778709]

Carmona-Saez P, Chagoyen M, Rodriguez A, Trelles O, Carazo JM, Pascual-Montano A. Integrated analysis of gene expression by Association Rules Discovery. BMC Bioinformatics. 2006; 7:54. [PubMed: 16464256]

Do K, Muller P, Tang F. A Bayesian mixture model for differential gene expression. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2005; 54(3):627–644.

Dutting S, Brachs S, Mielenz D. Fraternal twins: Swiprosin-1/EFhd2 and Swiprosin-2/EFhd1, two homologous EF-hand containing calcium binding adaptor proteins with distinct functions. Cell Communication and Signaling. 2011; 9:2. [PubMed: 21244694]

Efron B, Tibshirani B. On testing the significance of sets of genes. Annals of Applied Statistics. 2007; 1:107–129.

Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. Bioinformatics. 2007; 23:257–258. [PubMed: 17098774]

Fog A. Sampling methods for Wallenius' and Fisher's noncentral hypergeometric distributions. Communications in Statistics - Simulation and Computation. 2008; 37:241–257.

Fontanillo C, Nogales-Cadenas R, Pascual-Montano A, de Las Rivas J. Functional analysis beyond enrichment: Non-redundant reciprocal linkage of genes and biological terms. PLoS ONE. 2011; 6(9):e24289. [PubMed: 21949701]

Gelfand AE, Vounatsou P. Proper multivariate conditional autoregressive models for spatial data analysis. Biostatistics. 2003; 4:11–25. [PubMed: 12925327]

Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. Bayesian Data Analysis. Chapman & Hall; 2003.

The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. Nature Genetics. 2000; 25:25–29. [PubMed: 10802651]

Goeman JJ, Buhlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics. 2007; 23:980–987. [PubMed: 17303618]

Grossmann S, Bauer S, Robinson PN, Vingron M. Improved detection of overrepresentation of Gene-Ontology annotations with parent-child analysis. Bioinformatics. 2007; 23:3024–3031. [PubMed: 17848398]

Harkness WL. Properties of the extended hypergeometric distribution. The Annals of Mathematical Statistics. 1965; 36:938–945.

Hosack DA, Dennis G Jr, Sherman BT, Lane HC, Lempicki RA. Identifying biological themes within lists of genes with EASE. Genome Biology. 2003; 4:R70. [PubMed: 14519205]

Hsueh R, Scheuermann RH. Tyrosine kinase activation in the growth, differentiation and death responses initiated from the B cell antigen receptor. Advances in Immunology. 2000; 75:283–316. [PubMed: 10879287]

Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Research. 2009a; 37(1):1–13. [PubMed: 19033363]

Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols. 2009b; 4(1):44–57.

Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research. 2000; 28:27–30. [PubMed: 10592173]

Katz E, Deehan MR, Seatter S, Lord C, Sturrock RD, Harnett MM. B cell receptor-simulated mitochondrial phospholipase A2 activation and resultant disruption of mitochondrial membrane potential correlate with the induction of apoptosis in WEHI-231 B cells. The Journal of Immunology. 2001; 166:137–147. [PubMed: 11123286]

Khatri P, Drăghici S, Ostermeier GC, Krawetz SA. Profiling gene expression using Onto-Express. Genomics. 2002; 79:266–70. [PubMed: 11829497]

Khatri P, Drăghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics. 2005; 21:3587–3595. [PubMed: 15994189]

Lerman G, Shakhnovich BE. Defining functional distance using manifold embeddings of gene ontology annotations. Proceedings of the National Academy of Sciences. 2007; 104(27):11334–11339.

Lewin AM, Grieve IC. Grouping Gene Ontology terms to improve the assessment of gene set enrichment in microarray data. BMC Bioinformatics. 2006; 7:426. [PubMed: 17018143]

Lu Y, Rosenfeld R, Simon I, Nau GJ, Bar-Joseph Z. A probabilistic generative model for GO enrichment analysis. Nucleic Acids Research. 2008; 36(17):e109. [PubMed: 18676451]

Milne CD, Fleming HE, Zhang Y, Paige CJ. Mechanisms of selection mediated by interleukin-7, the preBCR, and hemokinin-1 during B-cell development. Immunological Reviews. 2004; 197:75–88. [PubMed: 14962188]

Muller, P.; Parmigiani, G.; Rice, K. FDR and Bayesian multiple comparisons rules. Proc. Valencia/ISBA 8th World Meeting on Bayesian Statistics; Benidorm (Alicante, Spain). 2006.

Newton M, Noueiry A, Sarkar D, Ahlquist P. Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics. 2004; 5(2):155–176. [PubMed: 15054023]

Newton M, Quintana FA, den Boon JA, Sengupta S, Ahlquist P. Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. The Annals of Applied Statistics. 2007; 1:85–106.

Nogales-Cadenas R, Carmona-Saez P, Vazquez M, Vicente C, Yang X, Tirado F, Caraso JM, Pascual-Montano A. GeneCodis: Interpreting gene lists through enrichment analysis and integration of diverse biological information. Nucleic Acids Research. 2009; 37(SUPPL 2):W317–W322. [PubMed: 19465387]

Schlicker A, Domingues FS, Rahnenfuhrer J, Lengauer T. A new measure for functional similarity of gene products based on gene ontology. BMC Bioinformatics. 2006; 7:302. [PubMed: 16776819]

Scott JG, Berger JO. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. Annals of Statistics. 2010; 38:2587–2619.

Stingo FC, Chen YA, Tadesse MG, Vannucci M. Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes. The Annals of Applied Statistics. 2011; 5(3):1978–2002. [PubMed: 23667412]

Storey JD. A direct approach to false discovery rate. Journal of the Royal Statistical Society, Series B. 2002; 64:479–498.

Subramanian AP, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005; 102(43):15545–15550.

Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. Proceedings of the National Academy of Sciences. 2001; 98(9):5116–5121.

Wang D, Feng J, Wen R, Marine JC, Sangster MY, Parganas E, Hoffmeyer A, Jackson CW, Cleveland JL, Murray PJ, Ihle JN. Phospholipase Cgamma2 is essential in the functions of B cell and several Fc receptors. Immunity. 2000; 13(1):25–35. [PubMed: 10933392]

Wang J, Huang Q, Liu ZP, Wang Y, Wu LY, Chen L, Zhang XS. NOA: A novel network ontology analysis method. Nucleic Acids Research. 2011; 39(13):e87. [PubMed: 21543451]

Wei P, Pan W. Bayesian joint modeling of multiple gene networks and diverse genomic data to identify target genes of a transcription factor. The Annals of Applied Statistics. 2012; 6(1):334–355. [PubMed: 22408712]

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. 2004; 5:16. [PubMed: 14975175]

Zhang S, Cao J, Kong M, Scheuermann RH. GO-Bayes: Gene Ontology-based enrichment analysis using a Bayesian approach. Bioinformatics. 2010; 26:905–911. [PubMed: 20176581]

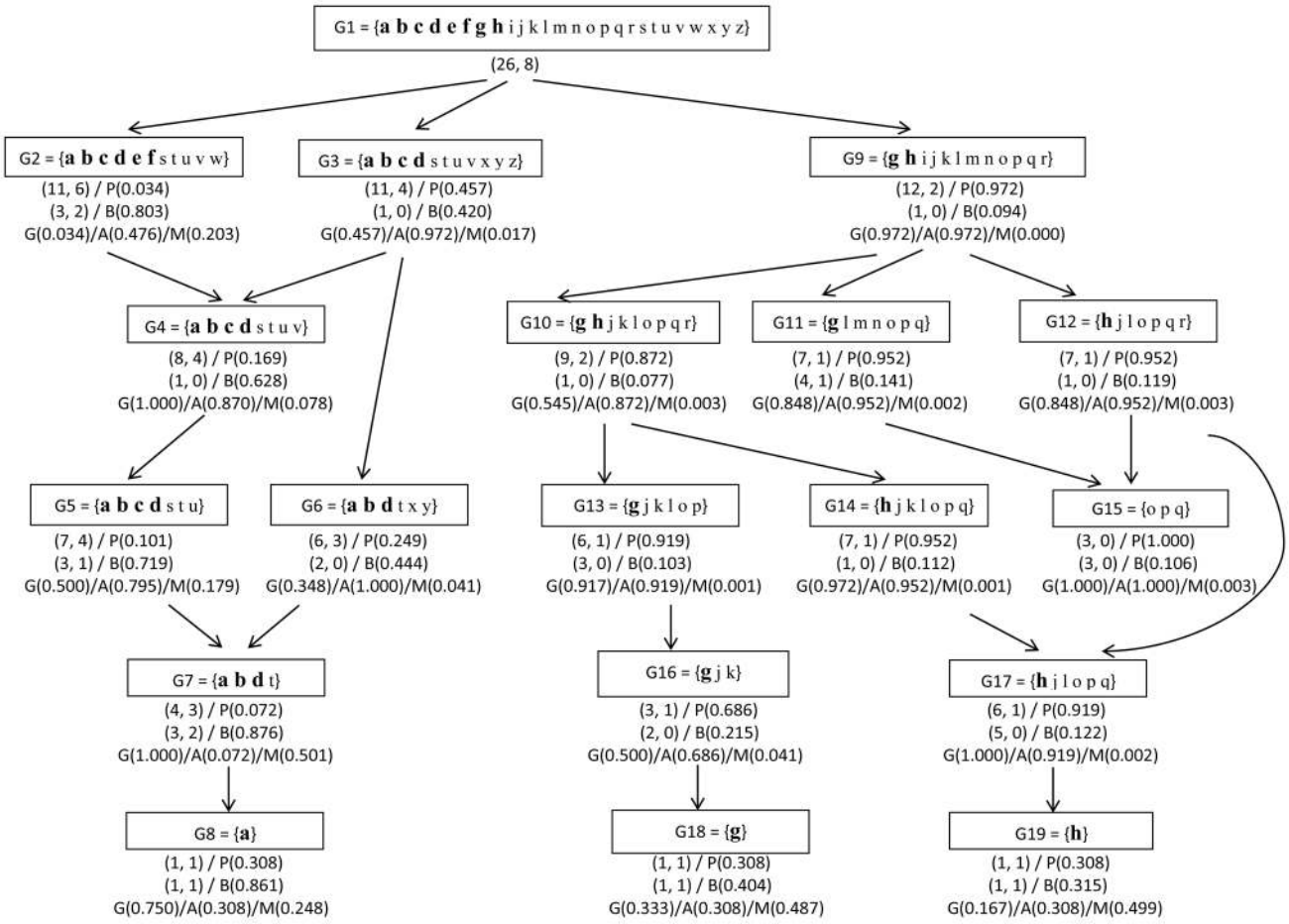

Figure 1.

Illustration of different enrichment measures based on a small region of GO DAG. The full list of genes (F) are denoted as lowercase letters; the genes in set D are marked by boldface. The rectangles contain the subset of genes annotated by each node. Under each rectangle, the first row lists (g_j, n_j) and the P-value presented in P(). The second row lists (g_j^*, n_j^*) and the B-value presented in B(). In the third row, the parent-child P-value (Grossmann *et al*., 2007) is presented in G() and the *elim* P-value (Alexa *et al*., 2006) is presented in A(), and the posterior probability under MGSA (Bauer *et al*., 2010) is presented in M(). Note that stronger enrichment is indicated by greater values of $B(\cdot)$ and $M(\cdot)$, and smaller values of $P(\cdot)$, $A(\cdot)$, and $G(\cdot)$.

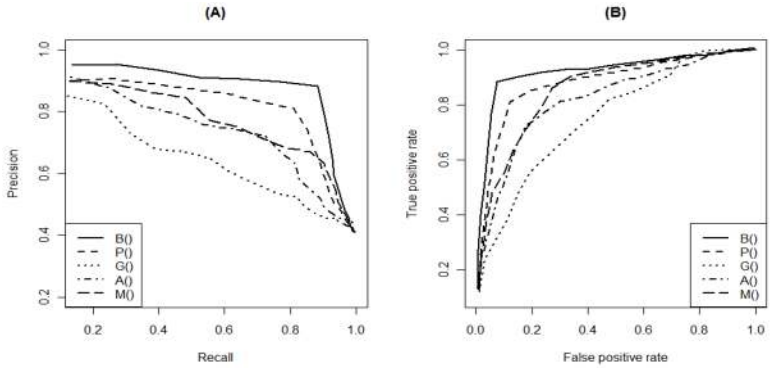

Figure 2.
Benchmarking on simulated datasets, where P() represents the hypergeometric *P*-value, B() the *B*-value, G() the parent-child *P*-value, A() the *elim P*-value, and M() the MGSA posterior probability. The left panel shows the precision/recall plot, and the right panel shows the ROC curve.

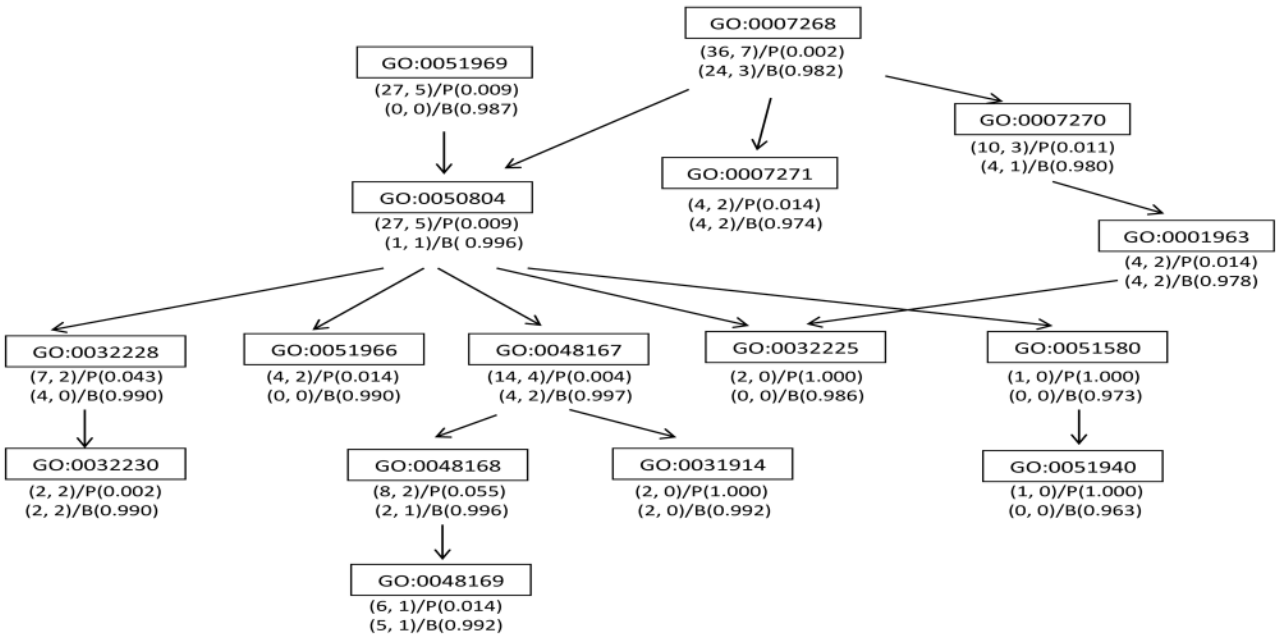

Figure 3.
The regional DAG of the biological functional group G. The rectangles in the DAG denote the GO terms in the group. The first row lists (g_j, n_j) and the *P*-value, an the second row lists (g_j^*, n_j^*) and the *B*-value.

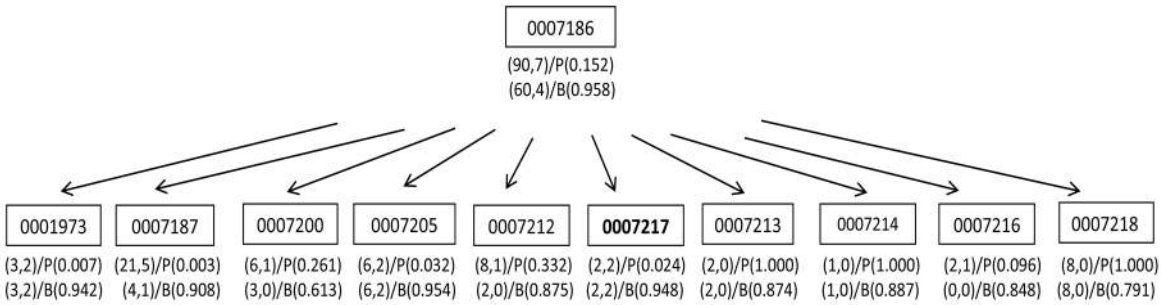

Figure 4.
The regional DAG of GO:0007217. The rectangles in the DAG denote the GO terms in the neighborhood of GO:0007217. The first row lists (g_j, n_j) and the *P*-value, an the second row lists (g_j^*, n_j^*) and the *B*-value.

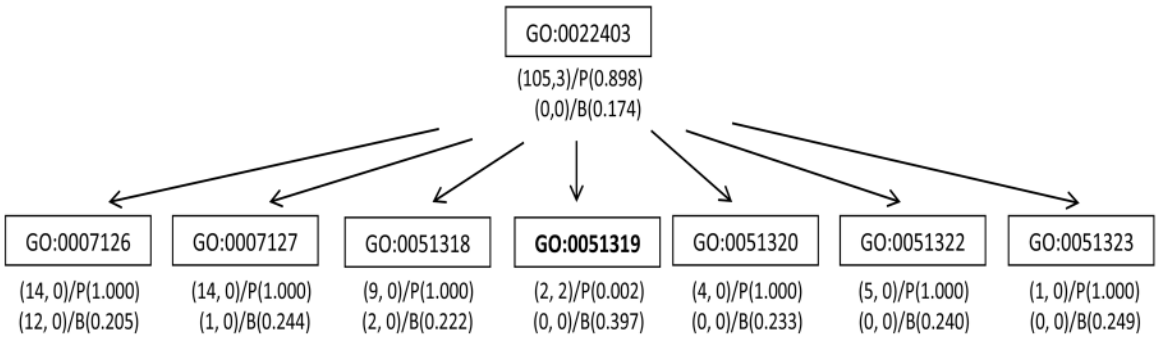

Figure 5.
The regional DAG of GO:0051319. The rectangles in the DAG denote the GO terms in the neighborhood of GO:0051319. The first row lists (g_j, n_j) and the *P*-value, an the second row lists (g_j^*, n_j^*) and the *B*-value.

Table 1

Comparison of GO-sets identified in the *P*-list and the *B*-list

Size of GO-sets	1	2	3	4	5	7	8	12	22	31
P-list	25	15	6	2	3	1	1	0	0	0
B-list	7	0	3	4	0	1	0	1	1	1

Table 2

Functional groups identified in the *P*-list and the *B*-list

Group	#(*P*)	#(*B*)	Representing GO term
A	3	3	GO:0005764 (lysosome)
B	15	31	GO:0051049 (regulation of transportation)
C	9	7	GO:0016021 (integral to membrane)
D	2	22	GO:0016757 (transferase activity, transferring glycosyl groups)
E	11	9	GO:0007188 (G-protein signaling, coupled to cAMP nucleotide second messenger)
F	0	7	GO:0051881 (regulation of mitochondrial membrane potential)
G	0	16	GO:0051969 (regulation of synaptic transmission)

The seven functional groups are represented by letters A to G. #(*P*) and #(*B*) represent the number of GO terms in the *P*-list and the *B*-list that appear in each functional group, respectively. The last column lists the GO terms that represent the corresponding functional groups, which are the root terms (the least-specific terms) in the groups, respectively.