GO-Bayes: Gene Ontology-based overrepresentation analysis using a Bayesian approach

Song Zhang1,*, Jing Cao2, Y. Megan Kong3 and Richard H. Scheuermann1,3,*

1Department of Clinical Sciences, U.T. Southwestern Medical Center, 5323 Harry Hines Boulevard Dallas, TX 75390-9072, 2Department of Statistical Science, Southern Methodist University, Dallas, TX 75275-0332, 3Department of Pathology, U.T. Southwestern Medical Center, 5323 Harry Hines Boulevard Dallas, TX 75390-9072, USA

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on August 4, 2009; revised on December 21, 2009; accepted on February 3, 2010

1 INTRODUCTION

In typical high-throughput experiments, such as gene expression microarrays, the first step in the analysis of the results is often to produce groups of genes based on certain criteria (e.g. genes that are differentially expressed). To gain more mechanistic insights into the underlying biology, overrepresentation analysis (ORA) is often conducted to investigate whether gene sets associated with particular biological functions, for example, as represented by Gene Ontology (GO) annotations, are statistically overrepresented in the identified gene groups. However, the standard ORA, which is based on the hypergeometric test, analyzes each GO term in isolation and does not take into account the dependence structure of the GO-term hierarchy.

Results: We have developed a Bayesian approach (GO-Bayes) to measure overrepresentation of GO terms that incorporates the GO dependence structure by taking into account evidence not only from individual GO terms, but also from their related terms (i.e., parents, children, siblings, etc.). The Bayesian framework borrows information across related GO terms to strengthen the detection of parent-child relationships. As a result, this method tends to identify sets of closely related GO terms rather than isolated GO terms. The advantage of the GO-Bayes approach is demonstrated with a simulation study and an application example.

Contact: song.zhang@utsouthwestern.edu; richard.scheuermann@utsouthwestern.edu

© The Author 2010. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org
Based on Figure 1, a hypergeometric test might be annotated at different specific levels (e.g., gene A exactly the same hypergeometric test in each node, where \( A_i \) are listed under each rectangle. The hypergeometric \( P \)-values are listed in blue and the \( B \)-scores are listed in black underneath the \( P \)-values.

GeneMerge, FuncAssociate, GOTree Machine, etc.) have been developed based on the hypergeometric \( P \)-value. Detailed review can be found in Khatir and Draghici (2005).

The hypergeometric \( P \)-value provides a straightforward measure of overrepresentation for each individual GO term. However, the major drawback of this approach is that it ignores the hierarchical structure in the GO DAG, which contains a substantial amount of information regarding the interactions among the GO terms.

We use an artificial DAG (Fig. 1) to illustrate this issue. It consists of 25 nodes \( A_{ij} \), each denoted by a circle. Let \( G = \{a, b, \ldots, z\} \) denote the full list of genes, and 10 genes in \( G \) are marked in red. The rectangles contain the subset of genes annotated by each node, where \( I(A) \) are listed under each rectangle. Figure 1 includes some important features of the GO DAG, such as multiple inheritance (e.g., \( A_{14} \) and \( A_{23} \) have two parent nodes) and that a gene might be annotated at different specific levels (e.g., gene \( k \) is annotated by \( A_2 \), but not by any of its children: \( A_3, A_4, \) and \( A_5 \)). Note that in this example \( S \) is overrepresented in the regions under \( A_2 \), while underrepresented in the regions under \( A_3 \) and \( A_4 \). We report the hypergeometric \( P \)-values estimated by (1) in blue.

First, given \( I \) and \( J \), GO terms with the same \( I(A) \) have exactly the same hypergeometric \( P \)-values. For example, with \( I(A) = \{2, 3\} \), the \( P \)-values for \( A_2 \) and \( A_3 \) are both 0.139. Based on Figure 1, \( A_{14} \) is more likely to be linked with \( S \) because of the stronger evidence of overrepresentation in its neighboring nodes (related biological functions). Using evidence only from individual terms, the hypergeometric \( P \)-value does not differentiate between \( A_{14} \) and \( A_{24} \). Second, given \( I \) and \( J \), the hypergeometric \( P \)-value has a lower limit determined by \( I \), which is denoted as \( L(I(A)) \). Specifically, \( L(I(A)) \) is achieved when \( n_A = I(A) \), i.e., the \( P \)-value reaches its lower limit when all the genes annotated by \( A \) are in \( S \). The smaller the \( I(A) \), the larger the \( L(I(A)) \). For example, we have \( L(3)=0.046 \) as in \( A_{12} \) and \( L(2)=0.139 \) as in \( A_{14} \). This observation suggests that if we set the threshold for the \( P \)-value at \( L(k) \), then the hypergeometric test could not identify any GO terms with \( I(A) < k \). In ORA, detecting more specific GO terms, which usually have a relatively smaller \( I(A) \), might be more desirable because they provide more detailed biological information. However, the hypergeometric test tends to identify less specific GO terms because of the constraint of \( L(I(A)) \). In Figure 1, the most significant term selected by the \( P \)-value is \( A_{24} \), a term next to the root. The more specific terms that are associated with \( S \) (i.e., \( A_{12} \) and \( A_{14} \)) are considered less significant compared with \( A_{24} \). All of these issues essentially stem from the limitation of the hypergeometric test in treating the GO terms as independent entities and ignoring their interrelated structure.

Recently some new methods have been proposed in ORA. Lewin and Grieve (2006) propose to group closely related GO nodes together and then compute a hypergeometric \( P \)-value for each group. In their approach, the graphical distance between nodes in the GO DAG is assumed to have some quantitative biological meaning. Alexe et al. (2006) calculate the overrepresentation of a GO term from leaf to root by downweighting the contribution of genes which are annotated by a child term that has been found to be significantly enriched. Grossmann et al. (2007) measure the overrepresentation of each GO term relative to its parent(s). From root to leaf, their method computes the significance of overrepresentation for a GO term conditional on the representation at the parent term(s). The last
two methods purposely remove the dependence between parent and child terms.

In this article, we develop a Bayesian hierarchical model to incorporate the dependence structure of the DAG in assessing GO term overrepresentation. It takes into account evidence not only from individual GO terms, but also from their related terms (i.e. parents, children, siblings, etc.). The Bayesian framework enables borrowing information across related GO terms to strengthen the detection of overrepresentation signals. As a result, this method tends to identify sets of closely related GO terms rather than individual unrelated GO terms. The utility of the method is demonstrated using a gene expression microarray dataset from a human B cell stimulation experiment.

2 METHOD

2.1 The Bayesian model

The proposed method is called GO-Bayes: GO-based ORA using a Bayesian approach. In the model, each GO term has a relevance parameter measuring its association with the selected genes in $S$. The novelty of the model is that the complex dependence structure in the GO DAG is incorporated via a hierarchical prior on the relevance parameters.

To introduce GO-Bayes, we define the following notations. Let $A = \{A_1, \ldots, A_J\}$ be the set of GO terms involved in the annotation of the full list of $I$ genes (i.e. $I_A > 0$, for $j = 1, \ldots, J$), among which $I_B$ genes are grouped in $S$ in the experiment. We use $A_j \rightarrow A_k$ to indicate that $A_k$ is a parent of $A_j$. We define $P_j = \{A_{j_1}, \ldots, A_{j_L}\}$ and $C_j = \{A_1, A_2 \rightarrow A_j\}$ to be the sets of parent and child nodes of $A_j$, respectively. We use $|U|$ to denote the cardinality of set $U$. Without loss of generality, let $A_1$ denote the root node and $|P_1| = 0$. For the non-root nodes, we have $|P_j| \geq 1$, where $|P_j| > 1$ indicates multiple inheritance. We call $A_j$ an end node if $|C_j| = 0$ and an inner node otherwise. For example, in Figure 1, $A_5$ is the root node, $P_{A_5} = \{A_1\}$, $C_{A_5} = \{A_2, A_3 \rightarrow A_5\}$ and $|P_{A_5}| = 2$ indicating multiple inheritance.

At the gene level, we use $g_j \in C_j$ to denote that gene $j$ is annotated by $A_j$. We further use $g_j \in A_k$ to indicate that $A_k$ is the most specific GO term that annotates gene $j$, with the formal definition being $g_j \in A_k$ and $g_j \notin A_l$ for any $A_l \notin C_k$. We define $R_k = \{A_{k_1}, \ldots, A_{k_M}\}$ to be the set of the most specific GO terms of gene $j$. The true-path rule implies that $R_k$ contains all the annotation information about gene $j$. In Figure 1, we have $R_1 = \{A_1\}$ and $R_{A_5} = \{A_2, A_3 \rightarrow A_5\}$. Let $y_j = (y_{j1}, \ldots, y_{jL})$ be the observed expression status of gene $j$, $y_{ji} = 1$ if gene $j$ is in $S$ and $y_{ji} = 0$ otherwise.

The binary $(y_{ji})$ is assumed to follow a Bernoulli distribution, $y_{ji} \sim \text{Bernoulli}(\pi_i)$, where $\pi_i$ is the probability that gene $j$ belongs to $S$. Using the idea that if $A_j$ is associated with $S$, the genes annotated by $A_j$ have a higher chance of being grouped in $S$, we construct the following logistic model.

\[
\log \left( \frac{p_{ij}}{1-p_{ij}} \right) = \sum_{k=1}^{L} a_{kj} \cdot \pi_{ik} + c_{ij},
\]

We specify $b_0$ as a constant and set $b_0 = \log(\pi_0/(1 - \pi_0))$ with $\pi_0 = 1/I$, where $p_0$ is the background probability that gene $j$ is grouped in $S$ by chance. The random error $\epsilon_j$ is assumed to have a normal distribution with mean 0 and variance $\sigma_j^2$, denoted by $\epsilon_j \sim \mathcal{N}(0, \sigma_j^2)$. Parameter $\epsilon_j$ characterizes the relevance of GO term $A_j$ to the set of identified genes $S$, where it modifies the odds of gene $j$ being grouped in $S$ by a factor of $e^{\epsilon_j}$ if $g_j \in A_k$. Thus, positive (negative) values of $\epsilon_j$ indicate over (under) representation. Based on the true-path rule, we only include $\epsilon_j$ from $R_k$, the most specific annotations, in model (2) to avoid repeated use of information. The $a_{kj}$'s from less-specific annotations are assumed to affect the odds indirectly via a hierarchical prior on $\mathbf{a} = \{a_{kj} \}_{j=1}^{I} \rightarrow \mathcal{J}$, constructed according to the dependence structure in the GO DAG. We set $a_{kj} = 0$ for the root node. Then the prior of $a_{kj}$ is specified conditionally given the relevance parameters of its parent nodes, denoted by $\mathbf{a}_j = \{a_{kj} : A_k \in P_j\}$. Specifically,

\[
a_{kj} | \mathbf{a}_j, \sigma_j^2 \sim \sum_{i=1}^{G_k} N(a_{kj}, \sigma_j^2),
\]

Prior (3) assumes that $a_{kj}$ arises from a mixture of $G_k$ components, each component being a normal distribution centered at the relevance parameter of one of its parents. The equal mixing probability $1/G_k$ in (3) assumes a prior that each parent has equal influence over $a_{kj}$. Parameter $\sigma_j^2$ characterizes the variability among the children nodes. The joint prior of $\mathbf{a}$ is obtained by the product of (3) over $j = 2, \ldots, J$. This prior provides a mechanism to share information among the GO terms based on the DAG structure. It also naturally accommodates multiple inheritance.

We assign an inverse gamma prior, $\mathcal{IG}(\delta_1, \delta_2)$ on $\sigma_j^2$, which has been used extensively in Bayesian models (Gelman et al., 2003). Similarly, an $\mathcal{IG}(\delta_3, \delta_4)$ prior is assumed for $\sigma_j^2$. We infer the relevance of GO term $A_j$ based on the posterior distribution of $a_{kj}$, denoted by $a_{kj} | Y$. Here, $Y = \{y_{ij}, i = 1, \ldots, I, j = 2, \ldots, J\}$ is the collection of observations. Specifically, we use $\tau_j = \log(\pi_j > 0) Y$ (denoted as the $B$-score) to measure the relevance of a GO term to $S$. It is the posterior probability of $A_j$ being positively associated with $S$. Making inferences based on posterior probabilities is a common practice in Bayesian analysis of microarray data (Newton et al., 2004; Do et al., 2005; Cao et al., 2009). Markov Chain Monte Carlo (MCMC) sampling algorithm is employed to simulate random samples from the joint posterior distribution. We implement the adaptive-rejection sampling method to take advantage of the log-concave property of the full conditional distributions (Gilks and Wild, 1992). The computation can be performed efficiently.

2.2 Demonstration of GO-Bayes with the artificial DAG

GO-Bayes is applied to the artificial DAG (Fig. 1) to detect overrepresentation of the terms. We present the B-score, $\tau_j$, below the hypergeometric $P$-value. In the following discussion, we use association to refer to positive association between a GO term and $S$.

By incorporating the dependence structure of the GO DAG, GO-Bayes shows distinctive advantages over the hypergeometric test. First, GO-Bayes is capable of distinguishing terms with the same $(I_A, n_A)$. For example, GO-Bayes produces $\tau_{A_3} = 0.958$ and $\tau_{A_4} = 0.772$, for $A_3$ and $A_4$, respectively, indicating that $A_4$ is more likely to be associated with $S$ based on the stronger evidence of overrepresentation of its neighboring nodes. Second, the B-score for all nodes has a range of 0–1 regardless of $I_A$. Thus, GO-Bayes can identify more specific GO terms as long as their neighboring nodes are consistently overrepresented. For example, the top two terms selected by GO-Bayes are $A_4$ and $A_5$, respectively, indicating that $A_4$ is more likely to be associated with $S$ based on the stronger evidence of overrepresentation of its neighboring nodes. In Figure 1, the paths with $n_A = 0$ have the same $P$-value of 1.0. Under all the branches of $A_1$, $A_2$, and $A_3$, there are terms with a $P$-value of 1.0. GO-Bayes suggests that it is the terms under $A_3$, whose $\tau_j$'s are closer to 0, that are most underrepresented in $S$.

We have also compared the GO-Bayes B-score with the $\text{elim}$ $P$-value (Alexa et al., 2006) and the parent-child (union) $P$-value (Grossmann et al., 2007) based on the artificial DAG. Due to the space limit, the results are presented in Supplementary Figure 3. In general, both the $\text{elim}$ method and the parent-child method address the ‘dependency problem’ caused by overlapping annotations between parents-child GO terms. The $\text{elim}$ method removes (or downweights) all genes annotated to a significantly enriched node from all its ancestors. By doing this, the method tends to identify strongly overrepresented GO terms that remain significant even after discounting evidence from their offsprings. For illustrative purpose, we set the $P$-value cutoff at 0.05 for the $\text{elim}$ method. Thus, $A_2$ is considered significantly overrepresented and genes $(a, b, c)$ are removed from its ancestor terms $A_1$ and $A_2$. Based on the $\text{elim}$ $P$-value, the term $A_12$ becomes the most significant, where $A_2$ ranks second and $A_1$ ranks sixth. The parent-child method computes the significance of a node conditional on the significance of its parents. It implements the idea by computing a hypergeometric $P$-value for each GO term in the context of its parent (treating the genes annotated by the parent term as the full
3 APPLICATION

3.1 Dataset

In order to test the utility of the GO-Bayes approach, we selected a gene expression microarray dataset in which a B-cell lymphoma cell line (Ramos) was stimulated either through the B-cell antigen receptor (BCR), CD40 or a combination of the two (Basso et al., 2005). Specifically, we used an Affymetrix gene expression dataset selected from the GSE2350 series (GSM44051 to GSM44074) downloaded from the NCBI GEO database (http://www.ncbi.nlm.nih.gov/projects/geo/). The expression values of all six replicates under four experimental conditions were normalized by rows and a list of 3952 differentially expressed genes ($I = 3952$) were selected using the Significance Analysis of Microarray (SAM) approach (Tusher et al., 2001). The differentially expressed genes were subdivided into 20 groups based on their expression pattern by K-means clustering using Euclidean distance as the similarity metric.

Gene Cluster #7 (GC7) was chosen for detailed analysis because it promised to reveal some interesting biology about B-cell responses to receptor signaling. The 196 genes ($I_7 = 196$) present in GC7 showed a particularly interesting expression pattern; these genes were upregulated in response to BCR signaling alone; however, this upregulation was suppressed when CD40 signaling was included (Fig. 2). These treatment conditions mimic important biological responses of immature B cells (Hsu and Scheremann, 2000). Immature B cells must learn to distinguish between signals delivered by authentic pathogen-derived antigens and signals delivered 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 a state of unresponsiveness or apoptotic cell death. The two-signal hypothesis is one mechanism proposed to elicit either responsiveness or non-responsiveness to antigen exposure, which states that B cells receiving only one signal, through the BCR, will proliferate and die, but B cells receiving two signals both through the BCR and a co-stimulatory receptor-like CD40 will proliferate and survive due to suppression of the cell death response. Thus, gene present in GC7 are those genes whose expression is suppressed with the addition of CD40 signaling and could thus be involved either in the cell death response or in the induction of unresponsiveness.

3.2 Result

For the full list of 3952 differentially expressed genes ($I = 3952$), 1968 GO terms ($I = 1968$) have been used to annotate their functions. Four groups of GO terms are found in the top 20 list of most significant terms based on the GO-Bayes approach (Table 1). The GO term with the highest $B$-score (0.9004) is 'transferrase activity, transferring hexosyl groups', all nine related terms (GO:0016758). Nine of the top 20 GO terms based on the $B$-score are related to this group. Five of the top 20 GO terms are related to 'G-protein signaling, coupled to cAMP nucleotide second messenger' (GO:0007188); six are related to 'lysosome' (GO:0005764); and two are in the group of 'negative regulation of transcription, DNA dependent' (GO:0045892). In the case of the 'transferrase activity, transferring hexosyl groups', all nine related terms (GO:0016758, GO:0016757, GO:0000330, GO:0003844, GO:0008375, GO:015020, GO:0042328, GO:0004703, GO:0045892) can be found near GO:0016758 in the GO hierarchy (see Supplementary Fig. 4). Even though some of these terms are poorly represented in this dataset (e.g. GO:0003844 represented by only a single gene in the dataset), all of these terms have relatively high $B$-scores due to the overrepresentation of other terms that are close relatives in the hierarchy.

In some cases, the association of GO terms highlighted by the GO-Bayes approach matches our expectations.
### Table 1. The top 20 lists GO terms associated with Gene Cluster #7 by CLASSIFI and GO-Bayes

<table>
<thead>
<tr>
<th>GO ID</th>
<th>GO ID</th>
<th>IA</th>
<th>nA</th>
<th>P-value</th>
<th>RankP</th>
<th>B-score</th>
<th>G10</th>
<th>RankG</th>
<th>Group</th>
<th>GO name</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0019933</td>
<td>GO:0005773</td>
<td>21</td>
<td>6</td>
<td>4.00E-04</td>
<td>1</td>
<td>0.7652</td>
<td>149</td>
<td></td>
<td>a</td>
<td>cAMP-mediated signaling</td>
</tr>
<tr>
<td>GO:0001726</td>
<td>GO:0007190</td>
<td>9</td>
<td>13</td>
<td>4.61E-04</td>
<td>2</td>
<td>0.7544</td>
<td>187</td>
<td></td>
<td>b</td>
<td>Vacuole</td>
</tr>
<tr>
<td>GO:0007188</td>
<td>GO:0045762</td>
<td>10</td>
<td>4</td>
<td>9.37E-04</td>
<td>5</td>
<td>0.6928</td>
<td>508</td>
<td></td>
<td>a</td>
<td>Activation of adenylate cyclase activity</td>
</tr>
<tr>
<td>GO:0003323</td>
<td>GO:0004786</td>
<td>10</td>
<td>11</td>
<td>1.74E-03</td>
<td>12</td>
<td>0.8762</td>
<td>5</td>
<td></td>
<td>b</td>
<td>Lytic vacuole</td>
</tr>
<tr>
<td>GO:0010324</td>
<td>GO:0006897</td>
<td>101</td>
<td>7</td>
<td>2.40E-03</td>
<td>4</td>
<td>0.5804</td>
<td>2957</td>
<td></td>
<td>a</td>
<td>G-protein-coupled receptor protein signaling</td>
</tr>
<tr>
<td>GO:0007187</td>
<td>GO:0032320</td>
<td>10</td>
<td>9</td>
<td>1.87E-03</td>
<td>14</td>
<td>0.5996</td>
<td>205</td>
<td></td>
<td>b</td>
<td>Membrane invagination</td>
</tr>
<tr>
<td>GO:0005764</td>
<td>GO:0031281</td>
<td>101</td>
<td>7</td>
<td>2.40E-03</td>
<td>4</td>
<td>0.5804</td>
<td>2957</td>
<td></td>
<td>a</td>
<td>G-protein-coupled receptor protein signaling</td>
</tr>
<tr>
<td>GO:0006898</td>
<td>GO:0001669</td>
<td>10</td>
<td>9</td>
<td>1.87E-03</td>
<td>14</td>
<td>0.5996</td>
<td>205</td>
<td></td>
<td>b</td>
<td>Membrane invagination</td>
</tr>
<tr>
<td>GO:0004786</td>
<td>GO:0001669</td>
<td>10</td>
<td>9</td>
<td>1.87E-03</td>
<td>14</td>
<td>0.5996</td>
<td>205</td>
<td></td>
<td>b</td>
<td>Membrane invagination</td>
</tr>
<tr>
<td>GO:0007187</td>
<td>GO:0045762</td>
<td>10</td>
<td>4</td>
<td>9.37E-04</td>
<td>5</td>
<td>0.6928</td>
<td>508</td>
<td></td>
<td>a</td>
<td>Activation of adenylate cyclase activity</td>
</tr>
<tr>
<td>GO:0003323</td>
<td>GO:0004786</td>
<td>10</td>
<td>11</td>
<td>1.74E-03</td>
<td>12</td>
<td>0.8762</td>
<td>5</td>
<td></td>
<td>b</td>
<td>Lytic vacuole</td>
</tr>
<tr>
<td>GO:0010324</td>
<td>GO:0006897</td>
<td>101</td>
<td>7</td>
<td>2.40E-03</td>
<td>4</td>
<td>0.5804</td>
<td>2957</td>
<td></td>
<td>a</td>
<td>G-protein-coupled receptor protein signaling</td>
</tr>
<tr>
<td>GO:0006898</td>
<td>GO:0001669</td>
<td>10</td>
<td>9</td>
<td>1.87E-03</td>
<td>14</td>
<td>0.5996</td>
<td>205</td>
<td></td>
<td>b</td>
<td>Membrane invagination</td>
</tr>
<tr>
<td>GO:0007187</td>
<td>GO:0045762</td>
<td>10</td>
<td>4</td>
<td>9.37E-04</td>
<td>5</td>
<td>0.6928</td>
<td>508</td>
<td></td>
<td>a</td>
<td>Activation of adenylate cyclase activity</td>
</tr>
<tr>
<td>GO:0003323</td>
<td>GO:0004786</td>
<td>10</td>
<td>11</td>
<td>1.74E-03</td>
<td>12</td>
<td>0.8762</td>
<td>5</td>
<td></td>
<td>b</td>
<td>Lytic vacuole</td>
</tr>
<tr>
<td>GO:0010324</td>
<td>GO:0006897</td>
<td>101</td>
<td>7</td>
<td>2.40E-03</td>
<td>4</td>
<td>0.5804</td>
<td>2957</td>
<td></td>
<td>a</td>
<td>G-protein-coupled receptor protein signaling</td>
</tr>
<tr>
<td>GO:0006898</td>
<td>GO:0001669</td>
<td>10</td>
<td>9</td>
<td>1.87E-03</td>
<td>14</td>
<td>0.5996</td>
<td>205</td>
<td></td>
<td>b</td>
<td>Membrane invagination</td>
</tr>
<tr>
<td>GO:0007187</td>
<td>GO:0045762</td>
<td>10</td>
<td>4</td>
<td>9.37E-04</td>
<td>5</td>
<td>0.6928</td>
<td>508</td>
<td></td>
<td>a</td>
<td>Activation of adenylate cyclase activity</td>
</tr>
<tr>
<td>GO:0003323</td>
<td>GO:0004786</td>
<td>10</td>
<td>11</td>
<td>1.74E-03</td>
<td>12</td>
<td>0.8762</td>
<td>5</td>
<td></td>
<td>b</td>
<td>Lytic vacuole</td>
</tr>
<tr>
<td>GO:0010324</td>
<td>GO:0006897</td>
<td>101</td>
<td>7</td>
<td>2.40E-03</td>
<td>4</td>
<td>0.5804</td>
<td>2957</td>
<td></td>
<td>a</td>
<td>G-protein-coupled receptor protein signaling</td>
</tr>
<tr>
<td>GO:0006898</td>
<td>GO:0001669</td>
<td>10</td>
<td>9</td>
<td>1.87E-03</td>
<td>14</td>
<td>0.5996</td>
<td>205</td>
<td></td>
<td>b</td>
<td>Membrane invagination</td>
</tr>
<tr>
<td>GO:0007187</td>
<td>GO:0045762</td>
<td>10</td>
<td>4</td>
<td>9.37E-04</td>
<td>5</td>
<td>0.6928</td>
<td>508</td>
<td></td>
<td>a</td>
<td>Activation of adenylate cyclase activity</td>
</tr>
<tr>
<td>GO:0003323</td>
<td>GO:0004786</td>
<td>10</td>
<td>11</td>
<td>1.74E-03</td>
<td>12</td>
<td>0.8762</td>
<td>5</td>
<td></td>
<td>b</td>
<td>Lytic vacuole</td>
</tr>
<tr>
<td>GO:0010324</td>
<td>GO:0006897</td>
<td>101</td>
<td>7</td>
<td>2.40E-03</td>
<td>4</td>
<td>0.5804</td>
<td>2957</td>
<td></td>
<td>a</td>
<td>G-protein-coupled receptor protein signaling</td>
</tr>
<tr>
<td>GO:0006898</td>
<td>GO:0001669</td>
<td>10</td>
<td>9</td>
<td>1.87E-03</td>
<td>14</td>
<td>0.5996</td>
<td>205</td>
<td></td>
<td>b</td>
<td>Membrane invagination</td>
</tr>
</tbody>
</table>

The columns IA and nA represent the cardinality of the relevant GO term and the number of genes in the GO term that also appear in Gene Cluster #7; P-value represents the hypergeometric P-value computed by CLASSIFI, and the rank of the GO terms based on the P-value; and B-score and RankB represent the GO-Bayes measure and the corresponding rank of the GO terms. Under the column Group, ‘a’ represents GO terms closely related to GO:0007188 (G-protein signaling, coupled to cAMP nucleotide second messenger) in the GO hierarchy; ‘b’ represents GO terms closely related to GO:0005764 (lysosome); ‘c’ represents GO terms closely related to GO:0016757 (transferase activity, transferring glycosyl groups); ‘d’ represents GO terms closely related to GO:0045892 (negative regulation of transcription, DNA-dependent); and ‘e’ represents GO terms unrelated to the above four groups. NA, not applicable.

It is well known that activation of B cells through the BCR induces endocytosis and the fusion of endocytic vesicles with lysosomes as a mechanism to capture antigen for presentation to T cells in order to stimulate the helper immune response (Lee et al., 2006). Thus, the presence of lysosome-related GO terms in the cluster of genes upregulated in response to BRC stimulation might be expected. In other cases, the association is not immediately expected, but subsequent investigations revealed that the association is supported by previous experiment data. For example, the majority of genes giving rise to the ‘negative regulation of transcription, DNA-dependent’ association, including ID3 (Pan et al., 1999), CTCF (Qi et al., 2003), SMAD3 (Ramesh et al., 2000), KLF12 (Roth et al., 2000), E2F6 (Xu et al., 2007), FoxS1 (Yin et al., 2007) and PA2G4 (Zhang et al., 2008), have been found to be upregulated in response to BCR stimulation in B cells or somehow involved in B-cell signaling responses. In the case of CTCF, Qi et al. (2003) showed that this upregulation is associated with the induction of apoptosis in B cells and can be suppressed by co-stimulation through
CD4 in agreement with the findings reported here. In still other cases, no direct corroborative evidence could be found (e.g. for ‘transferase activity, transferring hexosyl groups’). Thus, this finding serves as a hypothesis for future testing.

In order to compare the results of the GO-Bayes approach with the standard ORA based on the hypergeometric test, we processed the gene set in GC7 with the CLASSIFI algorithm (Lee et al., 2006) and selected the top 20 GO terms. The top 20 lists by CLASSIFI and GO-Bayes, respectively, are presented in Table 1. While four GO terms were ranked in the top 20 by both methods, the remaining top 20 terms from each method were distinct. Based on this comparison, several distinctions between the two approaches can be made. First, three of the four GO term groups are found in both top 20 term lists. Two of the groups, G protein signaling and lysosome, have multiple terms in both top 20 lists. One group, transferase activity, is only represented once in the CLASSIFI list, but multiple times in the GO-Bayes list. Given that little could be found in the literature about these genes in B cell biology, this association would likely be ignored from the hypergeometric analysis. One group, the negative regulation of transcription group was not found in the hypergeometric top 20, and yet is possibly the most interesting given the strong literature support described above.

Second, eight GO terms in GC7 with (\(\delta_{j,n} > 2,2\)) have the same hypergeometric P-value of 0.0024, and all of them are in the hypergeometric top 20. The GO-Bayes measure suggests that they are very different in their association with the cluster. For example, GO:0007217 (tachykinin signaling pathway) has a B-score of 0.8452 and it ranks in the GO-Bayes top 20. In contrast, GO:0016519 (gastric inhibitory peptide receptor activity) has a B-score of 0.1972 and its GO-Bayes rank is 6600. To shed light on the difference in the B-scores between these two terms, we compare their regional DAGs, which are shown in Supplementary Figures 5 and 6. The three direct ancestors of GO:0007217 have a stronger association with the cluster than those of GO:0016519. In addition, GO:0007217 has nine siblings, five of which have genes represented in the cluster (\(n_{S} > 0\)). By comparison, GO:0016519 has 13 siblings, of which only one sibling has genes represented in the cluster. The support from the related GO terms is substantially higher for GO:0007217 than for GO:0016519, and thus GO:0007217 is judged to be more likely associated with the cluster based on the GO-Bayes measure. While there is no evidence for the involvement of the gastric inhibitory peptide receptor activity in the regulation of B-cell function in the literature, tachykinin (also known as hemokinin-1) has been found to be secreted during the differentiation of B-cell precursors (Kemp et al., 2004). Thus, GO:0007217 does appear to be biologically associated with GC7.

Among the top 20 GO terms identified by GO-Bayes, 8 GO terms have two genes or fewer belonging to GC7 (GO:0007217, GO:0004703, GO:0000030, GO:0015020, GO:0003844, GO:42328 and GO:0008375) are all related to ‘transferase activity’, and all of them are in a close neighborhood in the GO DAG as shown in Supplementary Figure 4. As for GO:0042629, the fact that none of the two annotated genes were found in GC7 requires in-depth discussion on the term. Note that GO:0042629 is a direct child of ‘lysosome’ (GO:0005764), which ranked fifth in the top 20 GO-Bayes list. The two genes annotated with GO:0042629 are IL8 receptor beta (IL8RB) and serglycin (SRGN). IL8RB has been found to be involved in B-cell chemotaxis across the blood-brain barrier (Alter et al., 2003). Mice deficient for IL8RB show lymphadenopathy due to a specific expansion of the B-cell compartment (Cacalano et al., 1994). Importantly, IL8RB was found to be dramatically upregulated (+9-fold) in response to BCR stimulation of primary splenic B cells by microarray analysis (Sato et al., 2005) (see IL8RB expression profile in http://www.ncbi.nlm.nih.gov/sites/GDSBrowser?acc=GDS1467).

Although less is known about the function of SRGN in B cells, it was also found to be upregulated following BCR stimulation of B cells from TAK1-deficient mice in the same microarray experiment. Taken together, these findings suggest that the identification of GO:0042629 by GO-Bayes is significant and represents an example where the GO-Bayes approach was able to overcome false negative results generated by deficiencies in microarray data generation or processing. Thus, GO-Bayes will only identify a GO term with relatively few annotated genes belonging to S when there is strong evidence of overrepresentation from its neighboring GO terms.

4 DISCUSSION
In this article, we proposed GO-Bayes, a Bayesian approach for GO-based OR. The model has a relatively simple format, with the first level as a logistic regression model and straightforward prior specification. The key innovation is that it can incorporate the dependence structure of the GO DAG on a global scale. The resulting measure on the association between GO terms and selected genes borrows information across related GO terms to strengthen the detection of overrepresentation signals. We have given detailed comparison between the GO-Bayes approach and the hypergeometric test, which is widely used in ORA. Our analysis using an artificial dataset and a real microarray dataset suggests that the GO-Bayes approach can produce more biologically meaningful results than the hypergeometric test. Relying on individual GO terms, the hypergeometric P-value has a closed form, which only requires the information on I, IS and (\(\delta_{j,n} \leq 0\)) for each GO term. GO-Bayes, on the other hand, does not have a closed form and it needs the information on the structure of the GO DAG. We have developed a program in C to implement the GO-Bayes approach. For the B-cell lymphoma cell line example in the application section, the B-score calculations were completed in ~10 min on a MAC (OS X 10.4.11, 1.83 GHz Intel Core Duo processor) computer. The program is available upon request from S.Z.

The GO-Bayes measure, \(P_{i} = P(A_{i} > 0 | Y)\), is the posterior probability of a GO term being positively associated with S. As a reviewer pointed out, \(P_{i}\) behaves like a frequentist P-value under the null hypothesis (Bochkina and Richardson, 2007), and a frequentist-type estimator of false discovery rate (Storey, 2002) can be employed to assess the statistical significance.

Currently, the proposed method is based on binary outcomes (membership of genes in S). We are working on generalizing the GO-Bayes approach to cases with ordinal or continuous outcomes. Although we have focused on the use of this approach...
for the interpretation of gene expression microarray data and GO annotation, the general strategy can be applied to any circumstance in which groups of entities are annotated with terms derived from any ontology hierarchy or similar dependency structure.

ACKNOWLEDGEMENTS

The authors thank the three reviewers for their constructive comments and suggestions.

Funding: U.S. National Institutes of Health (N01AI40076 and UL1 RR024982, in part).

Conflict of Interest: none declared.

REFERENCES


ACKNOWLEDGEMENTS

The authors thank the three reviewers for their constructive comments and suggestions.

Funding: U.S. National Institutes of Health (N01AI40076 and UL1 RR024982, in part).

Conflict of Interest: none declared.

REFERENCES


ACKNOWLEDGEMENTS

The authors thank the three reviewers for their constructive comments and suggestions.

Funding: U.S. National Institutes of Health (N01AI40076 and UL1 RR024982, in part).

Conflict of Interest: none declared.

REFERENCES


ACKNOWLEDGEMENTS

The authors thank the three reviewers for their constructive comments and suggestions.

Funding: U.S. National Institutes of Health (N01AI40076 and UL1 RR024982, in part).

Conflict of Interest: none declared.

REFERENCES


ACKNOWLEDGEMENTS

The authors thank the three reviewers for their constructive comments and suggestions.

Funding: U.S. National Institutes of Health (N01AI40076 and UL1 RR024982, in part).

Conflict of Interest: none declared.

REFERENCES


ACKNOWLEDGEMENTS

The authors thank the three reviewers for their constructive comments and suggestions.

Funding: U.S. National Institutes of Health (N01AI40076 and UL1 RR024982, in part).

Conflict of Interest: none declared.