Potential Bias in GO::TermFinder

Robert M. Flight and Peter D. Wentzell

Submitted: 8th October 2008; Received (in revised form): 7th November 2008

Abstract
The increased need for multiple statistical comparisons under conditions of non-independence in bioinformatics applications, such as DNA microarray data analysis, has led to the development of alternatives to the conventional Bonferroni correction for adjusting \( P \)-values. The use of the false discovery rate (FDR), in particular, has grown considerably. However, the calculation of the FDR frequently depends on drawing random samples from a population, and inappropriate sampling will result in a bias in the calculated FDR. In this work, we demonstrate a bias due to incorrect random sampling in the widely used GO::TermFinder package. Both \( T^2 \) and permutation tests are used to confirm the bias for a test set of data, which leads to an overestimation of the FDR of about 10%. A simple fix to the random sampling method is proposed to remove the bias.

Keywords: false discovery rate; bias; gene ontology; GO::Termfinder; enrichment

INTRODUCTION
Statistical calculations are an essential part of systems biology and the veracity of those calculations is critical to the interpretation of results. Statistical tests are widely employed to evaluate the significance of observations, for example in the assessment of differential gene expression. Traditionally, statistical hypothesis testing uses the comparison of a test-statistic to a critical value, or alternatively, the calculation of a \( P \)-value for the corresponding test statistic. For cases of multiple comparisons, the Bonferroni correction is typically applied, but this is generally perceived to be too conservative for assessing differential expression in microarray studies [1]. An alternative approach in recent years has been the calculation of the false discovery rate (FDR), introduced into the broader microarray community through the well-known SAM (significance analysis of microarrays) approach [2]. The FDR provides a more reasonable estimate of the proportion of false positives present in a discovery set [3]. As the phenomenon of multiple testing has become more widespread, the use of an FDR value to supplement or replace \( P \)-values has become more common. Frequently, however, calculation of the FDR depends on taking multiple random subsets of genes from a background population and comparing the \( P \)-values obtained to the original \( P \)-values based on a statistical model. If the method of selecting the random samples is flawed, there exists the possibility of introducing a bias into the calculation of the FDR. This effect has been observed by others in different contexts [4, 5].

A popular resource for providing controlled annotations of gene products from many different organisms is provided by the Gene Ontology (GO) Consortium [6]. The GO annotations are frequently used to summarize commonality of function within groups of genes. Gene annotations are provided by various groups for the different organisms, and the combination of a file describing the GO structure and annotation files has allowed many different software packages to evolve, permitting researchers to examine their gene lists for enrichment of GO terms common to genes in the list. A popular software package for evaluating the statistical significance of GO terms represented in a set of genes extracted from a population is GO::TermFinder [7],
GO::TERMFINDER provides three options for evaluating the significance of terms: (i) Bonferroni-corrected P-values, (ii) simulation-corrected P-values and (iii) FDR estimates. Option (i) is generally regarded as too conservative and inappropriate in the case of the GO due to its hierarchical nature. Option (ii) uses 1000 randomly sampled gene lists, which is time-consuming, but should provide more accurate results than Bonferroni correction. Option (iii) estimates the FDR using 50 randomly generated gene lists. Both of the last two options assume that the selection of genes is random, and deviations from this can lead to bias in the estimated statistical parameters. In the results presented here, we demonstrate a significant bias in the FDR values calculated by GO::TERMFINDER. This bias could be quite large in certain circumstances and therefore its recognition and correction in this widely used software is important. We propose a simple change in the software source to remedy this problem.

BACKGROUND

The calculation of the FDR in GO::TERMFINDER has been described in the literature [7], but is briefly reviewed here. The goal of the calculation is to determine whether a subset of genes containing $N_{\text{gene}}$ entries, referred to here as the gene list, has significant enrichment of genes for a given GO term relative to a larger list of $N_{\text{total}}$ genes, referred to here as the background list, from which it was extracted. The gene list is typically obtained as the result of association in some biological study, such as a DNA microarray experiment. The null hypothesis is that the frequency of representation of a particular GO term in the gene list is no different from that in the background list.

The first step in the calculation for a given gene list is an assessment of conventional P-values that relate to the significance of each GO term represented by the genes in the gene list. In other words, for a given term, $t_i$ (1 ≤ $i$ ≤ $N_{\text{term}}$, where $N_{\text{term}}$ is the number of GO terms represented in the gene list), the value of $P_i$ is the estimated probability that the frequency of the term observed in the gene list is the same as the frequency of that term in the background list. This calculation is typically carried out using a hypergeometric distribution [7] and may or may not involve a Bonferroni correction. The list of terms is normally sorted in the order of decreasing significance and entries with a P-value above a certain cutoff are generally removed. The FDR calculation is based on repetitive random resampling of $N_{\text{gene}}$ genes from the background gene list. The resampling is repeated $r$ times and for each repeat, $j$, the P-values are calculated for all of the terms relevant for the new gene list. The FDR for each term in the original gene list is determined by enumerating the number of times terms with the same or smaller P-values were obtained in the resampled data. If we let $n_{ij}$ equal the number of terms found with $P_j ≤ P_i$ in replicate $j$ of the resampled data and $n_i$ equal the number of terms in the original list with $P_j ≤ P_i$, then the FDR is calculated as:

$$FDR_i = \frac{1}{n_i} \sum_{j=1}^{r} n_{ij}$$

(1)

The FDR is the average number of times a P-value less than or equal to $P_i$ appears in the random simulations, divided by the number of P-values in the original list that are less than or equal to $P_i$. The FDR estimates the fraction of false positives that will be obtained at a given P-value in the extracted list, and therefore can guide the investigator in assessing the true significance of terms extracted.

An important assumption of any resampling strategy is that it is being done randomly and there is no correlation among the genes that are selected. In order to perform this resampling, GO::TERMFINDER uses the rand command in PERL to randomly generate the indices of the genes to be selected in a serial fashion. If an index has already been selected for the current list, the algorithm increments the value of the index until one is found that is not on the list. The specific code is given below:

```perl
while (exists ($usedIndices{$n})){
    $n++;
    # wrap round to the beginning if we run off the end of the arrays from which we’re choosing
    $n = 0 if ($n >= $populationSize);
}
```
Prior to executing this sequence, the variable \( n \) has been set to a random integer between 1 and the number of genes in the background list. It should be apparent that this strategy will favor the selection of genes that are adjacent to one another within the list. If the arrangement of the list is such that function shows a correlation with position in the list, the potential exists for bias.

A calculation of FDRs has been included in software being developed in our laboratory. As part of the validation of this software, FDRs calculated using our software were compared to those obtained by GO::TERMINDER and results for 20 replicate runs for a given set of 237 genes (see Methods section) are shown in Figure 1. The error bars in the figure represent 1 SD (population) around the mean values and the line represents a slope of unity (perfect agreement). The figure suggests that there was a difference between the two methods and that the FDRs calculated by our in-house software were low compared to those provided by the GO::TERMINDER. This prompted the investigation which is described here.

**METHODS**

**Software**

Four different methods were used to evaluate the FDR in this study. Method 1 (GO1) was the original GO::TERMINDER (v. 0.81), which was written in PERL and run in a Windows XP® environment. The GO obo file and all organism association files were downloaded from the GO website [14] on 15 May 2008. Method 2 (GO2) consisted of a simple modification to the original code such that the sequence of commands in the preceding section was replaced with:

```perl
while (exists ($usedIndices{$n})){
    $n = int(rand($populationSize));
    # take another random number until unused one is found
}
```

This code was anticipated to remove any positional correlation in the extraction of random gene lists. (It should be noted that a more computationally efficient and unbiased PERL implementation was suggested by one of the reviewers of this article, but this has not been included here because it involves a more extensive revision of the original code.) Method 3 (MAT1) implemented the FDR calculations on a MATLAB® (v. 7.0, The MathWorks, Natick, MA, USA) platform using code written in our laboratory to emulate the random sampling method in GO1. Finally, Method 4 (MAT2) used a MATLAB® algorithm that employed the ‘randperm’ command to generate indices for the random lists, so, like the GO2 algorithm, it was not anticipated to exhibit a positional dependence. All four methods used 50 replicate resamplings in order to calculate the FDR in a single function call, and all algorithms were run in a Windows XP® environment.

**Data**

As a generic test case, the gene list used here was obtained simply through a search of the Saccharomyces Genome Database [8] for genes annotated to the biological process ‘transcription’. This yielded 237 genes from a background list of 6348 genes. In this case, the background list was sorted by gene name.

**Statistical analysis**

The statistical evaluation of bias among the different methods of FDR calculation is complicated by several factors. First, as evident in Figure 1, the variance in the calculated FDRs increases with the magnitude of the FDR, with the SD closely following a square root dependence over the useful range.

![Figure 1: Comparison of mean FDRs calculated from GO::TERMINDER (GO1) and in-house software (MAT2) using 20 replicates. Error bars represent 1 SD from the mean. The line has a slope of unity representing perfect agreement.](image-url)
This needs to be taken into account in any pairwise or regression-based comparison. Second, although not evident from Figure 1, the deviations of FDR values from the mean for a given run are correlated. This is a consequence of the fact that all of the calculations for a given iteration of the FDR computation are based on the same sampling of the background genes, leading to a relationship among the $P$-values. For example, if a particular sampling of the background genes leads to a fortuitous enrichment of one term, others are likely to be enriched due to the hierarchical nature of the ontology. This covariance adds additional complications to the statistical testing.

Several approaches can be used to develop a test statistic for such a data set. The calculation of a $T^2$-statistic for paired comparisons of multivariate data can be employed and referred to an $F$ or $\chi^2$ distribution [15, 16]. Alternatively, data from multiple variables can be pooled for use with a standard $t$-test, as long as they are appropriately standardized to account for non-uniform variance and the presence of covariance. Permutation testing, based on a variety of parameters, can also be used. All three of these approaches were used in this work and led to the same conclusions, but only the $T^2$ and permutation tests are described here, the former because it is fairly standard and sensitive, the latter because it is quite robust and not dependent on distributional assumptions.

For all statistical tests, the pairwise comparison of methods for FDR calculation was initiated as follows. First, let $X$ and $Y$ represent the results of FDR calculations for two methods applied to the same data set. The dimensions of these matrices are $N_{\text{term}} \times N_{\text{rep}}$, where $N_{\text{term}}$ represents the number of terms ($P$-values) in the original gene list being analyzed and $N_{\text{rep}}$ is the number of runs carried out for the statistical comparison. It should be noted that $N_{\text{rep}}$, which was set to 1000 in this work, is to be differentiated from $r$, described earlier as the number of replicate resamplings used in a single calculation of FDR values (set to 50 in this work). Also, for the purposes of comparison, only those rows for which the mean FDR was between 0.001 and 0.2 for both $X$ and $Y$ were retained in the analysis. This was considered to be the region of greatest interest in FDR calculations. Moreover, very small and very large FDR values are less reliable because the number of $P$-values counted, $n_i$, approaches zero or $n_i$, respectively. In this study, this restriction typically left about 62 terms out of the original 325, making the dimensions of $X$ and $Y$ $62 \times 1000$.

The first step in the statistical analysis is the calculation of a difference matrix, $D$:

$$D = X - Y \quad (2)$$

Under the null hypothesis that the two methods of FDR calculation are the same, these differences should be statistically equal to zero. Several methods were used to test this hypothesis.

For the $T^2$-test, the row averages of $D$ are calculated to give the vector $\bar{d}$ ($N_{\text{term}} \times 1$), where,

$$\bar{d}_i = \frac{1}{N_{\text{rep}}} \sum_{j=1}^{N_{\text{rep}}} d_{ij} \quad (3)$$

The covariance among the rows, $S_d$, is estimated as:

$$S_d = \frac{1}{N_{\text{rep}}} - 1 \sum_{j=1}^{N_{\text{rep}}} (d_j - \bar{d})(d_j - \bar{d})^T \quad (4)$$

where $d_j$ represents the $j$-th column of $D$. The $T^2$-value is calculated as:

$$T^2 = N_{\text{rep}} \bar{d}^T S_d^{-1} \bar{d} \quad (5)$$

Under the conditions of the null hypothesis, this statistic should be distributed as:

$$\left( \frac{N_{\text{rep}} - N_{\text{term}}}{N_{\text{rep}} - 1} \right) N_{\text{term}} T^2 \sim F_{N_{\text{term}}(N_{\text{rep}} - N_{\text{term}})} \quad (6)$$

Although this method should work in principle, a complication arises in the fact that the covariance matrix, $S_d$, is poorly conditioned (rank deficient) in some cases, so its inversion is unreliable. To remove this problem, $D$ was first decomposed by singular value decomposition and the loadings were truncated to a rank of 40 (a stable value), yielding a $40 \times 1000$ matrix of loadings, $V$, that were treated in the same way as the original $D$ matrix. For this calculation, $N_{\text{term}}$ was set to 40 for the appropriate degrees of freedom. Although the selection of 40 loading vectors was somewhat arbitrary, results were not substantially altered with other values which gave a favorable condition number.

In addition to the $T^2$-test, a permutation test, which is less dependent on distributional assumptions, was also carried out. Following the calculation of the mean vector, $\bar{d}$, the SDs of the rows are calculated as the vector $s_d$, where:

$$s_d = \sqrt{\frac{1}{(N_{\text{rep}} - 1)} \sum_{j=1}^{N_{\text{rep}}} (d_{ij} - \bar{d}_i)^2} \quad (7)$$
The vector of standard normal deviates for the mean values, \( z \), is then calculated in the usual way:

\[
  z_i = \frac{d_i \sqrt{N_{rep}}}{s_d}
\]  

(8)

This calculation assumes that the values within each row of \( D \) are independent, which should be a valid assumption. Since the row means have been standardized, they can, in principle, be combined to give an overall \( z \)-value:

\[
  z_{overall} = \frac{\bar{z} \sqrt{N_{term}}}{s_z}
\]  

(9)

where \( \bar{z} \) and \( s_z \) are the mean and SDs of the values in \( z \), respectively. In fact, because of the correlations that exist down the rows, the value of \( z_{overall} \) estimated in this way will be high compared to what it should be for independent measurements, but this is not critical for the permutation analysis.

In the permutation analysis, the procedure described above is repeated on data sets which have been permuted in such a way that the two methods are randomly assigned to each of the 2000 data sets available, with one method assigned to 1000 of these and the other assigned to the remaining 1000. The calculated values of \( z_{overall} \) then give a null distribution to which the original \( z_{overall} \) is compared, resulting in a \( P \)-value for a two-tailed comparison. This is calculated as the fraction of \( z \)-values that have an absolute value greater than that of the original comparison. For this work, 1000 random permutations were used.

**RESULTS AND DISCUSSION**

Table 1 summarizes the results of the pairwise comparisons of methods in the form of \( P \)-values for the two types of statistical tests carried out. Perhaps most important to note is the different results obtained with GO1 and GO2, suggesting a bias in the manner of random gene selection in the original algorithm. Figure 2 shows the distribution obtained in the permutation analysis for this comparison. Since the calculation was carried out such that the results of GO2 were subtracted from those of GO1, this shows that GO1 overestimates the FDR values, as anticipated from the manner of random sampling. Among the other comparisons, MAT1 should be equivalent to GO1 and MAT2 should be equivalent to GO2, which is consistent with the results shown in the table. Also shown for validation are self–self comparisons, where each set in the comparison was based on 1000 independent replicates. In all of these cases, no significant differences were detected. As anticipated, the results in Table 1 show that the \( T^2 \)-test is more sensitive to the differences than the permutation test.

**Table 1:** Statistical results of pairwise comparisons of methods for FDR calculations

<table>
<thead>
<tr>
<th></th>
<th>GO1</th>
<th>GO2</th>
<th>MAT1</th>
<th>MAT2</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO1</td>
<td>0.503</td>
<td>0.430</td>
<td>0.791</td>
<td>0.028</td>
</tr>
<tr>
<td>GO2</td>
<td>0.143</td>
<td>0.968</td>
<td>0.028</td>
<td>0.469</td>
</tr>
<tr>
<td>MAT1</td>
<td>0.429</td>
<td>0.980</td>
<td>0.041</td>
<td>0.920</td>
</tr>
<tr>
<td>MAT2</td>
<td>0.920</td>
<td>0.657</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The numbers given represent \( P \)-values for the pairwise comparison of the methods in the corresponding rows and columns. The upper value in each cell is for the \( T^2 \)-test and the lower value is for the permutation test.
While these results do not prove that the FDR calculations with the new methods (GO2 and MAT2) are correct (in the absence of a ‘true’ value), given the fairly obvious bias in random selection used in the original algorithm, it is highly likely that they are at least more correct. The only difference in the methods is the manner of random sampling, so the simplest and most logical conclusion is that the original algorithm is flawed.

The principal conclusion of this work is that the method of FDR calculation used in GO::TERM_FINDER v. 0.81 can lead to bias in some cases, resulting in an over-estimation of the FDR. The bias is likely to be largest when the percentage of genes in the gene list relative to the background list is high (since this increases the likelihood of duplicate random selections) and when the background list is sorted in a manner that correlates position in the list to gene function. The latter condition is most likely to be observed when background lists are sorted by gene name, but could also occur in other circumstances. This bias will lead to a decrease in the number of ontology terms retained at a particular cutoff level and therefore increase the probability of a Type II error, excluding terms that may be significant. The extent of this problem will depend on the conditions of the calculation, but in practical terms, it is not likely to have a significant impact on calculations carried out to date with GO::TERM_FINDER or related software, resulting only in a slight adjustment of FDRs obtained. In this work, it was found that FDRs in the critical region were over-estimated by a factor of about 10% with the current software. From a more fundamental perspective, however, the veracity of statistical calculations is important for those who rely on them. This is especially true where new methods are validated against existing standards as in this case. The authors of the GO::TERM_FINDER software have been made aware of this difficulty.

Acknowledgements
The authors would like to thank Dr David Hamilton for statistical advice.

FUNDING
Natural Sciences and Engineering Research Council of Canada.

References