GO-function: deriving biologically relevant functions from statistically significant functions

Jing Wang, Xianxiao Zhou, Jing Zhu, Yunyan Gu, Wenyuan Zhao, Jinfeng Zou and Zheng Guo

Abstract
In high-throughput studies of diseases, terms enriched with disease-related genes based on Gene Ontology (GO) are routinely found. However, most current algorithms used to find significant GO terms cannot handle the redundancy that results from the dependencies of GO terms. Simply based on some numerical considerations, current algorithms developed for reducing this redundancy may produce results that do not account for biologically interesting cases. In this article, we present several rules used to design a tool called GO-function for extracting biologically relevant terms from statistically significant GO terms for a disease. Using one gene expression profile for colorectal cancer, we compared GO-function with four algorithms designed to treat redundancy. Then, we validated results obtained in this data set by GO-function using another data set for colorectal cancer. Our analysis showed that GO-function can identify disease-related terms that are more statistically and biologically meaningful than those found by the other four algorithms.

Keywords: Gene Ontology; enrichment analysis; redundancy; statistically significant term; biologically relevant term

INTRODUCTION
To study a disease based on a high-throughput technology, a list of ‘interesting genes’ is usually extracted. Depending on the goal of a study, ‘interesting genes’ often represents genes related to a disease such as genes significantly differentially expressed in disease samples in comparison with normal controls or genes whose mutations elicit a phenotype of interest. Then, terms significantly enriched with the extracted interesting genes are routinely found using the Gene Ontology (GO) database [1], based on the assumption that terms in which the frequencies of...
interesting genes are significantly higher than expected by chance are likely to be relevant to the disease. We could also assume that the frequency of interesting genes in a GO term can measure the term’s relevance to a given disease. At least 68 enrichment analysis tools based on GO were developed prior to 2008 [2], and new tools continue to appear [3, 4]. The most often applied tools belong to a class of tools called singular enrichment analysis (SEA), which calculates enrichment P-values for all GO terms based on a list of interesting genes [2]. However, because each gene annotated by a term is also annotated by its ancestor terms according to GO’s ‘true path’ rule [1, 5], many terms found by SEA tools exhibit a redundant ancestor–offspring relationship. One popular approach to address this problem is to simply select the most specific GO terms from these terms [6–11], assuming that specific terms might be more biologically meaningful. However, a complex disease such as cancer could be a ‘systems disease’ with global molecular changes [12, 13]. Thus, whether a ‘specific’ or a ‘general’ term is more biologically relevant should be judged on the basis of the data rather than a priori assumption.

Several tools have been developed to address the redundancy problem [2]. For example, Grossmann et al. [14] developed the parent–child union (PCU) algorithm to analyse the significance of each GO term using genes in its parent term(s) as the background. However, the frequency of interesting genes in a term selected by PCU is not guaranteed to be significantly higher than the random background (i.e. the frequency of interesting genes in all genes under study) if the frequency of interesting genes in its parent term(s) happens to be lower than the random background (see examples in the ‘Results’ section). Another often used algorithm is called elim [15] which analyses terms iteratively from the most specific level to the most general level of GO: once a GO term is determined to be statistically significant, all genes associated with it are removed in the following analysis of its ancestor terms. However, elim may miss significant terms because excluding genes in significant terms from subsequent analysis affects the ranks of all terms in the FDR control procedure. These two algorithms combine the process of identifying statistically significant terms with the process of treating redundant terms, which may introduce the above mentioned problems. Thus, we believe that it is more appropriate to treat redundancy among the statistically significant terms pre-selected using a traditional enrichment analysis tool. As shown in Figure 1A, an ancestor term and an offspring term can both be detected as statistically significant in several situations. First, changes in genes in

![Figure 1](http://dx.doi.org/10.1093/bib/buj082)
only the offspring term may actually be related to the
disease such that non-random signals in this term
may lead to the entire ancestor term becoming de-
tected as statistically significant. In this situation, one
should only retain the offspring term (Case 1-1).
Second, changes in genes in the entire ancestor
term may be truly related to the disease, which
may result in one of its offspring terms being
detected as statistically significant. In this situa-
tion, two possible cases might be biologically mean-
ingful: the frequency of interesting genes in the
offspring term is not different from that in the
other genes in the entire ancestor term after remov-
ing genes in the offspring term (Case 1-2); it is dif-
f erent from that in the other genes in the ancestor
term (Case 1-3). In Case 1-2, we should only retain
the entire ancestor term. In Case 1-3, we could select
both terms because their relationship might be bio-
logically meaningful (see the Materials and Methods
section for details). Because PCU and elim are only
based on numerical considerations, they may pro-
duce results that do not account for these biologically
interesting cases.

Another concern in gene enrichment analysis is
the treatment of genes with multiple annotations
to terms with no ancestor–offspring relationship.
Some proposed methods penalize overlapping
terms to treat this problem, based on the assumption
that one of the overlapping terms could numerically
explain the subset of genes belonging to any of these
terms. For example, GENerative GO analysis
(GenGO) [3] maximizes a log-likelihood function
by penalizing the overall number of active terms.
Similarly, model-based gene set analysis (MGSA)
[4] uses the prior probabilities of terms being in the
active state to promote parsimonious explanations of
the data. Neither algorithm uses statistical tests for
term selection, thus, they may identify terms that
are not statistically relevant to the disease under
study. As shown in Figure 1B, two terms sharing
genes can both be identified as statistically significant
in several situations. First, if changes in genes in only
one term are truly related to a disease, another irrele-
vant term may be detected as statistically significant
(Case 2-1). Second, genes with multiple annotations
may actually play important roles in multiple path-
ways [7, 16–19], such that both terms are related to
the disease (Case 2-2) (see the Materials and Methods
section for details). As demonstrated in the ‘Results’
section, the two algorithms described above may miss
most biologically relevant terms.

In this work, we designed a tool called GO-
function for deriving biologically relevant, non-
redundant terms from statistically significant terms
for a disease. First, GO-function pre-selects statis-
tically significant GO terms at a false discovery rate
(FDR) control level. Then, from these statistically
significant terms, it searches for terms which are bio-
logically relevant to a disease according to several
explicit rules for the cases described in Figure 1.
Using one gene expression profile for colorectal
cancer, we compared GO-function with the four
above-mentioned algorithms. Compared with these
four tools, GO-function can find GO terms that
are more statistically and biologically meaningful.
Finally, we used another gene expression profile
for colorectal cancer to validate results obtained by
GO-function in the previous gene expression profile.
The GO-function R package can be found at its
project homepage at http://bioinfo.hrbmu.edu.cn/
fcensus/GO_function.jsp.

MATERIALS AND METHODS

Data sets

We mainly analysed a gene expression profile
(referred to as the Sabates-Bellver data set) based
on 32 pairs of colorectal adenomas tissue and adja-
cent normal mucosa [20]. Another gene expression
profile (referred to as the Hong data set) [21], includ-
ing 70 early stage colorectal cancer samples and
12 normal control samples, was used for validating
results obtained from the Sabates-Bellver data set.
Each of the two data sets was preprocessed using
the robust multi-array average (RMA) package
[22]. The SOURCE database [23] (April 2009) was
used for the translation of ProbeIDs to GeneIDs. For
each data set, differentially expressed (DE) genes be-
tween the disease samples and the normal samples
were selected by the significance analysis of micro-
arrays (SAM) algorithm [24] at a given FDR level.

In this work, we analysed the biological process
(BP) GO categories obtained from Bioconductor
GO.db on 21 March 2010 [25]. GO-function is cur-
rently applicable to 18 organisms including human
(see details in the Document at http://bioinfo.
hrbmu.edu.cn/fcensus/GO_function.jsp) and it will
be readily applicable to other organisms when their
gene annotation packages become available in
Bioconductor. GO-function uses the Rgraphviz
package [25] to output the hierarchy structure of
the identified terms. Because the Rgraphviz package
is dependent on the non-R tool graphviz (http://www.graphviz.org/), users should install the graphviz tool before running GO-function.

**Deriving biologically relevant terms from statistically significant terms**

At a given FDR control level, GO-function first selects statistically significant terms from all terms based on the hypergeometric distribution model [26], with enrichment probabilities (i.e., \( P \)-values) corrected by the Benjamini and Yekutieli FDR procedure [27]. For these terms, it treats local redundancy between ancestor and offspring terms produced by GO’s ‘true path’ rule [1, 5]. The comparison between an ancestor term and its offspring terms is to find whether the entire ancestor term or only its offspring terms is related to the disease. For all the three types of ancestor–offspring relationships, which are ‘Part-of’, ‘Is-a’, and ‘Regulate’, the genes in the offspring term are a subset of all the genes in the ancestor term according to the ‘true path’ rule (http://www.geneontology.org/GO.ontology.relations.shtml).

Finally, for the terms retained in the second step, GO-function treats the global redundancy between non-ancestor–offspring terms with overlapping genes.

**Local redundancy treatment**

To reduce local redundancy among the pre-selected, statistically significant GO terms, GO-function processes these terms iteratively from the most specific level to the most general level of GO according to three rules as described below. An ancestor term is removed after comparison with its significant offspring term(s) if there is no evidence that it is still related to the disease after excluding genes in its significant offspring terms. Formally, an ancestor term is removed if (i) after the removal of genes in its significant offspring term(s), the frequency of interesting genes in the remaining genes is not significantly higher than the random background at a given significance level (\( P \)-value cutoff), and (ii) this frequency is lower than that in its significant offspring terms (Rule 1-1 for Case 1-1 in Figure 1A). Otherwise, the ancestor term is retained. In this situation, the offspring terms are removed if at a given significance level these terms are not significantly enriched with interesting genes when using the ancestor term as the background (Rule 1-2 for Case 1-2 in Figure 1A). Otherwise, the offspring terms are also retained (Rule 1-3 for Case 1-3 in Figure 1A).

An offspring term that is considered redundant with one of its ancestor terms is still retained for comparison with other ancestor terms; it is finally removed after comparison with all ancestor–offspring terms. Notably, because the statistical power of detecting significance for a smaller set of genes tends to be lower, elim may miss some ancestor terms. GO-function remedies this problem to some extent with Rule 1-2 and 1-3 (see the ‘Results’ section for examples). However, some significant offspring terms may be removed simply due to insufficient power of using statistical testing for decision making.

**Global redundancy treatment**

GO-function offers an option for treating global redundancy among the terms retained after treating for local redundancy. For a pair of terms with overlapping genes, GO-function only retains one term if the following conditions apply (Rule 2-1 for Case 2-1 in Figure 1B): (i) there is evidence that the non-overlapping genes in a term may be related to the disease if at a given significance level the frequency of interesting genes in the non-overlapping genes is not significantly different from that in the overlapping genes or is significantly higher than the random background; and (ii) there is no such evidence for the non-overlapping genes of another term. In all other situations, GO-function retains both terms (Rule 2-2 for Case 2-2 in Figure 1B).

We highlight that we should consider local and global redundancy treatments with different confidence. When comparing an ancestor term with its offspring terms, we would rather be certain that the entire ancestor term is either redundant or non-redundant. On the other hand, for terms with overlapping genes, we would rather be uncertain to delete a term determined to be redundant only based on the numerical evidence because many genes with multiple annotations may truly play roles in multiple functions [7, 16–19]. Thus, to facilitate a comprehensive interpretation of the enrichment results, GO-function outputs the following: (i) all statistically significant terms, organized according to their ancestor–offspring relationship; (ii) any terms remaining after treating for local redundancy; and (iii) the terms remaining after treating for global redundancy.

**Comparison**

We compared GO-function with PCU [14] and elim [15]. The \( P \)-values obtained from both PCU
and elim were corrected using the Benjamini and Yekutieli FDR procedure [27]. We also compared GO-function with GenGO [3] and MGSA [4]. Because MGSA uses a Monte Carlo random walk algorithm, we considered that a term was significant only if it was significant in at least five of ten runs of MGSA.

We used the percentage of overlapping (PO) score [28] to measure the consistency between two lists of terms found by two algorithms: if \( k \) terms are shared by term list 1 with length \( L_1 \) and term list 2 with length \( L_2 \), then the PO score is calculated as \( \text{PO} = \left( \frac{(k/L_1 + k/L_2)}{2} \right) \times 100\% \).

**RESULTS**

**Comparing local redundancy test to PCU and elim**

In the analysis of the Sabates-Bellver microarray data on colorectal cancer, we used the SAM with an FDR of 1% to identify 9201 genes as DE genes between the colorectal adenomas tissue samples and the adjacent normal mucosa samples. Using GO-function, 108 GO terms were identified as statistically significant with a 5% FDR. We set \( P \)-value as 0.05 for local redundancy treatment with respect to ancestor–offspring relationships. After treatment for local redundancy, 33 GO terms remained (see Table 1 and Supplementary Figure S1).

Using PCU with a 5% FDR, 52 terms enriched with DE genes were found. Among these, seven terms were not statistically significant and were removed by GO-function, 17 were also identified by GO-function, and 28 were replaced by their offspring or ancestor terms identified by GO-function during local redundancy treatment. The detailed results of terms found by PCU are shown in Supplementary Table S1. The seven removed terms were not statistically relevant to the disease, but PCU detected them as significant because the frequencies of DE genes in their corresponding parent terms used as the background by PCU were lower than the random background. For example, the term ‘locomotory behavior’ (GO:0007626) contained 265 genes, among which 149 were DE genes, and its parent term ‘behavior’ (GO:0007610) contained 47.1% DE genes, which was lower than the 51.6% random background. Thus, the child term ‘locomotory behavior’ was significant when its parent term was set as the background by PCU (adjusted \( P = 4.9E-3 \)), whereas it was not significant when all of the annotated genes measured in the microarray data were set as the background (adjusted \( P = 1.0 \)). As shown in Supplementary Table S1, the other 27 terms found by PCU were removed by GO-function based on Rule 1-1 during local redundancy treatment. For each of these removed terms, after the removal of the genes in its significant offspring term(s), the frequency of DE genes in the remaining genes was lower than that in its corresponding offspring terms; moreover, this frequency was not significantly higher than the random background (\( P > 0.05 \)). For example, as shown in Figure 2A, ‘cellular response to stress’ (GO:0033554) was identified as a significant term by PCU (adjusted \( P = 9.1E-4 \)). However, after removing the genes in its child term ‘response to DNA damage stimulus’ (GO:0006974), the remaining genes associated with ‘cellular response to stress’ had a lower frequency of DE genes (54.2%) than it did in its child term (71.1%) and were not enriched with DE genes (\( P = 0.22 \)). Thus, the term ‘cellular response to stress’ was removed by GO-function. For another example, ‘organelle fission’ (GO:0048285) was found by PCU (adjusted \( P = 5.5E-6 \)). However, after removing the genes in its offspring term ‘mitosis’ (GO:0007067), the remaining genes associated with ‘organelle fission’ had lower frequencies of DE genes (62.5%) than it did in its offspring term (75.1%) and were not significantly enriched with DE genes (\( P = 0.40 \)). Thus, ‘organelle fission’ was removed by GO-function.

As shown in Table 1, among the 33 terms identified by GO-function, 17 terms were also found by PCU, and another 16 terms that were retained by GO-function according to Rule 1-1 were missed by PCU because the frequency of DE genes in their parent terms were higher than the random background. For example, the term ‘protein modification by small protein conjugation’ (GO:0032446) was annotated with 235 genes, including 159 DE genes; the term was significant when the background was set to include all annotated genes measured in the microarray data (adjusted \( P = 3.9E-4 \)). However, it was not significant when using its parent term ‘protein modification by small protein conjugation or removal’ (GO:0070647), which were annotated with 265 genes, including 178 DE genes, as the background (adjusted \( P = 1.0 \)). However, after the removal of genes associated with the term ‘protein modification by small protein conjugation’, the remaining genes associated with ‘protein modification
by small protein conjugation or removal’ had lower frequencies of DE genes (63.3%) than did ‘protein modification by small protein conjugation’ (67.7%) and were not enriched with DE genes ($P = 0.14$). Thus, only ‘protein modification by small protein conjugation’ was retained by GO-function.

Using elim with an FDR of 5%, 24 terms enriched with DE genes were found, among which 15 terms were also found by GO-function. Another nine terms were replaced by their ancestor terms identified by GO-function during local redundancy treatment according to Rule 1-2. All terms found by elim are described in Supplementary Table S2. Of the 33 terms (Table 1) identified by GO-function, 15 were also found by elim, and 18 were the ancestor terms of some terms found by elim and retained by GO-function according to Rules 1-2 and 1-3. For example, as shown in Figure 2B, when using elim, the term ‘positive regulation of ubiquitin–protein ligase activity during mitotic cell cycle’ (GO:0051437) was found to be significant, and its ancestor term, ‘positive regulation of ligase activity’ (GO:0051351), was removed due to redundancy. After removing the genes in the term ‘positive regulation of ubiquitin–protein ligase activity during mitotic cell cycle’, there remained only six genes associated with ‘positive regulation of ligase activity’, among which all genes were DE genes that were not...

<table>
<thead>
<tr>
<th>GOID</th>
<th>Name</th>
<th>Global redundancy</th>
<th>PCU</th>
<th>Elim</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0000075</td>
<td>Cell cycle checkpoint</td>
<td>N</td>
<td>Rule 1-1b</td>
<td>Overlapc</td>
</tr>
<tr>
<td>GO:0000226</td>
<td>Microtubule cytoskeleton organization</td>
<td>Y</td>
<td>Overlap</td>
<td>Rule 1-1</td>
</tr>
<tr>
<td>GO:0006260</td>
<td>DNA replication</td>
<td>N</td>
<td>Overlap</td>
<td>Rule 1-1</td>
</tr>
<tr>
<td>GO:0006261</td>
<td>DNA-dependent DNA replication</td>
<td>N</td>
<td>Overlap</td>
<td>Rule 1-1</td>
</tr>
<tr>
<td>GO:0006396</td>
<td>RNA processing</td>
<td>N</td>
<td>Overlap</td>
<td>Rule 1-2 and 1-3b</td>
</tr>
<tr>
<td>GO:0006412</td>
<td>Translation</td>
<td>N</td>
<td>Overlap</td>
<td>Rule 1-2 and 1-3</td>
</tr>
<tr>
<td>GO:0006974</td>
<td>Response to DNA damage stimulus</td>
<td>N</td>
<td>Overlap</td>
<td>Rule 1-2 and 1-3</td>
</tr>
<tr>
<td>GO:0007005</td>
<td>Mitochondrion organization</td>
<td>N</td>
<td>Rule 1-1</td>
<td>Rule 1-2 and 1-3</td>
</tr>
<tr>
<td>GO:0007049</td>
<td>Cell cycle</td>
<td>N</td>
<td>Overlap</td>
<td>Rule 1-2 and 1-3</td>
</tr>
<tr>
<td>GO:0007059</td>
<td>Chromosome segregation</td>
<td>Y</td>
<td>Overlap</td>
<td>Overlap</td>
</tr>
<tr>
<td>GO:0007067</td>
<td>Mitosis</td>
<td>Y</td>
<td>Rule 1-1</td>
<td>Overlap</td>
</tr>
<tr>
<td>GO:0007346</td>
<td>Regulation of mitotic cell cycle</td>
<td>Y</td>
<td>Rule 1-1</td>
<td>Overlap</td>
</tr>
<tr>
<td>GO:0009458</td>
<td>RNA modification</td>
<td>N</td>
<td>Overlap</td>
<td>Overlap</td>
</tr>
<tr>
<td>GO:0010202</td>
<td>Response to ionizing radiation</td>
<td>N</td>
<td>Overlap</td>
<td>Overlap</td>
</tr>
<tr>
<td>GO:0010564</td>
<td>Regulation of cell cycle process</td>
<td>N</td>
<td>Rule 1-1</td>
<td>Rule 1-2 and 1-3</td>
</tr>
<tr>
<td>GO:0022403</td>
<td>Cell cycle phase</td>
<td>N</td>
<td>Rule 1-1</td>
<td>Rule 1-2 and 1-3</td>
</tr>
<tr>
<td>GO:0022618</td>
<td>Ribonucleoprotein complex assembly</td>
<td>N</td>
<td>Rule 1-1</td>
<td>Rule 1-2 and 1-3</td>
</tr>
<tr>
<td>GO:0031445</td>
<td>Anaphase-promoting complex-dependent</td>
<td>Y</td>
<td>Rule 1-1</td>
<td>Overlap</td>
</tr>
<tr>
<td>GO:0032268</td>
<td>Regulation of cellular protein metabolic process</td>
<td>Y</td>
<td>Overlap</td>
<td>Rule 1-2 and 1-3</td>
</tr>
<tr>
<td>GO:0032446</td>
<td>Protein modification by small protein conjugation</td>
<td>Y</td>
<td>Rule 1-1</td>
<td>Rule 1-2 and 1-3</td>
</tr>
<tr>
<td>GO:0032508</td>
<td>DNA duplex unwinding</td>
<td>N</td>
<td>Rule 1-1</td>
<td>Rule 1-2 and 1-3</td>
</tr>
<tr>
<td>GO:0034621</td>
<td>Cellular macromolecular complex subunit organization</td>
<td>N</td>
<td>Overlap</td>
<td>Rule 1-2 and 1-3</td>
</tr>
<tr>
<td>GO:0042223</td>
<td>Response to chemical stimulus</td>
<td>N</td>
<td>Overlap</td>
<td>Rule 1-2 and 1-3</td>
</tr>
<tr>
<td>GO:0042254</td>
<td>Ribosome biogenesis</td>
<td>N</td>
<td>Rule 1-1</td>
<td>Overlap</td>
</tr>
<tr>
<td>GO:0043610</td>
<td>Proteasomal ubiquitin-dependent protein catabolic process</td>
<td>Y</td>
<td>Overlap</td>
<td>Rule 1-2 and 1-3</td>
</tr>
<tr>
<td>GO:0044419</td>
<td>Interspecies interaction between organisms</td>
<td>N</td>
<td>Overlap</td>
<td>Rule 1-2 and 1-3</td>
</tr>
<tr>
<td>GO:0048518</td>
<td>Positive regulation of biological process</td>
<td>N</td>
<td>Rule 1-1</td>
<td>Rule 1-2 and 1-3</td>
</tr>
<tr>
<td>GO:0050658</td>
<td>RNA transport</td>
<td>N</td>
<td>Rule 1-1</td>
<td>Overlap</td>
</tr>
<tr>
<td>GO:005301</td>
<td>Cell division</td>
<td>Y</td>
<td>Overlap</td>
<td>Overlap</td>
</tr>
<tr>
<td>GO:005351</td>
<td>Positive regulation of ligase activity</td>
<td>N</td>
<td>Overlap</td>
<td>Rule 1-2 and 1-3</td>
</tr>
<tr>
<td>GO:0054366</td>
<td>Negative regulation of ubiquitin–protein ligase activity during mitotic cell cycle</td>
<td>N</td>
<td>Rule 1-1</td>
<td>Overlap</td>
</tr>
<tr>
<td>GO:005726</td>
<td>Regulation of cell cycle</td>
<td>N</td>
<td>Overlap</td>
<td>Rule 1-2 and 1-3</td>
</tr>
<tr>
<td>GO:0065003</td>
<td>Macromolecular complex assembly</td>
<td>N</td>
<td>Rule 1-1</td>
<td>Rule 1-2 and 1-3</td>
</tr>
</tbody>
</table>

*N* represents that there was evidence that the significance of this term should not be simply due to the overlapping genes and *Y* represents no such evidence; ‘Rule 1-1’, ‘Rule 1-2’, and ‘Rule 1-3’ represent that a term was retained by GO-function according to the three rules described in ‘Materials and Methods’ section but missed by another method, respectively; ‘Overlap’ represents that a term found by GO-function was also found by another method.
significant according to elim (adjusted $P = 1.0$). However, the frequency of DE genes among the remaining genes in the ancestor term ‘ribosome biogenesis’ (GO:0042254) was 93.6%, which was higher than the frequency (88.3%) of DE genes in the offspring term ‘rRNA processing’. Meanwhile, the frequency of DE genes in the offspring term was not significantly different from that in the ancestor term ($P = 0.94$). Thus, GO-function retained only the ancestor term ‘ribosome biogenesis’.

Then, we analysed the data by GO-function using various $P$-value cutoffs for local redundancy treatment, and compared the results with the terms found by PCU and elim, respectively. Setting the $P$-value cutoff as 0.005, 0.01, 0.05 and 0.1, we found that the PO score (see ‘Materials and Methods’ section) took values ranging from 40.4 to 42.1% when comparing GO-function with PCU and from 54 to 57.8% when comparing GO-function with elim. Taken together, the above results indicated that terms found by PCU and elim are rather different from the terms found by GO-function.

Finally, we compared the terms found by GO-function using the $P$-value cutoff of 0.05 with the terms found by GO-function using other $P$-value cutoffs (0.005, 0.01 and 0.1) for local redundancy treatment. The results showed that the PO score took values ranging from 75.6% to 93.4%. When using various $P$-value cutoffs (0.005, 0.01 and 0.1) for both local and global redundancy treatments, the PO score between the final term lists took values ranging from 74.5% to 94.6%. These results indicated that GO-function is rather robust when using different $P$-value cutoffs.

**Comparing global redundancy test to MGSA and GenGO**

For two terms with overlapping genes, we set $P$-value as 0.05 to test whether the non-overlapping genes of each term are significantly enriched with DE genes and whether it is significantly different from the overlapping genes in the enrichment of DE genes. As shown in Table 1, for the 33 retained terms after local redundancy treatment, eight terms were removed according to Rule 2-1. For example, the terms ‘positive regulation of biological process’ (GO:0048518) and ‘cell cycle’ (GO:0007049) contained 319 overlapping genes, among which 69.6% were DE genes. After removing the overlapping genes from the term ‘positive regulation of biological process’, the frequency of DE genes among the
remaining genes (53%) was not significantly higher than the random background ($P = 0.11$) and was significantly lower than the frequency in the overlapping genes ($P = 2.4E-8$). However, the non-overlapping genes of the term ‘cell cycle’ were enriched with DE genes ($P < 2.2E-16$) and were not different from the frequency in the overlapping genes ($P = 0.19$). According to Rule 2-1, the term ‘cell cycle’ was retained, whereas the term ‘positive regulation of biological process’ was not because we could not find evidence that the significance of the latter was not due to the overlapping genes.

Using the DE genes found in the Sabates-Bellver data set, MGSA identified only two terms from all annotated BP terms, which were replaced by their significant offspring or ancestor terms identified by GO-function during local redundancy treatment (see Supplementary Table S3). GenGO found only five GO terms from all annotated BP terms (see Supplementary Table S4), among which (i) two terms were also found by GO-function, (ii) one term was replaced by its significant offspring terms identified by GO-function during local redundancy treatment, and (iii) another two terms were not significantly enriched with DE genes. The adjusted $P$-values of the latter two terms were 0.18 and 1.0, respectively, based on the hypergeometric distribution test with a 5% FDR. GenGO and MGSA tend to over-penalize overlapping terms and miss many biologically relevant terms, as shown in Table 1. Similarly, GenGO and MGSA missed most of the terms found by GO-function with other $P$-value cutoffs (0.005, 0.01 or 0.1) for treating local and global redundancy.

Because both GenGO and MGSA are based on pre-set optimization goals rather than statistical criteria, it is difficult to interpret their results statistically. We consider this a shortcoming of these two algorithms.

**Validation of application results using an independent experimental data**

One major advantage of GO-function is that the comparison between an ancestor term and its significant offspring terms can provide information for determining whether the entire ancestor term or only its an offspring term(s) is likely to be biologically related to the disease. This may provide a novel insight for understanding the disease mechanism at the systems biology level. Thus, we validated that the terms found in the Sabates-Bellver data set by GO-function, as well as the ancestor–offspring comparison results, could be reproduced in the Hong data set [21].

In the Hong data set, we used the SAM with an FDR of 1% to identify 7900 genes as DE genes between the colorectal cancer samples and the normal control samples. Using GO-function, 62 GO terms were identified as statistically significant with a 5% FDR. After treatment for local redundancy, 17 GO terms remained (see Table 2 and Supplementary Figure S2), among which eight terms (47.1%) were also found in the Sabates-Bellver data set. Then, we did a randomization experiment to test the null hypothesis that the number (eight) of overlapping terms would be expected for random genes unrelated to the disease. We randomly selected 7900 (i.e. the number of DE genes) genes from the Hong data set and used GO-function with an FDR of 5% to find terms enriched with these genes. Then, we counted the overlaps of the found terms with the terms found in the Sabates-Bellver data set. This process was repeated 10 000 times. The result showed that the average number of overlapping terms was 0.0001, which was significantly smaller than the number (eight) of overlapping terms observed for the same number of DE genes extracted from the Hong data set ($P < 1.0E-4$). Thus, we could reject the null hypothesis that the observed number (eight) of overlapping terms would be expected for random

**Table 2: Terms found by GO-function in the Hong microarray data for colorectal cancer**

<table>
<thead>
<tr>
<th>GOID</th>
<th>Name</th>
<th>Global redundancy</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0000226</td>
<td>microtubule cytoskeleton organization</td>
<td>Y</td>
</tr>
<tr>
<td>GO:0006260</td>
<td>DNA replication</td>
<td>N</td>
</tr>
<tr>
<td>GO:0006399</td>
<td>tRNA metabolic process</td>
<td>N</td>
</tr>
<tr>
<td>GO:0006403</td>
<td>RNA localization</td>
<td>N</td>
</tr>
<tr>
<td>GO:0006974</td>
<td>Response to DNA damage stimulus</td>
<td>N</td>
</tr>
<tr>
<td>GO:0007049</td>
<td>Cell cycle</td>
<td>N</td>
</tr>
<tr>
<td>GO:0007059</td>
<td>Chromosome segregation</td>
<td>N</td>
</tr>
<tr>
<td>GO:0008380</td>
<td>RNA splicing</td>
<td>N</td>
</tr>
<tr>
<td>GO:0012501</td>
<td>Programmed cell death</td>
<td>N</td>
</tr>
<tr>
<td>GO:0015931</td>
<td>Nucleobase, nucleoside, nucleotide and nucleic acid transport</td>
<td>N</td>
</tr>
<tr>
<td>GO:0016071</td>
<td>mRNA metabolic process</td>
<td>N</td>
</tr>
<tr>
<td>GO:0022618</td>
<td>Ribonucleoprotein complex assembly</td>
<td>N</td>
</tr>
<tr>
<td>GO:0033554</td>
<td>Cellular response to stress</td>
<td>N</td>
</tr>
<tr>
<td>GO:0034470</td>
<td>ncRNA processing</td>
<td>N</td>
</tr>
<tr>
<td>GO:0042245</td>
<td>Ribosome biogenesis</td>
<td>N</td>
</tr>
<tr>
<td>GO:0044267</td>
<td>Cellular protein metabolic process</td>
<td>N</td>
</tr>
<tr>
<td>GO:0051801</td>
<td>Cell division</td>
<td>Y</td>
</tr>
</tbody>
</table>

The notations are the same as described in the footnote of Table 1.
genes. After treatment for global redundancy, two terms were removed in the Hong data set, and both of them were also removed after treatment for global redundancy of the Sabates-Bellver data set.

In addition, most (83.1%) of all the 83 ancestor–offspring comparison results in the Sabates-Bellver data set could be reproduced in the Hong data set. For example, in the Sabates-Bellver data set, the term ‘cellular response to stress’ was redundant in comparison with its offspring term ‘response to DNA damage stimulus’ according to Rule 1-1. The same result was also observed in the Hong data set. This result indicated that the offspring term ‘response to DNA damage stimulus’ is more likely to be closely related to cancer [29, 30]. The second example is that, in both the Sabates-Bellver data set and Hong data set, the term ‘organelle fission’ was redundant in comparison with its offspring term ‘mitosis’ according to Rule 1-1. GO-function kept the term ‘mitosis’ which is also proved to be closely related to cancer [31]. Notably, the ancestor terms of the above two examples were found by PCU but removed by GO-function. As the third example, in both Sabates-Bellver data set and Hong data set, the term ‘positive regulation of ubiquitin–protein ligase activity during mitotic cell cycle’ was redundant in comparison with its ancestor term ‘positive regulation of ligase activity’ according to Rule 1-2. Many genes annotated in the ancestor term but not in the offspring term, such as PIN1 [32] and XRCC4 [33], are closely related to cancer. Thus, we could assume that the entire ancestor term is disturbed in the disease. The fourth example is that, in both the Sabates-Bellver data set and Hong data set, ‘rRNA processing’ was redundant in comparison with its ancestor term ‘ribosome biogenesis’ according to Rule 1-2. Studies also suggested that many genes annotated in the ancestor term but not in the offspring term, such as AATF [34] and MINA [35], are closely related to cancer. We note that the offspring terms of the above two examples were found by elim but removed by GO-function.

**DISCUSSION**

GO-function is designed to extract biologically relevant GO terms from statistically significant terms for a given disease. In comparison with GO-function, PCU tends to choose more ‘general’ terms whose significances might be simply due to their significant offspring terms. Elim tends to choose ‘specific’ terms, even if the frequencies of interesting genes in these terms are not different from or even lower than the frequencies in their corresponding ancestor terms. In such a case, changes in genes across the entire BPs, rather than only in the sub-processes, may actually be related to the disease. Both PCU and elim do not treat global redundancy. For the two algorithms that do treat global redundancy, GenGO and MGSA, we found that they tend to miss most biologically relevant terms and it is difficult to interpret their results statistically. In contrast to these other methods, GO-function can find GO terms that are more statistically and biologically meaningful. As shown in the ‘Results’ section, GO-function is rather robust when using different P-value cutoffs for local and global redundancy treatment. Notably, even after redundancy treatment that tends to exclude terms including many irrelevant genes, a term identified to be relevant to a disease is still based on the statistical evidence of the enrichment of interesting genes in this term. Thus, we still cannot infer that every gene in a selected term is relevant to the disease. However, we could assume that the changes of some ‘non-interesting genes’ in this term are also likely to be relevant to the disease because the power of statistical testing for determining ‘interesting genes’ may be insufficient in the presence of large biological variation.

GO-function belongs to the class of tools called modular enrichment analysis (MEA) [2], which calculates enrichment P-values of ‘interesting genes’ for all GO terms. One major limitation of this type of method is that they rely on thresholds for defining genes as ‘interesting genes’. In contrast, another type of method named threshold-free methods can handle this difficulty by using continuous measures [36]. However, the threshold-free methods also have their own limitations, as discussed in [2]. Taking the tool named GSEA (gene set enrichment analysis) [36] for example, a few highly changed genes in a term can introduce a small enrichment P-value even if most genes in this term are randomly changed [2]. When using GSEA, users still need to choose parameters which can affect the output. For example, for the Sabates-Bellver data set, no significant term could be found with a FDR of 5% when using GSEA with the ‘phenotype’ permutation but 94 terms could be found with the ‘gene_set’ permutation. The ‘Min size’ and ‘Max size’ parameters of GSEA can also affect the output. In general, as Huang et al. [2] commented, ‘current enrichment
analysis is more of an exploratory procedure, with the aid of enrichment P-value, rather than a pure statistical solution. The notion that the enriched terms should make sense based on a priori biological knowledge of the study is the most important guideline to help users in adjusting analytic thresholds. Finally, threshold-free methods also face the redundancy problem, which needs to be addressed in our future work.

One common problem in the evaluation of enrichment analysis tools is that there are no gold standards for phenotype-associated pathways by which to assess the results. Some studies [3, 4, 14, 15] have attempted to address this problem using simulated data. However, simulation studies are heavily dependent on the models themselves for generating the simulated data, and as such, they often lead to biased results. For example, we found that although PCU and elim performed well with certain simulated data [14, 15], the precisions of these two algorithms decreased to below 0.1 based on simulated data used to evaluate MGSA [4]. Thus, in this work, we did not use simulated data to compare GO-function with other methods. Instead, we used certain rules to compare results from different algorithms. The logic behind these rules is simple and can be interpreted by users. In addition, we have validated some interesting results obtained by GO-function in one data set using another independent laboratory data set for colorectal cancer.

Currently, a serious problem in studying disease-associated pathways is that the criteria for defining pathways vary greatly in different pathway databases [37–40]. For example, Kyoto Encyclopedia of Genes and Genomes (KEGG) database [41] tends to define general pathways, whereas BioCarta database (http://www.biocarta.com/) tends to define specific pathways [37]. Based on these pathway data sources, it would be hard if not impossible to find whether only a part of or the entire pathway is likely to be disturbed in a disease. In contrast, GO [1] defines functions at various specific levels in a hierarchical manner. As such, GO is relatively flexible to address the problem that whether only an offspring term of or the entire term is likely to be related to a disease. We recognize that BPs defined in GO are ‘series of events accomplished by one or more ordered assemblies of molecular functions’ [1], which usually cannot be simply mapped to traditional pathways. Nevertheless, some GO terms could be mapped to traditional pathways or parts of these pathways. For example, most of genes in the GO term ‘cell cycle’ can be mapped to the ‘cell cycle’ pathway collected in KEGG. GO contains not only the ‘cell cycle’ term but also its offspring terms such as ‘mitosis’ which can be largely mapped to the M phase of the ‘cell cycle’ pathway in KEGG (see Supplementary Figure S3). In the Sabates-Bellver data set for colorectal cancer, we found that the frequency of DE genes in the ‘mitosis’ term was significantly higher than that in the ‘cell cycle’ term. Therefore, these two terms were retained by GO-function according to Rule 1-3. This result indicated that ‘mitosis’ may play a key role in cancer mechanism [31]. Notably, this difference could not be found based on KEGG because no sub-pathways of ‘cell cycle’ pathway are described in KEGG. However, as a common problem of using current enrichment analysis tools based on GO, how to explain the selected GO terms in the context of traditional pathways deserves our future study.

SUPPLEMENTARY DATA
Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points
- Most current algorithms used to find significant GO terms cannot handle the redundancy that results from the dependencies of GO terms.
- Simply based on some numerical considerations, current algorithms developed for reducing this redundancy produce results that do not account for biologically interesting cases.
- GO-function extracts disease-related GO terms from pre-selected statistically significant terms and the results are interpretable based on several explicit rules.

FUNDING
The National Natural Science Foundation of China (grant numbers 91029717, 30770558, 30970668 and 81071646); Excellent Youth Foundation of Heilongjiang Province of China (grant number JC200808); Natural Science Foundation of Heilongjiang Province of China (grant number QC2010012); and Scientific Research Fund of Heilongjiang Provincial Education Department of China (grant number 11541156).
References


