BIOINFORMATICS ORIGINAL PAPER
Vol. 24 no. 2 2008, pages 265–271
doi:10.1093/bioinformatics/btm558

Databases and ontologies

Gaining confidence in biological interpretation of the microarray data: the functional consistence of the significant GO categories

Da Yang[1,†], Yanhui Li[2,†], Hui Xiao[1,†], Qing Liu[1], Min Zhang[1], Jing Zhu[2], Wencai Ma[1], Chen Yao[1], Jing Wang[1], Dong Wang[1], Zheng Guo[1,2,*] and Baofeng Yang[1]

[1]Department of Bioinformatics, Bio-pharmaceutical Key Laboratory of Heilongjiang Province-Incubator of State Key Laboratory, Harbin Medical University, Harbin 150086 and [2]Bioinformatics Centre and School of Life Science, University of Electronic Science and Technology of China, Chengdu, 610054, China

Received on July 21, 2007; revised on November 3, 2007; accepted on November 4, 2007

Advance Access publication November 15, 2007

Associate Editor: Martin Bishop

ABSTRACT

Motivation: In microarray studies, numerous tools are available for functional enrichment analysis based on GO categories. Most of these tools, due to their requirement of a prior threshold for designating genes as differentially expressed genes (DEGs), are categorized as threshold-dependent methods that often suffer from a major criticism on their changing results with different thresholds.

Results: In the present article, by considering the inherent correlation structure of the GO categories, a continuous measure based on semantic similarity of GO categories is proposed to investigate the functional consistence (or stability) of threshold-dependent methods. The results from several datasets show when simply counting overlapping categories between two groups, the significant category groups selected under different DEG thresholds are seemingly very different. However, based on the semantic similarity measure proposed in this article, the results are rather functionally consistent for a wide range of DEG thresholds. Moreover, we find that the functional consistence of gene lists ranked by SAM metric behaves relatively robust against changing DEG thresholds.

Availability: Source code in R is available on request from the authors.

Contact: guoz@ems.hrbmu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Currently, microarray data is routinely interpreted by using Gene Ontology (GO) categories (Ashburner *et al.*, 2000). Most commonly, genes are first ranked and selected by the evidence of differential expression according to some statistical metrics. Then, GO categories are examined to find the ones significantly overrepresented in the selected genes compared with the whole gene list. Various software tools for this analysis have been developed (Al-Shahrour *et al.*, 2004; Alexa *et al.*, 2006; Draghici *et al.*, 2003; Hosack *et al.*, 2003; Pehkonen *et al.*, 2005). Since they rely on thresholds for defining genes as 'differentially expressed genes (DEGs)', these tools can be referred to as threshold-dependent methods, for distinguishing from the threshold-free methods using continuous measures for gene expressions such as GSEA (gene set enrichment analysis) (Subramanian *et al.*, 2005) and global test (Goeman *et al.*, 2004). Although the threshold-dependent methods are commonly applied to investigate the concerted gene expression changes induced by various experimental conditions, they suffer from a major criticism on the choice of the thresholds (or the lengths of the lists) for selecting DEGs that might dramatically affect the identification of the functional categories significantly overrepresented in the gene lists (Ben-Shaul *et al.*, 2005; Nilsson *et al.*, 2007; Pan *et al.*, 2005). Notably, the threshold-free methods are efficient in finding categories containing high proportions of lowly differentially expressed genes. On the other hand, the threshold-dependent methods can be more powerful in detecting categories containing low proportions of highly differentially expressed genes (Nilsson *et al.*, 2007). Therefore, these two main streams of enrichment analysis methods would be mutually complementary if the threshold-dependent approaches could be largely free from the criticism on their 'instability'.

In previous studies for this problem (Ben-Shaul *et al.*, 2005; Nilsson *et al.*, 2007; Pan *et al.*, 2005), the similarity of two groups of categories were simply inspected by counting their overlaps, which might lead to inconclusive results because GO categories are hierarchically related and some closely related categories may be identified as being significant separately under different thresholds. Figure 1 depicts two groups of GO categories identified in a leukemia dataset: the hexagon categories are their overlaps; diamond and rectangle categories are respectively significant under two separate thresholds (see details in Methods and Results section). Intuitively, these two groups of categories are closely related because each rectangle category is within two steps (GO branches) of at least one of the diamond categories. The semantic similarity between such closely related category pairs (such as parent–child pairs 'glycosaminoglycan biosynthetic

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

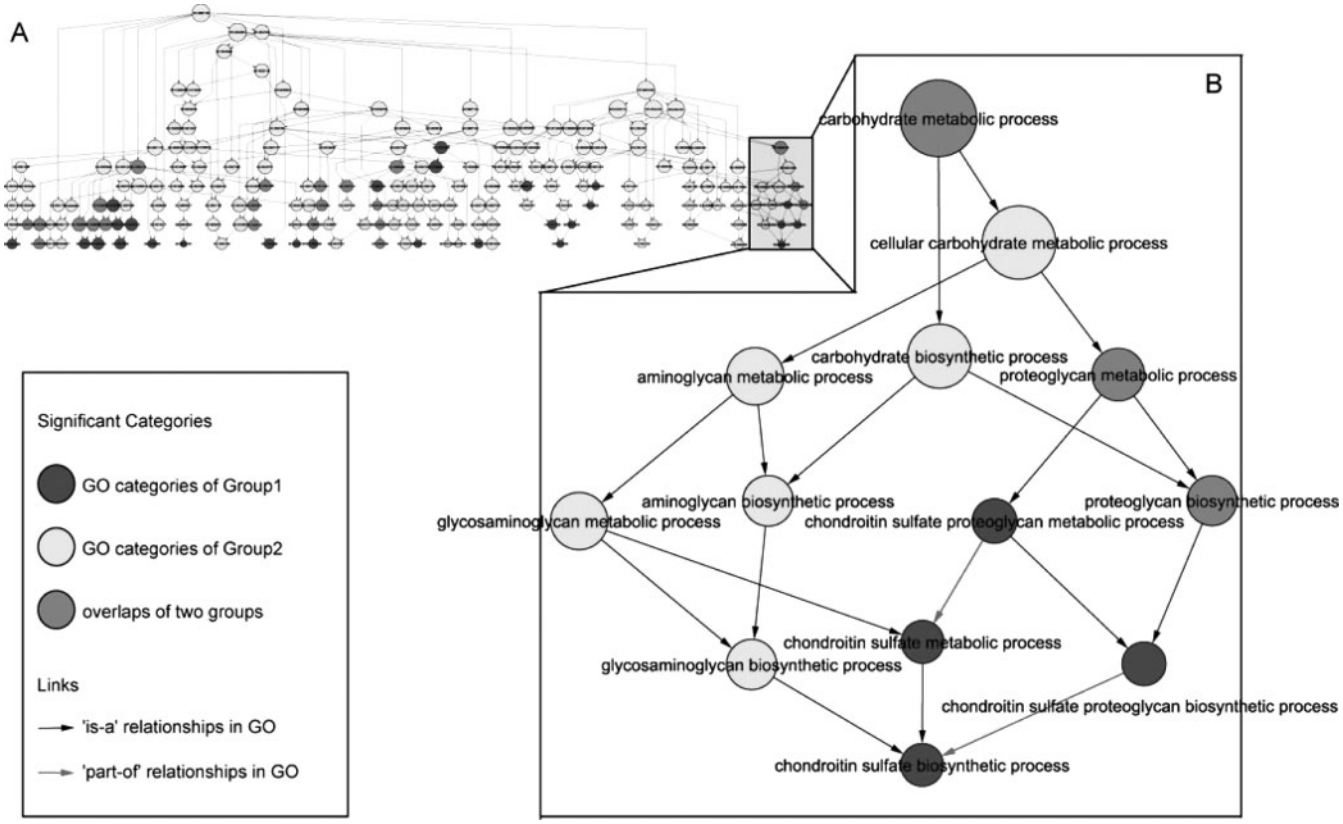

Fig. 1. Hierarchical graph of two groups of significant enriched GO categories in leukemia dataset. Group1 consists of 49 categories (rectangle and hexagon) enriched with the top 1000 genes ranked by SAM and Group2 consists of 67 categories (diamond and hexagon) enriched with the top 3000 genes by the same gene ranking method. The hexagon categories are the overlaps between these two groups of categories. Both groups of significant categories are selected with FDR ≤ 0.1. The size of each category is proportional to its total number of gene annotations. The sub-branch of the overall hierarchical graph (**A**) in the shadow region is zoomed in as (**B**). Hierarchical graph is visualized using Cytoscape.

process' and 'chondroitin sulfate biosynthetic process') are overlooked in some previous works (Ben-Shaul *et al.*, 2005; Nilsson *et al.*, 2007; Pan *et al.*, 2005). In addition to parent–child relation, other types of close relationships (such as siblings) exist among categories in GO.

Considering the semantic similarity of the related GO categories, we propose a continuous measure to investigate the functional consistence of the significant categories identified by threshold-dependent methods. The results based on four datasets show that, using the proposed semantic similarity measure, the threshold-dependent methods are actually rather robust to a wide range of DEG thresholds. Besides the choice of thresholds for selecting DEGs, various gene ranking metrics might also fundamentally influence the identification of significant categories.

Here, we further examine the impact of using four different gene ranking metrics on finding significant categories, including FC (fold change), SNR (Signal to Noise Ratio), *t*-test and SAM [Significance Analysis of Microarrays (Tusher *et al.*, 2001)]. However, previous studies only analyzed SNR (Pan *et al.*, 2005) or *t*-test statistics (Nilsson *et al.*, 2007). Our results show that the significant categories enriched with DEGs obtained by SAM metric tends to be more robust against changing thresholds.

2 METHODS

2.1 Datasets

Four publicly available datasets are analyzed in this study. The acute myeloid leukemia data (Valk *et al.*, 2004) consists of 19 samples for idt(16) and 23 t(15;17) karyotypes, generated by Affymetrix U133A GeneChips containing 22 283 probe sets. The prostate cancer oligonucleotide microarray data (Singh *et al.*, 2002) contains 52 tumors and 50 non-tumor samples, measured by Affymetrix U95Av2 GeneChips containing 12 625 probe sets. The liver cancer cDNA microarray data (Chen *et al.*, 2002) contains 82 primary hepato-cellular carcinoma (HCC) and 74 non-tumor liver tissues measured for 23 093 clones. The prostate cancer cDNA microarray data (Lapointe *et al.*, 2004) consists of 62 primary prostate tumors and 41 normal prostate specimens measured for 46 205 clones. For the last three carcinoma datasets, samples were grossly dissected, so the DEGs (and then the enrichment categories) may be derived from tumor stroma as well as cancer cells. We do not discriminate this difference just as most works did.

For the Affymetrix microarrays, the raw data is background corrected and normalized using the Bioconductor RMA package. This software implements the quantile normalization procedure carried out at the probe feature level (Bolstad *et al.*, 2003; Irizarry *et al.*, 2003). For the cDNA microarray data, each microarray is normalized by the global LOWESS normalization method (Yang *et al.*, 2002). The data for clones with missing value rate larger than 20% is deleted

and the remaining missing values are replaced using the KNNimpute imputation algorithm ($k = 15$) (Troyanskaya *et al*., 2001; Wang *et al*., 2006). Each expression value is base 2 logarithmically transformed. For each sample in a dataset, the measurement values of all the probes corresponding to a same Gene ID are averaged to a single value (the mean value). The probes mapping to multiple Gene IDs or having no known mapping to a Gene ID are deleted. The probe annotation data is from the SOURCE database (http://source.stanford.edu) in July 2007. Given the large amount of alternative splicing in human genome, we concede it is imprecise to refer the transcripts directly measured in the microarrays as genes. In the following text, we do not discriminate this difference just for simplicity.

2.2 Generation of DEG lists

DEG lists are produced using four methods. For the simple *ad hoc* fold change metric, the expression values for each gene are averaged across the samples in each group and their ratios are used to rank genes. For the SNR metric, the difference of the average values in each group is divided by the sum of the SDs of a gene in two groups. The *t*-statistic substitutes the denominator of SNR with a pooled SD, assuming an equal variance across two groups. When using the *t*-statistic, small per-gene variances can make small fold changes statistically significant. The SAM method (Tusher *et al*., 2001) tries to solve this problem by adding a small 'fudge factor' to the denominator of the test statistic.

2.3 Selection of significant categories

With a DEG list, a hypergeometric distribution statistics (Draghici *et al*., 2003) is applied to calculate the probability p of a GO (2007–9) category annotated with at least the observed number of the DEGs by random chance. For each DEG list, two criteria are used to select their corresponding significant categories: (1) according to their *P*-values ranked as the smallest n ones, and (2) by a given FDR level (Benjamini and Hochberg, 1995).

For simplicity, we only study categories from GO 'biological process'. There are two types of relation links in GO: 'IS-A' and 'PART-OF'. 'IS-A' is defined when a child category is a certain kind of its parent category, while 'PART-OF' is used when a parent has the child as its part. Many researches considered the two types as equal for estimating semantic similarity of GO categories (Lord *et al*., 2003; Wang *et al*., 2004), which could be partially justified under the consideration that most GO links are 'IS-A'. Here, we only consider 'IS-A' links for semantic similarity analysis (Lin, 1998; Resnik, 1999). This might result in 'semantically impoverishing' (Lord *et al*., 2003). In Supplementary Table S6, we also present the results considering the two links equally.

2.4 Semantic similarity between two categories

The semantic similarity for GO categories can be measured by the amount of information they share in common. The information content of a category c, based on information theory (Shannon, 2001), is defined as –log ($p(c)$), where $p(c)$ is the number of gene products annotated to this category divided by the number of all the gene products annotated to the GO taxonomy.

As suggested by Resnik (Resnik, 1999), the similarity measure of two categories relies on their minimum subsumer (i.e. most specific common ancestor) in the GO hierarchy. The semantic measure for the similarity between m and n can be defined as:

$$sim(m,n) = \max_{c \in P(m,n)} [-\log(p(c))] \tag{1}$$

where $P(m,n)$ represents the set of ancestor categories of m and n, and $p(c)$ is described as above. Thus the $\max_{c \in P(m,n)} [-\log(p(c))]$ corresponds to information content of the most specific category that subsumes m and n.

The above measure can take values varying between 0 and infinity, thus we adopt its normalized version proposed by Lin (Lin, 1998), which takes values varying between 0 and 1. Given categories m and n, the similarity is calculated as:

$$sim(m,n) = \frac{2 \times \max_{c \in P(m,n)} [-\log(p(c))]}{-\log(p(m)) - \log(p(n))} \tag{2}$$

2.5 Similarity between two category groups

We use Dice coefficient (DC) (Frakes and Baeza-Yates, 1992) and a semantic similarity score (SS) to investigate the similarity (consistence) of two category groups.

2.5.1 Discrete measure The consistence of two category groups can be evaluated by the DC. Let G_i and G_j denote two groups of categories, then:

$$\mathrm{DC}(G_i, G_j) = \frac{2 \times \#(G_i \cap G_j)}{\#G_i + \#G_j} \tag{3}$$

where $\#(G_i \cap G_j)$ is the number of overlaps between G_i and G_j, while $\#G_i$ and $\#G_j$ are the numbers of categories included in G_i and G_j, respectively.

The DC measure can take values ranging from 0 to 1, and it takes value 1 if and only if the two groups are identical. However, it can take only a few discrete values because it only counts overlaps of categories. For example, if two groups each with 30 categories are compared, then DC can take only 31 possible values. Suppose the number of overlaps between two groups is the same with that of another two groups, there is nothing DC can do for further discrimination. Under such a situation, the continuous semantic measure SS as described below tends to be more sensitive in exploring the similarity between the two group pairs by considering the functional relations of the categories.

2.5.2 Continuous measure Based on the semantic similarity metric for two GO categories, we propose a continuous metric for evaluating the functional similarity of two category groups (G_i and G_j). First, we compute $S(G_i, G_j)$ (semantic similarity from group G_i to group G_j) by assigning each term in G_i one most semantically similar term (best-match) in G_j and summing the semantic similarity measures of all the best-match pairs.

$$S(G_i, G_j) = \sum_{m \in G_i} \max_{n \in G_j} (sim(m,n)) \tag{4}$$

where $sim(m,n)$ is a semantic similarity measure for m and n.

Then, $S(G_j, G_i)$ is computed in the same way by swapping groups G_i and G_j. Finally, the overall similarity $\mathrm{SS}(G_i, G_j)$ between G_i and G_j is given by:

$$\mathrm{SS}(G_i, G_j) = \frac{S(G_i, G_j) + S(G_j, G_i)}{S(G_i, G_i) + S(G_i, G_i)} \tag{5}$$

From Equations (4) and (5), it is easy to prove that the SS metric has the following properties, which are helpful to understand this metric intuitively.

- $\mathrm{SS}(G_i, G_i) = 1$.
- $\mathrm{SS}(G_i, G_j) = \mathrm{SS}(G_j, G_i)$.
- $\mathrm{SS}(G_i, G_i) \in [0, 1]$.
- $\mathrm{SS}(G_i, G_i) = 0$ if and only if the most specific subsumer of any pair of categories from G_i and G_j is the GO root ('biological process').
- $\mathrm{SS}(G_i, G_i) = 1$ if and only if for each best-match pair from G_i and G_j, the two categories are identical or related as parent–child, where all the annotations in the parent category are from the child category.

Moreover, one may view SS measure as an extension of Dice coefficient: Equation (5) is equivalent to (3) if we replace the semantic similarity matrix with a binary diagonal matrix so that only the identical category pairs (the diagonals of the matrix) get value 1 while 0 for the other category pairs.

Two kinds of randomizations are used to obtain significance levels for SS measures. The first method is referred to as 'annotation randomization', where we randomly assign genes to GO categories while retaining the number of genes directly annotated to each GO category. Then, each annotation for the randomized set is associated to its ancestor GO categories using the GO graph relationships. This process generates randomized GO annotations while mimicking the effects of the GO graph relationships. The randomized sets are used to test the case when no prior biological information is used in the enrichment analysis. The second randomization method is referred to as 'genelist randomization', where we hold the GO annotations unchanged while randomly selecting DEG lists of the same length and keeping the overlaps between the lists the same. This randomization tests whether the functional consistence of the significant categories is mainly caused by the overlaps among gene lists. After 10 000 experiments for each randomization method, we can get a significance level for an observed SS score, by calculating the rate of the SS scores larger than the observed value in the empirical distribution.

3 RESULTS

3.1 Functional consistence of categories enriched with DEGs selected under different thresholds

First, using four ranking methods (fold change, SNR, *t*-test and SAM) respectively, six lists of n top-ranked genes, with n increasing from 500 to 3000 (adding 500 genes per step), are selected for each dataset. Then, by applying the hypergeometric statistics to each gene list, 30 most significant categories with the smallest significant values are selected. For each ranking method used in a dataset, similarity adjacent matrixes based on DC and SS respectively are constructed for the significant category groups identified by every two different DEG thresholds. Finally, the average values and the SSs for the elements in the upper triangle of each adjacent matrix are shown in Table 1. As described in Table 1, the means of DC measure take values ranging from 0.47 to 0.74, indicating that only 47–74% of the categories can be consistently identified under the varying thresholds. Analogous results are obtained when top 10, 20 and 40 categories are considered (see Table S1). Since the overlap-derived measure gives a rather negative signal, people may draw the conclusion that the threshold-dependent methods are not robust against varying thresholds.

On the other hand, through all datasets and gene ranking metrics, the means of SS measure take higher values (changing from 0.72 to 0.90), and generally, the SDs of the SS measure are lower. To test the performance of SS in the case that no prior biological information is used in the enrichment analysis, we carry out 10 000 'annotation randomization' (see Methods section) and find that, for all datasets and gene ranking metrics, the means of SS measures are significant ($p<0.005$). Moreover, to examine whether the observed functional consistence is mainly caused by the overlaps between gene lists, we carry out 10 000 'genelist randomization' (see Methods section) to estimate a significance level for each SS mean, and the results show that all the SS measures in Table 1 are significant

Table 1. The means and SDs of DC and SS when using top 30 ranked categories

		DC	SS
Leukemia	FC	0.65 ± 0.13	$0.84\pm0.07^{*}$
	SNR	0.59 ± 0.14	$0.83\pm0.09^{*}$
	t-test	0.58 ± 0.12	$0.80\pm0.11^{*}$
	SAM	0.56 ± 0.18	$0.79\pm0.14^{**}$
Prostate (oligo)	FC	0.74 ± 0.12	$0.90\pm0.05^{*}$
	SNR	0.48 ± 0.23	$0.73\pm0.15^{**}$
	t-test	0.47 ± 0.22	$0.72\pm0.15^{**}$
	SAM	0.60 ± 0.23	$0.82\pm0.12^{*}$
Liver	FC	0.66 ± 0.17	$0.86\pm0.11^{*}$
	SNR	0.61 ± 0.17	$0.82\pm0.12^{*}$
	t-test	0.66 ± 0.18	$0.82\pm0.14^{**}$
	SAM	0.67 ± 0.16	$0.85\pm0.12^{*}$
Prostate (cDNA)	FC	0.58 ± 0.13	$0.79\pm0.09^{**}$
	SNR	0.60 ± 0.13	$0.81\pm0.07^{**}$
	t-test	0.65 ± 0.15	$0.83\pm0.09^{*}$
	SAM	0.60 ± 0.12	$0.81\pm0.08^{*}$

The superfix '*' ('**') indicates that the corresponding significance levels are equal or smaller than 0.0001(0.005), by 'annotation randomization'.

($p<0.01$, see Table S2). Therefore, the functional consistence between the significant categories selected under different thresholds for choosing DEGs is not simply due to the overlaps between DEG lists but reflects the functional relations of the selected genes. To sum up, based on both large values of SS measures and their statistical significances, we suggest the threshold-dependent methods would actually produce functionally consistent results for a wide range of thresholds for selecting DEGs.

Furthermore, we adopt another commonly used criterion, FDR ≤ 0.1, to define significant categories and study their functional consistence. As shown in Table 2, the results are roughly the same with those in Table 1. In the prostate cancer data (oligo), according to gene lists ranked by SNR and *t*-test metrics, the average values of both SS and DC show conspicuous decreases. Similar results were also obtained in an early study. Actually, just using the SNR and the FDR criterion, Pan *et al.* (Pan *et al.*, 2005) claimed that GO categories enrich with DEGs (gene signature) highly depend on the thresholds used to select those genes. However, as shown in Table 2, the lack of robustness of some threshold-dependent methods such as SNR and *t*-test metrics in some cases could be actually ascribed to the improperness of some gene ranking metrics. In fact, for the prostate cancer data (oligo), according to the SAM and fold change metrics, the average of SS can reach high values as 0.72 and 0.92, respectively. We also present the results considering the two links equally (see Supplementary Table S6).

Two very closely related categories may share most of the annotated genes and thus share a comparable significance value. However, after applying a significance cutoff on the category list, some closely related but marginally significant

Table 2. The means and standard variances of DC and SS when using FDR ≤ 0.1 to define significant categories

		DC	SS
Leukemia	FC	0.64 ± 0.12	0.87 ± 0.06
	SNR	0.56 ± 0.14	0.78 ± 0.14
	t-test	0.53 ± 0.17	0.73 ± 0.22
	SAM	0.52 ± 0.18	0.75 ± 0.19
Prostate (oligo)	FC	0.76 ± 0.07	0.92 ± 0.03
	SNR	0.35 ± 0.22	0.58 ± 0.18
	t-test	0.37 ± 0.22	0.64 ± 0.16
	SAM	0.54 ± 0.26	0.72 ± 0.20
Liver	FC	0.74 ± 0.09	0.90 ± 0.05
	SNR	0.69 ± 0.14	0.86 ± 0.10
	t-test	0.66 ± 0.16	0.85 ± 0.10
	SAM	0.76 ± 0.09	0.90 ± 0.04
Prostate (cDNA)	FC	0.52 ± 0.16	0.71 ± 0.16
	SNR	0.57 ± 0.17	0.66 ± 0.16
	t-test	0.55 ± 0.27	0.68 ± 0.28
	SAM	0.56 ± 0.17	0.68 ± 0.19

categories can be identified as significant separately under two thresholds. Figure 1A exemplifies the differences when using DC and SS. The colored GO categories in Figure 1A are significant in the leukemia dataset. One group includes 49 categories (rectangle and hexagon) enriched with the top 1000 genes ranked by SAM and the other group includes 67 categories (diamond and hexagon) enriched with the top 3000 genes by the same ranking method. Both groups of significant categories are selected with FDR ≤ 0.1. The DC measure between these two groups is as low as 0.38. If we expend the criterion of overlap from identical category to including parent–child pairs, a DC-like measure can be calculated as 0.57. However, because the parent–child relationship is only one of the many relationships the GO categories can have, it is still unfair to truncate the other relationships to 0. As shown in Figure 1B, 'aminoglycan biosynthetic process' and 'chondroitin sulfate metabolic process' are marginally significant in both thresholds, while the former fails to reach FDR ≤ 0.1 by the top 1000 genes and the later is not identified by the top 3000 genes. For the DC measure, this pair will be scored as zero. However, the semantic similarity of the two categories is calculated to be as high as 0.80, which is comparable to the mean semantic similarity (as 0.84 ± 0.17) of the parent–child pairs in BP of GO. For the SS measure, the highly related functions of these two categories contribute much to the functional similarity of their corresponding groups. Accordingly, the SS measure between the two groups is 0.76, realistically reflecting their functional similarity.

Finally, because the six DEG thresholds are somewhat arbitrary, we further investigate the functional consistence of the significant categories identified using other 10 or 20 alternative DEG thresholds. Similar results can be observed (see Supplementary Tables S3 and S4). For two kinds of gene ranking metrics (SAM and *t*-test), we also use four FDR thresholds (0.1%, 1%, 5% and 10%) for selecting DEGs to carry out the above analysis. Analogous results are obtained (see Supplementary Table S5).

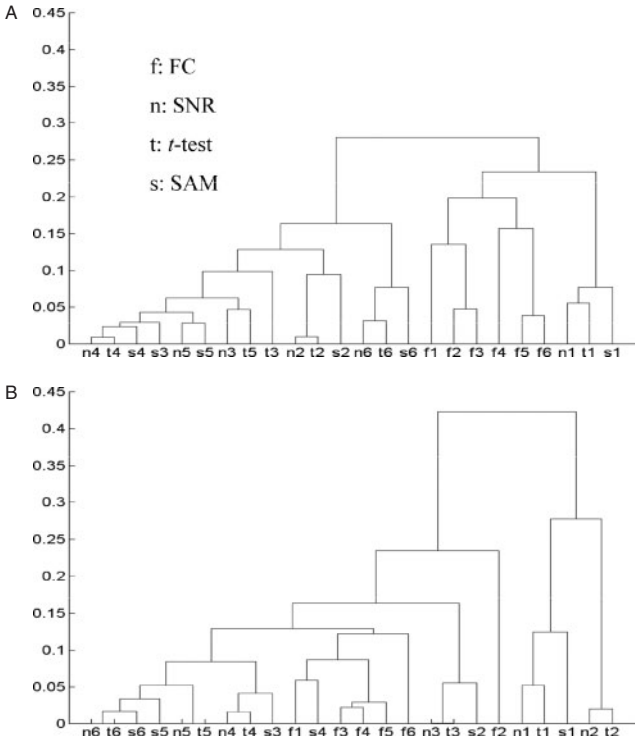

Fig. 2. Clustering significant categories produced by different gene ranking metrics and thresholds. Each gene ranking metric is applied to datasets of (**A**) leukemia (**B**) prostate cancer (oligo). Horizontal axis represents 24 groups being analyzed. Each group's label consists of two parts, a letter that stands for the metric used to rank the DEGs and a number for the grads corresponding to the number of top-ranked genes we selected out to do enrichment analysis.

3.2 Semantic similarity of categories enriched with DEGs selected by different gene ranking methods

In Table 1, we can observe that the fold change metric reaches the highest functional consistence through all the four datasets, while the SNR and *t*-test perform the worst. Here, we further study the consistence among four metrics by examining the similarity of categories enriched with DEGs selected by them, respectively.

We first calculate the SS similarity measure between every two groups of significant categories identified by gene lists from all ranking metrics and DEG thresholds. Then we apply the average linkage clustering algorithm, using (1–SS) as the distance measure, to cluster the significant categories corresponding to different ranking metrics and DEG thresholds. Taking the leukemia data for an example, three main clusters can be observed (Fig. 2A). The first cluster, sharing SS value around 0.92, consists of significant categories enriched with the genes selected under stringent thresholds (e.g. top 500 genes) for SNR, *t*-test and SAM. The second cluster, which shows

similarity value as high as 0.84, contains significant categories enriched with genes selected under more liberal thresholds (e.g. top 1000 to 3000 genes) for SNR, *t*-test and SAM. Categories identified by either SNR or *t*-test metrics tend to be inconsistent when the threshold moves from stringent to liberal. The similar result can also be found in liver and the two prostate cancer datasets. The third cluster with SS value of 0.8 contains results under most thresholds (top 1000 to 3000 genes) for the fold change metric. In the prostate (oligo, Fig. 2B), liver (Fig. S1A) and prostate (cDNA, Fig. S1B) cancer datasets, we can also observe the tendency that the results by fold change metric cluster together. Through four datasets, the fold change metric shares less consistence with other metrics except showing modest similarity to SAM under liberal DEG thresholds in prostate (oligo, Fig. 2B) as well as the liver and prostate (cDNA) cancer datasets. We also use 10 alternative DEG thresholds to carry out the above analysis. Analogous results are obtained (see Supplementary Fig. S1).

The above results suggest that (1) the SNR and *t*-test tend to characterize similar categories under similar thresholds for selecting DEG lists, while the categories characterized under liberal and stringent thresholds for either of these two gene ranking metrics are quite different; (2) although the fold change metric reaches the highest functional consistence across DEG threshold changes, it shows relative isolation to the other three metrics in the clustering dendrograms for all the four datasets. This result suggests the categories selected from the gene lists according to the fold change metric are less reproducible by other ranking metrics; (3) the categories characterized by SAM metric show acceptable functional consistence and at the same time tend to be more reproducible by fold change, SNR and *t*-test metrics.

4 DISCUSSION AND CONCLUSION

Functional analysis based on GO categories is now widely applied for knowledge discovery in microarray experiments. For the conventional threshold-dependent methods, simply counting overlapping categories suggest that the enrichment analyses may produce highly inconsistent results when using different DEG thresholds. However, from the view of semantic similarity values and their corresponding statistical significance levels, the sets of significant categories from varied DEG thresholds and even from some different gene ranking metrics are rather functionally similar. The essence of the difference of the two similarity measures for GO categories is that they differently address the nature of biologically relevance between ontological categories.

Some other semantic similarity measures could be used to evaluate the functional similarity between two groups of GO categories. For example, Frohlich *et al.* (Frohlich *et al.*, 2007) suggested three measures to estimate the functional similarity between two groups of GO categories separately annotated with two genes. One measure is calculated by the maximum or average of the semantic similarities for all the inter-group category pairs. On one hand, the maximum pairwise similarity may overestimate the functional similarity between two groups of categories as long as there is one overlap between the two groups, the measure reaches the maximum value as 1 regardless of the relations of the other category pairs. On the other hand, the average similarity tends to underestimate the similarity—the measure cannot reach the maximum even though the two groups of categories are identical. Another measure is to render each gene (corresponding to a category group) a feature vector by calculating its similarity to certain prototype genes (corresponding to category groups actually standing for a gold standard set). This measure is not proper for our problem since no gold standard set of categories is known for the data under this study. The third measure, optimal assignment gene similarity, is somewhat similar to our measure. However, its maximum value depends on the groups' sizes instead of being a constant value, which makes the measures for group pairs with different sizes incomparable. Therefore, the SS measure used in this article could be more suitable for functional consistence analysis.

In this article, for comparing gene ranking metrics, we perform most of our analyses using *n* top-ranked genes to keep the numbers of the overlaps between DEGs lists the same at any two thresholds. The main reason that we do not use FDR control levels is that, for different gene ranking metrics, the numbers of the overlaps between DEGs lists selected at two FDR levels are not the same, which will make it unfair to compare SS measures for different gene selection metrics. However, it should be noted that simply using *n* top-ranked genes for analyses lacks a sound statistical confidence measure. Therefore, in a real application to find the true set of associated GO categories that may be affected by gene expression changes, the FDR criterion that is more statistically sound should be applied for selecting DEGs. For the special purpose of this article to compare methods, in each dataset, we analyze at most 3000 top-ranked genes for each ranking metric. Based on the SAM algorithm, the FDRs of selecting 3000 DEGs are about 2.2%, 15%, 0.1% and 3% for the leukemia, prostate (oligo), liver and prostate (cDNA) datasets, respectively, which are within a reasonable range of FDR control levels for real applications.

As demonstrated in this study, although not conclusive, the threshold-dependent methods are actually rather robust to a wide range of gene selection thresholds for some gene ranking metrics such as SAM and fold change. Therefore, we could be more confident in believing that the threshold-dependent approaches can provide information about important functional aspects of changes in gene expression patterns. However, it should be noted that the functional consistence of the GO categories selected across a range of thresholds does not necessary prove that they are the true set of associated GO categories affected by gene expression changes. Consequently, some of the associated GO categories could be proven being false discoveries if gold-standard sets could be given. However, for a real application, the gold-standard set is usually unknown. Alternatively, for the categories significantly overrepresented in a gene list derived from a threshold-based approach, practical standard sets to compare with could be provided by other functional enrichment analyses approaches, such as the threshold-free GSEA, global test and some recently proposed methods aiming at reducing the high dependence between GO categories (Alexa *et al.*, 2006; Lewin and Grieve, 2006). It is worthy of studying whether the category sets derived

according to different criteria (such as those used in threshold-free approaches) tend to be functionally similar. Such a comparison study should be one interesting direction for our future works.

The proposed measure could be further applied to compare lists of associated GO categories obtained from different microarry datasets for a same disease. In this article, for the two prostate cancer datasets from different platforms, we take a preliminary study to see whether the functional module perspective can provide a more robust way to compare independently derived gene expression data. The data from these two studies show little in common. Using SAM at 10% FDR control level, the overlaps of the DEGs between these two datasets are only ~20%. Even in such a situation, the semantic similarity of the categories from these two datasets is still relatively high as 0.56, which is statistically significant ($P < 0.05$) according to 10 000 annotation permutations. In future work, we plan to apply a more comprehensive analysis addressing the functional consistence of DEG lists from different inter- and intra-platform datasets for a same disease.

Finally, the significant functional categories for cooperatively carrying out cellular functions in regulating diseases warrant future studies, e.g. by integrating other information such as protein–protein interaction (Guo *et al*., 2007; Ideker *et al*., 2002) and cellular localization data (Zhu *et al*., 2007).

ACKNOWLEDGEMENTS

We wish to thank the associate editor and two anonymous referees for their constructive advices and comments to improve this work. This work was supported in part by the National Natural Science Foundation of China (grant nos. 30370388 and 30670539).

Conflict of Interest: none declared.

REFERENCES

Al-Shahrour,F. *et al*. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.

Alexa,A. *et al*. (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, **22**, 1600–1607.

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*., **25**, 25–29.

Ben-Shaul,Y. *et al*. (2005) Identifying subtle interrelated changes in functional gene categories using continuous measures of gene expression. *Bioinformatics*, **21**, 1129–1137.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodological)*, **57**, 289–300.

Bolstad,B.M. *et al*. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Chen,X. *et al*. (2002) Gene expression patterns in human liver cancers. *Mol. Biol. Cell*, **13**, 1929–1939.

Draghici,S. *et al*. (2003) Global functional profiling of gene expression. *Genomics*, **81**, 98–104.

Frakes,W.B. and Baeza-Yates,R. (1992) *Information Retrieval: Data Structures and Algorithms*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA.

Frohlich,H. *et al*. (2007) GOSim – an R-package for computation of information theoretic GO similarities between terms and gene products. *BMC Bioinformatics*, **8**, 166.

Goeman,J.J. *et al*. (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, **20**, 93–99.

Guo,Z. *et al*. (2007) Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network. *Bioinformatics*, **23**, 2121–2128.

Hosack,D.A. *et al*. (2003) Identifying biological themes within lists of genes with EASE. *Genome Biol*., **4**, R70.

Ideker,T. *et al*. (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18** (Suppl. 1), S233–240.

Irizarry,R.A. *et al*. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Lapointe,J. *et al*. (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc. Natl Acad. Sci. USA*, **101**, 811–816.

Lewin,A. and Grieve,I.C. (2006) Grouping Gene Ontology terms to improve the assessment of gene set enrichment in microarray data. *BMC Bioinformatics*, **7**, 426.

Lin,D. (1998) An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*. San Francisco, CA, pp. 296–304.

Lord,P.W. *et al*. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**, 1275–1283.

Nilsson,B. *et al*. (2007) Threshold-free high-power methods for the ontological analysis of genome-wide gene-expression studies. *Genome Biol*., **8**, R74.

Pan,K.H. *et al*. (2005) Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays. *Proc. Natl Acad. Sci. USA*, **102**, 8961–8965.

Pehkonen,P. *et al*. (2005) Theme discovery from gene lists for identification and viewing of multiple functional groups. *BMC Bioinformatics*, **6**, 162.

Resnik,P. (1999) Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res*., **11**, 95–130.

Shannon,C.E. (2001) A mathematical theory of communication. *ACM SIGMOBILE Mob. Comput. Commun. Rev*., **5**, 3–55.

Singh,D. *et al*. (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203–209.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Troyanskaya,O. *et al*. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

Tusher,V.G. *et al*. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Valk,P.J. *et al*. (2004) Prognostically useful gene-expression profiles in acute myeloid leukemia. *N. Engl. J. Med*., **350**, 1617–1628.

Wang,D. *et al*. (2006) Effects of replacing the unreliable cDNA microarray measurements on the disease classification based on gene expression profiles and functional modules. *Bioinformatics*, **22**, 2883–2889.

Wang,H. et al. (2004) Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships. In *Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*. San Diego, CA, pp. 25–31.

Yang,Y.H. *et al*. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*., **30**, e15.

Zhu,J. *et al*. (2007) GO-2D: identifying 2-dimensional cellular-localized functional modules in Gene Ontology. *BMC Genomics*, **8**, 30.