Efficient Generation of Biologically Relevant Enriched Gene Sets

Igor Trajkovski and Nada Lavrač

Department of Knowledge Technologies, Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
{igor.trajkovski,nada.lavrac}@ijs.si

Abstract. Gene set enrichment analysis is a microarray data analysis method that uses predefined gene sets and ranks of genes to identify significant biological changes in microarray data sets. In this paper we present a novel method integrating gene interaction information with Gene Ontology (GO) for the construction of new interesting enriched gene sets. The experimental results show that the introduced method improves over traditional methods that compute the enrichment of a single GO terms, i.e. that it is capable to find new statistically relevant descriptions of the biology governing the experiments not detectable by the existing methods.

1 Introduction

High-throughput technologies such as DNA microarrays and proteomics are revolutionizing biology and medicine. Global gene expression profiling using microarrays monitors changes in the expression of thousands of genes simultaneously. The large amounts of data acquired must then be reduced or “translated” to a smaller set of genes representing meaningful biological differences between control and test systems and validated in an experimental or clinical setting.

In a typical experiment, mRNA expression profiles are generated for thousands of genes from a collection of samples belonging to one of two classes - for example, tumors that are sensitive vs. those resistant to a drug. The genes can be ordered in a ranked list L, according to the difference of expression between the classes. The challenge is to extract the meaning from this list.

A common approach involves focusing on a handful of genes at the top of L (genes showing the largest difference in its expression between the classes), to extract the underlying biology responsible for the phenotypic differences. This approach has a few major limitations:

- After correcting for multiple hypotheses testing, no individual gene may meet the threshold for statistical significance, because the relevant biological differences are small relative to the noise inherent to the microarray technology.
- The opposite situation, one may be left with a long list of statistically significant genes without any common biological function.

I. Mandoiu and A. Zelikovsky (Eds.): ISBRA 2007, LNBI 4463, pp. 248–259, 2007.

- Single-gene analysis may miss important effects on pathways. Cellular processes often affect sets of genes acting jointly. An increase of 20% in all genes encoding members of a biological process may dramatically alter the execution of that process, and its impact on other processes, than a 10-fold increase in a single gene.
- It is not a rare case when different groups studying the same biological system, report a list of statistically significant genes from the two studies that have a significantly small overlap.

To overcome these analytical challenges, recently method was developed, called Gene Set Enrichment Analysis (GSEA) [1], that evaluates microarray data at the level of gene sets. Biologically interesting sets of genes, for example genes that belong to a pathway or genes known to have the same biological function, are good examples of such gene sets. The most popular choice for gene sets are genes annotated with some GO term. The goal of GSEA is to determine whether members of a gene set S tend to occur toward the top of the list L, in which case the gene set is correlated with the phenotypic class distinction.

In this work we propose a method for generating new gene sets that have relevant biological interpretations, by combining the existing gene sets, and by inclusion of gene-gene interaction information available from the public gene annotation databases. The experimental results show that our method can find descriptions of interesting enriched gene sets, that traditional methods are unable to discover. We applied the proposed method to three gene expression data sets with the following respective sets of sample classes: (i) acute lymphoblastic leukemia (ALL) vs. acute myeloid leukemia (AML), (ii) seven subtypes of ALL, and (iii) fourteen different types of cancers. Significant number of discovered gene sets have description which highlights the underlying biology that is responsible for distinguishing one class from the other classes.

The paper is organized as follows. In Section 2 we give background information about Gene Ontology and differentially expressed genes. Section 3 provides details of the Gene Set Enrichment Analysis. Section 4 presents the idea of our approach, and the steps taken in the construction of interesting gene sets. Section 5 presents the results of the experiments. In Section 6 we draw some final conclusions.

2 Background

2.1 Gene Ontologies

One of the most important tools for the representation and processing of information about gene products and functions is the Gene Ontology (GO)[1]. It provides a controlled vocabulary for the description of cellular components, molecular functions, and biological processes.

[1] http://www.geneontology.org

As of December 2006, GO contains 1864 cellular component, 7513 molecular function and 12549 biological process terms. Terms are organized in parent-child hierarchies (see Fig. 1), indicating either that one term is more specific than another (is_a) or that the entity denoted by one term is part of the entity denoted by another (part_of). Typically, such associations (or "annotations") are first of all established electronically and later validated by a process of manual verification which requires the annotator to have expertise both in the biology of the genes and gene products and in the structure and content of GO. GO, in spite of its name, is not an ontology in the sense accepted by computer scientists, in that it does not deal with axioms and formalized definitions associated to terms. It is rather a taxonomy, or, as the GO Consortium puts it, a "controlled vocabulary" providing a practically useful framework for keeping track of the biological annotations applied to gene products.

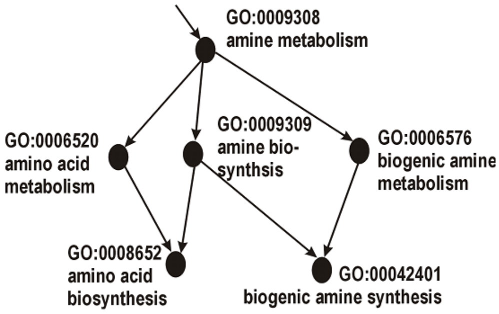

Fig. 1. This figure shows a part of GO providing the annotations concerning amine metabolism

2.2 Differentially Expressed Genes

Differentially expressed genes are genes that are expressed differently (relative. to the reference) between the conditions of interest. In the context of finding differentially expressed genes, the null hypothesis for each gene is that it is not differentially expressed between two conditions, usually against the two-sided alternative hypothesis that the gene is up- or down regulated. The most commonly used statistical test in this setting has been the two-sample t-test [3] [4], although other statistics such as the signal-to-noise ratios [2], or Pearson's correlation [5], have often been used.

Let $T(g, c)$ denote the t-test score of gene g for a target class c, which is computed by the following procedure: $[\mu_1(g), \sigma_1(g)]$ and $[\mu_2(g), \sigma_2(g)]$ denote the means and standard deviations of the logarithm of the expression levels of gene g for the samples in class c and samples in $C \setminus c$, respectively. Also, let $N_1 = |c|$ and $N_2 = |C \setminus c|$. $T(g, c)$ is computed by the following formula:

$$T(g, c) = \frac{\mu_1(g) - \mu_2(g)}{\sqrt{\frac{\sigma_1(g)}{N_1} + \frac{\sigma_2(g)}{N_2}}} \qquad (1)$$

which reflects the difference between the classes relative to the standard deviation within the classes. Large values of $|T(g,c)|$ indicate a strong correlation between the gene expression and the class distinction, while the sign of $T(g,c)$ being positive or negative corresponds to g being more highly expressed in class c or in other classes.

3 Gene Set Enrichment Analysis (GSEA)

GSEA considers experiments with gene expression profiles from samples belonging to two classes. First, genes are ranked based on the correlation between their expression and the class distinction by using any suitable metric, for instance computed by signal-to-noise ratios or Pearson's correlation. In our experiments we ranked the genes according to their t-score value.

Given a predefined set of genes S (e.g. genes involved in some specific biological process) and ranked gene list L, the goal of GSEA is to determine whether the members of S are randomly distributed throughout L or primarily found at the top of the list.

There are two major steps in the GSEA method:

1. **Calculation of an Enrichment Score.** Enrichment score (ES) reflects the degree to which a set S is overrepresented at the top of the ranked list L. The score is calculated by walking down the list L, increasing a running-sum statistic when encountering a gene in S and decreasing it when gene is not in S. The magnitude of the increment depends on the size of S and the total number of genes N. The enrichment score is the maximum deviation from zero encountered in the random walk. If $L = (g_1, g_2, ..., g_N)$ is a ranked list of genes, according to their t-score, enrichment score ES is calculated as:

$$Hit(S,i) = \sum_{\substack{g_j \in S \\ 1 \le j \le i}} \frac{1}{|S|} \qquad Miss(S,i) = \sum_{\substack{g_j \in S \\ 1 \le j \le i}} \frac{1}{N - |S|}$$

$$ES(S) = \max_{1 \le i \le N} |Hit(S,i) - Miss(S,i)| \tag{2}$$

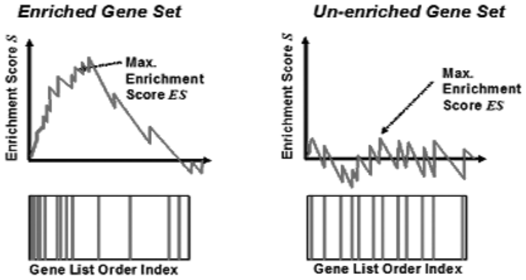

Fig. 2. 'Spectral lines' show the position of gene set members on the ranked gene list

2. **Estimation of Significance Level of ES.** The statistical significance of the ES is computed by using an empirical phenotype-based permutation test procedure that preserves the complex correlation structure of the gene expression data. Specifically, one permutes the phenotype labels and recomputes the ES of the gene set for the permuted data, which generates a null distribution for the ES. The empirical, p-value of the observed ES is then calculated relative to this null distribution. Importantly, the permutation of class labels preserves gene-gene correlations and, thus, provides a more biologically reasonable assessment of significance than would be obtained by permuting genes.

4 Generation of New Gene Sets

Methods that test for enrichment of GO terms have been proposed by [6], [7], [8] and [9]. A comparative study of commonly used tools for analyzing GO term enrichment was presented by [10]. [11] presented two novel algorithms that improve GO term scoring using the underlying GO graph topology.

None of the papers includes the gene interaction information, and none of them presents a method for the construction of novel gene sets, but rather they just calculate the enrichment of an a-priory given list of gene sets.

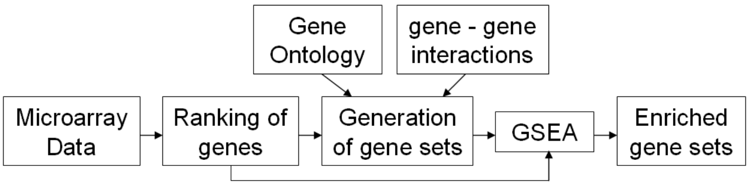

Fig. 3. Data flow of the proposed method for generation of enriched gene sets

First, let us mention some properties of the gene annotations with GO terms:

- one gene can be annotated with several GO terms,
- a GO term may have thousands of genes annotated to it,
- if a gene is annotated with a GO term A then it is annotated with all ancestors of A.

From this information, we can conclude that each GO term defines one gene set, that one gene can be member of several gene sets, and that some gene sets are subsets of other gene sets.

Second, let *Func*, *Proc* and *Comp* denote the sets of gene sets that are defined by the GO terms that are a subterm of the term “molecular function”, “biological process” and “cellular component”, respectively.

Our method relies on two ideas, that are used in the construction of new gene sets:

- **Inclusion of gene interaction information.** There are cases when some abrupted processes are not detectable by the enrichment score, one reason can be that the genes had a slight increase/decrease in their expression, but had a much larger effect on the interacting genes. Therefore we think that it is reasonable to construct gene set whose members interact with another gene set. Gene-gene interaction information is provided for pairs of genes for which there is an evidence that their expression levels are correlated, determined by analysis of microarray data or other experimental methods.

  Formally: if $G_1 \in Func$ (or $Proc, Comp$, respectively), then $G_2 = \{g_2 | g_2$ is a gene, and g_2 interacts with $g_1 \in G_1$ $\}$ was added to $Func$ (or $Proc, Comp$).

- **Intersection of gene sets.** There are cases where two or three given gene sets are not significantly enriched, but their intersection is significantly enriched.

  Formally: if $G_1 \in Func, G_2 \in Proc$ and $G_3 \in Comp$, then $G_4 = G_1 \bigcap G_2 \bigcap G_3$ is a new defined gene set.

  For example, it can happen that gene set defined by the molecular function F is not enriched, because a lot of genes in different parts of the cell execute it and one can not expect that all of them will be over/under expressed, but if genes with that function in some specific part of the cell C_{part} are abnormally active, then it can be elegantly captured by the following gene set:

  $$\text{Func(F)} \bigcap \text{Comp}(C_{part}).$$

  Note that all genes are annotated with the top three GO terms: “molecular function”, “biological process” and “cellular component”, which means that top three GO terms contain all the genes.

The newly defined gene sets are interpreted very intuitively. For example, the gene set defined as intersection of a “functional” term A and “process” term B:

$$\text{function(A), process(B)} \equiv \text{Func(A)} \bigcap \text{Proc(B)}$$

is interpreted as: *Genes that are part of the process B and have function A*, or:

$$\text{int(process(A))} \equiv \{g_2 | g_2 \text{ is a gene, and } g_2 \text{ interacts with } g_1 \in \text{Proc(A)}\}$$

is interpreted as: *Genes that interact with genes that are members of process A*, or:

$$\text{int(process(A),component(B))} \equiv \text{int(Proc(A))} \bigcap \text{int(Comp(B))}$$

is interpreted as: *Genes that interact with genes that are members of process A, and genes that operate in cellular component B.*

The number of the newly defined gene sets is huge. In December 2006, $|Func|$ = 7513, $|Proc|$ = 12549 and $|Comp|$ = 1846. After the inclusion of the gene interaction information, the size of these sets is doubled. Then the number of newly generated gene sets is:

$$2^3 \times |Func| \times |Proc| \times |Comp| \approx 1.4 \times 10^{12}$$

For each of these sets we need to compute its enrichment score, ES, that takes linear time in the number of genes ($\approx 2 \times 10^4$), we get $\approx 3 \times 10^{16}$ floating operations. If we want to statistically validate founded enriched gene sets, usually with 1000 permutation tests, we get $\approx 10^{20}$ operations, that is well above the average performance of today PC's. Therefore we need to efficiently search the space of newly generated gene sets for possible enriched gene sets.

The first idea for improvement is that we are not interested in generating all possible gene sets, but only those that are potentially enriched, and have some minimum number of genes at the top of the list, for example 5 in the first 100, or 10 in the first 300 genes of the list (this was the constraint used in our experiments). That is a weak constraint concerning the biological interpretation of the results, because we are not really interested in the gene sets that do not have this number of genes at the top of the list, but it is a hard constraint concerning the pruning of the search space of all gene sets. By having this constraint we can use the GO topology to efficiently generate all gene sets that satisfy it.

GO is a directed acyclic graph, the root of the graph is the most general term, which means that if one term (gene set) does not satisfy our constraint, than all its descendants will also not satisfy it, because they cover a subset of the genes covered by the given term. In this way we can significantly prune the search space of possible enriched gene sets. Therefor, we first try to construct gene sets from the top nodes of the GO, and if we fail we do not refine the last added term that did not satisfy our constraint.

After the construction of the gene sets that satisfy our constraint, we calculate their ES value, and statistically validate this values using the permutation testing.

In the original version of Kolmogorov-Smirnov test, used by GSEA, the ES statistic used equal weights at every step, which yielded high scores for sets clustered near the middle of the ranked list. These sets do not represent biologically relevant correlation with the phenotype. We addressed this issue by weighting the steps according to each genes correlation with a phenotype. Like in the original version we first rank the N genes to form L = $(g_1, g_2, \ldots, g_N)$ according to their t-score, $t(g_j) = t_j$, of their expression profiles with class c. Then the running sum Hit is computed by following formula:

$$Hit(S, i) = \sum_{\substack{g_j \in S \\ 1 \le j \le i}} \frac{|t_j|}{\sum_{g_j \in S} |t_j|}$$

The other running sum, $Miss$, and ES statistic were calculated with the original formulas.

5 Experiments

We applied the proposed methodology to three classification problems from gene expression data, with the aim to describe the most important biological processes that are responsible for class differentiation.

The first problem was introduced in [2] and aims at distinguishing between samples of ALL and AML from gene expression profiles. The second problem was described in [12] and aims at distinguishing different subtypes of ALL (6 recognized subtypes plus a separate class 'other' containing the remaining samples). The third problem was defined in [13]. Here one tries to distinguish among 14 classes of cancers from gene expression profiles. Gene annotations and interaction data was downloaded from Entrez database ftp://ftp.ncbi.nlm.nih.gov/gene/.

Note that this paper does not address the learning task of discriminating between the classes. Instead, for the given target class we aim at finding relevant enriched gene sets that can capture the underlying biology characteristic for that class.

Table 1. Some of the enriched gene sets in the first dataset, with p-value ≤ 0.001

Class	Gene Set	ES
ALL	1. int(Func('zinc ion binding'), Comp('chromosomal part'), Proc('interphase of mitotic cell cycle'))	0.60
	2. Proc('DNA metabolism')	0.59
	3. int(Func('RNA polymerase II transcription factor activity'), Proc('ubiquitin cycle'), Comp('intracellular non-membrane-bound organelle'))	0.56
	4. int(Func('ATP binding'), Comp('chromosomal part'), Proc('DNA replication'))	0.55
AML	1. int(Func('metal ion binding'), Comp('cell surface'), Proc('response to pest, pathogen or parasite'))	0.54
	2. int(Comp('lysosome'))	0.53
	3. Proc('inflammatory response')	0.51
	4. int(Proc('inflammatory response'), Comp('cell surface'))	0.51

5.1 Experimental Results

To illustrate the straightforward interpretability of the enriched gene sets found by our approach, we provide the best-scoring gene sets for some of the target classes in the mentioned three classification problems (see Table 1, 3 and 4). We should mention that enriched gene sets that include too general GO terms (i.e. "biological function", "protein binding", "cellular physiological process", "cytoplasm", etc.), were removed from the result list.

For comparison of enrichment of the found gene sets with the gene sets defined by a single GO term, in Table 2 we list the most enriched gene sets defined by a single GO term, for the first dataset. We can see that ES of the single GO terms is much smaller then the ES of the newly constructed gene sets, and most

Table 2. Summary of GSEA results for the first dataset, with p-value ≤ 0.005. Gene sets constructed from a single GO term.

Class	Gene Set	ES
ALL	1. Proc(’DNA metabolism’)	0.59
	2. Comp(’intracellular non-membrane-bound organelle’)	0.35
	3. Proc(’development’)	0.22
	4. Comp(’cytoplasmic part’)	0.22
	4. Proc(’transport’)	0.22
AML	1. Proc(’inflammatory response’)	0.51
	2. Proc(’response to chemical stimulus’)	0.41
	3. Proc(’proteolysis’)	0.38
	4. Proc(’cell communication’)	0.33

Table 3. Some of the enriched gene sets in the second data set, with p-value ≤ 0.001

CLASS	Gene Set	ES
BRC	1. Proc(’cell adhesion’),Comp(’integral to membrane’)	0.56
	2. int(Func(’zinc ion binding’), Proc(’cell surface receptor linked signal transd.’), Comp(’endoplasmic reticulum’))	0.54
	3. int(Func(’metal ion binding’), Proc(’cell migration’),Comp(’membrane’))	0.53
E2A	1. int(Func(’calc. ion bind.’), Proc(’protein kinase casc.’), Comp(’intracellular membrane-bound organelle’))	0.57
	2. Proc(’cell adhesion’),Comp(’membrane’)	0.57
	3. int(Func(’ATP binding’), Comp(’integral to membrane’), Proc(’reg. of transcr. from RNA poly. II promoter’))	0.55
MLL	1. int(Func(’protein kinase activity’), Comp(’mem. fraction’), Proc(’transmem.rec.protein.tyros.kinase.sig.path.’))	0.63
	2. int(Func(’SH3/SH2 adaptor activity’), Proc(’apoptosis’), Comp(’cytoskeleton’))	0.61
T_ALL	1. int(Func(’protein-tyrosine kinase activity’), Proc(’positive regulation of T cell proliferation’), Comp(’immunological synapse’))	0.92
	2. Proc(’antigen presentation’), Comp(’integral to membrane’)	0.88
TEL	1. int(Proc(’cell adhesion’), Comp(’cell junction’))	0.65
	2. int(Proc(’synaptic transmission’), Comp(’cytoskeleton’))	0.57

importantly, the found gene sets are constructed from not enriched GO terms. Similar results we got for the other two datasets.

5.2 Statistical Validation

The following procedure calculated the significance of an observed ES by comparing it with the set of scores ES_{NULL} computed with randomly assigned phenotypes:

1. Randomly assign the original phenotype labels to samples, reorder genes according to their t-score values, and re-compute ES(S).
2. Repeat step 1 for 1,000 permutations, and create a histogram of the corresponding **maximum** enrichment scores ES_{NULL}.
3. Estimate the p-value for the ES value of the gene set S from ES_{NULL} by using the histogram computed at step 2. If there was not a case where random labeling of the examples give bigger ES value, then p-value < 0.001.

We use class labeled permutation because it preserves gene-gene correlations and, thus, provides a more biologically reasonable assessment of significance than would be obtained by permuting genes. Importantly, the permutation of class labels preserves gene-gene correlations and, thus, provides a more biologically reasonable assessment of significance than would be obtained by permuting genes.

Table 4. Some of the enriched gene sets in the third data set, with p-value ≤ 0.001

CLASS	Gene Set	ES
BREAST	1. `Func('RNA binding')`	1.03
	2. `int(Func('zinc ion binding'),Comp('nuclear part'))`	0.79
	3. `int(Func('RNA.polym.II.trans.fact.act.'),Comp('nucleus'), Proc('reg. of transcr. from RNA poly. II promoter'))`	0.76
CNS	1. `Func('struc.const.of.ribosome'),Proc('protein biosynth.')`	2.06
	2. `int(Func('actin binding'),Comp('cytoskeletal part'))`	1.33
COLO.	1. `int(Comp('extracellular matrix (sensu Metazoa)'))`	0.62
LYMPH.	1. `int(Func('transmem.rec.activ.'),Comp('int.to.plasma.mem.') Proc('posit.reg. of I-kappaB kinase/NF-kappaB casc.'))`	1.78
MELAN.	1. `int(Func('transcription cofactor activity'), Proc('muscle development'), Comp('nucleus'))`	1.20
MET	1. `int(Proc('MAPKKK cascade'),Comp('membrane'))`	0.45
OVARY	1. `int(Func('zinc ion binding'),Comp('membrane fraction') Proc('phosphorylation'))`	0.45
	2. `int(Func('zinc ion binding'),Comp('integral to membrane') Proc('cell growth'))`	0.42
PANCR.	1. `Proc('proteolysis')`	0.51
	2. `Comp('ribonucleoprotein complex')`	0.50
PROST.	1. `int(Func('androgen receptor binding'),Comp('nucleus'), Proc('reg. of transcr. from RNA poly. II promoter'))`	0.50
	2. `Comp('cytoskeletal part')`	0.49
RENAL	1. `int(Proc('insulin receptor signaling pathway'), Comp('intracellular membrane-bound organelle'))`	0.43
	2. `int(Func('protein-tyr. kinase activ.'),Comp('cyto. part') Proc('regulation of cell growth'))`	0.43

6 Conclusion

We addressed the problem of finding enriched functional groups of genes based on gene expression data. We proposed a novel method for integrating the gene interaction information into the construction of new interesting relevant gene sets. The experimental results show that the introduced method improves over existing methods, and we base our conclusion on the following facts:

- ES of the newly constructed sets are higher then the ES of any single GO terms.
- Newly constructed sets are composed of non-enriched GO terms, which means that we are extracting additional biological knowledge that can not be found by single GO term GSEA.
- This method is generalization of the traditional methods. If we turn-off gene-gene interactions and combination of GO terms, we will get classical single GO term GSEA.

We believe that the strength of the proposed method will be even bigger through the expected increase in both the quality and quantity of gene annotations and gene-gene interaction information in the near future.

Acknowledgment

This research was supported by the Slovenian Ministry of Higher Education, Science and Technology.

References

1. Subramanian A., et al. (2005) Gene set enrichment analysis: A knowledgebased approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. of the U.S.A., 102(43):15545-15550.
2. Golub T. R., Slonim D. K., Tamayo P., Huard C., Gaasenbeek M. et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science, 286:5439, 531-537.
3. Snedecor G. W. and Cochran W. G. (1989) Statistical Methods, Eighth Edition, Iowa State University Press.
4. Tsai C. A., Chen Y. J. and Chen J. J. (2003) Testing for differentially expressed genes with microarray data. Nucleic Acids Res 31, e52.
5. Troyanskaya O. G., et al. (2002) Nonparametric methods for identifying differentially expressed genes in microarray data. Bioinformatics 18(11):1454-61.
6. Draghici S., , et al. (2003) Global functional profiling of gene expression. Genomics, 81:98-104.
7. Zeeberg B. R., et al. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biology, 4(4):R28.
8. 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.
9. Beissbarth T. and Speed T. (2004) GOstat: Find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics, 1(1):1-2.

10. Khatri P. and Draghici S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics, 21(18):3587-3595.
11. Alexa A., et al. (2006) Improved Scoring of Functional Groups from Gene Expression Data by Decorrelating GO Graph Structure. Bioinformatics, 22(13):1600-1607.
12. Ross M. E., et al. (2003) Classification of pediatric acute lymphoblastic leukemia by gene expression profile. BLOOD, pp. 2951-2959.
13. Ramaswamy S., et al. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. Proc. Natl. Acad. Sci. of the U.S.A. 18;98(26):15149-54.