Available online at www.sciencedirect.com

ScienceDirect

Journal of Biomedical Informatics xxx (2008) xxx–xxx

Journal of Biomedical Informatics

www.elsevier.com/locate/yjbin

ELSEVIER

SEGS: Search for enriched gene sets in microarray data

Igor Trajkovski [a,*], Nada Lavrač [a,b], Jakub Tolar [c]

[a] Department of Knowledge Technologies, Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[b] University of Nova Gorica, Vipavska 13, Nova Gorica, Slovenia
[c] Division of Hematology-Oncology and Blood and Marrow Transplantation, University of Minnesota, USA

Received 22 May 2007

Abstract

Gene Ontology (GO) terms are often used to interpret the results of microarray experiments. The most common approach is to perform Fisher's exact tests to find gene sets annotated by GO terms which are over-represented among the genes declared to be differentially expressed in the analysis of microarray data. Another way is to apply Gene Set Enrichment Analysis (GSEA) that uses predefined gene sets and ranks of genes to identify significant biological changes in microarray data sets. However, after correcting for multiple hypotheses testing, few (or no) GO terms may meet the threshold for statistical significance, because the relevant biological differences are small relative to the noise inherent to the microarray technology. In addition to the individual GO terms, we propose testing of gene sets constructed as intersections of GO terms, Kyoto Encyclopedia of Genes and Genomes Orthology (KO) terms, and gene sets constructed by using gene–gene interaction data obtained from the ENTREZ database. Our method finds gene sets that are significantly over-represented among differentially expressed genes which cannot be found by the standard enrichment testing methods applied on individual GO and KO terms, thus improving the enrichment analysis of microarray data.

Keywords: Microarray data analysis; Ontology; Gene set enrichment

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 outcome of such studies is usually a list of genes whose expression varies between different conditions and therefore may be of interest for further analysis. Lately, databases of other information about genes are used in order to provide additional inference. Two of the most used are Gene Ontology (GO) [1], and Kyoto Encyclopedia of Genes and Genomes (KEGG) [2].

Gene Ontology (GO) is a controlled vocabulary of standardized biological terms used to annotate gene products. It comprises several thousand terms, divided in three branches: Molecular Function, Biological Process and Cellular Component. KEGG Orthology (KO) is a collection of manually drawn pathway maps representing the knowledge on the molecular interaction and reaction networks for Metabolism, Genetic Information Processing, Environmental Information Processing, Cellular Processes and Human Diseases.

Tests for gene set enrichment compare lists of differentially expressed (DE) genes and non-DE genes to find which gene sets annotated by GO and KO terms are over- or under-represented amongst the DE genes. Several research groups have developed software to carry out Fisher's exact tests to find which gene sets are over-represented among the genes found to be differentially expressed, e.g., [4,5] and other works cited in [6]. The Fisher's test for term T essentially compares the proportion of DE genes

[*] Corresponding author.
E-mail addresses: igor.trajkovski@ijs.si (I. Trajkovski), nada.lavrac@ijs.si (N. Lavrač), tolar003@umn.edu (J. Tolar).

1532-0464/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2007.12.001

annotated by term *T* with the proportion of non-DE genes annotated by term *T*. Since there is a test for each of several thousands of GO nodes, and hundreds of KO nodes, multiple hypothesis testing must be taken into account. This is usually done by the Bonferroni correction or a more sophisticated correction controlling the False Discovery Rate (FDR). Benjamini and Hochberg's method [7] gives valid control of the FDR even when the different tests are dependent.

Approaches based on Fisher's exact testing have some major limitations:

- After correcting for multiple hypothesis testing, in selecting DE genes, no individual gene may meet the threshold for statistical significance, because the relevant biological differences are small relative to the inherent microarray technology noise.
- The opposite situation, one may be left with a long list of statistically significant genes without any common biological function, so none of the gene sets annotated by GO and KO terms is significantly enriched.
- Single gene analysis may miss important effects on pathways. Biological pathways often affect sets of genes acting jointly. An increase of 20% in the expression of all gene members of a biological pathway can alter the execution of that pathway, and its impact on other processes, significantly more than a 10-fold increase in a single gene [8].
- It is not rare that different research groups studying the same biological system report lists of DE genes they found to be statistically significant which have just a small overlap [11].
- Since all genes annotated by a given GO term are also annotated by all of its parents, closely related nodes may be found separately significant [15].
- Specific GO terms have few genes annotated, so there is often not enough statistical evidence to find these terms as statistically significant. The more general the GO term, the more genes are annotated by it, but the less useful the term is as an indication of the function of the differentially expressed genes [12].

The described problems have recently triggered the development of numerous methods described below.

1.1. Related work

Several methods have been developed recently to overcome the analytical challenges presented in the previous section. For improving the sensitivity of enrichment detection, Gene Set Enrichment Analysis (GSEA) [9] and Parametric Analysis of Gene Set Enrichment (PAGE) [13] were developed. GSEA calculates an enrichment score (ES) for a given gene set using ranks of genes and infers the statistical significance of ES against the ES-background distribution calculated by permutating the labels of the original data set. In the new version of GSEA, GSEA-P [10], there is an option for importing gene sets from MSigDB (Molecular Signatures Database) and testing them for enrichment, by that increasing the probability for finding enriched gene sets.

In contrast, PAGE calculates a *Z*-score for a given gene set from a parameter such as *t*-score value calculated on the basis of two experimental groups and infers statistical significance of the *Z*-score against the standard normal distribution. These two methods are capable to find enriched gene sets, not detectable by the standard Fisher's exact test.

Grossmann et al. [14] take into account the hierarchical structure of the GO by measuring the over-representation of each term relative to its parent terms. Alexa et al. [15] downweight the contribution of genes to the calculation of over-representation of a term if the children of that term have already been found significantly enriched. These two methods do not improve the statistical power, as the number of genes in each hypothesis test will be smaller than in the usual term-by-term tests, as double counting is penalized. However, they do help to improve the interpretation, since they produce just one (or at least not too many) significant *p*-values for each significant region of the graph. Levin et al. [12] use grouping of similar GO terms (which are close in the GO graph) in order to increase the statistical power. The reason is that the lower terms in the GO have few genes annotated by it, and can not be found statistically significantly enriched. Therefore, the authors of [12] group several terms to increase the size of the gene sets tested for enrichment. This approach is useful and can find enriched gene sets not detectable by standard screening of GO terms, but it is different form ours: we construct new gene sets as intersection of gene sets defined by Molecular Function, Biological Processes and Cellular Component terms of GO and KO terms, whereas [12] create new gene sets by making union of similar terms in GO. Concerning the usage of KO term in enrichment analysis, the work of Mao et al. [3] uses KO terms for automated annotation of large sets of genes, including whole genomes, and automated identification of pathways. This is done by identifying both the most frequent and the statistically significantly enriched pathways.

1.2. The proposed SEGS approach

In this work, we propose a novel approach for searching of enriched gene sets (SEGS) which proves to further improve the gene set enrichment results and by that the interpretation of gene expression data. Our approach is based on the efficient generation of new biologically relevant gene sets, that are tested for possible enrichment. The new gene sets are generated as intersections of GO and KO terms and gene sets defined with the help of gene–gene interaction data. Testing the enrichment of these gene sets with the standard methods (Fisher's exact test, GSEA and PAGE) shows that our method finds gene sets constructed from GO and KO terms significantly over-represented amongst differentially expressed genes, while these

GO and KO terms are not found to be enriched by Fisher's test, GSEA or PAGE, thus improving the enrichment analysis of microarray data.

The paper is organized as follows. Section 2 gives some background information about the publicly available resources of biological knowledge, followed by the methods for finding DE genes and methods for testing the gene set enrichment: Fisher's exact test, GSEA and PAGE. Section 3 presents the main idea of our SEGS approach, and the methodological steps taken in the construction of the new gene sets. Section 4 presents the results of the experiments and in Section 5 we draw the main conclusions and plans for further work.

2. Background

In this section, we first provide background information about the resources of biological knowledge, distributed across several publicly available databases. Then we present the most popular methods for finding the differentially expressed genes and calculating the gene set enrichment.

2.1. Resources of biological knowledge

2.1.1. Gene Ontology

Gene Ontology (GO)[1] is a database of standardized biological terms used to annotate gene products. In total it comprises about 23,000 terms,[2] divided in three branches: Molecular Function, Biological Process and Cellular Component. Each branch can be represented as a directed acyclic graph (DAG) relating terms (or nodes) of different degrees of specificity, with directed links from less specific to more specific terms. Each node in the graph can have several parents (broader related terms) and children (more specific related terms). See Fig. 1 presenting a small section of the GO graph. Annotation of a gene by any node A implies its automatic annotation by all ancestors of A (the set of broader terms related to A by directed paths). Genes can be annotated by several terms, however note that many genes have not been annotated at all.

2.1.2. Kyoto Encyclopedia of Genes and Genomes Ortology

Kyoto Encyclopedia of Genes and Genomes (KEGG) includes KEGG orthology (KO)s[3] that is a database of manually drawn pathway maps representing the knowledge on the molecular interaction and reaction networks. A metabolic pathway is a series of chemical reactions occurring within a cell, catalyzed by enzymes (genes), resulting in either the formation of a metabolic product to be used or stored by the cell, or the initiation of another metabolic pathway. That for each KEGG pathway (KO term) defines a set of genes that can be considered for statistical enrichment testing and by that detecting disrupted pathways. The KO is structured as a DAG hierarchy of four flat levels. The top level consists of the following five categories: Metabolism, Genetic Information Processing, Environmental Information Processing, Cellular Processes and Human Diseases. The second level divides the five functional categories into finer sub-categories. The third level corresponds directly to the KEGG pathways, and consists of 272 terms.[4] See Fig. 2 for an example of a small section of the KO hierarchy. Note that some of the KO terms appear also as process terms in the GO.

2.1.3. ENTREZ

ENTREZ[5] is a database that provides various information about genes and their products, including gene annotations with GO and KO terms (see Fig. 3) and gene-gene interaction data. As the collection of interaction data in a consistent, well-annotated format is essential for discovering of gene functions and benchmarking of high throughput interaction studies, a number of gene–gene interaction databases were developed. The examples of such a databases are BIND,[6] BioGRID,[7] EcoCyc[8] and HPRD.[9] ENTREZ (among other functionality) is a repository of these interaction databases, to house and distribute comprehensive collections of gene–gene interactions. The number of all gene-gene interactions in ENTEZ is about 118,000.[10]

2.2. Methods for finding differentially expressed genes

Selection of DE genes is the first step performed in the functional interpretation of microarray data. DE genes are the genes that are expressed differently (relative to the reference) between the given classes of microarray data. The most frequently used algorithms for the selection of DE genes are presented below. Mathematical definitions used by these methods are given in Fig. 4.

2.2.1. Fold change method

The simplest, non-statistical test method used for the selection of DE genes is the fold change method. In this method, the ratios between expression levels in two conditions are evaluated. All genes with a ratio of expression level higher than an arbitrary cut-off value are considered to be differentially expressed. The fold change method in its original form can be strongly biased by an inappropriate normalization. This problem has been addressed by the development of intensity-specific thresholds [16]. However, as this simple method is not a statistical test, it has no asso-

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

[2] This number of terms was available in September 2007.

[3] http://www.genome.jp/kegg/pathway.html

[4] This number of terms was available in KO in September 2007.

[5] ftp://ftp.ncbi.nlm.nih.gov/gene/

[6] http://www.bind.ca

[7] http://www.thebiogrid.org

[8] http://www.ecocyc.org

[9] http://www.hprd.org

[10] This number of interactions was available in ENTEZ in September 2007.

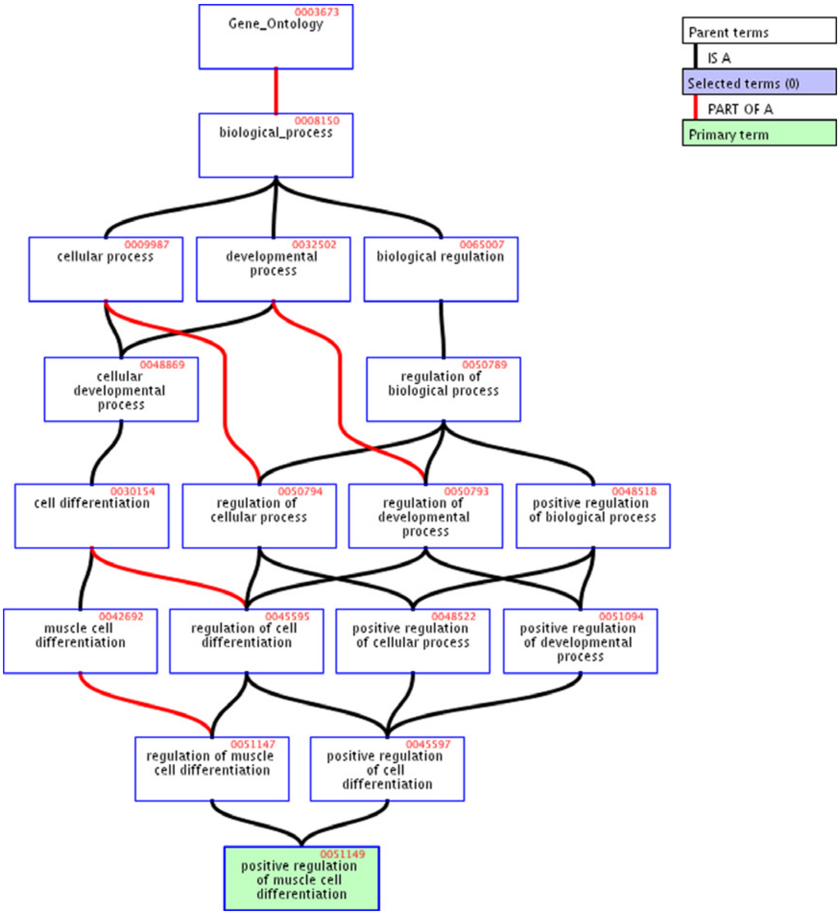

0003673
Gene_Ontology
0008150
biological_process
0009987
cellular process
0032502
developmental process
0065007
biological regulation
0048869
cellular developmental process
0050789
regulation of biological process
0030154
cell differentiation
0050794
regulation of cellular process
0050793
regulation of developmental process
0048518
positive regulation of biological process
0042692
muscle cell differentiation
0045595
regulation of cell differentiation
0048522
positive regulation of cellular process
0051094
positive regulation of developmental process
0051147
regulation of muscle cell differentiation
0045597
positive regulation of cell differentiation
0051149
positive regulation of muscle cell differentiation
Parent terms
IS A
Selected terms (0)
PART OF A
Primary term

Fig. 1. A part of GO providing the annotations concerning positive regulation of muscle cell differentiation.

ciated value indicating a level of confidence in the designation of genes as being differentially expressed.

2.2.2. *Signal to noise ratio test*

The signal to noise ratio (SNR) test identifies genes with large difference in the mean level of expression between two groups and at the same time have small variation of expression within each group. This test does not assume the equality of standard deviations (variances). The SNR combined with different feature selection methods has been the method of choice in most classification studies performed at the Whitehead Institute, MIT [20–22], as well as by several other groups.

2.2.3. *Student's t test*

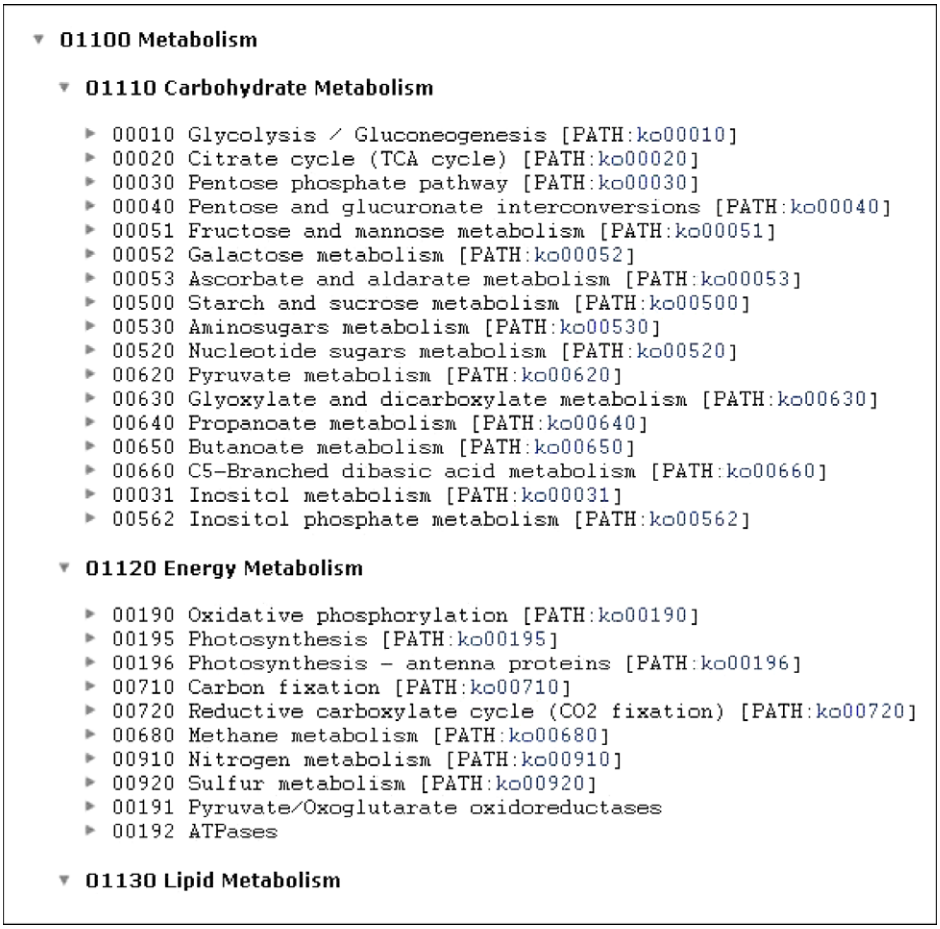
The Student's t test is one of the simplest statistics-based methods used in microarray analysis, both for estimating the accuracy of results from replicated experiments and for the selection of DE genes. The t test for independent samples (Student's t test) allows for the determination of an expression pattern that has a maximal difference in the mean levels of expression between two groups of independent samples with a minimal variation of expression within each group. Therefore, the t test has been used frequently for the selection of DE genes in microarray experiments [17–19]. The difference in gene expression between sample types is expressed as the p value which evaluates the probability that random sampling would result in the observed difference. As the Student's t test determines the significance of the difference between the means of two independent samples, it is a good choice when: (i) the two samples are independently and randomly drawn from the source population(s); (ii) the measurements for both samples have an equal interval; and (iii) the source population(s) can be reasonably assumed to have a normal distribution.

2.3. *Methods for evaluating gene set enrichment*

Here, we present three methods for evaluating gene set enrichment. The first one, Fisher's exact test, is a threshold based procedure. It accept two lists of genes: differentially expressed and all other genes. The next two are from the family of threshold-free procedures. They accept only one list of genes, ranked by some criterion (e.g., the t score value of the genes).

2.3.1. *Fisher's exact test*

When using Fisher's exact test, the score for a gene set annotated by GO term S is the degree of independence between the two properties:



```
▾ 01100 Metabolism

  ▾ 01110 Carbohydrate Metabolism

    ▸ 00010 Glycolysis / Gluconeogenesis [PATH:ko00010]
    ▸ 00020 Citrate cycle (TCA cycle) [PATH:ko00020]
    ▸ 00030 Pentose phosphate pathway [PATH:ko00030]
    ▸ 00040 Pentose and glucuronate interconversions [PATH:ko00040]
    ▸ 00051 Fructose and mannose metabolism [PATH:ko00051]
    ▸ 00052 Galactose metabolism [PATH:ko00052]
    ▸ 00053 Ascorbate and aldarate metabolism [PATH:ko00053]
    ▸ 00500 Starch and sucrose metabolism [PATH:ko00500]
    ▸ 00530 Aminosugars metabolism [PATH:ko00530]
    ▸ 00520 Nucleotide sugars metabolism [PATH:ko00520]
    ▸ 00620 Pyruvate metabolism [PATH:ko00620]
    ▸ 00630 Glyoxylate and dicarboxylate metabolism [PATH:ko00630]
    ▸ 00640 Propanoate metabolism [PATH:ko00640]
    ▸ 00650 Butanoate metabolism [PATH:ko00650]
    ▸ 00660 C5-Branched dibasic acid metabolism [PATH:ko00660]
    ▸ 00031 Inositol metabolism [PATH:ko00031]
    ▸ 00562 Inositol phosphate metabolism [PATH:ko00562]

  ▾ 01120 Energy Metabolism

    ▸ 00190 Oxidative phosphorylation [PATH:ko00190]
    ▸ 00195 Photosynthesis [PATH:ko00195]
    ▸ 00196 Photosynthesis - antenna proteins [PATH:ko00196]
    ▸ 00710 Carbon fixation [PATH:ko00710]
    ▸ 00720 Reductive carboxylate cycle (CO2 fixation) [PATH:ko00720]
    ▸ 00680 Methane metabolism [PATH:ko00680]
    ▸ 00910 Nitrogen metabolism [PATH:ko00910]
    ▸ 00920 Sulfur metabolism [PATH:ko00920]
    ▸ 00191 Pyruvate/Oxoglutarate oxidoreductases
    ▸ 00192 ATPases

  ▾ 01130 Lipid Metabolism
```

Fig. 2. A part of KO providing the annotations concerning metabolism.

A = gene is in the list of DE genes
B = gene is annotated by GO term S

Testing the independence of these two properties corresponds to the Fisher's exact test [6], and is computed by the following procedure:

(1) Let N be the number of genes on a microarray.
(2) S is a GO term.
  (a) M genes $\in S$.
  (b) $N - M$ genes $\notin S$.

(3) Let k be the number of DE genes.
(4) The probability of having exactly x, out of k DE genes, annotated by S is computed as follows:

$$P(X = x \mid N, M, k) = \frac{\binom{M}{x}\binom{N-M}{k-x}}{\binom{N}{k}}$$

(5) The Fisher's score determines the probability of having at least x genes, out of k DE genes, annotated by S:

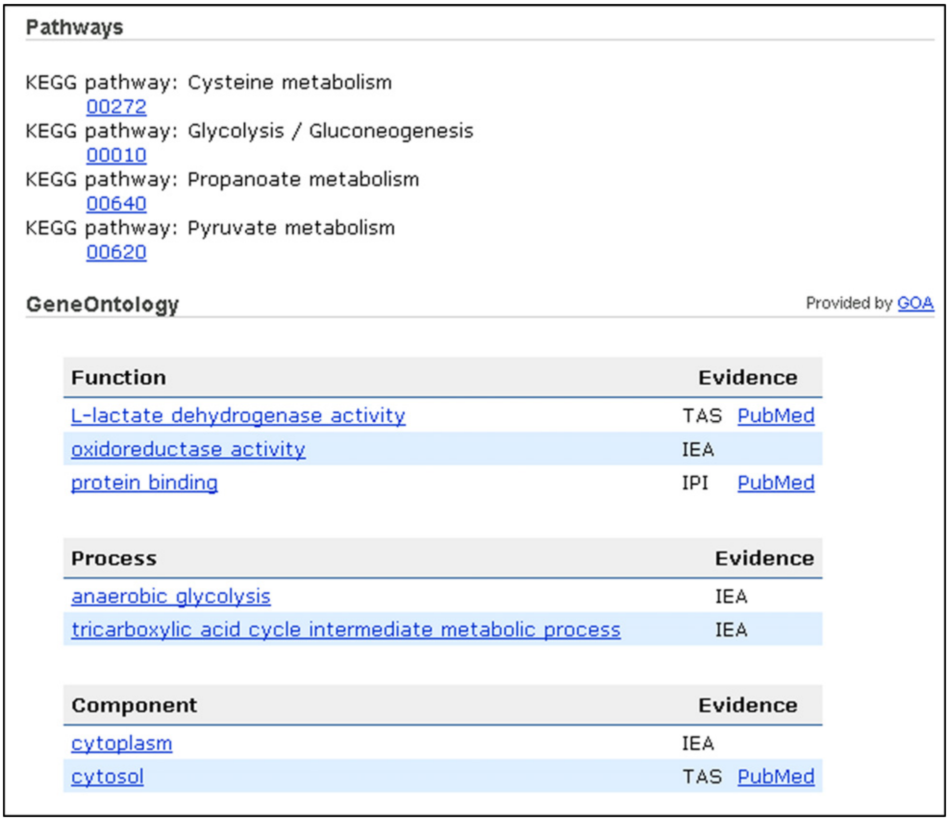
$$p = 1 - \sum_{i=0}^{x-1} \frac{\binom{M}{i}\binom{N-M}{k-i}}{\binom{N}{k}}$$

2.3.2. Gene Set Enrichment Analysis (GSEA)

GSEA [9] considers experiments with gene expression profiles from samples belonging to two classes. First, genes are ranked based on their t-score values. Given a predefined set of genes S (e.g., genes involved in some biological process) the goal of GSEA is to determine whether the members of S are randomly distributed throughout the ranked gene list (L) or primarily found at the top of the list.

There are two major steps of the GSEA method:

(1) Calculation of the enrichment score. The 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 a gene is not in S. The magnitude of the increment depends on the size of S, $|S| = M$, and the total number of genes N. The



Pathways

KEGG pathway: Cysteine metabolism
00272
KEGG pathway: Glycolysis / Gluconeogenesis
00010
KEGG pathway: Propanoate metabolism
00640
KEGG pathway: Pyruvate metabolism
00620

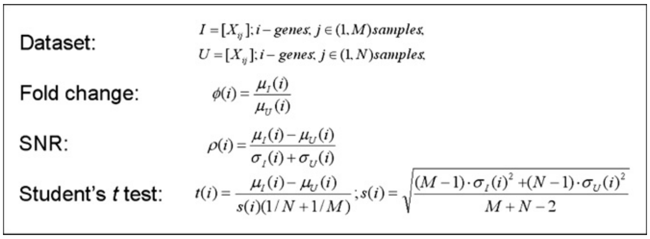
GeneOntology Provided by GOA

Function	Evidence
L-lactate dehydrogenase activity	TAS PubMed
oxidoreductase activity	IEA
protein binding	IPI PubMed

Process	Evidence
anaerobic glycolysis	IEA
tricarboxylic acid cycle intermediate metabolic process	IEA

Component	Evidence
cytoplasm	IEA
cytosol	TAS PubMed

Fig. 3. A part of data providing the annotation of gene LDHA lactate dehydrogenase with KO and GO terms, contained in the ENTREZ database.



Dataset:	$I = [X_{ij}]; i - genes, j \in (1, M) samples,$ $U = [X_{ij}]; i - genes, j \in (1, N) samples,$
Fold change:	$\phi(i) = \frac{\mu_I(i)}{\mu_U(i)}$
SNR:	$\rho(i) = \frac{\mu_I(i) - \mu_U(i)}{\sigma_I(i) + \sigma_U(i)}$
Student's *t* test:	$t(i) = \frac{\mu_I(i) - \mu_U(i)}{s(i)(1/N + 1/M)}; s(i) = \sqrt{\frac{(M-1)\cdot\sigma_I(i)^2 + (N-1)\cdot\sigma_U(i)^2}{M+N-2}}$

Fig. 4. Mathematical definitions of the selected statistical methods used for the selection of DE genes. I and U are two sets of microarray data that define two separate classes, I and U, respectively. X_{ij} is the expression of gene i in sample j, $\mu_C(i)$ is the mean of the expression of gene i in class C, $\sigma_C(i)$ is the standard deviation of the expression of gene i in class C.

enrichment score is the maximum deviation from zero encountered in a random walk (see Fig. 5). If $L = [g_1, g_2, \ldots, g_N]$ is a ranked list of genes, according to their t-score, enrichment score ES is calculated as:

$$\mathrm{ES}(S) = \max_{1\leqslant i\leqslant N} \mid \mathrm{Hit}(S,i) - \mathrm{Miss}(S,i) \mid \qquad (1)$$

where

$$\mathrm{Hit}(S,i) = \sum_{\substack{g_j \in S \\ 1\leqslant j\leqslant i}} \frac{1}{M}, \quad \mathrm{Miss}(S,i) = \sum_{\substack{g_j \notin S \\ 1\leqslant j\leqslant i}} \frac{1}{N-M}$$

(2) Estimation of the 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.

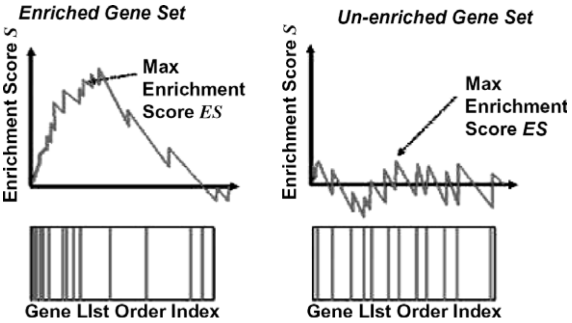


Fig. 5. The 'spectral line's show the positions of genes members of a gene set S on the ranked gene list. This figure is borrowed from the supplementary material of [9].

2.3.3. Parametric Analysis of Gene set Enrichment (PAGE)

According to the Central Limit Theorem in statistics [23], the distribution of the average of randomly sampled n observations tends to follow the normal distribution as the sampling size n becomes larger, even when the parent distribution from which the average is calculated is not normal. In other words, when the mean and variance of the parent distribution (whether it is normally distributed or not) are μ and σ^2, the average of n observations from the parent distribution will follow a normal distribution of mean μ and variance $\frac{\sigma^2}{n}$ when the sampling size n is large enough.

In PAGE [13], the parent distribution is a distribution of any numerical values (also termed parameters here) that describe differential expression of genes among samples in a microarray data set. In most cases, the distribution of a parameter, i.e., t-score values of the genes, is not normally distributed. However, as the Central Limit Theorem states, when we sample n observations from the parent distribution of a parameter, the average of the sampled observations tends to follow the normal distribution as our sampling size n becomes larger. Here, we define sampled observations as parameter values for the genes within predefined gene sets, groups of genes having similar functions, genes in the same biological pathway, and so on. If we define a gene set of sufficiently large size, i.e., 30, we can use the normal distribution to test the statistical significance of that gene set.

The following procedure is used for p value calculation of a gene set S:

(1) From input data containing t-score values for each gene, mean of all t-score values (μ) and standard deviation of all t-score values (σ) are calculated (this is a common step for the calculation of p values of all genes).
(2) The mean of t-scores (μ_S) of gene members of S is calculated.
(3) If M is the size of S then the Z-score is calculated as

$$Z = \frac{(\mu_S - \mu) \cdot \sqrt{M}}{\sigma} \tag{2}$$

Gene set p value is computed from the Z-score, using numerical methods.

3. SEGS: Construction of new gene sets

Methods that test for enrichment of GO terms[11] have been proposed by [4,5,24,25]. A comparative study of commonly used tools for analyzing GO term enrichment was presented by [6]. Papers [14,15] present two novel algorithms that improve GO term scoring using the underlying GO graph topology. None of the papers includes the gene-gene interaction data, and none of them presents a method for the construction of novel gene sets; they only calculate the enrichment of an a-priory given list of gene sets.

We propose a method that additionally to the testing of the enrichment of individual GO and KO terms, tests the enrichment of newly defined gene sets constructed by the combination of GO terms, KO terms and gene sets defined by taking into account the gene–gene interaction data from ENTREZ.

3.1. Properties of GO and KO terms

First, let us state some properties of gene annotations by GO and KO terms:

- one gene can be annotated by several terms,
- if a gene is annotated by a term T then it is annotated by all the ancestors of T, and
- a term may have thousands of genes annotated by it.

From this we can conclude that:

- each GO and KO term defines a gene set,
- one gene can be a member of several gene sets, and
- some gene sets are subsets of other gene sets.

Second, let *Func* (or *Proc*, *Comp*, respectively) denote the set of gene sets that are defined by the GO terms that are subterms of the term Molecular Function (or Biological Process, Cellular Component, respectively), and let *Path* denote the set of gene sets defined by the KO terms.

3.2. Basic operations for gene set construction using GO, KO and ENTREZ

Our method relies on two ideas for the construction of new gene sets: inclusion of gene–gene interactions, and construction by the intersection of gene sets.

3.2.1. Gene–gene interactions

There are cases when some abrupted processes are not detectable by the enrichment score. One of the reasons can be that gene members of that process have a slight increase/decrease in their expression, but this increase/decrease can have a much larger effect on the genes that interact with them. Therefore, we propose to construct a gene set whose members interact with members of another gene set (see Fig. 6). The gene–gene interaction data can be

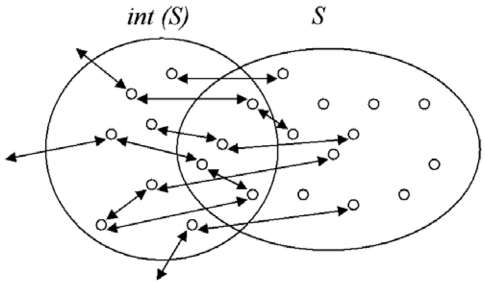


Fig. 6. Construction of a new gene set, *int (S)*, from existing gene set S. All $g_i \in int\ (S)$ are interacting with some $g_j \in S$. Gene sets S and int (S) do not need to intersect.

[11] In the rest of the paper, GO or KO term enrichment is used, meaning the enrichment (i.e. differential expression) of a set of genes, annotated by the given GO or KO term.

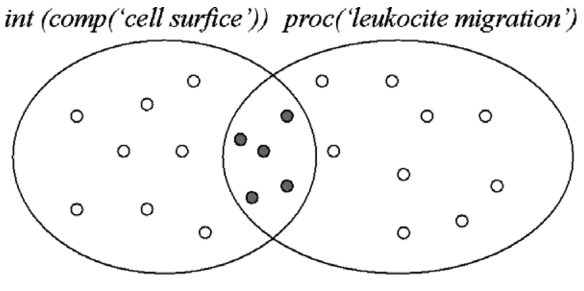


Fig. 7. Construction of a new gene set, consisting of the members of the "leukocyte migration" process which interact with genes on the cell surface.

found in the ENTREZ database. Gene set construction is formally described as follows:

if $S \in Func$ (or $Proc, Comp, Path$, respectively) **then** $int(S) = \{g_j \mid g_j \text{ interacts with } g_i \in S\}$ is added to $Func$ (or $Proc, Comp, Path$).

3.2.2. Intersection of gene sets

There are cases where some gene sets are not significantly enriched, but their intersection is significantly enriched. For example, it can happen that a gene set defined by 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/underexpressed, but if genes with that function in a specific part of the cell (C_{part}) are abnormally active, then this can be elegantly described by defining the following gene set:

$$S = func(F) \bigcap Comp(C_{part}) = S_F \bigcap S_{C_{part}}.$$

Gene set construction due to gene sets intersection is formally described as follows:

if $S_1 \in Func$, $S_2 \in Proc$, $S_3 \in Comp$ and $S_4 \in Path$,**then** $S_{new} = S_1 \bigcap S_2 \bigcap S_3 \bigcap S_4$ is a newly defined gene set.

An example of this type of construction is presented in Fig. 7.

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

$$S = func(A), proc(B) \equiv S_A \bigcap S_B$$

is interpreted as: genes that are part of process B and have function A.

The number of potentially newly defined gene sets is huge. It is currently[12] estimated at:

$$\mid Func \mid \times \mid Proc \mid \times \mid Comp \mid \times \mid Path \mid \approx 47 \times 10^{12}$$

If for each of these sets we compute its enrichment score, which in case of GSEA takes linear time in the number of genes ($\approx 2 \times 10^4$), then we need $\approx 10^{18}$ numeric operations. If we want to statistically validate discovered enriched gene sets, usually with 1000 permutation tests, we get $\approx 10^{21}$ operations, that is well above the average performance of today's PCs. Therefore, we need to efficiently search the gene set space for potentially enriched gene sets, as proposed below.

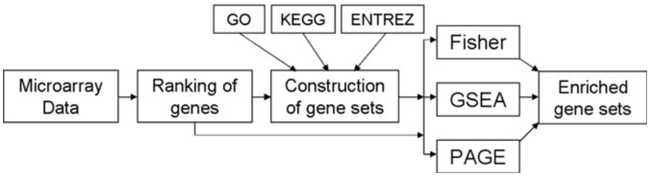


Fig. 8. Data flow of the proposed SEGS method for the generation of enriched gene sets.

3.3. Pruning the search space for 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. This can be achieved by generating gene sets that have some predefined minimum number of genes at the top of the ranked list, i.e. according to the genes t-scores, for example, 3 in the first 100, or 10 in the first 300 genes of the list. That is a weak constraint concerning the biological interpretation of the results, because we are not really interested in gene sets that do not have some minimum 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 and KO topology to efficiently generate all gene sets that satisfy the constraint.

As the GO is a directed acyclic graph, with the root of the graph being the most general term, this 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 potentially enriched gene sets. Therefore, we first construct gene sets from the top nodes of the GO and KO, and if we fail to satisfy the given constraint we do not refine the last added term.

The pseudo code, presented in Appendix A, implements the basic idea for efficient construction of potentially enriched gene sets.

The proposed method has the data flow model shown in Fig. 8.

4. Experiments

Note that this paper does not address the problem 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 the class.

4.1. Brief description of datasets

We applied the proposed SEGS methodology to three classification problems: leukemia [20], diffuse large B-cell

[12] In September 2007, $\mid Func \mid = 7513, \mid Proc \mid = 12,549, \mid Comp \mid = 1846$ and $\mid Path \mid = 272$.

lymphoma (DLBCL) [26] and prostate tumor [27]. All of them are binary classification problems. The leukemia data includes 48 acute lymphoblastic leukemia (ALL) samples and 25 acute myeloid leukemia (AML) samples, each with 7074 gene expression values. The DLBCL data set includes 7070 gene expression profiles for 77 patients, 58 with DLBCL and 19 with follicular lymphoma (FL). The prostate tumor data set includes 12,533 genes measured for 52 prostate tumor and 50 normal tissue samples. The data for these three data sets were produced from Affymetrix gene chips and are available at http://www.genome.wi.mit.edu/cancer/.

4.2. Experimental results

To illustrate the straightforward interpretability of the enriched gene sets found by our approach, we provide the most enriched gene sets for all classes in the three mentioned classification problems (see Tables 1–3). Because we use three statistical tests, which give three different rankings for the enrichment of the gene sets, we calculated the aggregate rank for each gene set by summing its ranks from the separate rankings.

Concerning the number of generated gene sets, for the leukemia data set we generated 210,762 (ALL) and 127,187 (AML) gene sets, for DLBCL data set we generated 158,152 (DLBCL) and 78,048 (FL) gene sets, and for the prostate data set we generated 28,027 (tumor) and 62,567 (normal) gene sets, that satisfied the constraint to have at least three genes in the first 100, or 10 in the first 300 most differentially expressed genes. We also set an additional constraint needed for the PAGE algorithm, the size of the generated gene sets, which was chosen to be larger than 30.

Table 1
Five most enriched gene sets (according to the aggregate ranking) found in the leukemia dataset by using GO, KO and ENTREZ

Gene set Enriched in ALL	Set size	Gene set Enriched in AML	Set size
func(‘DNA binding’), int(comp(‘nucleoplasm’)), int(proc(‘histone modification’))	41	int(comp(‘lysosome’)), int(proc(‘response to ext. stimulus’)), int(path(‘Immune System’))	37
int(func(‘transcrip. repressor activ.’)), comp(‘nucleus’), int(proc(‘histone modification’)), int(path(‘Long-term potentiation’))	50	int(comp(‘membrane part’)), proc(‘inflammatory response’), int(path(‘Human Diseases’))	38
int(func(‘acetyltransferase activity’)), int(comp(‘nucleus’)), int(proc(‘ubiquitin cycle’)), int(path(‘Signal Transduction’))	45	int(func(‘peptidase activity’)), int(comp(‘integral to pl. membrane’)), proc(‘defense response’)	31
int(func(‘nucleotidyltransferase activ.’)), comp(‘nucleus’), int(proc(‘DNA repair’)), int(path(‘Cell cycle’))	84	int(func(‘metal ion binding’)), int(comp(‘integral to membrane’)), proc(‘inflammatory response’)	39
int(func(‘zinc ion binding’)), comp(‘intracellular organelle part’), int(proc(‘protein complex assembly’)), int(path(‘Wnt signaling pathway’))	64	int(func(‘endopept. inhibitor act.’)), int(comp(‘integral to pl. membrane’)), int(proc(‘response to pest.path.par.’)), int(path(‘Cell adhesion molecules’))	43

Table 2
Five most enriched gene sets (according to the aggregate ranking) found in the DLBCL dataset by using GO, KO and ENTREZ

Gene set Enriched in DLBCL	Set size	Gene set Enriched in FL	Set size
int(func(‘transf.phosph.cont.groups’)), int(comp(‘nuclear part’)), proc(‘biopolymer metabolism’)	33	comp(‘integral to membrane’), proc(‘humoral immune response’)	47
int(func(‘transf.phosph.cont.groups’)), comp(‘nucleus’), proc(‘DNA metabolism’), int(path(‘Cell cycle’))	46	comp(‘plasma membrane’), path(‘Hematopoietic cell lineage’)	40
int(func(‘DNA binding’)), int(comp(‘nucleus’)), proc(‘DNA replication’), int(path(‘Cancers’))	35	func(‘transmembrane receptor act.’), int(comp(‘membrane’)), int(proc(‘immune response’)), int(path(‘Immune System’))	83
int(func(‘DNA binding’)), int(comp(‘nucleus’)), proc(‘biopolymer metabolism’), int(path(‘Pancreatic cancer’))	50	func(‘transmembrane receptor act.’), comp(‘integral to membrane’), int(proc(‘immune response’)), int(path(‘Env. Inf. Processing’))	100
int(func(‘transcrip. factor act.’)), int(comp(‘nucleus’)), proc(‘biopolymer metabolism’), int(path(‘Cell Growth and Death’))	64	proc(‘humoral immune response’), int(path(‘Sign. Molec. & Inter.’))	48

Table 3
Five most enriched gene sets (according to the aggregate ranking) found in the prostate dataset by using GO, KO and ENTREZ

Gene set	Set size	Gene set	Set size
Enriched in prostate cancer		Enriched in normal	
func(‘struct. constituent of ribosome’),comp(‘intracellular organelle part’), proc(‘protein biosynthesis’), path(‘Ribosome’)	52	int(func(‘receptor binding’)), comp(‘integral to membrane’) int(proc(‘+ regul. of cell prolif.’)), int(path(‘Human Diseases’))	143
func(‘RNA binding’), comp(‘ribosome’), proc(‘protein biosynthesis’)	45	int(func(‘protein kinase act.’)),int(comp(‘integral to membrane’)), int(proc(‘Ras protein sig. transd.’)), int(path(‘Fc eps. RI sig. path.’))	162
func(‘RNA binding’), comp(‘cytoplasmic part’), path(‘Genetic Information Processing’)	51	int(func(‘protein kinase act.’)), int(comp(‘integral to membrane’)), int(proc(‘Ras protein sig. transd.’)), int(path(‘Focal adhesion’))	172
func(‘struct. constituent of ribosome’), comp(‘cytost. ribosome (s. Eukaryota)’), proc(‘protein biosynthesis’)	62	int(func(‘receptor binding’)), int(comp(‘cytosol’)), int(proc(‘+ regul. of cell prolif.’)), int(path(‘Colorectal cancer’))	178
func(‘RNA binding’), comp(‘intracellular organelle part’)	120	int(func(‘protein kinase activity’)), int(comp(‘integral to membrane’)), int(proc(‘Ras protein sig. transd.’)), int(path(‘Nat.kill.cell.medi.cyt.’))	170

4.3. Statistical validation

The following procedure was used to calculate the significance of the observed enrichment of a gene set by comparing it with the set of maximal enrichment scores computed from the same datasets but with randomly assigned phenotypes (class labels):

(1) Randomly assign the original phenotype (class) labels to samples, reorder genes according to their *t*-score values, and re-compute the enrichment scores.
(2) Repeat step 1 for 1000 permutations, and create a histogram of the corresponding best enrichment scores for all three tests.
(3) Estimate the *p*-value for the calculated enrichment score value of the gene set *S* using the histogram computed at step 2. If there was not a case where random labeling of the examples gives a better enrichment score, 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 the significance than the one obtained by randomly permuting the genes.

After the calculation of the gene sets enrichment, we remove gene sets that have too general descriptions. For example, if gene set S_1 is more enriched then gene set S_2, and S_1 has a more specific description than S_2, then S_2 is eliminated. Note that $S_1 = T_{11} \bigcap T_{12} \bigcap T_{13} \bigcap T_{14}$ is more specific than $S_2 = T_{21} \bigcap T_{22} \bigcap T_{23} \bigcap T_{24}$ if T_{1j} is a subterm of T_{2j} for $j = 1 \ldots 4$.

Table 4 provides the results of the empirical comparison of SEGS with single GO and KO term analysis for the ALL class of the leukemia dataset. Extensive results for all three datasets are given in the supplementary material.[13] We can see that on all tests the best constructed gene sets are found to be more enriched than the most enriched gene sets defined by taking into account only single GO and KO terms.

Concerning the joint coverage of the five most enriched gene sets, for the ALL class of the first problem, we found that their union consists of 179 genes. The sum of the cardinalities of these five sets is 284. This means that we did not find five different descriptions of the same gene set, but these descriptions cover quite different sets of genes. Similar results were obtained for all the classes of the other two datasets.

4.4. Biomedical significance of the discovered enriched gene sets

The goal of this study is to provide a better understanding of the biology of malignancies through the use of the background knowledge encoded in GO, KO and ENTREZ. To do so, we have examined biological functions of genes using the entire pathway changes which are more likely (than the changes in the expression of individual genes) to represent meaningful alterations of cellular metabolism in cancers. In its overall design this study fills in the gap of knowledge represented by the common reductionist approach to the interpretation of microarray data whereby increased or decreased expression of a single gene, rather than behavior of a functionally linked group of genes (a pathway), is used as a readout. In this way, discovered enriched gene sets (described in Tables 1–3) for ALL vs. AML, DLBCL vs. follicular lymphoma, and prostate cancer vs. normal tissue, expand our understanding of predictors of clinical behavior of these cancers. Expert interpretation of several found enriched gene sets for each of the three problems is given below.

4.4.1. ALL vs. AML

Acute leukemias strike 3–4 people per 100,000 every year. Two major classes of acute leukemias exist: acute lymphoblastic leukemia (ALL) and acute myelogenous leukemia (AML). The peak incidence of ALL is in childhood

[13] http://kt.ijs.si/igor-trajkovski/SEGS/supplement.html

Please cite this article in press as: Trajkovski I et al., SEGS: Search for enriched gene sets in microarray data, J Biomed Inform (2008), doi:10.1016/j.jbi.2007.12.001

Table 4
Comparison of the most enriched gene sets constructed using GO, KO and ENTREZ compared to the most enriched gene sets defined by singe GO and KO terms, for the ALL class in the leukemia data set

Gene set	Set size	Fisher p-value (adj p-value)	GSEA ES score (adj p-value)	PAGE Z-score(adj p-value)	Aggregate rank (ranks)
Enriched gene sets in ALL (the same as in Table 1)					
func('DNA binding'), int(comp('nucleoplasm')),int(proc('histone modification'))	41	4.18×10^{-18}(0.001)	0.33(0.001)	8.92(0.001)	5(2 + 2 + 1)
int(func('transcrip. repressor activ.')), comp('nucleus'), int(proc('histone modification')), int(path('Long-term potentiation'))	50	4.96×10^{-19}(0.001)	0.31(0.001)	7.37(0.001)	9(1 + 3 + 5)
int(func('acetyltransferase activity')), int(comp('nucleus')), int(proc('ubiquitin cycle')), int(path('Signal Transduction'))	45	1.38×10^{-17}(0.001)	0.21(0.005)	5.110.015)	16(3 + 6 + 7)
int(func('nucleotidyltransf. activ.')), comp('nucleus'), int(proc('DNA repair')), int(path('Cell cycle'))	84	1.16×10^{-15}(0.004)	0.25 (0.002)	5.90(0.002)	17(6 + 5 + 6)
int(func('zinc ion binding')), comp('intracellular organelle part'), int(proc('protein complex assembly')), int(path('Wnt signaling pathway'))	64	5.70×10^{-16}(0.002)	0.28(0.001)	5.05(0.021)	19(5 + 4 + 10)
Enriched gene sets in ALL (using single GO and KO terms analysis)					
proc('DNA metabolic process')	314	9.14×10^{-7}(0.031)	0.14(0.018)	4.47(0.003)	8(3 + 4 + 1)
comp('nucleus')	1461	3.51×10^{-9}(0.012)	0.13(0.020)	3.29(0.045)	11(1 + 5 + 5)
comp('chromosome')	139	5.28×10^{-7}(0.025)	0.19(0.004)	3.11(0.061)	15(2 + 1 + 12)
path('pyrimidine metabolism')	48	9.21×10^{-6}(0.072)	0.15(0.010)	4.13(0.009)	16(11 + 3 + 2)
func('DNA binding')	810	1.15×10^{-6}(0.048)	0.10(0.071)	3.89(0.011)	18(7 + 8 + 3)
proc('nucleobase, nucleoside, nucleotide & nucleic acid met. proc.')	1321	4.31×10^{-6}(0.050)	0.08(0.125)	3.65(0.022)	23(9 + 10 + 4)
path('nucleotide metabolism')	101	1.02×10^{-6}(0.040)	0.07(0.144)	3.19(0.053)	28(5 + 13 + 10)

(and children account for one quarter of all acute leukemia cases) and it is rare in older adults. In contrast, the median age of AML patients is 60 years and its incidence increases gradually with age. Therefore, as ALL and AML are distinct in clinical presentation, we expected that there would be correlative differences in their biology, as evidenced by microarray expression data.

In fact, the results of our analysis show that functionally linked groups of genes involved in DNA binding (a process whereby transcription factors exert their positive or negative effects on the first phase of protein expression, i.e., transcription of DNA sequence into RNA) and in histone modification (a process whereby transcription machinery is either allowed or prohibited from the access to DNA in the first place) are prominent in ALL cellular pathways, with 41 genes and 50 genes in the first and second ALL gene sets, respectively [31,32].

This is in agreement with the current understanding of the role of transcriptional activators and repressors in ALL, as is the role of ubiquitin (the third ALL gene set with 45 genes) and DNA repair in this condition (the fourth ALL gene set with 84 genes). Ubiquitin cascade is the major cellular mechanism for recycling proteins, thus regulating their activity and permanence (half-life) in the cell. DNA repair is a key regulator of survival of the cell, normal or malignant, as the unrepaired DNA typically precludes cellular division and proliferation. Lastly, the fifth ALL gene set (64 genes) identifies the evolutionarily conserved Wnt-signaling pathway as active in ALL [33]. This is relevant, since Wnt-dependent cellular processes have been shown to be critical for solid organ malignancies, and as therapeutics are already in development for application in solid neoplasms, most notably heaptocellular and colon carcinomas [34,35], it is plausible that they would have a role in chemotherapy for ALL as well.

Terms identified as relevant in AML include those of immune and inflammatory response, cell adhesion and metal ion binding processes. This perhaps gives extra weights to a recently identified, yet not completely understood, property of AML to be more susceptible to eradication by immune means than ALL [36]. In fact, the success of hematopoietic stem cell transplantation for AML maybe in a large part a result of graft vs. leukemia effect, i.e., immune mediated [37].

4.4.2. DLBCL vs. follicular lymphoma

Follicular and diffuse large B-cell lymphomas are two common classes of lymphoma, malignancy that typically involves lymph nodes, spleen, but can originate at other sites, such as gastrointestinal tract, liver, throat, bone, and brain. As expected, immune response pathways (for follicular lymphoma), and DNA binding and replication (key processes in transcriptional regulation of cell division and proliferation in diffuse large B-cell lymphoma) dominate the expression patterns [29,30].

4.4.3. Prostate cancer vs. normal tissue

Prostate cancer is the most common, non-dermatologic male cancer. It represents 33% of cancers and is the third leading cause of cancer deaths in men [28]. Thus, the impact on public health is dramatic and any insights with a potential of translation into viable preventive or therapeutic interventions are urgently needed. In this work, the pathways active in gene transcription (upregulated in any rapidly dividing cells, e.g., malignant cell) have been identified: gene sets 1, 2 and 3 in prostate cancer (with 52, 45 and 51 genes, respectively in Table 3).

In addition, the investigations of normal cells of prostate point, as expected in normal glandular tissue of prostate, discovered groups of genes involved in cell adhesion, Ras oncogene signal transduction, protein regulation (phosphorylation by kinases), including surface membrane receptors (gene sets 1–5 on normal prostate tissue in Table 3).

5. Conclusion and further work

This paper addresses the problem of finding enriched functional groups of genes based on gene expression data. The proposed SEGS method allows integration of GO and KO gene annotations as well as the gene–gene interaction data from ENTREZ into the construction of new interesting relevant gene sets. The experimental results show that the introduced method improves the statistical significance and the functional interpretation of gene expression data, and we base our conclusion on the following facts:

- Enrichment scores of the newly constructed sets are better then the enrichment scores of any single GO and KO term.
- Newly constructed enriched gene sets can be described by non-enriched GO and KO terms, which means that we are extracting additional biological knowledge that can not be found by single term enrichment analysis.
- This method is a generalization of traditional methods. If we turn-off gene–gene interactions and intersections of GO and KO terms, we get the classical single term enrichment analysis.

This paper provides strongly suggesting evidence that the proposed SEGS method indeed finds biologically relevant terms not found by single term analysis (see the examples of terms commented by the medical expert in Section 4.4). The expert interpretation of the results of this study shows that meaningful analysis of gene products acting jointly in biologically relevant ways is possible and that this and future studies can provide support for transferring of this new technology to clinic. An extensive study about the relevance of the found terms (percentage of false positives) is planed in the future. Next, further work will also aims at using discovered enriched gene sets as features for classification of microarray data. We believe that some of these features will turn out to be statistically significant markers of specific diseases.

We believe that the impact of the proposed method will be even greater given the expected increase in both the quality and quantity of gene annotations and gene–gene interaction data in the near future.

Acknowledgments

This research was supported by the Slovenian Ministry of Higher Education, Science and Technology of Slovenia. We are grateful to Filip Železný for previous joint work on the use of GO and gene-gene interaction data in the analysis of microarray data which has largely stimulated this work. We also thank two anonymous reviewers who made numerous useful suggestions that have enabled us to improve the quality of this paper.

Appendix A. SEGS procedures for generating gene sets

The pseudo code presented below is the part of the SEGS algorithm. These procedures are generating all gene sets that contain predefined minimal number of genes (e.g., 3) located at the top (e.g., first 100) of the provided input gene list.

```
topTerm = ['molecular_function', 'biological_process',
    'cellular_component', 'kegg_pathway']

function GENERATE-GENE-SETS(GeneList)
  input: GeneList
  output: gene_sets

  gene_sets= []
  BUILD-CLAUSE(0, [], GeneList[1:100], topTerm[0], gene_sets)
  return gene_sets

procedure BUILD-CLAUSE (depth, clause, genes, term, gene_sets)
  input: depth, clause, gene_set, term
  output: gene_sets

    new_genes=INTERSECTION (genes, TERM_TO_GENES[term])
  IF LENGTH(new_genes)>3 THEN # minimal support ?
    ADD(clause, term)
    ADD(gene_sets, clause)
```

Please cite this article in press as: Trajkovski I et al., SEGS: Search for enriched gene sets in microarray data, J Biomed Inform (2008), doi:10.1016/j.jbi.2007.12.001

```
20 IF depth<4 THEN # add more terms
21   BUILD-CLAUSE(depth+1, clause, new_genes,
22          topTerm[depth+1], gene_sets)
23 REMOVE(clause, term)
24 FOR EACH child IN CHILDREN(term)DO # refine
25   BUILD-CLAUSE(depth, clause, new_genes,
26          child, gene_sets)
```

The main function of the algorithm is the recursive function BUILD-CLAUSE. It tries to add a new term to the given input clause (conjunction of terms). If the new clause cover enough top genes (line 17) then it is added to the resulting list of clauses that describe the new gene sets. After the term is added the procedure recursively call itself in order to add more terms in the clause (line 21) or to refine the added term (line 25). The provided code will generate all gene sets that have at least three genes in the top 100 genes of the GENELIST.

References

[1] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25(1):25–9.

[2] Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 2000;28:27–30.

[3] Mao X, Cai T, Olyarchuk JG, Wei L. Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. Bioinformatics 2005;21(19):3787–93.

[4] Al-Shahrour F, Diaz-Uriarte R, Dopazo J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics 2004;20:578–80.

[5] Beissbarth T, Speed T. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics 2004;1(1):1–2.

[6] Khatri P, Draghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics 2005;21(18):3587–95.

[7] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 1995;57:289–300.

[8] Patti ME, Butte AJ, Crunkhorn S, Cusi K, Berria R, Kashyap S, et al. Coordinated reduction of genes of oxidative metabolism in humans with insulin resistance and diabetes: potential role of PGC1 and NRF1. Proc Natl Acad Sci USA 2003;100(14): 8466–71.

[9] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledgebased approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 2005;102(43):15545–50.

[10] Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP., 2007. GSEA-P: a desktop application for Gene Set Enrichment Analysis. Bioinformatics July 20:(Epub ahead of print)..

[11] Fortunel NO, Otu HH, Ng HH, Chen J, Mu X, Chevassut T, et al. Comment on ‘Stemness’: “transcriptional profiling of embryonic and adult stem cells” and “a stem cell molecular signature”. Science 2003;302(5644). author reply 393.

[12] Lewin A, Grieve IC. Grouping Gene Ontology terms to improve the assessment of gene set enrichment in microarray data. BMC Bioinformatics 2006;3(7):426.

[13] Kim SY, Volsky D. PAGE: parametric analysis of gene set enrichment. BMC Bioinformatics 2005;6:144 .

[14] Grossmann S, Bauer S, Robinson PN, Vingron M. An improved statistic for detecting over-represented gene ontology annotations in gene sets. In: RECOMB 2006. Berlin: Springer; 2006. p. 85–98.

[15] Alexa A, Rahnenfuhrer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 2006;22(13):1600–7.

[16] Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S, et al. Within the fold: assessing differential expression measures and reproducibility in microarray assays. Genome Biol 2002;3.

[17] Ma X, Salunga R, Tuggle JT, Gaudet J, Enright E, McQuary P, et al. Gene expression profiles of human breast cancer progession. Proc Natl Acad Sci USA 2003;100:5974–9.

[18] Bueno R, Loughlin KR, Powell MH, Gordon GJ. A diagnostic test for prostate cancer from gene expression profiling data. J Urol 2004;171:903–6.

[19] Amatschek S, Koenig U, Auer H, Steinlein P, Pacher M, Gruenfelder A, et al. Tissue-wide expression profiling using cDNA substraction and microarrays to identify tumor-specific genes. Cancer Res 2004;64:844–56.

[20] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 1999;286(5439):531–7.

[21] Savage KJ, Monti S, Kutok JL, Cattoretti G, Neuberg D, De Leval L, et al. The molecular signature of mediastinal Large B-Cell lymphoma differs from that of other Diffuse Large B-Cell lymphomas and shares features with classical Hodkin lymphoma. Blood 2003;102:3871–9.

[22] Yeang CH, Ramaswamy S, Tamayo P, Mukherjee S, Rifkin RM, Angelo M, et al. Molecular classification of multiple tumor types. Bioinformatics 2001;17(Suppl. 1):316–22.

[23] Rosner B. Fundamentals of biostatistics. Duxbury Press; 2000.

[24] Draghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA. Global functional profiling of gene expression. Genomics 2003;81: 98–104.

[25] Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol 2003;4(4):R28.

[26] Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, et al. Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med 2002;8:68–74.

[27] Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, et al. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 2002;1:203–9.

[28] Walczak JR, Carducci MA. Prostate cancer: a practical approach to current management of recurrent disease. Mayo Clin Proc 2007;82(2):243–9.

[29] Bende RJ, Smit LA, van Noese CJM. Molecular pathways in follicular lymphoma. Leukemia 2007;21(1):18–29.

[30] De Paepe P, De Wolf-Peeters C. Diffuse large B-cell lymphoma: a heterogeneous group of non-Hodgkin lymphomas comprising several distinct clinicopathological entities. Leukemia 2007;21(1):37–43.

[31] Crazzolara R, Bernhard D. CXCR4 chemokine receptors, histone deacetylase inhibitors and acute lymphoblastic leukemia. Leuk Lymphoma 2005;46(11):1545–51.

[32] Einsiedel HG, Kawan L, Eckert C, Witt O, Fichtner I, Henze G, et al. Histone deacetylase inhibitors have antitumor activity in two NOD/SCID mouse models of B-cell precursor childhood acute lymphoblastic leukemia. Leukemia 2006;20(8):1435–6.

[33] Weerkamp F, van Dongen JJ, Staal FJ. Notch and Wnt signaling in T-lymphocyte development and acute lymphoblastic leukemia. Leukemia 2006;20(7):1197–205.

[34] Gregorieff A, Clevers H. Wnt signaling in the intestinal epithelium: from endoderm to cancer. Genes Dev 2005;19(8):877–90.

[35] Lee HC, Kim M, Wands JR. Wnt/Frizzled signaling in hepatocellular carcinoma. Front Biosci 2006;11:1901–15.

[36] Baron F, Storb R. The immune system as a foundation for immunologic therapy and hematologic malignancies: a historical perspective. Best Pract Res Clin Haematol 2006;19(4): 637–53.

[37] Ruggeri L, Mancusi A, Burchielli E, Aversa F, Martelli MF, Velardi A, et al. Natural killer cell alloreactivity in allogeneic hematopoietic transplantation. Curr Opin Oncol 2007;19(2):142–7.

Please cite this article in press as: Trajkovski I et al., SEGS: Search for enriched gene sets in microarray data, J Biomed Inform (2008), doi:10.1016/j.jbi.2007.12.001