Available online at www.sciencedirect.com
ScienceDirect
Journal of Biomedical Informatics 41 (2008) 588–601
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

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
Available online 15 December 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