BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 22 2007, pages 3098–3099
doi:10.1093/bioinformatics/btm445

Gene expression

Functional profiling of microarray experiments using text-mining derived bioentities

Pablo Minguez[1], Fátima Al-Shahrour[1], David Montaner[1,2] and Joaquín Dopazo[1,2,*]

[1]Department of Bioinformatics and [2]Functional Genomics Node, (INB), Centro de Investigación Príncipe Felipe (CIPF), Valencia, E46013, Spain

Received on January 17, 2007; revised on July 11, 2007; accepted on August 21, 2007

Advance Access publication September 13, 2007

Associate Editor: John Quackenbush

ABSTRACT

Motivation: The increasing use of microarray technologies brought about a parallel demand in methods for the functional interpretation of the results. Beyond the conventional functional annotations for genes, such as gene ontology, pathways, etc. other sources of information are still to be exploited. Text-mining methods allow extracting informative terms (bioentities) with different functional, chemical, clinical, etc. meanings, that can be associated to genes. We show how to use these associations within an appropriate statistical framework and how to apply them through easy-to-use, web-based environments to the functional interpretation of microarray experiments. Functional enrichment and gene set enrichment tests using bioentities are presented.

Availability: Marmite and MarmiteScan can be found in the Babelomics suite: http://www.babelomics.org

Contact: jdopazo@cipf.es

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

1 BACKGROUND

A major challenge in microarray and, in general, in any genome-scale experiment is to provide a functional explanation that links the results found at molecular level to the macroscopic observation or to the hypothesis that generated the experiment. This is commonly achieved by means of 'functional enrichment' analysis that first selects genes of interest based on the experimental values (e.g. genes differentially expressed between patients and healthy controls) and then studies the enrichment in functional terms (e.g. gene ontology —GO— annotations) in them (Al-Shahrour *et al.*, 2004; Khatri and Draghici, 2005). Conceptually newer approaches avoid the first step of gene selection, where much information is lost because the functional interactions between genes are ignored (Dopazo, 2006), and directly focus on functionally related blocks of genes. Thus, functional profiling methods such as GSEA (Mootha *et al.*, 2003) or FatiScan (Al-Shahrour *et al.*, 2005) report blocks of genes belonging to different functional categories (GO, KEGG pathways, etc.) displaying a cooperative significant over- or under-expression when comparing two classes of microarray experiments. Genes can be grouped in many different ways that contain some biological or functional significance by using different repositories or information sources. To this end information coming from GO, KEGG pathways, Swissprot keywords, chromosomal position, Interpro functional motifs, transcription factor binding sites, etc. has been used for the functional profiling of microarray experiments (Al-Shahrour *et al.*, 2006).

Text-mining methods (Krallinger and Valencia, 2005) offer the possibility of extracting different functional aspects of the genes beyond the ones covered by the 'traditional' repositories (GO, KEGG, etc.) that can be further used for functional profiling purposes. We present two tools that use functional terms (essentially chemical and clinical terms) obtained using text-mining techniques which can be used within a statistical framework that covers both types of tests previously mentioned: tests of functional enrichment in pre-selected sets of genes (Marmite tool) or tests for blocks of functionally related genes (MarmiteScan tool).

2 DEFINITION OF BIOENTITIES

In the approach described here two types of terms (bioentities) have been used: those referred to chemical products and those related to diseases. The terms were extracted from PubMed abstracts, and are related to human genes by a score derived from the frequency of gene-term co-occurrences and depending on their proximity within the text. The scores are derived from a Z-statistic that estimates how unlikely it is to observe a certain level of co-occurrences to happen by chance (Andrade and Valencia, 1998). The gene-bioentity correspondence tables with the respective scores were obtained using the AKS software (available at: http://www.bioalma.com/aks2/) and are freely available in the GeneCards (Safran *et al.*, 2003) database (http://www.genecards.org/). Contrarily to the case of GO and other similar functional categories, bioentities are not discrete classes. The membership of a gene to a given bioentity is conditioned through the scores.

3 TESTS FOR FUNCTIONAL ENRICHMENT OF BIOENTITIES

Given a set of genes selected by some experimental measurement (e.g. because they are differentially expressed between two types of experiments), Marmite checks for significant enrichments in

*To whom correspondence should be addressed.

bioentity annotations in this set with respect to the background. The functional enrichment test carried out by Marmite is in many ways conceptually similar to the tests used for classical repositories such as GO, KEGG, etc. (Al-Shahrour *et al*., 2004; Khatri and Draghici, 2005). The difference in this case is that the functional category to be tested, the bioentity, is considered to be a continuous class. Membership of a gene to a bioentity is therefore defined by a score value, which would reflect the strength of the real relationship gene-bioentity. Therefore, instead of the usual Fisher's or hypergeometric (or similar) tests, we use a Kolmogorov–Smirnov test to compare the distributions of the scores of the co-occurrences between genes and bioentities for each bioentity studied to the background distribution of scores. Since all the bioentities are tested, the *P*-values assigned to them are adjusted by False Discovery Rate (Benjamini and Hochberg, 1995).

4 GENE SET ENRICHMENT ANALYSIS USING BIOENTITIES

Likewise, the way of studying the behavior of blocks of genes defined by bioentities is carried out by means of a segmentation test similar to the one used in FatiScan (Al-Shahrour *et al*., 2005). A pre-selection of genes is not necessary, only a ranked list is used in this test. Thus, given a list of genes arranged by any biological characteristic of the experiment (e.g. by differential expression between two types of experiments), a segmentation test is used to detect significant asymmetrical distributions of bioentities across it. Again, given the continuous nature of the bioentities, a Kolmogorov–Smirnov test is used to detect blocks of genes constitutively skewed to the extremes of the ranking and, consequently related to the biological criteria used for producing the ranking. This test is implemented in the MarmiteScan program and can be used in combination with the t-rex tool from the GEPAS (Montaner *et al*., 2006), which produces the ranked lists of genes for distinct microarray experimental designs.

5 A CASE STUDY OF AML

A recent study (Stegmaier *et al*., 2004) described a high-throughput screening methodology to test whether the action of a number of compounds in the transcriptome of cells with acute myeloid leukemia (AML) reproduce the gene signature characteristic of AML differentiation to normal cells. Additional material show different chemical products significantly associated to high expression values. These results should be understood as co-activations of blocks of genes, which have been related to chemical products throughout the biomedical literature, when two experimental conditions are compared (treated AML cells versus different controls). The nature of the chemical products found provide a new perspective on the biochemical processes acting in AML cells with the different treatments received (see Supplementary Material for an explanation).

6 CONCLUSIONS

In the last years, several proposals that make use of text-mining methods in the context of microarrays have been made such as GEISHA (Oliveros *et al*., 2000), MedMiner (Tanabe *et al*., 1999), ConceptMaker (Kuffner *et al*., 2005) or others (Krallinger and Valencia, 2005). Nevertheless, although such programs provide biological terms related to the query gene(s), they do not implement a robust statistical framework to assess the significance of the results found beyond simple measurements of enrichment. And especially, there is nothing like the functional profiling method presented in MarmiteScan that, similarly to FatiScan (Al-Shahrour *et al*., 2005), directly tests the behavior of blocks of functionally related genes, and does not require of a previous step of gene selection. It is worth mentioning that the segmentation test implemented in this tool does not depend on the original data for obtaining *P*-values, but only on the gene ranking. As a result, many different experimental designs (two-class comparisons, survival, correlation to any parameter, etc.) can be tested providing these produce a gene ranking.

To our knowledge, Marmite and MarmiteScan are the only applications in which functional profiling, based on text-mining, is performed in user-friendly environment within the proper statistical framework.

ACKNOWLEDGEMENTS

This work is supported by grants from project BIO 2005-01078 from the MEC, NRC Canada-SEPOCT Spain, INDIGO EU project and National Institute of Bioinformatics (www.inab.org), a platform of Genoma España.

Conflict of Interest: none declared.

REFERENCES

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

Al-Shahrour,F. *et al*. (2005) Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. *Bioinformatics*, **21**, 2988–2993.

Al-Shahrour,F. *et al*. (2006) BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Res*., **34**, W472–W476.

Andrade,M.A. and Valencia,A. (1998) Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics*, **14**, 600–607.

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

Dopazo,J. (2006) Functional interpretation of microarray experiments. *Omics*, **10**, 398–410.

Khatri,P. and Draghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.

Krallinger,M. and Valencia,A. (2005) Text-mining and information-retrieval services for molecular biology. *Genome Biol*., **6**, 224.

Kuffner,R. *et al*. (2005) Expert knowledge without the expert: integrated analysis of gene expression and literature to derive active functional contexts. *Bioinformatics*, **21** (**Suppl. 2**), ii259–ii267.

Montaner,D. *et al*. (2006) Next station in microarray data analysis: GEPAS. *Nucleic Acids Res*., **34**, W486–W491.

Mootha,V.K. *et al*. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet*., **34**, 267–273.

Oliveros,J.C. *et al*. (2000) Expression profiles and biological function. *Genome Inform. Ser. Workshop Genome Inform*., **11**, 106–117.

Safran,M. *et al*. (2003) Human gene-centric databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Res*., **31**, 142–146.

Stegmaier,K. *et al*. (2004) Gene expression-based high-throughput screening (GE-HTS) and application to leukemia differentiation. *Nat. Genet*., **36**, 257–263.

Tanabe,L. *et al*. (1999) MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques*, **27**, 1210–1214, 1216–1217.