Computational Statistics and Data Analysis 71 (2014) 588–598

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Incorporation of gene exchangeabilities improves the reproducibility of gene set rankings

Charlotte Soneson *, Magnus Fontes

Centre for Mathematical Sciences, Lund University, Sweden

ARTICLE INFO

Article history:
Received 30 May 2011
Received in revised form 26 July 2012
Accepted 27 July 2012
Available online 13 August 2012

Keywords:
Bioinformatics
Exchangeability
Gene set ranking
Microarray

ABSTRACT

Gene set-based analysis methods have recently gained increasing popularity for analysis of microarray data. Several studies have indicated that the results from such methods are more reproducible and more easily interpretable than the results from single gene-based methods. A new method for ranking gene sets with respect to their association with a given predictor or response, using a new framework for robust gene list representation, is proposed. Employing the concept of exchangeability of random variables, this method attempts to account for the functional redundancy among the genes. Compared to other evaluated methods for gene set ranking, the proposed method yields rankings that are more robust with respect to sampling variations in the underlying data, which allows more reliable biological conclusions.

1. Introduction

The development of high-throughput genetic measurement techniques, for example the microarray, has made it possible to monitor genome-wide expression patterns and genetic aberrations in a routine fashion. The results of high-level analysis of such data sets are often represented by *gene lists* containing the genes that are associated with a given clinical or experimental factor, possibly ranked according to their level of association. The obtained lists must then be carefully interpreted to generate biologically valid hypotheses. However, different studies testing the same hypothesis generally return lists with a very small overlap (Fortunel et al., 2003; Miklos and Maleszka, 2004; Irizarry et al., 2005; Michiels et al., 2005; Ein-Dor et al., 2006; Fan et al., 2006). Moreover, even if a good set of predictive genes for a certain condition has been found it is often possible to find other, equally powerful gene collections (Ein-Dor et al., 2005; Reyal et al., 2008). Part of this instability is likely due to a certain level of redundancy, for example among genes belonging to the same pathway and encoding similar biological functions. Other factors contributing to the instability are the presence of noise in the measurements and the often relatively small sample sizes (He and Yu, 2010).

It has been noted that by studying the differential expression of *gene sets*, that is, collections of genes (often with similar functions), instead of individual genes, the resulting conclusions are more reproducible and more subtle effects can be detected (Hosack et al., 2003; Subramanian et al., 2005; Abraham et al., 2010). The gene sets can be obtained, for example, from large publicly available annotation databases such as the Gene Ontology (GO) (Ashburner et al., 2000), the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Ogata et al., 1999) or the Molecular Signatures Database (MSigDB) (Subramanian et al., 2005). In this paper, we will study the robustness of different methods for ranking such gene sets by their association with a response, and propose a new method which provides an advantageous ranking stability with respect to sampling variations in the underlying data.

* Correspondence to: Swiss Institute of Bioinformatics, Bâtiment Génopode, Quartier Sorge, University of Lausanne, CH-1015 Lausanne, Switzerland. Tel.: +41 21 692 40 91.
E-mail addresses: Charlotte.Soneson@isb-sib.ch (C. Soneson), fontes@maths.lth.se (M. Fontes).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.07.026

Soneson and Fontes (2012) proposed a general framework for representation of gene lists, and suggested to use the concept of exchangeability of random variables to quantify the functional redundancy among the genes with respect to a specific experiment. In that paper, the proposed framework was used to obtain *gene* rankings that were more robust than the original ranking of the genes with respect to sampling variations, without compromising the biological relevance of the top-ranked genes. In the present paper, we use the proposed framework to study a different problem, namely that of obtaining robust rankings of gene *sets* with respect to their association with a given response. We explore two different ways of using the stabilized gene ranking to compute a gene set ranking, and the results suggest that the stabilized gene rankings provide improved reproducibility also of gene set rankings compared to existing methods, without compromising the biological relevance of the top-ranked gene sets. First, we describe a method, called Exchangeability-Stabilized Gene set Ranking (ESGR), to rank gene sets according to their association with a response. Second, we show that the stabilized gene ranking, or the modified gene ranking scores, obtained by the method described in Soneson and Fontes (2012), can be used as input to any gene set analysis method defined only in terms of gene rankings or gene ranking scores, to potentially give a more robust version of the original method.

2. Related work

A large number of methods for gene set analysis exist in the literature, and a general modular framework is provided by Ackermann and Strimmer (2009). Reviews and discussions are also given by, for example, Goeman and Bühlmann (2007) and Song and Black (2008). Perhaps the most straightforward methods to test the differential expression of a gene set are the overrepresentation analysis methods, where the size of the overlap between a gene set and an unordered collection of differentially expressed genes is examined for significance using tests based on the hypergeometric distribution, or approximations based on the binomial or χ^2 distributions (Drăghici et al., 2003; Hosack et al., 2003; Khatri and Drăghici, 2005). Overrepresentation is also commonly quantified by the percentage of overlapping genes (POG) between two lists (Ein-Dor et al., 2006; MAQC Consortium, 2006). These methods have recently been extended to take into account correlated molecular changes or known functional relationships (Zhang et al., 2009; Gong et al., 2010).

The gene set enrichment methods, where the genes from a gene set are investigated for enrichment in one of the extreme ends of a ranking of all genes from an experiment, are used for example by Mootha et al. (2003) and Subramanian et al. (2005). Other methods combine single-gene statistics to assess the differential expression of a gene set (Tian et al., 2005; Dinu et al., 2007; Efron and Tibshirani, 2007; Tintle et al., 2008). Finally, the multivariate and global test methods directly test the genes in the gene set for significant association to the response, without using individual gene statistics (Goeman et al., 2004; Mansmann and Meister, 2005; Kong et al., 2006; Tsai and Chen, 2009; Shen et al., 2011).

Note that while most methods indicated above are mainly developed to test whether or not a single gene set is significantly associated with a given response, we focus on the *ranking* of the gene sets according to their level of association with the response. The question of gene set ranking stability has previously been studied by Abraham et al. (2010), who ranked gene sets based on their ability to discriminate between different sample groups, and studied the stability of the resulting ranking.

3. Methods

3.1. Data

Here, we describe the two publicly available gene expression data sets and the collection of gene sets that will be used to evaluate the performance of the studied gene set ranking methods.

3.1.1. Boston lung cancer data

This microarray data set (Bhattacharjee et al., 2001) contains gene expression profiles from 62 lung cancer patients. The data set was downloaded from http://www.broadinstitute.org/gsea/datasets.jsp. The patients are stratified into two groups; those with poor outcome and those with good outcome. Each group contains 31 patients. In the downloaded data file, the original Affymetrix probe IDs have already been replaced by the corresponding gene symbols and probes mapping to the same gene have been summarized by the largest value for each sample (Subramanian et al., 2005). The final data set contains 5217 variables. We centered and scaled each variable to zero mean and unit variance before the analysis.

3.1.2. Hedenfalk breast cancer data

This data set was described by Hedenfalk et al. (2001), and contains gene expression measurements for 15 breast cancer patients, with either a BRCA1 mutation ($N = 7$) or a BRCA2 mutation ($N = 8$). We downloaded the data set from http://research.nhgri.nih.gov/microarray/NEJM_Supplement/. The intensity ratios in the downloaded data set were log-transformed before the analysis. The original data set contains 3226 clones, which were selected by Hedenfalk et al. (2001). We further removed all clones without a corresponding gene symbol, as well as all clones which mapped to several UniGene cluster IDs. Finally, we collapsed clones corresponding to the same gene by replacing them with the largest value for each sample, as was done for the Boston lung cancer data. The final data set contains 2224 variables. Before the analysis, we centered and scaled each variable to zero mean and unit variance.