Gene 518 (2013) 179–186

Contents lists available at SciVerse ScienceDirect

Gene

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

Random forests-based differential analysis of gene sets for gene expression data

Huey-Miin Hsueh [a], Da-Wei Zhou [a], Chen-An Tsai [b,*]

[a] Department of Statistics, National Chengchi University, Taiwan
[b] Department of Agronomy, National Taiwan University, Taiwan

ARTICLE INFO

Article history:
Accepted 27 November 2012
Available online 6 December 2012

Keywords:
DNA microarray
Gene-set analysis
Random Forests

ABSTRACT

In DNA microarray studies, gene-set analysis (GSA) has become the focus of gene expression data analysis. GSA utilizes the gene expression profiles of functionally related gene sets in Gene Ontology (GO) categories or priori-defined biological classes to assess the significance of gene sets associated with clinical outcomes or phenotypes. Many statistical approaches have been proposed to determine whether such functionally related gene sets express differentially (enrichment and/or deletion) in variations of phenotypes. However, little attention has been given to the discriminatory power of gene sets and classification of patients.

In this study, we propose a method of gene set analysis, in which gene sets are used to develop classifications of patients based on the Random Forest (RF) algorithm. The corresponding empirical *p*-value of an observed out-of-bag (OOB) error rate of the classifier is introduced to identify differentially expressed gene sets using an adequate resampling method. In addition, we discuss the impacts and correlations of genes within each gene set based on the measures of variable importance in the RF algorithm. Significant classifications are reported and visualized together with the underlying gene sets and their contribution to the phenotypes of interest.

Numerical studies using both synthesized data and a series of publicly available gene expression data sets are conducted to evaluate the performance of the proposed methods. Compared with other hypothesis testing approaches, our proposed methods are reliable and successful in identifying enriched gene sets and in discovering the contributions of genes within a gene set. The classification results of identified gene sets can provide an valuable alternative to gene set testing to reveal the unknown, biologically relevant classes of samples or patients.

In summary, our proposed method allows one to simultaneously assess the discriminatory ability of gene sets and the importance of genes for interpretation of data in complex biological systems. The classifications of biologically defined gene sets can reveal the underlying interactions of gene sets associated with the phenotypes, and provide an insightful complement to conventional gene set analyses.

1. Introduction

Biological phenomena often occur through the interactions of multiple genes via signaling pathways, genetic networks, or other functional relationships. In DNA microarray studies, single gene analyses can only take into account a small portion of genetic variation in the complex biological system. In contrast to single gene analyses, a gene-set analysis (GSA) is used to evaluate the association between the expression of biological pathways, or a priori defined gene sets, and a particular phenotype. Genes that serve a common molecular function, a biological process, or a cellular component are annotated to the same term and grouped together into sets. The annotation terms can be obtained from public-domain web-libraries such as Gene Ontology (GO), KEGG, BioCarta and the Broad Institute. See Pang et al. (2006) and Delongchamp et al. (2006). This helps biologists to interpret the selected sets of genes in a manner of gene regulation mechanism from the microarray data.

Many statistical methods are proposed for gene-set data analysis in literatures. Mootha et al. (2003) and Subramanian et al. (2005), first proposed the Gene Set Enrichment Analysis (GSEA), in which they consider the distributions of entire genes in a gene set, rather than a subset from the list of differential expression genes, and then use some statistic to assess the significance of predefined gene sets. These existing approaches are not only distinct in terms of the test statistic used, but also differ in terms of the null hypothesis and hence differ in the problem of research interest. Tian et al. (2005) and Goeman and Buhlmann (2007) summarize the methods into two types: competitive and self-contained tests. The null hypothesis of a competitive test is that the specific gene set is not differentially expressed when compared to other gene sets. This method involves not only the gene set of research interest but also the full data set. The sampling unit in construction of the empirical null distribution for calculating a *p*-value is the gene. A

Abbreviations: GSA, gene-set analysis; GO, Gene Ontology; RF, Random Forest algorithm; OOB, out-of-bag error rate.

* Corresponding author. Tel.: +886 2 3366 4775; fax: +886 2 2362 0879.
E-mail addresses: hsueh@nccu.edu.tw (H.-M. Hsueh), 98354014@nccu.edu.tw (D.-W. Zhou), catsai@ntu.edu.tw (C.-A. Tsai).

0378-1119/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2012.11.034

positive finding can be obtained only when the gene set contributes to the statistical variation of a phenotype of interest more than other gene sets. On the other hand, the self-contained test is utilized to determine whether the gene set is differentially expressed. The analysis is conducted with respect to the specific gene set data alone. This approach evaluates *p*-values by permuting samples using the conventional approach. Following the idea of GSEA, many statistical methods have been proposed, such as the global test (Goeman et al., 2004), approaches similar to the two-sample t-test (Tian et al., 2005), the ANCOVA test (Mansmann and Meister, 2005), the Hotelling's T2 test (Kong et al., 2006), the MaxMean approach (Efron and Tibshirani, 2007), the SAM-GS test (Dinu et al., 2007), the global statistics approach (Chen et al., 2007), the Random-sets method (Newton et al., 2007), the Logistic Regression (LRpath) approach (Sartor et al., 2009), and the MANOVA test (Tsai and Chen, 2009) amongst others. These approaches rely on different statistical assumptions and consider different data structures, which then usually leads to different findings even when they are applied to the same data set. A comprehensive review of these methodologies can be found in, for example, Goeman and Buhlmann (2007) and Nam and Kim (2008). Fridley et al. (2010) also provided intensive empirical comparisons on the self-contained analysis. As referenced above, none of their methods have addressed the feasibility of a pre-defined gene set in discriminating different phenotypes.

On the other hand, various machine learning-type algorithms, which take various biological information into consideration, have been developed from a classification perspective. For example, Lin et al. (2006) demonstrated that accuracy and robustness of a classification in analyzing microarray data can be improved by considering the existing biological annotations. Wei and Li (2007) applied a boosting-based method for a nonparametric pathways-based regression (NPR) analysis. Although the NPR generates an improved prediction, there is not a selection criterion to identify differential gene sets. Tai and Pan (2007a, 2007b) proposed a group penalization method that incorporates biological information to build a penalized classifier. Lottaz and Spang (2005) provided a biologically focused classifier, such as StAM, based on the GO hierarchical structure. This method has a limitation that only the genes annotated in the leaf nodes of the GO tree can be used as the predictors, while other genes (relevant, but not annotated yet) cannot be used. However the biological information of gene sets may come from different databases, such as KEGG or BioCarta, and are not limited to the GO annotation only.

Recently, the Random Forest algorithm, developed by Breiman (2001), has gained popularity for use in microarray data analysis due to its flexibility in terms of the type and the dimension of the input data, the absence of overfitting, and a predictive performance comparable to other machine learning methods. See Huang et al. (2005), Díaz-Uriarte and Alvarez de Andrés (2006), Statnikov et al. (2008), and Boulesteix et al. (2008). Pang et al. (2006) and Pang and Zhao (2008) used the Random Forest classification and regression based on pathway information. They proposed a rank analysis of pathways in terms of the predictive performance of the Random Forest built on the pathway. However, no biological variation is taken into account, and hence no confirmatory conclusion can be made based on the evidence. Here, the Random Forest algorithm will be employed to link a gene set and the phenotypic response. The correspondent predictive performance will be used to reveal the strength of the association between the gene set and the phenotype. In addition, the resultant statistical evidence, considering the biological variation, will be obtained.

In this paper, we propose a self-contained GSA method that can not only identify differential gene sets, which are significantly associated with the variation of phenotypes, but can also assess the impacts of individual genes on a prediction model. The Random Forests algorithm will be applied to develop a classifier based on the gene set. The empirical *p*-value of the performance of the classifier will be obtained by using the permutation test to evaluate the statistical significance of the gene set. In addition, during the analysis, we integrate the classification results from the identified gene sets to uncover potential associations between gene sets and phenotypes. Our proposed approach is compared with some existing GSA approaches based on the performance of synthesized data sets and a series of publicly available microarray data.

2. Materials and methods

Consider a microarray study of size n and one k-class phenotype. Assume the gene set or pathway S including m genes is of interest. In contrast to the competitive test, where a relative conclusion is made upon a comparison with the whole gene set, the self-contained test, which seeks an absolute association between the gene set and the phenotype, is studied here. The null hypothesis of a self-contained test of the gene set S is given as

$\mathbf{H_0^S}$. The gene set S is independent with the phenotype variable.

To collect more information on multiple genes in a gene set, a complex classifier is constructed and its test set error rate is recorded. The lower the error rate, the more the evidence shows that the gene set is associated with the phenotypes. Hence the test set error rate is utilized as a test statistic of the self-contained test, and the correspondent *p*-value is applied to draw a statistical conclusion.

We consider using the Random Forests classifier (Breiman, 2001). The Random Forest is based on an ensemble of many classification trees, in which every one of them is constructed based on a bootstrap sample out of the original dataset, which is then split. For each tree, the algorithm randomly selects input variables as potential predictors. The observations outside the bootstrap sample are called the out-of-bag (OOB) data and are used for calculating a test set error rate of the tree. Every subject is likely to be OOB in one-third of the bootstrappings and is predicted under those circumstances. When the specified numbers of trees are added to the forest, there is a final prediction for each subject by aggregating these predictions. Typically, the classification with the most votes (majority vote) over all the trees in the forest is considered. Summarizing the deviations between the observed phenotypes and their predictions produces the OOB test set error rate. This error rate reveals the association between the gene set and the phenotype. A gene-set with a lower error rate is regarded to have a better predicting power with regard to the phenotype variable and hence has a greater significance. Breiman (2001) indicated that unlike the cross-validation, the OOB error rate provides an unbiased estimate of the error rate. Moreover, applying classification trees makes the method time-efficient.

Given an observed OOB test set error rate e_0 in the Random Forest, a permutation-based p-value can be obtained as following,

$$p-\text{value}=\frac{\sum_{k=1}^{N} I\left\{e^{(k)}\leq e_0\right\}}{N}, \tag{1}$$

Table 1
Type I error rate comparisons in the simulation study. Type I error rates of eight GSA methods: RF, Hotelling's T2, PCA, SAM-GS, ANCOVA, Global, GSEA, and MaxMean tests.

Method	$\rho=0$	$\rho=0.3$	$\rho=0.5$	$\rho=0.9$
Hotelling's T^2	0.050	0.039	0.038	0.050
PCA	0.053	0.042	0.052	0.062
SAM-GS	0.046	0.042	0.038	0.055
ANCOVA	0.042	0.038	0.034	0.052
Global	0.001	0.009	0.016	0.034
GSEA	0.059	0.058	0.052	0.048
MaxMean	0.093	0.094	0.107	0.098
RF	0.040	0.034	0.027	0.036