Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Environmental Sciences 8 (2011) 693 – 699

Procedia Environmental Sciences

ICESB 2011: 25-26 November 2011, Maldives

Statistical and Biological Evaluation of Different Gene Set Analysis Methods

Wenjun Cao[a,b,+], Yunming Li[c,+], Danhong Liu[a], Changsheng Chen[a,*], Yongyong Xu[a]

[a]*Department of Health Statistics, Fourth Military Medical University, Xi'an, China*
[b]*Department of Mathmatics, Chang Zhi Medical College, Changzhi, China*
[c]*Department of Quality Management, Military General Hospital of Chengdu PLA, Chengdu, China*

Abstract

Gene-set analysis (GSA) methods have been widely used in microarray data analysis. Owing to the unusual characteristics of microarray data, such as multi-dimension, small sample size and complicated relationship between genes, no generally accepted methods have been used to detect differentially expressed gene sets (DEGs) up to now. Our group assessed the statistical performance of some commonly used methods through Monte Carlo simulation combined with the analysis of real-world microarray data sets. Not only did we discover a few novel features of GSA methods during experiences, but also we find that some GSA methods are effective only if genes were assumed to be independent. And we also detected that model-based methods (GlobalTest and PCOT2) performed well when analyzing our simulated data sets in which the inter-gene correlation structure was incorporated into each gene set separately for more reasonable. Through analysis of real-world microarray data, we found GlobalTest is more effective. Then we concluded that GlobalTest is a more effective gene set analysis method, and recommended using it with microarray data analysis.

Keywords: GSA; Monte Carlo simulation; differentially expressed gene sets (DEGs); statistical inference

1. Introduction

* Corresponding authors. Tel.: 13772051370
E-mail: changshengchen@126.com
[+] Wenjun Cao, Yunming Li contributed equally to this article.

doi:10.1016/j.proenv.2011.10.106

Microarrays are at the center of a revolution in biotechnology, allowing researchers to monitor simultaneously the expression of tens of thousands of genes. Extracting useful information from such gene expression profiles and then implementation of biological interpretation are the main challenges faced by researchers.

In recent years, GSA methods, which are based on pre-defined gene set [1], have been proposed for testing the coordinated association of gene sets with a phenotype of interest. Such an analysis allows researchers to make full use of previously biological knowledge and make the results more interpretable. Gene annotations have been used to define the gene sets, such as Gene Ontology [2] or KEGG [3]. Many authors have reviewed and compared the test power of some of proposed GSA methods from different perspective. For example, Dinu et al. [4] compared the biological performance of six GSA methods, and pointed out the advantages of SAM-GS, GlobalTest and ANCOVA [5] methods. Various GSA methods were reviewed insightfully and a practical guideline was given to researchers by Nam and Kim [6]. The study by Liu et al. [7] compared the performance of three methods (GlobalTest, ANCOVA GlobalTest and SAM-GS) and concluded that the results achieved by the three approaches investigated were broadly similar. A review paper by Song and Black [8] studied the performance of a subset of commonly used approaches through the analysis of both simulated and real microarray data. We confirmed that those GSA methods, which incorporate correlation structures, were promising in detecting DEGs.

In this paper, we mainly explore and compare the performance of some usually used GSA methods from the point of statistical view. Our aim is to find those more effective approaches. We study the sensitivity and specificity of those GSA methods, and then compare their performances through the simulated data in which we incorporate the correlation structure between genes within each gene set, respectively.

2. The statistical performance of GSA methods

2.1. Sensitivity of the methods

Under the assumption that genes are independent, we examined the sensitivity of different DEGs detection methods on simulated data sets. Here, we simulated 1000 genes, and 50 samples in each of two groups (A, B), control and treatment. The genes were assigned to 50 gene-sets, each with 20 genes. All measurements were generated as N (0, 1) before the treatment effect was added. There were five different scenarios:

- All 20 genes of gene-set 1 are 0.20 units higher in group B. $\mu_{B1} - \mu_{A1} = 0.20$
- All 20 genes of gene-set 1 are 0.25 units higher in group B. $\mu_{B1} - \mu_{A1} = 0.25$
- All 20 genes of gene-set 1 are 0.30 units higher in group B. $\mu_{B1} - \mu_{A1} = 0.30$
- All 20 genes of gene-set 1 are 0.35 units higher in group B. $\mu_{B1} - \mu_{A1} = 0.35$
- All 20 genes of gene-set 1 are 0.40 units higher in group B. $\mu_{B1} - \mu_{A1} = 0.40$

To avoid the bias in results, we replicated 20 times in each case. The seeds of the normal distribution were taken from 1 to 20. In every one of these scenarios, 50 gene sets was analyzed simultaneously, and the statistical significance for each gene set was adjusted for multiple testing of many hypotheses. In the simulated data sets, only the first gene-set was of interest. We compared five different GSA methods: GlobalTest-per [9], PCOT2 [10], Efron's GSA [11], SAFE-boot, and SAFE-per [12].

Our study reveals that Efron's GSA has the highest power (Table1, 2), but the test statistic used in this method is re-standardized by gene resample. By doing so, the correlation between genes is entirely ignored in the analysis. The results also show that SAFE-per has a similar performance compared

to GlobalTest-Asy, and the two methos have slightly higher power than SAFE-boot. In addition, the PCOT2 method is not easy to get a smaller P-value. Its effectiveness will be tested in later research, in which the members of a gene set exhibiting strong cross-correlation.

TABLE 1. $\bar{P} \pm S_{\bar{P}}$ FOR THE FIRST GENE SET FOR THE FIVE Different APPROACHES

Methods	$\mu_{B1} - \mu_{A1}$				
	0.20	**0.25**	**0.30**	**0.35**	**0.40**
Efron's GSA	0.075 ± 0.245	0.017 ± 0.075	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
SAFE-boot*	0.191 ± 0.042	0.077 ± 0.024	0.024 ± 0.011	0.005 ± 0.003	0.000 ± 0.000
SAFE-per*	0.143 ± 0.051	0.039 ± 0.021	0.002 ± 0.000	0.001 ± 0.000	0.001 ± 0.000
GlobalTest-Asy*	0.103 ± 0.205	0.034 ± 0.101	0.006 ± 0.022	0.000 ± 0.002	0.000 ± 0.000
PCOT2	0.253 ± 0.269	0.317 ± 0.733	0.076 ± 0.170	0.037 ± 0.139	0.023 ± 0.103

* Globaltest-Asy: using the asymptotic distribution of the test statistic; SAFE-boot: bootstrap-based tests and SAFE-per: permutation-based tests.

TABLE 2. DETECTION RATES FOR THE FIVE GENE SET ANALYSIS METHODS (CONCERNED ONLY THE FIRST GENE-SET)

Methods	$\mu_{B1} - \mu_{A1}$				
	0.20	**0.25**	**0.30**	**0.35**	**0.40**
Efron's GSA	95%	95%	100%	100%	100%
SAFE-boot	20%	60%	85%	95%	100%
SAFE-per	55%	80%	100%	100%	100%
GlobalTest-Asy	65%	85%	95%	100%	100%
PCOT2	45%	55%	80%	95%	95%

2.2. Specificity of the methods

According to Nam et al. [6], we also generated expression profiles of 2000 genes with two groups, each owning 20 samples. The expression values were sampled from a standard normal distribution in both groups. We selected randomly 600 genes and added a random value between 0.5 and 1 to the second group to generate DEGs. The genes were divided into 100 gene sets orderly, each of which contained 20 genes. We compared the detection rates of five GSA approaches. Because the DEGs were chosen uniformly at random, no gene sets were expected to be 'enriched'. Indeed, Efron's GSA method recognized no DEGs, while SAFE-boot detected one DEGs and SAFE-per found four. The false discovery rates of these three methods were all smaller than 0.05. However, GlobalTest-Asy identified 38 DEGs, and the PCOT2 method detected a large part of the gene sets (57%) as differentially expressed with a cutoff of 0.05 (Fig. 1).

From this result, we find that Efron's GSA, SAFE-boot, and SAFE-per agree closely and show higher specificity under the assumption that the genes are independent. However, GlobalTest and PCOT2, which based on self-contained show lower specificity. This is the shortcoming of all self-contained methods because in these methods only a single differential expressed gene can make the whole gene set significant [6].

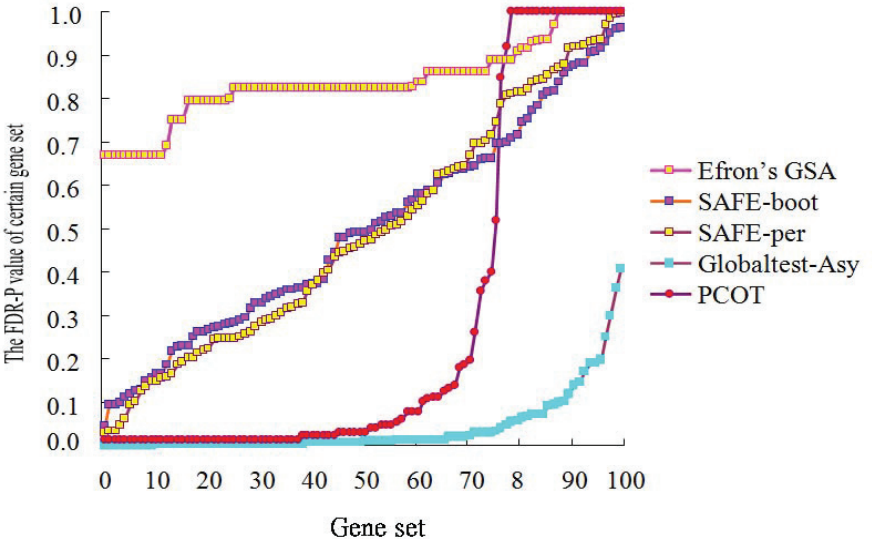

Fig.1. The P-value of 100 no-enriched gene sets

2.3. Taking into account the correlation between genes in simulated experimental data

We usually define gene sets based on prior biological knowledge. Those genes with the same functional or pathway are grouped as a gene set. So there is strongly correlation within a gene set. In the following we will compare the test power of GSA methods through the simulated data which incorporate a correlation structure within each gene set respectively. The simulated data set contained 2000 genes and 40 microarrays, each consisting of 20 control and 20 treated samples. We divided the simulated gene expression profiles into 100 gene sets, each containing 20 genes. Each gene set were generated from multivariate normal distribution with mean μ and variance-covariance symmetric matrix Σ, based on the following density function:

$$f(x)=\frac{1}{(2\pi)^{r/2}\left|\sum\right|^{1/2}}e^{-(x-\mu)'\Sigma^{-1}(x-\mu)/2} \qquad \left(\sum=\begin{bmatrix}1 & r & \dots & r\\ r & 1 & & r\\ \dots & & \dots & \dots\\ r & r & \dots & 1\end{bmatrix}\right).$$

In this formula, we set $\mu=1$ and r=trunc{[(n-1)/10]+1}/10 for the nth gene sets, and observed the adjusted P-values of every gene set by simulating repeatedly 20 times in each case.

Our results showed that the model-based methods (GlobalTest and PCOT2) identified all 100 DEGs. However, the Efron's GSA didn't identify any DEGs. To our surprise, the detection rate of SAFE was 0, 0.08 and 1, when the correlation coefficient between genes (r) was set lower than 0.6, equal to 0.6 and bigger than 0.6, respectively. For further research, we decreased the mean difference between the two

groups to 0.05, and we found the detection rate of SAFE became 0, 0.4 and 1, when the correlation coefficient between genes (r) was set lower than 0.7, equal to 0.7 and bigger than 0.7, respectively. The reason for this phenomenon is that SAFE interprets incorrectly the strong correlation among genes as the differences between the groups.

3. Analysis of real-world microarray data

In this section, we use two well-known gene expression datasets to evaluate the biological performance of the three methods, namely GlobalTest, PCOT2 and SAFE. The first two tests incorporate correlation structures into the analysis process, while the last test performs well in the above simulation data sets relative to other tests.

3.1. Classification of acute leukemia

Golub (1999) [13] Initially, the experimental data contained 7129 genes. In the data pretreatment step we excluded those genes that had very low expression levers. Thus, the data set turned into a matrix of 3051 × 38. Then we used cellular component (CC) branch from the collection of Gene Ontology (GO) Consortium to define gene sets, and obtained 159 functional gene sets. We obtained P-values of these gene set through 2000 permutations of the sample labels, and then adjusted them by BY (Benjamini & Yekutieli) method (Benjamini & Yekutieli) method [14]. The results indicate that GlobalTest identify 151 DEGs, while PCOT2 observe 133 DEGs which are all detected by GlobalTest. And SAFE find only 23 DEGs which are all discovered by GlobalTest and PCOT2.

3.2. Molecular pathways of type II diabetes mellitus

Another data set that we chose was the diabetes data published by Mootha et al [15]. These data contain expression information on 22283 genes and consist of 34 samples: 17 with type II diabetes mellitus (DM2) and 17 with normal glucose tolerance (NGT). The expression information on these genes can available online at http://www.broad.mit.edu/gsea/datasets.jsp. The Bioconductor annotation package hgu133aPATH is used to define 172 gene sets relation to KEGG pathways. Based on the adjusted *P*-values of these gene sets, both the PCOT2 and GlobalTest methods do detect no DEGs in these data information. However, SAFE detects 7 DEGs (Table 3). Combination of single-gene analysis results, we conclude that PCOT2 and GlobalTest methods are more consistent.

TABLE 3. DETECTION RATES FOR THE FIVE GENE SET ANALYSIS METHODS (CONCERNED ONLY THE FIRST GENE-SET)

Set size	KEGGID	KEGG-Name	FDR_*P*		
			SAFE	PCOT2	Globaltest
59	KEGG:03420	Nucleotide excision repair	.0005	.9173	1.0000
37	KEGG:03430	Mismatch repair	.0105	.9182	1.0000
155	KEGG:00190	Oxidative phosphorylation	.0115	.9173	1.0000
50	KEGG:00260	Glycine, serine and threonine metabolism	.0190	.9182	1.0000
44	KEGG:03440	Homologous recombination	.0360	.9182	1.0000

12	KEGG:00900	Terpenoid biosynthesis	.0480	.9173	1.0000
22	KEGG:04614	Renin-angiotensin system	.0480	.9173	1.0000

4. Discussion

In this paper, we have investigated the performance of commonly used gene set approaches through the simulation study. In the absence of gold standards for diagnosis of these approaches, we resorted to simulated data which can be set difference based on requirements. Although we have solved part of the problems to some extent through simulated experiment, it was generally unable to meet the inherent complexity of real microarray data. To reduce the impact of the gene set size in the simulation study, we assumed that all he gene set are the same size in our simulation setting.

In microarray data studies, most researchers are keenly aware of the potentially biological correlation between genes and the distributions of gene expressional levels which is always unknown. The special inter-dependence relationship between information on gene expression and its influence factors is not clear, which is usually complex linear relationship, rather than a simple linear relationship. In addition, the high-dimensional and small sample sizes of microarray data make it have conservative bias for any basic statistical model in data analysis. If we still use the general statistical model to make an inference, then the two limitations may weaken the reliability of records, or even gain wrong conclusions.

Mixed effect model has been developed rapidly in recent years as a statistical analysis method, which overcomes the limitations of traditional linear model. Not only can it handle the non-equilibrium data with random missing, but also it is able to analyze various types of non-independent data, which has been widely used in the analysis of hierarchical data, longitudinal observational data, repeated measurement data and the growth curve data. It has become a hot topic in statistical theory model, called "21st century model". At present, mixed-effects models have been used for single-gene analysis and it is also feasible for gene set analysis. The improvement to microarray data analysis methods will be able to take full advantage of the information contained in the data, and accelerate the development process of functional genomics research.

This paper did not discuss the well-known gene set enrichment analysis (GSEA) method, because doubts were raised in the literature [4, 8, 16]. Among our considered methods, SAFE and Efron's GSA are the improvement of GSEA method, which have better statistical properties. However, GSEA is still a user-friendly, independent software package that based on sample permutation. It is different from other analytical tool through R programming language, which is why GSEA is more convenient to use for majority of biologists.

Acknowledgements

This work was supported by National Science and Technology Support Project Grant from the Ministry of Science and Technology of the People’s Republic of China [2008BAI52B01]; and the science and technology projects in Shaanxi Province [2008K04-02].

References

[1] Subramanian A, Tamayo P, Mootha VK, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Nat Acad S*ci 2005;**102**:15545-50.

[2] Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000; **25**:25-9.

[3] Kanehisa M. Goto S: KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 2000; **28**:27-30.

[4] Dinu I, Liu Q, Potter JD, *et al*. A biological evaluation of six gene set analysis methods for identification of differentially expressed pathways in microarray data. Cancer Inform 2008; **6**:357-68

[5] Mansmann U, Meister R. Testing differential gene expression in functional groups. Goeman's global test versus an ANCOVA approach. Methods Inf Med 2005; **44**:449-53.

[6] Nam D, Kim SY. Gene-set approach for expression pattern analysis. Brief Bioinform 2008; **9**: 189-97.

[7] Liu Q, Dinu I, Adewale AJ, Potter JD, Yutaka Y. Comparative evaluation of gene-set analysis methods. BMC Bioinformatics 2007; **8**:431(1)-(13).

[8] Song S, Black MA. Microarray-based gene set analysis: a comparison of current methods. BMC Bioinformatics 2008; **9**:502(1)-(14).

[9] Goeman JJ, van de Geer SA, de KF, van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics 2004; **20**:93-9.

[10] Kong SW, Pu WT, Park PJ. A multivariate approach for integrating genome-wide expression data and biological knowledge. Bioinformatics 2006; **22**:2373-80.

[11] Efron B, Tibshirani R. On testing the significance of sets of genes. Annals of Applied Statistics 2007; **1**:107-29.

[12] Barry WT, Nobel AB, Wright FA. Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinformatics 2005; **21**:1943–49.

[13] Golub TR, Slonim DK, Tamayo P, *et al*. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 1999; **286**:531-37.

[14] Benjamini Y, Yekuticli D. The control of the false discovery rate in multiple testing under dependency. Ann Stat, 2001; **29**: 1165-1188.

[15] Mootha VK, Lindgren CM, Eeiksson KF, *et al.* PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 2003; **34**:267-73.

[16] Dinu I, Potter JD. Improving gene set analysis of microarray data by SAM-GS. BMC Bioinformatics 2007; **8**:242(1)-(13).