Journal of Zhejiang University SCIENCE B
ISSN 1673-1581 (Print); ISSN 1862-1783 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Science Letters:

A robust statistical procedure to discover expression biomarkers using microarray genomic expression data*

ZOU Yang-yun, YANG Jian, ZHU Jun†‡
(*Institute of Bioinformatics, Zhejiang University, Hangzhou 310029, China*)
†E-mail: jzhu@zju.edu.cn

Received Apr. 4, 2006; revision accepted May 31, 2006

Abstract: Microarray has become increasingly popular biotechnology in biological and medical researches, and has been widely applied in classification of treatment subtypes using expression patterns of biomarkers. We developed a statistical procedure to identify expression biomarkers for treatment subtype classification by constructing an *F*-statistic based on Henderson method III. Monte Carlo simulations were conducted to examine the robustness and efficiency of the proposed method. Simulation results showed that our method could provide satisfying power of identifying differentially expressed genes (DEGs) with false discovery rate (FDR) lower than the given type I error rate. In addition, we analyzed a leukemia dataset collected from 38 leukemia patients with 27 samples diagnosed as acute lymphoblastic leukemia (ALL) and 11 samples as acute myeloid leukemia (AML). We compared our results with those from the methods of significance analysis of microarray (SAM) and microarray analysis of variance (MAANOVA). Among these three methods, only expression biomarkers identified by our method can precisely identify the three human acute leukemia subtypes.

Key words: Microarray, Biomarker, Henderson method III, Gene expression pattern, Mixed linear model
doi:10.1631/jzus.2006.B0603 **Document code:** A **CLC number:** Q332

INTRODUCTION

Microarray technique is a powerful laboratory tool for simultaneously monitoring genome-wide expression in different conditions. One of the most important applications of microarray technique is the classification of tumor subtypes or different disease states to facilitate clinical researchers in diagnostic, therapeutic or prognostic decisions for patients (Golub *et al*., 1999; Alizadeh *et al*., 2000; Spindler, 2006). The generally used approaches, such as cluster analysis and supervised grouping, only focus on the similarity of the data structure, and fail to guarantee that the used class predictors are biologically associated with class distinction. Therefore, prior to cluster analysis, a central need is to explore whether there are some expression biomarkers strongly correlated with specific classes. Some statistical methods such as significance analysis of microarray (SAM) (Tusher *et al*., 2001), fixed ANOVA method (Kerr *et al.*, 2000; Kerr and Churchill, 2001) and mixed linear model method (Wolfinger *et al*., 2001; Jin *et al*., 2001; Lu *et al*., 2005) have been proposed to serve this goal.

Based on Henderson method III, we developed a statistical method under the mixed linear model framework (Zhu, 2000) to objectively identify expression biomarkers for treatment classification. Monte Carlo simulations were conducted to validate the robustness and efficiency of the present method, and a real dataset of leukemia was analyzed to assess the utility of the method.

METHOD FRAMEWORK

As in most analysis methods of microarray, we

‡ Corresponding author
* Project partly supported by the National Basic Research Program (973) of China (No. 2004CB117306) and the National Natural Science Foundation of China (No. 2002AA234031)

first use the normalization procedure to minimize the global systematical variations involved in the experiment from the original fluorescence measurements. The normalization model can be written as

$$y_{ijklm}=\mu+A_i+D_j+P_k+T_l+\gamma_{ijklm}, \quad (1)$$

where μ represents the mean expression level over all genes, fixed effect; A_i is the array effect, random effect, $A_i \sim (0,\sigma_A^2)$; D_j is the dye effect, random effect, $D_j \sim (0,\sigma_D^2)$; P_k is the pin effect, random effect, $P_k \sim (0,\sigma_P^2)$; γ_{ijklm} is residual error, $\gamma_{ijklm} \sim (0,\sigma_\gamma^2)$. γ_{ijklm} is obtained by subtracting the fitted values of the effects in model (1) from base 2 logarithm of background-corrected measurements (y_{ijklm}) using least square estimation (LSE) method, and will subsequently be used as the inputs for the gene-specific models

$$\gamma_{ijglm}=\mu_g+A_{gi}+D_{gj}+T_{gl}+\varepsilon_{ijglm}, \quad (2)$$

where μ_g represents the mean expression level of gene g; A_{gi} is gene specific array effect, $A_{gi} \sim (0,\sigma_{A_g}^2)$; D_{gj} is gene specific dye effect, $D_{gj} \sim (0,\sigma_{D_g}^2)$; ε_{ijglm} is gene-dependent residual error, $\varepsilon_{ijglm} \sim (0,\sigma_{\varepsilon_g}^2)$, which is different from γ_{ijklm} in model (1). Under the null hypothesis H_0: $T_{g1}=T_{g2}=\ldots=0$, Henderson method III is employed to construct the F-statistic to test the significance of treatment effects (Searle, 1971). Since analysis of microarray data involves multiple statistical tests, we use false discovery rate (FDR) (Benjamini and Hochberg, 1995) to control the experimental-wise type I error. The identified differentially expressed genes (DEGs) are ranked by their statistic scores which can provide more information and choice for biologists.

Finally, the potentially DEGs detected by model (2) are fitted in the following full model to estimate the variance components and effects interested.

$$y_{ijkglm}=\mu+G_g+A_i+D_j+P_k+T_l+GA_{gi}+GD_{gj}+GT_{gl}+\varepsilon_{ijkglm}, \quad (3)$$

where μ and G_g are fixed effects, and A_i, D_j, P_k, GA_{gi}, GD_{gj}, ε_{ijkglm} are random effects normally distributed with zero means and variance components σ_A^2, σ_D^2, σ_P^2, σ_{GA}^2, σ_{GD}^2, σ_ε^2, respectively. The terms T_l, T_{gl} and GT_{gl} in models can be regarded as fixed effects or random effects according to the experimental intent. Since different pins have different characteristics and surface properties with different amounts of target cDNA, we include the P_k effect in our model. These models are extensible to more complex situations such as N-dyes, multiple factors decomposed from the treatment effect and other variations like fluctuations due to mRNA extraction, cDNA synthesis. Markov Chain Monte Carlo (MCMC) method (Wang *et al*., 1994) is used to estimate the variance components of random effects in the model, to estimate fixed effects, and to predict random effects as well.

RESULTS

Simulation analysis

Different variation magnitudes were set according to the results from the previously analyzed real dataset available in Stanford microarray database. We assigned the residual variance as 1 with the proportion of gene by treatment interaction effect (V_{GT}) variance to the residual variance (V_{GT}/V_ε) ranging from 1 to 10. Two assumptions were adopted: (1) All parameters in the model follow independent and identical normal distribution, denoted as NormAspt; (2) The observations from different array do not share the same variance, denoted as ArrayHetero. Simulation datasets were generated from different experimental designs, loop design with spots replicated within single array (denoted as LOOPREPIN), loop design with spots replicated between arrays (denoted as LOOPREPOUT), reference design with spots replicated within single array (denoted as REFREPIN), reference design with spots replicated between arrays (denoted as REFREPOUT). All the simulation datasets were run with 200 replicates, with powers of identifying DEGs with FDR control at 0.05 being shown in Fig.1. Simulation results revealed that our method was appealing for identifying DEGs validated with high power when V_{GT}/V_ε exceeded 2, especially in the case of LOOPREPIN. In addition, it was shown that our method could offer intriguing stability under different assumptions, which would be important for our method for analyzing microarray data when array

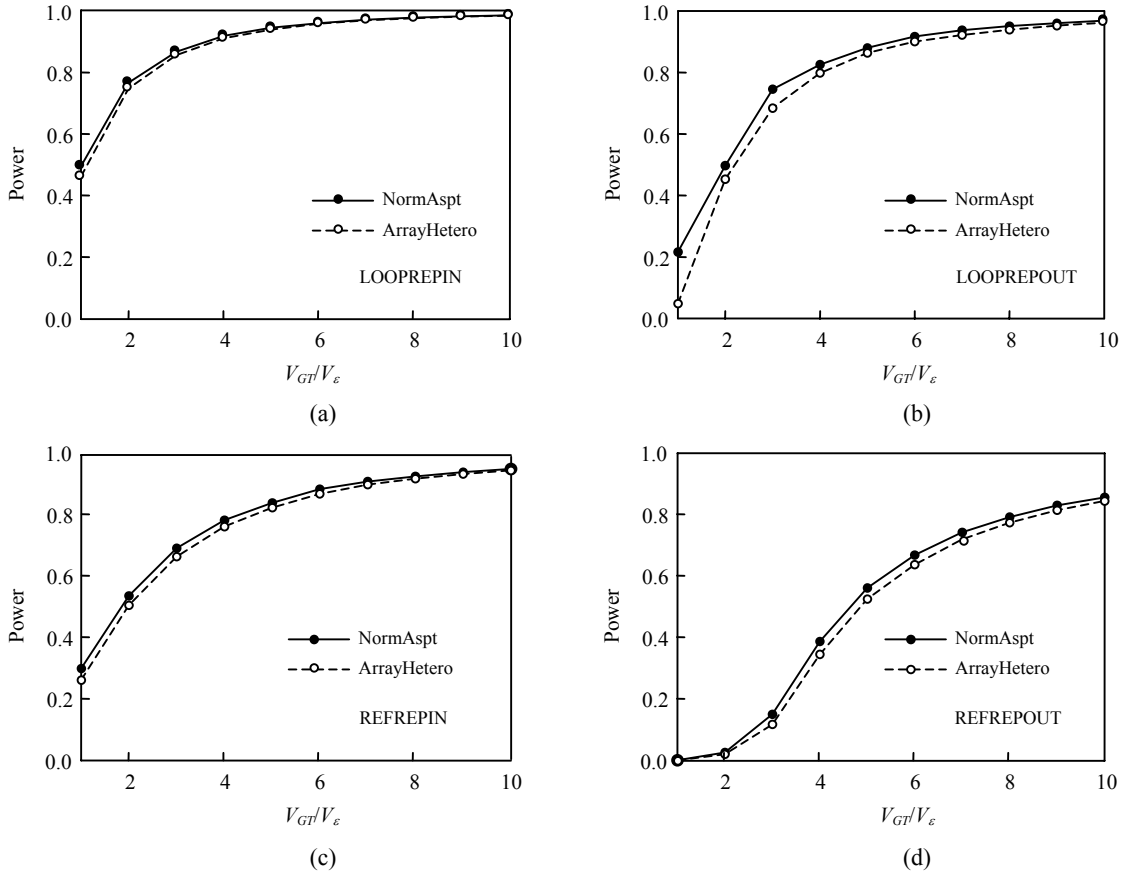

Fig.1 Powers of identifying DEGs under assumptions of NormAspt (solid lines) and ArrayHetero (dotted lines) with varied proportion of *GT* variance component to the residual variance in different experimental designs. (a) Powers of identifying DEGs in loop design with spots replicated within array; (b) Powers of identifying DEGs in loop design with replications between arrays; (c) Powers of identifying DEGs in reference design with spots replicated within array; (d) Powers of identifying DEGs in reference design with replications between arrays

variance heterogeneity is quite common in microarray data. However, in the case of REFREPOUT, the result was not perfect, due to the severe confounding between other variations like *GA* or *GD* and variance of *GT* effect in this experimental design. Therefore, it is strongly recommended that appropriate experimental design should be used, such as loop design with spots replicated within an array or multi-color microarray design (Woo *et al*., 2005).

Worked example

Previous study of leukemia (Golub *et al*., 1999) monitored expression patterns from 38 leukemia patients (with clinically predefined of T-cell ALL, B-cell ALL and AML) to develop an expression-based molecular classification method for acute leukemia as an assistant tool of clinical diagnosis. Affymetrix Hu6800 GeneChips were used. Since the sample grouping of these datasets has been clinically verified, we could use it to validate the utilization of our method. And we also used the methods of SAM and MAANOVA to analyze this dataset for comparison. By a default in the configurations, MAANOVA could only identify 102 marker genes, less than those identified by the other two methods. Thus, we used the top ranked 102 genes from our method and SAM, as well as 102 significant genes from MAANOVA to classify the samples by hierarchical cluster using Pearson correlation distance with UPGMA-linkage criterion. In distinguishing two predefined classes of leukemia ALL and AML, our method yielded accurate classifications, while SAM confused to classify 3 ALL samples into the AML samples and MAANOVA incorrectly placed 8 ALL

samples and 5 AML samples to each other. Expression biomarkers identified by our method were also sensitive for partitioning ALL samples into T-cell ALL and B-cell ALL subclasses with only two B-cell samples and one T-cell sample wrongly classified, while SAM and MAANOVA entirely failed to do it.

Since putative cluster labels are available for leukemia data, external indices of adjusted rank index (Hubert and Arabie, 1985), Jaccard index (Jain and Dubes, 1988) and FM index (Fowlkes and Mallows, 1983) were computed to evaluate the quality of cluster results. These external indices have the property that the higher the score, the better the cluster solution, with a score of 1.0 indicating a perfect solution. Compared cluster results showing dissimilar structure of expression data were due to the different biomarkers identified by our method, SAM, and MAANOVA, respectively, were summarized in Table 1. Our method showed the highest scores in these three cluster validation measurements, indicating that the biomarkers discovered by our method were very close to these different classes of leukemia.

Table 1 Comparison results of three methods

Method	Two classes			Three classes		
	Adjusted rank	Jaccard	*FM*	Adjusted rank	Jaccard	*FM*
Our method	1.000	1.000	1.000	0.776	0.747	0.856
SAM	0.893	0.911	0.954	0.533	0.568	0.731
MAANOVA	0.012	0.383	0.554	0.076	0.252	0.403

Note: Two classes: ALL and AML subclasses; Three classes: T-cell ALL, B-cell ALL and AML

DISCUSSION

The recognition of objective expression biomarkers plays a crucial role in correct classification of tumor subtypes which is valuable for assisting in clinical diagnosis. Many standard statistical methods have been used to address the issue, but none has yet obtained widespread acceptance because of the high rates of false discovery. In the present study, we implement a novel statistical approach in three interconnected steps, normalization (model (1)), gene-specific model fitting (model (2)) and multiple genes model fitting (model (3)). In the second step (model (2)), an *F*-statistic is constructed via Henderson method III to scale the expression change among different treatments of each gene. This strategy is quite efficient in terms of statistical power and computation. Simulation results in multiple configurations and the real leukemia data analysis showed that our method can improve the ability to correctly identify DEGs or expression biomarkers in expression profiles analysis.

Meanwhile, technical and stochastic variations such as mRNA extraction, cDNA synthesis, labelling reactions and print or hybridization efficiency, are usually involved in microarray experiments, and inevitably lead to the noise in raw expression measurements and bias interpretation of class distinction. So it is recommended to do the third step—multi-gene model fitting (model (3)) using the MCMC method which can give unbiased prediction of *GT* interaction effects for cluster analysis or discriminant analysis. Besides, the estimates of various sources of variation can provide some feedback on the quality of the experiment to researchers, which is essential for improving the laboratory protocols for further experiments. For example, if variations of *A* and *GA* effects is large, it is essential to re-select appropriate array with inherently less variations, or use finer experimental design so as to construct appropriate statistical model to screen these variations.

This is only a preliminary study, with the detailed research carried out in our following paper.

ACKNOWLEDGEMENT

We thank Guobo Chen and Lide Han for providing helpful discussion and comments on this manuscript.

References

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., *et al*., 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**(6769):503-511. [doi:10.1038/35000501]

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, **57**(1):289-300.

Fowlkes, E.B., Mallows, C.L., 1983. A method for comparing two hierarchical clusterings. *J. American Statistical Association*, **78**(383):553-569. [doi:10.2307/2288117]

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., *et al*., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**(5439):531-537. [doi:10.1126/science.286.5439.531]

Hubert, L., Arabie, P., 1985. Comparing partitions. *Journal of Classification*, **2**(1):193-218. [doi:10.1007/BF01908075]

Jain, A.K., Dubes, R.C., 1988. Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, NJ.

Jin, W., Riley, R.M., Wolfinger, R.D., White, K.P., Passador-Gurgel, G., Gibson, G., 2001. The contributions of sex, genotype and age to transcriptional variance in drosophila melanogaster. *Nature Genetics*, **29**(4):389-395. [doi:10.1038/ng766]

Kerr, M.K., Churchill, G.A., 2001. Experimental design for gene expression microarrays. *Biostatistics*, **2**(2):183-201. [doi:10.1093/biostatistics/2.2.183]

Kerr, M.K., Martin, M., Churchill, G.A., 2000. Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, **7**(6):819-837. [doi:10.1089/10665270050514954]

Lu, Y., Zhu, J., Liu, P., 2005. A two-step strategy for detecting differential gene expression of cDNA microarray data. *Current Genetics*, **47**(2):121-131. [doi:10.1007/s00294-004-0551-3]

Searle, S.R., 1971. Linear Models. John Wiley & Sons, New York.

Spindler, S.R., 2006. Use of microarray biomarkers to identify longevity therapeutics. *Aging Cell*, **5**(1):39-50. [doi:10.1111/j.1474-9726.2006.00194.x]

Tusher, V.G., Tibshirani, R., Chu, G., 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Nat. Acad. Sci. USA*, **98**(9):5116-5121. [doi:10.1073/pnas.091062498]

Wang, C.S., Rutledge, J.J., Gianola, D., 1994. Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs. *Genetics Selection Evolution*, **26**:91-115.

Wolfinger, R.D., Gibson, G., Wolfinger, E.D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., Paules, R.S., 2001. Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology*, **8**(6):625-638. [doi:10.1089/10665270 1753307520]

Woo, Y., Krueger, W., Kaur, A., Churchill, G., 2005. Experimental design for three-color and four-color gene expression microarrays. *Bioinformatics*, **21**(Suppl. 1):i459-i467. [doi:10.1093/bioinformatics/bti1031]

Zhu, J., 2000. Mixed linear model approaches for analyzing genetic models of complex quantitative traits. *Journal of Zhejiang University SCIENCE*, **1**(1):78-90.