浙江大学学报(农业与生命科学版) 30(5): 492～494, 2004
Journal of Zhejiang University (Agric. & Life Sci.)

Article ID: 1008-9209(2004)05-0492-03 【Research notes】

A new method for analyzing gene expression data

PAN Hai-yan[1,2], ZHU Jun[1], HAN Dan-fu[2]

(1. *Institute of Bioinformatics, Zhejiang University, Hangzhou* 310029, *China*;
2. *Department of Mathematics, Zhejiang University, Hangzhou* 310027, *China*)

分析基因表达谱数据的新方法

潘海燕[1,2],朱 军[1], 韩丹夫[2]

(1. **浙江大学 生物信息研究所,浙江 杭州** 310029; 2. **浙江大学 数学系,浙江 杭州** 310027)

Microarray technique provides a systematic genome-wide approach to solve a wide range of problems such as gene functions, gene regulations, and the disease diagnoses and treatments. A key step in the analysis of gene expression data is to identify biologically relevant groups of genes or tissue samples that have similar expression patterns. However, systematic and stochastic fluctuations are usually involved in microarray experiments[1], so the raw measurements have inherent 'noise' within microarray experiments. In current, logarithmic ratios are usually analyzed directly by various clustering methods, which may introduce bias interpretation in identifying groups of genes or samples. In the present study, a new method based on mixed model approaches is proposed for cluster analysis of gene expression data. It is expected to minimize or eliminate inherent 'noise' in microarray experiments and to make sure the inputs of cluster analysis are more biologically meaningful. Meanwhile, we present a windows-interface software, called ClusterProject, for gene expression analysis and visualization.

1 Materials and Methodologies

1.1 Statistical framework

The basis of this method is to construct a statistical model for a gene expression data. Let y_{ijkl} is the measurement from array i, variety j, dye k, and gene l, an overall ANOVA model is

$$y_{ijkl}=\mu+A_i+V_j+D_k+G_l+GA_{li}+GV_{lj}+GD_{lk}+\varepsilon_{ijkl} \quad (1)$$

where the generic term "variety" refers to the mRNA samples under study which could be

Received date: 2004-06-18
Foundation item: National Technology Research and Development Program (863) of China (2002AA234031).
Biography: PAN Hai-yan, female, born in 1979 in TaiZhou, Zhejiang, P.G. student, specializing in Bioinformatics.
Corresponding author: ZHU Jun, male, professor, specializing in Bioinformatics and Quantitative Genetics. Tel: 0571-86971731; E-mail: jzhu@zju.edu.cn.

treatments, tissue types or time points in a biological process. The detailed interpretations of these effects were described in Kerr *et al*[2]. The effects of the interactions between genes and varieties are biological interest among these effects. These terms reflect differences in expression for particular variety and gene combinations that are not explained by the average effects of those varieties and genes.

We use AUP method[3] to predict the *GV* effects and the *t*-test based on jackknife procedures to test for the significance of *GV* effects. Hypotheses can be made about each *GV* interaction effect, H_0: $GV_{lj}=0$ vs. H_1: $GV_{lj}\neq 0$. If H_0 about GV_{lj} in the null hypothesis is accepted, the effect of GV_{lj} is set as zero. The significant level is set at 0.05 for each *GV* interaction effect. The predicted *GV* effects can be used as the inputs of cluster analysis.

1.2 Hierarchical clustering methods

Hierarchical clustering methods are mostly used by biologists to produce a hierarchical tree of clusters. This hierarchical tree provides potentially useful information about the relationships between clusters and can be broken into the desired number of clusters by cutting across the tree at a particular height. Four hierarchical clustering methods (complete-linkage[4], UPGMA-linkage[4], UPGM-linkage[5] and Diana[6]) with one minus Pearson correlation are employed to analyze for the phenotypic values of $\log_2$(Ratios) and the predicted effects of GV_{lj}, respectively.

1.3 Assessment of solutions

Assessing and interpreting the clustering results are as important as generating the clusters. Different measures are applicable in different situations, depending on the information available such as whether a partial true solution is known or not. Because the true cluster labels are available for the gene expression data used, the *Jaccard coefficient*[7] is adopted to evaluate the quality of cluster results. This index has a property: the higher the score, the better the solution. Especially, score one suggests a perfect solution.

1.4 A worked example

We use the B-cell lymphoma data[8] to elucidate the utility of our approach. The final data analyzed in our approach consists of 45 DLBCL tumor samples (22 GC B-like DLBCL and 23 Activated B-like DLBCL). Among these samples, 23 samples have two replicated arrays, one sample has three arrays and the others have only one array. Thus, total 70 arrays are used for the experimental analysis. There are missing values in this data. The original data and information can be available at http://llmpp.nih.gov/lymphoma. We use a *t*-test to select 100 genes that are differentially expressed between DLBCL subtypes for further cluster analysis.

Varieties are completely confounded with dyes because each variety is labeled with only one dye. So the dye effects and *GA* interaction effects are excluded from full model (1). The model for this data is

$$y_{ijk}=\mu+A_i+V_j++G_k+GV_{kj}+\varepsilon_{ijk} \qquad (2)$$

where y_{ijk} is the base-2 logarithms of ratio, and $i=1,\cdots,70$ arrays; $j=1,\cdots,45$ varieties (tissue samples); and $k=1,\cdots,100$ genes. We use the methods described in the part of statistical framework to predict and test for the *GV* effects in model (2).

2 Results and Discusses

Four clustering algorithms under consideration are applied to clustering for the phenotypic values of $\log_2$(Ratios) and the predicted *GV* effects, respectively. *Jaccard coefficient* is computed to compare the cluster results. The implementation results are displayed in Table 1.

Table 1 The *Jaccard* values of four clustering methods with $\log_2$(Ratios) and *GV* effect

	$\log_2$(Ratios)	*GV* effects
Complete-linkage	0.552	0.913
UPGMA-linkage	0.913	1.000
UPGM-linkage	0.837	1.000
Diana	0.713	1.000

From Table 1, when clustering for log2 (Ratios), UPGMA-linkage produces the best result than the others. It misclassified only one sample, UPGM-linkage misclassifies two samples, Diana misclassifies four samples, and complete-linkage misclassifies eight samples. As clustering for the *GV* effects, the performance of complete-linkage has been greatly improved, it misclassifies only one sample. The other three methods all properly classify the subtypes of DLBCL.

In the present study, a statistical method based on mixed model approaches is proposed to attempt to minimize or eliminate inherent 'noise' in microarray experiments. The underlying basic principle of this method is to partition the total observed gene expression into various variations caused by different factors and to predict the genetic effects. The predicted *GV* effects are more biologically meaningful than the raw $\log_2$(Ratios). The results show that using the predicted *GV* effects to construct clusters may improve the quality of cluster result. Therefore, the clustering algorithms may be benefited especially when the noise of the employed data is high.

References:

[1] Schuchhardt J, Beule D, Malik A, *et al*. Normalization strategies for cDNA microarrays[J]. **Nucl. Aci. Res.**, 2000, 28: 47.

[2] Kerr M K, Martin M, Churchill, G A. Analysis of variance for gene expression microarray data[J]. **J. Comput. Biol.**, 2000, 7: 819-837.

[3] Zhu J, Weir B S. Diallel analysis for sex-linked and maternal effects[J]. **Theor. Appl. Genet.**, 1996, 92: 1-9.

[4] Spath H. **Cluster Analysis Algorithm**[M]. Chichester: Ellis Horwood, UK, 1989.

[5] Sokal R R, Michener C D. A statistical method for evaluating systematic relationships[J]. **Univ. Kans. Sci. Bull.**, 1958, 38: 1409-1438.

[6] Kaufman L, Rousseeuw P J. **Finding Groups in Data** [M]. John Wiley, New York, 1990.

[7] Jain A K, Dubes R C. **Algorithms for Clustering Data** [M]. Englewood Cliffs: Prentice Hall, NJ, 1988.

[8] Alizadeh A A., Eisen M B, Davis R E, *et al*. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling[J]. **Nature**, 2000, 403: 503-511.