Clustering Gene Expression Data Based on Predicted Differential Effects of GV Interaction

Hai-Yan Pan[1,2], Jun Zhu[1]*, and Dan-Fu Han[1,2]

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

Microarray has become a popular biotechnology in biological and medical research. However, systematic and stochastic variabilities in microarray data are expected and unavoidable, resulting in the problem that the raw measurements have inherent "noise" within microarray experiments. Currently, logarithmic ratios are usually analyzed by various clustering methods directly, which may introduce bias interpretation in identifying groups of genes or samples. In this paper, a statistical method based on mixed model approaches was proposed for microarray data cluster analysis. The underlying rationale of this method is to partition the observed total gene expression level into various variations caused by different factors using an ANOVA model, and to predict the differential effects of GV (gene by variety) interaction using the adjusted unbiased prediction (AUP) method. The predicted GV interaction effects can then be used as the inputs of cluster analysis. We illustrated the application of our method with a gene expression dataset and elucidated the utility of our approach using an external validation.

Key words: gene expression, clustering analysis, predicting GV interaction effects

Introduction

Monitoring tens of thousands of genes in parallel under different experimental environments or across different tissue types provides a systematic genome-wide approach to help in understanding a wide range of problems, such as gene functions in various cellular processes, gene regulations in different cellular signaling pathways, the diagnose of disease conditions, and the effects of medical treatments. A key step in the analysis of gene expression data is the identification of biologically relevant groups of genes or tissue samples that have similar expression patterns. When clustering genes across many samples, unknown gene function may be inferred from clusters of genes similarly expressed (*1–3*). By clustering samples over the expression levels of multiple genes, novel disease subgrouping may be identified (*4–6*).

Gene expression data can be analyzed by various clustering methods, including hierarchical clustering (*3*, *7*), self-organizing maps (*8*), K-means (*9*, *10*), and graph theoretic approaches of CAST (*11*), HCS (*12*), and CLICK (*13*). However, systematic and stochastic fluctuations are usually involved in microarray experiments (*14*). Therefore, the raw measurements have inherent "noise" within microarray experiments, which may introduce bias in identifying groups of genes or tissue samples and result in the false interpretation of expression patterns. It needs an approach to minimize or eliminate inherent "noise" in microarray experiments and to make the inputs of cluster analysis more biologically meaningful. Mixed model approaches are widely used to partition the sources of variation of observed phenotypes. They have the flexibility to handle a wide variety of experimental designs and data shapes (including balanced and unbalanced data), and to be easily extended to more complicated biological models. Mixed model approaches have been applied to detect significantly differential expression genes (*15*, *16*).

*** Corresponding author.**
E-mail: jzhu@zju.edu.cn

Hierarchical clustering methods

Hierarchical clustering methods are popularly used by biologists to produce a hierarchical tree of clusters (*3*, *7*). The dendrogram provides potentially useful information about the relationships among clusters and can be broken into the desired number of clusters by cutting across the tree at a desired height. Accord-

ing to the methods that produce clusters, hierarchical clustering algorithms can be further divided into agglomerative algorithms and divisive algorithms.

Agglomerative clustering starts with the points as individual clusters and then iteratively merges the two closest clusters together. This iterative merging procedure continues until only one cluster is remaining. Different criteria of measuring the similarity between a pair of clusters yield different cluster algorithms. All algorithms are based on distance metrics for measuring the similarity of a pair of points. Euclidian distance and one minus the Pearson correlation coefficient are two commonly used distance metrics to measure the proximity of a pair of points in clustering expression profiles. The Pearson correlation coefficient $R(\mathrm{X,Y})$ and Euclidian distance $D(\mathrm{X,Y})$ between X and Y are defined as follows,

$$\text{Pearson correlation } R(X,Y) = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

$$\text{Euclidian distance } D(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i-Y_i)^2}$$

Using these two distance metrics, the distance between a pair of clusters can be computed in two ways. In complete-linkage criterion, the distance between two clusters is simply the maximum metric between a point in one cluster and a point in another cluster. In UPGMA-linkage criterion, the distance between two clusters is calculated as the average distance of the pairwise distances between the points in one cluster and the points in another cluster.

Divisive clustering starts with one, all-inclusive cluster and, at each step, the biggest group is broken down into two smaller groups until each cluster contains only a single sample. In this case, we need to decide which cluster to split and how to split the bigger one into two smaller ones at each step. Different decisions and split criteria can generate different divisive clustering algorithms such as DIANA (DIvisive ANAlysis). The detailed implementation of DIANA can be seen in Kaufman and Rousseeuw (*17*).

Assessment of clustering results

A key question in the design and analysis of clustering techniques is how to evaluate the clustering results. Different measures are applicable in different situations, depending on the information available such as whether a partial true solution is known or not. Jain and Dubes (*18*) divided cluster evaluation indices into two main categories: internal and external criterion. Internal criterion measures the quality of clusters based only upon the data, whereas external criterion measures the agreement between the derived clusters and some external gold standards. The external criterion analysis has the strong capacity of providing an independent, hopefully unbiased assessment of cluster quality. Because the inputs of the predicted GV (gene by variety) effects for a cluster method are not the same as compared with the raw $\log_2$(Ratios), in this situation, internal criteria such as figure of merit (FOM; ref. *19*) or silhouette width (*20*) are not suitable to assess the quality of cluster results. Since the putative cluster labels have been available for the gene expression dataset used, an external index of *Jaccard coefficient* has been adopted to evaluate the quality of clustering results. The *Jaccard coefficient* is defined as the proportion of correctly identified mates in the derived solution to the sum of correctly identified mates plus the total number of disagreements between the derived solution and the putative solution (a disagreement is a pair that are mates in one solution and non-mates in the other). The higher the score, the better the solution, and a score of 1.0 suggests a perfect solution. Sharan *et al* (*13*) applied this index to evaluate the clustering results.

We presented a statistical method based on mixed model approaches for cluster analysis of microarray data. The objective of this method was to partition the observed total gene expression level into various variations caused by different factors using an ANOVA model, and to predict the differential effects of GV interaction using the adjusted unbiased prediction (AUP) method (*21*, *22*). Then we applied three hierarchical clustering methods: complete-linkage (*23*), UPGMA-linkage (*23*), and DIANA (*17*) to clustering for the phenotypic values of $\log_2$(Ratios) and the predicted differential effects of GV interaction, respectively. The utility of our method on the task of clustering genes was judged by *Jaccard coefficient* (*18*).

We developed a windows-interface software (ClusterProject) for analysis and visualization of gene expression data. This software with a graphical user interface contains various clustering methods, similarity metrics, and the evaluation metrics, as well as multi-variant analysis including PCA (principal component analysis) and the mixed model approach. It can visualize the raw expression data and the cluster results in several ways. The software is available at http://ibi.zju.edu.cn/software/clusterproject/.

Model

Prediction of GV interaction effects

A microarray experiment is a multi-step process and each step may introduce a potential source of variation. The variation of the measured gene expression data can be generally classified into three generic categories: biological variation, technical variation, and residual variation (*24*, *25*). Biological variation in measured gene expression accounts for the variation from different mRNA sources, such as different animals, cell lines, or tissues. Technical variation refers to the variation coming from the use of the microarray system, such as the sample preparation procedures, the hybridization and washing procedures, the detection method of gene expressions, and laboratory environmental conditions. Residual variation accounts for sampling or experimental error or other unexplainable factors. The variation in a measured gene expression is the sum of these three variations.

Our approach centered around the ANOVA model of Kerr *et al* (*26*) for the analysis of microarray data. The ANOVA model is a popular statistical approach to account for different sources of variation. It can consider all possible sources of variation in a microarray experiment and use one equation to summarize them. The exact form of the ANOVA model depends on the particular experiment. That is, one should determine which sources of variation are present in each experiment individually and construct the model accordingly. Let y_{ijkl} be the observed gene expression measurement from gene i, dye j, array k, and variety l, then an overall ANOVA model is

$$y_{ijkl} = \mu + G_i + D_j + A_k + V_l + GD_{ij} + GA_{ik} + GV_{il} + \varepsilon_{ijkl} \quad (1)$$

where μ is the average of overall expression levels, a fixed effect; G_i is the fixed effect of the *i-th* gene. The effects of dye D_j, array A_k, variety V_l, gene by dye interaction GD_{ij}, gene by array interaction GA_{ik}, gene by variety interaction GV_{il}, and residual ε_{ijkl} are all random variables with zero means and variance components σ^2_D, σ^2_A, σ^2_V, σ^2_{GD}, σ^2_{GA}, σ^2_{GV}, σ^2_ε, respectively. The generic term "variety" refers to the effect of treatments, tissue types, or time points in a biological process (*26*). The interaction effects of genes by varieties interaction (GV_{il}) are biologically interesting among these effects. These terms reflect differences in expression of genes to particular varieties that are not explained by the marginal effects of genes and varieties.

Variance components of the aforementioned models can be estimated using restricted maximum likelihood estimation (REML) and minimum norm quadratic unbiased estimation (MINQUE) (*27*). The random effects of GV can be predicted by the best linear unbiased prediction (BLUP; ref. *28*) and the AUP method (*21*, *22*). We used MINQUE (1) to estimate the variance components and the AUP method to predict the random effects. MINQUE (1) is a MINQUE method with all the prior values setting as 1.0. The predicted differential effects of GV interaction were used as the inputs for further cluster analysis.

Application to yeast sporulation data

We applied our method to the analysis of gene expression data on the transcriptional program of sporulation in budding yeast collected and analyzed by Chu *et al* (*2*). The data set is publicly available at http://cmgm.stanford.edu/pbrown/sporulation. In this experiment, cDNA microarrays containing 97% of the totally 6,118 known and predicted genes of yeast were used to study gene expression during meiosis and spore formation. The mRNA samples were taken at seven time points: 0, 0.5, 2, 5, 7, 9, and 11.5 h. For each time point, the researchers prepared a "red"-labeled cDNA pool. Meanwhile, time-0 sample was served as a reference pool for all of the samples taken from seven time points and was labeled with "green" fluorescent dye. Seven microarrays were used in the study, and each array was probed with the green-labeled sample mixed with one of the seven red-labeled samples.

Each spot contained four measurements: red signal, red background, green signal, and green background. The background-normalized ratio (red signal − red background)/(green signal − green background) was used as respective expression level of a gene at each time point. In addition, Chu *et al* (*2*) described a small set of hand-picked representative genes from each of the seven temporal classes that were expressed during sporulation. Two genes (MRD1 and NAB4) for profile 3, and two genes (KNR4 and EXO1) for profile 4 could not be found at the publicly available data file. The remaining 36 genes were used for next analysis.

We modified the preceding full mixed linear model to support this specific data. The modification of model (1) is

$$y_{ijkl} = \mu + G_i + D_j + A_k + V_l + GA_{ik} + GV_{il} + \varepsilon_{ijkl} \quad (2)$$

where y_{ijkl} is the background-corrected base-2 logarithm of individual intensity measurement, and $i = 1, \ldots, 36$ genes; $j = 1, 2$ dyes; $k = 1, \ldots, 7$ arrays; and $l = 1, \ldots, 7$ varieties (time points). It is not possible to fit the full model (1) that includes GD interaction effects to this experimental design because 0 residual degree of freedom remains. So we excluded the GD effect from the full model. Model (2) is similar to the model of Kerr and Churchill (*29*) of this gene expression data, which used AD effect instead of variety effect.

Results

The proportions of variance components to the total variance in model (2) were summarized in Table 1. Variety effects and gene by variety effects contributed largely to the variation of gene expression (45.2% and 36.4%, respectively). It elucidated that the variation of gene expression was mainly determined by variety (time point) and gene by variety interaction. There is strong evidence that the expression levels of genes vary from different time points. The proportion of residual variation to the total variance was small (3.9%).

Furthermore, the GV effects in model (2) were predicted by the AUP method. Each of the three clustering algorithms with one minus Pearson correlation metric and Euclidian metric was applied to clustering for the phenotypic values of $\log_2$(Ratios) and the predicted GV effects, respectively. *Jaccard coefficient* was computed for each run to assess the quality of each obtained cluster result. The comparisons of cluster results for the yeast sporulation data were shown in Table 2. It was obvious that the cluster results of three clustering methods have been improved when using the predicted GV effects as the inputs of cluster analysis instead of the phenotypic values of $\log_2$(Ratios) except UPGMA-linkage with Euclidian. Three clustering methods with Euclidian all correctly discovered the genes of Metabolic class for $\log_2$(Ratios) and GV effects. These clustering methods with Pearson correlation also correctly discovered the metabolic genes for GV effects except DIANA. However, when clustering for $\log_2$(Ratios), the Metabolic class was partitioned into two subclasses (one group includes SIP4, CAT2, YOR100C, CAR1, AGA2, and YPR192W, another includes ACS1 and PYC1). DIANA with Euclidian produced the best performance using GV effects, and it accurately discovered three classes (including Metabolic, Early I, and Late).

Table 1 Variance Component Estimates and Their Proportions to Total Variance for Yeast Sporulation Data

Parameter	Estimate	Proportion
σ^2_D/σ^2_T	0.002	0.001
σ^2_A/σ^2_T	0.119	0.045
σ^2_V/σ^2_T	1.194	0.452
σ^2_{GA}/σ^2_T	0.262	0.099
σ^2_{GV}/σ^2_T	0.962	0.364
$\sigma^2_\varepsilon/\sigma^2_T$	0.102	0.039

Table 2 Comparisons of Three Clustering Methods with $\log_2$(Ratios) and GV Effects for Yeast Sporulation Data (Model 2)

Method	Pearson		Euclidian	
	$\log_2$(Ratios)	GV effects	$\log_2$(Ratios)	GV effects
Complete-linkage	0.369	0.420	0.390	0.487
UPGMA-linkage	0.291	0.395	0.338	0.315
DIANA	0.301	0.311	0.391	0.500

Discussion

Microarray technologies provide an overall, simultaneous view on the expression levels of tens of thousands of genes under different conditions or processes. Large numbers of valuable datasets have been produced to serve biological and biomedical researches (*30*). Finding structure in a large dataset is a venerable, well-studied problem that is routinely implemented as a first step of data mining. Finding groups of similarly expressed genes or tumors in a microarray data is very valuable to help in understanding gene functions and gene regulations, and to assist in clinical treatments.

However, since there are many different variations induced in different stages of microarray experiments, the identification and estimation of different sources of variation are fundamental to the design of cost-efficient microarray experiments.

Genes, dyes, arrays, varieties (treatments, time points or disease types), and their interactions are well known as the source of effects contributing to variation in the micaroarray experiments (*24*, *26*). In the present study, a statistical method based on mixed model approaches was proposed to assist the cluster analysis of gene expression data. The underlying basic principle of this method is to use the constructed model to partition the total gene expression variation into various components caused by different factors and then predict the differential effects of GV interaction in the model by the AUP method. The mixed model method provides an automatic correction for the nuisance effects in estimating the relative expression of genes across experimental samples. GV interaction effects capture the departure from the overall averages that reflect the biologically relative expressions for the specific combination of the gene and the variety. These effects exclude the contributions of the genes, dyes, arrays, their interactions effects and random error effects on the gene expression, so it is more biologically meaningful than the raw expression measurements. Using predicted GV interaction effects as the inputs of cluster analysis to construct clusters can decrease the noise blight on the cluster result. The result of the yeast sporulation data elucidated the utility of using GV effects as inputs.

Replications allow for assessment of the variability of expression data (for example, in RNA isolation, labeling efficiency, or in chip quality), so that formal statistical analysis methods can be applied. Replication is an important aspect in microarray design. Two basic types of replications can be incorporated within or between arrays: 1) biological replication in which mRNA samples taken from multiple populations can be used on multiple arrays; and 2) technical replication in which the same mRNA samples can be repeated on multiple arrays, or multiple clones or probes of the same gene can be spotted multiple times on the array. So replication can minimize technical artifacts and assess biological variability and is the key to the accuracy and reliability of the data. Whether biological or technical replication or both of the two are used in microarray experiments depends on the relative magnitude of biological and technical variability in the sample. Replication of the same genes on an array can reduce array effects due to the quality of robot-fabricated immobilized cDNA probes within the same array. However, replicated spots should be well spaced so that the true variability within an array can be estimated. The yeast sporulation experiment used a replication of making a self-comparison of the time-0 sample. Although this was adequate for providing error degrees of freedom, it was not an ideal situation. All of the nonzero residuals from the ANOVA model come from the self comparison array and all other data points are exactly fit because they are not replicated (*29*).

Good experimental design will likely provide the greatest amount of satisfaction and the least amount of frustration in executing a microarray project. The reference design and loop design are two common experiment designs in microarray experiments. Kerr and Churchill (*29*) suggested a more effective loop design for yeast sporulation experiment. Fitting model (1) with this design, residuals are obtained from every array and GV effects can be estimated more precisely. In addition, dye swap is also a common design. Furthermore, the technique with more than two dyes has been proposed to decrease the experimental expenses (*31*). Our method can be easily applied to these designs and their modifications with replications. A common problem in microarray experiments is missing data. In microarray experiments, each array may contain a number of genes with fluorescence intensity measurements that are flagged by the experimenter and recorded as missing data. Due to noise and missing values in data sets, many statistic methods may result in estimates quite different from the real values. Mixed model approach has advantages of handling both unbalanced data and of predicting the random effects.

Acknowledgements

This research was partially supported by the National Natural Science Foundation of China (No. 30470916).

References

1. Cho, R.J., *et al.* 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* 2: 65-73.
2. Chu, S., *et al.* 1998. The transcriptional program of sporulation in budding yeast. *Science* 282: 699-705.
3. Eisen, M.B., *et al.* 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl.*

Acad. Sci. USA 95: 14863-14868.

4. Alizadeh, A.A., *et al.* 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403: 503-511.
5. Golub, T.R., *et al.* 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286: 531-537.
6. Pomeroy, S.L., *et al.* 2002. Prediction of central nervous system embryonal tumor outcome based on gene expression. *Nature* 415: 436-442.
7. Alon, U., *et al.* 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide array. *Proc. Natl. Acad. Sci. USA* 96: 6745-6750.
8. Tamayo, P., *et al.* 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* 96: 2907-2912.
9. Gasch, A.P. and Eisen, M.B. 2002. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol.* 3: research0059.
10. Tavazoie, S., *et al.* 1999. Systematic determination of genetic network architecture. *Nat. Genet.* 22: 281-285.
11. Ben-Dor, A. and Yakhini, Z. 1999. Clustering gene expression patterns. In *Proceedings of the Third Annual International Conference on Research in Computational Molecular Biology.* Lyon, France.
12. Hartuv, E. and Shamir, R. 2000. A clustering algorithm based on graph connectivity. *Inform. Process. Lett.* 76: 175-181.
13. Sharan, R., *et al.* 2003. CLICK and EXPANDER: a system for clustering and visualizing gene expression data. *Bioinformatics* 19: 1787-1799.
14. Schuchhardt, J., *et al.* 2000. Normalization strategies for cDNA microarrays. *Nucleic Acids Res.* 28: E47.
15. Lu, Y., *et al.* 2005. A two-step strategy for detecting differential gene expression in cDNA microcarray data. *Curr. Genet.* 47: 121-131.
16. Wolfinger, R.D., *et al.* 2001. Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.* 8: 625-637.
17. Kaufman, L. and Rousseeuw, P.J. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley & Sons, New York, USA.
18. Jain, A.K. and Dubes, R.C. 1988. *Algorithms for Clustering Data.* Prentice Hall, Englewood Cliffs, USA.
19. Yeung, K.Y., *et al.* 2001. Validating clustering for gene expression data. *Bioinformatics* 17: 309-318.
20. Rousseeuw, P.J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20: 53-65.
21. Zhu, J. 1993. Methods of predicting genotype value and heterosis for offspring of hybrids. *J. Biomath.* 8: 32-44.
22. Zhu, J. and Weir, B.S. 1996. Diallel analysis for sex-linked and maternal effects. *Theor. Appl. Genet.* 92: 1-9.
23. Spath, H. 1989. *Cluster Analysis Algorithm.* Ellis Horwood, Chichester, UK.
24. Churchill, G.A. 2002. Fundamentals of experimental design for cDNA microarrays. *Nat. Genet.* 32: 490-495.
25. Novak, J.P., *et al.* 2002. Characterization of variability in large-sacle gene expression data: implications for study design. *Genomics* 79: 104-113.
26. Kerr, M.K., *et al.* 2000. Analysis of variance for gene expression microarray data. *J. Comput. Biol.* 7: 819-837.
27. Searle, S.R., *et al.* 1992. *Variance Components.* John Wiley & Sons, New York, USA.
28. Henderson, C.R. 1963. Selection index and expected genetic advance. In *Statistical Genetics and Plant Breeding* (eds. Hanson, W.D. and Robinson, H.E.), pp. 141-163. National Academy of Sciences, Washington DC, USA.
29. Kerr, M.K., and Churchill, G.A. 2002. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc. Natl. Acad. Sci. USA* 98: 8961-8965.
30. Brazma, A., and Vilo, J. 2000. Gene expression data analysis. *FEBS Lett.* 480: 17-24.
31. Forster, T., *et al.* 2004. Triple-target microarray experiments: a novel experimental strategy. *BMC Genomics* 5: 13.