BIOINFORMATICS

Vol. 19 no. 4 2003, pages 459–466
DOI: 10.1093/bioinformatics/btg025

Comparisons and validation of statistical clustering techniques for microarray gene expression data

Susmita Datta [1,*] *and Somnath Datta* [2]

[1] *Department of Mathematics and Statistics and Department of Biology, Georgia State University, Atlanta, GA 30303, USA and* [2] *Department of Statistics, University of Georgia, Athens, GA 30602, USA*

Received on May 30, 2002; revised on October 18, 2002; accepted on October 21, 2002

ABSTRACT

Motivation: With the advent of microarray chip technology, large data sets are emerging containing the simultaneous expression levels of thousands of genes at various time points during a biological process. Biologists are attempting to group genes based on the temporal pattern of their expression levels. While the use of hierarchical clustering (UPGMA) with correlation 'distance' has been the most common in the microarray studies, there are many more choices of clustering algorithms in pattern recognition and statistics literature. At the moment there do not seem to be any clear-cut guidelines regarding the choice of a clustering algorithm to be used for grouping genes based on their expression profiles.

Results: In this paper, we consider six clustering algorithms (of various flavors!) and evaluate their performances on a well-known publicly available microarray data set on sporulation of budding yeast and on two simulated data sets. Among other things, we formulate three reasonable validation strategies that can be used with any clustering algorithm when temporal observations or replications are present. We evaluate each of these six clustering methods with these validation measures. While the 'best' method is dependent on the exact validation strategy and the number of clusters to be used, overall *Diana* appears to be a solid performer. Interestingly, the performance of correlation-based hierarchical clustering and model-based clustering (another method that has been advocated by a number of researchers) appear to be on opposite extremes, depending on what validation measure one employs. Next it is shown that the group means produced by *Diana* are the closest and those produced by UPGMA are the farthest from a model profile based on a set of hand-picked genes.

Availability: S+ codes for the partial least squares based clustering are available from the authors upon request. All other clustering methods considered have S+ implementation in the library MASS. S+ codes for calculating the validation measures are available from the authors upon request. The sporulation data set is publicly available at http://cmgm.stanford.edu/pbrown/sporulation.

Supplementary information: http://www.mathstat.gsu.edu/~matsnd/clustering/supp.htm

Contact: sdatta@mathstat.gsu.edu

INTRODUCTION

Motivation

One of the central goals in microarray or expression data analysis is to identify the changing and unchanging levels of gene expression and to correlate these changes to identify sets of genes with similar profiles. In some cases (DeRisi *et al.*, 1997; Chu *et al.*, 1998; Cho *et al.*, 1998), a mainly visual analysis has been successful in grouping genes into functionally relevant classes; however, this method is labor intensive, very subjective and may not be suitable in more complicated and large scale studies. In subsequent studies, simple sorting of expression ratios and some form of 'correlation distance' were used to identify genes (Spellman *et al.*, 1998; Roth *et al.*, 1998; Eisen *et al.*, 1998). Since the influential paper by Eisen *et al.* and the availability of corresponding free software, the gold standard in microarray studies has been to use hierarchical clustering (UPGMA) with correlation 'distance' (or dissimilarity). Waddell and Kishino (2000) recommended using partial correlations instead of correlations as measures of closeness. Here, we introduce a new dissimilarity based on partial least squares modeling (Datta, 2001). Model-based clustering is another technique that has recently been used for grouping microarray data (McLachlan *et al.*, 2002). This technique is based on modeling the expression profiles by mixtures of multivariate normal distributions.

The literature on statistical clustering is fairly vast,

[*] To whom correspondence should be addressed.

offering many other choices of clustering methods, notably partition methods such as *K-means*, divisive clustering method *Diana* and fuzzy logic based method *Fanny*. At the moment, there do not seem to exist any objective guidelines regarding the choice of a clustering algorithm to be used for grouping genes based on their expression profiles. In this paper, we have selected six clustering algorithms of various types and evaluated their performance on a well known publicly available microarray data set on sporulation of budding yeast, as well as on two simulated data sets which are introduced in the next section. Of course, one can extend and modify this list of competing clustering algorithms to include his/her favorite algorithm. At least five of these algorithms are chosen to represent different classes of methods. Thus, well known algorithms such as *Pam* and *Clara*, both of which fall under partition methods, are not included in favor of including the *K-means* algorithm. Also, another noteworthy, but more complex algorithm called SOM (self-organizing maps, Kohonen, 1997), whose implementation requires careful selection of various tuning parameters is not included here.

Related work

Kerr and Churchill (2001) used a linear model (ANOVA) and residual based resampling to access the reliability of clustering algorithms. Chen *et al.* (2002) compared the performances of a number of clustering algorithms by physical characteristics of the resulting clusters such as the homogeneity and separation. Yeung *et al.* (2001) introduces the concept of *Figure of Merit* (FOM) in selecting between competing clustering algorithms. FOM resembles the *Error sum of squares* (ESS) criterion of model selection.

Outline and summary

The **Algorithm and implementation** section introduces three different validation criteria, following a novel approach, each of which can be used for an objective basis of checking the consistency of the groupings produced by a clustering algorithm. These criteria can be used for any microarray data set that includes temporal observations. We assume that the biologists have a rough prior assessment of the number of clusters to be used. We compare the stability or consistency of the results produced by deleting one set of temporal observations at a time.

Often in microarray experiments, the biologists have prior ideas concerning the various groups of interesting expression patterns to be expected and a number of representative genes from each group. We compare, for each clustering method, the average expression patterns of all genes in each group with the model profiles. The paper ends with a discussion of our findings.

SYSTEMS AND METHODS

Sporulation data

We consider microarray data on the transcriptional program of sporulation in budding yeast collected and analyzed by Chu *et al.* (1998). The data set is publicly available at http://cmgm.stanford.edu/pbrown/sporulation. They used DNA microarrays containing 97% of the known and predicted genes involved, 6118 in total. The mRNA levels were measured at seven time points during the sporulation process. Further details can be found in the article by Chu *et al.*

The ratio of each gene's mRNA level (expression) to its mRNA level in vegetative cells just before transfer to sporulation medium is measured, and the ratio data are then log-transformed. 1143 genes whose expression levels did not change 'significantly' during the sporulation processes are deemed uninteresting and were dropped from further analysis. Chu *et al.* (1998) determined significance by using a threshold level of 1.13 for the root mean squares of the $\log_2$-transformed ratios. Overall, 513 genes were (positively) expressed during the process[†].

Simulated data

We consider two simulated data sets built around the same nine distinct temporal patterns over ten time points as used by Quackenbush (2001). The data values are then generated by adding independent random noises at each of these mean expression-ratio values. Overall $10 \times 9 \times 50 = 4500$ random variates were generated to create the expression profiles (i.e. $\log_2$-transformed ratios) of 50 genes around each of the nine temporal patterns (for a total sample size of 450 genes). In the first simulated data set, normal variates with mean zero and standard deviation 0.2 were used; in the second data set half of them were normal with zero mean and standard deviation 0.4, and the remaining half were generated from an exponential distribution with location -0.2 and scale 0.2.

Clustering methods

The following clustering techniques were considered. S+ implementation of all these techniques (with the exception of partial least squares) are available in the library MASS and are described in Venables and Ripley (1998).

(i) *Hierarchical clustering with correlation:* This algorithm produces a hierarchy of clusters rather than a set number of clusters fixed in advance. At the base or initial level, each observation forms its own cluster. At each subsequent level, the two 'nearest' clusters are combined to form one bigger cluster. We use *method* = '*average*' which means the 'distance' between clusters is the average

[†] A gene is considered to be expressed during the process if $\sum \log R > 0$, where the sum is over all time points.

of the ‘distances’ between the points in one cluster and the points in the other cluster. This is perhaps the most common and the simplest tree method and is popularly known as UPGMA (Unweighted Pair Group Method with Arithmetic Mean). The ‘distance’ between genes x and y were taken to be $d(x, y) = 1 - |\operatorname{corr}(x, y)|$, where $\operatorname{corr}(x, y)$ is the statistical correlation between the expression profiles of x and y.

(ii) *Clustering by* K-means*:* Under this scheme, one needs to fix the number of clusters in advance. Usually another clustering algorithm, such as the one above, is run to determine the initial cluster centers to be used in the *K-means* algorithm. The algorithm then assigns the observations into various clusters in order to minimize the total within-class sum of squares. A complex iterative numerical algorithm (see Hartigan and Wong, 1979) is used to find this minimum (or rather a local minimum).

(iii) *Diana:* This is a divisive clustering method where initially all the observations are clustered together. Subsequently, the bigger groups are broken down into smaller groups so that genes with larger distance or dissimilarity are placed in different clusters. Suppose at a given stage a big cluster C with cardinality $n(C)$ needs to be split into two. For each member $x_1 \in C$, one computes the ‘distance’ from the rest of C by

$$\Delta_{1,x_1} = (n(C) - 1)^{-1} \sum_{y \in C \backslash \{x_1\}} d(x_1, y)$$

and identifies the gene x_1^*, say, for which Δ_{1,x_1} is the largest. Continue the procedure to identify genes $x_1^*, \ldots, x_{k+1}^*$ iteratively till $\Delta_{k+1}, x_{k+1}^* < 0$ where

$$\Delta_{k+1,x_{k+1}} = (n(C) - k - 1)^{-1} \sum_{y \in C \backslash \{x_1^*, \ldots, x_k^*\}} d(x_{k+1}, y)$$
$$-k^{-1} \sum_{i=1}^{k} d(x_i^*, x_{k+1}),$$

and x_{k+1}^* maximizes $\Delta_{k+1,x_{k+1}}$. Then the split will be $\{x_1^*, \ldots, x_k^*\}$ and $C \backslash \{x_1^*, \ldots, x_k^*\}$. See Kaufman and Rousseeuw (1990) for further details and examples. We used the standard Euclidean distance for $d(x, y)$.

(iv) *Fanny:* This method uses fuzzy logic and produces a probability vector for each observation. A hard cluster is determined by assigning an observation to a group which has the highest probability. Like distance-based methods, one has a choice of using a general *dissimilarity* measure. We have used the L_1 distance (also known as Manhattan distance) which is more robust than the Euclidean distance.

Letting K denote the total number of desired clusters, *Fanny* computes the probability vectors (called membership coefficients; $u_{x1}, \ldots, u_{xK}$) for all genes x that minimize the objective function

$$\sum_{k=1}^{K} \frac{\sum_{x,y} u_{xk}^2 u_{yk}^2 d(x, y)}{\sum_x u_{xk}^2}.$$

Hard clusters are then produced, if needed, by assigning genes to the group with the highest probability. See Kaufman and Rousseeuw (1990) for further details. Typically, relatively fewer hard clusters are produced by this method.

(v) *Model-based clustering:* The idea behind model based clustering is to regard the data as coming from a mixture distribution. Suppose for the ith observation γ_i gives the true, but unknown, group level for that observation. Then letting $f_j(.; \theta_j)$ denote the density function for a typical observation from group j, where θ_j denotes an unknown parameter, the resulting likelihood of genes with expression profiles $x_1, \ldots, x_n$ is given by

$$L(\gamma, \theta) = \prod_{i=1}^{n} f_{\gamma_i}(x_i, \theta_{\gamma_i}).$$

The unknown group levels γ are obtained by method of maximum likelihood that maximizes L jointly in γ and θ. We use the S+ procedure *mclust* with the default option that allows for multivariate normals with different centers and orientations, but of constant pre-specified shape (the ratio of the axes of the ellipsoid). See Banfield and Raftery (1993) for further details.

(vi) Hierarchical clustering with partial least squares: The usefulness of partial least squares in identifying gene relationships through their expression profiles has recently been demonstrated by Datta (2001). For a more detailed account on partial least squares the reader may consult Stone and Brooks (1990) and Brown (1993).

In the case of microarray data, x_i will be the vector of expression ratios (log-transformed, and normalized) for the ith gene (or ORF). Let us suppose, we fit a partial least squared model of x_1 on $x_2, \ldots, x_M$, of the form

$$x_i = \sum_{l=1}^{p} \widehat{\beta}_{il} t_i^{(l)},$$

where p will typically be a small integer (much smaller than M; we used $p = 2$), $t_i^{(l)} = \sum_{k \neq i}^{M} c_{ik}^{(l)} x_k$, $\{c_{ik}^{(l)}\}$ are defined in a special way through the x. Then for any gene pair (i, j), $i \neq j$, the symmetrized coefficient

$$s_{ij} = [\widehat{\beta}_{i1} c_{ij}^{(1)} + \widehat{\beta}_{i2} c_{ij}^{(2)} + \cdots + \widehat{\beta}_{ip} c_{ij}^{(p)} + \widehat{\beta}_{j1} c_{ji}^{(1)}$$
$$+ \widehat{\beta}_{j2} c_{ji}^{(2)} + \cdots + \widehat{\beta}_{ji} c_{ji}^{(p)}]/[2].$$

represents the closeness or prediction power one gene has towards the expression level of the other gene. s was further normalized by the maximum value of observed s, and s_{xx} was taken to be 1 for all x. We use hierarchical clustering with $method$ = '*average*' and this similarity measure s.

ALGORITHM AND IMPLEMENTATION

First we implement each of these clustering techniques using S+ for the sporulation data. Chu *et al.* (1998) advocated grouping the expressed genes into seven temporal classes on biological grounds. Following Chu *et al.* (1998), the number of clusters was set to seven in each case. As expected, there are some differences in the results of the various algorithms. Overall, *K-means* and *Diana* seem to be most effective in achieving good separation and almost distinct class boundaries. One potential problem with *Fanny* is that it typically produces only few distinct hard clusters. For this data set, only three clusters were produced, even though seven were desired. Further details, including pictures, can be obtained from the supplementary website.

Validation

Even before the recent surge of microarray data there have been a number of clustering techniques in the existing statistical/pattern-recognition literature. With the growing availability of more and more microarray data sets, newer algorithms are being proposed. This may pose a potential problem for a practitioner, since, at the moment, there do not seem to be any suitable guidelines regarding the choice of a clustering algorithm to be used for grouping genes based on their expression profiles. The problem is particularly difficult, since no single algorithm is expected to be the winner in every case.

Let K be the number of classes we set a clustering algorithm to produce. As in the case of the sporulation data, often times the biologists have some prior ideas of what a good choice of K is, at least approximately. We assume this to be the case for our strategy. However, we suggest that the performance of an algorithm be investigated over an entire range of nearby or usable K values.

The idea behind the validation approach is that an algorithm should be rewarded for consistency. We are envisioning a setup where expression (ratio) data are collected over all the genes under study at various time points say $T_1, T_2, \ldots, T_l$. In the case of the sporulation data, K was around 7 (Chu *et al.* used $K = 7$) and $l = 7$. Thus our data values are points in the l dimensional Euclidean space $\Re^l$. For each $i = 1, 2, \ldots, l$, repeat the clustering algorithms for each of the l data set in $\Re^{l-1}$ obtained by deleting the observations at time T_i. For each gene $1 \leqslant g \leqslant M$ let $C^{g,i}$ denote the cluster containing gene g in the clustering based on the data set with time T_i observations deleted. Let $C^{g,0}$ be the cluster in the original data containing gene g. Each of the following validation measures seems to be a reasonable choice. For a good clustering algorithm, we would expect these values to be small.

(I) The average proportion of non-overlap measure

$$V_1(K) = \frac{1}{Ml}\sum_{g=1}^{M}\sum_{i=1}^{l}\left(1 - \frac{n(C^{g,i} \cap C^{g,0})}{n(C^{g,0})}\right).$$

This measure computes the (average) proportion of genes that are not put in the same cluster by the clustering method under consideration on the basis of the full data and the data obtained by deleting the expression levels at one time point at a time.

(II) The average distance between means measure

$$V_2(K) = \frac{1}{Ml}\sum_{g=1}^{M}\sum_{i=1}^{l} d\left(\overline{x}_{C^{g,i}}, \overline{x}_{C^{g,0}}\right),$$

where $\overline{x}_{C^{g,0}}$ denotes the average expression profile for genes across cluster $C^{g,0}$ and $\overline{x}_{C^{g,i}}$ denotes the average expression profile for genes across cluster $C^{g,i}$. This measure computes the (average) distance between the mean expression ratios (log transformed) of all genes that are put in the same cluster by the clustering method under consideration on the basis of the full data and the data obtained by deleting the expression levels at one time point at a time.

(III) The average distance measure

$$V_3(K) = \frac{1}{Ml}\sum_{g=1}^{M}\sum_{i=1}^{l}\frac{1}{n(C^{g,0})n(C^{g,i})} \times \sum_{g\in C^{g,0}, g'\in C^{g,i}} d(x_g, x_{g'}),$$

where $d(x_g, x_{g'})$ is a distance (e.g. Euclidean, Manhattan, etc.) between the expression profiles of genes g and g'. This measure computes the average distance between the expression levels of all genes that are put in the same cluster by the clustering method under consideration on the basis of the full data and the data obtained by deleting the expression levels at one time point at a time.

Results for sporulation data For each of the six clustering algorithms under consideration, we compute the three validation measures over a range of K values around seven (4–12). The results are displayed in Figures 1, 2 and 3, respectively. In each plot, a profile closer to the horizontal axis indicates better performance over the usable range of K values.

The *Average Proportion of Non-overlap Measure* and the *Average Distance Between Means Measure* produce

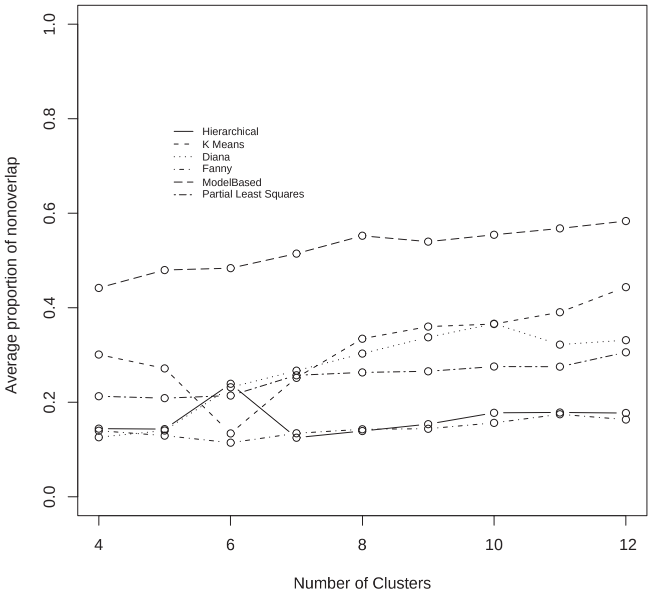

Fig. 1. The average proportion of non-overlap measure for various clustering algorithms applied to the sporulation data.

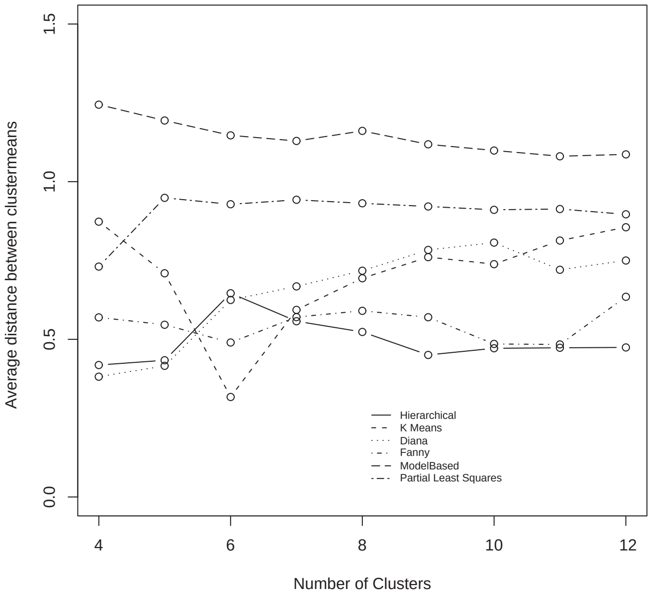

Fig. 2. The average distance between means measure for various clustering algorithms applied to the sporulation data.

similar results. A somewhat surprising finding is that the performance of model-based clustering appears to be the worst as judged by these measures. The standard hierarchical clustering (with absolute correlation dissimilarity and method = average) and *Fanny* appear to be the best as judged by these measures. As mentioned earlier, *Fanny* has its own problem in the sense that the true number of hard clusters it would produce may be smaller than K. On the other hand, according to the *Average Distance Measure*, performance of hierarchical clustering seems to be the worst. Overall, *Diana* appears to be a solid and robust performer under all three validation measures.

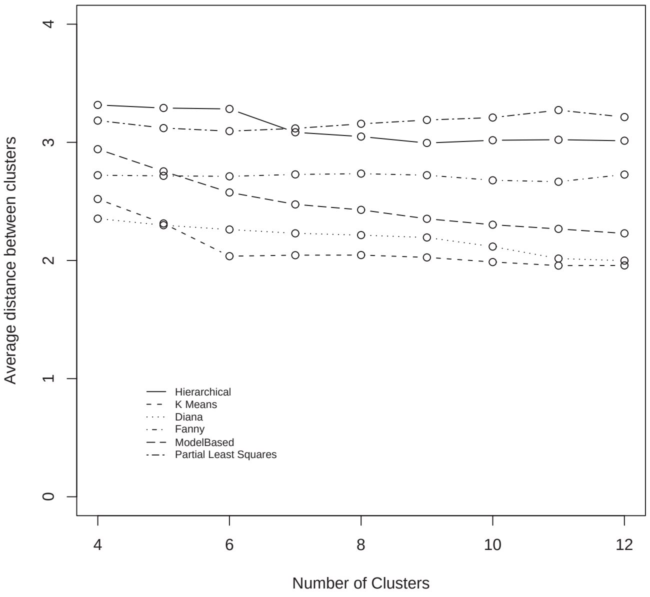

Fig. 3. The average distance measure for various clustering algorithms applied to the sporulation data.

Results for simulated data For the sake of brevity we would only discuss the performance of the first validation measure. Hierarchical clustering (UPGMA) with both correlation-based and partial least squares-based similarity measures perform poorly. There is a simple reason why correlation-based and partial least squares-based similarity measures are not appropriate here. These measures are invariant under location and scale transformations, and, thus, they cannot distinguish between the patterns that are related by location and/or scale changes. On the other hand, hierarchical clustering with Euclidean distance is able to distinguish between the target patterns (results not shown). *K-means*, *Diana*, *Fanny* and *Model based* all performed quite well with *Model based* and *Diana* being slightly better than *K-means* and *Fanny*. Interestingly, all three have a local minima at nine for the first data set. The plots are available on the supplementary website.

Comparisons with model profiles

Sporulation data Chu *et al.* (1998) described a small training set of hand-picked genes from each of the seven temporal classes that were expressed during sporulation.

Table 1. Total distance (*dist*) from the model profile

Hierarchical	*K-Means*	*Diana*	Model based	Partial least squares
17.58	12.48	10.47	11.80	14.55

We use essentially[‡] the same training set to construct our model temporal profile, which can serve as a benchmark for the end result produced by seven clusters using each of these clustering methods.

For each class, the average of the log-expression ratio of all the genes in that class are plotted over the seven time points during sporulation. The resulting curves are termed model temporal profiles and are given in the last plot in Figure 4. Next, we run the five clustering algorithms (all but *Fanny*[§]) for each set to generate $K = 7$ classes. For each algorithm, the average of the log-expression ratio of all the genes in each of the seven classes produced by that algorithm are plotted over the seven time points during sporulation. All these plots together with the model profiles constitute Figure 4. An overall visual comparison of these plots with the model profiles show that the *Diana* plots are perhaps the closest to the model profiles.

An objective way to measure the closeness of any of these plots to the model profiles will be to compute the following distance measure

$$\text{dist} = \min_\pi \sum_{i=1}^{K} d(\overline{x}_i^m, \overline{x}_{\pi(i)}),$$

where the minimum is taken over all permutation π of the integers $\{1, \ldots, K\}$, $\overline{x}_i^m$ is the value of the (average) model profile for the ith cluster, and d is the Euclidean (or any other) distance between points in $\Re^K$. The reason for using various permutations π is to find the 'best' match between the two sets of clusters. These values for the five methods are given in Table 1.

This confirms our visual impression that *Diana* group means come closest to the model profile based on the hand-picked genes. Hierarchical seems to be the farthest from the model profile.

Simulated data We set the number of clusters to nine in each of the six algorithms and plotted the average temporal profiles in each group (supplementary website). A comparison with the target model profile shows that all four of *K-means*, *Diana*, *Fanny* and Model based performed extremely well for data set 1, but, for data set 2, *K-means* was unable to pick at least one pattern. Hierarchical clustering with correlation or partial least squares based similarity were unable to reproduce the target patterns (for the same reasons as explained before).

Further analysis of sporulation data using *Diana*

We have seen that *Diana* has been a consistent performer, as evidenced by the analysis done so far. We provide a brief re-analysis of the sporulation data using this particular clustering algorithm. Of course, the groups represented by a clustering algorithm are unlikely to be linked only through their time of first induction (which was used to define the biological groups in (Chu *et al.*, 1998). However, the labels seem to be identifiable from the group average profiles which match the model profiles surprisingly well (Figure 4).

The number of genes clustered into the metabolic, early 1, early 2, early–middle, middle, mid-late and late induction groups were 29, 34, 100, 232, 73, 38 and 7, respectively. Overall, fewer genes were declared to be metabolic, early 1 and mid-late. Many more genes were identified as early II and, especially, early-mid than previously reported (Chu *et al.*, 1998). For example, the metabolic group (cluster 1) contained known metabolic genes such as ASC1 and PYC1, whereas genes such as SIP4 and CAT2 that were placed in the metabolic group in the model profile were actually grouped in cluster 5 by *Diana*. A closer inspection shows that, though the later two genes did induce rapidly, their expression is less transient than what was typical of a metabolic gene. Also *Diana* found seven genes that were induced late. Six of these genes had the typical expression profile of a late gene. The other gene (ORF=YJL015C) clustered into this group by *Diana* was first induced seven hours after transfer to the sporulation medium. At time nine hours its expression diminished, and it was re-expressed at the last time point (time 11.5) samples were taken.

DISCUSSION

Cluster analysis programs are routinely run as a first step of data summary and grouping genes in a microarray data analysis. As we show here with the sporulation data on budding yeast of Chu *et al.* (1998), the end result is very much dependent on what clustering method is employed. This is particularly disturbing since, at the moment, there do not seem to exist clear guidelines in this regard. While the standard hierarchical clustering (UPGMA) with correlation (or absolute correlation) similarity seems to be the most commonly employed clustering technique in microarray data analysis, its optimality or superiority over other methods have not been demonstrated in microarray literature.

This paper offers some guidelines in the choice of a clustering technique to be used in connection with a particular microarray data set. The first and the obvious

[‡] KNR4 was dropped from the list since it was not significantly expressed. Also the gene PDS1 appears to have been incorrectly listed as Early-Mid.
[§] *Fanny* was dropped because it was unable to produce seven hard clusters.

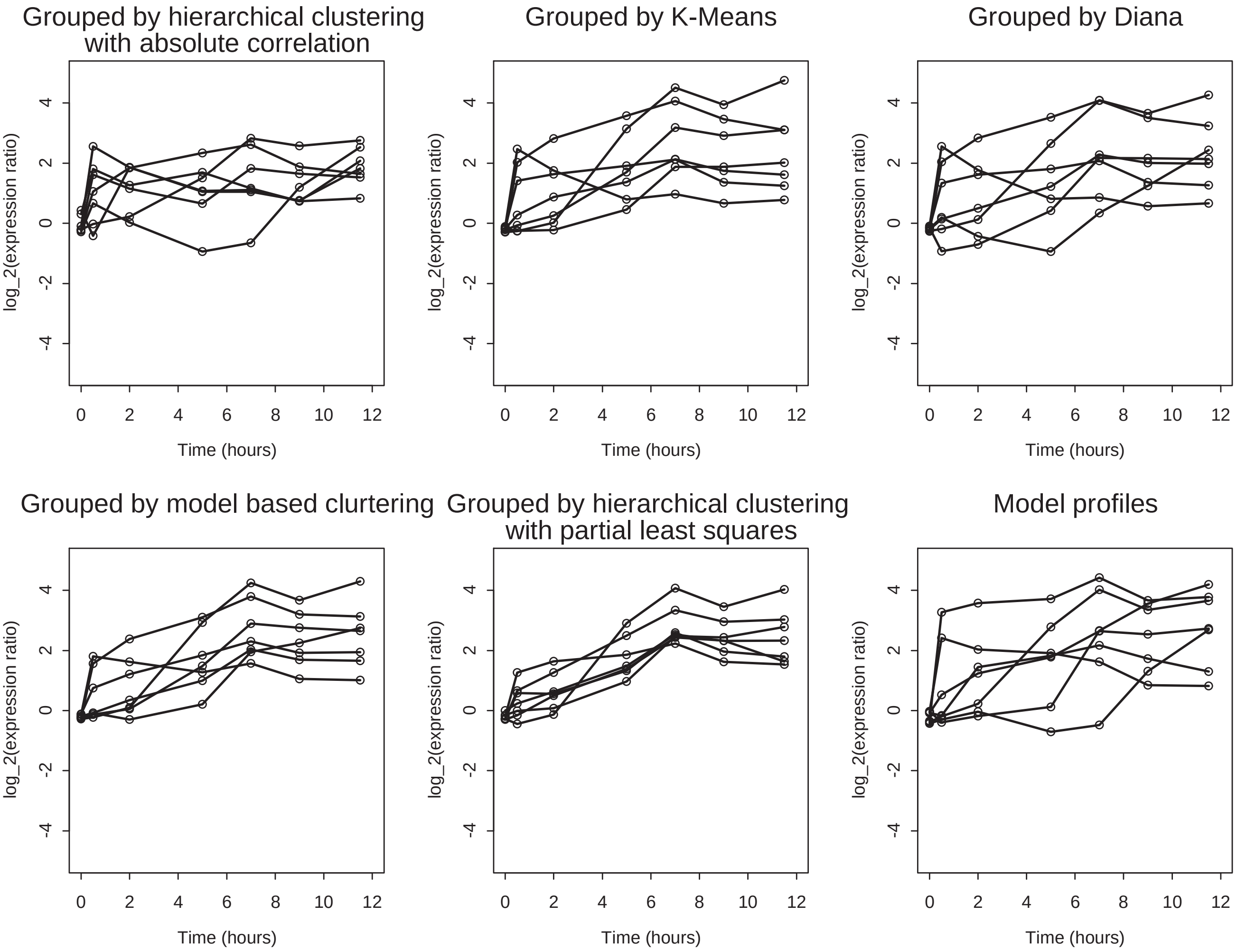

Fig. 4. Average temporal profiles of seven groups obtained using various clustering algorithms and the hand-picked genes for the sporulation data.

one is a visual plot in the space of the first two principal components to see which method offers the most separation among the various groups. A more objective method that checks for consistency is a novel validation technique that is demonstrated with the temporal observations in the sporulation data and in two sets of simulated data. An adaptation of this method for data sets with full replicates should not be difficult to formulate. Three different validation measures, each of which has some appeal, have been proposed with respect to this validation scheme. The third method we advocate is to compare the average group temporal profiles with the model profiles constructed from a set of hand-picked genes (such as a training set) whose group memberships are relatively well understood. We propose a distance that can be computed for each method, which, in turn, can be used to order their performance. We have also examined this method applied to two simulated data sets by comparing the average temporal profile with the known target patterns used in generating the data. Though relatively unknown in the microarray literature, we conclude that the divisive clustering technique *Diana* is the overall 'winner', amongst those we have considered, for the data sets we have examined. Perhaps it should receive more attention in future microarray analyses.

Other important aspects of clustering techniques, such as computational stability and computational time, have not been emphasized in this paper. There is a sizable literature dealing with these issues and other relative advantages and disadvantages of most of these standard methods (see for example, Hartigan, 1975, Kaufman and Rousseeuw, 1990). Correlation and partial least squares based similarity measures used in conjunction with hierarchical clustering may not adequately locate all the features in an expression profile. We should point out, however, that we have considered only the most common type of hierarchical clustering, namely UPGMA. The class of hierarchical clustering methods is certainly very broad and the resulting trees produced by different hierarchical methods could look very different (Waddell and Kishino, 2000). In certain applications, hierarchical clustering is preferable as an exploratory tool since it does not require a pre-specification of the number of clusters. Instead, the resulting dendrogram (or tree) can be inspected at various 'heights'. Divisive hierarchical clustering methods have received less attention, partly because optimal division requires higher computing time. *Diana* avoids this problem by using an intelligent splitting strategy. If a specific number of hard clusters is desired, *Fanny* may not be a suitable algorithm as we have seen for the sporulation data. *K-means* is a popular algorithm that uses a reasonable objective criterion. However, it could be sensitive to the choice of the initial cluster centers. For the simulated data sets, we have observed that the *K-means* algorithm failed when the initial centers were taken to be the cluster means obtained using UPGMA, and, subsequently, we had used the cluster means from *Diana* for it to work. Model-based clustering has several modeling and optimization options, and it worked reasonably well in the simulated data.

Optimal selection of the number of clusters K is a difficult problem and perhaps should be addressed elsewhere in its own right. There are limited statistical tools for this purpose that are applicable to all clustering algorithms. Variants of the model selection criteria such as the BIC are available for certain likelihood based algorithms (e.g. model-based clustering). Perhaps clustering, when used as an exploratory tool, should be carried out at a number of plausible K values, and existing scientific knowledge about the problem be combined with the resulting clustering outputs. Echoing the same sentiment, we suggest examining the graphs of the validation measures over an entire range of K values around the assumed number of clusters. For the sporulation data set we ultimately used the same number of clusters as in Chu *et al.* (1998), in order to keep the analysis compatible and for the use of the model profile which was constructed on biological grounds.

ACKNOWLEDGEMENTS

Susmita Datta's research was supported in part by US National Science Foundation grant DBI-0074642. We are grateful to Missy Cody and Jaxk Reeves for their careful readings of the manuscript. We thank the referees for many constructive comments, including a suggestion of using the simulated data sets, which improved the manuscript substantially.

REFERENCES

Banfield,J.D. and Raftery,A.E. (1993) Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**, 803–822.

Brown,P.J. (1993) *Measurement, Regression, and Calibration*. Oxford University Press, New York.

Chen,G. *et al.*, (2002) Evaluation and comparison of clustering algorithms in analyzing ES cell gene expression data. *Statistica Sinica*, **12**, 241–262.

Cho,R.J. *et al.*, (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell.*, **2**, 65–73.

Chu,S., DeRisi,J. *et al.*, (1998) The transcriptional program of sporulation in budding yeast. *Science*, **282**, 699–705.

Datta,S. (2001) Exploring relationships in gene expressions: a partial least squares approach. *Gene Expression*, **9**, 257–264.

DeRisi,J.L., Iyer,V.R. and Brown,P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Hartigan,J.A. (1975) *Clustering Algorithms*. Wiley, New York.

Hartigan,J.A. and Wong,M.A. (1979) A *k*-means clustering algorithm. *Applied Statistics*, **28**, 100–108.

Kaufman,L. and Rousseeuw,P.J. (1990) *Fitting Groups in Data. An Introduction to Cluster Analysis*. Wiley, New York.

Kerr,M.K. and Churchill,G.A. (2001) Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc. Natl Acad. Sci. USA*, **98**, 8961–8965.

Kohonen,T. (1997) *Self-Organizing Maps*, 2nd edn, Springer, Berlin.

McLachlan,G.J., Bean,R.W. and Peel,D. (2002) A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**, 1–10.

Quackenbush,J. (2001) Computational analysis of microarray data. *Nat. Rev. Genet.*, **2**, 418–427.

Roth,F.P., Hughes,J.D., Estep,P.W. and Church,G.M. (1998) Finding DNA regulatory motifs within unaligned non-coding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, **16**, 939–945.

Spellman,P.T. *et al.*, (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell.*, **12**, 3273–3297.

Stone,M. and Brooks,R.J. (1990) Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal component regression. *J. Roy. Statist. Soc. B, 52, 237–269, Corrigendum (1992)*, **54**, 906–907.

Venables,W.N. and Ripley,B.D. (1998) *Modern Applied Statistics with S-Plus, (3rd corrected printing)*, 2nd edn, Springer, New York.

Waddell,P. and Kishino,H. (2000) Cluster inference methods and graphical models evaluated on NC160 microarray gene expression data. *Genome Informatics*, **11**, 129–140.

Yeung,K., Haynor,D.R. and Ruzzo,W.L. (2001) Validating clustering for gene expression data. *Bioinformatics*, **17**, 309–318.