BIOINFORMATICS

Vol. 18 no. 9 2002
Pages 1194–1206

Bayesian infinite mixture model based clustering of gene expression profiles

Mario Medvedovic[1,*] *and Siva Sivaganesan*[2]

[1]Center for Genome Information, Department of Environmental Health, University of Cincinnati Medical Center, 3223 Eden Av. ML 56, Cincinnati, OH 45267-0056, USA and [2]Mathematical Sciences Department, University of Cincinnati, 4221 Fox Hollow Dr, Cincinnati, OH 45241, USA

Received on November 9, 2001; revised on January 22, 2002; March 4, 2002; accepted on March 12, 2002

ABSTRACT

Motivation: The biologic significance of results obtained through cluster analyses of gene expression data generated in microarray experiments have been demonstrated in many studies. In this article we focus on the development of a clustering procedure based on the concept of Bayesian model-averaging and a precise statistical model of expression data.

Results: We developed a clustering procedure based on the Bayesian infinite mixture model and applied it to clustering gene expression profiles. Clusters of genes with similar expression patterns are identified from the posterior distribution of clusterings defined implicitly by the stochastic data-generation model. The posterior distribution of clusterings is estimated by a Gibbs sampler. We summarized the posterior distribution of clusterings by calculating posterior pairwise probabilities of co-expression and used the complete linkage principle to create clusters. This approach has several advantages over usual clustering procedures. The analysis allows for incorporation of a reasonable probabilistic model for generating data. The method does not require specifying the number of clusters and resulting optimal clustering is obtained by averaging over models with all possible numbers of clusters. Expression profiles that are not similar to any other profile are automatically detected, the method incorporates experimental replicates, and it can be extended to accommodate missing data. This approach represents a qualitative shift in the model-based cluster analysis of expression data because it allows for incorporation of uncertainties involved in the model selection in the final assessment of confidence in similarities of expression profiles. We also demonstrated the importance of incorporating the information on experimental variability into the clustering model.

Availability: The MS Windows$^{\text{TM}}$ based program implementing the Gibbs sampler and supplemental material is available at http://homepages.uc.edu/∼medvedm/BioinformaticsSupplement.htm

Contact: medvedm@email.uc.edu

INTRODUCTION

The ability of the DNA microarray technology to produce expression data on a large number of genes in a parallel fashion has resulted in new approaches to identifying individual genes as well as whole pathways involved in performing different biologic functions. One commonly used approach in making conclusions from microarray data is to identify groups of genes with similar expression patterns across different experimental conditions through a cluster analysis (D'haeseleer *et al.*, 2000). The biologic significance of results of such analyses has been demonstrated in numerous studies. Various traditional clustering procedures, ranging from simple agglomerative hierarchical methods (Eisen *et al.*, 1998) to optimization-based global procedures (Tavazoie *et al.*, 1999) and self organizing maps (Tamayo *et al.*, 1999), to mention a few, have been applied to clustering gene expression profiles (in this paper we use the term profile to represent the set of observed expression values for a gene across several experiments). In identifying patterns of expression, such procedures depend on either a visual identification of patterns (hierarchical clustering) in a color-coded display or on the correct specification of the number of patterns present in data prior to the analysis (k-means and self organizing maps).

Expression data generated by DNA arrays incorporates different sources of variability present in the process of obtaining fluorescence intensity measurements. Choosing a statistically optimal experimental design and the appropriate statistical analysis of produced data is increasingly important aspect of such experiments (Kerr and Churchill, 2001; Baldi and Long, 2001). In the situation when statistically significant differential expression needs to be detected, various modifications of traditional statistical

[*]To whom correspondence should be addressed.

approaches have been proposed (Kerr *et al.*, 2000; Newton *et al.*, 2001; Ideker *et al.*, 2000; Wolfinger *et al.*, 2001; Rocke and Dubin, 2001; Lonnstedt and Speed, 2001). A common goal of all these approaches is the precise treatment of different sources of variability resulting in precise estimates of differential expression and their statistical significance. In this article we suggest Bayesian hierarchical infinite mixture model as a general framework for incorporating the same level of precision in the cluster analysis of gene expression profiles.

Generally, in a model-based approach to clustering, the probability distribution of observed data is approximated by a statistical model. Parameters in such a model define clusters of similar observations. The cluster analysis is performed by estimating these parameters from the data. In a Gaussian mixture model approach, similar individual profiles are assumed to have been generated by the common underlying 'pattern' represented by a multivariate Gaussian random variable. The confidence in obtained patterns and the confidence in individual assignments to particular clusters are assessed by estimating the confidence in corresponding parameter estimates. Recently, a Gaussian finite mixture model (McLachlan and Basford, 1987) based approach to clustering has been used to cluster expression profiles (Yeung *et al.*, 2001a). Assuming that the number of mixture components is correctly specified, this approach offers reliable estimates of confidence in assigning individual profiles to particular clusters. In the situation where the number of clusters is not known, this approach relies on ones ability to identify the correct number of mixture components generating the data. All conclusions are then made assuming that the correct number of cluster is known. Consequently, estimates of the model parameters do not take into account the uncertainty in choosing the correct number of clusters.

We developed a statistical procedure based on the Bayesian infinite mixture (also called the Dirichlet process mixture) model (Ferguson, 1973; Neal, 2000; Rasmussen, 2000) in which conclusions about the probability that a set of gene expression profiles is generated by the same pattern are based on the posterior probability distribution of clusterings given the data. In addition to generating groups of similar expression profiles at a specified confidence, it identifies outlying expression profiles that are not similar to any other profile in the data set. In contrast to the finite mixture approach, this model does not require specifying the number of mixture components, and the obtained clustering is a result of averaging over all possible numbers of mixture components. Consequently, the uncertainty about the true number of components is incorporated in the uncertainties about the obtained clusters. Advantages of such an approach over the finite mixture approach that relies on the specification of the 'correct' number of mixture components are demonstrated by a direct comparison of the two approaches. We also extended the model to incorporate experimental replicates. In a simulation study, we demonstrated that such an extension is beneficial when the expression levels of different genes are measured with different precision.

STATISTICAL MODEL

Suppose that T gene expression profiles were observed across M experimental conditions. If x_{im} is the expression level of the ith gene for the mth experimental condition, then $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iM})$ denotes the complete expression profile for the ith gene. Each gene expression profile can be viewed as being generated by one out of Q different underlying expression patterns. Expression profiles generated by the same pattern form a cluster of similar expression profiles. If c_i is the classification variable indicating the pattern that generates the ith expression profile ($c_i = j$ means that the ith expression profile was generated by the jth pattern), then a 'clustering' is defined by a set of classification variables for all expression profiles $\boldsymbol{c} = (c_1, c_2, \ldots, c_T)$. Values of classification variables are meaningful only to the extend that all observed expression profiles having the same value for their classification variable were generated by the same pattern and form a cluster. In our probabilistic model, underlying patterns generating clusters of expression profiles are represented by multivariate Gaussian random variables. That is, profiles clustering together are assumed to be a random sample from the same multivariate Gaussian random variable.

The following hierarchical model defines the probability distribution generating observed expression profiles. This model implicitly defines the posterior distribution of the classification set $\boldsymbol{c}$ and consequently the number of clusters (patterns) in the data, Q.

Level 1: The distribution of the data given the parameters of underlying patterns $[(\boldsymbol{\mu}_1, \sigma_1^2), \ldots, (\boldsymbol{\mu}_Q, \sigma_Q^2)]$, and the classification variables $\boldsymbol{c}$ is given by:

$$p(\boldsymbol{x}_i | c_i = j, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_Q, \sigma_i, \ldots, \sigma_Q) = f_N(\boldsymbol{x}_j | \boldsymbol{\mu}_j, \sigma_j^2 \boldsymbol{I}) \tag{I}$$

where $p(x|\boldsymbol{\theta})$ denotes the marginal probability distribution function of x given parameters $\boldsymbol{\theta}$; and $f_N(.|\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ is the probability density function of a multivariate Gaussian random variable with mean $\boldsymbol{\mu}$ and variance–covariance matrix $\sigma^2\boldsymbol{I}$ and $\boldsymbol{I}$ denotes the identity matrix.

Level 2: Prior distributions for $[(\boldsymbol{\mu}_1, \sigma_1^2), \ldots, (\boldsymbol{\mu}_Q, \sigma_Q^2)]$ and $(c_1, \ldots, c_T)$, given hyperparameters $\boldsymbol{\lambda}$, r, β and w are given by:

$$p(c_i | \pi_1, \ldots, \pi_Q) = \prod_{j=1}^{Q} \pi_j^{\boldsymbol{I}(c_i = j)}$$

$$p(\boldsymbol{\mu}_j | \boldsymbol{\lambda}, r) = f_N(\boldsymbol{\mu}_j | \boldsymbol{\lambda}, r^{-1}\boldsymbol{I})$$

$$p(\sigma_j^{-2}|\beta, w) = f_G\left(\sigma_j^{-2}\middle|\frac{\beta}{2}, \frac{\beta w}{2}\right) \qquad \text{(II)}$$

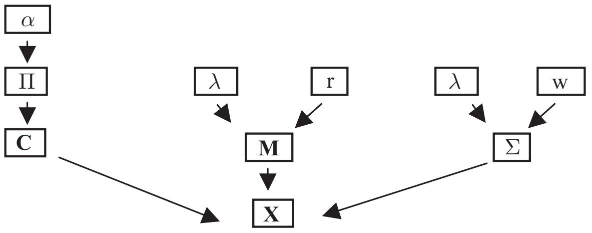
where $f_G(.|\theta, \tau)$ is the probability density function of a Gamma random variable with the shape parameter $\boldsymbol{\theta}$ and the scale parameter τ and $\boldsymbol{I}(c_i = j) = 1$ whenever $c_i = j$ and it is zero otherwise.

Level 3: Prior distributions for $(\pi_1, \ldots, \pi_Q)$ and hyperparameters $\boldsymbol{\lambda}$, r, β and w are given by:

$$\begin{aligned} p(\pi_1, \ldots, \pi_Q|\alpha, Q) &= f_D\left(\pi_1, \ldots, \pi_Q\middle|\frac{\alpha}{Q}, \ldots, \frac{\alpha}{Q}\right) \\ p(w|\sigma_x^2) &= f_G\left(w\middle|\frac{1}{2}, \frac{\sigma_x^{-2}}{2}\right) \\ p(\beta) &= f_G\left(\beta\middle|\frac{1}{2}, \frac{1}{2}\right) \qquad \text{(III)} \\ p(r|\sigma_x^2) &= f_G\left(r\middle|\frac{1}{2}, \frac{\sigma_x^2}{2}\right) \\ p(\lambda|\mu_x, \sigma_x^2) &= f_N(\lambda|\mu_x, \sigma_x^2\boldsymbol{I}) \qquad \text{(IV)} \end{aligned}$$

where $f_D(\cdot|\theta_1, \ldots, \theta_Q)$ represents the probability density function of a Dirichlet random variable with parameters $(\theta_1, \ldots, \theta_Q)$. $\boldsymbol{\mu}_x$ is the average of all profiles analyzed and σ_x^2 is the average of gene-specific sample variances based on all profiles analyzed (see the appendix), and α was set to be equal to 1.

Dependences in this model can be represented using the usual directed acyclic graph (DAG) representation and the directed Markov assumption (Cowell *et al.*, 1999)



$$\begin{aligned} \boldsymbol{\Pi} &= (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_Q) \\ \mathbf{C} &= (c_1, \ldots, c_T) \\ \mathbf{X} &= (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T) \\ \mathbf{M} &= (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_Q) \\ \boldsymbol{\Sigma} &= (\sigma_1^2, \ldots, \sigma_Q^2) \end{aligned}$$

As Q approaches infinity, the model above is equivalent to the corresponding Dirichlet process prior mixture model (Neal, 2000). This model is similar to the models described by Rasmussen (2000) and Escobar and West (1995). Previously, we used a similar approach to model multinomial data (Medvedovic, 2000). This hierarchical model can be easily generalized to include more general covariance structures, to accommodate experimental replicates, more flexible or more specific priors, missing data, etc.

The cluster analysis based on this model proceeds by approximating the joint posterior distribution of classification vectors given data, $p(\boldsymbol{c}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$. This distribution is implicitly specified by this hierarchical model but cannot be written in the closed form. However, it can be shown (Neal, 2000) that the posterior marginal distribution of classification variables given all model parameters and the data, when Q approaches infinity, is fully specified by following two equations:

$$\begin{aligned} &p(c_i = j|\boldsymbol{c}_{-i}, \boldsymbol{x}_i, \boldsymbol{\mu}_j, \sigma_j^2) \\ &\quad = b\frac{n_{-i,j}}{T - 1 + \alpha} f_N(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \sigma_j^2\boldsymbol{I}) \qquad \text{(V)} \\ &p(c_i \neq c_j, j \neq i|\boldsymbol{c}_{-i}, \boldsymbol{x}_i, \boldsymbol{\mu}_x, \sigma_x^2) \\ &\quad = b\frac{\alpha}{T - 1 + \alpha}\int f_N(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \sigma_j^2\boldsymbol{I}) \\ &\qquad \times p(\boldsymbol{\mu}_j, \sigma_j^2|\lambda, r^{-1})\mathrm{d}\boldsymbol{\mu}_j\mathrm{d}\boldsymbol{\sigma}_j^2 \qquad \text{(VI)} \end{aligned}$$

where, $n_{-i,c}$ is the number of expression profiles classified in $\boldsymbol{c}$, not counting the ith profile, $\boldsymbol{c}_{-i}$ is the classification vector for all except the ith profile and b is just a normalizing constant ensuring that all classification probabilities add up to one. Posterior conditional distributions for other model parameters are given in the Appendix. Having all complete conditional distributions specified, we can employ a Gibbs sampler to estimate joint posterior distribution of all model parameters. It is important to note that although the number of mixture components is technically assumed to be infinite, the number of nonempty components is finite and ranges between 1 and T.

While the above expression for the marginal probability of classification variables could be somewhat simplified, a closed form expression for the (unconditional) probability for the classification variables is not possible in light of the (non-conjugate) form of the priors we have chosen for $\boldsymbol{\mu}_j$s and $\boldsymbol{\sigma}_j$s. In the current set-up, we did not find any problems with mixing aspects of the Gibbs sampler, or the convergence. However, with the use of a conjugate prior and the resulting simplification, one may obtain a more efficient Gibbs sampler.

Prior knowledge was only used in the selection of the Dirichlet process precision parameter α. In the cell cycle data analysis, this parameter was set to 1 representing the prior belief that six to seven clusters are expected to transpire (Escobar, 1994). The prior belief about the number of clusters is based on the fact that we expected to see 6 different clusters related to different cell cycle phases (Cho *et al.*, 1998) and one cluster to allow for genes not related to cell cycle. In the sensitivity analysis, we demonstrated that the posterior distribution of clusterings was not sensitive to doubling and halving of α which is approximately equivalent to doubling and halving the prior expected number of mixture components (Escobar, 1994). An alternative approach to specifying the prior belief on the number of clusters could be to specify a prior distribution for α and treat it like all other parameters in the model (Escobar and West, 1995; Rasmussen, 2000). All other hyper-parameters in the model are given prior distributions that are centered around the overall profiles mean and standard deviation. This, we believe, is a reasonable 'default' choice in the absence of any specific prior information about the cluster locations.

EXPERIMENTAL REPLICATES MODEL

When the signal in the data is drowned in the experimental variability, the most obvious way to improve the signal to noise ratio is to perform experimental replicates. The cluster analysis of such data can be performed by first averaging gene expression measurements observed at different experimental replicates and applying an usual clustering procedure. Since the standard deviation of such average expression levels will decrease with the square root of the number of replicates, it is likely that weaker associations will be detected than in the case when a single replicate is performed. However, such an approach is not using all information present in such data. The information on the variability between experimental replicates is discarded and only the information about the mean expression level is utilized. In the situation when experimental variability varies from gene to gene, such information can be crucial for the proper assessment of confidence in our final clustering results. On the other hand, it is a well documented phenomenon that log-transformed expression measurements of low-expressed genes vary more than expression measurements of highly expressed genes. Baggerly *et al.* (2001) presented a probabilistic model of hybridization dynamics that explains this fact. Furthermore, it has been indicated that expression levels of some genes might be more variable than others (Hughes *et al.*, 2000). In the context of the mixture model based approach, experimental replicates and different between—replicates variability can be accommodated by modifying Level 1 of our hierarchical model (I–IV):

$$\begin{aligned} &p(\boldsymbol{x}_{ik} \mid \boldsymbol{y}_i, \psi_i) = f_N(\boldsymbol{x}_{ik} \mid \boldsymbol{y}_i, \psi_i^2 \boldsymbol{I}),\\ &p(\boldsymbol{y}_i \mid \boldsymbol{c}_i = j, \boldsymbol{\mu}_j, \boldsymbol{\sigma}_j) = f_N(\boldsymbol{y}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\sigma}_j^2 \boldsymbol{I}). \qquad \text{(VII)}\\ &\quad k = 1, \ldots, G,\ i = 1, \ldots, T,\ j = 1, \ldots, Q \end{aligned}$$

In this new formulation, $\boldsymbol{x}_{ik}$ represents the expression profile in the kth replicate for the ith gene and $\boldsymbol{y}_i$ is the mean expression profile for the ith gene. ψ_i^2 represents the between replicates variance for the ith gene which is assumed to be homogeneous across all experimental conditions. After integrating out the mean expression profiles for individual genes ($\boldsymbol{y}_i$s), the joint probability distribution of all observations for a single gene is given by

$$\begin{aligned} &p(\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iG} \mid c_i = j, \boldsymbol{\mu}_j, \sigma_j^2, \psi_i) =\\ &\prod_k f_N\left(\bar{\boldsymbol{x}}_{i\bullet} \mid \boldsymbol{\mu}_j, \left(\sigma_j^2 + \frac{\psi_i^2}{G}\right)\boldsymbol{I}\right)\\ &i = 1, \ldots, T,\\ &j = 1, \ldots, Q; \quad \bar{\boldsymbol{x}}_{i\bullet} = \frac{\sum_k \boldsymbol{x}_{ik}}{G}. \qquad \text{(VIII)} \end{aligned}$$

Prior distribution for the within profile variance ψ_{i}^2 and conditional posterior distributions for parameters in this model are also given in the Appendix.

COMPUTING POSTERIOR PROBABILITIES AND CLUSTER FORMATION

Gibbs sampler

The Gibbs sampler (Gelfand and Smith, 1990) is a general procedure for sampling observations from a multivariate distribution. It proceeds by iteratively drawing observations from complete conditional distributions of all components. Under mild conditions, the distribution of generated multivariate observations converges to the target multivariate distribution. The Gibbs sampler for generating a sequence of clusterings $\boldsymbol{c}^1, \boldsymbol{c}^2, \boldsymbol{c}^3, \ldots, \boldsymbol{c}^S$ proceeds as follows

Initialization phase: The algorithm is started by assuming that all profiles are clustered together. That is c^0 is initialized as:

$$\boldsymbol{c}_1^0 = \boldsymbol{c}_2^0 = \cdots = \boldsymbol{c}_{\mathrm{T}}^0 = 1.$$

Consequently, Q_0 is set to one. Corresponding pattern parameters $\boldsymbol{\mu}_1$ and σ_1^2 are generated as random samples from their prior distribution.

Iterations: Given parameters after the kth step $(c^k, Q_k, \mu_1, \ldots, \mu_{Qk}, \sigma_1, \ldots, \sigma_k)$, the $(k+1)$th set of parameters is generated by first updating classification variables, that is drawing c^{k+1} according to their posterior conditional distributions given by (V) and (VI). This also determines the value of Q_{k+1}. New $\boldsymbol{\mu}$s and σ^2s, as well as hyperparameters λ, r, w, and β are then generated according to their posterior conditional distributions (Appendix). Whenever the number of profiles in a cluster falls to zero, the cluster is removed from the list. A new cluster is created whenever a $c_i \neq c_j$ for all $i \neq j$ is selected.

It can be shown (Neal, 2000) that this algorithm, in the limit, generates clusterings from the desired posterior distribution of clusterings. Therefore, it can be assumed that the empirical distribution of generated clusterings c^B, $c^{B+1}, \ldots$, after B 'burn-in' samples, approximates the true posterior distribution of clusterings. Groups of genes that had common assignments in a large proportion of generated clusterings are likely to have been generated by the same underlying pattern. That is, the proportion of clusterings in which a group of genes had common assignments approximates the probability that they are generated by the same underlying pattern. In addition to the posterior distribution of clusterings averaged over all possible number of clusters, the results of this sampling procedure enable us to estimate the posterior distribution of the number of clusters present in the data.

Cluster Formation

Given the sequence of clusterings $(\boldsymbol{c}^B, \boldsymbol{c}^{B+1}, \ldots, \boldsymbol{c}^S)$ generated by the Gibbs sampler after B 'burn-in' cycles, pair-wise probabilities for two genes to be generated by the same pattern are estimated as:

$$P_{ij} = \frac{\text{\# of samples after 'burn-in' for which } c_i = c_j}{S - B}$$

Using these probabilities as a similarity measures, pair-wise 'distances' are created as

$$D_{ij} = 1 - P_{ij}.$$

Based on these distances, clusters of similar expression profiles are created by the complete linkage approach (Everitt, 1993) utilizing the Proc Cluster and Proc Tree of the SAS© statistical software package (1999). It is important to notice that using the posterior pairwise probabilities and the complete linkage approach are just one of many possible approaches to processing the posterior distribution of clusterings generated by the Gibbs sampler. An alternative approach to identifying the optimal clustering could be to identify the most probable clustering using the traditional maximum a posteriori probability (MAP) approach. That is, identifying the clustering that occurred more times in the Gibbs sampler generated sequence than any other clustering. The problem with that type of approach is that many different clusterings have a very similar yet very small posterior probability. Choosing the 'best' out of a slew of very similar clusterings without summarizing the other almost equally probable alternatives would likely offer an incomplete picture about the observed distribution of clusterings. Other approaches to summarizing the posterior distribution of clusterings are complicated by the complex nature of the clustering space. Finding an optimal approach to summarizing the posterior distribution of clusterings is a challenging task that needs further improvements.

DATA ANALYSIS

Yeast cell cycle data

The utility of this procedure is illustrated by the analysis of the cell-cycle data described and analyzed by Cho *et al.* (1998). This data set has been routinely used to demonstrate effectiveness of various clustering approaches (Tamayo *et al.*, 1999; Yeung *et al.*, 2001a,b; Tavazoie *et al.*, 1999; Lukashin and Fuchs, 2001) The complete data set was downloaded from http://genomics.stanford.edu/. Data was processed in a similar fashion as described by (Tamayo *et al.*, 1999). 893 genes that showed significant variation in both cell cycles were identified. Data was log-transformed and normalized separately for each cell cycle. Clusters of similar expression profiles were identified based on 100 000 Gibbs sampler generated clusterings after 100 000 'burn-in' cycles. The distribution of the posterior pairwise distances is shown in the Figure 1a. Approximately 75% of all posterior pairwise distances were equal to 1. In other words, the majority of posterior probabilities of two profiles being generated by the same underlying Gaussian distribution were equal to zero.

Before we created clusters of similar profiles, we removed 'outlying' profiles that were defined to be those profiles for which all posterior pairwise probabilities were less than 0.5. Thirty 'outlying' profiles were identified and removed from the analysis (the list and the plot are given at the supporting web page). Clusters were generated by separating profiles in groups with the maximum possible complete linkage distance. That is, for any pair of clusters formed, there was at least one profile in the first cluster that had a zero posterior pairwise probability with at least one profile in the second cluster. Twelve clusters in the Figure 2 with total of 302 expression profiles were selected, from the total of 43 clusters generated, based on their periodicity. Additional five outlying profiles were identified separately as correlating with the cell-cycle stages (see additional material). These profiles represent five groups of cell cycle related genes identified by both Cho *et al.* (1998) and Tamayo *et al.* (1999). Further examination of the profiles identified in our analysis reveals that out of 389 genes identified by Cho *et al.*, 1998 as peaking at specific portion of the cell cycle, 251 also satisfied our selection criteria. Out of these 251 genes, 201 were identified by our analysis to be associated with cell-cycle events. An additional nine genes out of 50 that were not identified by our analysis showed a significant association with genes in Cluster 43 (see supplemental material), meaning that the average pairwise posterior probability with genes in this cluster were greater than 0.1. Two other genes from this group had a significant association with one of the 12 clusters in Figure 1. Although Cluster 43 does exhibit a cell-cycle associated

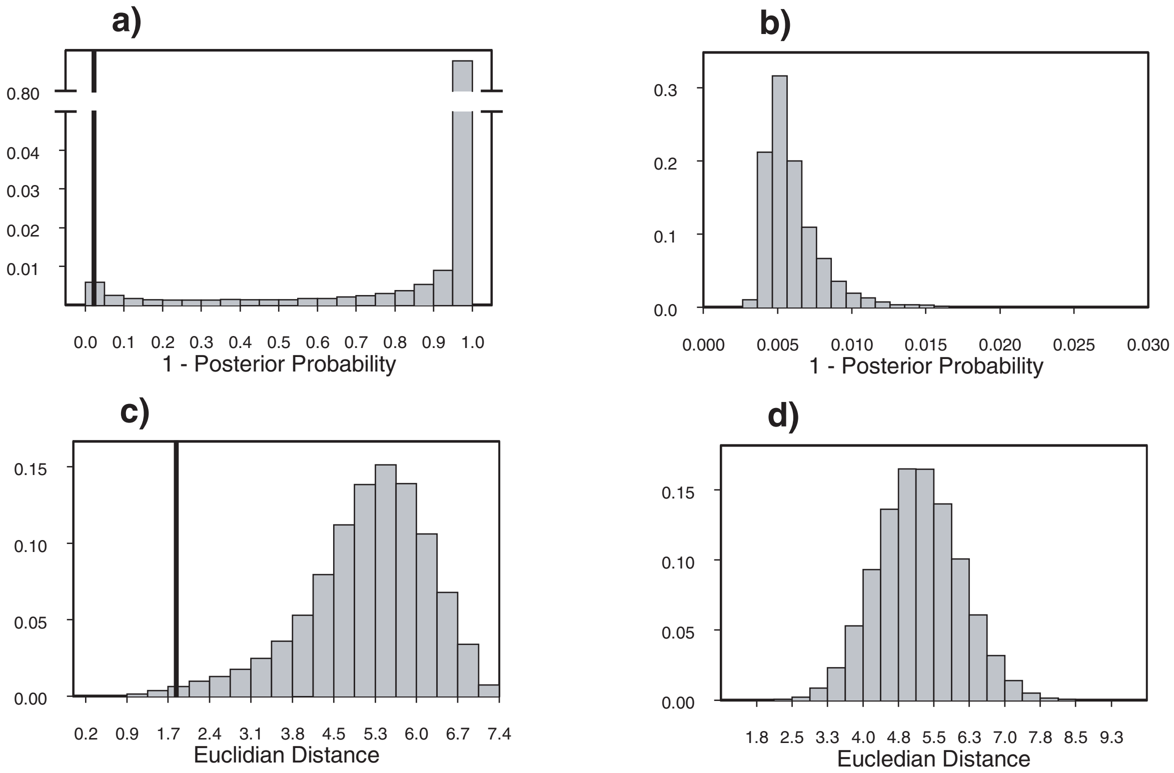


Fig. 1. (a) The distribution of posterior pairwise distances. The vertical line represents the maximum posterior distance observed by analyzing the bootstrapped data set. (b) The distribution of posterior pairwise distances observed by analyzing the bootstrapped data set. (c) The distribution of pairwise Eucledian distances. The vertical line denotes the minimum pairwise Euclidian distance in the bootstrapped data set. (d) The distribution of pairwise Euclidian distances in the bootstrapped data set.

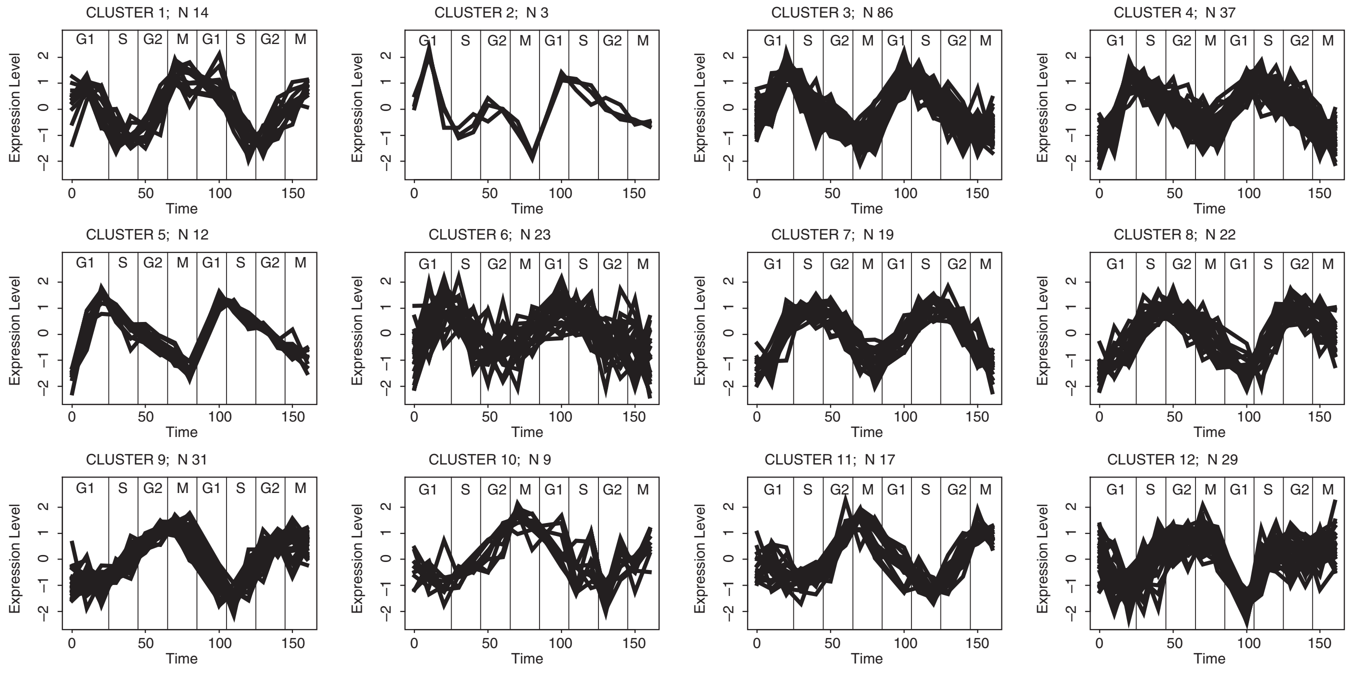


Fig. 2. Twelve Cell-cycle associated clusters out of total 43 clusters identified.

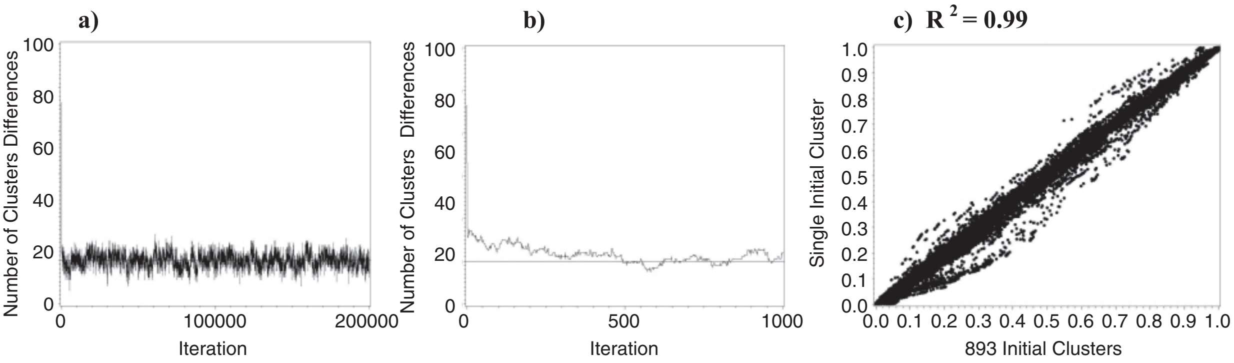


Fig. 3. Mixing properties of the Gibbs sampler. (a) The path of the difference in the number of clusters in 200 000 generated samples for two independent runs of the Gibbs sampler. The first Gibbs sampler was initialized by assuming that each profile was generated by a separate mixture component and the second Gibbs sampler was initialized with the number of mixture components set to 1. (b) Same as (a) except only first 1000 are shown. (c) Scatter plot of posterior pairwise probabilities for these two independent runs of the Gibbs sampler.

behavior, due to the large variability, it was not selected as being cell-cycle related. By adding this cluster, the number of genes being implied as cell-cycle related would increase to a total of 355. While visual inspection of other generated clusters reveal obvious similarities between expression profiles, they do not seem to be correlated with cell-cycle events. Plots of all 43 clusters can be viewed at the supporting web page. By a closer examination of 106 genes that were identified by our analysis but not by the initial examination performed by Cho *et al.* (1998), we established that majority of them are functionally classified, according to MIPS database (Mewes *et al.*, 1999), in categories previously shown to be associated with cell-cycle (Table 1).

This shows that our analysis is capable of providing additional insight not accessible by a visual inspection of hierarchically ordered data as performed by Cho *et al.* (1998). The complete list of these 106 genes along with their functional classification is included in the supplemental material.

Convergence of the Gibbs sampler

We assessed the appropriateness of our 'burn-in' period by comparing posterior distributions obtained from two independent runs of the Gibbs sampler (Figure 3). First Gibbs sampler was initialized with the number of mixture components set to one while the second Gibbs sampler was initialized by assuming that each profile was generated by a separate mixture component. As it can be seen from Figures 3a and 3b, the number-of-clusters parameter (Q) quickly converged to the same stationary distribution. Its posterior distribution, as well as the posterior distributions of other model parameters generated by the two runs were indistinguishable after the 'burn-in' period we used (100 000 samples) (data not shown). Furthermore, posterior pairwise probabilities, our key parameters used for creating clusters, correlated extremely well for two independent runs (Figure 3c). These results led us to believe that the 'burn-in' period we used was appropriate. In this paper we based our clustering on the data from a single run. An alternative approach would be to perform several independent runs and average the results. Potential benefits of running several shorter runs instead of a single long Gibbs sampler, as well as advantages of using potentially more efficient Gibbs sampling scheme (e.g. Jain and Neal, 2000; MacEachern, 1994) need to be further investigated.

Importance of Model Averaging

To demonstrate the importance of model-averaging in the process of identifying groups of similar profiles, we performed a finite mixture analysis of the cell cycle data. To do this we used the MCLUST software described by Fraley and Raftery (1999). In concordance with our infinite mixture model we used the unequal volume spherical model. We first identified the optimal number of clusters using the Bayesian Information criterion (BIC) (Schwarz 1978). Plot of BICs for models having between one and 100 clusters is shown in Figure 4a. The model with 23 clusters had the highest BIC and was assumed to be 'optimal'. In the finite mixture approach, model parameters in the finite mixture model are first calculated based on maximum likelihood principle. Based on these estimates, clusters are then formed based on the posterior probabilities of individual profiles of being generated by a particular mixture component (McLachlan and Basford, 1987; Fraley and Raftery, 1999; Yeung *et al.*, 2001a). Results of this analysis were compared to the results of our infinite mixture model. The goal of the comparison was to compare the reliability of clusterings obtained by two different approaches. In this respect, we posed the following question: How robust are the probabilities of concluding that two expression profiles are generated

Table 1. Functional classification of 106 genes implicated by our analysis but not implicated in the original paper (Number of new genes) compared to functional categorization of 387 genes implicated in the original paper (number of old genes). Since some of the genes belong to multiple categories, totals in the table are larger than total number of genes

Functional category	Number of new genes	Number of old genes
Cell cycle and dna processing	16	158
Cell fate	11	70
Cell rescue, defense and virulence	7	14
Cellular communication/signal transduction mechanism	1	5
Cellular transport and transport mechanisms	5	23
Classification not yet clear-cut	2	13
Control of cellular organization	1	16
Energy	3	15
Metabolism	20	60
Protein fate (folding, modification, destination)	6	18
Protein synthesis	0	4
Regulation of/interaction with cellular environment	1	16
Subcellular localisation	28	196
Transcription	10	44
Transport facilitation	5	14
Unclassified protein	51	99

by the same probability distribution (i.e. that they will cluster together) in the two approaches? To assess the robustness of the finite mixture approach we repeated the analysis assuming that the number of clusters is one less or one more than the number indicated by BIC. That is, we fitted the finite mixture models with 22 and 24 mixture components and calculated the pairwise posterior probabilities of two profiles being clustered together for all pairs of profiles and for all three finite mixture models. Suppose that p^q_{ik} is the posterior probability that the ith profile is generated by the kth mixture component (McLachlan and Basford, 1987) in the model with q mixture components. Than, the pairwise posterior probability of two profiles (profile i and profile j) being generated by the same mixture component can be estimated as

$$P^q_{ij} = \sum_{k=1}^{q} p^q_{ik} p^q_{jk}$$

Scatter plots of pairwise probabilities for mixture models with 22 and 24 mixture components against the model with 23 mixture components, along with the corresponding coefficients of linear determination (R^2) are shown in Figure 4b and 4c. It is obvious from these plots that there is a strong dependence of these probabilities on the number of mixture components used in the analysis. A small deviation in the number of clusters could have a strong influence on the quality of clusters as well as the estimated confidence in the clustering. That is, the results from the finite mixture approach can be highly non-robust to the selected ‘optimal’ number of clusters. This means the correct specification of the number of clusters is a paramount in the finite mixture approach, which would typically be difficult in practice without substantial amount of prior information to justify such specification.

When the number of mixture components is averaged out, as it is in the infinite mixture approach, the posterior distribution of clusterings, and consequently clusters created based on this distribution, is robust with respect to the specification of the prior number of clusters. Doubling or halving (setting $\alpha = 2$ or 0.5) of the mean prior number of clusters (Escobar, 1994) does affect the posterior distribution of the number of clusters (Figure 4d) but has little effect on the posterior pairwise probabilities (Figure 4e and 4f).

Although regularity conditions required for BIC to be asymptotically correct are not satisfied in the finite mixture model, empirical studies have shown that the criterion works quite well in identifying the correct number of mixture components (Biernacki and Govaert, 1999). Our goal here was not to propose a better approach to obtaining the number of clusters in the data. However, we believe that there are shortcomings with the finite mixture approach that are remedied by the infinite mixture approach. In the cluster analysis based on the finite mixture model, calculated confidence in a particular clustering does not take into account the uncertainties related to the choice of the right number of clusters based on the BIC criterion. Consequently, they are valid only under the model where a specific number of clusters is assumed known. On the other hand, when there is uncertainty about the number of clusters, as is often the case in practice, we demonstrated that ensuing results can be highly sensitive to the chosen number of clusters. These

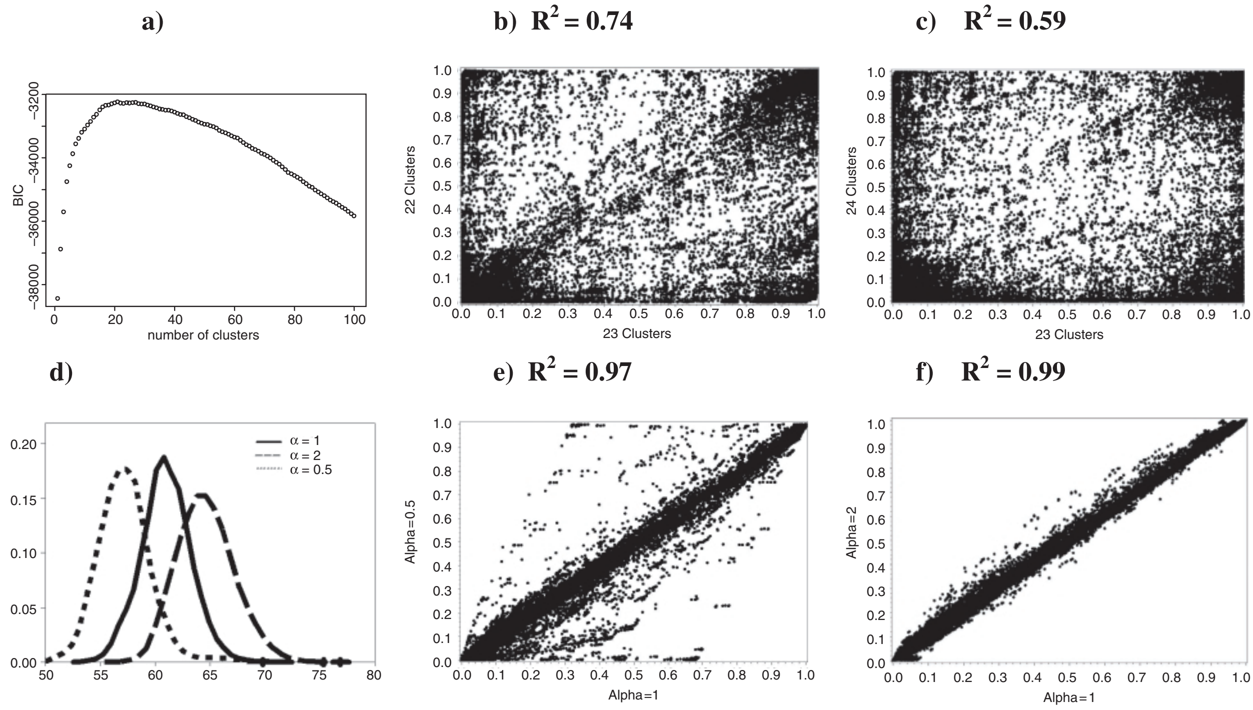


Fig. 4. Effect of changes in the assumed number of mixture components (clusters) for finite and infinite mixture models. (a) BIC criterion for number of clusters ranging from 1 to 100. (b) Scatter plot of pairwise probabilities of two profiles being clustered together for the 'optimal' with 23 mixture components versus the model with 22 components. (c) Scatter plot of pairwise probabilities of two profiles being clustered together for the 'optimal' with 23 mixture components vs the model with 24 components. (d) Posterior distribution of number of mixture components for three different levels of α. (e) Scatter plot of posterior pairwise probabilities of two profiles being clustered together for $\alpha = 1$ versus $\alpha = 0.5$. (f) Scatter plot of posterior pairwise probabilities of two profiles being clustered together for $\alpha = 1$ versus $\alpha = 2$.

are compelling reasons, in our view, for circumventing the whole process of identifying the number of clusters altogether by integrating it out, as it is done in the infinite mixture approach.

Simulated Replicates data

We performed a simulation study to assess the advantage of using full replicates-model (VII) in situations when the experimental variability varies from gene to gene. Data was simulated to represent a two-cluster situation with sixty observations from Gaussian distributions in each cluster. The number of replicates was two or ten. 1000 data sets were simulated for each, two and ten replicates situation. Following the structure of the model (VII), data was generated in two stages:

Stage 1:

120 mean expression profiles ($\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{120}$) were generated. Sixty of them ($\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{60}$) were randomly sampled from the Gaussian distribution with the mean of 0 and the variance 0.005, and other 60 ($\boldsymbol{y}_{61}, \ldots, \boldsymbol{y}_{120}$) from the Gaussian distribution with the mean of 1.5 and same variance. Between-replicates variances for each mean expression profile ($\psi_1^2, \ldots, \psi_{120}^2$) were randomly sampled from the Inverse Gamma distribution with the shape parameter 2 and the scale parameter 0.4 for the two-replicates scenario, and with the scale parameter two and the location parameter two for the ten-replicates scenario. These parameters reflect the scenario in the which majority of between profiles variability within the same cluster is due to the experimental between-replicates variability, and the total variability in averaged profiles ($\bar{x}_{\mathrm{i}\bullet}$) was comparable for two and ten replicates situation.

Stage 2:

For each mean expression profile $\boldsymbol{y}_i$, $i = 1, \ldots, 120$, $G = 2$ or 10 expression profile replicates ($\boldsymbol{x}_{i,1}, \ldots, \boldsymbol{x}_{i,G}$) were randomly sampled from the Gaussian distribution with the mean y_i and the variance ψ_i^2.

The clustering procedure based on the full model showed an improvement in the ability to correctly separate observations from different clusters. The difference is particularly evident in the situation when the between replicates variability is high. Scatter plots in Figure 4 represent observed differences between average posterior pairwise probabilities for a profile and the rest of the profiles in the same cluster and in the opposite cluster (R_i, $i = 1, \ldots, 120$) for all 1000 simulated samples. That

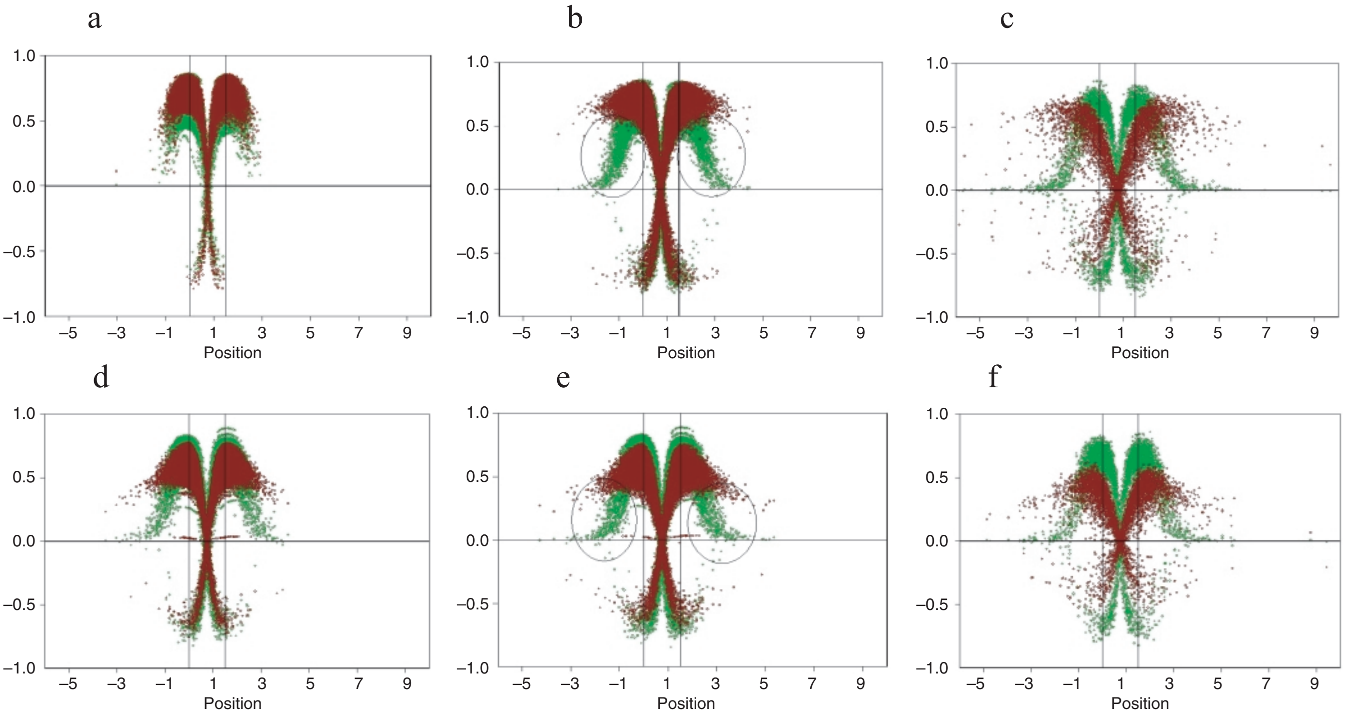


Fig. 5. Scatter plot of the differences between average posterior pairwise probabilities for a profile with the rest of the profiles in the same cluster and in the opposite cluster (R_i, $i = 1, \ldots, 120$) for all 1000 simulated data sets. Red dots represent the full model-based analysis and green dots represent the analysis based on the averaged profiles. Vertical lines denote two cluster centers at 0 and 1.5. (a) Profiles with the below median between replicates standard deviation. (b) Profiles with the between replicates standard deviation being between 50th and 95th percentile. (c) Profiles with the higher than 95th percentile between replicates standard deviation. (d), (e) and (f) are same as (a), (b) and (c) respectively only for 2 replicates situation.

is, the Rs are defined as:

$$R_i = \sum_{j=1}^{60} \frac{P_{ij}}{60} - \sum_{j=61}^{120} \frac{P_{ij}}{60}, \quad \text{for } i = 1, \ldots, 60, \quad \text{and}$$

$$R_i = \sum_{j=61}^{120} \frac{P_{ij}}{60} - \sum_{j=1}^{60} \frac{P_{ij}}{60}, \quad \text{for } i = 61, \ldots, 120.$$

The more positive this difference is, it is more likely that the profile will be correctly clustered. The more negative this difference is, it is more likely that the profile will be misclassified.

As the between replicates variability increases, the difference in the performance of the full replicates model vs. the simple model becomes more obvious. For the below median variable profiles in the ten replicates situation (Figure 5a), both methods perform similarly with the full model showing somewhat higher confidence in the correct clustering. For variabilities between the median and the 95th percentile of all variabilities (Figure 5b and 5e), the full model still behaves as expected while the probability of correctly clustering profiles that are relatively far from cluster centers using the simple model is seriously diminished (circled observations). This trend is evident in the two-raplicates situation even in the lowest variability group (Figure 5d). In the extreme variability situations (Figure 5c and 5f), in addition to this trend, there is also an increased chance of incorrectly classifying profiles that are lying between or close to two cluster centers. These observations suggest that, when experimental replicates are available, there is a clear advantage of using the full model instead of using a simple model with averaged data.

DISCUSSION

Assuming a few basic principles, the Bayesian approach is the optimal approach to drawing uncertain conclusions from noisy data (Baldi and Brunak, 1998). The utility of Bayesian approach in computational biology has been demonstrated in a multitude of situations ranging from the analysis of biological sequences (Liu and Lawrence, 1999), to identification of differentially expressed genes. (Newton *et al.*, 2001; Baldi and Long, 2001; Long *et al.*, 2001; Lonnstedt and Speed, 2001). In this article we developed a model-based clustering procedure utilizing the Bayesian paradigm for knowledge discovery from noisy data. We postulated a reasonable probabilistic model for generation of gene expression profiles and performed clustering by estimating the posterior distribution of clusterings given data. The clustering and the measures of uncertainty about the clustering are assessed by model averaging across models with all possible number of clusters. In this respect, the model-averaging approach represents a qualitative shift in the problem of identifying groups of co-expressed genes.

Previously, Schmidler *et al.* (2000) used the Bayesian model-averaging approach to estimate probabilities of a particular secondary structure after averaging over all possible segmentation. Furthermore, Zhu *et al.* (1998) used a similar approach to estimate distances between a pair of sequences after averaging over all possible alignments. In both situations, taking into account uncertainties in the model selection proved beneficial for the overall performance of the computational procedure. Precise quantification of uncertainties related to clustering of expression profiles is especially important when obtained clustering is used as the starting point for identifying common regulatory motifs (e.g. Tavazoie *et al.*, 1999) for genes that cluster together. The inclusion of genes that have a low probability of actually belonging to a cluster in such an analysis is likely to seriously undermine ones ability to detect relevant commonalities in regulatory regions of clustered genes. One appealing way of taking into account uncertainties about created clusters in identifying common regulatory sequences is to simultaneously model both gene expression and corresponding promoter sequence data. Holmes and Bruno (2000) recently proposed a ‘sequence-expression’ model that incorporates both, the Bayesian finite mixture model for expression data and the sequence model of Lawrence *et al.* (1993) for promoter sequence data. Extending this model to incorporate model-averaging over models with different number of clusters and for incorporating experimental replicates seems to be a logical next step in developing such models.

The procedure we used to create clusters based on the estimated posterior distribution of clusterings is the usual complete linkage nearest-neighbor approach with pairwise posterior distances between two profiles as the distance measure (Everitt, 1993). However, in contrast to usual distance measures (e.g. the Euclidian distance) that are based only on two profiles at a time, our measure incorporates information from the whole data set in assessing the distance between any two profiles. The effect of such ‘information pulling’ is demonstrated by comparing the distribution of our posterior distances to the distribution of the commonly used Euclidian distances (e.g. Lukashin and Fuchs, 2001) calculated for the same data set (Figure 1c). The high concentration of posterior distances around the maximum distance (equal to 1) allows for an easy identification of unrelated profiles. On the other hand, the distribution of corresponding Euclidian distances is rather smooth throughout the range and there is no obvious separation of clearly unrelated genes. This point is further emphasized by observing distributions of these two distance measures for the structure-less data set obtained by bootstrapping data (Efron and Tibshirani, 1998) from the original data set (Figure 1b and 1d). The distribution of posterior pairwise distances is drastically different for the bootstrapped data set when compared to the original data and it indicates that all data in the bootstrapped set form a single cluster. In contrast to this, differences between distributions of Euclidean distances for two data sets is rather subtle and it is unclear what distance indicates a high-confidence separation between two profiles. Furthermore, posterior pairwise probabilities are themselves meaningful measures of uncertainty about similarity of two expression profiles and they can be used to assess confidence and uncertainty about clusters and the membership of individual profiles in a cluster. On the other hand, usual distance measures such as Euclidian distance and similarity measures such as correlation coefficient are meaningful only in comparison to their distribution under the no-structure assumption (Lukashin and Fuchs, 2001; Herrero *et al.*, 2001).

The issue of performing experimental replicates to model various sources of experimental variability will likely play an increasingly important role in modeling microarray data. Currently, information about experimental variability is commonly used when the goal of the analysis is to identify differentially expressed genes. Results of our simulation study show that using this information when clustering expression data is likely to improve results of the analysis. Failing to take into account differences in precisions with which expression levels of different genes are measured can result in both, a failure to identify existing and/or inferring non-existing relationships.

A major assumption of mixture model, finite and infinite, is that, given the classification variables and the parameters of individual mixture components, expression profiles for individual genes are independently distributed as multivariate Gaussian random variables. It has been previously shown (Yeung *et al.*, 2001a) that the normality assumption is rather reasonable for properly transformed microarray data sets. In our own experience with both Affymetrix oligonucleotide arrays and two-color cDNA microarrays, after a proper normalization, distribution of log-transformed data closely resembles Gaussian distributions with different variances for different genes (data not shown). Conditional independence assumption is more difficult to assess. Typically, in the process of normalizing microarray data, an attempt is made to remove systematic correlations that can occur due to the microarray-to- microarray variability that systematically affects all genes on the microarray (Yang *et al.*, 2001). More subtle correlations due to, for example, cross-hybridizations resulting from sequence similarities are much more difficult to assess. Even if such correlations can be quantified, there is no obvious way to incorporate them in the traditional mixture model.

Finally, the model based approach described here is a flexible platform for dealing with various other issues in the cluster analysis of expression profiles. In addition to model-averaging and outlier detection that we demonstrated, Gibbs sampler described here can be easily modified to handle incomplete profiles (profiles missing some data points). In such a situation, it is again important to take into account additional source of variability introduced by imputing missing data points. At this point we are unaware of another clustering approach capable of dealing with all these problems at once

ACKNOWLEDGEMENTS

This work was partially supported by NIEHS grants P30-ES06096 and P42-ES04908. We are also thankful for very insightful comments and suggestions offered by the referees.

REFERENCES

Baggerly,K.A., Coombes,K.R., Hess,K.R., Stivers,D.N., Abruzzo,L.V. and Zhang,W. (2001) Identifying differentially expressed genes in cDNA microarray experiments. *J. Comput. Biol.*, **8**, 639–659.

Baldi,P. and Brunak,S. (1998) *Bioinformatics: The Machine Learning Approach*. The MIT Press, Cambridge, MA.

Baldi,P. and Long,A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509–519.

Biernacki,C. and Govaert,G. (1999) Choosing models in model-based clustering and discriminant analysis. *J. Statis. Comput. Simul.*, **64**, 49–71.

Cho,R.J., Campbell,M.J., Winzeler,E.A., Steinmetz,L., Conway,A., Wodicka,L., Wolfsberg,T.G., Gabrielian,A.E., Landsman,D., Lockhart,D.J. and Davis,R.W. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.

Cowell,R.G, Dawid,P.A., Lauritzen,S.L. and Spiegelhalter,D.J. (1999) *Probabilistic Networks and Expert Systems*. Springer, New York.

D'haeseleer,P., Liang,S. and Somogyi,R. (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, **16**, 707–726.

Efron,B. and Tibshirani,R.J. (1998) *An Introduction to the Bootstrap*. Chapman & Hall/CRC, Boca Raton.

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.

Escobar,M.D. (1994) Estimating normal means with a Dirichlet process prior. *J. Amer. Stat. Assoc.*, **89**, 268–277.

Escobar,M.D. and West,M. (1995) Bayesian density estimation and inference using mixtures. *J. Amer. Stat. Assoc.*, **90**, 577–588.

Everitt,B.S. (1993) *Cluster Analysis*. Edward Arnold, London.

Ferguson,T.S. (1973) A Bayesian analysis of some nonparametric problems. *Ann. Stat.*, **1**, 209–230.

Fraley,C. and Raftery,A.E. (1999) Mclust: Software for Modelbased Cluster Analysis. *Journal of Classification*, **16**, 297–306.

Gelfand,E.A. and Smith,F.M.A. (1990) Sampling-based approaches to calculating marginal densities. *J. Amer. Stat. Assoc.*, **85**, 398–409.

Herrero,J., Valencia,A. and Dopazo,J. (2001) A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, **17**, 126–136.

Holmes,I. and Bruno,W.J. (2000) Finding regulatory elements using joint likelihoods for sequence and expression profile data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 202–210.

Hughes,T.R., Marton,M.J., Jones,A.R., Roberts,C.J., Stoughton,R., Armour,C.D., Bennett,H.A., Coffey,E., Dai,H., He,Y.D., Kidd,M.J., King,A.M., Meyer,M.R., Slade,D., Lum,P.Y., Stepaniants,S.B., Shoemaker,D.D., Gachotte,D., Chakraburtty,K., Simon,J., Bard,M. and Friend,S.H. (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Ideker,T., Thorsson,V., Siegel,A.F. and Hood,L.E. (2000) Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J. Comput. Biol.*, **7**, 805–817.

Jain,S. and Neal,R. (2000) A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Technical Report No. 2003*, Department of Statistics, University of Toronto.

Kerr,K.M. and Churchill,G.A. (2001) Experimental design for gene expression microarrays. *Biostatistics*, **2**, 183–201.

Kerr,K.M., Martin,M. and Churchill,G.A. (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, **7**, 819–837.

Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Liu,J.S. and Lawrence,C.E. (1999) Bayesian inference on biopolymer models. *Bioinformatics*, **15**, 38–52.

Long,A.D., Mangalam,H.J., Chan,B.Y., Tolleri,L., Hatfield,G.W. and Baldi,P. (2001) Improved statistical inference from dna microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in *Escherichia coli* K12. *J. Biol. Chem.*, **276**, 19937–19944.

Lonnstedt,I. and Speed,T.P. (2001) Replicated microarray data. *Statistical Sinica*, **12**, 31–46.

Lukashin,A.V. and Fuchs,R. (2001) Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters. *Bioinformatics*, **17**, 405–414.

MacEachern,S.N. (1994) Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics: Simulation and Computation*, **23**, 727–741.

McLachlan,J.G. and Basford,E.K. (1987) *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York.

Medvedovic,M. (2000) clustering multinomial observations via finite and infinite mixture models and MCMC Algorithms. *Proceedings of the Joint Statistical Meeting 2000: Statistical Computing Section*. pp. 48–51.

Mewes,H.W., Heumann,K., Kaps,A., Mayer,K., Pfeiffer,F., Stocker,S. and Frishman,D. (1999) MIPS: a database for protein sequences and complete genomes. *Nucleic Acids Res.*, **27**, 44–48.

Neal,R.M. (2000) Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, **9**, 249–265.

Newton,M.A., Kendziorski,C.M., Richmond,C.S., Blattner,F.R. and Tsui,K.W. (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.*, **8**, 37–52.

Rasmussen,C.A. (2000) The infinite gaussian mixture model. *Advances in Neural Information Processing Systems*, **12**, 554–560.

Rocke,D.M. and Dubin,B. (2001) A Model for Measurement Error for Gene Expression Arrays. *J. Comput Biol.*, **8**, 557–569.

SAS/STAT User's Guide, (1999) SAS Institute, Cary, NC.

Schmidler,S.C., Liu,J.S. and Brutlag,D.L. (2000) Bayesian segmentation of protein secondary structure. *J. Comput. Biol.*, **7**, 233–248.

Tamayo,P., Slonim,D., Mesirov,J., Zhu,Q., Kitareewan,S., Dmitrovsky,E., Lander,E.S. and Golub,T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.

Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nature Genet.*, **22**, 281–285.

Wolfinger,R.D., Gibson,G., Wolfinger,E.D., Bennett,L., Hamadeh,H., Bushel,P., Afshari,C. and Paules,R.S. (2001) Assessing Gene Significance From CDNA Microarry Expression Data Via Mixed Models. *J. Comput. Biol.*, **8**, 625–637.

Yang,Y.H., Dudoit,S., Luu,P. and Speed,T. (2001) Normalization for cDNA microarray data. *SPIE BiOS 2001*. San Jose, California.

Yeung,K.Y., Fraley,C., Murua,A., Raftery,A.E. and Ruzzo,W.L. (2001a) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977–987.

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

Zhu,J., Liu,J.S. and Lawrence,C.E. (1998) Bayesian adaptive sequence alignment algorithms. *Bioinformatics*, **14**, 25–39.

APPENDIX

Posterior probability distributions for the single replicate case

$$p(\boldsymbol{\mu}_j|\boldsymbol{c},\sigma_j^2,\boldsymbol{X},\boldsymbol{\lambda},r) = f_N\left(\boldsymbol{\mu}_j\left|\frac{r^{-1}\bar{\boldsymbol{x}}_{j\bullet}+\frac{\sigma_j^2}{n_j}\boldsymbol{\lambda}}{r^{-1}+\frac{\sigma_j^2}{n_j}},\frac{r^{-1}\frac{\sigma_j^2}{n_j}}{r^{-1}+\frac{\sigma_j^2}{n_j}}I\right.\right)$$

$$p(\sigma_j^{-2}|\boldsymbol{X},\boldsymbol{M},\beta,w) = f_G\left(\sigma_j^{-2}\left|\frac{Mn_j+\beta}{2},\frac{s_j^2+\beta w}{2}\right.\right)$$

$$f(\boldsymbol{\lambda}|\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_Q,r) = f_N\left(\boldsymbol{\lambda}\left|\frac{\sigma_x^2\frac{\sum_i\boldsymbol{\mu}_i}{Q}+\frac{r^{-1}}{Q}\boldsymbol{\mu}_x}{\sigma_x^2+\frac{r^{-1}}{Q}},\frac{\sigma_x^2\frac{r^{-1}}{Q}}{\sigma_x^2+\frac{r^{-1}}{Q}}I\right.\right)$$

$$f(r|\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_Q,\boldsymbol{\lambda}) = f_G\left(r\left|\frac{MQ+1}{2},\frac{\sum_i(\boldsymbol{\mu}_i-\boldsymbol{\lambda})(\boldsymbol{\mu}_i-\boldsymbol{\lambda})'+\sigma_x^2}{2}\right.\right)$$

where

$$\boldsymbol{\mu}_x = \frac{\sum_{i=1}^T\boldsymbol{x}_i}{T}\quad \sigma_x^2 = \frac{\sum_{i=1}^T(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})'}{TM-1}$$

$$\bar{\boldsymbol{x}}_{j\bullet} = \frac{\sum_{c_i=j}^T\boldsymbol{x}_i}{n_j}\quad s_j = \sum_{c_i=j}^T(\boldsymbol{x}_i-\boldsymbol{\mu}_j)(\boldsymbol{x}_i-\boldsymbol{\mu}_j)'.$$

$$f(w|\sigma_1^{-2},\ldots,\sigma_Q^{-2},\beta) = f_G\left(w\left|\frac{Q\beta+1}{2},\frac{\beta\sum_j\sigma_j^{-2}+\sigma_x^{-2}}{2}\right.\right)$$

$$f(\beta|\sigma_1^{-2},\ldots,\sigma_Q^{-2},w) \propto \Gamma\left(\frac{\beta}{2}\right)\left(\frac{\beta}{2}\right)^{\frac{(Q\beta-3)}{2}}\exp\left(-\frac{\beta^{-1}}{2}\right)$$

$$\times\prod_j\left[(w\sigma_j^{-2})^{\frac{\beta}{2}}\exp\left(-\frac{\sigma_j^{-2}w\beta)}{2}\right)\right].$$

Posterior conditional probability distributions of parameters in the multiple replicates model

The prior distribution for the within profile variance is given by

$$p(\psi_i^{-2}) = f_G\left(\psi_i^{-2}\left|\frac{1}{2},\frac{\sigma_w^2}{2}\right.\right)$$

where

$$\sigma_w^2 = \frac{\sum_{i=1}^T\sum_{k=1}^G(\boldsymbol{x}_{ik}-\bar{\boldsymbol{x}}_{i\bullet})(\boldsymbol{x}_{ik}-\bar{\boldsymbol{x}}_{i\bullet})'}{(G-1)TM}\ \text{and}\ \bar{\boldsymbol{x}}_{i\bullet} = \frac{\sum_{k=1}^G\boldsymbol{x}_{ik}}{G}.$$

The posterior conditional distribution for parameters in the model for experimental replicates are:

$$f(\boldsymbol{y}_i|c_i=j,\boldsymbol{\mu}_j,\sigma_j^2,\psi_i^2,\boldsymbol{x}_{i1},\ldots,\boldsymbol{x}_{iG})$$

$$= f_N\left(\boldsymbol{y}_i\left|\frac{\sigma_j^2\bar{\boldsymbol{x}}_{i\bullet}+\frac{\psi_i^2}{G}\boldsymbol{\mu}_j}{\sigma_j^2+\frac{\psi_i^2}{G}},\frac{\frac{\psi_i^2}{G}\sigma_j^2}{\sigma_j^2+\frac{\psi_i^2}{G}}I\right.\right)$$

$$f(\psi_i^{-2}|\boldsymbol{y}_i,\sigma_w^2,\boldsymbol{x}_{i1},\ldots,\boldsymbol{x}_{iG})$$

$$= f_G\left(\psi_i^{-2}\left|\frac{GM+1}{2},\frac{s_{iw}^2+\sigma_w^2}{2}\right.\right)$$

$$s_{iw} = \sum_{k=1}^G(\boldsymbol{x}_{ik}-\boldsymbol{y}_i)(\boldsymbol{x}_{ik}-\boldsymbol{y}_i)'$$

$$s_j = \sum_{c_i=j}^T(\boldsymbol{y}_i-\boldsymbol{\mu}_j)(\boldsymbol{y}_i-\boldsymbol{\mu}_j)'\quad \bar{\boldsymbol{y}}_{j\bullet} = \frac{\sum_{c_i=j}^T\boldsymbol{y}_i}{n_j}$$

$$p(\boldsymbol{\mu}_j|\boldsymbol{c},\sigma_j^2,\boldsymbol{Y},\boldsymbol{\lambda},r) = f_N\left(\boldsymbol{\mu}_j\left|\frac{r^{-1}\bar{\boldsymbol{y}}_j+\frac{\sigma_j^2}{n_j}\boldsymbol{\lambda}}{r^{-1}+\frac{\sigma_j^2}{n_j}},\frac{r^{-1}\frac{\sigma_j^2}{n_j}}{r^{-1}+\frac{\sigma_j^2}{n_j}}I\right.\right)$$

$$p(\sigma_j^{-2}|\boldsymbol{Y},\boldsymbol{M},\beta,w) = f_G\left(\sigma_j^{-2}\left|\frac{Mn_j+\beta}{2},\frac{s_j^2+\beta w}{2}\right.\right).$$

Conditional posterior distributions for other parameters remain unchanged.