Clustering of time-course gene expression data using a mixed-effects model with B-splines

Yihui Luan¹, ² and Hongzhe Li¹,*

¹Rowe Program in Human Genetics, Department of Medicine, University of California, Davis, CA 95616-8500, USA and ²Department of Statistics, Peking University, Beijing, People's Republic of China

Received on April 20, 2002; revised on October 8, 2002; accepted on October 10, 2002

ABSTRACT

Motivation: Time-course gene expression data are often measured to study dynamic biological systems and gene regulatory networks. To account for time dependency of the gene expression measurements over time and the noisy nature of the microarray data, the mixed-effects model using B-splines was introduced. This paper further explores such mixed-effects model in analyzing the time-course gene expression data and in performing clustering of genes in a mixture model framework.

Results: After fitting the mixture model in the framework of the mixed-effects model using an EM algorithm, we obtained the smooth mean gene expression curve for each cluster. For each gene, we obtained the best linear unbiased smooth estimate of its gene expression trajectory over time, combining data from that gene and other genes in the same cluster. Simulated data indicate that the methods can effectively cluster noisy curves into clusters differing in either the shapes of the curves or the times to the peaks of the curves. We further demonstrate the proposed method by clustering the yeast genes based on their expression profiles over time or over different samples. Clear periodic patterns and varying times to peaks are observed for different clusters of the cell-cycle regulated genes. Results of the analysis of the human fibroblasts data show seven distinct transcriptional response profiles with biological relevance.

Availability: Matlab programs are available on request from the authors.

Supplementary Information: http://dna/ucdavis.edu/~hli/bioinforsupp.pdf.

Contact: hli@ucdavis.edu

INTRODUCTION

In an attempt to understand complex biological regulatory networks, both cDNA and oligonucleotide microarrays have been used to simultaneously measure the gene expression levels of thousands of genes. Because of the large number of genes involved and the complexity of biological processes, clustering of genes is one of the most commonly used statistical method for analyzing such data. Clustering genes with similar expression profiles can potentially be utilized to predict the functions of gene products with unknown functions, and to identify sets of genes that are regulated by the same mechanism. Popular non-model-based clustering methods include hierarchical clustering (Spellman et al., 1998), the graph-based CAST algorithm (Ben-Dor and Yakhini, 1999), self-organizing maps (Tamayo et al., 1999). Recently, clustering method based on multivariate normal mixture model, in which each gene is assumed to have come from a mixture of multivariate normal densities with different means and certain parameterization of the covariance matrix (Fraley and Raftery, 1998), was also used to cluster genes based on their expression profiles over time or over different samples (Yeung et al., 2001).

Since many biological systems are dynamic systems, temporal profiles of gene expression levels during a given biological process can often provide more insight about how gene expression levels evolve in time and how genes are dependent during a given biological process. Examples include the yeast cell cycle process (Spellman et al., 1998) and circadian rhythms (Claridge-Chang et al., 2001). One important feature of such time-course gene expression data is the possible dependency of gene expression levels across times for a given gene. In addition, as gene expression levels evolve over time, time can be an important factor that affects the gene expression levels. Methods which can preserve the time sequence and the time dependence of the observed data are needed for analyzing the time-course gene expression data, including methods for clustering analysis.

Motivated by the work of Rice and Wu (2001) and James and Suger (2002), Luan and Li (2002) and Li et al. (2002) developed the mixed-effects model for the time-course gene expression data using B-splines (De Boor, 1978), treating gene expression level as a continuous function...
of time. Bar-Joseph et al. (2002) independently developed the same model using the cubic splines. In this modelling framework, the observed time-course expression data are treated as samples taken from underlying continuous smooth processes. Such mixed-effects analysis represents repeated measures of each gene as the sum of the smooth population mean spline function dependent on time and gene cluster, a spline function with random coefficients for individual gene effects, and gaussian measurement noise. Under this mixed-effects model, the EM algorithm (Dempster et al., 1976) can be employed to cluster genes.

To further explore the mixed-effects model for the time-course gene expression data, in this paper, we present the results of a small simulation study to compare this method with the normal-mixture model based clustering method. We also present the results of the detailed analysis of two real data sets, including both periodic expression profiles with even-s spaced sampling times and aperiodic expression profiles with unevenly spaced sampling times. In addition, we demonstrate the use of the Bayesian inference criteria (BIC) (Schwarz, 1978) for determining the number of clusters in the data set. The rest of the paper is organized as follows: we first present the mixed-effects model for time-course gene expression data as defined in Luan and Li (2002) and Li et al. (2002). We then give some details on mixture model formulation of clustering and parameter estimation using the EM algorithm. We follow that by comparing the results of different methods using simulated data sets. We further demonstrate the method by analyzing the yeast cell cycle gene expression data set (Spellman et al., 1998) and the data set of the response of human fibroblasts to serum (Iyer et al., 1999). Finally, we conclude with a brief discussion.

MODELS AND METHODS

A mixed-effects model for the time-course gene expression data

Let \( Y_{ij} \) be the gene expression level of the \( i \)th gene at time \( t_{ij} \), for \( i = 1, \ldots, n, j = 1, \ldots, T \), where \( n \) is the number of genes, and \( T \) is the number of time points. Let \( Y_i = \{ Y_{i1}, \ldots, Y_{iT} \} \) be the vector of the time-course gene expression data for the \( i \)th gene. We assume that these \( n \) genes are from \( C \) different gene clusters, indexed by \( c = 1, \ldots, C \). Let \( Z_i \) be the cluster indicator for the \( i \)th gene, which takes one value from \( \{1, \ldots, C\} \). For the \( i \)th gene in the \( c \)th gene cluster, Luan and Li (2002) and Li et al. (2002) proposed the following mixed-effects model for the observed expression level at time \( t_{ij} \).

\[
Y_{ij} = \left( \sum_{l=1}^{p} \beta_l^{(c)} B_l(t_{ij}) \right) + \sum_{l=1}^{q} \gamma_{il} B_l(t_{ij}) + \epsilon_{ij} \quad (1)
\]

In this model, the first term is used to model the mean or population average gene expression profile of the \( c \)th cluster, where the sum is over a fixed knots sequence, and \( B = \{ B_l(), l = 1, \ldots, p \} \) is a basis for the spline function on \([t_1, t_T]\). We use the B-spline basis with equal spaced knots in this paper (De Boor, 1978). Here \( \beta_l^{(c)} \) is the \( p \)-vector of coefficients corresponding to the cluster \( c \). Note that we assume the same spline basis for all the \( C \) clusters. The second term in model (1) is used to model the random effect curve for the \( q \)th gene, where \( \gamma = \{ \gamma_{il}, l = 1, \ldots, q \} \) is a basis vector for a possibly different space of spline function on \([t_1, t_T]\), and \( \gamma_{il} \) are normal random coefficients with mean 0, and covariance matrix \( Cov(\gamma_{il}) = \Gamma \), varying cross genes. This term is used to model the gene-specific deviation from the ‘population’ average gene expression profile. Notice that in order to avoid too many parameters in the model, we assume that the gene-specific random curves are independent of the cluster indicators. Finally, the last term \( \epsilon_{ij} \) is used to model the uncorrelated normal measurement errors \( \epsilon_{ij} \), with \( E(\epsilon_{ij}) = 0 \), \( Var(\epsilon_{ij}) = \sigma^2 \).

The random-effects term induces the following covariance kernel for the random curve \( Y_i(t) \) at the two different time points \( t \) and \( s \),

\[
Cov(Y(s), Y(t)) = \sum_{k=1}^{q} \sum_{l=1}^{q} \Gamma_{kl} B_k(s) B_l(t) + \sigma^2 \delta(s-t),
\]

where \( \delta() \) is the Dirac function. Note that model (1) is a conditional model (i.e., conditioning on the unknown gene cluster \( c \)). This is a generalization of the model used in Rice and Wu (2001) and a special case of the model by James and Sugar (2002) for modelling noisy curves. Bar-Joseph et al. (2002) independently developed a similar model using the same cubic spline bases for both mean and random effects.

Clustering based on a mixture model using an EM algorithm

For the purpose of clustering genes, the cluster indicator vector \( Z = \{ Z_1, Z_2, \ldots, Z_n \} \) for the \( n \) genes is unknown. Treating the clustering problem as a mixture model problem, we assume that the observed gene expression levels over time come from a mixture of \( C \) probability distributions with the \( c \)th probability defined as in model (1). The EM algorithm can be used for estimating the parameters associated with this mixture model treating the vector \( Z \) and the random coefficients \( \gamma = \{ \gamma_1, \ldots, \gamma_n \} \) as missing data. We further assume that conditioning on the cluster membership, the gene expression profiles are independent across all the genes considered. Let \( \pi_c \) be the unknown prior probability that a given gene belongs to cluster \( c \), so that \( \sum_{c=1}^{C} \pi_c = 1 \). The EM algorithm involves calculation of the expected values of the complete data log-likelihood, and the maximization of the expected
complete data log-likelihood over the model parameters. Details of the derivations and the closed forms of the expectations and the EM equations for updating the model parameters are given in Luan and Li (2002) and can be found in the web supplementary materials.

After the convergence of the EM iterations, we obtain the parameter estimates of $\hat{\beta}_j^{(c)}$, $\hat{\gamma}$, $\delta^2$, and also the posterior probability of the $i$th gene belonging to the $c$th cluster, $\tilde{\pi}_{c|i} = E[I(Z_i = c)]$ for $i = 1, 2, \ldots, n$, and $c = 1, 2, \ldots, C$. Based on these probability estimates, we then cluster the $i$th gene into the $c$th cluster if $\pi_{c|i}$ is the largest for $c = 1, \ldots, C$. The $\tilde{\pi}_{c|i}$ also provides a measurement of uncertainty about the resulting clustering. We can also only cluster those genes with the max$_i \pi_{c|i}$ greater than a pre-specified cutoff value, and declare the other genes as not belonging to any of the $C$ clusters.

**Determination of the number of the clusters**

In the mixed-effects model framework, we can compute the BIC scores for selecting the number of clusters. BIC, which can be regarded as an approximation to the Bayes factor, is defined by

$$ \text{BIC}(C) = 2L(C) - m_C \log n, $$

where $L(C)$ is the maximized log-likelihood for the model with $C$ clusters, $m_C$ is the number of independent parameters to be estimated in the $C$-clusters model and $n$ is the sample size. Although the regularity conditions for BIC do not hold for mixture models, there is growing support for the use of BIC in the context of mixture models for analysis of microarray data (Yeung et al., 2001; Fraley and Raftery, 2002). As suggested by Fraley and Raftery (2002), in practice, one should pick a good candidate in the region where the rate of change of BIC drops significantly.

**Estimation of gene expression trajectory**

After the number of clusters is determined and the parameter estimates are obtained, for each gene, we can obtain the estimate of the gene expression trajectory using gene expression data from the same cluster. For the $i$th gene in the $c$th cluster, the best linear unbiased predictor or BLUP (Robinson, 1991) of the spline coefficients of the random effects is given by

$$ \hat{\gamma}_i = E(\gamma_i|Y, Z_i = c) = \hat{\gamma} B'(\hat{\beta}' B + \delta^2 I)^{-1}(Y_i - \hat{\beta}^{(c)}), $$

(see Luan and Li, 2002). The corresponding estimate of the individual gene expression trajectory for gene $i$ in the $c$th cluster at time $t$ is then the smooth curve

$$ \hat{Y}_i(t) = \left( \sum_{l=1}^{p} \hat{\beta}_l^{(c)} B_l(t) \right) + \sum_{l=1}^{q} \hat{\gamma}_{il} B_l(t). $$

This estimation is different from the approaches that only use the data of a single gene. Since the microarray data are often measured with large amount of noises, by pooling data from the same cluster, we expect to obtain more reliable estimate of the gene expression profile than the methods which utilize only the single gene data. The estimate of the gene expression trajectory can be used for estimating the gene expression level at any given time $t$ and for estimating the times to peak gene expression levels.

**RESULTS**

**Simulated data examples**

To demonstrate the proposed methods, we first present examples of simulated data sets. In the following simulation, we chose $C = 4$ and simulated 200 observations measured evenly over 18 time points for each cluster. For all the simulations, the initial values for the EM iterations are chosen as follows: first, simple linear regression is fitted using the data and the B spline design matrix while ignoring the random effects. The cluster specific $\beta$ are chosen by adding a random vector whose entries vary between -1 and 1. Second, variance of the residuals from the linear fit is used as the initial value for noise variance. The covariance matrix of random effect is initialized as unity matrix times $\sigma^2$, which is the initial noise variance. Finally, the prior for membership probabilities is chosen from a uniform distribution. The EM iterations are stopped when the difference of successive parameter estimates is less than 0.0001 for all parameters.

**SIMULATION 1.** For the first simulation, we simulated the data from the true model (1), with the following cluster-specific parameters for the mean curves,

$$ \begin{pmatrix} \beta^{(1)} \\ \beta^{(2)} \\ \beta^{(3)} \\ \beta^{(4)} \end{pmatrix} = \begin{pmatrix} -2.3 & -0.5 & 1.5 & -1.0 & 0.5 & 0.4 & -0.8 & 1.2 \\ -0.5 & -0.3 & -1.5 & 1.7 & -0.5 & -0.4 & 1.2 & 0.5 \\ 2.5 & -1.5 & 2.4 & -0.6 & -1.3 & 0.5 & 1.5 & -1.0 \\ 1.2 & -2.5 & 0.8 & -0.6 & -1.4 & 1.5 & -0.4 & 2.4 \end{pmatrix}. $$

Cubic B-splines with four equally spaced knots are used for both mean and random effects. For the random effects, we generated data from a multivariate normal distribution with correlations between the time points ranging from -0.44 to 0.53, and the variances ranging from 0.21 to 0.61. We chose $\sigma^2 = 0.25$. The solid lines in the left panel of Figure 1 show the true mean curves for each of the four clusters, and the curves in the right panel are the simulated 200 curves for each of the four corresponding clusters. Distinguished patterns of expression over time and large amount of noises can be seen in different clusters.

The BIC score plot obtains its maximum value at $C = 4$, indicating four different clusters (see Web supplement Figure S.1). The dashed lines in the left panel of Figure 1
show the estimated smooth mean curves. It can be seen that the estimated smooth mean curves agree with the true mean curves well. The correct clustering rates range from 193/200 to 196/200 (see web supplement Table S.1).

SIMULATION 2. In order to demonstrate that the B-spline basis functions also work well for periodic expression patterns, for the second simulation, we generated data from the following model.

\[
Y(t) = (\beta_1^{(c)} + \gamma_1) + (\beta_2^{(c)} + \gamma_2) \cos\left(\frac{2\pi t}{T}\right) + (\beta_3^{(c)} + \gamma_3) \sin\left(\frac{2\pi t}{T}\right) + \epsilon
\]

for the \(c\)th cluster, where \(c = 1, 2, 3, 4\),

\[
\begin{pmatrix}
\beta_1^{(1)} \\
\beta_2^{(1)} \\
\beta_3^{(1)} \\
\beta_4^{(1)}
\end{pmatrix} = \begin{pmatrix}
0 & 0.5 & 0.87 \\
0 & -0.81 & 0.59 \\
0 & -0.81 & -0.59 \\
0 & 0.31 & -0.95
\end{pmatrix},
\]

and \(\{\gamma_1, \gamma_2, \gamma_3\}\) is the vector of random effects, generated from a multivariate normal distribution with variances of 0.065, 0.073, 0.17, and pair-wise correlations of 0.70, -0.17, -0.62. We chose \(\sigma^2 = 0.09\). Again 200 curves are simulated for each cluster. This model simulates curves in different clusters with the same shape, but different time to peaks. The BIC score obtains its maximum values at \(C = 4\), indicating that there are four different clusters (see Web supplement Figure S.1). Again, we found that the fitted mean curves are very close to the true mean curves (details not shown). The correct clustering rates range from 188/200 to 200/200. This example demonstrates that the cubic B-spline basis functions work well for modelling periodic curves such as the sine or cosine curves that we simulated in this example.

As a comparison, we also performed the clustering analysis using the normal mixture model with different specifications of the covariance matrix using the \textit{mclust} function as provided in the Splus package. Fraley and Raftery (1998) proposed different parameterizations for the covariance matrix, \(\Sigma_c = \lambda_c D_c A_c D_c^T\), including the model with \(\Sigma_c = \lambda I\), the model with \(\Sigma_c = \lambda_c I\), the model with \(\Sigma_c = \lambda_c D_c A_c D_c^T\), and the model with \(\Sigma_c = \lambda_c D_c A D_c^T\). We found that our proposed method outperformed these normal-mixture models in clustering observations into their respective true clusters. The clustering results based on these models for both simulated data sets are given in the web supplement (Table S.1).
Using the proposed method, we classified these genes into five clusters. We found that the five clusters obtained by Spellman et al. (1998) show clear periodic patterns for times to peak expression levels (in minutes) for the cell-cycle regulated yeast genes.

<p>| Table 1. Distribution of the 612 cell-cycle regulated genes of five different phases defined by Spellman et al. (1998) over the five estimated genes clusters using the proposed method |
|---------------------------------|---------------|-------------|-------------|---------------|-------------|</p>
<table>
<thead>
<tr>
<th>Cluster</th>
<th>M/G₁</th>
<th>G₁</th>
<th>S</th>
<th>S/G₂</th>
<th>G₂/M</th>
</tr>
</thead>
<tbody>
<tr>
<td>I (62)</td>
<td>38</td>
<td>12</td>
<td>0</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>II (228)</td>
<td>40</td>
<td>176</td>
<td>5</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>III (97)</td>
<td>5</td>
<td>27</td>
<td>28</td>
<td>23</td>
<td>14</td>
</tr>
<tr>
<td>IV (176)</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>59</td>
<td>102</td>
</tr>
<tr>
<td>V (49)</td>
<td>7</td>
<td>3</td>
<td>6</td>
<td>8</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>92</td>
<td>223</td>
<td>47</td>
<td>92</td>
<td>158</td>
</tr>
</tbody>
</table>

<p>| Table 2. A summary (mean and standard deviation) of the estimates of times to peak expression levels (in minutes) for the cell-cycle regulated yeast genes in four different clusters |
|---------------------------------|---------------|-------------|-------------|---------------|</p>
<table>
<thead>
<tr>
<th>Cluster</th>
<th>Time to peak expression, mean(sd)</th>
<th>First peak</th>
<th>Second peak</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>9.9 (3.6)</td>
<td>76.9 (6.6)</td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>18.4 (2.0)</td>
<td>81.8 (6.6)</td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>46.3 (2.9)</td>
<td>100.5 (5.2)</td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>51.7 (7.4)</td>
<td>113.1 (3.2)</td>
<td></td>
</tr>
</tbody>
</table>

**Application to the yeast cell cycle data set**

We first applied the method to a data set with equally spaced sampling time points. Using cDNA arrays, Spellman et al. (1998) measured the genome-wide mRNA levels for 6108 yeast ORFs simultaneously over approximately two cell-cycle periods in a yeast culture synchronized by α factor relative to a reference mRNA from an asynchronous yeast culture. The yeast cells were sampled at 7 min intervals for 119 min with a total of 18 time points after synchronization. Among these genes, about 800 are characterized as cell cycle regulated genes by Spellman et al. (1998). Among these 800 genes, 798 genes had gene expression data available from the web site and 612 genes had no missing gene expression data across all the 18 time points. In the following analysis, we used these 612 genes for illustrating the method. Spellman et al. (1998) classified these genes into five different cell-cycle phases, M/G₁, G₁, S, S/G₂ and G₂/M phases.

Since we are mainly interested in the differences of the phases at which the gene expression levels approach peak values, we first normalized the expression data by subtracting the mean and dividing by the standard deviation for each gene. To compare the results with the five clusters obtained by Spellman et al. (1998), we first cluster these genes into five clusters. We found that these five clusters have different peak times, and four clusters show clear periodic patterns (see Web supplement Figure S.2 for plots). Table 1 presents the distribution of the genes in each cluster in terms of the five cell-cycle phases defined by Spellman et al. (1998). Most of the genes in a given cluster are from one phase group or two nearby phase groups. Genes in cluster I are mostly those expressed in M/G₁ and G₁ phases. Genes in cluster II are mostly those expressed in the G₁ phase. Cluster III includes genes expressed in S and S/G₂ phases, and cluster IV includes genes expressed in S/G₂ and G₂/M phases. Finally, cluster V mainly includes genes expressed in G₂/M phase. This indicates that the clustering results indeed have biological relevance.

We further obtained the estimated smooth gene expression trajectories for genes in clusters I-IV (see web supplement Figure S.3). For each cluster, the estimated individual gene expression profiles deviate from the mean expression profile in either the horizontal or the vertical directions. Large variation of gene expression trajectories are also observed for genes in the same cluster. Based on these estimated gene expression curves, we can obtain estimates of the times at the peak expression for each gene, indicating five clusters might not be the optimal number of the clusters. Table 2 gives a summary of the estimates for times to the two peak expression levels for genes in the first four clusters. The times to peak expressions are clearly separated among the four different clusters and the estimated times between the two peaks are 67.0, 63.4, 54.2 and 61.4 for the four clusters, respectively.

We next used the BIC score to determine the number of clusters in the data set. The BIC score plot (Fig. 2 plot I) indicates that there are seven clusters, although the difference of BIC scores for five and seven clusters is not large. Plots I–VII of Figure 2 show the average of the observed time-course gene expression levels (in log scale) and the estimated mean gene expression profile for each of the seven gene clusters. The closeness of the fitted and the observed values suggests that the model fits the data well. Again, except for the cluster III, clear periodic expression patterns with different phases are observed across different clusters.

**Application to the response of human fibroblasts to serum data set**

Iyer et al. (1999) reported the temporal program at 12 time points of gene expression during the physiological response of fibroblasts to serum using cDNA microarrays representing 8613 genes over 24 h. They further identified 517 genes whose expression levels changed in response to serum stimulation. The sampling times are 0, 0.25, 0.15, 1, 2, 4, 6, 8, 12, 16, 20, 24 h after serum stimulation. The purpose of this application is to demonstrate the applicability of the methods to unevenly sampled time points by clustering these 517 genes.
The BIC score plot indicates that these 517 genes can be clustered into seven different gene clusters, as compared to ten groups of genes defined by visual inspection by Iyer et al. (1999). Plots I–VII of Figure 3 show the average of the observed time-course gene expression levels (in log scale) and the estimated mean gene expression profile for each of the seven gene clusters, each with distinct gene expression pattern. The closeness of the fitted and the observed values suggests that the model fits the data well. Cluster I (44 genes) and cluster III (six genes) have similar expression profile, but the peak expression level of genes in cluster III is higher. Genes in these two clusters are induced hours after the serum stimulus, including genes related to inflammation (e.g. IL6, IL8, IL1β, ICAM1, and MIP2α), and genes related to re-epithelization (e.g. FGF2, FGF7) (Iyer et al., 1999). Cluster II (24 genes) includes genes with rapid induction in 15 min after serum stimulation (e.g. c-FOS, JUN B, and MKP1). These genes are likely to play a role in signal transduction or acting as immediately-early transcription factors (Iyer et al., 1999). Genes in cluster VI (36 genes) show sharply diminished expression 4–6 h after serum stimulation. This cluster includes genes related to cholesterol biosynthesis (e.g. lipoprotein lipase, IPP isomerase, apolipoprotein D). Genes in cluster VII (26 genes) were induced at about 16 to 20 h after serum addition. This cluster include regulators of passage through the S phase and the transition from G2 to M phase (e.g. Cyclin A, Cyclin B1, CDC28 kinase, cell division cycle 2). The largest cluster is cluster IV, including 285 genes. This cluster includes most of the genes in clusters A and B defined by Iyer et al. (1999). The nadir in the expression levels of genes in this cluster occurred between 6 and 12 h after serum stimulation, before onset of M phase. Genes in this cluster include p27Kip1, p57Kip2, CDK6 inhibitor p18, and cyclin-dependent kinase inhibitor 1, and the transcript encoding the WEE1-like protein kinase. The gene products of these genes inhibit progression of cell division cell cycle (Iyer et al., 1999). Genes in cluster V (90 genes) had their expression induced and remained for longer period. Function of most of the genes in this cluster is unknown.

To get an idea how certain our clustering result is, plot VIII of Figure 3 plots the maximum of the posterior
probabilities of cluster memberships for each of the 517 genes in the seven clusters. Higher probabilities of most of the genes indicate high certainty on our clustering results.

**DISCUSSION AND CONCLUSIONS**

The analysis of time-course gene expression data can potentially provide more insights about the dynamic biological systems. In this paper, we have further studied the methods of clustering genes based on the time-course gene expression profiles in the framework of a mixture model using the mixed-effects model with B-splines. We have compared the methods with the mixture-normal based methods by simulations. Simulated data indicate that our method works well in clustering noisy curves into respective clusters which are different in terms of either curve shapes or times to expression peaks, and the method outperforms the normal-mixture model-based method in correctly clustering samples into their respective true clusters. We further demonstrated the methods using two real data sets, a real yeast cell cycle time-course gene expression data which are sampled at even time points and show periodic expression patterns, and a real temporal responses of fibroblasts to serum which are sampled at unevenly-spaced time points. For both data sets, the biological relevance of the results can be recognized.

Recently, Ramoni et al. (2002) proposed a method for clustering genes by assuming an autoregressive (AR) model for time course gene expression data in a Bayesian framework, and applied their method to the response of human fibroblasts to serum data set of Iyer et al. (1999). It worths comparing these two methods. Comparing to the B-spline based clustering methods presented in this paper, the clustering method based on the AR model proposed by Ramoni et al. (2002) has the following limitations. First, simple low order AR model is not appropriate for modelling possible non-linear relationship between the gene expression levels at different time points. Second, when the gene expression are measured at unevenly-spaced time points, the interpretation of the coefficients in the AR models is not very clear. Third, it is well known that simple AR model as used in Ramoni et al. cannot be used for modelling time-trend, which can sometimes be precisely what differentiates among different gene clusters. In contrast, the B-spline based method for clustering studied in this paper provides flexible way of modelling possible non-linear relationship between gene
expression levels and works equally well for time-course data measured at unevenly spaced time points. Finally, it is also interesting to compare the results of analysis of the human fibroblasts data set obtained by these two methods. Ramoni et al. (2002) clustered the 517 genes into four clusters, including two clusters of three and five genes respectively. In contrast, the B-spline based method identified seven clusters of distinct expression profiles (see Fig. 3) with the smallest cluster having six genes which include the three genes in cluster 1 identified by Ramoni et al. (2002). Our clustering results are more close to those originally obtained by Iyer et al. (1999).

There are several issues that deserve further study. First, we used the BIC in this paper for estimating the number of clusters. This method seems to work well for our simulated data sets and the two real data sets we considered. Determining the number of the clusters is still an active research. Methods such as those based on the predictivestrength or re-sampling (Tibshirani et al., 2001) can be similarly adapted for our proposed mixed-effects models. Second, current approach clusters all observations into several clusters. However, it is possible that some observations, such as noises and outliers, do not belong to any clusters. We can handle these observations in the same framework by adding a term or terms to the mixture to represent ‘nonconforming’ data (Fraley and Raftery, 2002). An alternative is to examine the final estimates of the cluster probabilities and to only cluster those observations with high probabilities. We can also isolate these outliers by iterated sampling, in which observations with low probabilities are removed from clusters and the clustering/removal process is repeated until all remaining observations have high probabilities (e.g. Fraley and Raftery, 2002). Third, for practical application, the number and locations of the knots for the B-splines corresponding to the mean function and the random effects have to be specified. Small number of measurements for typical time-course data limits the numbers of knots to only small numbers. From our analysis of simulated and real data sets, we observed that the results are rather insensitive to the specification; a relatively small number of equally-spaced knots has generally been sufficient. More formal guidance using BIC or AIC or cross-validation can be employed for the proposed model. We are investigating these possibilities. Finally, in order to calculate the complete data log-likelihood, we made the assumption of conditional independence of the gene expression profiles conditioning on the cluster membership variables. Violation of such assumption can result in smaller variance estimates. However, since the inferences on the model parameters are not the main goal of the proposed methods, mild violation of this assumption should not greatly affect the clustering results.

In summary, we have conducted further studies of the mixed-effects model using B-splines for clustering genes based on the time-course gene expression data. Results from the clustering analysis can be used for further downstream analysis such as identifying gene regulatory motifs (Tavazoie et al., 1999). Furthermore, the estimated gene expression trajectory provides estimates of missing gene expression levels for any time point utilizing all the data in the same cluster, not just the gene expression data of this particular gene with missing data points (Li et al., 2002; Bar-Joseph et al., 2002). Bar-Joseph et al. (2002) also presented some other interesting applications in this modelling framework. The mixed-effects model can potentially be extended to a regression setting in order to identify DNA sequence motifs in the promoter regions which determine the time-course gene expression profiles, and eventually help in identifying gene regulatory networks.

Acknowledgements

We thank for the referees for many helpful comments which greatly improved the presentation of this paper.

REFERENCES


