Multilevel modeling for inference of genetic regulatory networks

Shu-Kay Ng[a], Kui Wang[b] and Geoffrey J. McLachlan[a,b,c]

[a]Department of Mathematics, University of Queensland, Brisbane, QLD 4072, Australia;
[b]ARC Centre for Complex Systems, University of Queensland, Brisbane, QLD 4072, Australia;
[c]Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD 4072, Australia

ABSTRACT

Time-course experiments with microarrays are often used to study dynamic biological systems and genetic regulatory networks (GRNs) that model how genes influence each other in cell-level development of organisms. The inference for GRNs provides important insights into the fundamental biological processes such as growth and is useful in disease diagnosis and genomic drug design. Due to the experimental design, multilevel data hierarchies are often present in time-course gene expression data. Most existing methods, however, ignore the dependency of the expression measurements over time and the correlation among gene expression profiles. Such independence assumptions violate regulatory interactions and can result in overlooking certain important subject effects and lead to spurious inference for regulatory networks or mechanisms. In this paper, a multilevel mixed-effects model is adopted to incorporate data hierarchies in the analysis of time-course data, where temporal and subject effects are both assumed to be random. The method starts with the clustering of genes by fitting the mixture model within the multilevel random-effects model framework using the expectation-maximization (EM) algorithm. The network of regulatory interactions is then determined by searching for regulatory control elements (activators and inhibitors) shared by the clusters of co-expressed genes, based on a time-lagged correlation coefficients measurement. The method is applied to two real time-course datasets from the budding yeast (Saccharomyces cerevisiae) genome. It is shown that the proposed method provides clusters of cell-cycle regulated genes that are supported by existing gene function annotations, and hence enables inference on regulatory interactions for the genetic network.

Keywords: EM algorithm, mixture models, multilevel mixed-effects model, genetic regulatory networks, time-course data

1. INTRODUCTION

In recent times, there has been an explosion in the development of comprehensive, high-throughput methods for molecular biology experimentation. The advent in DNA microarray technologies, such as complementary DNA (cDNA) arrays and oligonucleotide arrays, provides means for measuring tens of thousands of genes simultaneously under different conditions. The complete description of the human genome and those of other organisms has been a major achievement of modern science. Microarray technologies promise to revolutionize our approaches in biomedical research and further our understanding of biological processes and the evaluations of gene expression patterns, regulation, and interactions.[1] The study of the dynamics of gene interactions is amongst the latest research directions in the post-genomic era in biology. It is well-known that the information encoded in DNA leads to the expression of certain phenotypes, or characteristics, in the organism.[2] The determination of genetic regulatory networks (GRNs) provides useful information on how genes influence each other in cell-level development of organisms. It can therefore provide important insights for the fundamental biological processes such as growth and is useful in disease diagnosis and genomic drug design.[2, 3] Time-course experiments with microarrays are used to measure gene expressions at several points in time as a serial study.

Further author information: (Send correspondence to S.K. Ng)
S.K. Ng: E-mail: skn@maths.uq.edu.au, Telephone: +617 3365 6139
K. Wang: E-mail: kwang@maths.uq.edu.au, Telephone: +617 3346 2623
G.J. McLachlan: E-mail: gjm@maths.uq.edu.au, Telephone: +617 3365 2150

For example, Cho et al.[4] published a 17-point time series data set measuring the expression levels of 6,220 different genes from the budding yeast (Saccaromyces cerevisiae) genome. These temporal profiles of expression levels reflect genes interactions in pathways. The inference from these time-course gene expression data thus allows researchers to explore genes interactions from a causal perspective. This inference procedure is referred to as "reverse engineering", which attempts to determine the cause (the network of genes interactions) from the outcome (the observed effects on the gene expression profiles).[5]

As genes sharing the same expression pattern are likely to be involved in the same regulatory process, the inference of gene regulation can be accomplished via the clustering of time-course data into groups of co-expressed genes.[5,6] However, due to the experimental design, multilevel data hierarchies are often present in time-course gene expression data. For example, gene expressions obtained at the same time point are often interdependent and tend to be more alike in characteristics than data chosen at random from the population as a whole. Most existing methods, however, ignore the dependency of the expression measurements over time and the correlation among gene expression profiles.[4,7–9] Such assumptions of independence violate regulatory interactions and can result in overlooking certain important subject and temporal effects, and lead to spurious inference for regulatory networks or mechanisms. In this paper, we propose a clustering-based method for the inference of GRNs, where a mixed-effects model is adopted to incorporate data hierarchies in the clustering of time-course data. In the context of multilevel analysis,[10] the temporal (subject) effects are assumed to be random (random effects) and shared among data collected at the same time/condition. The method starts with the clustering of genes by fitting the mixture model within the multilevel random-effects model framework using the expectation-maximization (EM) algorithm.[11,12] The clustering of time-course data allows us to extract groups of genes that are co-expressed over certain cell-cycle periods and hence infer shared regulatory inputs and functional pathways. The network of regulatory interactions is then determined by searching for regulatory control elements (activators and inhibitors) shared by the clusters of co-expressed genes, based on a time-lagged correlation coefficients measurement method.[13] The method is applied to two real time-course datasets from the budding yeast (Saccharomyces cerevisiae) genome.

2. CLUSTERING OF CO-EXPRESSED GENES

The method starts with the clustering of genes into groups in which they have similar expression profiles. This process helps to identify sets of genes that are potentially regulated by the same mechanism.[3] Other potential applications of co-expressed gene clusters include the inference of functional annotation and the identification of molecular signatures that are potential predictors of disease outcome.[5,14] As described in Section 1, popular clustering methods adopted for the analysis of gene expression data, such as hierarchical agglomerative techniques[8] and self-organizing maps,[15] do not account for the correlation among gene expression levels. In this paper, we consider the clustering of genes by fitting a normal mixture model within the multilevel mixed-effects model framework.

2.1. Normal Mixtures and Mixed-effects Models

Let y_{ij} be the gene expression level of the ith gene at time j ($i = 1, \dots, N$; $j = 1, \dots, T$), where N is the number of genes and T is the number of time points. The time-course microarray data are thus given by $\boldsymbol{y} = (\boldsymbol{y}'_1, \dots, \boldsymbol{y}'_N)'$, where $\boldsymbol{y}_i = (y_{i1}, \dots, y_{iT})'$ is the vector of the time-course gene expression data for the ith gene. With a normal mixture model, the observed data are assumed to have come from a mixture of an initially specified number g of multivariate normal densities in some unknown proportions $\pi_1, \dots, \pi_g$, which sum to one.[16] That is, each observed vector $\boldsymbol{y}_i$ is taken to be a realization of the mixture probability density function,

$$f(\boldsymbol{y}_i; \boldsymbol{\Psi}) = \sum_{h=1}^{g} \pi_h \phi(\boldsymbol{y}_i; \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h), \tag{1}$$

where $\phi(\boldsymbol{y}_i; \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)$ denotes the T-dimensional multivariate normal density with mean $\boldsymbol{\mu}_h$ and covariance matrix $\boldsymbol{\Sigma}_h$. The vector $\boldsymbol{\Psi}$ denotes unknown parameters in the model.

The extension of normal mixtures to incorporate data hierarchies via multilevel (linear) mixed-effects models is formulated as follows. Let the vectors $\boldsymbol{\nu} = (\boldsymbol{\nu}'_1, \dots, \boldsymbol{\nu}'_g)'$ and $\boldsymbol{\xi} = (\boldsymbol{\xi}'_1, \dots, \boldsymbol{\xi}'_g)'$ denote the random effects that occur in the data vector $\boldsymbol{y}$. Conditional on its membership of the hth component of the normal mixture, the linear mixed-effects model[17] specifies the mean of $\boldsymbol{y}_i$ conditional on the realized $\boldsymbol{\nu}_h$ and $\boldsymbol{\xi}_h$ as

$$E(\boldsymbol{y}_i \mid \boldsymbol{\nu}_h, \boldsymbol{\xi}_h) = \boldsymbol{X}\boldsymbol{\beta}_h + \boldsymbol{V}\boldsymbol{\nu}_{hi} + \boldsymbol{W}\boldsymbol{\xi}_h, \tag{2}$$

where elements of $\boldsymbol{\beta}_h$ are fixed effects (unknown constants) modeling the conditional mean of $\boldsymbol{y}_i$. The vectors $\boldsymbol{\nu}_{hi}$ (p-dimensions) and $\boldsymbol{\xi}_h$ (q-dimensions) represent the unobservable random effects which have zero mean ($E(\boldsymbol{\nu}_{hi}) = \boldsymbol{0}$; $E(\boldsymbol{\xi}_h) = \boldsymbol{0}$) and are taken, respectively, to be i.i.d. $N(\boldsymbol{0}, \theta_{h1}\boldsymbol{I}_p)$ and $N(\boldsymbol{0}, \theta_{h2}\boldsymbol{I}_q)$, where $\boldsymbol{I}_p$ and $\boldsymbol{I}_q$ are identity matrices with dimensions being specified by the subscripts. Also, it is assumed that the random effects $\boldsymbol{\nu}_h$ and $\boldsymbol{\xi}_h$ are mutually independent ($h = 1, \ldots, g$). In Equation (2), $\boldsymbol{X}$, $\boldsymbol{V}$, and $\boldsymbol{W}$ are known design matrices of the fixed effects and random effects parts, respectively. Under this formulation, the vector of unknown parameters $\boldsymbol{\Psi}$ now consists of $\pi_1, \ldots, \pi_{g-1}$, $\boldsymbol{\beta}_h$, θ_{h1}, θ_{h2}, and σ_h^2 for $h = 1, \ldots, g$, where σ_h^2 is the hth component-variance. McLachlan and Krishnan[18] described the learning of single component mixed-effects models via the EM algorithm. This approach can be extended to the present context where a normal mixture of mixed-effects model is to be learned.

2.2. Learning via the EM Algorithm

The EM algorithm is a popular tool for iterative maximum likelihood (ML) estimation of mixture models.[16] It has a number of desirable properties including its simplicity of implementation and reliable global convergence.[12,18] Within the EM framework, each $\boldsymbol{y}_i$ is conceptualized to have arisen from one of the g components. We let $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N$ denote the unobservable component-indicator vectors, where the hth element z_{hi} of $\boldsymbol{z}_i$ is taken to be one or zero according as $\boldsymbol{y}_i$ does or does not come from the hth component. We put $\boldsymbol{z} = (\boldsymbol{z}_1', \ldots, \boldsymbol{z}_N')'$.

With the use of the EM algorithm to learn mixtures of mixed-effects model, the unobservable component indicator variables $\boldsymbol{z}$ and the random effects $\boldsymbol{\nu}$ and $\boldsymbol{\xi}$ are treated as missing data in the EM framework. The complete data are then given by $(\boldsymbol{y}', \boldsymbol{z}', \boldsymbol{\nu}', \boldsymbol{\xi}')'$. On each iteration of the EM algorithm, there are two steps called the expectation (E) step and the maximization (M) step. On the $(k+1)$th iteration, the E-step computes the so-called Q-function, which is the conditional expectation of the complete-data log likelihood, given the observed data $\boldsymbol{y}$ and the current estimate $\boldsymbol{\Psi}^{(k)}$. With reference to Eqs. (1) and (2), it follows that the E-step involves the calculation of the following conditional expectations

$$E_{\boldsymbol{\Psi}^{(k)}}(z_{hi}|\boldsymbol{y}),\ E_{\boldsymbol{\Psi}^{(k)}}(\boldsymbol{\nu}_{hi}|\boldsymbol{y}),\ E_{\boldsymbol{\Psi}^{(k)}}(\boldsymbol{\xi}_h|\boldsymbol{y}),\ E_{\boldsymbol{\Psi}^{(k)}}(\boldsymbol{y}_i - \boldsymbol{X}\boldsymbol{\beta}_h - \boldsymbol{V}\boldsymbol{\nu}_{hi} - \boldsymbol{W}\boldsymbol{\xi}_h|\boldsymbol{y}). \tag{3}$$

The conditional expectations in Equation (3) are directly obtainable as shown in Appendix A. In particular, the conditional expectation, $E_{\boldsymbol{\Psi}^{(k)}}(z_{hi}|\boldsymbol{y})$, is given by

$$E_{\boldsymbol{\Psi}^{(k)}}(z_{hi}|\boldsymbol{y}) = \tau_{hi}^{(k)} = \frac{\pi_h^{(k)} f(\boldsymbol{y}_i|z_{hi} = 1; \boldsymbol{\Psi}^{(k)})}{\sum_{l=1}^{g} \pi_l^{(k)} f(\boldsymbol{y}_i|z_{li} = 1; \boldsymbol{\Psi}^{(k)})}, \tag{4}$$

which is the current estimated posterior probability that $\boldsymbol{y}_i$ belongs to the hth component, and where

$$\log f(\boldsymbol{y}_i|z_{hi} = 1; \boldsymbol{\Psi}^{(k)}) = -\tfrac{1}{2}\Big\{ \log|\boldsymbol{A}_h^{(k)} + \boldsymbol{B}_h^{(k)}| \ + (\boldsymbol{y}_i - \boldsymbol{X}\boldsymbol{\beta}_h^{(k)})'[\boldsymbol{A}_h^{(k)} + \boldsymbol{B}_h^{(k)}]^{-1}(\boldsymbol{y}_i - \boldsymbol{X}\boldsymbol{\beta}_h^{(k)}) \Big\}, \tag{5}$$

apart from an additive constant, is the marginal (log) density of $\boldsymbol{y}_i$ given that it belongs to the hth component. In Equation (5), we have

$$\boldsymbol{A}_h^{(k)} = \sigma_h^{2(k)}\boldsymbol{I}_T + \theta_{h1}^{(k)}\boldsymbol{V}\boldsymbol{V}'$$

and

$$\boldsymbol{B}_h^{(k)} = \theta_{h2}^{(k)}\boldsymbol{W}\boldsymbol{W}'.$$

The M-step updates the estimates that maximize the Q-function with respect to $\boldsymbol{\Psi}$. It follows from Eqs. (1) and (2) that the updating formulae for $\boldsymbol{\Psi}^{(k+1)}$ are

$$\pi_h^{(k+1)} = \sum_{i=1}^{N} \tau_{hi}^{(k)}/N, \tag{6}$$

$$\boldsymbol{\beta}_h^{(k+1)} = \boldsymbol{\beta}_h^{(k)} + \sigma_h^{2\,(k)} \boldsymbol{D}_h \sum_{i=1}^{N} \tau_{hi}^{(k)} (\boldsymbol{y}_i - \boldsymbol{X}\boldsymbol{\beta}_h^{(k)}) / \sum_{i=1}^{N} \tau_{hi}^{(k)}, \tag{7}$$

$$\theta_{h1}^{(k+1)} = (\sum_{i=1}^{N} \tau_{hi}^{(k)} \boldsymbol{\nu}_{hi}^{(k)\prime} \boldsymbol{\nu}_{hi}^{(k)} + E_h)/(p \sum_{i=1}^{N} \tau_{hi}^{(k)}), \tag{8}$$

$$\theta_{h2}^{(k+1)} = (\boldsymbol{\xi}_h^{(k)\prime} \boldsymbol{\xi}_h^{(k)} + F_h)/q, \tag{9}$$

and

$$\sigma_h^{2\,(k+1)} = (\sum_{i=1}^{N} \tau_{hi}^{(k)} \boldsymbol{\epsilon}_{hi}^{(k)\prime} \boldsymbol{\epsilon}_{hi}^{(k)} + G_h)/(T \sum_{i=1}^{N} \tau_{hi}^{(k)}), \tag{10}$$

where

$$\boldsymbol{D}_h = [\boldsymbol{X}'\boldsymbol{X}]^{-1} \boldsymbol{X} [\boldsymbol{A}_h + \sum_{i=1}^{N} \tau_{hi}^{(k)} \boldsymbol{B}_h]^{-1}$$

and $\boldsymbol{\nu}_{hi}^{(k)}$, E_h, $\boldsymbol{\xi}_h^{(k)}$, F_h, $\boldsymbol{\epsilon}_{hi}^{(k)}$, and G_h are given in Appendix A.

The E- and M-steps are alternated repeatedly until convergence of the EM sequence of iterations.[12] An outright or hard clustering of the genes[16] is obtained by assigning each $\boldsymbol{y}_i$ to the component of the mixture (Equation (1)) to which it has the highest estimated posterior probability of belonging.

3. INFERENCE FOR GENE REGULATORY INTERACTIONS

Once the co-expressed genes are clustered into groups with similar pattern of expression levels, the second step of the method "predicts" the gene expression profile for each gene based on the estimated parameters $\hat{\boldsymbol{\Psi}}$ obtained in the first step. Given that the ith gene belongs to the hth cluster, its gene expression profile at different time points can be expressed as

$$\hat{\boldsymbol{y}}_i = \boldsymbol{X}\hat{\boldsymbol{\beta}}_h + \boldsymbol{V}\hat{\boldsymbol{\nu}}_{hi} + \boldsymbol{W}\hat{\boldsymbol{\xi}}_h, \tag{11}$$

where

$$\hat{\boldsymbol{\nu}}_{hi} = E_{\boldsymbol{\Psi}}(\boldsymbol{\nu}_{hi} | \boldsymbol{y}, \hat{\boldsymbol{z}}; \hat{\boldsymbol{\Psi}})$$

and

$$\hat{\boldsymbol{\xi}}_h = E_{\boldsymbol{\Psi}}(\boldsymbol{\xi}_h | \boldsymbol{y}, \hat{\boldsymbol{z}}; \hat{\boldsymbol{\Psi}})$$

correspond to the best linear unbiased predictor[19,20] (BLUP) for random effects $\boldsymbol{\nu}_{hi}$ and $\boldsymbol{\xi}_h$, respectively. The average gene expression profile for each cluster, $\hat{\boldsymbol{y}}^h = (\hat{y}_1^h, \ldots, \hat{y}_T^h)'$, is then obtained as

$$\hat{\boldsymbol{y}}^h = \sum_{i=1}^{N} \hat{z}_{hi} \hat{\boldsymbol{y}}_i / \sum_{i=1}^{N} \hat{z}_{hi} \qquad (h = 1, \ldots, g).$$

Based on the least square approach considered by Booth et al,[21] the expression profiles for genes from the same cluster may be used to estimate the times to peak expression levels and the periods.

The network of regulatory interactions is then determined in the final step by searching for candidate regulatory control elements (activators and inhibitors) shared by the clusters of co-expressed genes. In the analysis of time-course data, Li et al.[13] have adopted time-lagged correlation coefficients to characterize the time-delayed dependency between genes. With our method, the time-lagged correlation coefficients are calculated using the average gene expression profiles for the clusters. In particular, the correlation coefficient between the hth cluster at time t and the lth cluster at time $t + \kappa$ is defined as

$$R_{hl}(\kappa) = \text{corr}\{(\hat{y}_1^h, \ldots, \hat{y}_{(T-\kappa)}^h)', (\hat{y}_\kappa^l, \ldots, \hat{y}_T^l)'\}, \tag{12}$$

where $\kappa = 1, 2, 3, \ldots$ is a constant integer and the function corr(.,.) is the standard correlation coefficient between the two $(T - \kappa)$ dimensional vectors.[13] A large absolute value of $R_{hl}(\kappa)$ thus indicates a pair-wise dependence between the genes expression of the hth cluster at time t and that of the lth cluster at time $t+\kappa$. In practice, the value of κ should be so chosen that $T-\kappa$ is not too small, such as $T-\kappa > 10$ suggested by Li et al.[13] The examination of the significance of these $g*(g-1)$ time-lagged correlation coefficients, after adjusting for multiple testing using the Bonferroni procedure, helps to determine potential regulation between genes that interact with each other directly or via some intermediates.[6] For example, upstream DNA sequence patterns specific to each cluster may be identified in the promoter region of gene clusters, through which co-regulation of the genes within the cluster is achieved.[22] With our method, we identify the most significant correlations (if present) for a range of values for κ, for each pair of gene clusters. These identified pairs of gene clusters are then considered to be potential elements within the network of regulatory interactions.

4. RESULTS

The method is illustrated using two well-known yeast Saccharomyces cerevisiae datasets.[4, 8] By analyzing cDNA microarrays from yeast cultures synchronized by three independent methods over approximately two cell-cycle periods, Spellman et al.[8] identified 800 yeast genes that are characterized as cell-cycle regulated. In our study, we consider the 18 α-factor (pheromone) synchronization, in which the yeast cells were sampled at 7 minutes intervals for 119 minutes, for each of 104 yeast genes that are determined to be cell cycle-regulated by traditional methods.[8, 21] These 104 gene-expression profiles were used by Spellman et al.[8] to synchronize the initial phases of gene expression profiles obtained in different experiments.[21] A list of these genes (with the relevant list of references) and a complete data sets are available online at `http://genome-www.stanford.edu/cellcycle/data/rawdata`.

For the clustering of the cell cycle-regulated genes based on the microarray data of $N = 104$ genes and $T = 18$ time points, we take the design matrix $\boldsymbol{X}$ be an 18×2 matrix with the l-th row ($l = 1, \ldots, 18$)

$$(\cos(2\pi l/\omega + \Phi) \ \ \sin(2\pi l/\omega + \Phi)), \tag{13}$$

where ω is the period of the cell cycle and Φ is the phase offset.[8] We adopt here the least squares estimates $\omega = 62$ and $\Phi = 0$ obtained by Booth et al.[21] in the analysis of the full dataset of Spellman data. For the design matrices of the random effects parts, we take $\boldsymbol{V} = \boldsymbol{1}_{18}$ and $\boldsymbol{W} = \boldsymbol{I}_{18}$, where $\boldsymbol{1}_{18}$ is a vector of ones with dimension being specified by the subscript. That is, we assume that there exists random gene effects ν_{hi} with $p = 1$ and random temporal effects $\boldsymbol{\xi}_h$ with $q = 18$ for $h = 1, \ldots, g$ and $i = 1, \ldots, N$. Normal mixture models with mixed effects as described in Section 2.1 were fitted to the data with $g = 4$ to $g = 10$. The number of components g was determined using the Bayesian information criterion (BIC) of Schwarz[23] for model selection.[16] With BIC, the number of components is determined by selecting the value of g that provides the minimum negative penalized log likelihood. It indicated here that there are eight clusters. The clustering results for $g = 8$ are given in Figure 1, where the expression profiles for genes in each cluster are presented. For comparison, we also depict the predicted average gene expression profile for each cluster. It can be seen that, in some clusters, large variation of expression profiles are observed for genes within a cluster. In addition, from Figure 1, the times to peak expression and/or the periods are found to be separated among the eight clusters. Genes in the same cluster do show biological relevance. For example, genes that expressed in the G1 phase such as $CLN2$, $RNR1$, $CDC9$, $RAD27$, and $MNN1$[8] are clustered into the same group (cluster 3). Genes that expressed in the M phase are clustered into two groups with different periods, such as cluster 1 containing $CLB2$, $CDC5$ and $SWI5$, and cluster 4 containing $DBF2$ and $CDC20$.

For the computation of the time-lagged correlation coefficients, we considered $\kappa = 1$ to 7. Significant correlation coefficients and the time lags for each pair of gene clusters are given in Table 1, where the first column represents the leading gene clusters. From Table 1, it can be seen that, for small time lag ($\kappa = 1$), genes expressed in the same cell-cycle phase tend to have significant positive correlation; see, for example, gene clusters 1 and 4 that correspond to the genes that expressed in the M phase. Significant correlation coefficients with respect to larger time lags are more relevant in the inference for the regulatory network, as these indicate a correlation between genes that expressed in different phases. For example, the "CLN2" group (cluster 3) is found to be activated by gene cluster 4 (such as the gene $CLN3$) and inhibited by gene cluster 1 (such as the gene $CLB2$).[8] On the other hand, the "CDC46" group that expressed at about the M/G1 boundary (cluster 5) is found to be activated by gene cluster 1 (such as the gene $CLB2$) and inhibited by gene cluster 4 (such as the gene $CLN3$).[8]

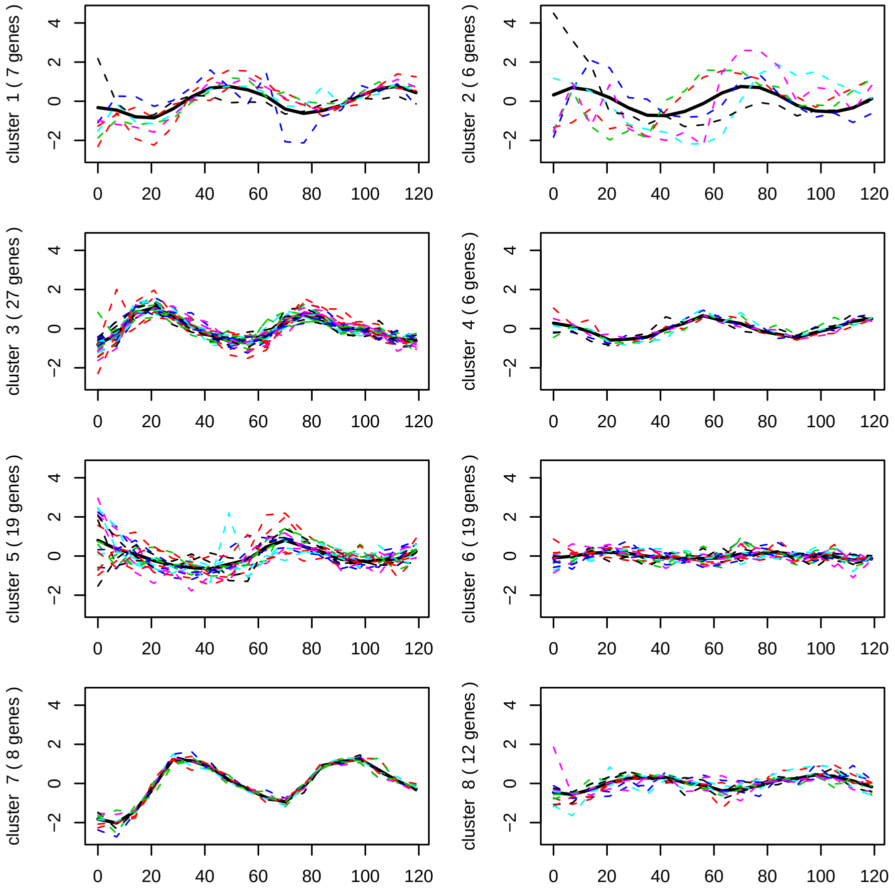

Figure 1. Clustering results for Spellman data. For all the plots, the x-axis is the time point and the y-axis is the gene expression level. The gene expression profiles are given in dashed lines, while the predicted average gene expression profiles are presented in solid lines.

We considered another dataset gathered by Cho et al.[4] It used Affymetrix oligonucleotide microarrays to query the abundances of 6,220 mRNA species in synchronized Saccharomyces cerevisiae batch cultures, using a temperature-sensitive mutation of cdc28. The data provided us with 17 time points across two cell cycles, where the samples were taken at 10 minutes apart. In our study, we worked with a subset of 384 genes that have non-negative normalized fluorescence readings across the time points and were assigned to one of the five phases of cell cycle given at the website `http://yscdp.stanford.edu/yeast_cell_cycle/functional_categories.html` of the Cho data. The data set was normalized as in Tamayo et al.,[15] where the 17 time points were divided into two panels that correspond to two cell cycles and were normalized to have mean 0 and variance 1 within each panel.[24] This standardized dataset is available at `http://faculty.washington.edu/kayee/cluster/`. With this time course data of $N = 384$ and $T = 17$, we take the design matrix $\boldsymbol{X}$ be an 17×2 matrix where the l-th row ($l = 1, \ldots, 17$) is again specified as in Equation (13). Here we adopted the least squares estimates $\omega = 85$ and $\Phi = 0.17\pi$ obtained by Booth et al.[21] in the analysis of the full dataset of Cho data. The design matrices for the random effects were taken to be $\boldsymbol{V} = \mathbf{1}_{17}$ and $\boldsymbol{W} = \boldsymbol{I}_{17}$, respectively. Model selection via BIC indicated that there are six clusters. The clustering results for $g = 6$ are given in Figure 2. Again, it can be seen that the six clusters are different in terms of either times to peak expression or the periods. Genes in the same cluster do show similar functions. For example, genes that expressed at the M/G1 boundary such as $CDC6$, $CDC46$, $CDC54$, $MCM2$, and $MCM3$ involved in DNA replication[4] are clustered into the same group (cluster 4). Genes that expressed in the M phase such as $HDR1$, $MYO1$, $NUF2$, $ASE1$, and $IQG1$ involved in

Table 1. Significant Correlation Coefficients and the Time Lags for Each Pair of Gene Clusters (Spellman data).

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
Cluster 1	—	0.923 $(\kappa = 3)$	-0.861 $(\kappa = 1)$	0.929 $(\kappa = 1)$	0.939 $(\kappa = 3)$	-0.855 $(\kappa = 1)$	0.949 $(\kappa = 7)$	0.964 $(\kappa = 7)$
Cluster 2	-0.968 $(\kappa = 1)$	—	-0.942 $(\kappa = 6)$	0.987 $(\kappa = 7)$	—	0.884 $(\kappa = 1)$	0.902 $(\kappa = 3)$	0.942 $(\kappa = 4)$
Cluster 3	0.936 $(\kappa = 4)$	0.968 $(\kappa = 7)$	—	-0.883 $(\kappa = 1)$	0.983 $(\kappa = 7)$	—	0.954 $(\kappa = 2)$	0.876 $(\kappa = 2)$
Cluster 4	-0.934 $(\kappa = 3)$	0.983 $(\kappa = 2)$	0.867 $(\kappa = 4)$	—	-0.810 $(\kappa = 6)$	—	0.916 $(\kappa = 5)$	0.947 $(\kappa = 6)$
Cluster 5	—	0.852 $(\kappa = 1)$	—	—	—	—	—	—
Cluster 6	0.960 $(\kappa = 4)$	-0.933 $(\kappa = 3)$	—	-0.928 $(\kappa = 1)$	0.940 $(\kappa = 7)$	—	0.933 $(\kappa = 2)$	0.890 $(\kappa = 2)$
Cluster 7	0.954 $(\kappa = 2)$	-0.872 $(\kappa = 1)$	-0.914 $(\kappa = 2)$	0.937 $(\kappa = 3)$	0.949 $(\kappa = 5)$	-0.901 $(\kappa = 2)$	—	—
Cluster 8	0.970 $(\kappa = 2)$	-0.897 $(\kappa = 1)$	-0.951 $(\kappa = 2)$	0.943 $(\kappa = 3)$	0.931 $(\kappa = 5)$	-0.948 $(\kappa = 2)$	-0.924 $(\kappa = 4)$	—

chromosome segregation[4] are clustered into the same group (cluster 5). For the computation of the time-lagged correlation coefficients, we considered $\kappa = 1$ to 6. Significant correlation coefficients and the time lags for each pair of gene clusters are given in Table 2. It can be seen that genes in the cluster 2 (such as $HPR5$, $CDC11$, $NNF1$, $CSE4$, $YKR010C$, $STU2$, $KEX2$, and $YNR009W$) are potential activators[25] for gene clusters 3 to 6.

5. DISCUSSION

We have presented a three-step clustering-based method for inference about GRNs on the basis of time-course gene expression data. The method starts with the clustering of co-expressed genes by fitting normal mixture models with multilevel random-effects framework via the EM algorithm. The prediction of the average gene expression profile for each cluster is then performed on the second step. The network of regulatory interactions is then determined in the final step by searching for regulatory control elements based on the measurement of time-lagged correlation coefficients between clusters. Our aim is to identify a small set of candidate regulatory control elements, which will enable biologists to identify and visualize interesting features from time course expression data. If the promoter regions of the genes are known, as is the case for yeast, it may be possible to identify the cis-regulatory control elements shared by the co-expressed genes.[5]

The significance of the time-lagged correlation coefficients is determined with the adjustment for multiple hypothesis testing using the Bonferroni procedure. The Bonferroni procedure is the most commonly used method for dealing with multiple testing. It controls the family-wise error rate (FWER), which is the probability that at least one false positive error (a type I error) will be committed. However, it is known to be conservative especially when the number of hypothesis testing is very large[1]; for example, if the number of clusters so chosen is as large as that described in Chen et al.[25] In this case, it may be more appropriate to control the expected number of false positives among the rejected hypotheses (the false discovery rate (FDR)).[26, 27] Alternatively, the bootstrap resampling technique[28] can be adopted to check the reliability of regulatory elements obtained from the time-lagged correlation coefficients.[1, 6]

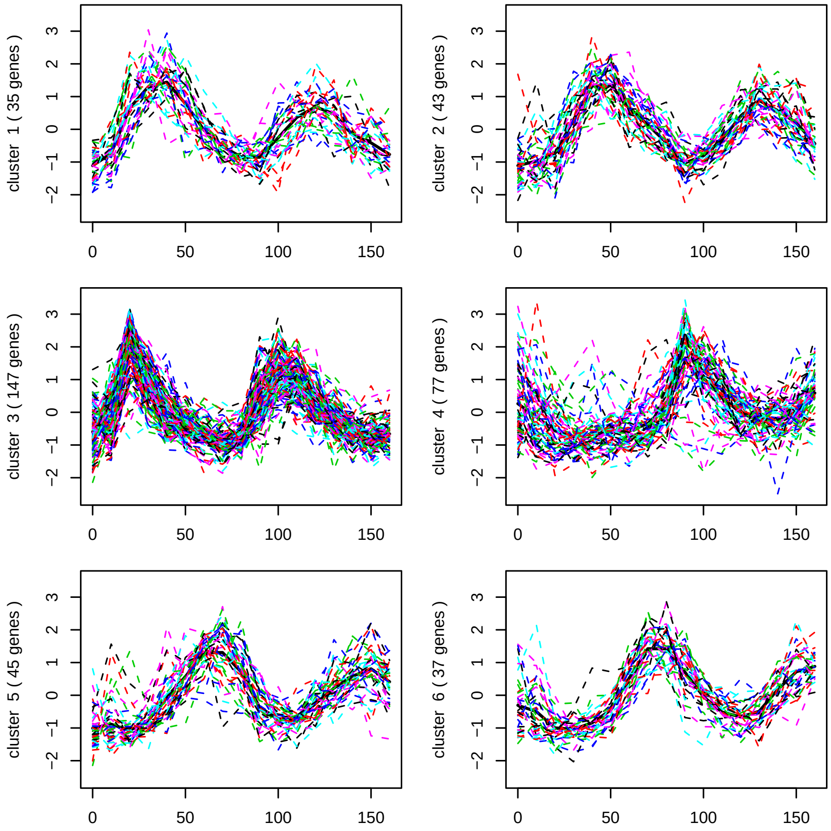

Figure 2. Clustering results for Cho data. For all the plots, the x-axis is the time point and the y-axis is the gene expression level. The gene expression profiles are given in dashed lines, while the predicted average gene expression profiles are presented in solid lines.

With the ML fitting of mixture models via the EM algorithm, it can be seen in Section 2.2 that an initial parameter value of $\mathbf{\Psi}$, $\mathbf{\Psi}^{(0)}$, has to be specified. The monograph of McLachlan and Peel[16] provides an in-depth account of the choice of initial values and the effects of different starting strategies on parameter estimates. Briefly, with mixture models the likelihood typically will have multiple maxima. Hence in practice the EM algorithm needs to be started from a variety of initial values for the parameter vector $\mathbf{\Psi}$ or for a variety of initial partitions of the data into g components. The latter can be obtained by randomly dividing the data into g groups corresponding to the g components of the mixture model. With random starts, the effect of the central limit theorem tends to have the component parameters initially being similar at least in large samples. Nonrandom partitions of the data can be obtained via some clustering procedure such as k-means. In our method, we adopt the automatic approach of the EMMIX program[16] (available online at `http://www.maths.uq.edu.au/~gjm/emmix/emmix.html`) to obtain the initial values $\mathbf{\Psi}^{(0)}$. With the automatic approach, the EMMIX algorithm is run for ten starts corresponding to classifications of the data by k-means (k-means clustering-based starts). The parameter estimates, which correspond to the clustering that produces the highest log likelihood, are adopted as the initial estimates $\mathbf{\Psi}^{(0)}$ for the purposes of starting the EM algorithm.

In Section 4, the number of components g was determined using Schwarz's BIC.[23] This criterion, which is based on a penalized form of the log likelihood, has growing support in the literature for selecting the value of g in the context of mixture model-based clustering of microarray data.[3, 9] In practice, other information criteria in model selection,[16] such as the commonly used Akaike's information criterion (AIC),[29] can also be adopted. An empirical comparison of these criteria with some of the more recently suggested criteria is provided by McLachlan

Table 2. Significant Correlation Coefficients and the Time Lags for Each Pair of Gene Clusters (Cho data).

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Cluster 1	—	0.977 ($\kappa = 1$)	-0.842 ($\kappa = 3$)	—	0.931 ($\kappa = 3$)	0.979 ($\kappa = 4$)
Cluster 2	-0.909 ($\kappa = 3$)	—	0.972 ($\kappa = 6$)	0.892 ($\kappa = 5$)	0.923 ($\kappa = 2$)	0.976 ($\kappa = 3$)
Cluster 3	0.822 ($\kappa = 1$)	—	—	—	0.857 ($\kappa = 4$)	0.894 ($\kappa = 5$)
Cluster 4	—	—	—	—	—	—
Cluster 5	-0.924 ($\kappa = 2$)	-0.921 ($\kappa = 3$)	—	0.931 ($\kappa = 3$)	—	0.967 ($\kappa = 1$)
Cluster 6	-0.831 ($\kappa = 1$)	-0.848 ($\kappa = 2$)	0.839 ($\kappa = 3$)	0.924 ($\kappa = 2$)	—	—

and Peel.[16] When the information criteria scores are compared for a range of values of g, consideration has to be given to the problem of relatively large local maxima of log likelihood that occur as a consequence of a fitted component having a very small (but nonzero) generalized variance (the determinant of the covariance matrix). Such a component corresponds to a cluster containing a few data points either relatively close together or almost lying in a lower-dimensional subspace. There is thus a need to monitor the relative size of the fitted mixing proportions and of the generalized component variances in an attempt to identify these spurious local maximizers.[1]

APPENDIX A. EVALUATIONS OF CONDITIONAL EXPECTATIONS IN THE E-STEP

As described in Section 2.2, the conditional expectations in Equation (3) are directly obtainable within the E-step. With reference to Eqs. (1) and (2), the complete-data log likelihood is given by

$$\log L_c(\boldsymbol{\Psi}) = \sum_{h=1}^{g} \Bigg[\sum_{i=1}^{N} z_{hi} \log \pi_h - \tfrac{1}{2} \Bigg\{ \sum_{i=1}^{N} z_{hi} p \log \theta_{h1} + q \log \theta_{h2} + \sum_{i=1}^{N} z_{hi} T \log \sigma_h^2 + \frac{\boldsymbol{\nu}'_h \boldsymbol{\nu}_h}{\theta_{h1}} + \frac{\boldsymbol{\xi}'_h \boldsymbol{\xi}_h}{\theta_{h2}} + \frac{\boldsymbol{\epsilon}'_h \boldsymbol{\epsilon}_h}{\sigma_h^2} \Bigg\} \Bigg], \tag{14}$$

apart from an additive constant, where

$$\boldsymbol{\nu}'_h \boldsymbol{\nu}_h = \sum_{i=1}^{N} z_{hi} \boldsymbol{\nu}'_{hi} \boldsymbol{\nu}_{hi}$$

and

$$\boldsymbol{\epsilon}'_h \boldsymbol{\epsilon}_h = \sum_{i=1}^{N} z_{hi} \boldsymbol{\epsilon}'_{hi} \boldsymbol{\epsilon}_{hi}$$

with

$$\boldsymbol{\epsilon}_{hi} = (\boldsymbol{y}_i - \boldsymbol{X}\boldsymbol{\beta}_h - \boldsymbol{V}\boldsymbol{\nu}_{hi} - \boldsymbol{W}\boldsymbol{\xi}_h).$$

Since $\boldsymbol{\nu}_h'\boldsymbol{\nu}_h$, $\boldsymbol{\xi}_h'\boldsymbol{\xi}_h$, $\boldsymbol{\epsilon}_h'\boldsymbol{\epsilon}_h$, and $(\boldsymbol{y}_i - \boldsymbol{V}\boldsymbol{\nu}_{hi} - \boldsymbol{W}\boldsymbol{\xi}_h)$ are sufficient statistics[17, 30] for the complete model (Equation (14)), the conditional expectation of the complete-data log likelihood is effected simply by replacing these sufficient statistics in Equation (14) by their conditional expectations.[31, 32] That is, we have

$$E_{\boldsymbol{\Psi}^{(k)}}(z_{hi}|\boldsymbol{y}) = \tau_{hi}^{(k)} = \frac{\pi_h^{(k)} f(\boldsymbol{y}_i|z_{hi}=1;\boldsymbol{\Psi}^{(k)})}{\sum_{l=1}^{g}\pi_l^{(k)} f(\boldsymbol{y}_i|z_{li}=1;\boldsymbol{\Psi}^{(k)})},$$

as given in Equation (4), and

$$\begin{aligned} E_{\boldsymbol{\Psi}^{(k)}}(\boldsymbol{\nu}_h'\boldsymbol{\nu}_h|\boldsymbol{y}) &= \sum_{i=1}^{N}\tau_{hi}^{(k)}\boldsymbol{\nu}_{hi}^{(k)'}\boldsymbol{\nu}_{hi}^{(k)} + E_h, \\ E_{\boldsymbol{\Psi}^{(k)}}(\boldsymbol{\xi}_h'\boldsymbol{\xi}_h|\boldsymbol{y}) &= \boldsymbol{\xi}_h^{(k)'}\boldsymbol{\xi}_h^{(k)} + F_h, \end{aligned}$$

and

$$E_{\boldsymbol{\Psi}^{(k)}}(\boldsymbol{\epsilon}_h'\boldsymbol{\epsilon}_h|\boldsymbol{y}) = \sum_{i=1}^{N}\tau_{hi}^{(k)}\boldsymbol{\epsilon}_{hi}^{(k)'}\boldsymbol{\epsilon}_{hi}^{(k)} + G_h,$$

where

$$\boldsymbol{\nu}_{hi}^{(k)} = E_{\boldsymbol{\Psi}^{(k)}}(\boldsymbol{\nu}_{hi}|\boldsymbol{y}) = \theta_{h1}^{(k)}\left[\boldsymbol{V}'\boldsymbol{A}_h^{-1}(\boldsymbol{y}_i - \boldsymbol{X}\boldsymbol{\beta}_h^{(k)}) - \boldsymbol{V}'\boldsymbol{A}_h^{-1}\boldsymbol{B}_h[\boldsymbol{A}_h + \sum_{l=1}^{N}\tau_{hl}^{(k)}\boldsymbol{B}_h]^{-1}\sum_{l=1}^{N}\tau_{hl}^{(k)}(\boldsymbol{y}_l - \boldsymbol{X}\boldsymbol{\beta}_h^{(k)})\right],$$

$$E_h = \sum_{i=1}^{N}\tau_{hi}^{(k)}p\theta_{h1}^{(k)} - (\sum_{i=1}^{N}\tau_{hi}^{(k)} - 1){\theta_{h1}^{2}}^{(k)}\text{trace}(\boldsymbol{V}'\boldsymbol{A}_h^{-1}\boldsymbol{V}) - {\theta_{h1}^{2}}^{(k)}\text{trace}(\boldsymbol{V}'[\boldsymbol{A}_h + \sum_{i=1}^{N}\tau_{hi}^{(k)}\boldsymbol{B}_h]^{-1}\boldsymbol{V}),$$

$$\boldsymbol{\xi}_h^{(k)} = E_{\boldsymbol{\Psi}^{(k)}}(\boldsymbol{\xi}_h|\boldsymbol{y}) = \theta_{h2}^{(k)}\boldsymbol{W}'[\boldsymbol{A}_h + \sum_{l=1}^{N}\tau_{hl}^{(k)}\boldsymbol{B}_h]^{-1}\sum_{l=1}^{N}\tau_{hl}^{(k)}(\boldsymbol{y}_l - \boldsymbol{X}\boldsymbol{\beta}_h^{(k)}),$$

$$F_h = q\theta_{h2}^{(k)} - \sum_{i=1}^{N}\tau_{hi}^{(k)}{\theta_{h2}^{2}}^{(k)}\text{trace}(\boldsymbol{W}'[\boldsymbol{A}_h + \sum_{i=1}^{N}\tau_{hi}^{(k)}\boldsymbol{B}_h]^{-1}\boldsymbol{W}),$$

$$\boldsymbol{\epsilon}_{hi}^{(k)} = \boldsymbol{y}_i - \boldsymbol{X}\boldsymbol{\beta}_h^{(k)} - \boldsymbol{V}\boldsymbol{\nu}_{hi}^{(k)} - \boldsymbol{W}\boldsymbol{\xi}_h^{(k)},$$

and

$$G_h = \sum_{i=1}^{N}\tau_{hi}^{(k)}T{\sigma_h^{2}}^{(k)} - (\sum_{i=1}^{N}\tau_{hi}^{(k)} - 1){\sigma_h^{4}}^{(k)}\text{trace}(\boldsymbol{A}_h^{-1}) - {\sigma_h^{4}}^{(k)}\text{trace}([\boldsymbol{A}_h + \sum_{i=1}^{N}\tau_{hi}^{(k)}\boldsymbol{B}_h]^{-1}).$$

ACKNOWLEDGMENTS

This work was supported in part by grants from the Australian Research Council and its Centre for Complex Systems, University of Queensland.

REFERENCES

1. G. J. McLachlan, K. A. Do, and C. Ambroise, *Analyzing Microarray Gene Expression Data*, Wiley, New Jersey, 2004.
2. M. P. Styczynski and G. Stephanopoulos, "Overview of computational methods for the inference of gene regulatory networks," *Comput. Chem. Eng.* **29**, pp. 519–534, 2005.
3. Y. Luan and H. Li, "Clustering of time-course gene expression data using a mixed-effects model with B-splines," *Bioinformatics* **19**, pp. 474–482, 2003.

4. R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis, "A genome-wide transcriptional analysis of the mitotic cell cycle," *Mol. Cell* **2**, pp. 65–73, 1998.
5. P. D'haeseleer, S. Liang, and R. Somogyi, "Genetic network inference: from co-expression clustering to reverse engineering," *Bioinformatics* **16**, pp. 707–726, 2000.
6. H. Toh and K. Horimoto, "Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modelling," *Bioinformatics* **18**, pp. 287–297, 2002.
7. G. J. McLachlan, R. W. Bean, and D. Peel, "A mixture model-based approach to the clustering of microarray expression data," *Bioinformatics* **18**, pp. 413–422, 2002.
8. P. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization," *Mol. Biol. Cell* **9**, pp. 3273–3297, 1998.
9. K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo, "Model-based clustering and data transformations for gene expression data," *Bioinformatics* **17**, pp. 977–987, 2001.
10. H. Goldstein, *Multilevel Statistical Models (second edition)*, Arnold, London, 1995.
11. A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," *J. R. Stat. Soc. B* **39**, pp. 1–38, 1977.
12. S. K. Ng, T. Krishnan, and G. J. McLachlan, "The EM algorithm," in *Handbook of Computational Statistics Vol. 1*, J. Gentle, W. Hardle, and Y. Mori, eds., pp. 137–168, Springer-Verlag, New York, 2004.
13. H. Li, Y. Luan, F. Hong, and Y. Li, "Statistical methods for analysis of time course gene expression data," *Front. Biosci.* **7**, pp. 90–98, 2002.
14. L. Ben-Tovim Jones, S. K. Ng, C. Ambroise, K. Monico, N. Khan, and G. J. McLachlan, "Use of microarray data via model-based classification in the study and prediction of survival from lung cancer," in *Methods of Microarray Data Analysis IV*, J. S. Shoemaker and S. M. Lin, eds., pp. 163–173, Springer, New York, 2005.
15. P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Domitrovsky, E. S. Lander, and T. R. Golub, "Interpreting patterns of gene expression pattern with self-organizing maps: methods and application to hematopietic differentiation," *Proc. Natl Acad. Sci. USA* **96**, pp. 2907–2912, 1999.
16. G. J. McLachlan and D. Peel, *Finite Mixture Models*, Wiley, New York, 2000.
17. C. E. McCulloch and S. R. Searle, *Generalized, Linear, and Mixed Models*, Wiley, New York, 2001.
18. G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, Wiley, New York, 1997.
19. C. R. Henderson, "Best linear unbiased estimation and prediction under a selection model," *Biometrics* **31**, pp. 423–447, 1975.
20. G. K. Robinson, "That BLUP is a good thing: the estimation of random effects (with discussion)," *Stat. Sci.* **6**, pp. 15–51, 1991.
21. J. G. Booth, G. Casella, J. E. K. Cooke, and J. M. Davis, "Statistical approaches to analysing microarray data representing periodic biological processes: a case study using the yeast cell cycle," Technical report, Department of Biological Statistics and Computational Biology, Cornell University, 2004.
22. S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture," *Nature Genet.* **22**, pp. 281–285, 1999.
23. G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.* **6**, pp. 461–464, 1978.
24. K. Y. Yeung, D. R. Haynor, and W. L. Ruzzo, "Validating clustering for gene expression data," *Bioinformatics* **17**, pp. 309–318, 2001.
25. T. Chen, V. Filkov, and S. S. Skiena, "Identifying gene regulatory networks from experimental data," *Parallel Comput.* **27**, pp. 141–162, 2001.
26. Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. R. Stat. Soc. B* **57**, pp. 289–300, 1995.
27. S. Dudoit, J. P. Shaffer, and J. C. Boldrick, "Multiple hypothesis testing in microarray experiments," *Stat. Sci.* **18**, pp. 71–103, 2003.
28. B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, London, 1993.

29. H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Second International Symposium on Information Theory*, B. N. Petrov and F. Csaki, eds., pp. 267–281, Akadémiai Kiadó, Budapest, 1973.

30. S. R. Searle, G. Casella, and C. E. McCulloch, *Variance Components*, Wiley, New York, 1992.

31. G. Celeux, O. Martin, and C. Lavergne, "Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments," *Stat. Model.* **5**, pp. 243–267, 2005.

32. S. K. Ng and G. J. McLachlan, "Using the EM algorithm to train neural networks: Misconceptions and a new algorithm for multiclass classification," *IEEE T. Neural Networ.* **15**, pp. 738–749, 2004.