Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 313747, 9 pages
http://dx.doi.org/10.1155/2014/313747

Research Article

Nonlinear-Model-Based Analysis Methods for Time-Course Gene Expression Data

Li-Ping Tian,[1] Li-Zhi Liu,[2] and Fang-Xiang Wu[2,3]

[1] *School of Information, Beijing Wuzi University, No. 1 Fuhe Street, Tongzhou District, Beijing 101149, China*
[2] *Department of Mechanical Engineering, University of Saskatchewan, 57 Campus Drive, Saskatoon, SK, Canada S7N 5A9*
[3] *Division of Biomedical Engineering, University of Saskatchewan, 57 Campus Drive, Saskatoon, SK, Canada S7N 5A9*

Correspondence should be addressed to Fang-Xiang Wu; faw341@mail.usask.ca

Received 27 August 2013; Accepted 16 October 2013; Published 2 January 2014

Academic Editors: B. Shen, J. Wang, and J. Wang

Microarray technology has produced a huge body of time-course gene expression data and will continue to produce more. Such gene expression data has been proved useful in genomic disease diagnosis and drug design. The challenge is how to uncover useful information from such data by proper analysis methods such as significance analysis and clustering analysis. Many statistic-based significance analysis methods and distance/correlation-based clustering analysis methods have been applied to time-course expression data. However, these techniques are unable to account for the dynamics of such data. It is the dynamics that characterizes such data and that should be considered in analysis of such data. In this paper, we employ a nonlinear model to analyse time-course gene expression data. We firstly develop an efficient method for estimating the parameters in the nonlinear model. Then we utilize this model to perform the significance analysis of individually differentially expressed genes and clustering analysis of a set of gene expression profiles. The verification with two synthetic datasets shows that our developed significance analysis method and cluster analysis method outperform some existing methods. The application to one real-life biological dataset illustrates that the analysis results of our developed methods are in agreement with the existing results.

1. Background

To understand the mechanisms of dynamic biological processes, DNA microarray experiments have been employed to produce gene expression profiles at a series of time points, for example, the cell division cycle processes of yeast *Saccharomyces cerevisiae* [1, 2], bacterium *Caulobacter crescentus* [3], and human being [4]. Such time-course gene expression data provides a dynamic snapshot of most (if not all) of the genes related to the biological development process and thus can be useful in genomic disease diagnosis and genomic drug design. The challenge is how to uncover useful information from such data by proper analysis methods [5].

Although the behaviours of genome-wide genes can be monitored simultaneously with the current DNA microarray technology, not are all of monitored genes closely related to the biological process being studied or of interest. In addition, gene expression data are often contaminated by various noises or noisy genes. It is impossible to uncover some useful information without any preprocessing. Either excluding genes of interest or including noisy genes could degrade the significance of any analysis results. Therefore, it is critical to select the genes which are closely relevant to a biological process from gene expression profiles measured during the biological process. The selection of genes can be performed by the so-called significance analysis of gene expression profiles. Much attention has been paid to the significant analysis of static gene expression data over the past years. For gene expression data obtained from a pair of conditions (e.g., normal versus abnormal, or control versus treatment) with multiple replicates, one of the widely used approaches in early years is called the R-fold change method [6, 7]. The "R-fold change" method determines a gene to be significantly expressed if the ratio of expression values under two different conditions is greater than R or less than $1/R$, where R is a user-preset positive number. This approach has been improved by a resampling (bootstrapping) method called SAM [8, 9]. Another approach to the significance analysis is the use of

t-test, for example, on logarithm of the expression levels. In a t-test, the means and variances of gene expressions from a pair of conditions are used to compute a normalized distance so-called t-value. When the t-value exceeds a certain threshold depending on the confidence level selected, gene expression data from a pair of conditions are considered to be significantly different. Although R-fold and t-test approaches can be extended to apply for the analysis of gene expression data with multiple conditions, for example, SAM [8, 9] and RIT [10], these approaches need the assumption that multiconditional values are statistically independent. Therefore, it is not applicable to time-course gene expression profiles as they are not statistically independent but dynamically dependent. In recent year, we have developed several methods for significance analysis of time-course gene expression data. In [11, 12], we employ linear regression models to detect the significantly differentially expressed genes. In [13, 14], we employ nonlinear models to detect the periodically expressed genes.

Besides the significance analysis, the cluster analysis is another class of analysis methods to uncover the useful information from gene expression data [5]. A number of clustering methods have been proposed for cluster analysis of gene expression data. These include distance/correlation-based clustering methods (e.g., hierarchical clustering [15], k-means clustering [16], and self-organizing maps [17]) and static-model-based clustering methods [18, 19]. In these methods, gene expression profiles are viewed as multidimensional vectors. Distance/correlation-based clustering methods cluster genes based on the distance/correlation among their expression profiles. Static-model-based clustering methods assign genes to one of the clusters if their expression profiles are generated by a multivariate normal distribution. These methods do not take the dynamic of time-course gene expression data into account and thus are not efficient for periodically expressed gene data. Some dynamic-model-based clustering methods have been proposed to analyze time-course gene expression data [20, 21]. These methods employ linear autoregressive models to describe the dynamics of time-course gene expression data. Recently we propose the nonlinear-model-based method for clustering periodically expressed genes [22, 23].

As measured from typical nonlinear biological systems, time-course gene expression profiles should display the nonlinear properties. In this paper, we propose nonlinear-model-based methods for significance analysis and cluster analysis of time-course gene expression data. The proposed nonlinear model can be viewed as a generalization of many existing models [13, 14, 20–23]. A two-step method is proposed to estimate the model parameter. An F-test is employed to determine if a gene expression profile is significantly different from noisy data. A relocation-iteration algorithm is employed to assign each gene to an appropriate cluster. A bootstrapping method and an average adjusted Rand index (AARI) are employed to measure the quality of clustering. We employ two synthetic datasets to evaluate the performance of the proposed methods and apply them to one real-life biological dataset.

2. Methods

2.1. Nonlinear Model for Time-Course Gene Profiles. Let $x(t)$ $(t = 1, 2, \ldots, m)$ be a time-course gene expression profile generated from a dynamic biological process, where m is the number of time points at which gene expression is measured. Many nonlinear gene expression profiles contain a periodic component and a long-term decrease or increase component. In this study, we employ the following nonlinear model to describe time-course gene expression data:

$$x(t) = e^{\alpha t}\left[a\cos(\omega t) + b\sin(\omega t)\right] + ct + d + \varepsilon(t), \quad (1)$$

where parameter α represents the degradation rate of periodicity; parameters a and b are the coefficients of sine and cosine functions, respectively; parameter ω is the frequency of periodic expression data; parameters c and d are the coefficients of linear function; and $\varepsilon(t)$ represents random errors. This study assumes that the errors have a normal distribution independent of time with the mean of 0 and the variance of σ^2. This model generalizes several existing models; for example, setting $\alpha = c = d = 0$, model (1) is reduced to sinusoidal function model [24–30]:

$$x(t) = A\sin(\omega t + \Phi) + \varepsilon(t), \quad (2)$$

which is widely used to generate the synthetic periodic gene expression profiles [24] and to detect the periodically expressed genes [27–29]. In model (2), $A = \sqrt{a^2 + b^2}$ is called the magnitude and $\Phi = \arctan(a/b)$ is called the phase. Setting $\alpha = 0$, model (1) is reduced to a model used in [13], while, setting $c = d = 0$, model (1) is reduced to a model used in [14, 22]. As model (1) is the generalization of several existing models, it is expected that the analysis results based on this model are better than those reduced models.

To construct model (1) six parameters need to be estimated from a time-course gene expression profile $x(t)$ $(t = 1, 2, \ldots, m)$. Obviously estimating those parameters in model (1) is a nonlinear estimation problem as α and ω are nonlinear in the model. In general, all nonlinear optimization programs can be used to estimate parameters in model (1), for example, Gauss-Newton iteration method and its variants such as Box-Kanemasu interpolation method, Levenberg damped least squares methods, and Marquardt's method [31–33]. However, these iteration methods are sensitive to initial values. Another main shortcoming is that these methods may converge to the local minimum of the least squares cost function and thus cannot find the true values of the parameters.

Our observation is that noise free model (1)

$$x(t) = e^{\alpha t}\left[a\cos(\omega t) + b\sin(\omega t)\right] + ct + d \quad (3)$$

can be viewed as the general solution of the following second order ordinary differential equation:

$$\ddot{x}(t) + A\dot{x}(t) + Bx(t) = Ct + D, \quad (4)$$

which is independent of a and b and

$$\alpha = -\frac{A}{2}, \qquad \omega = \frac{\sqrt{4B - A^2}}{2}, \qquad c = \frac{C}{B}, \qquad d = \frac{DB - AC}{B^2}. \quad (5)$$

Now we can see that constant parameters A, B, C, and D are linear in (4). As long as we get the first and second derivatives, we can easily estimate the parameters A, B, C, and D by the linear least squares method. Then we can get the estimation of α, ω, c, and d from equations in (5). Finally we can use (3) to estimate the rest of parameters a and b. Therefore, we propose the following two-step parameter estimation methods to estimate all six parameters in model (1).

Step 1. Numerically calculate the first and second derivatives of $x(t)$. As time-course gene expression data are discrete, the first and second derivatives of $x(t)$ can be estimated by the central (second order) finite difference formulas as follows:

$$\dot{x}(t) = \frac{x(t+1) - x(t-1)}{2\Delta} \quad \text{for } t = 2, \ldots, m-1, \tag{6}$$

$$\ddot{x}(t) = \frac{x(t+1) + x(t-1) - 2x(t)}{\Delta^2} \quad \text{for } t = 2, \ldots, m-1, \tag{7}$$

respectively, where Δ is the time difference between two consecutive gene expression data points. If the number of data points in a gene expression profile is enough, one can choose a high order finite difference formula to get more accurate estimation of these derivatives.

Then, based on model (4), we use the linear least squares method to estimate parameter ω^2. In detail, let

$$Y = \begin{bmatrix} \ddot{x}(1) \\ \ddot{x}(2) \\ \vdots \\ \ddot{x}(l) \end{bmatrix}, \qquad X = \begin{bmatrix} -\dot{x}(1) & -x(1) & t_2 & 1 \\ -\dot{x}(2) & -x(2) & t_3 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ -\dot{x}(l) & -x(l) & t_{m-1} & 1 \end{bmatrix}. \tag{8}$$

From (6) and (7), we have $l = m-2$. Then by the least squares method, the paramters A, B, C, and D in model (4) can be estimated as

$$\begin{bmatrix} \widehat{A} \\ \widehat{B} \\ \widehat{C} \\ \widehat{D} \end{bmatrix} = \left(X^T X\right)^{-1} X^T Y. \tag{9}$$

Note that if the value of $4\widehat{B} - \widehat{A}^2$ calculated by (5) for a gene is negative, the expression of this gene will be judged not to be described by model (1).

Step 2. Substitute the estimated values of α, ω, c, and d into (3). Then we apply the least squares method to model (1) to estimate parameters a and b. In detail, let

$$Z = [z(1), \ldots, z(m)],$$
$$E = \begin{bmatrix} \cos(\Delta\widehat{\omega}), \ldots, \cos(m\Delta\widehat{\omega}) \\ \sin(\Delta\widehat{\omega}), \ldots, \sin(m\Delta\widehat{\omega}) \end{bmatrix}; \tag{10}$$

by the least squares method, a and b can be estimated as

$$\begin{bmatrix} \widehat{a} \\ \widehat{b} \end{bmatrix} = \left(EE^T\right)^{-1} \left(EZ^T\right), \tag{11}$$

where

$$z(t) = e^{-\widehat{\alpha}t} \left[x(t) - \widehat{c}t - \widehat{d}\right] \quad \text{for } t = 1, 2, \ldots, m. \tag{12}$$

2.2. Nonlinear-Model-Based Significance Analysis. Significance analysis of gene expression data is to determine if a gene expression profile is significantly different from noisy data. This issue is not easy to answer through statistical inference [29, 30] yet, especially for time-course gene expression profiles as their data points are not statistically independent. However, a practical way in the literature [27–30] is to perform a statistical hypothesis test whether the gene expression profile is pure normal white noise or it fits a certain model as specified by (1). Along with this way, this study tests the null hypothesis of

(H_0)

$$x(t) = d + \varepsilon(t) \tag{13}$$

versus the alternative hypothesis of

(H_1) (see (1)).

Let

$$S_0^2 = \sum_{i=1}^{m} \left(x(t_i) - \widehat{d}\right)^2, \qquad \widehat{d} = \frac{1}{m} \sum_{i=1}^{m} x(t_i), \tag{14}$$

where S_0^2 is the residual of model (13) with estimated parameters, and

$$S_1^2 = \sum_{i=1}^{m} \left\{x(t_i) - e^{\widehat{\alpha}t} \left[a\cos\left(\widehat{\omega}t_i\right) + \widehat{b}\sin\left(\widehat{\omega}t_i\right)\right] - \widehat{c}t_i - \widehat{d}\right\}^2, \tag{15}$$

where S_1^2 is the residual of model (1) with estimated parameters. As the noise model (13) can be viewed as a special case of model (1), the statistic

$$F = \frac{\left(S_0^2 - S_1^2\right)/5}{S_1^2/(m-6)} = \frac{m-6}{5}\left(\frac{S_0^2}{S_1^2} - 1\right) \tag{16}$$

follows the F-distribution with the degrees of freedom $(5, m-6)$, according to statistics theory [21, 23].

When the value of F-statistic is large enough (greater than a specified threshold), model (13) is rejected; that is, the gene expression profile is not pure normal white noise, and otherwise the gene expression profile appears as white noises. According to degrees of freedom (which are related to the length of time-course data m and the number of parameters in the models) and a significance level (typically, 0.01, 0.05, 0.1, 0.2, or the like) specified by a user, the threshold

value can be determined from F-distribution table or by using a f-distribution table or a standard MATLAB function $icdf(\text{“}f\text{”}, 1-\gamma, 5, m-6)$, where γ is the significance level. A significance level is the probability that the null hypothesis is true. Therefore, the rejection of the null hypothesis at a smaller significance level indicates the more favourable to alternative hypothesis. That is, the smaller the significance level is, the more confidence one accepts that genes are not noises if its corresponding value of F-statistic is greater than the threshold.

2.3. Nonlinear-Model-Based Cluster Analysis

2.3.1. The Mixture Model. In this study, it is assumed that a time-course gene expression dataset is a collection of time series which belongs to several clusters and time series in each cluster can be described by model (1) with different parameters. Let $\theta_k = [\alpha_k, \omega_k, a_k, b_k, c_k, d_k]$ be parameters of model (1) for the kth cluster. Then the task of nonlinear-model-based clustering is as follows: for a given number of cluster K, divide a time-course gene expression dataset into a partition $C = \{C_1, \ldots, C_k, \ldots, C_K\}$ using model (1) with parameters $\theta_k = [\alpha_k, \omega_k, a_k, b_k, c_k, d_k] (k = 1, \ldots, K)$ which minimize

$$f(C \mid \Theta) = \sum_{k=1}^{K} \sum_{x \in C_k} \sum_{i=1}^{m} \left\{ x(i) - e^{\alpha_k \Delta i} \left[a_k \cos\left(i \Delta \omega_k\right) + b_k \sin\left(i \Delta \omega_k\right) \right] - c_k \Delta i - d_k \right\}^2, \tag{17}$$

where the parameters Θ consist of $\{\theta_k,\ k = 1, \ldots, K\}$.

2.3.2. Estimation of Cluster Parameters. According to the parameter estimation method proposed in previous section for single time-course expression profiles, for the kth cluster, parameters $\theta_k = [\alpha_k, \omega_k, a_k, b_k, c_k, d_k]$ can be estimated as

$$\widehat{\alpha}_k = -\frac{\widehat{A}_k}{2}, \qquad \widehat{\omega}_k = \frac{\sqrt{4\widehat{B}_k - \widehat{A}_k^2}}{2}, \qquad \widehat{c}_k = \frac{\widehat{C}}{\widehat{B}}, \qquad \widehat{d}_k = \frac{\widehat{D}_k \widehat{B}_k - \widehat{A}_k \widehat{C}_k}{\widehat{B}_k^2}, \tag{18}$$

where

$$\begin{bmatrix} \widehat{A}_k \\ \widehat{B}_k \\ \widehat{C}_k \\ \widehat{D}_k \end{bmatrix} = \left(\sum_{x \in C_k} X^T X \right)^{-1} \sum_{x \in C_k} X^T Y, \qquad \begin{bmatrix} \widehat{a}_k \\ \widehat{b}_k \end{bmatrix} = \left(\sum_{x \in C_k} E E^T \right)^{-1} \sum_{x \in C_k} E Z^T, \tag{19}$$

where $|C_k|$ represents the number of time series in cluster C_k, $\sum_{k=1}^{K} |C_k| = N$.

2.3.3. Algorithm for Clustering. This study employs the following relocation-iteration algorithm to estimate the parameters such that the cost function (17) is minimized:

(1) select an initial partition for the given number of clusters, K;

(2) iterate ($s = 1, 2, \ldots$):

(a) estimate the parameter Θ^s based on the current partition by using (18)-(19);

(b) generate a new partition by assigning each sequence x to cluster k where

$$k = \arg\min_{1 \le j \le K} \sum_{i=1}^{m} \left\{ x(i) - e^{\alpha_j^s \Delta i} \left[a_j^s \cos\left(i \Delta \omega_j^s\right) + b_j^s \sin\left(i \Delta \omega_j^s\right) \right] - c_j^s \Delta i - d_j^s \right\}^2; \tag{20}$$

(3) stop if the improvement of the cost function (17) is below a given threshold, or the cluster memberships of time series do not change significantly.

In 2(a), $\Theta^s = \{\theta_k^s, 1 \le k \le K\}$ represents the estimated parameters in cost function (17) at iteration s while in 2(b), parameters α_j^s, ω_j^s, a_j^s, b_j^s, c_j^s, and d_j^s represent the parameters of model j at iteration s.

3. Evaluation

In this section, we use two synthetic datasets to evaluate our proposed significance analysis method and cluster analysis method, respectively. To evaluate the significance analysis method, we generate one synthetic dataset that consists of 2000 noisy gene expression profiles based on model (13) and 2000 time-course gene expression profiles based on model (1). All 4000 expression profiles are depicted in Figure 1, from which one can hardly differentiate time-course gene expression profiles from noisy ones. To measure the performance of significance analysis, we employ two widely used indices: sensitivity and specificity, which can be defined as follows [34]:

$$\text{Sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}},$$

$$\text{Specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}, \tag{21}$$

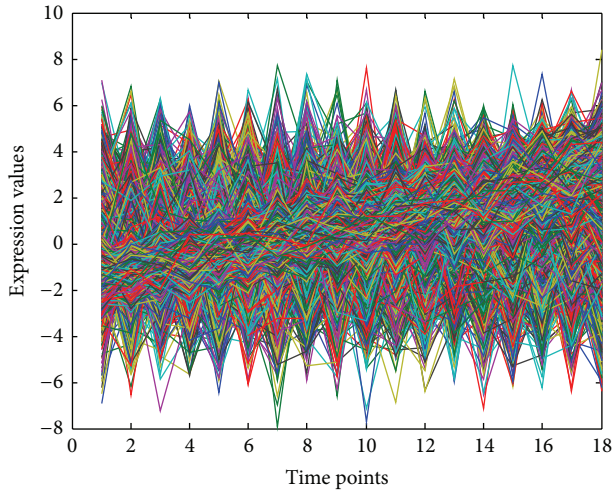

Figure 1: Plot of 4000 expression profiles for evaluating significance analysis method.

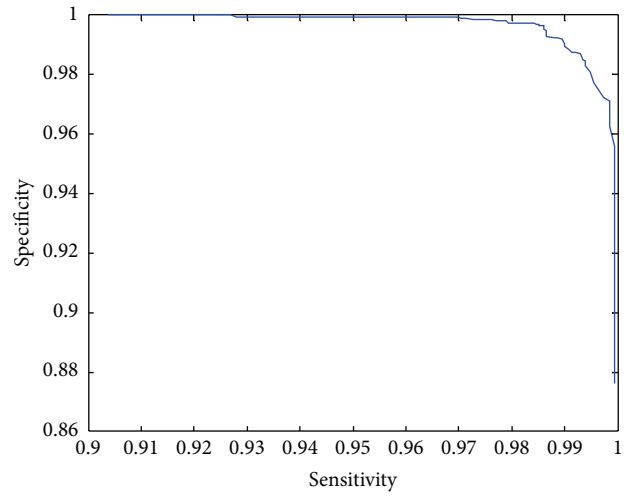

Figure 2: Plot of sensitivity versus specificity.

where

true positive is a time-course gene expression profile identified as it is;

false positive is a time-course gene expression profile identified as it is noisy;

true negative is a noisy gene expression profile identified as it is;

false negative is a noisy gene expression profile identified as it is time-course.

The sensitivity and the specificity depend on thresholds which determine if an expression profile is time-course or noisy. In general, the sensitivity is increasing, while the specificity is decreasing and vice versa. However, a good method is expected to have both high sensitivity and specificity. Figure 2 depicts the curves of sensitivity versus specificity over various thresholds. From this figure, we can see that both sensitivity and specificity can be as high as of 99% for a specific threshold, which indicates that our proposed significance analysis methods are excellent.

To evaluate our proposed cluster analysis method, another synthetic dataset consisting of six clusters is generated from model (1), where different clusters have different randomly selected parameters with some large variances. In each cluster, all profiles are generated with model parameters for this cluster with some random perturbations. All generated profiles are plotted in Figure 3, from which one can see that all time-course gene expression profiles are mixed up. To measure the quality of clustering results, we use the adjusted Rand index (ARI) [35], which originally is to measure the degree of agreement between two partitions of the same set of objects. Given two partitions of N objects, the r-cluster partition $U = \{u_1, \ldots, u_r\}$ and the s-cluster partition $V = \{v_1, \ldots, v_s\}$, the ARI is defined as follows [35]:

$$\text{ARI} = \frac{\sum_{i=1}^{r}\sum_{i=1}^{s}\binom{n_{ij}}{2} - 1/T\sum_{i=1}^{r}\binom{n_{i.}}{2}\sum_{i=1}^{s}\binom{n_{.j}}{2}}{1/2\left[\sum_{i=1}^{r}\binom{n_{i.}}{2} + \sum_{i=1}^{s}\binom{n_{.j}}{2}\right] - (1/T)\sum_{i=1}^{r}\binom{n_{i.}}{2}\sum_{i=1}^{s}\binom{n_{.j}}{2}}, \quad (22)$$

where T is the number of pairs of N objects, n_{ij} is the number of objects that are both in clusters u_i and v_j, $i = 1, \ldots, r$, $j = 1, \ldots, s$, and $n_{i.}$ is the number of objects in cluster u_i, while $n_{.j}$ is the number of objects in cluster v_j. From these definitions, we have

$$T = \frac{N(N-1)}{2}, \qquad n_{i.} = \sum_{j=1}^{s} n_{ij}, \qquad n_{.j} = \sum_{i=1}^{r} n_{ij}. \quad (23)$$

The expected value of ARI is 1 when two partitions agree perfectly and 0 when they are selected at random.

As the results of clustering are sensitive to the initial partition, we run our proposed clustering algorithm and competing clustering algorithms 30 times on the synthetic dataset and calculate the average ARI (AARI) for each algorithm. Figure 4 depicts the AARI of three algorithms named "algorithm with random initial," "algorithm with k-means initial," and "k-means" over several different numbers of clusters, where "algorithm with random initial" means our proposed clustering algorithm with randomly chosen initial partition, "algorithm with k-means initial" means our proposed clustering algorithm with k-means result as initial partition, and "k-means" is an algorithm coded in the MATLAB software for k-means clustering method. Those values of AARI are also listed in Table 1.

From Figure 4 and Table 1, one can see that our algorithm with random chosen initial partitions outperforms the other two algorithms. Particularly, at the correct number of clusters,

Table 1: The values of AARI for different clustering methods on synthetic data.

No. of clusters	2	3	4	5	6	7	8	9	10
Random initial	0.2915	0.5741	0.6636	0.7549	0.9787	0.9516	0.8862	0.826	0.7944
k-means initial	0.2915	0.4875	0.6741	0.7168	0.7732	0.7668	0.7666	0.7739	0.753
k-means	0.2915	0.5099	0.6352	0.7047	0.8001	0.7635	0.8189	0.7849	0.7827

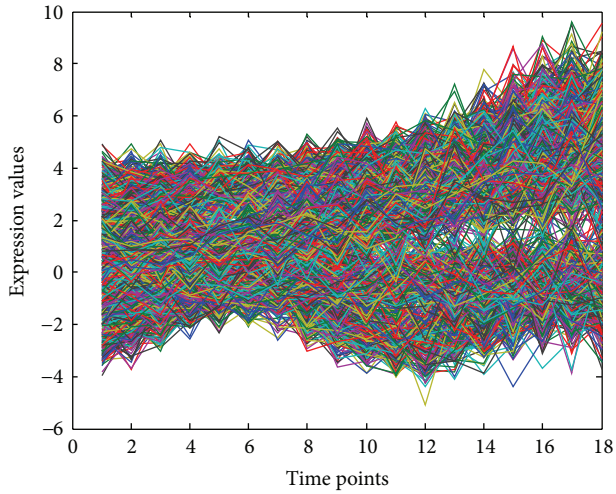

Figure 3: Plot of expression profiles for evaluating cluster analysis method.

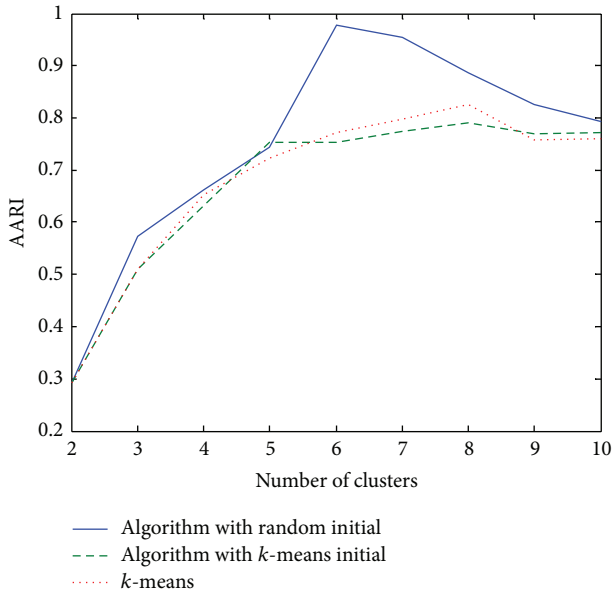

Figure 4: Plot of AARI with different numbers of clusters.

the ARRI from our algorithm with random chosen initial partitions reaches its maximum. The quality of our algorithm with k-means result as initial partitions is comparable with that of k-means, which indicates that after k-means falls in a local optimum, our algorithm cannot escape from that local optimum and thus inherits the drawbacks of k-means. This also suggests that our developed algorithm should combine with random chosen initial partitions.

4. Applications to a Real-Life Gene Expression Data

In this section, we apply our proposed significance analysis and cluster analysis method to a real-life gene expression dataset which is collected from the alpha-synchronized experiment [2]. To study the mitotic cell division cycle of yeast, Spellman et al. [2] have monitored more than 6000 genes of yeast (*Saccharomyces cerevisiae*) at 18 equally spacing time points in the alpha-synchronized experiment. The original dataset is publicly available at http://genome-www.stanford.edu. Genes with missing data are excluded in this study. The resultant dataset contains the expression profiles of 4489 genes.

We first apply our proposed significance analysis method to this dataset and set the significance level $\gamma = 0.1$. As a result, 846 genes are determined to be involved in the alpha-synchronized cell division cycle process, while the other 3643 genes are determined to be noises with respect to this process. Figure 5(f) depicts these 3643 expression profiles. From Figure 5(f), most of the expression profiles look like noises and are not related to the alpha-synchronized cell division cycle process according to the results in [2]. Then we apply our proposed clustering algorithm to the subset consisting of 846 genes involved in the alpha-synchronized cell division cycle process. According to the biological meaning of this process [2], the reasonable number of clusters is 5. The model parameters identified for each of the five clusters are listed in Table 2. From Table 2, for all clusters the values of parameter α_k are negative numbers, which are reasonable. As the cell division cycle is a stable biological system, after a perturbation such as the alpha synchronization, the system tends to its stable attractor. Therefore the degradation rate represented by α_k should be negative.

Furthermore, the values of model parameters a_k and b_k determine the importance of periodic components. From Table 2, the module of parameters a_k and b_k is the largest, while the absolute value of parameter α_k is small for Cluster 3. This indicates that 17 genes in Cluster 3 are periodically expressed in this process, which can be verified from Figure 5(c). Actually all 17 genes in this cluster have also been identified as periodically expressed genes in [2]. The module of parameters a_k and b_k is the second largest for Cluster 5, while the absolute value of parameter α_k is very large for Cluster 5. As a result, gene expression profiles in Cluster 5 are quickly degrading while hardly displaying periodicity as

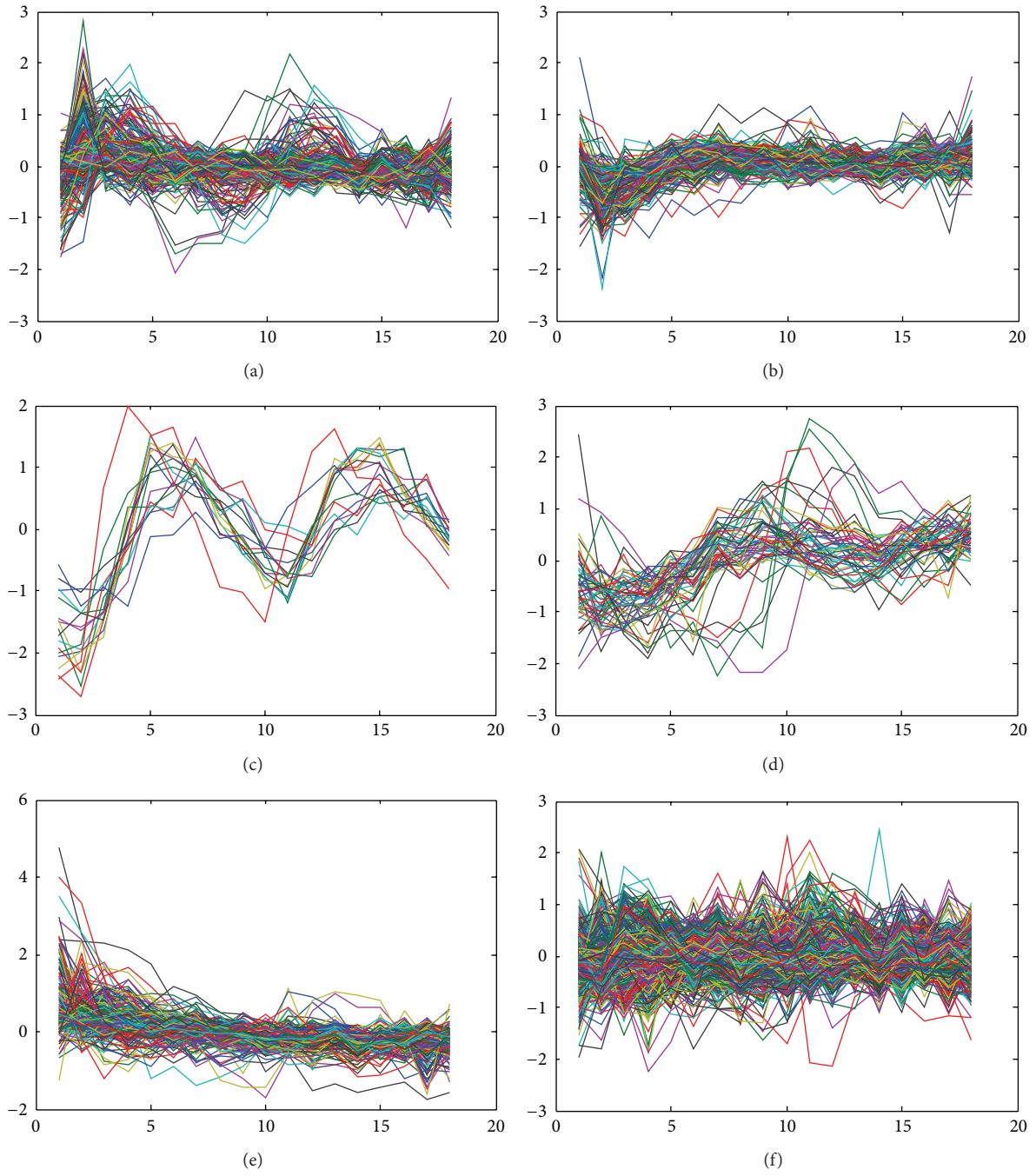

Figure 5: Plot of gene expression profiles. (a)–(e) show gene expression profiles for one of five clusters. (f) shows gene expression profiles which are determined as noises.

Table 2: The model parameters for each cluster.

Parameters	Cluster 1 (315)	Cluster 2 (233)	Cluster 3 (17)	Cluster 4 (53)	Cluster 5 (228)
α_k	−1.1543	−1.7033	−0.6612	−0.5111	−1.8483
ω_k	9.8108	9.8673	7.1631	7.0517	8.736
a_k	0.0234	0.2675	1.0948	0.0024	0.4427
b_k	0.1389	0.033	−1.2261	0.1248	−0.6807
c_k	−0.1287	0.1422	0.3353	0.5748	−0.3738
d_k	0.1383	−0.1372	−0.2723	−0.6011	0.3946

shown in Figure 5(e). According to the estimated values of model parameters, expression profiles in other clusters can similarly be explained.

5. Conclusions

In this paper, we have presented a significance analysis method and a cluster analysis method for time-course gene expression profiles. In these methods, time-course gene expression profiles are modeled by a nonlinear model, which is a generalization of several existing models. To estimate the parameters, which is key to the developed significance analysis method and a cluster analysis method, we propose a two-step linear least squares method. One synthetic dataset has been employed to verify our developed significance analysis method in terms of sensitivity and specificity, while another synthetic dataset has been employed to verify our developed cluster analysis method in terms of AARI. The results have shown that both of our developed methods outperform some existing methods. The application to one real-life biological dataset illustrates that the analysis results of our developed methods are in agreement with the existing results. The reconstruction of gene regulatory network from time-course gene expression data is a very important issue in systems biology [36]. Obviously, noisy genes should be excluded from gene expression data for reconstructing gene regulatory networks. In the future, we may combine our method with other methods as in [36] to reconstruct gene regulatory networks.

Conflict of Interests

The authors declare that there are no competing interests.

Acknowledgments

This research is supported by the Special Fund of Ministry of Education of Beijing for Distinguishing Professors and the Science and Technology Funds of Beijing Ministry of Education (SQKM201210037001) through the first author and Natural Sciences and Engineering Research Council of Canada (NSERC) through the other authors.

References

[1] R. J. Cho, M. J. Campbell, E. A. Winzeler et al., "A genome-wide transcriptional analysis of the mitotic cell cycle," *Molecular Cell*, vol. 2, no. 1, pp. 65–73, 1998.

[2] P. T. Spellman, G. Sherlock, M. Q. Zhang et al., "Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization," *Molecular Biology of the Cell*, vol. 9, no. 12, pp. 3273–3297, 1998.

[3] M. T. Laub, H. H. McAdams, T. Feldblyum, C. M. Fraser, and L. Shapiro, "Global analysis of the genetic network controlling a bacterial cell cycle," *Science*, vol. 290, no. 5499, pp. 2144–2148, 2000.

[4] M. L. Whitfield, G. Sherlock, A. J. Saldanha et al., "Identification of genes periodically expressed in the human cell cycle and their expression in tumors," *Molecular Biology of the Cell*, vol. 13, no. 6, pp. 1977–2000, 2002.

[5] P. Baldi and G. W. Hatfield, *DNA Microrarrays and Gene Expression: From Experiments to Data Analysis and Modeling*, Cambridge University Press, New York, NY, USA, 2002.

[6] A. A. Alizadeh, M. B. Elsen, R. E. Davis et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, no. 6769, pp. 503–511, 2000.

[7] J. M. Claverie, "Computational methods for the identification of differential and coordinated gene expression," *Human Molecular Genetics*, vol. 8, no. 10, pp. 1821–1832, 1999.

[8] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 9, pp. 5116–5121, 2001.

[9] G. K. Smyth, J. Michaud, and H. S. Scott, "Use of within-array replicate spots for assessing differential expression in microarray experiments," *Bioinformatics*, vol. 21, no. 9, pp. 2067–2075, 2005.

[10] R. Nilsson, J. M. Pena, J. Bjorkegren, and J. Tegner, "Detecting multivariate differentially expressed genes," *BMC Bioinformatics*, vol. 8, article 150, 2007.

[11] F. X. Wu and W. J. Zhang, "Dynamic-model-based method for selecting significantly expressed genes from time-course expression profiles," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 1, pp. 16–22, 2010.

[12] F. X. Wu, Z. H. Xia, and L. Mu, "Finding significantly expressed genes from time-course expression profiles," *International Journal of Bioinformatics Research and Applications*, vol. 5, no. 1, pp. 50–63, 2009.

[13] L. P. Tian, L. Z. Liu, and F. X. Wu, "Identification of pseudo-periodic gene expression profiles," in *Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP '11)*, pp. 42–46, Las Vegas, Nev, USA, July 2011.

[14] L. P. Tian, L. Z. Liu, and F. X. Wu, "Detecting nearly periodically expressed genes from their microarray time-course profiles," in *Proceedings of the IASTED International Conference on Computational Bioscience (CompBio '10)*, pp. 612–619, November 2010.

[15] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 25, pp. 14863–14868, 1998.

[16] 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*, vol. 17, no. 10, pp. 977–987, 2001.

[17] P. Törönen, M. Kolehmainen, G. Wong, and E. Castrén, "Analysis of gene expression data using self-organizing maps," *FEBS Letter*, vol. 451, no. 2, pp. 142–146, 1999.

[18] D. Ghosh and A. M. Chinnaiyan, "Mixture modelling of gene expression data from microarray experiments," *Bioinformatics*, vol. 18, no. 2, pp. 275–286, 2002.

[19] G. J. McLachlan, R. W. Bean, and D. Peel, "A mixture model-based approach to the clustering of microarray expression data," *Bioinformatics*, vol. 18, no. 3, pp. 413–422, 2002.

[20] M. F. Ramoni, P. Sebastiani, and I. S. Kohane, "Cluster analysis of gene expression dynamics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 14, pp. 9121–9126, 2002.

[21] F. X. Wu, W. J. Zhang, and A. J. Kusalik, "Dynamic model-based clustering for time-course gene expression data," *Journal*

of Bioinformatics and Computational Biology, vol. 3, no. 4, pp. 821–836, 2005.

[22] L. P. Tian, H. Y. Lu, and F. X. Wu, "Nonlinear model-based clustering for periodically expressed gene profiles," in *Proceedings of the 7th International Symposium on Bioinformatics Research and Applications (ISBRA '11)*, pp. 62–66, Hunan, China, May 2011.

[23] L. P. Tian, L. Z. Liu, Q. W. Zhang, and F. X. Wu, "Nonlinear model-based method for clustering periodically expressed genes," *TheScientificWorldJOURNAL*, vol. 11, pp. 2051–2061, 2011.

[24] S. L. Harmer, J. B. Hogenesch, M. Straume et al., "Orchestrated transcription of key pathways in Arabidopsis by the circadian clock," *Science*, vol. 290, no. 5499, pp. 2110–2113, 2000.

[25] S. Wichert, K. Fokianos, and K. Strimmer, "Identifying periodically expressed transcripts in microarray time series data," *Bioinformatics*, vol. 20, no. 1, pp. 5–20, 2004.

[26] R. A. Fisher, "Test of significance in harmonic analysis," *Proceedings of the Royal Society A*, vol. 125, no. 796, pp. 54–59, 1929.

[27] J. Chen, "Identification of significant periodic genes in microarray gene expression data," *BMC Bioinformatics*, vol. 6, article 286, 2005.

[28] E. F. Glynn, J. Chen, and A. R. Mushegian, "Detecting periodic patterns in unevenly spaced gene expression time series using Lomb-Scargle periodograms," *Bioinformatics*, vol. 22, no. 3, pp. 310–316, 2006.

[29] J. Chen and K. C. Chang, "Discovering statistically significant periodic gene expression," *International Statistical Review*, vol. 76, no. 2, pp. 228–246, 2008.

[30] A. W. C. Liew, N. F. Law, X. Q. Cao, and H. Yan, "Statistical power of Fisher test for the detection of short periodic gene expression profiles," *Pattern Recognition*, vol. 42, no. 4, pp. 549–556, 2009.

[31] J. V. Beck and K. J. Arnold, *Parameter Estimation in Engineering and Science*, John Wiley & Sons, New York, NY, USA, 1977.

[32] A. van den Bos, *Parameter Estimation for Scientists and Engineers*, John Wiley & Sons, New York, NY, USA, 2007.

[33] I. Chou and E. O. Voit, "Recent developments in parameter estimation and structure identification of biochemical and genomic systems," *Mathematical Biosciences*, vol. 219, no. 2, pp. 57–83, 2009.

[34] P. Baldi and S. Brunak, *Bioinformatic: The Machine Learning Approach*, The MIT Press, Cambridge, Mass, USA, 2nd edition, 2001.

[35] A. M. Krieger and P. E. Green, "A generalized rand-index method for consensus clustering of separate partitions of the same data base," *Journal of Classification*, vol. 16, no. 1, pp. 63–89, 1999.

[36] W. Yan, H. Zhu, Y. Yang, J. Chen, Y. Zhang, and B. Shen, "Effects of time point measurement on the reconstruction of gene regulatory networks," *Molecules*, vol. 15, no. 8, pp. 5354–5368, 2010.