Research Article
TheScientificWorldJOURNAL (2011) 11, 2051–2061
ISSN 1537-744X; doi:10.1100/2011/520498

TheScientificWorldJOURNAL
www.thescientificworld.com

Nonlinear Model-Based Method for Clustering Periodically Expressed Genes

Li-Ping Tian,[1] Li-Zhi Liu,[2] Qian-Wei Zhang,[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 Dr., Saskatoon, SK, Canada S7N 5A9*
[3] *Division of Biomedical Engineering, University of Saskatchewan, 57 Campus Dr., Saskatoon, SK, Canada S7N 5A9*

Received 15 September 2011; Accepted 15 October 2011

Academic Editor: Akhmad Sabarudin

Clustering periodically expressed genes from their time-course expression data could help understand the molecular mechanism of those biological processes. In this paper, we propose a nonlinear model-based clustering method for periodically expressed gene profiles. As periodically expressed genes are associated with periodic biological processes, the proposed method naturally assumes that a periodically expressed gene dataset is generated by a number of periodical processes. Each periodical process is modelled by a linear combination of trigonometric sine and cosine functions in time plus a Gaussian noise term. A two stage method is proposed to estimate the model parameter, and 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. One synthetic dataset and two biological datasets were employed to evaluate the performance of the proposed method. The results show that our method allows the better quality clustering than other clustering methods (e.g., *k*-means) for periodically expressed gene data, and thus it is an effective cluster analysis method for periodically expressed gene data.

KEYWORDS: Gene expression data, nonlinear model, periodicall expressed genes, clustering, average adjusted Rand index

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

Published by TheScientificWorldJOURNAL; http://www.tswj.com/

1. BACKGROUND

Many biological processes such as cell-cycle division exhibit periodic behaviors. To understand the mechanisms of these 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. It is believed that clustering periodically expressed gene from their time-course expression data could help understand the molecular mechanisms of those biological processes.

In past decade, a number of methods have been proposed for identifying and clustering periodically expressed genes. The discrete Fourier transform method is the earliest method for identifying and clustering periodically expressed genes [1–4]. In these papers, the discrete Fourier transform is applied to gene expression data to get a two-dimensional vector. One component of the vector is the sum of all coefficients of sine functions while another component is the sum of all coefficients of cosine functions. Then the magnitude of the two-dimensional vector is used to measure periodicity of time-course gene expression profile. The rather subjective cut-off value is taken to determine if a gene is periodically expressed. By this way, Spellman et al. determine that 800 genes are periodically expressed out of more 6000 gene expression profiles from yeast *Saccharomyces cerevisiae*. After performing cluster analysis, these 800 genes are divided into five groups [2]. However, microarray experiments typically generate short time-course data. As pointed in [5, 6], the frequency resolution obtained on such short time-course data by the discrete Fourier transform is often not adequate for resolving periodicities of interest.

Authors in [7] propose a method called CORRCOS to find periodically expressed genes. CORRCOS generates totally 101000 periodic synthetic models. Each gene expression profile is compared to each of these 101000 models. Although it can identify periodically expressed gene, CORRCOS is too time consuming and the cross-correlation is not real metric. In [6], authors develop another algorithm named RAGE for detecting periodically expressed genes. Like CORRCOS, RAGE is a synthetic model-based method. Compared with CORRCOS, RAGE is less time consuming [6]. Wichert et al. [8] propose a statistical method to identify periodically expressed genes from their time-course gene expression profiles. The method models gene expression profiles also as sine functions use the Fisher g-test for statistical analysis. Given a time-course gene expression profile y_t $(t = 1, 2, \ldots, m)$, the g-static is defined as

$$g = \frac{\max_k I(\omega_k)^2}{\sum_{k=1}^{[m/2]} I(\omega_k)}, \tag{1.1}$$

where

$$I(\omega) = \frac{1}{m}\left|\sum_{t=1}^{m} y_t \exp(-i\omega t)\right|^2, \quad \omega \in [0, \pi] \tag{1.2}$$

is called the periodogram. It is assumed that if a time-course gene expression profile has a significant sinusoidal component with frequency $\omega_0 \in [0, \pi]$, the periodogram exhibits a peak at that frequency with a high probability. On the other hand, if a time-course gene expression profile is purely random, the periodogram reduces to a straight line. Based on Fisher g-test [9], Chen [10] proposes a C&G procedure to identify periodically expressed genes from their time-course expression profiles. The g-statistic is effective only for evenly spaced gene expression profiles. For unevenly spaced gene expression profiles, Chen et al. propose to use Lomb-Scargle periodograms to discover statistically significant periodic gene expression [11, 12]. However, a recent research [13] has concluded that the Fisher g-test is poor if the time-course data is short and/or that data length is not an integer number of periods. Therefore, one can not expect to get a good clustering based on periodically expressed genes identified from these methods.

On the other hand, a number of clustering methods have been proposed for cluster analysis on gene expression data. These include distance/correlation-based clustering methods (e.g., hierarchical clustering [14], k-means clustering [15], and self-organizing maps [16]) and static model-based clustering methods [17, 18]. 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 clusters if their expression profiles may be generated by a multivariate normal distribution. These methods do not take into account the dynamic of time-course gene expression data and thus are not efficient for periodically expressed gene data.

Recently, some dynamic model-based clustering methods have been proposed to analyze time-course gene expression data [19, 20]. These methods employ autoregressive models to describe the dynamics of time-course gene expression data. As periodically expressed genes are associated with periodic biological processes, it is natural to model a periodically expressed gene data by periodic (nonlinear) function. This paper proposes a nonlinear model based method for clustering periodically expressed genes from their time-course expression profiles. The proposed method assumes that a periodically expressed gene dataset is generated by a number of periodical processes which are modelled by a linear combination of trigonometric sine and cosine functions in time plus a Gaussian noise term. A two-stage method is proposed to estimate the model parameters, and 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. One synthetic dataset and two biological datasets were employed to evaluate the performance of the proposed method.

2. METHODS

2.1. Model for Periodically Expressed Gene Profiles

Let $x(t)$ $(t = 1, 2, \ldots, m)$ be a time-course gene expression profile generated from a periodical biological process, where m is the number of time points at which gene expression is measured. After shifting the mean of gene expression profiles to 0, the periodicity of this time-course gene expression profile can be modeled by a linear combination of trigonometric sine and cosine functions in time plus a Gaussian noise term as follows [21]

$$x(t) = a\cos(\omega t) + b\sin(\omega t) + \varepsilon(t), \tag{2.1}$$

where a and b are the coefficients of sine and cosine function, respectively; ω is the frequency of periodic expression data, and $\varepsilon(t)$ represent 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 is equivalent to sinusoidal function model [7, 8, 10–13]

$$x(t) = A\sin(\omega t + \Phi) + \varepsilon(t) \tag{2.2}$$

which are widely used to generate the synthetic periodic gene expression profiles [7] and to detect the periodically expressed genes [2, 8, 10–12]. In model (2.2), $A = \sqrt{a^2 + b^2}$ is called magnitude and $\Phi = \arctan(a/b)$ is called the phase.

Given a time-course gene expression profile $x(t)$ $(t = 1, 2, \ldots, m)$, estimating parameters a, b, and ω in model (2.1) is a nonlinear estimation problem as ω is nonlinear in the model. In general, all nonlinear optimization programs can be used to estimate parameters in model (2.1), for example, Gauss-Newton iteration method and its variants such as Box-Kanemasu interpolation method, Levenberg damped least squares methods, and Marquardt's method [22]. 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 (2.1)

$$x(t) = a\cos(\omega t) + b\sin(\omega t) \tag{2.3}$$

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

$$\ddot{x}(t) + \omega^2 x(t) = 0 \tag{2.4}$$

and that ω^2 is linear in equation (2.4) which is independent of a and b. Therefore, we propose the following two-step parameter estimation methods to estimate parameters a, b, and ω in model (2.2).

Step 1. Numerically calculate the second derivative of $x(t)$. Then, based on equation (2.4), use linear least squares method to estimate parameter ω^2. In details, let

$$X2 = [\ddot{x}(1), \ldots, \ddot{x}(l)]^T, \qquad X1 = [x(1), \ldots, x(l)]^T, \tag{2.5}$$

then, by the least squares method, ω^2 is estimated as

$$\widehat{\omega}^2 = -\frac{X1^T X2}{X1^T X1}, \qquad \widehat{\omega} = \sqrt{\widehat{\omega}^2}, \tag{2.6}$$

as time-course gene expression data are discrete, the second derivative $\ddot{x}(t)$ is estimated by the central finite difference formula as follows:

$$\ddot{x}(t) = \frac{x(t+1) + x(t-1) - 2x(t)}{\Delta^2} \quad \text{for} \quad t = 2, \ldots, m-1, \tag{2.7}$$

where Δ is time difference between two consecutive gene expression data points. From (2.7), the length of vectors $X2$ and $X1$ is $m-2$. Note that if the value of $\widehat{\omega}^2$ calculated by (2.6) for a gene is negative, this gene will be judged not to be periodically expressed.

Step 2. Substitute the estimated value of ω into (2.2). Apply the maximum likelihood method to model (2.1) to estimate parameters a and b. In detail, let

$$X = [x(1), \ldots, x(m)], \qquad A = \begin{bmatrix} \cos(\Delta\widehat{\omega}), \ldots, \cos(m\Delta\widehat{\omega}) \\ \sin(\Delta\widehat{\omega}), \ldots, \sin(m\Delta\widehat{\omega}) \end{bmatrix} \tag{2.8}$$

by the least squares method, a and b are estimated as

$$\begin{bmatrix} \widehat{a} \\ \widehat{b} \end{bmatrix} = \left(AA^T\right)^{-1}\left(AX^T\right). \tag{2.9}$$

2.2. Nonlinear Model-Based Clustering

2.2.1. The Mixture Model

In this study, it is assumed that a time-course gene expression dataset is a collection of periodically expressed gene profiles which belong to several clusters, and profiles in each cluster can be described by model (2.1) or (2.2) with different parameters. Let $\theta_k = [a_k, b_k, \omega_k, \sigma_k^2]$ be parameters of model (2.1) 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 (2.1) with parameters $\theta_k = [a_k, b_k, \omega_k, \sigma_k^2]$ $(k = 1, \ldots, K)$ which minimize

$$f(\Theta) = \sum_{k=1}^{K} \sum_{x \in C_k} \sum_{i=1}^{m} [x(i) - a_k \cos(i\Delta\omega_k) - b_k \sin(i\Delta\omega_k)]^2, \tag{2.10}$$

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

2.2.2. Estimation of Model Parameters

According to the parameter estimation method proposed in previous section for a single time-course expression profile, for the kth cluster parameters, $\theta_k = [a_k, b_k, \omega_k, \sigma_k^2]$ can be estimated as

$$\begin{aligned}
\widehat{\omega}_k^2 &= -\frac{\sum_{x \in C_k} X1^T X2}{\sum_{x \in C_k} X1^T X1}, \qquad \widehat{\omega}_k = \sqrt{\widehat{\omega}_k^2}, \\
\begin{bmatrix} \widehat{a}_k \\ \widehat{b}_k \end{bmatrix} &= \frac{1}{|C_k|} \sum_{x \in C_k} \left(AA^T\right)^{-1} \left(AX^T\right), \\
\widehat{\sigma}_k^2 &= \frac{1}{m|C_k|} \sum_{x \in C_k} \sum_{i=1}^{m} [x(i) - \widehat{a}_k \cos(i\Delta\widehat{\omega}_k) - b_k \sin(i\Delta\widehat{\omega}_k)]^2,
\end{aligned} \tag{2.11}$$

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

2.2.3. Algorithm

This study employs a relocation-iteration algorithm as shown in Algorithm 1 to estimate the parameters such that the cost function (2.10) is minimized. In 2(a) of Algorithm 1, Θ^t represents the estimated parameters in cost function (2.10) at iteration t while, in 2(b), parameters $\widehat{a}_k^t$, $\widehat{b}_k^t$, and $\widehat{\omega}_k^t$ represent the parameters of model k at iteration t.

2.3. Evaluation

In this study, we use the adjusted Rand index (ARI) [23] to evaluate the quality of the clustering. Consider two partitions of N objects: the r-cluster partition $U = \{u_1, \ldots u_r\}$ and the s-cluster partition $V = \{v_1, \ldots, v_s\}$. One may construct a contingency table (matrix) as in Table 1.

In Table 1, entry n_{ij} is the number of objects that are both in clusters u_i and v_j, $i = 1, \ldots, r$, $j = 1, \ldots, s$. Let $n_{i.} = \sum_{j=1}^{s} n_{ij}$ and $n_{.j} = \sum_{i=1}^{r} n_{ij}$ denote the sum of row i $(i = 1, \ldots, r)$ and the

(1) Select an initial partition for given the number of clusters, K
(2) Iteration ($t = 1, 2, \ldots$):
(a) estimate the parameter Θ^t based on the present partition by using (2.11);
(b) generate a new partition by assigning each sequence x to cluster k for which the value of $s^2 = \sum_{i=1}^{m} [x(i) - \widehat{a}_k^t \cos(i\,\Delta\widehat{\omega}_k^t) - \widehat{b}_k^t \sin(i\,\Delta\widehat{\omega}_k^t)]^2$ is minimum.
(3) Stop if the improvement of the cost function (2.10) is below a given threshold, the cluster memberships of time series do not change.

Algorithm **1:** Algorithm for nonlinear model-based clustering.

TABLE 1: Contingency table for two partitions of n objects.

	v_1	v_2		v_s	Total
u_1	n_{11}	n_{12}	$\cdots$	n_{1s}	$n_{1.}$
u_2	n_{21}	n_{22}	$\cdots$	n_{2s}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
u_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.s}$	$n_{..} = n$

sum of column j ($j = 1, \ldots, s$) in the contingency matrix, respectively, and let $V = N(N-1)/2$ (the number of pairs of N objects). Based on the contingency matrix of two partitions, the ARI is defined as [23]

$$\mathrm{ARI} = \frac{\sum_{i=1}^{r}\sum_{i=1}^{s}\binom{n_{ij}}{2} - (1/V)\sum_{i=1}^{r}\binom{n_{i.}}{2}\sum_{i=1}^{s}\binom{n_{.j}}{2}}{(1/2)\left[\sum_{i=1}^{r}\binom{n_{i.}}{2} + \sum_{i=1}^{s}\binom{n_{.j}}{2}\right] - (1/V)\sum_{i=1}^{r}\binom{n_{i.}}{2}\sum_{i=1}^{s}\binom{n_{.j}}{2}}. \tag{2.12}$$

The expected value of ARI is 1 when they matched perfect and 0 when the two partitions are selected at random.

If the true cluster labels for some dataset are known, the proposed clustering methods can be applied these datasets to obtain new cluster labels. Then ARI can be calculated for these two partitions. If ARI is close to 1, one can say that the proposed clustering method is in agreement with the true clusters. However, for real-life gene expression datasets, the true cluster labels are typically unknown. For this case, this study adopts a bootstrapping approach as shown in Algorithm 2 [20] to evaluate the proposed clustering methods. For the given number of clusters, K, the average ARI (AARI) reports the quality of the clustering result obtained from the evaluated clustering methods. Accordingly, the larger AARI, the better the quality of the clustering is, that is, the better the performance of the clustering method is.

3. EXPERIMENTAL RESULTS AND DISCUSSION

This study employs a synthetic dataset and two biological datasets to investigate the performance of the proposed method in different aspects.

(1) Repeat the following B times (where B is a preset integer number).
(a) Randomly divide the original dataset into two nonoverlapping sets, a learning set L, and a test set T.
(b) Apply the evaluated method to the learning set L to obtain a partition $P(\circ, L)$.
(c) Construct a predictor (classifier) $C(\circ, L)$ using the cluster labels from the partition $P(\circ, L)$.
(d) Apply the predictor C$(\circ, L)$ to the test set T to get the predicted partition $\widetilde{P}(\circ, T)$.
(e) Apply the evaluated method to the test set T to obtain a partition $P(\circ, T)$.
(f) Calculate the ARI of partitions $\widetilde{P}(\circ, T)$ and $P(\circ, T)$.
(2) Calculate the average ARI (AARI) over the B times as the measure index of the proposed clustering method.
(3) For the various number of clusters, K, repeat the procedure described in steps (1) and (2) above to get AARI(K), and then plot AARI(K) with respect to K.

Algorithm **2:** The procedure for evaluating proposed clustering method.

TABLE 2: Parameters for synthetic data.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
a_k	3.4397	6.9227	9.9126	12.1470	14.8819
b_k	2.3705	6.0603	8.9280	12.2379	14.8195
w_k	5.1516	3.1085	1.9359	1.5344	1.2413
σ_k	0.4000	0.4000	0.4000	0.4000	0.4000
n_k	136	300	279	239	120

3.1. Synthetic Dataset (SYN)

The synthetic dataset is generated by model (2.1). Let x_{it} be the simulated expression (log-ratio) values of gene i at time point t in the dataset, that is,

$$x_{it} = a_k \cos(w_k t) + b_k \sin(w_k t) + \varepsilon_{it}, \qquad i = 1, 2, \ldots, n,\ t = 1, 2, \ldots, m;\ k = 1, 2, \ldots, K, \tag{3.1}$$

where n is the number of genes, m is the number of time points, and K is the number of clusters.

In this study, parameters for synthetic data a_k, b_k, and w_k are randomly chosen as follows:

$$\begin{gathered} a_k \sim N\left(3k, d^2\right), \qquad b_k \sim N\left(3k, d^2\right), \qquad w_k \sim U\left(\frac{2\pi}{k+1}, \frac{2\pi}{k}\right), \qquad \varepsilon \sim N\left(0, \sigma^2\right), \\ K = 5,\ n = 20, \qquad n_k \sim U(100,\ 300), \end{gathered} \tag{3.2}$$

where n_k is the number genes in the kth cluster. The resulted parameters for synthetic data are shown in Table 2.

For various numbers of clusters, we run the proposed method described in Algorithm 1 with randomly chosen initial partitions, with the initial partitions from k-means results as and to the k-means

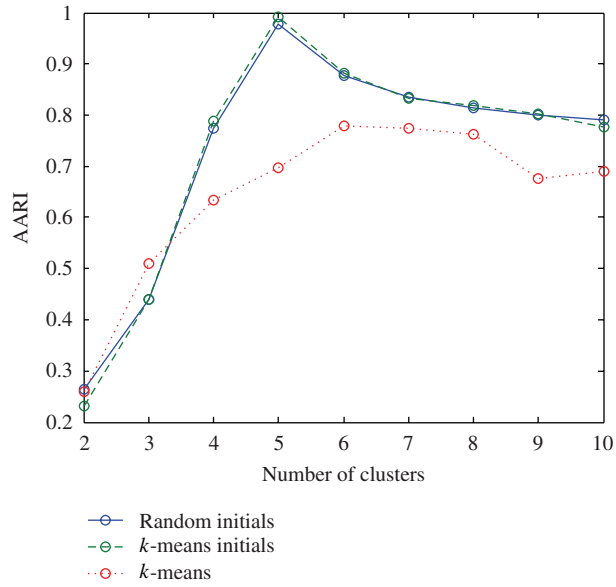

FIGURE 1: Plot of ARRI over different number of clusters for synthetic dataset.

TABLE 3: 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	0.2638	0.4391	0.7734	0.9778	0.8766	0.8361	0.8140	0.7991	0.7895
k-means	0.2586	0.5088	0.6335	0.6970	0.7794	0.7731	0.7620	0.6758	0.6904
k-means initial	0.2310	0.4391	0.7889	0.9913	0.8824	0.8328	0.8185	0.8027	0.7771

TABLE 4: The number of genes after filtering.

α	0.10	0.20
ELU	691	1207
BAC	471	658

methods. The ARI between clustering results and the known true cluster labels is calculated. The values of AARI are calculated over 20 runs and shown in the Table 3 and Figure 1.

From Figure 1, the proposed method with both initial partitions randomly chosen and those from k-means results has greater value of AARI than k-means when the number of clusters is greater than 3. Furthermore, when the number of clusters is the true value of 5, the AARI of the proposed method with both initial partitions reaches its maximum, which makes sense. However, the AARI of k-means method did not reach its maximum when the number of clusters is 5. Therefore, we can conclude that the proposed method outperforms the k-means in terms of AARI.

3.2. Real-Life Datasets

In this study, two real-life datasets are employed to illustrate the proposed method: ELU and BAC. ELU consist of expression profiles of 4304 genes without missing data. Expression profiles are obtained from yeast cell cycle division process through Eluration-synchronized experiments conducted by Spellman et al. [2]. Each expression profile has 14 equally spacing time points. BAC consists of expression profiles of 1590

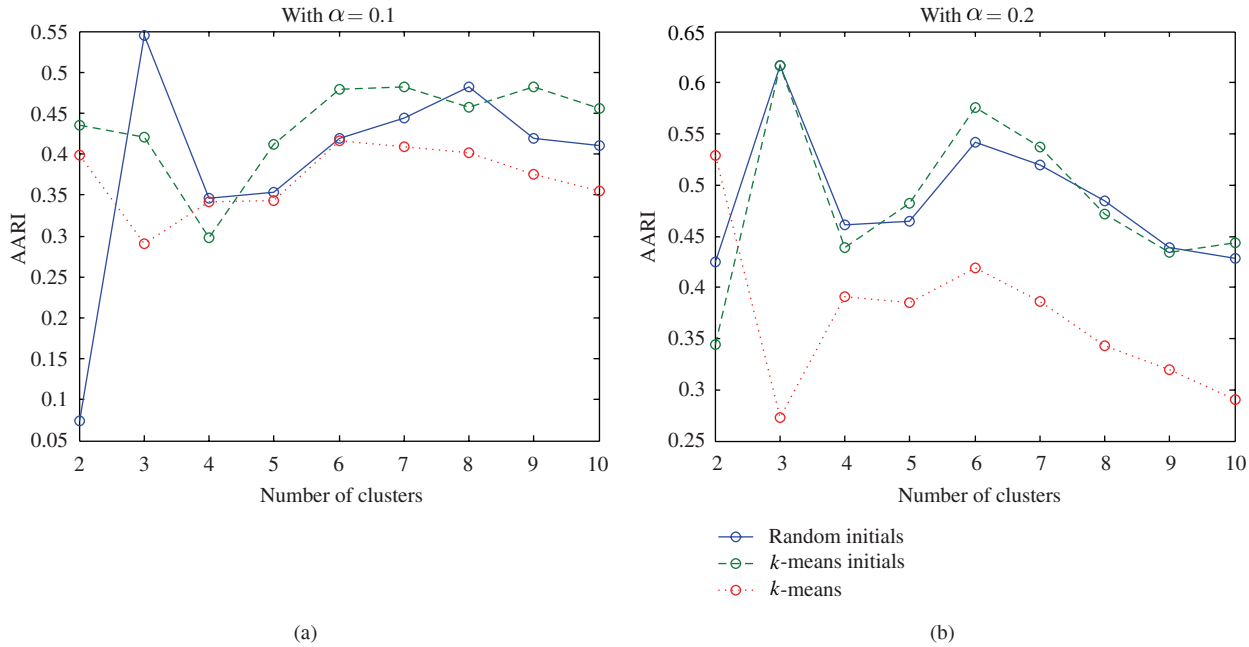

FIGURE 2: Plot of ARRI over different number of clusters for ELU after filtering.

genes without missing data. Expression profiles are measured during the cell cycle division process of the bacterium Caulobacter crescentus [3]. The measurements were taken at 11 equally spaced time points over 150 minutes. Both datasets are preprocessed in the following two steps.

Step 1. Shift the mean of each gene expression profile to 0.

Step 2. Filter the dataset with F-test at the significance level α, that is,

$$H_0 : x_i(t) = \varepsilon_{it}, \qquad H_1 : x_{it} = a\cos(wt) + b\sin(wt) + \varepsilon_t,$$
$$F = \frac{(R_{H_0} - R_{H_1})/2}{R_{H_1}/(m-2)} \sim F(2, m-2), \tag{3.3}$$

where R_{H_i} is the sum of squared errors under the specific hypothesis and m is the number of time points. Keep the genes which reject the null hypothesis (show periodical behaviours) [21].

After these two steps, the number of genes remains for different significant level as in Table 4. Then we run the evaluation procedure proposed in Algorithm 2 on these selected gene expression profiles. The AARIs of the proposed method and k-means over various numbers of clusters are plotted in Figures 2 and 3 for dataset ELU and BAC, respectively. From Figures 2 and 3 the results from both real-life datasets show that the proposed method outperforms the k-means in terms of AARI.

4. CONCLUSIONS

This paper has presented a nonlinear model-based method for clustering periodically expressed genes from their time-course expression profiles. In this method, profiles of periodically expressed genes and thus the cluster of profiles are modelled by a linear combination of trigonometric sine and cosine functions in time plus a Gaussian noise term which is equivalent to a sinusoidal function model [1–4, 6–13, 17–19]. Although

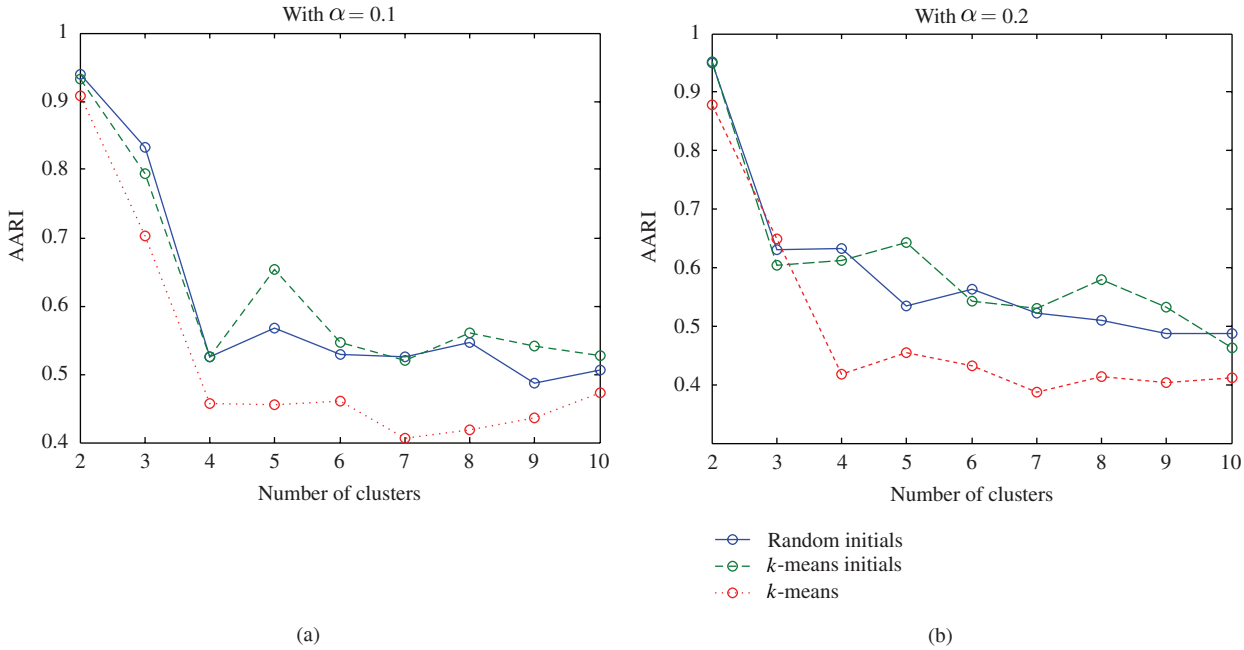

FIGURE 3: Plot of ARRI over different number of clusters for BAC after filtering.

this model is not new, the existing methods are not based on parameter estimation technique, especially not estimating the frequency in the model as it is nonlinear in parameter. In the presented method, a two step linear least squares method is proposed to estimate all model parameters including the frequency for each clusters. Computational experiments on one synthetic dataset and two biological datasets show that the proposed method outperforms the traditional clustering methods such as k-means in terms of AARI, which indicate that the proposed method can effectively cluster periodically expressed genes from their time-course expression profiles.

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

ACKNOWLEDGMENTS

This research is supported by 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 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, S. L. Chen, L. Shapiro, and H. H. McAdams, "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] V. Filkov, S. Skiena, and J. Zhi, "Analysis techniques for microarray time-series data," in *Proceedings of the 5th Annual Internatinal Conference on Computational Biology*, pp. 124–131, May 2001.

[6] C. J. Langmmead, A. K. Yan, C. R. McCung, and B. R. Donald, "Phase-independent Rhythmic analysis of genome-wide expression patterns," in *Proceedings of the Sixth Annual International Conference on Computational Biology*, pp. 1–11, 2011.

[7] 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.

[8] S. Wichert, K. Fokianos, and K. Strimmer, "Identifying periodically expressed transcripts in microarray time series data," *Bioinformatics*, vol. 20, pp. 5–20, 2004.

[9] R. A. Fisher, "Test of significance in harmonic analysis," *Proceedings of the Royal Society A*, vol. 125, pp. 54–59, 1929.

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

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

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

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

[18] 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.

[19] 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.

[20] 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.

[21] F. X. Wu, "Identification of periodically expressed genes from their time-course expression profiles," in *Proceedings of the International Symposium on Bioinformatics Research and Applications, (ISBRA '10)*, pp. 12–15, May 2010.

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

[23] 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.

This article should be cited as follows:

Li-Ping Tian, Li-Zhi Liu, Qian-Wei Zhang, and Fang-Xiang Wu, "Nonlinear Model-Based Method for Clustering Periodically Expressed Genes," *TheScientificWorldJOURNAL*, vol. 11, pp. 2051–2061, 2011.