Identification of Pseudo-Periodic Gene Expression Profiles

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, P.R. China
[2]Department of Mechanical Engineering, [3]Division of Biomedical Engineering,
University of Saskatchewan, 57 Campus Dr., Saskatoon, SK S7N 5A9, CANADA
*Corresponding author: faw341@mail.usask.ca

Abstract— **Time-course gene expression profiles associated with periodic biological processes should appear periodic. However, because of inherit problems with the experimental protocols measured gene expression data are actually pseudo-periodic, not exactly periodic. Therefore, identifying pseudo-periodically expressed gene from their time-course data could help understand the molecular mechanism of periodic biological processes. This paper proposes a method for identifying pseudo-periodic gene expression profiles. In the proposed method, a pseudo-periodic gene expression profile is modeled by a linear combination of trigonometric and exponential functions in time plus a Gaussian noise term. A two-step parameter estimation method is employed for estimating parameters in the model. On the other hand, non-pseudo periodic gene expression profiles are model by a constant plus a Gaussian noise term. The statistic F-testing is used to make a decision if a gene is pseudo-periodically expressed or not. Three biological datasets were employed to evaluate the performance of the proposed method. The results show that the proposed method can effectively identify pseudo-periodically expressed genes.**

Keywords: time-course gene expression profiles, pseudo-periodically expressed gene, parameter estimation, F-testing

I. INTRODUCTION

DNA microarray experiments have been employed to produce gene expression profiles at a series of time points. Such time-course gene expression data provides a dynamic snapshot of most (if not all) of the genes related to the biological development process. The analysis of such time-course gene expression data is helpful in understanding the mechanism of their associated biological process. Many time-course gene expression datasets have been collected from periodic biological processes. For periodic biological process, Furthermore, identifying periodically expressed gene from their time-course expression data could help understand the molecular mechanism of those biological processes [1,2].

In past decade, a number of methods have been proposed to identify periodically expressed genes. The discrete Fourier transform method is the earliest method for identifying periodically expressed genes [1, 2]. However, microarray experiments typically generate short time-course data. As pointed in [3], the frequency resolution by the discrete Fourier transform is often not adequate for resolving periodicities of interest. Recently periodic (trigonometric) functions are used to model periodic gene expression data.

There are typically two ways to match the models with data. In one way, many models with known parameters are created, and searching datasets is performed to find the expression profiles which match well with some of created models. For example, Authors in [4] proposed a method called CORRCOS which generates 101000 periodic synthetic models with different frequencies and phases. Each gene expression profile is compared to each of these 101000 models. The cross-correlation is used to measure the similarity between the synthetic model and gene expression profiles. The frequency and phase of the model most similar to the expression profile is assigned to the corresponding gene. Although it can identify periodically expressed gene, CORRCOS is too time-consuming and the cross-correlation is not real metric. Authors in [3] developed another algorithm named RAGE for detecting periodically expressed genes. Like

CORRCOS, RAGE is a synthetic model-based method. RAGE first estimates the frequency of expression profiles using autocorrelations of both the synthetic model and gene data. Then, RAGE generates a number of models with the estimated frequency over a variety of phases. The similarity between the synthetic model and gene expression profile is measured by a real metric called Hausdorff distance. Compared with CORRCOS, RAGE is less time-consuming [3].

These methods lack the statistical analysis. Wichert et al [5] proposed 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. Instead of estimating nonlinear parameters (frequency) in the model, they used the Fisher g-test to find the best frequency. Based on Fisher g-test, several similar method were also developed for identifying [6,7,8]. However, a recent research [9] concludes 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. In [9], the data length is said to be short if it is less than 40 data points. By this criterion, most gene expression profiles are too short to use Fisher g-test. In addition, it is hard in practice to obtain gene expression profiles with an integer number of periods as the period might be unknown before collecting the data.

In another way, models with unknown parameters are employed and unknown parameters are estimated based on the data such that the models with estimated parameters match well with the data. However, it is challenging to estimate parameters which are nonlinear in a model such as trigonometric function. Recently we proposed a two-step parameter estimation method to estimate all parameters in trigonometric) function models from gene expression profiles [10, 11]

In principle, expression profiles associated with periodic processes should appear periodic. However, because of inherit problems with gene expression experimental protocols [1,12, 13], measured gene expression data are actually pseudo-periodic, not exactly periodic. In this paper, a method is proposed for identifying pseudo-periodic gene expression profiles. In the proposed method, a pseudo-periodic gene expression profile is modeled by a linear combination of trigonometric and exponential functions in time plus a Gaussian noise term. This model is more complex than the one in [10, 11]. A new two-step parameter estimation method is employed for estimating parameters in the model. On the other hand, non-pseudo periodic gene expression profiles are model by a constant plus a Gaussian noise term. The statistic F-testing is used to make a decision if a gene is pseudo-periodically expressed or not. Three biological datasets were employed to evaluate the performance of the proposed method.

II. METHODS

In this section, we first propose the model for pseudo-periodic gene expression profiles and then describe a two-step parameter estimation method for the proposed model. Finally a hypothesis testing is described to make a decision whether a gene expression profile is pseudo-periodic or not.

2.1 Model for pseudo periodic gene expression profiles

Let $x(t)$ (t=1,2,…, m) be a time-course gene expression profile generated from a periodic biological process, where m is the number of time points at which gene expression is measured. In this study, we always shift the mean of gene expression profiles to 0. To model pseudo-periodic gene expression profile, we adopt the linear combination of trigonometric and exponential functions plus a Gaussian noise term as follows:

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

where a and b are the coefficients of sine and cosine function, respectively; α is the decrease (increase) rate; 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. When α =0, model (1) becomes

$$x(t) = a\cos(\omega t) + b\sin(\omega t) + \varepsilon(t)$$

$$\text{or} \quad x(t) = A\sin(\omega t + \Phi) + \varepsilon(t) \quad (2)$$

which are widely used to generate the synthetic periodic gene expression profiles[1-9]. However, because of inherit problems with gene expression experimental protocols we believed that model (1) is more reasonable.

Given a time-course gene expression profile $x(t)$ (t=1, 2,…, m), estimating parameters a, b, α and ω in model (1) is a nonlinear estimation problem as α and ω is nonlinear in the model. Nonetheless, our observation is that noise-free model (1)

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

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

$$\ddot{x}(t) + 2\alpha\dot{x}(t) + \gamma^2 x(t) = 0 \quad (4)$$

where $\gamma^2 = \omega^2 + \alpha^2$ and equation (4) is independent of a and b. Note that α and γ^2 are linear in equation (4) while a and b are linear in model (1). Therefore, we propose the following two-step parameter estimation methods to estimate parameters a, b, α and ω in model (1):

Step1: Based on equation (4), use linear least squares method to estimate parameters α and γ^2, thus α and ω. In detail, let

$$X2 = \begin{bmatrix} \ddot{x}(1) \\ \vdots \\ \ddot{x}(l) \end{bmatrix}, \text{ and } \quad X1 = \begin{bmatrix} \dot{x}(1) & x(1) \\ \vdots & \vdots \\ \dot{x}(l) & x(l) \end{bmatrix}$$

then by the least squares method, α and γ^2 are estimated as

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\gamma}^2 \end{bmatrix} = -(X1^T X1)^{-1} X1^T X2 \tag{5}$$

and thus ω is estimated

$$\hat{\omega} = \sqrt{\hat{\gamma}^2 - \hat{\alpha}^2} \tag{6}$$

As time-course gene expression data are discrete, the first and second derivatives $\dot{x}(t)$ and $\ddot{x}(t)$ are estimated by the central finite difference formula, respectively, as follows

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

$$\ddot{x}(t) = \frac{x(t+1) + x(t-1) - 2x(t)}{\Delta^2} \quad \text{for t =2,…, m-1} \tag{8}$$

where Δ is time difference between two consecutive gene expression data points. From equations (7) and (8), *l*=m-2. Note that equations (7) and (8) are for evenly spaced time-course data. For unevenly spaced time-course data, equation (7) and (8) should be replaced by a modified formula which can be found in any numerical method textbooks. If the value of $\hat{\gamma}^2 - \hat{\alpha}^2$ calculated by (5) for a gene is negative, this gene will be judged not to be periodically expressed.

Step2: Substitute the estimated values of α and ω in Step 1 into equation (1). Apply the least squares method to model (1) to estimate parameters a and b. In detail, let

$$X = \begin{bmatrix} x(1) \\ \vdots \\ x(m) \end{bmatrix}, \text{and} \quad A = \begin{bmatrix} \cos(\Delta\hat{w}) & \sin(\Delta\hat{w}) \\ \vdots & \vdots \\ \cos(m\Delta\hat{w}) & \sin(m\Delta\hat{w}) \end{bmatrix}$$

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

$$\beta = \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = (A^T A)^{-1} (A^T X) \tag{9}$$

2.2 Hypothesis testing

To determine if a gene is pseudo-periodically expressed, we test the null hypothesis of

$$H_0: \quad x(t) = \varepsilon(t) \tag{10}$$

versus the alternative hypothesis of

$$H_a: \quad x(t) = e^{\alpha t}[a\cos(\omega t) + b\sin(\omega t)] + \varepsilon(t) \tag{1}$$

In terms of the following F-statistic

$$F = \frac{m-2}{2}\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2} - 1\right) \tag{11}$$

where $\hat{\sigma}_0^2$ is the estimated variance of white noise in model (10) and is calculated as

$$\hat{\sigma}_0^2 = \frac{1}{m-1} XX^T \tag{12}$$

and $\hat{\sigma}_1^2$ is the estimated variance of white noise in model (1) and is calculated as

$$\hat{\sigma}_1^2 = \frac{1}{m-1}\left[X^T - A^T\beta\right]^T\left[X - A^T\beta\right] \tag{13}$$

As noise terms in both model (1) and (10) are normal white noise, F-statistic (11) follows the F-distribution with the degrees of freedom (2, m-2), according to statistics theory. When the value of F-statistic is large enough (greater than a threshold), model (10) is rejected, i.e., the gene expression profile exhibits periodic behaviour, and otherwise the gene expression profile appears white noises. According to degrees of freedom (i.e., the length of time-course data m) 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 standard MatLab function icdf('f', 1-α, 2,m-2), where α is the significance level. If a significance level associated with a gene is smaller than the preset significant level, the genes are judged to be pseudo-periodic, and otherwise it is not.

III. EXPERIMENTAL RESULTS AND DISCUSSION

This study employs the following three biological datasets to investigate the performance of the proposed method.

Eluration-synchronized gene expression data of the yeast (ELU): Spellman et al. [1] studied the mitotic cell division cycle of yeast and monitored more than 6000 genes of yeast (*Saccharomyces cerevisiae*) at 14 equally-spacing time points in the eluration-synchronized experiment. Genes with missing data were excluded in this study. The resultant dataset contains the expression profiles of 5766 genes.

Alpha-synchronized gene expression data of the yeast (ALPHA): Spellman et al. [1] studied the mitotic cell division cycle of yeast and monitored more than 6000 genes of yeast (*Saccharomyces cerevisiae*) at 18 equally-spacing time points in the alpha-synchronized experiment Genes with missing data were excluded in this study. The resultant dataset contains 4489 expression profile of 4489 genes.

Bacterial cell cycle (BAC): This dataset contains gene expression measurements during the bacterial cell cycle division process for about 3000 predicted open reading frames, representing about 90% of all bacterium Caulobacter crescentus genes [2]. The measurements were taken at 11 equally-space time points over 150 minutes. Genes with missing data were excluded in this study. The resultant dataset contains the expression profile of 1593 genes.

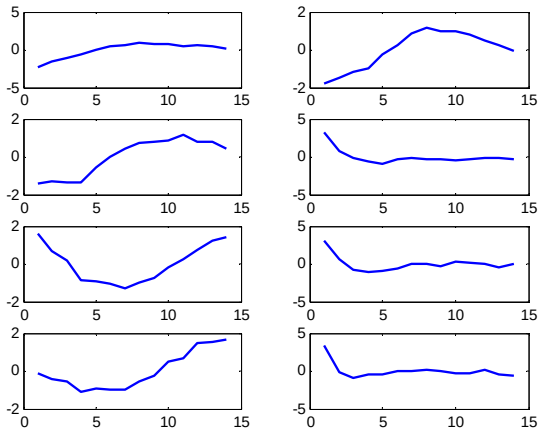

Figure 1. 8 gene profiles identified to be pseudo-periodically expressed in ELU dataset

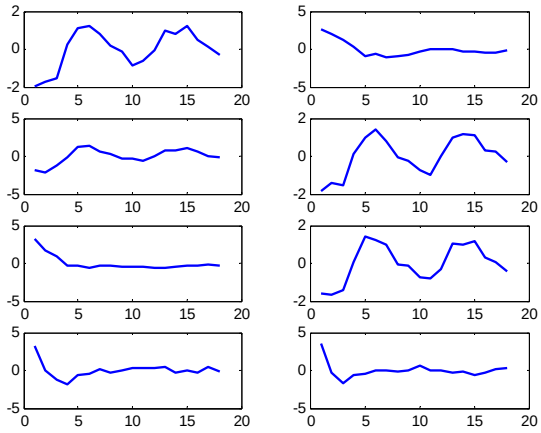

Figure 2. 8 gene profiles identified to be pseudo-periodically expressed in ALPHA dataset

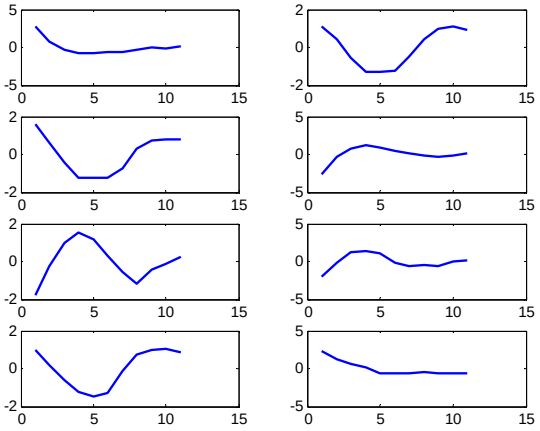

Figure 3. 8 gene profiles identified to be pseudo-periodically expressed in BAC dataset

The proposed method is applied to these three datasets. Figures 1-3 show the 8 gene profiles identified to be pseudo-periodically expressed from these datasets, respectively. From these figures, we can see these gene expression profiles appear pseudo-periodic. Most of gene expression profiles look more periodic, in whose models the values of the decrease (increase) rate α is small. Others look less, in whose models the values of the decrease (increase) rate α is dominant.

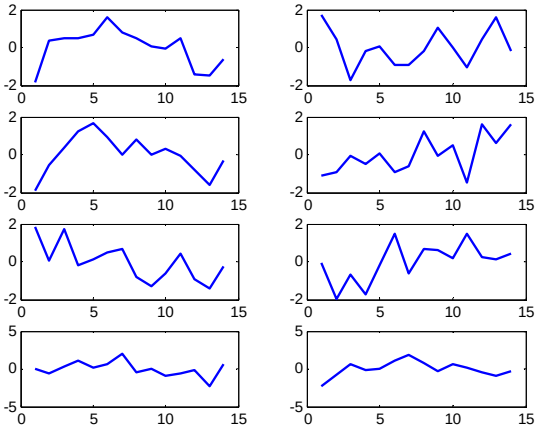

Figure 4. 8 gene profiles identified not to be periodically expressed in ELU dataset

Figures 4-5 shows show the 8 gene profiles identified to be non-pseudo-periodically expressed from ELU and ALPHA datasets (Figure for BAC is omitted because of space limitation), respectively. These gene expression profiles really look random noises.

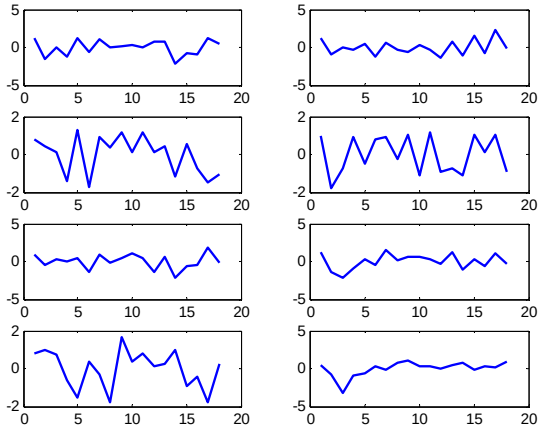

Figure 5. 8 gene profiles identified not to be periodically expressed in ALPHA dataset

IV. Conclusion and Future Work

The linear combination of trigonometric and exponential functions has proposed to model pseudo-periodic gene expression profiles. A two step linear least squares method is proposed to estimate all model parameters. In addition, the proposed method uses F-test to determine if a gene expression profile appears pseudo-periodic or not. Computational experiments on three biological datasets have showed that the proposed method can effectively identify periodically expressed genes from their time-course expression profiles.

In this paper, the performance of the propose method is evaluated by manually checking some of results, for example, showing the profiles identified to be pseudo-periodic or those identified not to be pseudo-periodic. In the future, more objective criteria should be used to evaluate from both bioinformatic and biological view of points. In addition, this paper does not evaluate the proposed method on gene expression profiles. Another direction of feature work is to perform cluster analysis of gene expression data based proposed models.

Acknowledgment: This study was supported by Base Fund of Beijing Wuzi University and Fund for Beijing Excellent Team for Teaching Mathematics through the first author and by Natural Science and Engineering Research Council of Canada (NSERC) through the second and third authors.

REFERENCES

[1] Spellman, PT, et al.: Comprehensive identification of cell cycle-regulated genes of the Yeast Saccharomyces cerevisiae by microarray hybridization. Molecular Biology of the Cell 9 (1998) 3273-3297

[2] Laub, MT, et al: Global analysis of the genetic network controlling a bacteria cell cycle. Science 290(2000) 2144-2148

[3] Langmmead CJ, Yan A K, McCung C R, and Donald B. R.: Phase-independent Rhythmic analysis of genome-wide expression patterns, Proceedings of the 6th Annual Internat-ional Conference on Computational Biology (2002) 1-11

[4] Harmer S, et al.: Orchesterated transcription of key pathways in Arabidopsis by the circadian clock. Science 290 (2000) 2110-2113

[5] Wichert S, Fokianos K and Strimmer K: Identifying periodically expressed transcripts in microarray time series data. Bioinformatics 20 (2004) 5-20

[6] Chen J: Identification of significant period genes in microarray gene expression data. BMC bioinformatics 6 (2005) 286

[7] Glynn EF, Chen J, and Mushegian AR: Detecting periodic patterns in unevenly spaced gene expression time series using Lomb-Scargle Periodogrms. Bioinformatics 22 (2006) 310-316

[8] Chen J and Chang KC: Discovering Statistically significant period Gene expression. International Statistical Review 76 (2008) 228-246

[9] Liew AWC, et al.: Statistical power of fisher test for the detection of short periodic gene expression profiles. Pattern Recognition 42 (2009) 549-556

[10] FX Wu (2010): Identification of Periodically Expressed Genes from Their Time-Course Expression Profiles, ISBRA10(short paper): 12-15

[11] LP Tian, LZ Liu, FX Wu (2010): Parameter estimation method for Periodical Gene Identification, iCBBE2011, accepted

[12] Duggan D.J., et al. (1999) Expression profiling using cDNA microarrays. Natural Genetics 21(Sup1): 10-14.

[13] Eisen M.B. and Brown, P.O. (1999) DNA arrays for analysis of gene expression. Methods Enzymol 303: 179-205.