Analysis of Mouse Periodic Gene Expression Data Based on Singular Value Decomposition and Autoregressive Modeling

Tsz Yan Tang, Wee Chung Liew, and Hong Yan

***Abstract*—Each DNA microarray experiment generates a large amount of gene expression profiles and it remains a challenge for biologists to robustly identify periodic gene expression profiles with certain noise level in the data. In this paper, we propose a new scheme with noise filtering technique to analyze the periodicity of gene expression base on singular value decomposition (SVD), singular spectrum analysis (SSA) and autoregressive (AR) model-based spectrum estimation. With the combination of these methods, noise can be filtered out and over 85% of periodic gene expression can be identified in mouse presomitic mesoderm transcriptome data set.**

***Index Terms*—Autoregressive (AR) model, DNA microarray gene expression data, singular value decomposition (SVD), singular spectrum analysis (SSA), time series analysis.**

I. Introduction

DNA sequence is a succession of letters which carries the genetic information of living organism. A DNA microarray or DNA chip consists of an arrayed series of spots, each of which conveys a partition of DNA sequence. In the microarray experiment, thousands of gene expression levels are recorded simultaneously to study the functions of genes, the effects of certain therapy, illness, and developmental processes [1]-[2]. With microarray technology, genome gene expression data are been generated at rapid rate. Biologists are interested in identifying the characteristics, trends, and patterns of the gene expression profiles. However, each gene expression profile usually contains certain amount of noise. It remains a main challenge to identify periodic gene expression profiles especially when the number of data points is small and the level of noise is high.

The microarray data used in this paper is recorded from the mouse presomitic mesoderm transcriptome [2] which is generated to study the developmental process of mouse embryo. Presomitic mesoderm (PSM) is the embryonic tissue composed of mesoderm which is the source of muscle and bone and is divided into somites later during the segmentation process. According to [2], this process involves a molecular oscillator, the segmentation clock, which produces time series signal in PSM rhythmically [3]. PSM samples from 40 mouse embryos are collected and the lunatic fringe (*Lfng*) expression patterns are used as a proxy to select 17 samples of different time points which involving an entire oscillation period. Based on this dataset, a research study is carried out to compare the pattern detection ability of several mathematical approaches, which included Lomb-Scargle (L) periodogram, Phase consistency (P), Address reduction (A), Cyclohedron test (C), and Stable persistence (S). The probe sets were ranked based on the power ratio using these five methods and the results show that the Stable persistence (S) method has the best performance by identifying most of the benchmark probe sets within the top 300 probe sets [4]. Nevertheless, microarray data usually contain a high level of noise and the performance is degraded with most pattern analysis algorithms. Therefore, we need to develop a useful method to process the noisy time series data.

We propose an effective method in this paper to detect the periodicity of microarray time series data by combining singular value decomposition (SVD), singular spectrum analysis (SSA) and autoregressive (AR) model-based spectral analysis. By considering the singular values of time series data, trend component is extracted effectively [5]; and using AR modeling, more accurate results are generated [6]. About 85% of genes expression profiles in a mouse PSM dataset are found to be periodic. A comparison is made to investigate the effectiveness of noise reduction using SVD, SSA and AR modeling.

II. Methods

A. Dataset

In this research, we use the dataset downloaded from http://www.ebi.ac.uk/microarray-as/ae/ under the accession ID of E-TABM-163. It contains 22,690 probe sets, each of them have 17 samples [2]. The data pre-processing is applied before the research carry on. The data is first normalized to zero mean and then filtered out based on three criteria: the detection call (taking out the probe sets called “absent” and “marginal”), the maximum signal intensity (removing the genes with expression level less than 50), and the peak-to-peak amplitude (less than 1.65). After the data pre-processing operations, 10025 probe sets remain to be analyzed.

Manuscript received December 8, 2009. This work is supported by the Hong Kong Grant Research Council (Project CityU 122607).

T. Y. Tang is with the Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong (email: vtytang@student.cityu.edu.hk).

W. C. Liew is with the School of Information & Communication Technology, Gold Coast Campus, Griffith University, QLD 4222, Australia (email: a.liew@griffith.edu.au).

H. Yan is with the Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong, and also with the School of Electrical and Information Engineering, University of Sydney, Dydney, N.S.W. 2006, Australia (email: h.yan@cityu.edu.hk).

B. Noise Reduction Using Singular Spectrum Analysis

We will introduce an SSA based algorithm to reduce the noise of microarray time series data. The time series data from microarray is usually short and noisy. Although we can analyze the periodicity of the gene expression profiles directly, the results will be degraded by noise significantly. Therefore, before performing periodic detection, a pre-processing technique is needed to reduce the noise level.

SSA is proposed for the purpose of reconstructing the attractive component from the experimental data [7]. It is a model free approach because it decomposes an original time series to trend and noise according to the singular value decomposition (SVD) [8]. Assume there is a time series data ($y_1,\ldots y_p,\ldots y_n$), which is reorganized as an AR(p) model representation, where p is the order of the AR model, and n is equal to 17, which is the number of samples of gene expression profiles.

The order p determines the number of equations we can have [9]. Usually, the more the number of equations, the more accurate of the results we can have. We took p equals to 8 since we can form the largest number of linear equations this way. However, we have only 17 time points in the dataset and it provides 9 equations which are not enough to estimate the AR coefficients reliably. To solve this problem, we use the forward-backward linear prediction method instead of forward or backward prediction to double the number of equations. Thus, the resultant AR coefficients can be estimated accurately [10]. The matrix form of the forward-backward linear system can be written as:

$$\begin{bmatrix} y_{p+1} \\ y_p \\ \vdots \\ y_n \\ y_1 \\ y_2 \\ \vdots \\ y_{n-p} \end{bmatrix} = -\begin{bmatrix} y_p & y_{p-1} & \cdots & y_1 \\ y_{p+1} & y_p & \cdots & y_2 \\ \vdots & \vdots & \ddots & \vdots \\ y_{n-1} & y_{n-2} & \cdots & y_{n-p} \\ y_2 & y_3 & \cdots & y_{p+1} \\ y_3 & y_4 & \cdots & y_{p+2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n-p+1} & y_{n-p+2} & \cdots & y_n \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{p-1} \\ a_p \end{bmatrix} \qquad (1)$$

The upper part in (1) is called forward prediction and the lower part in (1) is the backward prediction. By combining these two linear prediction method, the linear system in (1) becomes more stable and reliable.

A common limitation of the AR modeling method is the high bias if the prediction order is low; and the presence of false peaks in these frequency spectrums when a high prediction order is used. The problem can be solved by using SVD, which is the foundation of SSA [11]. We apply SVD to the rectangular matrix of reorganized gene expression profile (defined as **Y**, where$\mathbf{Y} \in \mathbb{R}^{2(n-p)\times p}$). Rewriting (1) as below,

$$\mathbf{y} = -\mathbf{Ya} \qquad (2)$$

in which both **Y** and **y** are known [12]. SVD adopts the computing method of least square and the pseudo inverse of matrix **Y**. **Y** can be decomposed to

$$\mathbf{Y} = \mathbf{USV}^{\mathrm{T}} \qquad (3)$$

where $\mathbf{U} \in \mathbb{R}^{2(n-p)\times 2(n-p)}$, $\mathbf{V} \in \mathbb{R}^{p\times p}$, and **S** has the same dimension matrix as **Y**. **S** has non-zero values only in its diagonal entries which is called the singular value of **Y** and equal to the square roots of eigenvalues from $\mathbf{YY}^{\mathrm{T}}$ or $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$. The singular values are always real positive numbers and arranged in descending order.

$$\mathbf{S} = \begin{pmatrix} s_1 & 0 & \cdots & 0 & 0 \\ 0 & \ddots & 0 & 0 & 0 \\ \vdots & 0 & s_k & 0 & \vdots \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}, \text{ where } s_1>s_2\ldots>s_k \qquad (4)$$

SVD is a powerful tool to separate the data of interest and the noise. Typically, the large singular values represent the interesting information in the time series signal. Therefore, by zeroing the small singular values, we can extract the trend component and noise component [13].

The important consideration in SSA is the selection of the number of singular values. Usually, the first few numbers of singular values are large and followed by some very small singular values. The leading singular values of **Y** contain the most amount of energy and the small singular values are considered as noise level [14]-[15].

By choosing a different number of leading singular values from matrix **Y**, we can compute (3) again from right to left to get a new matrix $\mathbf{Y}'$. The new matrix contains the useful information of the time series data only. The number of singular values of **Y** retained is varied since every gene expression profile carries a different amount of noise to achieve the best noise filtering ability.

The SSA based procedure is performed six times on each genome gene expression using a different number of leading singular values from 3 to 8, and the best result is recorded. Then the time series data ($y_1,\ldots y_p,\ldots y_n$) is reconstructed by averaging the elements of matrix $\mathbf{Y}'$ over the diagonal.

C. The AR Model-Base Power Spectrum Estimation

Power spectral density (PSD) is simply the spectrum of the time series sequences which describes the power distribution of the signal. In genome-wide gene expression cell-cycle identification, PSD analysis is one of the useful techniques. If the time series signal is highly periodic, the resultant power spectrum has sharp peaks at the corresponding frequency [9], [16]. PSD can be easily found by applying Fast Fourier Transform (FFT) to the time series data. However, if the time series data is too short, the FFT power spectrum estimation PSD estimation will be degraded due to the so-called windowing artifacts. Therefore, FFT is not suitable for microarray data analysis. Instead of using it, the AR model-based spectrum estimation is adopted in our research.

According to (3), the AR coefficient $\mathbf{a}^{\mathrm{T}}= (a_1, a_2\ldots, a_p)^{\mathrm{T}}$, is given by

$$\mathbf{a} = -\mathbf{VS}^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{y} \qquad (7)$$

where $\mathbf{S}^{-1}$ is the pseudo inverse of **S** .According to the basic properties of the matrix inverse, **S** is a diagonal matrix with diag ($s_1, s_2,\ldots, s_k$), then $\mathbf{S}^{-1}$ is equal to diag ($s_1^{-1}, s_2^{-1},\ldots, s_k^{-1}$) [17]. Once the AR coefficient is estimated, the spectrum of the time series data is given by

$$\mathrm{P}(\omega) = \frac{T\sigma^2}{\left|1 + \sum_{r=1}^{p} a_r \mathrm{e}^{-j\omega r T}\right|^2} \qquad (8)$$

where ω is the angular frequency in the range (0, π), T is the sampling interval, σ^2 is the variance of the noise, p is the order of the AR model and a_r are the AR coefficients. The AR(p) spectral estimator is consistent if the given process is truly autoregressive of order p [18]-[19]. The power spectrum density function is normalized and is bounded in the range of [0, 1].

D. The Periodicity Detection using Power Spectrum Width

A periodic time series signal gives a peak spectrum in its frequency domain power spectral density at its corresponding frequency. The width of the peak is used as the criterion to detect the periodicity of gene expression profiles. Consider a sharp spectrum located at f_i, the width of the frequency band $[f_{i-1}, f_{i+1}]$ is estimated, where f_{i-1} and f_{i+1} are the frequencies at 90% decay of the peak. If the width is sufficiently small, the time series signal is said to be highly periodic. According to the width of each spectrum, we can rank the whole microarray dataset and determine how many genes are periodic. We restrict the estimated width to be 30% the of total width which is equal to 0.3π, since the spectrum with large width is considered as lacking in periodicity and can be discarded. We normalize the width of the power spectrum to [0, 1]. Thus, if the normalized width of the power spectrum is less than 0.1, the corresponding profile is detected as periodic.

To summarize, our method consists of two parts. The first part is to filter out the noise of the time series data and the second part is to detect the periodicity. First, each gene expression profiles are formed as an AR model in matrix form. The forward and backward linear prediction is used to increase the number of equations in the AR model. SVD is performed to obtain the singular values of the system. Noise is filtered by zeroing some of the singular values which are small enough. The noise filtered time series data is reconstructed based on the remaining singular values. In the second part, the AR coefficients and power spectrum density are calculated according to (7) and (8) respectively. Finally, the width of the power spectrum is ranked to detect the periodic time series signal.

III. RESULTS

A. Noise Reduction using SSA

We have tested our algorithm with 10025 gene expression profiles of mouse Presomitic Mesoderm Transcriptome Data. We have utilized the SVD, SSA and AR modeling to do the noise filtering. In order to verify the performance of noise reduction using our algorithm, the spectrum width of each gene expression profiles which without using SSA are calculated first. Then, we apply our algorithm to the entire dataset. The power spectrum width of reconstructed expression profiles is recorded. We compare the spectrum width of the gene expression profiles which using SSA and without using SSA.

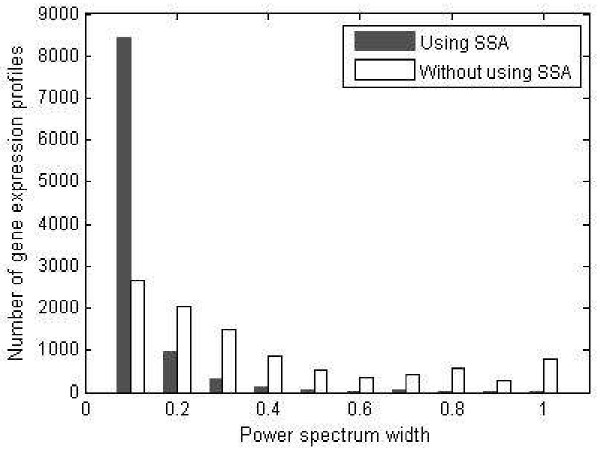

Fig. 2. Comparison of the spectrum width between using SSA and without using SSA.

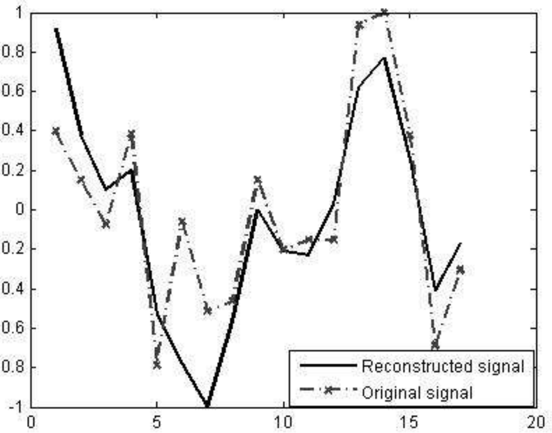

Fig. 1. The signal intensity of probe set named 1450818_a_at before and after SVD and SSA based reconstruction.

Fig. 1 shows the comparison of the spectrum width distribution of gene expression profiles between using SSA and without using SSA. The number of gene expression profiles with spectrum width less than 0.1 is 2663 if SSA is not applied. However, after we adopt our algorithm to do the noise filtering, the total number of spectrum width less than 0.1 increases from 2663 to 8445. The width of the power spectrum of mouse presomitic mesoderm profiles is mainly located within the range of 0 to 0.1.

By applying our algorithm, it can be seen that a total of 8445 genes are determined to be periodic, which is about 85% of the total microarray dataset. Fig. 2 shows an example of the reconstructed expression profile compared with the original signal without using SSA noise filtering technique. The solid black line representing the reconstructed signal indeed looks like sinusoidal where the dotted line on behalf of the original signal has few ripples which considered as the noise level. By applying our algorithm, noise is removed using SVD and SSA. After the noise is removed, the periodic gene expression profiles are detected easily. Therefore, we can conclude that using AR modal based power spectrum estimation can more effectively filter out the noise and detect the periodic genes.

B. Selection of Leading Singular Values

In computing the SVD, one critical criterion is the selection

TABLE I
The number of leading singular values to reconstruct the periodic gene expression level in murine presomitic mesoderm

No. of leading singular values	No. of expression profiles
3	739
4	3011
5	1974
6	1397
7	852
8	472
Total periodic genes	8445

of the number of leading singular value to reconstruct the signal. The singular values contain the power of the trend components and the noise signal, where the expression profiles in the microarray dataset may consist of a different amount of noise. Table I shows the number of leading singular value we used to reconstruct the periodic signal in Mouse Presomitic Mesoderm Transcriptome Data. It indicates that difference profiles require different number of singular values to preserve trend while suppressing noise. There are totally 8445 periodic gene expression levels. Within these 8445 periodic genes, we observe that if we choose the number of leading singular values equals to four and five, about 60% of expression profiles produce the best result, which implies that over half of the periodic genome profiles contain most of their energy in the first four and five leading singular values.

IV. Conclusion

In general, gene expression profiles in microarray dataset have short length and the signal contain varying amounts of noise. In order to extract the interesting component from the short noisy time series signal, we have proposed a new algorithm which combines with SVD, SSA and AR modeling. After applying our algorithm, the noise can be effectively reduced and periodic trend component can be detected easily. We have considered the presence of sharp spectral peak from the AR spectrum density to detect the periodic genome expression profiles. From the results, we observed that our proposed method can detect over 85% of periodic genes from the murine presomitic mesoderm expression profiles.

References

[1] J. DeRisi, *et al.*, "Use of a cDNA microarray to analyse gene expression patterns in human cancer," *Nature Genetics,* vol. 14, pp. 457-460, Dec 1996.

[2] M. L. Dequeant, *et al.*, "A complex oscillating network of signaling genes underlies the mouse segmentation clock," *Science,* vol. 314, pp. 1595-1598, Dec 8 2006.

[3] M. L. Dequeant and O. Pourquie, "Segmental patterning of the vertebrate embryonic axis," *Nat Rev Genet,* vol. 9, pp. 370-82, May 2008.

[4] M. L. Dequeant, *et al.*, "Comparison of pattern detection methods in microarray time series of the segmentation clock," *PLoS One,* vol. 3, p. e2856, 2008.

[5] D. S. Watkins, *Fundamentals of matrix computations*, 2nd ed. ed. New York ; [Great Britain]: Wiley-Interscience, 2002.

[6] H. Yan and T. D. Pham, "Spectral estimation techniques for DNA sequence and microarray data analysis," *Current Bioinformatics,* vol. 2, pp. 145-156, May 2007.

[7] D. S. Broomhead and G. P. King, "Extracting Qualitative Dynamics from Experimental-Data," *Physica D,* vol. 20, pp. 217-236, Jun-Jul 1986.

[8] N. K. Myung, "Singular Spectrum Analysis," Master of Science, Department of Statistics, Univarsity of California, Los Angeles, 2009.

[9] N. Golyandina, *et al.*, *Analysis of time series structure : SSA and related techniques*. London: Chapman & Hall, 2001.

[10] M. K. Choong, *et al.*, "Autoregressive-Model-Based Missing Value Estimation for DNA Microarray Time Series Data," *Ieee Transactions on Information Technology in Biomedicine,* vol. 13, pp. 131-137, Jan 2009.

[11] D. W. Tufts and R. Kumaresan, "Estimation of Frequencies of Multiple Sinusoids - Making Linear Prediction Perform Like Maximum-Likelihood," *Proceedings of the Ieee,* vol. 70, pp. 975-989, 1982.

[12] J. Lee, "Autoregressive parameter estimation with embedded order selection in arbitrary noise," *Dissertation Abstracts International,* vol. 66-10, p. 5588, 2005.

[13] L. Du, *et al.*, "Spectral analysis of microarray gene expression time series data of Plasmodium falciparum," *Int J Bioinform Res Appl,* vol. 4, pp. 337-49, 2008.

[14] R. Vautard and M. Ghil, "Singular Spectrum Analysis in Nonlinear Dynamics, with Applications to Paleoclimatic Time-Series," *Physica D,* vol. 35, pp. 395-424, May 1989.

[15] L. Liu, *et al.*, "Robust singular value decomposition analysis of microarray data," *Proc Natl Acad Sci U S A,* vol. 100, pp. 13167-72, Nov 11 2003.

[16] A. W. Liew, *et al.*, "Spectral estimation in unevenly sampled space of periodically expressed microarray time series data," *BMC Bioinformatics,* vol. 8, p. 137, 2007.

[17] J. R. Schott, *Matrix analysis for statistics*, 2nd ed. Hoboken, N.J.: Wiley-Interscience, 2005.

[18] B. Porat, *Digital processing of random signals : theory and methods*. Englewood Cliffs, N.J.: Prentice Hall ; London : Prentice-Hall International, 1994.

[19] L. K. Yeung, *et al.*, "Dominant spectral component analysis for transcriptional regulations using microarray time-series data," *Bioinformatics,* vol. 20, pp. 742-U575, Mar 22 2004.