Statistics and Its Interface Volume 3 (2010) 413–418

Periodicity analysis of DNA microarray gene expression time series profiles in mouse segmentation clock data

Vivian, Tsz-Yan Tang*, Alan Wee-Chung Liew and Hong Yan

With microarray technology, gene expression profiles are produced at a rapid rate. It remains a challenge for biologists to robustly identify periodic gene expression profiles when the time series have short data length and contain a high level of noise. An effective method is proposed in this paper to analyze the periodicity of gene expression time series using singular value decomposition (SVD), singular spectrum analysis (SSA) and autoregressive (AR) model-based spectral estimation. Using these procedures, noise can be filtered out and over 85% of periodic gene expression can be identified in the mouse segmentation clock data set.

AMS 2000 subject classifications: Primary 60K35, 60K35; secondary 60K35.
Keywords and phrases: Singular value decomposition (SVD), Singular spectrum analysis (SSA), Segmentation clock, Periodicity analysis, Microarray time series analysis.

1. INTRODUCTION

A DNA microarray, or DNA chip, consists of an arrayed series of spots, each of which carries a segment of a DNA sequence. In a microarray experiment, thousands of gene expression levels are recorded simultaneously to study the effects of certain illnesses, therapy, and developmental processes. With microarray technology, gene expression data are produced at a rapid rate. Biologists are interested in detecting important patterns hidden in the gene expression profiles. However, it is still a challenge to identify periodic gene expression profiles reliably because the gene expression time series are usually short in length and are corrupted by a high level of noise.

In this paper, we study microarray data from the mouse presomitic mesoderm transcriptome generated to investigate the segmentation process during embryogenesis [1]. Presomitic mesoderm (PSM) is the embryonic tissue composed of mesoderm in the region of the embryo that will be divided into somites later by the segmentation process. This process involves a molecular oscillator, the segmentation clock, which produces periodic time series in PSM [2]. Transcription profiling of mouse presomitic mesoderm with 17 samples at different time points is carried out to identify periodic genes of the segmentation clock [1]. Based on this dataset, Dequeant [3] carried out a research study comparing the pattern detection performance of several mathematical approaches, which included the Lomb-Scargle (L) periodogram, Phase consistency (P), Address reduction (A), Cyclohedron test (C), and Stable persistence (S). The top three hundred ranked probe sets from these five methods were found and the results show that the Stable persistence (S) method performs best by identifying most of the benchmark probe sets within the top 300 probe sets. However, the data contain a high level of noise, which will degrade the performance of most data analysis algorithms. Therefore, we need to develop an effective method to process the noisy time series data.

In this paper, an effective method is developed to identify 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, the noise can be reduced [4]. By using AR modeling, more accurate spectral estimation results are obtained [5]. In our work, about 85% of gene expression profiles in the mouse segmentation clock dataset are found to be periodic.

2. METHODS

2.1 Dataset

The microarray dataset used in this research is downloaded from http://www.ebi.ac.uk/microarray-as/ae/ with the accession ID E-TABM-163. The dataset contains about 22 thousand probe sets and each of which contains 17 time points and is normalized to have zero mean [1]. The data are filtered based on three criteria: by detection call (removing the probe sets called "absent" and "marginal"), by maximum signal intensity (taking out the genes with an expression level less than 50), and by peak-to-peak amplitude (less than 1.65). After these filtering operations, the dataset contains 10,025 probe sets. All computations in our work are performed using software developed in-house. We use the MATLAB programming language as it provides many toolboxes for matrix operations.

*Corresponding author.

2.2 Singular spectrum analysis for noise reduction

Microarray time series data are usually short in length and contain a high level of noise. Although we can analyze the periodicity of gene expression levels directly based on the original data, the accuracy of the results will be degraded significantly by noise. Thus, before performing periodicity detection, a pre-processing technique is needed to reduce the noise level.

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

We introduce an SSA based algorithm to reduce the noise of microarray time series data. SSA is regarded as a model free approach as it decomposes an original time series into trend and noise components according to the SVD [6]. Assume there is a time series represented as $y^T = (y_1, \ldots, y_p, \ldots, y_n)^{\mathrm{T}}$, which is reorganized in an AR(p) model representation, where p is the order of the AR model, and n is equal to 17, the length of gene expression time series. The matrix form of the AR(p) model can be written as Equation (1). In this model, we consider that a time series is produced by a linear system. Given one or more time series as outputs of the system, our task is to identify the parameters of the AR coefficients of the model. In the AR(p) model, the number of unknowns is p, the model order, and the number of linear equations is equal to $2(n-p)$. The relationship between the number of equations and the number of unknowns is important. Since each microarray time series only contains 17 samples, the number of equations is small even for a moderate value of p. This may cause the linear system in Equation (1) to become unstable and error-prone computationally. To deal with this problem, the forward-backward linear prediction method is used instead of only forward or backward prediction to double the number of equations. Thus, the AR coefficients can be evaluated more reliably [7]. We set p to 8, so there are 8 AR coefficients $\mathbf{a}^{\mathrm{T}} = (a_1, a_2, \ldots, a_8)^{\mathrm{T}}$ in 18 linear equations.

In the matrix form, Equation (1) can be rewritten as,

(2)
$$\mathbf{y} = -\mathbf{Y}\mathbf{a}$$

where both $\mathbf{Y} \in \Re^{2(n-p)\times p}$ and $\mathbf{y} \in \Re^{2(n-p)\times 1}$ are known [8]. $\mathbf{Y}$ is often called the trajectory matrix [7]. In order to compute the AR coefficients $\mathbf{a}$, we apply the SVD to matrix $\mathbf{Y}$ and decompose it to

(3)
$$\mathbf{Y} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}}$$

where $\mathbf{U} \in \Re^{2(n-p)\times 2(n-p)}$, $\mathbf{V} \in \Re^{p\times p}$ and $\mathbf{S}$ has the same dimension as $\mathbf{Y}$. $\mathbf{S}$ has non-zero values only in its diagonal entries which are called the singular values of $\mathbf{Y}$ and are equal to the square root of the eigenvalues of $\mathbf{Y}\mathbf{Y}^{\mathrm{T}}$ or $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$. The singular values are always real positive numbers and can be arranged in descending order along the diagonal direction, as shown in the following equation.

(4)
$$\begin{pmatrix} s_1 & 0 & \cdots & 0 & 0 \\ 0 & \ddots & 0 & 0 & 0 \\ \vdots & 0 & s_k & 0 & \vdots \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}, \text{where } s_1 > s_2 > \cdots > s_k.$$

The SVD provides a powerful tool to remove noise from the data. Typically, large singular values represent the trend pattern in the time series signal and small singular values correspond to noise [9]. Therefore, by zeroing the small eigenvalues, we can extract the trend component and remove the noise component in the gene expression data. After setting small singular values to zero, we can reconstruct the trajectory matrix according to Equation (3) and obtain a new matrix **Y'**. This new matrix mainly contains the trend component of the time series data. Since every gene expression profile carries a different amount of noise, the number of singular values of $\mathbf{Y}$ retained is varied to achieve the best noise filtering result.

Figure 1 shows the singular values of gene expression data from probe set 1415717_at. We can see that the first four leading singular values contain most of the energy and therefore the remaining eigenvalues can be considered as noise contamination [10, 11]. The SSA based procedure is performed six times on each gene expression profile using a different number of leading singular values, varying from 3 to 8, and results are recorded. Then the time series data $(y_1, \ldots, y_p, \ldots, y_n)$ is reconstructed by averaging the elements of matrix **Y'** over the diagonals.

2.3 The AR power spectrum estimation

The power spectral density (PSD) of a time series sequence can be used to detect periodic components in the sequence. It is especially useful for genome-wide gene expression cell-cycle identification. If a time series signal is highly periodic, the resultant power spectrum has sharp peaks at the corresponding frequency [12, 13]. PSD can be easily found by applying the Fast Fourier Transform (FFT) to the time series data. However, microarray gene expression time series are usually short, which produces the so-called data truncation or windowing artifacts in the power spectrum. The AR model-based spectral estimation method can overcome this problem and is used in this work [9, 12, 14].

Figure 1. The singular values of the data matrix for probe set 1415717_at

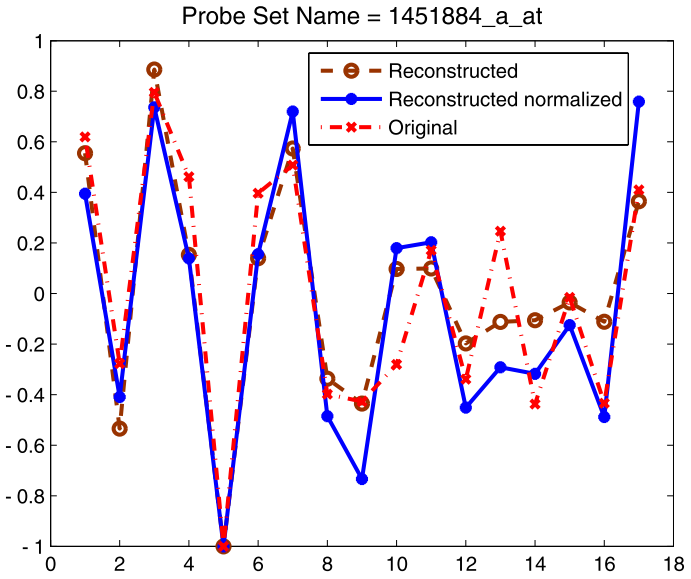

Figure 2. The reconstructed signal normalized to reduce the damping factor

A basic sinusoidal waveform can be defined by the expression below,

$$y(t) = A\sin(2\pi ft) \tag{5}$$

where A is the amplitude that affects the peak of the signal and f is the frequency. In AR power spectrum estimation, a time series signal is called periodic if a sharp peak exists at the frequency f. Assume that $A(t)$ is an exponential function as shown below,

$$A(t) = \exp(-\alpha t) \tag{6}$$

where α is a constant. Equation (6) causes damping in the time series signal. To reduce the effect of the damping factor, we normalize the signal amplitude along the time direction by dividing the sample values by the local averages of the signal intensity. The resultant signal of an example is shown in Figure 2. Through this normalization process, the large peak at the beginning of the expression profile is reduced. The amplitude of the normalized signal is more uniform than both the original and noise filtered signals.

From Equation (3), the AR coefficients can be found as

$$\mathbf{a} = -\mathbf{V}\mathbf{S}^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{y} \tag{7}$$

where $\mathbf{S}^{-1}$ contains the inverse of the singular values. That is, if $\mathbf{S} = \mathrm{diag}(s_1, s_2, \ldots, s_k)$, which is a diagonal matrix with diagonal elements s_1, $s_2, \ldots, s_k$, then $\mathbf{S}^{-1} = \mathrm{diag}(s_1^{-1}, s_2^{-1}, \ldots, s_k^{-1})$ [15]. Once the AR coefficients are estimated, the spectrum of the time series is given by

$$P(\omega) = \frac{T\sigma^2}{|1 + \sum_{r=1}^{p} a_r \exp(-j\omega rT)|^2} \tag{8}$$

where ω is the angular frequency in the range $(0, \pi)$, 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 [14, 16]. The power spectrum density function is normalized to the range [0,1].

2.4 The periodicity detection

A periodic signal produces a peak at its corresponding frequency in the spectrum. Considering a peak 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 from the peak. These two values indicate how sharp a spectral peak is. If the width $[f_{i-1}, f_{i+1}]$ is sufficiently small, the time series signal is said to be highly periodic. According to the width of the dominant peak in the spectrum of each time series, we can rank the expression profiles in the entire microarray dataset and determine how many genes are periodic. Large spectral width is considered as lacking in periodicity and can be discarded. We normalize the frequency range of the power spectrum to [0,1]. Periodicity is determined according to the width of the dominant peak in a spectrum. If the normalized peak width is less than 0.1, the corresponding gene expression profile is detected as highly periodic. We have applied our algorithm six times using a different number of leading singular values. The power spectrum of each gene expression is computed during each experiment and the widths of the spectral peaks are ranked. Only the minimum power spectrum widths with the corresponding number of leading singular value are considered for the periodicity detection of gene expression profiles.

To summarize, our method consists of two parts. The first filters out the noise of the time series data and the second detects the periodicity. First, each gene expression

profile is represented using an AR model. 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 small singular values. The noise filtered time series data are reconstructed based on the remaining singular values. In the second part, the damping in the microarray time series data is corrected. Then, the AR coefficients and power spectrum density are computed according to Equations (7) and (8) respectively. Finally, the width of the dominant peak in the power spectrum is used to detect the periodic component in a time series signal.

3. RESULTS

We have applied our algorithm to the expression data from the transcription profiling of mouse presomitic mesoderm [1]. After data pre-processing, 10,025 gene expression profiles remain. As discussed above, the first part of our algorithm involves noise filtering based on signal reconstruction using the SVD, SSA and the AR model. The second part consists of periodicity detection using the AR power spectrum of the reconstructed signal. When an expression profile is highly periodic, its power spectrum has a sharp peak at the corresponding frequency.

3.1 The periodicity detection using power spectrum width

We have utilized the spectral peak width as the score to analyze the periodicity of a time series. The widths for the entire dataset are ranked to detect the genes which are most likely to be periodic. The number of mouse presomitic mesoderm expression profiles counted as periodic is shown in Figure 3.

We can see that the widths of spectral peaks of mouse presomitic mesoderm profiles are mainly within the range of 0 to 0.1. By applying our algorithm, we have found that a total of 8,445 (85% of the entire microarray dataset) genes have spectral widths of less than 0.1 which are periodic. Without performing SSA and SVD, only 2,992 genes can be counted as periodic. Thus, the filtering technique is indeed effective for removing noise in the data.

After our algorithm is applied, the reconstructed signals look sinusoidal and the dominant spectral peak is located around frequency value 0.4π. Compared with the original signal, noise is reduced in the filtered signal by SVD and SSA and the power spectral peaks become sharper. As another comparison, the FFT is also applied to the reconstructed signal to estimate the power spectrum. The results are shown in Figure 4 and 5. Due to the windowing artifacts, the FFT provides a poor spectral resolution as it is incapable of producing sharp peaks for resolving different frequency components. In our algorithm, noise filtering is integrated in AR modelling which improves the performance of the periodic gene detection procedure. The five methods described in [3]

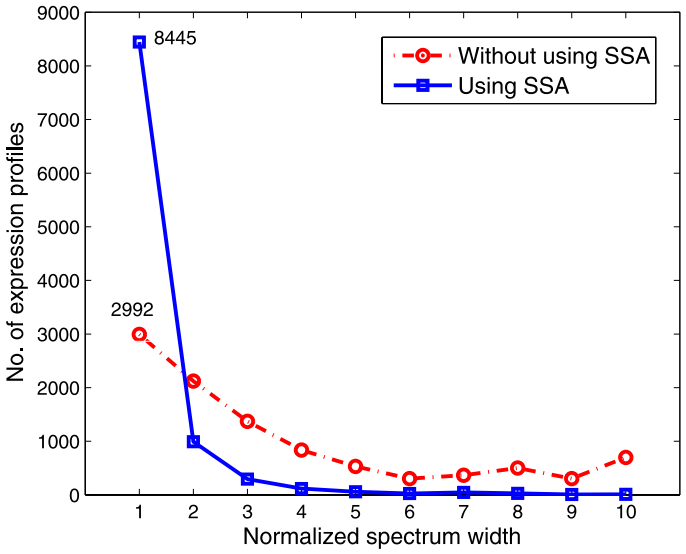

Figure 3. The number of expression profiles counted as periodic in the mouse segmentation dataset

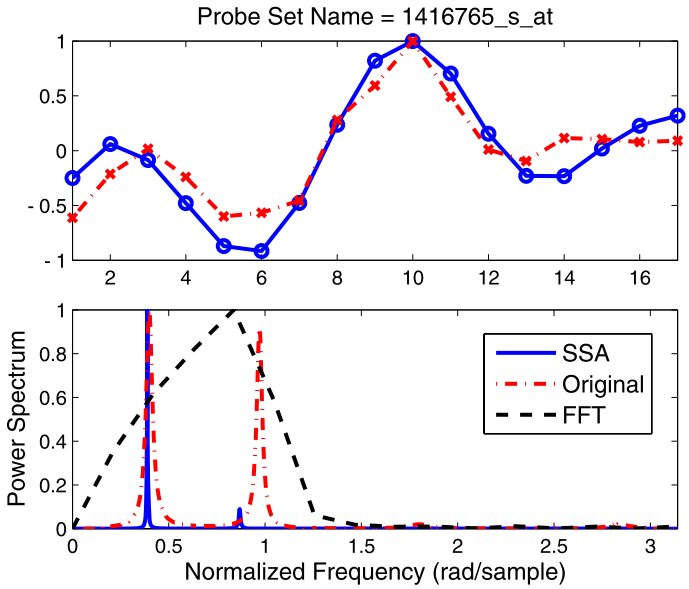

Figure 4. Reconstructed signal and its power spectrum density of the expression profile from probe set 1416765_s_at

deal with the original noisy data directly and may miss some periodic profiles. The probe set in Figure 4 is ranked 255 and 274 among the top 300 periodic profiles using the Cyclohedron test (C) and the Lomb-Scargle (L) periodogram respectively. This probe set is not ranked among the top 300 by three other methods described in [3]. The probe set in Figure 5 is not ranked among the top 300 by any of the five methods. The periodicity of the two profiles is evident after noise removal using our method. The noise reduction in Figure 5 is especially significant.

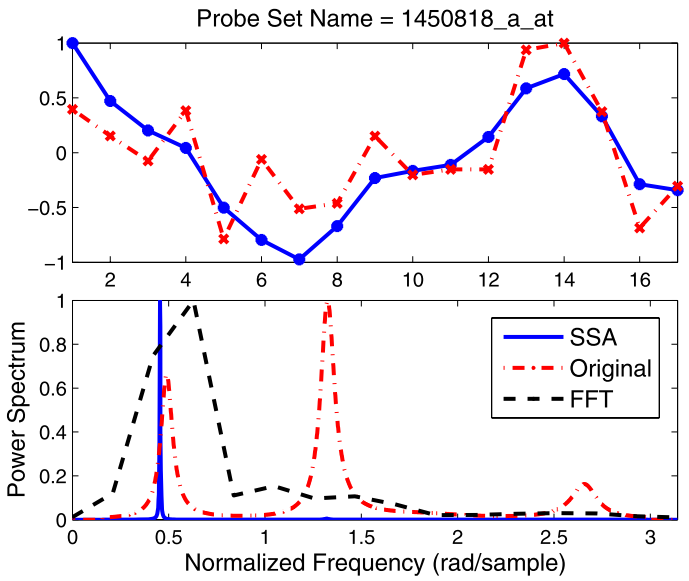

Figure 5. Reconstructed signal and its power spectrum density of the expression profile from probe set 1450818_a_at

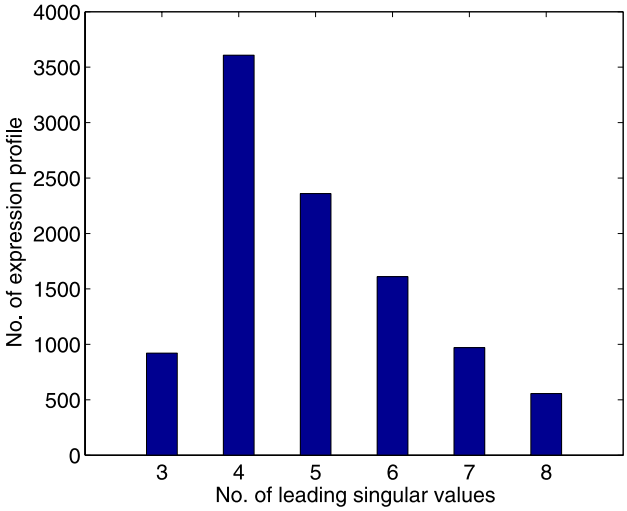

Figure 6. The histogram of the number of leading singular values in the gene expression profiles of murine presomitic mesoderm

We use spectral peak width rather than the power ratio (the detection method in [13]) to detect periodic genes because a highly periodic gene expression profile may also have a sharp peak with a smaller amplitude (see Figure 4). For example, the power spectrum of probe set 1416765_s_at has a sharp peak at about 0.4π, as well as another smaller peak at about 0.8π. In this case the power ratio at the larger peak will be affected by the smaller peak, which reduces the accuracy of periodicity detection.

3.2 The number of singular values

In computing the SVD, one important criterion is how to select the number of singular values to reconstruct the signal. The singular values represent the contributions of the trend components and the noise. In general, each expression profile in the microarray dataset may contain a different amount of noise and require the use of different numbers of leading singular values in noise filtering and signal reconstruction.

Figure 6 shows the histogram of the number of singular values we have used in the noise filtering algorithm. We observe that if we keep three to four leading singular values and set the remaining to be zero when performing SVD, about 60% of expression profiles contains a spectral peak with the minimum power width, which implies that over half the time series contain most of their energy within the first four leading singular values.

4. CONCLUSION

In this paper, we have proposed an effective algorithm to perform noise filtering and periodicity detection in microarray datasets. In general, microarray gene expression profiles are short in length and contain a high level of noise. By combining SVD, SSA and AR modeling, we can effectively reduce the noise and detect the periodic trend component. We have considered the presence of sharp spectral peaks in 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.

ACKNOWLEDGEMENTS

This work is supported by the Hong Kong Grant Research Council (Project CityU 122607).

Received 27 November 2009

REFERENCES

[1] Dequeant, M. L., Glynn, E., Gaudenz, K., Wahl, M., Chen, J., Mushegian, A. and Pourquie, O., "A complex oscillating network of signaling genes underlies the mouse segmentation clock," *Science*, vol. 314, no. 5805, pp. 1595–1598, 2006.

[2] Dequeant, M. L. and Pourquie, O., "Segmental patterning of the vertebrate embryonic axis," *Nat Rev Genet*, vol. 9, no. 5, pp. 370–382, 2008.

[3] Dequeant, M. L., Ahnert, S., Edelsbrunner, H., Fink, T. M., Glynn, E. F., Hattem, G., Kudlicki, A., Mileyko, Y., Morton, J., Mushegian, A. R., Pachter, L., Rowicka, M., Shiu, A., Sturmfels, B. and Pourquie, O., "Comparison of pattern detection methods in microarray time series of the segmentation clock," *PLoS One*, vol. 3, no. 8, p. e2856, 2008.

[4] Watkins, D. S., *Fundamentals of matrix computations*. Pure and applied mathematics, New York; [Great Britain]: Wiley-Interscience, 2nd ed. ed., 2002. MR1899577

[5] Yan, H. and Pham, T. D., "Spectral estimation techniques for DNA sequence and microarray data analysis," *Current Bioinformatics*, vol. 2, no. 2, pp. 145–156, 2007.

[6] Myung, N., *Singular Spectrum Analysis*. PhD thesis, 2009.

[7] Choong, M. K., Charbit, M., and Yan, H., "Autoregressive-model-based missing value estimation for DNA microarray time series data," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 1, pp. 131–137, 2009.

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

[9] Du, L., Wu, S., Liew, A. W., and Smith, D. K. "Spectral analysis of microarray gene expression time series data of plasmodium falciparum," *Int J Bioinform Res Appl*, vol. 4, no. 3, pp. 337–349, 2008.

[10] Vautard, R., and Ghil, M., "Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time-series," *Physica D*, vol. 35, no. 3, pp. 395–424, 1989. MR1004204

[11] Liu, L., Hawkins, D. M., Ghosh, S., and Young, S. S., "Robust singular value decomposition analysis of microarray data," *Proc Natl Acad Sci USA*, vol. 100, no. 23, pp. 13167–13172, 2003. MR2016727

[12] Liew, A. W., Xian, J., Wu, S., Smith, D. and Yan, H., "Spectral estimation in unevenly sampled space of periodically expressed microarray time series data," *BMC Bioinformatics*, vol. 8, p. 137, 2007.

[13] Golyandina, N., Nekrutkin, V. and Zhigljavsky, A., *Analysis of time series structure: SSA and related techniques*. Monographs on statistics and applied probability, London: Chapman & Hall, 2001. MR1823012

[14] Yeung, L. K., Szeto, L. K., Liew, A. W. C. and Yan, H., "Dominant spectral component analysis for transcriptional regulations using microarray time-series data," *Bioinformatics*, vol. 20, no. 5, pp. 742–U575, 2004.

[15] Schott, J. R., *Matrix analysis for statistics*. Wiley series in probability and statistics, Hoboken, N.J.: Wiley-Interscience, 2nd ed., 2005. MR2111601

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

Vivian, Tsz-Yan Tang
Department of Electronic Engineering,
City University of Hong Kong,
Tat Chee Avenue, Kowloon, Hong Kong
E-mail address: yan.tang.ty@gmail.com

Alan Wee-Chung Liew
School of Information & Communication Technology,
Gold Coast Campus, Griffith University,
QLD 4222,Australia
E-mail address: a.liew@griffith.edu.au

Hong Yan
Department of Electronic Engineering,
City University of Hong Kong,
Tat Chee Avenue, Kowloon, Hong Kong
E-mail address: h.yan@cityu.edu.hk