Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Reliable Detection of Short Periodic Gene Expression Time Series Profiles in DNA Microarray Data

Alan Wee-Chung Liew
School of Information and Communication Technology
Gold Coast Campus, Griffith University
QLD4222, Australia
a.liew@griffith.edu.au

Hong Yan
Department of Electronic Engineering
City University of Hong Kong, Kowloon, Hong Kong
School of Electrical and Information Engineering
University of Sydney, NSW2006, Australia
h.yan@cityu.edu.hk

***Abstract*—Many cellular processes exhibit cyclic behaviors. Hence, one important task in gene expression data analysis is to detect subset of genes that exhibit periodicity in their gene expression time series profiles. Unfortunately, gene expression time series profiles are usually of very short length, with very few periods, unevenly sampled, and are highly contaminated with noise. This makes detection of periodic profiles a very challenging problem. In this paper, we present several effective computational techniques developed recently in our research group for the reliable detection of short periodic gene expression time series profiles.**

***Keywords*—periodic gene expression profile, short time series, spectrum estimation, signal reconstruction, Fisher statistical test**

I. Introduction

Oscillation arises in genetic and metabolic networks as a result of various modes of cellular regulation. These rhythmic processes occur at all levels of biological organization with widely varying periods, i.e., from fractions of seconds to decades [1]. Well known examples of biological rhythms include cell division [2-4] and circadian rhythms [5, 6].

The cell-division cycle is fundamental to the proliferation of all organisms. In mitotic cell division, a single cell going through a sequence of events yields two identical daughter cells. Four phases are usually distinguished [7]. In the presynthetic G1 phase (Gap 1), the cell prepares itself for subsequent DNA synthesis. Enzymes and proteins required for initiating and carrying out DNA synthesis are synthesized late in the G1 phase and early in S phase. The G1 phase is followed by the S phase (synthetic phase) in which the cell replicates its DNA. The postsynthetic G2 phase (Gap 2) is the phase after completion of DNA synthesis in which the cells controls whether DNA replication has been completed and prepares for cell division by synthesizing molecules required in mitotic operation. The M phase (mitotic phase) following the G2 phase is characterized by the disappearance of nuclear membranes and nucleoli, appearance of the spindle apparatus (prophase), condensation of chromatin into chromosomes, parallel alignment of chromosomes in the equatorial plane of the spindle (metaphase), separation of chromosomes into pairs, and simultaneous movement of chromosomes to opposite poles (anaphase), the reassembly of two nuclei (telophase), and cytoplasmic division (cytokinesis), i.e. the segmentation and separation of the cytoplasm, resulting in the formation of two separate cells.

Rhythmic cellular processes are regulated by different gene products and can be measured through a series of DNA microarray experiments. If the expression patterns of a group of genes are measured over a number of time points, we obtain a time series gene expression profiles describing the rhythmic behaviors of the genes under study.

A well known set of gene expression time series datasets is that of the Yeast (Saccharomyces cerevisiae) from Spellman et al. [3]. In this set of data, the genome-wide mRNA levels for 6178 yeast ORFs are monitored simultaneously using several different methods of synchronization including an alpha-factor-mediated G1 arrest which covers approximately two cell-cycle periods with measurements at 7 minute intervals for 119 minutes with a total of 18 time points, a temperature-sensitive cdc15 mutation to induce a reversible M-phase arrest (24 time points taken every 10 minutes covering approximately 3.5 cell-cycle periods), and a temperature-sensitive cdc28 mutation to arrest cells in G1 phase reversibly (17 time points taken every 10 minutes covering approximately 2 cell-cycle periods), and finally, an elutriation synchronization to produce the elutriation dataset of 14 time points taken every 30 minutes covering approximately 1 cell-cycle period. Figure 1 shows some periodic (top panel) and random (bottom panel) profiles from this data set.

Gene expression time series data are uniquely different from many traditional time series data in several aspects. First, gene expression time series profile usually contains very few time points. It is not uncommon to see expression profiles that are less than ten time points long. Second, the number of cycles within a profile is usually very few. For example, the 14 time point elutriation dataset of [3] contains only 1 cell-cycle. Third, gene expression data may contain missing values and they usually need to be estimated from the dataset in advance [8]. Fourth, the time points need not be spaced at regular interval, giving rise to the problem of detecting periodicity in unevenly sampled time series data [9]. Finally, gene expression data is notoriously noisy. All the above makes the detection of

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

periodic gene expression time series profiles an extremely challenging problem and conventional signal processing methods such as FFT-based techniques do not perform adequately.

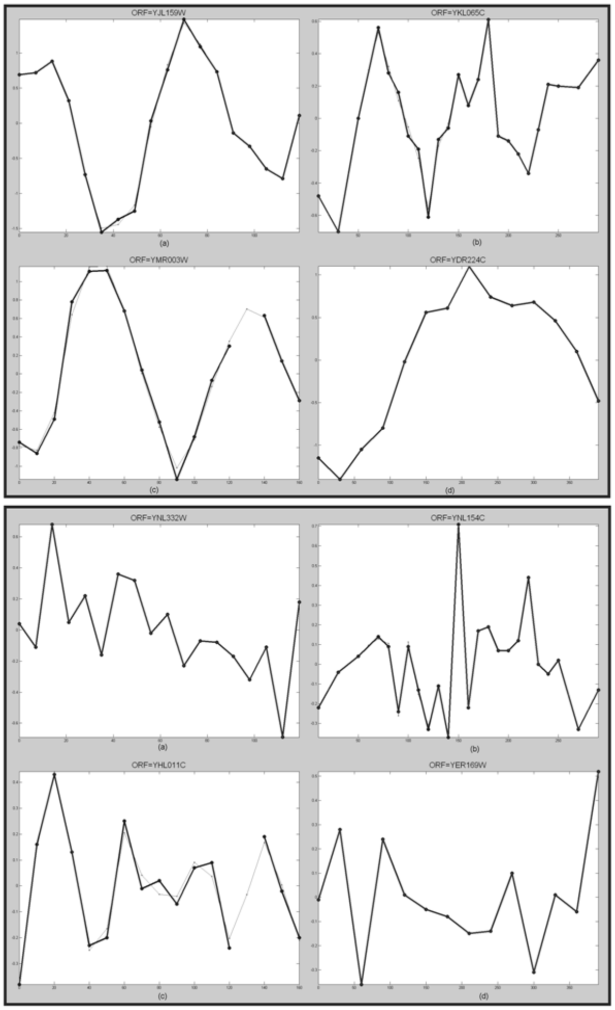

Figure 1. Top panel: highly periodic expression profiles. Bottom panel: random profiles from Yeast datasets of Spellman et al. [3]. Profiles (a), (b), (c), (d) correspond to the alpha, cdc15, cdc28, elutriation datasets, respectively. The thin curves in the figure are the interpolated profiles with missing values filled in. The x-axis shows the time points, the y-axis shows the measured expression values.

In our work on gene expression profile analysis, we have developed effective computational techniques to address: (1) Missing value estimation [8], (2) Periodicity detection [9, 10], and (3) Cluster and bicluster analysis [11-17]. They constitute a suit of useful algorithms that can be used to identify interesting genes that are involved in certain cellular processes. In this paper, we will concentrate on the problem of detection of short noisy periodic gene expression time series profiles. Specifically, we describe several effective computational techniques for periodicity detection based on: (i) Singular Spectrum Analysis (SSA) and Autoregressive (AR) based spectral estimation, (ii) Spectral estimation of short unevenly sampled profiles by signal reconstruction, and (iii) Statistical hypothesis testing for periodic signal detection.

II. Singular Spectrum Analysis (SSA) and Autoregressive (AR) Spectral Estimation

In [10], we proposed a parametric spectral estimation technique for short time series profiles based on SSA and AR modeling. The AR model for a time series s(n) is given by

$$s(n) = -\sum_{p=1}^{P} a_p s(n-p) + u(n) \tag{1}$$

where a_p are the AR coefficients, P is the order of the AR model, and $u(n)$ is a white noise sequence. AR model based power spectrum estimation allows better frequency resolution to be obtained by fitting a relatively high order AR model to the data sequence. However, the AR spectrum is sensitive to noise. When the signal to noise ratio is low, the accuracy of the parameter estimation in Equation (1) would be reduced substantially. A higher order AR model has to be used to improve the frequency resolution, but this would induce the appearance of spurious peaks. To remedy this problem, we preprocess the profiles using SSA. The main idea of SSA is to extract the underlying trend from short and noisy time series.

SSA performs a Singular Value Decomposition (SVD) on the trajectory matrix obtained from the original time series [18]. The singular values can then be grouped into two separate components: trend component and noise component. With the proper selection of singular values, the trend curve that represents the dominant spectral component can be reconstructed from the original expression profile. Let each expression profiles be a time series $\{s_1, s_2, ..., s_n, ..., s_N\}$, the SSA can be performed as follows:

1. Construct the trajectory matrix $X_{M,K}$ from the original series by sliding a window of length M $(M \le N/2)$, $K = N - M + 1$

$$X_{M,K} = (x_{ij} = s_{i+j-1}) = \begin{bmatrix} s_1 & s_2 & \cdots & s_K \\ s_2 & s_3 & \cdots & s_{K+1} \\ \vdots & \vdots & \cdots & \vdots \\ s_M & s_{M+1} & \cdots & s_N \end{bmatrix} \tag{2}$$

2. Perform the SVD of the matrix $R = XX^T$. The eigenvalues are ranked in decreasing order $\lambda_i, (1 < i < M)$ and $(\lambda_1 > \lambda_2 > \cdots \lambda_M)$, and the trajectory matrix is decomposed into a series of components $X_i = \sqrt{\lambda_i} U_i V_i^T$, $(i = 1, 2, ..., M)$, which are rank-one biorthogonal matrices, the U_i and V_i are the left and right singular vectors of the matrix X , respectively.

3. Group a specified number of leading eigenvalues λ_i and sum the corresponding components X_i, then the resultant matrix is $X'_{M,K} = (x'_{ij})$.
4. Reconstruct the data series $\{s'_1, s'_2, ..., s'_n, ..., s'_N\}$ by averaging the elements of matrix X' over the "diagonals" $i + j = n + 1$. The choice n = 1 gives $s'_1 = x'_{11}$, for n = 2 we have $s'_2 = (x'_{12} + x'_{21})/2$ and so on.

Using SSA and AR, periodic profiles can be detected as follows:

- First, each expression data is reconstructed using SSA. Only the expression profiles with the eigenvalue ratio (sum of first two eigenvalues over the sum of all eigenvalues) greater than 0.6 are to be reconstructed.
- Second, the AR spectrum is calculated for each reconstructed expression profiles. Then the frequency f_i at peak value point and the ratio of the power in f_i 's ROI (i.e. the frequency band $[f_{i-1}, f_{i+1}]$) to the total power are calculated according to $S = \frac{power_i}{power_{total}}$.
- Finally, the profiles are screened according to following rule: if the power ratio S calculated previously is larger than 0.7, the corresponding profile would be selected as periodic, otherwise, we consider it lacking of periodicity and discard it.

The technique is applied to the expression data of the IDC of Plasmodium falciparum. The data contains the expression profiles of 5080 oligonucleotides measured at 46 time points spanning 48 hours during the IDC with one hour time resolution for the HB3 strain [19]. For the majority of the transcriptome of IDC of Plasmodium falciparum, the leading two singular values of expression profiles contain most of the energy and correspond to the signal. Figure 2 shows the singular values for Dihydrofolate Reductase-Thymidylate Synthase, where the first two singular values contain most of the energy in the time series profile. As shown in Figure 3, due to the removal of noise using the SSA, spurious peaks are eliminated from the spectrum. Using this method, we are able to detect more functional genes compared with the Fourier analysis technique used in [19]. By using our periodicity detection method, 4496 periodic profiles are found to be periodic. Compared to [19], an additional 777 periodic oligonucleotides are detected using our algorithm. Our method using the SSA has helped to smooth out noise and makes the dominant spectral component much more evident.

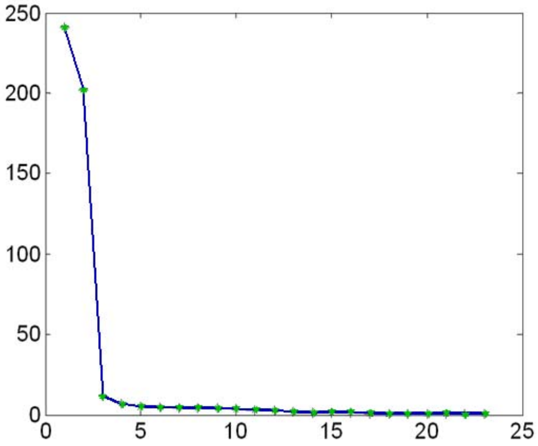

Figure 2. The singular values of the data matrix for Dihydrofolate Reductase-Thymidylate Synthase (DHFS-TS).

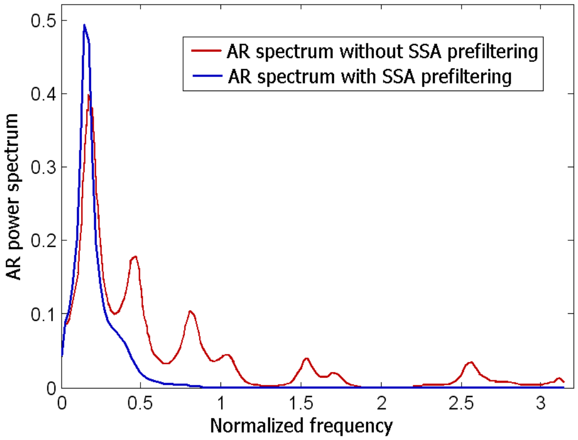

Figure 3. The AR spectra of the expression profile of DHFR-TS with and without SSA filtering.

Since the function of a gene is related to the initial phase of its expression profile, we have ordered the expression profiles of the 4496 periodic oligonucleotides according to their peak time points of expression profiles as in Figure 4. The phaseogram shows a continuous cascade of gene expressions, which correspond to the developmental stages throughout the IDC, that is, ring, trophozoite and schizont stages. According to the sharp transitions of ring-to-trophozoite (at the 17h time point), trophozoite-to-schizont (at the 29h time point) and schizont-to-rings stages (at the 45h time point), the 4496 periodic genes could be categorized into four stages on the basis of the peak time points of their expression profiles. The comparison of the classification results of the oligonucleotides assigned to these stages by our method with those in [19] is shown in Table 1.

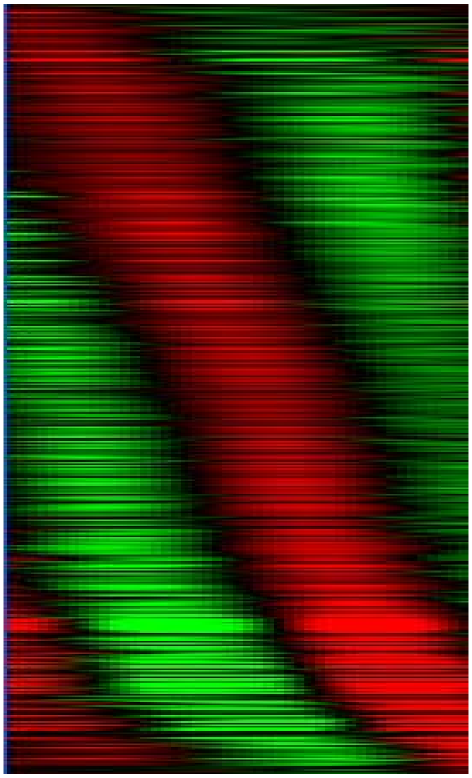

Figure 4. The phaseogram of the transcriptome of the IDC of P. falciparum. 4496 genes are ordered along the y axis in the order of the time of their peak expression.

TABLE I. CLASSIFICATION RESULTS OF OLIGONUCLEOTIDES IN THREE DIFFERENT STAGES. MORE GENES CAN BE IDENTIFIED USING OUR METHOD.

Stages	Our method	Bozdech et al. [19]
Ring/early Trophozoite	1970	1563
Trophozoite/early Schizont	1524	1296
Schizont	709	625
Early ring	293	235

III. SPECTRAL ESTIMATION BY SIGNAL RECONSTRUCTION

In many microarray time series data, the microarray experiments are not carried out at regular sampling intervals [3, 20]. Moreover, missing values are a common occurrence in microarray data. Time series profiles with missing values can be viewed as unevenly sampled. The unevenly sampled profiles make spectral analysis a challenging task. In [9], we proposed a new spectral estimation algorithm for unevenly sampled gene expression data. The method is based on signal reconstruction in a shift-invariant signal space where a direct spectral estimation procedure is developed using the B-spline basis.

Let $V(\phi)$ be the shift-invariant (also called time-invariant) signal space:

$$V(\phi) = \{f : f(x) = \sum_{k\in \mathrm{Z}} c_k \phi(x-k) : (c_k) \in \ell^2\} \quad (3)$$

where the coefficients $\{c_i\}$ are related to the choice of basis function ϕ. In our work, ϕ is chosen to be the set of B-spline functions. Unlike the traditional *sinc* interpolating functions, the B-spline function has compact support with smooth decay. We showed that under certain condition, a signal $f \in V(\phi)$ can be uniquely reconstructed from its unevenly sampled values $\{f(x_i)\}$, where $\{x_i\}$ is the sampling point set. For the B-spline interpolating functions, we can obtain an explicit formulation of the power spectrum density (PSD) as follows:

$$P_{xx}(\omega) = \frac{1}{A_2 - A_1} \left| \sum_{k=A_1-\Omega+1}^{A_2+\Omega-1} c_k e^{-i2\pi\omega k} \hat{\phi}(\omega) \right|^2 \quad (4)$$

where $$\hat{\phi}(\omega) = \left[\frac{\sin(\pi\omega)}{\pi\omega}\right]^{N+1}$$

Our method allows the PSD of an unevenly sampled signal to be computed directly from Equation (4). Since the periodogram of a microarray time series profile from a periodically expressed gene must contain a peak corresponding to its dominant frequency, we can perform statistical test (see next section) on the PSD to determine whether a time series profile is periodic or random. We applied the method on the gene expression dataset of Plasmodium falciparum and showed that it gives a good estimate of the number of periodic genes. Interested readers are referred to [9] for detailed analysis of the experimental results.

IV. STATISTICAL HYPOTHESIS TESTING FOR PERIODIC PROFILE DETECTION

The problem of deciding whether a time series is random or periodic can be cast as a statistical decision problem using hypothesis testing. The Fisher test can be used to determine whether a peak in the periodogram is significant or not. The test proceeds as follows. Given a time series *y* of length *N*, the periodogram $I(\omega)$ is first computed as

$$I(\omega) = \frac{1}{N} \left| \sum_{n=1}^{N} y_n e^{-j\omega n} \right|^2, \quad \omega \in [0, \pi] \quad (5)$$

and $I(\omega)$ is evaluated at the discrete normalized frequencies $\omega_l = \frac{2\pi l}{N}, \quad l = 0,1,\ldots,a$, where $a = [(N-1)/2]$ and $[x]$ denotes the integer part of *x*. If a time series has a significant sinusoidal component with frequency ω_k, then the periodogram will exhibit a peak at that frequency ω_k. An exact test of the significance of the spectral peak can be done by using the Fisher *g*-statistic [21].

$$g = \frac{\max_l I(\omega_l)}{\sum_{l=1}^{a} I(\omega_l)} \qquad (6)$$

Under the Gaussian noise assumption, the exact distribution of the *g*-statistic under the null hypothesis (that the spectral peak is insignificant) is given by

$$P(g > x) = \sum_{k=1}^{b} (-1)^{k-1} \frac{a!}{k!(a-k)!} (1 - kx)^{a-1} \qquad (7)$$

where *b* is the largest integer less than 1/*x* and *x* is the observed value of the g-statistic. Equation (7) yields a p-value that allows us to test whether a given time series behaves like a random sequence. Large value of *g* indicates a strong periodic component and leads us to reject the null hypothesis.

Although the exact distribution of the Fisher *g*-statistic is available analytically, we found that care must be taken when applying it in practice. Figure 5 shows the exact distribution and the empirical distribution for signal length *N* = 10. The deviation from exact distribution can be clearly seen. For larger value of *N* (i.e. *N*>40), no significant deviation can be observed.

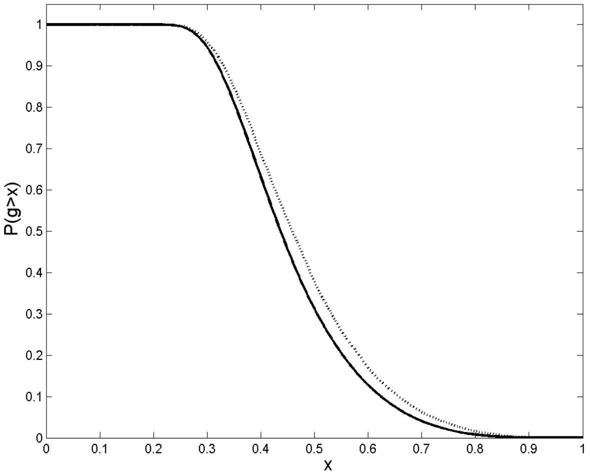

Figure 5. Empirically computed distribution (dashed curve) versus theoretical distribution (solid curve) for short length signal (*N* = 10).

In [22], we performed a series of experiments with simulated signals to investigate the statistical power of the Fisher test as a function of noise distribution, signal length, SNR, and the false discovery rate. We found that the deviation from the theoretical null distribution can be significant when signal length is shorter than 40 time points. Moreover, when the signal does not cover an integer number of periods, significant drop in the statistical power of the test was observed. In this case, a much longer signal is needed for the test to return reliable result. These findings indicate that in high likelihood, the number of periodic gene expression profiles can be severely underestimated for short length signal (<< 40 time points) as is the case with many of the publicly available gene expression datasets. Although our study shows that the Fisher test may be unreliable for short signal, the Fisher g-statistic, on the other hand, has been observed to provide a useful ranking of periodic signals. Strongly periodic signals are found to rank highly while random sequences have low ranking. In [9], we use this ranking to discover the periodic gene expression profiles in the Plasmodium falciparum dataset and showed that the number of periodic profiles in the complete dataset should be around 3700 to 4000. This estimate is based on analyzing the trend of the sorted G-statistic as shown in Figure 6. The intersection of the two distinct slopes points indicates a sudden change in the G-statistic trend.

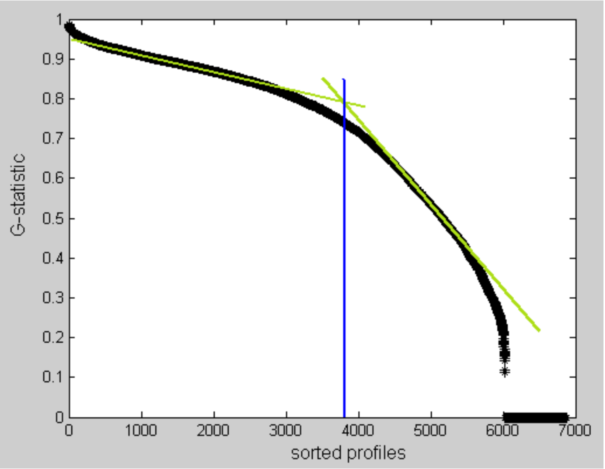

Figure 6. Sorted G-statistic values of Plasmodium falciparum. There is a change in the trend of the ranked G-statistic values at around the 4000 sorted profiles, indicating that two classes of profiles, i.e., periodic/aperiodic, are present in the dataset.

V. Conclusions

Detection of periodicity in gene expression time series profiles is an important step in the study of many cyclic cellular processes. Unfortunately, gene expression time series profiles are usually of very short length, with very few periods, unevenly sampled, and are highly contaminated with noise. This makes detection of periodic gene expression profiles a very challenging problem. In this paper, we present several effective computational techniques for the detection of periodic gene expression profiles, namely, (i) Singular Spectrum Analysis (SSA) and Autoregressive (AR) based spectral estimation, (ii) Spectral estimation of short unevenly sampled profiles by signal reconstruction, and (iii) Statistical hypothesis testing for periodic signal detection. We show that with proper care and appropriate techniques, reliable detection of periodic gene expression profiles is still possible despite the peculiarity of the time series gene expression profiles.

ACKNOWLEDGMENT

This work is support by a grant from the Hong Kong Research Grant Council (Project CityU122607).

REFERENCES

[1] A. Goldbeter, "Computational approaches to cellular rhythms", Nature, vol. 420, pp. 238-245, 2002.

[2] J.M. Mitchison, "Growth during the cell cycle", International review of cytology, vol. 226, pp. 165-258, 2003.

[3] T.S. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisia by microarray hybridization", Mol. Biol. Cell, vol. 9, pp. 3273-3297, 1998.

[4] G. Rustici, J. Mata, K. Kivinen, P. Lió, C. J. Penkett, G. Burns, J. Hayles, A. Brazma, P. Nurse, and J. Bähler, "Periodic gene expression program of the fission yeast cell cycle", Nature Genetics, vol. 36, pp. 809 – 817, 2004.

[5] S.K. Crosthwaite, "Circadian clocks and natural antisense RNA", FEBS Lett., vol. 567, pp. 49-54, 2004.

[6] U. Schibler, and F. Naef, "Cellular oscillators: rhythmic gene expression and metabolism", Current Opinion in Cell Biology , vol. 17(2), pp.223-229, 2005.

[7] A. Maton, D. Lahart, J. Hopkins, M.Q. Warner, S. Johnson, J.D. Wright, Cells: Building Blocks of Life, New Jersey: Prentice Hall, 1997.

[8] X. Gan, A.W.C. Liew, and H. Yan, "Microarray Missing Data Imputation based on a Set Theoretic Framework and Biological Consideration", Nucleic Acids Research, vol. 34(5), pp.1608-1619, 2006, doi:10.1093/nar/gkl047.

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

[10] L. Du, S. Wu, A.W.C. Liew, D.K. Smith, and H. Yan, "Spectral analysis of microarray gene expression time series data of Plasmodium Falciparum", International Journal of Bioinformatics Research and Applications , vol. 4(3), pp.337-349, 2008.

[11] K.O. Cheng, N.F. Law, W.C. Siu, and A.W.C. Liew, "Identification of coherent patterns in gene expression data using an efficient biclustering algorithm and parallel coordinate visualization", BMC Bioinformatics, vol. 9:210, April 2008, doi.10.1186/1471-2105-9-210.

[12] X. Gan, A.W.C. Liew, and H. Yan, "Discovering biclusters in gene expression data based on high-dimensional linear geometries", BMC Bioinformatics, vol. 9:209, April 2008, doi:10.1186/1471-2105-9-209.

[13] H. Zhao, A.W.C. Liew, X. Xie, and H. Yan, "A new geometric biclustering algorithm based on the Hough transform for analysis of large-scale microarray data", Journal of Theoretical Biology, vol. 251(2), pp.264-274, March 2008.

[14] B.S.Y. Lam, A.W.C. Liew, D. Smith, and H. Yan, "A regularized clustering algorithm based on calculus of variations", Journal of Signal Processing Systems, vol. 50(3), pp. 281-292, March 2008, doi 10.1007/s11265-007-0119-9.

[15] A.W.C. Liew, L.K. Szeto, S.S. Tang and H. Yan, "A computational approach to gene expression data extraction and analysis", special issue on Genomic Signal Processing, Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology, vol. 38(3), pp.237-258, November 2004.

[16] S. Wu, A.W.C. Liew and H. Yan, "Cluster Analysis of Gene Expression Data Based on Self-Splitting and Merging Competitive Learning", IEEE Transactions on Information Technology in Biomedicine, vol. 8(1), pp.5-15, March 2004.

[17] L.K. Szeto, A.W.C. Liew, H. Yan and S.S. Tang, "Gene Expression data clustering and visualization based on a binary hierarchical clustering framework", special issue on Biomedical Visualization for Bioinformatics, Journal of Visual Languages and Computing, vol. 14, pp.341-362, August 2003.

[18] R. Vautard, P. Yiou and M. Ghil, "Singular-spectrum analysis: A toolkit for short, noisy chaotic signals", Physica D, vol. 58, pp.95-126, 1992.

[19] Z. Bozdech, M. Llinas, B.L. Pulliam, E.D. Wong, J.C. Zhu, and J.L. DeRisi, "The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum", Plos Biology, vol. 1, pp.1-16, 2003.

[20] S. Chu, J.L. DeRisi, M.B. Eisen, J. Mulholland, D. Botstein, P.O. Brown, I. Herskowitz, "The transcriptional program of sporulation in budding yeast", Science, vol. 282, pp. 699-705, 1998

[21] R.A. Fisher, "Tests of significance in harmonic analysis", Proc. R. Soc. A, vol. 125, pp. 54–59, 1929

[22] 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(4), pp. 549-556, Apr. 2009.