IDENTIFYING DROSOPHILA CELL-CYCLE REGULATED GENES FROM IRREGULAR MICROARRAY DATA

Wentao Zhao[1]*, Kwadwo Agyepong*[1]*,Erchin Serpedin*[1]*, and Edward R. Dougherty*[1,2]

[1]Texas A&M University
Dept. of Electrical and Computer Engineering
College Station, Texas, 77843-3128

[2]Translational Genomics Research Institute
400 North Fifth Street, Suite 1600
Phoenix, Arizona 85004

ABSTRACT

Due to experimental constraints, most microarray observations are obtained through irregular sampling. In this paper three popular spectral analyzing schemes, i.e., Lomb-Scargle, Capon and the missing data amplitude and phase estimation (MAPES), are compared in terms of their ability and efficiency to recover the periodically expressed genes. The *in silico* experiments based on microarray measurements of Drosophila Melanogaster not only verify half of the published cell-cycle genes, but also corroborate genes that behave periodically in Human Hela time series experiments.

Index Terms— Spectral analysis, Genetics, Biological Systems

1. INTRODUCTION

The eukaryotic cell cycle is an echelon of molecular-level events that lead to a cell dividing into two daughter cells. The transcriptional events in the cell cycle can be quantitatively observed by measuring the concentration of messenger RNA (mRNA) via microarray experiments. The sampled time series data obtained from microarray are: of small sample size,unevenly sampled, characterized by missing time points, highly corrupted by experimental noise. These demand treatment of the data with robust stochastic analysis.

Quantitative analysis of microarray experiments reveals the genes involved in cell-cycle. Giurcaneanu [1] explored the stochastic complexity of the detection mechanism of periodically expressed genes. Ahdesmaki [2] implemented a robust periodicity testing procedure based on the non-Gaussian noise assumption. Bowles [3] compared Capon and robust Capon methods in terms of their ability to identify a predetermined frequency. The majority of current works deal with evenly sampled data, and missing data points are usually filled by interpolation in the time domain or are disregarded from the analysis when a large proportion of samples are missing.

The last decades have witnessed an increased interest in analysis of unevenly sampled data sets. The harmonics exploited in DFT are no longer orthogonal for uneven sampling. However, Lomb and Scargle [4] demonstrated that a phase shift suffices to make the sine and cosine terms orthogonal again. Stoica [5] updated the traditional Capon method to cope with the irregularly sampled data. Wang [6] developed missing-data amplitude and phase estimation (MAPES) by employing the Expectation Maximization (EM). In this paper, we analyze the performance of Lomb-Scargle, Capon, and the MAPES and answer the following questions: do technically more sophisticated schemes, like MAPES, achieve a better performance on real biological data sets, and is the sacrifice in efficiency by advanced methods justifiable?

Drosophila will be serving as our research target. It is a simple organism, though with 75% of human diseases, and 50% of its proteins have human analogs. These make it an ideal model for human diseases. In the literature, most of the computational methods for discovering periodic genes have been targeted either to yeast or human, mainly due to earlier publications of their data sets. However, in the case of Drosophila most works were conducted through experimental biological methods, and the computational analysis have not been fully explored for the detection of periodically expressed genes. Through intense computer simulations, for each of the investigated spectral estimation methods, we have identified 150 cyclic genes each for embryonic and pupal stages. 50% of published genes that are involved in cell cycle were verified. The detected cyclic genes in Drosophila were also discovered to be periodic in Human Hela. Our results not only illustrate the strength of the investigated spectral estimation methods for unevenly sampled data sets but also shed light on cross species genomic research on the cell cycle.

2. METHODS

In this section the Lomb-Scargle periodogram, Capon method and MAPES approach are introduced and compared. Detailed derivations are omitted. As a general notational convention, matrices and vectors are represented in bold characters, while scalars are denoted in regular fonts.

This work was supported by the National Cancer Institute (CA-90301) and the National Science Foundation (ECS-0355227 and CCF-0514644).

2.1. Lomb-Scargle Periodogram

Given N time-series observations $(t_l, y_l), l = 0, \ldots, N-1$, where t stands for the time tag and y denotes the sampled expression of a specific gene, the normalized Lomb-Scargle periodogram for that gene at angular frequency ω is

$$\begin{aligned}\Phi_{LS}(\omega) &= \frac{1}{2\hat{\sigma}^2}\frac{\left(\sum_{l=0}^{N-1}[y_l-\bar{y}]cos[\omega(t_l-\tau)]\right)^2}{\sum_{l=0}^{N-1}cos^2[\omega(t_l-\tau)]} \\ &+ \frac{1}{2\hat{\sigma}^2}\frac{\left(\sum_{l=0}^{N-1}[y_l-\bar{y}]sin[\omega(t_l-\tau)]\right)^2}{\sum_{l=0}^{N-1}sin^2[\omega(t_l-\tau)]},\end{aligned}$$

where $\bar{y}$ and $\hat{\sigma}^2$ stand for the mean and variance of the sampled data, respectively, and τ is defined as:

$$\tau = \frac{1}{2\omega}atan\left(\left(\sum_{l=0}^{N-1}sin(2\omega t_l)\right)/\left(\sum_{l=0}^{N-1}cos(2\omega t_l)\right)\right).$$

For evenly sampled data, the sampling interval Δ can be expressed as

$$\Delta = t_{l+1} - t_l = (t_{N-1} - t_0)/(N-1), \quad l = 0, \ldots, N-2.$$

The highest frequency, i.e. the Nyquist frequency, is $1/(2\Delta)$. Beyond this limit, the computed spectra repeat. For unevenly sampled data, let δ be the greatest common divisor (gcd) for all intervals $t_k - t_l$ $(k \neq l)$, the highest frequency that should be searched is $f_{max} = \omega_{max}/(2\pi) = 1/(2\delta)$. The number of probing frequencies is $\tilde{N} = (t_{N-1} - t_0)/\delta$, and the frequency grid can be defined as $\omega_l\delta = 2\pi l/\tilde{N}, l = 0, \ldots, \tilde{N}-1$. Notice further that the spectra on the front and rear halves of the frequency grid are symmetric since the microarray experiments output real values.

2.2. Capon Method

Recently, the Capon method has been updated to cope with the presence of irregular samples [5]. The order of autoregressive model or the bandwidth of the capon filter is assumed to be N_0. The ancillary vector is defined as $\mathbf{a}(\omega) = \left(1\ e^{j\omega}\cdots e^{j\omega(N_0-1)}\right)^T$. The largest value for N_0 is $\lfloor(\tilde{N}-1)/2\rfloor$ for the Capon method to be solvable. An estimate of the autocorrelation matrix $\hat{\mathbf{R}}$ can be obtained from the Lomb-Scargle periodogram. It can be represented by

$$\hat{\mathbf{R}} = \frac{1}{\tilde{N}\delta}\sum_{l=0}^{\tilde{N}-1}\mathbf{a}(\omega_l\delta)\mathbf{a}^H(\omega_l\delta)\Phi_{LS}(\omega_l).$$

The Capon power spectral estimate at frequency ω is given by

$$\Phi_C(\omega) = 1/\left(\mathbf{a}^H(\omega\delta)\hat{\mathbf{R}}^{-1}\mathbf{a}(\omega\delta)\right).$$

2.3. MAPES Method

Irregular sampling can be treated as a case of missing data as long as the sampling time tags share a greatest common divisor. This constraint is satisfied in most biological experiments and published data sets. The missing-data amplitude and phase estimation (MAPES) method, proposed in [6], is a non-parametric spectral estimation approach. It is robust to model errors and achieves a better spectral resolution. However, the exploitation of the expectation maximization (EM) algorithm sacrifices its computational efficiency.

The data, $y_l, l = 0, \ldots, \tilde{N}$, are assumed to be sampled uniformly, however, only N data points are available and there are $\tilde{N} - N$ missing data points. Noticeably $\tilde{N}$ still conforms to the previous definition. The gene expression signal with frequency ω can be modeled as

$$y_l = \alpha(\omega)e^{j\omega l} + \varepsilon_l(\omega), \quad l = 0, \ldots, \tilde{N}-1, \ \ \omega \in [0, 2\pi],$$

where $\alpha(\omega)$ represents the complex amplitude of the sinusoidal component and $\varepsilon_l(w)$ denotes the residual term. Employing the EM algorithm, MAPES tries to iteratively assess the missing data, and meanwhile to update the estimation of spectra and error.

The data vector $\mathbf{y} = (y_0, \cdots, y_{\tilde{N}-1})^T$ can be split into L overlapping subvectors, each with dimension $M \times 1$, and $L = \tilde{N} - M + 1$. These subvectors constitute the enhanced data vector $\tilde{\mathbf{y}}$ $(LM \times 1)$, which assumes the following expression

$$\tilde{\mathbf{y}} = (\tilde{\mathbf{y}}_0 \cdots \tilde{\mathbf{y}}_{L-1})^T = \mathbf{U}\boldsymbol{\gamma} + \mathbf{V}\boldsymbol{\mu},$$

where $\boldsymbol{\gamma}$ $(N \times 1)$ and $\boldsymbol{\mu}$ $((\tilde{N}-N)\times 1)$ represent the available and missing data, respectively, and $\mathbf{U}$ $(LM \times N)$ and $\mathbf{V}$ $(LM \times (\tilde{N}-N))$ denote their selection matrices, respectively. Alternatively, given $\mathbf{U}$, $\mathbf{V}$ and $\tilde{\mathbf{y}}$, the data vectors $\boldsymbol{\gamma}$, $\boldsymbol{\mu}$ can be computed in the least-squares (LS) sense as follows

$$\begin{aligned}\boldsymbol{\gamma} &= (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\tilde{\mathbf{y}} = \tilde{\mathbf{U}}^\dagger\tilde{\mathbf{y}}, \quad \text{where} \quad \tilde{\mathbf{U}}^\dagger = (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T, \\ \boldsymbol{\mu} &= (\mathbf{V}^T\mathbf{V})^{-1}\mathbf{V}^T\tilde{\mathbf{y}} = \tilde{\mathbf{V}}^\dagger\tilde{\mathbf{y}}, \quad \text{where} \quad \tilde{\mathbf{V}}^\dagger = (\mathbf{V}^T\mathbf{V})^{-1}\mathbf{V}^T.\end{aligned}$$

The residual vector and its covariance matrix are next defined

$$\mathbf{e}_l(\omega) = (\varepsilon_l(\omega)\ \varepsilon_{l+1}(\omega)\cdots\varepsilon_{l+M-1}(\omega))^T,$$

$$\mathbf{Q}(\omega) = E\left(\mathbf{e}_l(\omega)\mathbf{e}_l^H(\omega)\right),$$

where $E(\cdot)$ denotes the expectation operator, and in practice is replaced by a sample mean estimator. The following two notations are also required:

$$\boldsymbol{\rho}(\omega) = \begin{pmatrix} e^{j\omega 0}\mathbf{a}(\omega) \\ \vdots \\ e^{j\omega(L-1)}\mathbf{a}(\omega)\end{pmatrix}, \mathbf{D}(\omega) = \begin{pmatrix}\mathbf{Q}(\omega) & & 0 \\ & \ddots & \\ 0 & & \mathbf{Q}(\omega)\end{pmatrix}.$$

In the ith EM iteration, the probability density function (PDF) of the missing data vector $\boldsymbol{\mu}$ conditioned on the available data $\boldsymbol{\gamma}$ and other context parameters is complex Guassian

with mean and variance denoted by $(\mathbf{b}, \mathbf{K})$ as follows

$$\begin{aligned}\mathbf{b}_i(\omega) &= \tilde{\mathbf{U}}^T\boldsymbol{\rho}(\omega)\alpha_i(\omega) \\ &\quad +\tilde{\mathbf{U}}^T\mathbf{D}_i(\omega)\tilde{\mathbf{V}}\Big(\tilde{\mathbf{V}}^T\mathbf{D}_i(\omega)\tilde{\mathbf{V}}\Big)^{-1}\Big(\boldsymbol{\gamma}-\tilde{\mathbf{V}}^T\boldsymbol{\rho}(w)\alpha_i(w)\Big), \\ \mathbf{K}_i(\omega) &= \tilde{\mathbf{U}}^T\mathbf{D}_i(\omega)\tilde{\mathbf{U}} \\ &\quad -\tilde{\mathbf{U}}^T\mathbf{D}_i(\omega)\tilde{\mathbf{V}}\left(\tilde{\mathbf{V}}^T\mathbf{D}_i(\omega)\tilde{\mathbf{V}}\right)^{-1}\tilde{\mathbf{V}}^T\mathbf{D}_i(\omega)\tilde{\mathbf{U}}.\end{aligned}$$

Then the estimates for spectral magnitude $\alpha(\omega)$ and residual matrix $\mathbf{Q}$ are updated in terms of equations

$$\begin{aligned}\alpha_{i+1}(\omega) &= \frac{\mathbf{a}^H(\omega)\mathbf{S}^{-1}(\omega)\mathbf{Z}(\omega)}{\mathbf{a}^H(\omega)\mathbf{S}^{-1}(\omega)\mathbf{a}(\omega)}, \\ \mathbf{Q}_{i+1}(\omega) &= \mathbf{S}(\omega)+(\alpha_{i+1}(\omega)\mathbf{a}(\omega)-\mathbf{Z}(\omega))(\alpha_{i+1}(\omega)\mathbf{a}(\omega)-\mathbf{Z}(\omega))^H\end{aligned}$$

where the auxiliary matrices are defined as follows

$$\begin{aligned}(\mathbf{z}_0\cdots\mathbf{z}_{L-1})^T &= \mathbf{U}\boldsymbol{\gamma}+\mathbf{V}\mathbf{b}(\omega), \quad \mathbf{Z}(\omega)=\tfrac{1}{L}\textstyle\sum_{l=0}^{L-1}\mathbf{z}_l e^{-j\omega l}, \\ \mathbf{S}(\omega) &= \tfrac{1}{L}\textstyle\sum_{l=0}^{L-1}\boldsymbol{\Gamma}_l+\tfrac{1}{L}\sum_{l=0}^{L-1}\mathbf{z}_l\mathbf{z}_l^H-\mathbf{Z}(\omega)\mathbf{Z}^H(\omega).\end{aligned}$$

where $\boldsymbol{\Gamma}_\mathbf{0}, \cdots, \boldsymbol{\Gamma}_{\mathbf{L-1}}$ are $M \times M$ sub-block matrices located on the main diagonal of matrix $\mathbf{UKU^T}$. Finally, the MAPES estimator can be expressed as $\Phi_{MAPES}(\omega) = |\alpha(\omega)|^2/\tilde{N}$.

2.4. Periodicity Test

Based on the obtained power spectral density, each gene is classified as either cyclic or non-cyclic gene. Null hypothesis is formed to assume that the measurements are generated by a Gaussian noise. For a general periodogram or power spectral density estimator $\Phi(\omega)$, Fisher's test can be exploited to examine the significance of the detected peak. The Fischer's test statistic is defined as

$$T = \left(\max_{1\le k\le N_0}\Phi(\omega_k)\right) / \left(N_0^{-1}\sum_{1\le k\le N_0}\Phi(\omega_k)\right),$$

where $N_0 = \lfloor(\tilde{N}-1)/2\rfloor$ since the spectra on the defined frequency grid are symmetric. The asymptotic p-value for detecting the largest peak is given by $P(T > t) = 1 - e^{-N_0 e^{-t}}$.

A rejection of the null hypothesis based on a p-value threshold implies the power spectral density contains a frequency with magnitude substantially greater than the average value. This indicates the time series data represent a periodic signal and the corresponding gene is cyclic in expression. Notice also that more accurate estimation methods for the p-values exist. However, we will exploit the asymptotic values. A universal p-value threshold is impossible to be determined for variable sample sizes. The rank of genes ordered by their p-values is of additional importance and it helps to hedge the risk of dichotomous decisions. The asymptotic p-values will not change gene ranks and will be computed just as a valuable reference information.

For the Lomb-Scargle periodogram, $\Phi_{LS}(\omega)$ is exponentially distributed under the null hypothesis [4]. However, this exponential distribution is not applicable for a general power spectral density. Therefore, Fisher's test is employed to perform the comparison among different spectral schemes. Our simulation results also show that, for Lomb-Scargle periodogram, the gene ranks generated by Fisher's test do not differ much from that produced by the exponential distribution. Finally, we remark that other periodicity detection tests exist,such as the robust Fisher test, the likelihood ratio test and the χ^2 test.

3. RESULTS

Our *in silico* experiments are performed on the Drosophila data published by Arbeitman [7]. The RNA of 4028 genes were measured with 75 sequential samples through embryonic, larval, pupal and adulthood. The pupal and adult stages are excluded from analysis because they present too small sample sizes to be of considerable value.

The simulation recognized several patterns. For example the expression of gene CG8199, shown in Fig. 1, implies obvious periodicity in both the embryonic stage with a frequency about 0.05/hour (period 20 hours) and the pupal stage with a frequency 0.02/hour (period 50 hours). The expression period is elongated along the development. This is true since the life activity slows down when it grows up. All three schemes successfully detect the periodicity, however, Capon and MAPES show questionable biases for peak frequency, as indicated in Fig. 1D. Other common patterns include the facts that the gene expression keeps increasing or decreasing. Their spectra possess strong peaks near or at frequency zero.

There are 97 experimentally-verified cell-cycle genes [8]. Among these 97 genes, 41 were measured in Arbeitman's experiment [7]. For each scheme, our simulation associate all 4028 genes with their periodicity p-values. The top 150 genes with the smallest p-values are selected and conferred to be periodic with the highest confidence. The ability of different schemes to detect periodic genes are examined by comparing the identified 150 cyclic genes with the published 41 genes. As illustrated in Fig. 2, Lomb-Scargle and Capon share more than 38% of all identified genes. The overlapping between any two schemes is larger for embryonic data because of the larger sample size. However, for the pupal data with small sample size, MAPES differs significantly from Lomb-Scargle and Capon schemes. This illustrates that the sample size plays a key role in the decisions made by different schemes. In total, out of the 41 published cell-cycle genes, 15 were recognized from the embryonic data set, while 9 were spotted from the pupal data. We also identified 6 published genes that remain cyclic in both embryonic and pupal stages. For both embryonic and pupal data sets, Lomb-Scargle is the best for recovering cell-cycle genes, while Capon achieves a relatively good performance, and MAPES exhibits poor performance at discerning periodicity from the available data. This is mainly

attributed to the very small sample size.

Immediately after fertilization, each of the first 10 division cycles of the Drosophila spends around 10 minutes. However, in the available microarray data, the first 13 samples in the embryonic stage were taken every 30 minutes. Based on the available measurements, it is impossible to identify genes that play major roles in the early embryonic stage. Therefore, combined with a stringent p-value threshold, a portion of published 41 genes could not be recovered.

A gene's periodicity need not to hold throughout its life span, in other words, an embryo-periodic gene can lose its periodicity in the pupal stage and vice versa. Actually, it has been discovered that different genes control the cell cycle at different developmental stages. Our simulation results also verify this property. A comparative genomic study is also performed. By comparing the Drosophila data set with the Human Hela cyclic gene list published in [9], we find that 77 cyclic Human Hela genes also appear in the Drosophila genome. Out of these 77 genes, 34 of them were measured in our data sets and the schemes were able to detect 12 of these genes as being cyclic. Lomb-Scargle and Capon methods still achieve the best performance. These results show that the cell-cycle involved genes, together with their functions, are preserved along the evolution. Therefore, Drosophila does represent a good model for exploring diseases occurring in humans.

4. REFERENCES

[1] C. D. Giurcaneanu, "Stochastic complexity for the detection of periodically expressed genes," in *Proceedings of IEEE International Workshop on Genomic Signal processing and Statistics (GENSIPS)*, Tuusula, Finland, Jun. 2007.

[2] M. Ahdesmaki, M., H. Lahdesmaki, et al., "Robust detection of periodic time series measured from biological systems," *BMC Bioinformatics*, vol. 6:117, 2007.

[3] T. Bowles, A. Jakobsson and J. Chambers, "Detection of cell-cyclic elements in mis-sampled Gene expression Data Using a Robust Capon Estimator," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, V417–420, Montreal, Quebec, Canada, May 2004.

[4] J.D. Scargle, "Statistical aspects of spectral analysis of unevenly spaced data," *The Astrophysics Journal*, vol. 263, pp. 835–853, 1982.

[5] P. Stoica and N. Sandgren, "Spectral analysis of irregularly-sampled data: parallelling the regularly-sampled Data Approaches," *Digital Signal Processing*, vol. 16, pp. 712–734, 2006.

[6] Y. Wang, P. Stoica, et al, "Nonparametric spectral analysis with missing data via the EM algorithm," *Digital Signal Processing*, vol. 15, pp. 191–206, 2005.

[7] M. N. Arbeitman, E. M. Furlong, et al., "Gene expression during the life cycle of Drosophila melanogaster," *Science*, vol. 297, pp. 2270–2275, 2003.

[8] "Cell Cycle involved genes," available at: http://www.sdbonline.org/fly/aignfam/cellcycle.htm.

[9] M. L. Whitfield, G. Sherlock, et al., "Identification of genes periodically expressed in the human cell cycle and their expression in tumors," *Molecular Biology of the Cell*, vol. 13, pp. 1977–2000, 2002.

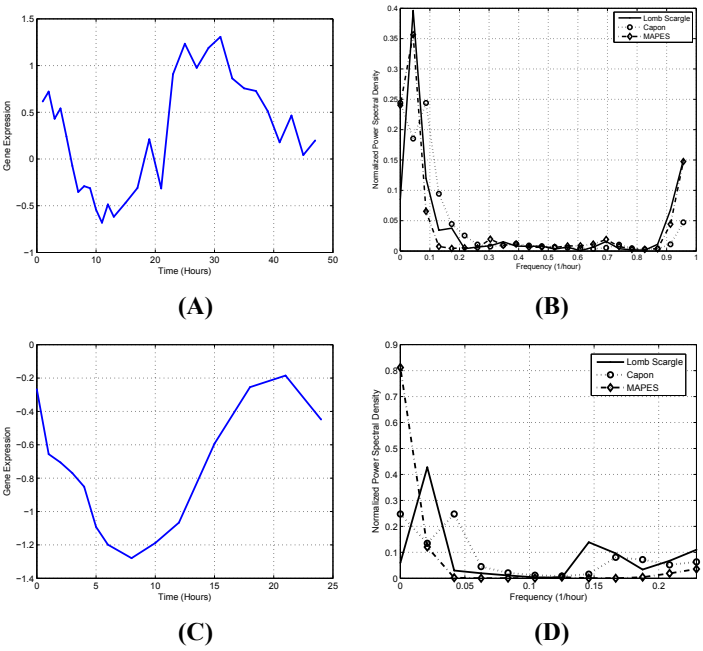

Fig. 1. Gene CG8199: (A) embryonic time series; (B) embryonic spectral. (C) pupal time series; (D) pupal spectral. The spectral density is normalized over the summation of spectra at all probed frequencies.

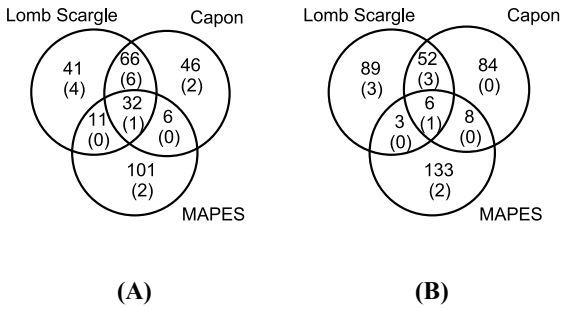

Fig. 2. (A) Embryonic Venn Graph; (B) Pupal Venn Graph. Each scheme preserves 150 genes with the lowest periodic p-values. The number within the parentheses indicates the number of cell cycle genes that have been published, while the number on top of the parentheses implies the number of identified genes by the corresponding scheme.