ANALYSIS OF TEMPORAL GENE EXPRESSION PROFILES USING TIME-DEPENDENT MUSIC ALGORITHM

Nidhal Bouaynaya[1]*, Dan Schonfeld*[2] *and Radhakrishnan Nagarajan*[3]

[1]Dept. of Systems Engineering, University of Arkansas at Little Rock,
[2]Dept. of Electrical and Computer Engineering, University of Illinois at Chicago,
[3] Dept. of Biostatistics, University of Arkansas for Medical Sciences

ABSTRACT

Identifying periodically expressed genes and their subsequent transcriptional circuitry can shed new lights in studying the molecular basis of many diseases including cancer; and subsequently provide potential drug targets to treat them. Classical approaches for detecting periodically expressed transcripts in paradigms such as cell-cycle implicitly assume the given data to be stationary. However, it has been experimentally shown that modulation in the magnitude of gene expression is ubiquitous and defy stationary assumptions. In this paper, we formulate the problem of estimating the frequencies of multicomponent amplitude modulated (AM) signals as a hypothesis testing problem based on a time-dependent extension of the MUSIC algorithm. We subsequently propose a test statistic to detect periodic components in AM time-series. The power of the proposed algorithm is assessed in synthetic test signals and in real cell-cycle gene profiles extracted from microarray data.

1. INTRODUCTION

High-throughput techniques such as microarrays have been used recently to capture the temporal expression of several thousand genes simultaneously across distinct biological paradigms including cell cycle [1–3]. Identifying periodically regulated genes is an important and challenging problem. It is important because (i) it can help us detect the deleterious genes in cancerous cells, for which there is a discrepancy in either the shape or the period of the genes' expressions between normal and cancerous cells; and (ii) it can shed new lights in studying the molecular basis of many diseases, and subsequently provide potential drug targets to treat them. For example, symptoms in Parkinsons disease tend to fluctuate in a circadian manner [4]. The challenge of the problem stems from (i) the large number (thousands) of genes that have to be simultaneously measured; (ii) the small number (3 to 20) of measurements taken per gene; and (iii) the highly non-Gaussian nature of the noise embedded in the data [1].

Interestingly, most of the efforts dedicated to finding periodicity in microarray data sets have relied on Fisher's maximum periodogram method and other *ad hoc* variations of it [1–3]. However, besides its well-known bias, poor performance for short time-series typical of microarray time course data, and spectral leakage, the periodogram is only valid for stationary data series. Per contra, genomic data series have been shown to be non-stationary [5]. In particular, many genes known to be cyclically expressed (also called "clock" genes) have been recently found to exhibit considerable amplitude modulation in the magnitude of their expressions across the time points [6]. Such amplitude modulation has been attributed to "master" genes that are involved in the control of the circadian phase and amplitude of clock genes [6]. Also, external enzymes, such as drugs, were found to modulate the amplitude of clock genes. Finally, the inherent heterogeneity and noisiness characteristic of transcriptional regulation in cell populations may also explain the observed amplitude modulation. Thus, it is crucial to develop techniques that accommodate such non-stationarities to reach meaningful biological interpretation of the results.

The present study provides a systematic approach for identifying genes that show amplitude modulated periodic patterns during the time course of a biological process. Specifically, we extend the MUltiple Signal Classification (MUSIC) method [7] to amplitude modulated (AM) signals. We show that the peaks of the time-dependent MUSIC (TD-MUSIC) pseudo-spectrum correspond to the signal frequencies. We, subsequently, propose a statistic to determine whether these peaks are significant or not. Finally, unlike the maximum periodogram method, [1–3], our approach does not assume Gaussian noise characteristics.

2. PERIODICITY DETECTION

The problem of detecting periodicity in a time-series can be formulated as a statistical decision problem using hypothesis testing as follows: Given N observations $x[1], x[2], \cdots, x[N-1]$, consider the model

$$x[n] = \sum_{i=1}^{p} A_i[n] e^{j(\omega_i n + \phi_i)} + w[n], \qquad (1)$$

where $w[n]$ is an additive white noise process, $\{\omega_i\}$, $\{\phi_i\}$ and $\{A_i[n]\}$ are the unknown angular frequencies, initial phases, and time-dependent amplitudes, respectively. Observe that we do not make any assumptions on the noise distribution. To test for periodic components, we define

the test

$$\underline{H_0} : A_i[n] = 0,\ \forall i, \forall n;$$
$$\underline{H_1} : \exists\ A_i[n] \neq 0,\ \text{for some } i = 1, \cdots, M.$$

where the symbols $\forall$ and $\exists$ denote "for all" and "there exists", respectively.

3. TIME-DEPENDENT MUSIC

Consider first the monocomponent AM signal

$$x[n] = A[n]\ e^{j(n\omega_0+\phi_0)} + w[n], \quad n = 0, 1, \cdots, N-1. \tag{2}$$

Let σ^2 be the variance of the noise $w[n]$. We assume that the initial phase ϕ_0 is a random variable uniformly distributed on $[-\pi, \pi]$. We further make the assumption that the time-dependent amplitude $A[n]$ can be expressed as a linear combination of some basis functions $\{f_k[n]\}_{k=0}^{M}$,

$$A[n] = \sum_{k=1}^{M} c_k f_k[n], \tag{3}$$

where $M \leq N$. Observe that if the basis $\{f_k[n]\}_{k=0}^{M}$ is orthonormal, then the coefficients c_k are found by taking the dot product of $A[n]$ with the basis vector f_k, i.e., $c_k = \sum_{n=0}^{N-1} A[n] f_k^*[n]$, where * denotes complex conjugate. The basis functions $\{f_k[n]\}_{k=1}^{M}$ do not have to be limited to the standard choices of polynomial functions, Legendre or Fourier, but can also take advantage of any prior information, such as the presence of a jump in the coefficients at a known instant, circadian effects, etc. The autocorrelation sequence is given by

$$\begin{aligned} r[n,m] &= E[x[n]x^*[m]] \\ &= \sum_{k,l=1}^{M} c_k c_l^* f_k[n] f_l^*[m] e^{j(n-m)\omega_0} + \sigma^2 \delta[n-m], \end{aligned} \tag{4}$$

where $\delta[n-m]$ is the Kronecker delta function. Let ^{T} and ^{H} denote transpose and conjugate transpose, respectively, and consider the data vector $\mathbf{x} = [x[0], \cdots, x[N-1]]^T$. The autocorrelation matrix, $\mathbf{R}_x = E[\mathbf{x}\mathbf{x}^H]$, may then be written as

$$\mathbf{R}_x = \mathbf{R}_s + \mathbf{R}_n = \sum_{k=1}^{M}\sum_{l=1}^{M} c_k c_l^* \mathbf{\Gamma}(k,l) + \sigma^2 \mathbf{I}, \tag{5}$$

where $\mathbf{I}$ denotes the identity matrix, and

$$\mathbf{\Gamma}(k,l) = \begin{pmatrix} f_k[0]f_l^*[0] & f_k[0]f_l^*[1]e^{-j\omega_0} & \cdots & f_k[0]f_l^*[N-1]e^{-j(N-1)\omega_0} \\ f_k[1]f_l^*[0]e^{j\omega_0} & f_k[1]f_l^*[1] & \cdots & f_k[2-N]f_l^*[N-1]e^{-j(N-2)\omega_0} \\ \vdots & \vdots & \vdots & \vdots \\ f_k[N-1]f_l^*[0]e^{j(N-1)\omega_0} & f_k[N-2]f_l^*[1]e^{j(N-2)\omega_0} & \cdots & f_k[N-1]f_l^*[N-1] \end{pmatrix}. \tag{6}$$

The matrix $\mathbf{\Gamma}(k,l)$ can be concisely written as

$$\mathbf{\Gamma}(k,l) = \mathbf{e}_k(\omega_0)\mathbf{e}_l^H(\omega_0), \tag{7}$$

where

$$\mathbf{e_i}(\omega) = [f_i[0], f_i[1]e^{j\omega}, \cdots, f_i[N-1]e^{j(N-1)\omega}]^T. \tag{8}$$

Subsequently, the signal autocorrelation matrix reduces to

$$\mathbf{R}_s = \sum_{k=1}^{M}\sum_{l=1}^{M} c_k\ c_l^*\ \mathbf{e}_k(\omega_0)\ \mathbf{e}_l^H(\omega_0). \tag{9}$$

Let

$$\mathbf{E}(\omega) = [\mathbf{e_1}(\omega), \mathbf{e_2}(\omega), \cdots, \mathbf{e_M}(\omega)], \tag{10}$$

and $\mathbf{C} = \mathbf{c}\ \mathbf{c^H}$, where $\mathbf{c} = [c_1, c_2, \cdots, c_M]^T$. Then, Eq. (9) can be rewritten as

$$\mathbf{R}_s = \mathbf{E}(\omega_0)\ \mathbf{C}\ \mathbf{E^H}(\omega_0) = (\mathbf{E}(\omega_0)\mathbf{c})(\mathbf{E}(\omega_0)\mathbf{c})^H. \tag{11}$$

Thus, the signal autocorrelation matrix $\mathbf{R}_s$ is positive definite of rank 1. Moreover, its unique positive eigenvalue is given by $||\mathbf{E}(\omega_0)\mathbf{c}||^2$, and its corresponding eigenvector is $\mathbf{E}(\omega_0)\mathbf{c}$. The autocorrelation matrix of the data can then be concisely expressed as

$$\mathbf{R}_x = \mathbf{E}(\omega_0)\ \mathbf{C}\ \mathbf{E^H}(\omega_0) + \sigma^2\mathbf{I}. \tag{12}$$

and its eigenvalues, λ_i, are given by

$$\lambda_i = \lambda_i^s + \sigma^2,\ i = 1, \cdots, N, \tag{13}$$

where $\lambda_1^s = ||\mathbf{E}(\omega_0)\mathbf{c}||^2$ is the positive eigenvalue of $\mathbf{R}_s$. Arranging the eigenvalues of $\mathbf{R}_x$ in non-decreasing order leads to the following result:

$$\begin{cases} \lambda_1 > \sigma^2 \\ \lambda_i = \sigma^2, \quad \text{for } i = 2, \cdots, N. \end{cases}$$

Since $\mathbf{R}_x$ is Hermitian, its eigenvectors, $\mathbf{v}_i$, are orthogonal. In particular, the signal subspace, consisting of the signal eigenvector, $\mathbf{E}(\omega_0)\mathbf{c}$, and the noise subspace, corresponding to the remaining eigenvectors, are orthogonal. Therefore, the frequency ω_0 may be estimated by the location of the highest peak of the TD-MUSIC pseudo-spectrum

$$P_{\text{TD-MUSIC}}(\omega) = \frac{1}{\sum_{k=2}^{N} |(\mathbf{E}(\omega)\mathbf{c})^H \mathbf{v}_k|^2}. \tag{14}$$

Equation (14) is referred to as a "pseudo-spectrum" because it indicates the presence of sinusoidal components in the studied signal, but it is not a true Power Spectral Density. The extension of the above derivation to multicomponent AM signals is summarized in the following proposition.

Proposition 1 (Multicomponent AM signals) *Consider the multicomponent AM signal*

$$x[n] = \sum_{i=1}^{p} A_i[n] e^{j(\omega_i n+\phi_i)} + w[n], \tag{15}$$

where the initial phases $\{\phi_i\}$ are independent random variables uniformly distributed on $[-\pi, \pi]$. We assume that the time-dependent amplitudes $A_i[n]$ can be expressed as a linear combination of some basis functions $\{f_k[n]\}_{k=1}^{M}$;

$$A_i[n] = \sum_{k=1}^{M} c_{i,k} f_k[n]. \tag{16}$$

Then the signal frequencies $\omega_1, \cdots, \omega_p$ may be estimated by the location of the p highest peaks of the following TD-MUSIC pseudo-spectrum

$$P_{\text{TD-MUSIC}}(\omega) = \frac{1}{\sum_{i=1}^{p} \sum_{k=p+1}^{N} |(\mathbf{E}(\omega)\mathbf{c_i})^H \mathbf{v_k}|^2}, \quad (17)$$

where $\mathbf{c_i} = [c_{i,1}, c_{i,2}, \cdots, c_{i,M}]$, $\mathbf{E}(\omega)$ is given by Eq. (10), and the $\mathbf{v}_k$'s are the noise eigenvectors (corresponding to the eigenvalue σ^2) of the autocorrelation matrix

$$\mathbf{R}_x = \sum_{i=1}^{p} (\mathbf{E}(\omega_i)\mathbf{c_i})\ (\mathbf{E}(\omega_i)\mathbf{c_i})^H + \sigma^2 \mathbf{I}. \quad (18)$$

4. TEST FOR PERIODIC COMPONENTS

The TD-MUSIC pseudo-periodogram is evaluated at the frequencies $\omega_j = \frac{2\pi j}{N},\ j = 0, 1, \cdots, N-1$. Let $P_1 < P_2 < \cdots < P_N$ denote the TD-MUSIC pseudo-spectrum ordinates in ascending order. Define the statistic

$$U(r) = \frac{P_{N-r+1}}{\sum_{k=1}^{N} P_k}. \quad (19)$$

If r peaks are indicated by the statistics $U(r)$, then the estimates of the frequencies are taken to be the frequencies associated with the r TD-MUSIC pseudo-spectrum ordinates $P_{N-r+1}, \cdots, P_N$. Hence, if $U^*(r)$ is the observed value of $U(r)$, then Eq. (19) yields a p-value, $P(U(r) > U^*(r))$, that allows to test whether the r maximum peaks in the TD-MUSIC pseudo-spectrum $P(\omega)$ are significant. To obtain the p-value for the observed sequence, we generate a set of random time-series, evaluate the test statistic for each one of the time-series, and use the obtained $U(r)$-values to compute an estimate of the distribution of the $U(r)$-statistic under the null-hypothesis; thus, obtaining the p-value of the original test-statistic $U^*(r)$ as an estimate of $P(U(r) > U^*(r))$. We reject the null hypothesis if the p-value is smaller than a chosen significance level α.

5. SIMULATION RESULTS

We first consider a synthetic test signal given by

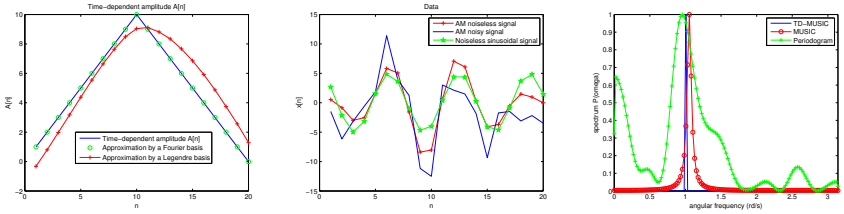
$$x[n] = A[n]\cos(\frac{41}{128}\pi n) + w[n],\ n = 1, \cdots, N, \quad (20)$$

where $A[n]$ is a triangular signal given by $A[n] = \begin{cases} n, & 1 \le n \le \frac{N}{2}; \\ -n+N, & \frac{N}{2} \le n \le N. \end{cases}$, and $w[n]$ is a random Gaussian noise. the signal length is set to $N = 20$ (typical for microarray data). Figure 1(a) shows the time-dependent amplitude $A[n]$, its Fourier basis representation with $M = 40$, and its Legendre basis representation with $M = 6$. Observe that the Fourier basis provides an exact representation of the time-dependent amplitude, whereas the Legendre basis provides a close approximation. Figure 1(b) shows the pure sinusoidal signal $\cos(\frac{41\pi}{128}n)$, its noiseless AM modulated version by $A[n]$, and its noisy AM modulated version with a Signal to Noise Ratio (SNR) equal to 0 dB. Figure 1(c) displays the TD-MUSIC pseudo-spectrum, the MUSIC pseudo-spectrum, and the periodogram of the noisy AM modulated signal. For display clarity, the three spectrums have been scaled to $[0, 1]$. It is clear that, with the sharpest main lobe, the TD-MUSIC provides the most accurate estimation of the frequency $\omega_0 = \frac{41\pi}{128}$. The broad main lobe of the periodogram, and its extensive leakage are mainly due to the non-stationarity of the signal, and the short length of the data.

We then tested the proposed TD-MUSIC algorithm on microarray real data. Specifically, we investigated temporal expression profiles from high-throughput microarray data generated in the alpha-factor synchronization yeast cell-cycle experiment consisting of 6178 genes across 18 time points [8]. We first eliminated all genes whose values are missing even at a single time point resulting in a reduced set consisting of 4491 genes. Spellman et al. [8] identified 104 sinusoidal genes as well-documented through extensive literature survey of the cell-cycle paradigm. Out of these 104 genes, 72 genes had values across all 18 time points (i.e. overlapped with the list of 4491 genes). We assessed the time-dependent amplitude of each gene by dividing the signal by its MUSIC estimated sinusoidal component. We, subsequently, applied the TD-MUSIC algorithm using the Fourier basis representation, and a significance value $\alpha = 0.05$. The TD-MUSIC algorithm detected all 72 ground truth genes. The stationary MUSIC algorithm missed 13 (18%) periodic genes, and the periodogram missed 35 of them (more than 48%). Interestingly, all the 14 genes missed by the stationary MUSIC visually exhibited amplitude modulated profiles. The expression plots (versus time) of three of these genes is given in Fig. 2 along with their corresponding TD-MUSIC pseudo-spectrums, MUSIC pseudo-spectrums, and periodograms. We observe that the time-dependent amplitude seems to consist of jumps at certain time points for the gene YDL179W (PCL9), whereas it clearly shows more complex patterns for the genes YCL055W (KAR4), and YCL027W (FUS1). While YDL179W and YCL027W are regulated during the M/G1 boundary, YCL055W is regulated during the G1/SCB (i.e. SwI4, 6-dependent cell cycle box, CACGAAA) phase of the cell-cycle. In all cases, the TD-MUSIC pseudo-spectrum estimates the frequency of each gene profile with a higher resolution than the MUSIC and the periodogram.

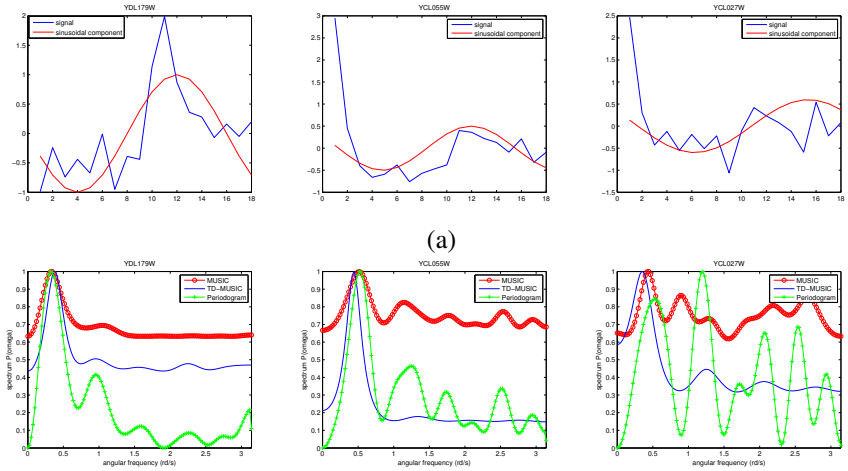
6. CONCLUSION

Classical approaches for detecting periodically expressed transcripts in paradigms such as cell-cycle implicitly assume the given data to be stationary. However, many genes known to be cyclically expressed have been recently found to exhibit considerable amplitude modulation in the magnitude of their expressions [6]. Such modulations can be an outcome of biological as well as non-biological factors. In this paper, we extended the MUSIC algorithm to multicomponent AM signals, and showed its power in detecting periodic components in AM modulated signals, specifically, in cell-cycle gene profiles. Our algorithm as-



(a) (b) (c)

Figure 1. (a) The Time-dependent amplitude $A[n]$ and its Fourier and Legendre basis representations; (b) The pure sinusoidal signal, its noiseless AM modulated version, and its noisy AM modulated version with SNR = 0 dB; (c) The TD-MUSIC pseudo-spectrum, the MUSIC pseudo-spectrum, and the periodogram of the noisy AM modulated signal.



(a)

(b)

Figure 2. (a) The expression plot (versus time) of the genes YDL179W, YCL055W, and YCL027W, respectively; (d) The TD-MUSIC pseudo-spectrum, the MUSIC pseudo-spectrum, and the periodogram of the gene profiles in (a).

sumes, however, that the shape of the time-dependent amplitude is known a priori. This assumption can be circumvented by estimating the time-dependent amplitude using the stationary MUSIC algorithm. The TD-MUSIC can, therefore, be applied sequentially, where each iteration provides a better estimate of the time-dependent amplitude and hence of the signal frequencies. The performance analysis of the sequential TD-MUSIC will be the subject of our future work, which will also extend the TD-MUSIC to amplitude and frequency modulated (AM-FM) signals.

7. REFERENCES

[1] S Wichert, K Fokianos, and K Strimmer, "Identifying periodically expressed transcripts in microarray time series data," *Bioinformatics*, vol. 20, pp. 5–20, 2004.

[2] M Ahdesmki, H Lhdesmki, R Pearson, H Huttunen, and O Yli-Harja, "Robust detection of periodic time series measured from biological systems," in *BMC Bioinformatics 2005*, 2005, pp. 117–135.

[3] J Chen, "identification of significant periodic genes in microarray gene expression data," *BMC Bioinformatics*, vol. 6, pp. 286–297, 2005.

[4] J G Nutt, J H Carter, E S Lea, and W R Woodward, "Motor fluctuations during continuous levodopa infusions in patients with parkinsons disease," *Movement Disorders*, vol. 12, pp. 285292, 1997.

[5] N Bouaynaya and D Schonfeld, "Non-stationary analysis of coding and non-coding regions in nucleotide sequences," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 3, pp. 357 – 364, 2008.

[6] A Nakashima, T Kawamoto, K Honda, T Ueshima, M Noshiro, T Iwata, K Fujimoto, H Kubo, S Honma, N Yorioka, N Kohno, and Y Kato, "DEC1 modulates the circadian phase of clock gene expression," *Molecular and Cellular Biology*, vol. 28, no. 12, pp. 4080–4092, June 2008.

[7] P Stoica and A Nehorai, "Music, maximum likelihood, and cramer-rao bound," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 5, pp. 720–741, May 1989.

[8] P T Spellman, G Sherlock, M Q Zhang, V R Iyer, K Anders, M B Eisen, P O Brown, D Botstein, and B Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization," *Molecular Biology of the Cell*, vol. 9, pp. 3273–3297, 1998.