Hindawi Publishing Corporation
Advances in Bioinformatics
Volume 2013, Article ID 171530, 10 pages
http://dx.doi.org/10.1155/2013/171530

Research Article

Spectral Analysis on Time-Course Expression Data: Detecting Periodic Genes Using a Real-Valued Iterative Adaptive Approach

Kwadwo S. Agyepong,[1] Fang-Han Hsu,[1] Edward R. Dougherty,[1,2] and Erchin Serpedin[1]

[1] *Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843-3128, USA*
[2] *Computational Biology Division, Translational Genomics Research Institute, Phoenix, AZ 85004-2101, USA*

Correspondence should be addressed to Erchin Serpedin; serpedin@ece.tamu.edu

Received 26 October 2012; Accepted 23 January 2013

Academic Editor: Mohamed Nounou

Time-course expression profiles and methods for spectrum analysis have been applied for detecting transcriptional periodicities, which are valuable patterns to unravel genes associated with cell cycle and circadian rhythm regulation. However, most of the proposed methods suffer from restrictions and large false positives to a certain extent. Additionally, in some experiments, arbitrarily irregular sampling times as well as the presence of high noise and small sample sizes make accurate detection a challenging task. A novel scheme for detecting periodicities in time-course expression data is proposed, in which a real-valued iterative adaptive approach (RIAA), originally proposed for signal processing, is applied for periodogram estimation. The inferred spectrum is then analyzed using Fisher's hypothesis test. With a proper p-value threshold, periodic genes can be detected. A periodic signal, two nonperiodic signals, and four sampling strategies were considered in the simulations, including both bursts and drops. In addition, two yeast real datasets were applied for validation. The simulations and real data analysis reveal that RIAA can perform competitively with the existing algorithms. The advantage of RIAA is manifested when the expression data are highly irregularly sampled, and when the number of cycles covered by the sampling time points is very reduced.

1. Introduction

Patterns of periodic gene expression have been found to be associated with essential biological processes such as cell cycle and circadian rhythm [1], and the detection of periodic genes is crucial to advance our understanding of gene function, disease pathways, and, ultimately, therapeutic solutions. Using high-throughput technologies such as microarrays, gene expression profiles at discrete time points can be derived and hundreds of cell cycle regulated genes have been reported in a variety of species. For example, Spellman et al. applied cell synchronization methods and conducted time-course gene expression experiments on *Saccharomyces cerevisiae* [2]. The authors identified 800 cell cycle regulated genes using DNA microarrays. Also, Rustici et al. and Menges et al. identified 407 and about 500 cell cycle regulated genes in *Schizosaccharomyces pombe* and *Arabidopsis*, respectively [3, 4].

Signal processing in the frequency domain simplifies the analysis and an emerging number of studies have demonstrated the power of spectrum analysis in the detection of periodic genes. Considering the common issues of missing values and noise in microarray experiments, Ahdesmäki et al. proposed a robust detection method incorporating the fast Fourier transform (FFT) with a series of data preprocessing and hypothesis testing steps [5]. Two years later, the authors further proposed a modified version for expression data with unevenly spaced time intervals [6]. A Lomb-Scargle (LS) approach, originally used for finding periodicities in astrophysics, was developed for expression data with uneven sampling [7]. Yang et al. further improved the performance using a detrended fluctuation analysis [8]. It used harmonic regression in the time domain for significance evaluation. The method was termed "Lomb-Scargle periodogram and harmonic regression (LSPR)." Basically, these methods consists of two steps: transferring the signals into the frequency

(spectral) domain and then applying a significance evaluation test for the resulting peak in the spectral density.

While numerous methods have been developed for detecting periodicities in gene expression, most of these methods suffer from false positive errors and working restrictions to a certain extent, particularly when the time-course data contain limited time points. In addition, no algorithm seems available to resolve all of these challenges. Microarray as well as other high-throughput experiments, due to high manufacturing and preparation costs, have common characteristics of small sample size [9], noisy measurements [10], and arbitrary sampling strategies [11], thereby making the detection of periodicities highly challenging. Since the number and functions of cell cycle regulated genes, or periodic genes, remain greatly uncertain, advances in detection algorithms are urgently needed.

Recently, Stoica et al. developed a novel nonparametric method, termed the "real-valued iterative adaptive approach (RIAA)," specifically for spectral analysis with nonuniformly sampled data [12]. As stated by the authors, RIAA, an iteratively weighted least-squares periodogram, can provide robust spectral estimates and is most suitable for sinusoidal signals. These characteristics of RIAA inspired us to apply it to time-course gene expression data and conduct an examination on its performance. Herein, we incorporate RIAA with a Fisher's statistic to detect transcriptional periodicities. A rigorous comparison of RIAA with several aforementioned algorithms in terms of sensitivities and specificities is conducted through simulations and simulation results dealing with real data analysis are also provided.

In this study, we found that the RIAA algorithm can provide robust spectral estimates for the detection of periodic genes regardless of the sampling strategies adopted in the experiments or the nonperiodic nature of noise present in the measurement process. We show through simulations that the RIAA can outperform the existing algorithms particularly when the data are highly irregularly sampled, and when the number of cycles covered by the sampling time points is very few. These characteristics of RIAA fit perfectly the needs of time-course gene expression data analysis. This paper is organized as follows. In Section 2, we begin with an overview of RIAA. In Section 3, a scheme for detecting periodicities is proposed, and simulation models for performance evaluation and a real data analysis for validation purposes are presented. A complete investigation of the performance of RIAA and a rigorous comparison with other algorithms are provided in Section 4.

2. RIAA Algorithm

RIAA is an iterative algorithm developed for finding the least-squares periodogram with the utilization of a weighted function. The essential mathematics involved in RIAA is introduced in this section with the algorithm input being time-course expression data; for more details regarding RIAA, the readers are encouraged to check the original paper by Stoica et al. [12].

2.1. Basics. Suppose that the signals associated with the periodic gene expressions are composed of noise and sinusoidal components. Let $y_h(t_i)$, $i = 1, \ldots, n$, denote the time-course expression ratios of gene h at instances $t_1, \ldots, t_n$, respectively; $y_h(t_i)$ are real numbers; $\sum_{i=1}^{n} y_h(t_i) = 0$. The least-squares periodogram Φ_{lsp} is given by

$$\Phi_{lsp} = |\widehat{\alpha}(\omega)|^2, \tag{1}$$

where $\widehat{\alpha}(\omega)$ is the solution to the following fitting problem:

$$\widehat{\alpha}(\omega) = \arg\min_{\alpha(\omega)} \sum_{i=1}^{n} \left[y_h(t_i) - \alpha(\omega) e^{j\omega t_i} \right]^2. \tag{2}$$

Let $\alpha(\omega) = |\alpha(\omega)| e^{j\phi(\omega)} = \beta e^{j\theta}$, where $\beta = |\alpha(\omega)| \geq 0$ and $\theta = \phi(\omega) \in [0, 2\pi]$ refer to the amplitude and phase of $\alpha(\omega)$, respectively. The criterion in (2) can then be rewritten as

$$\sum_{i=1}^{n} \left[y_h(t_i) - \beta \cos(\omega t_i + \theta) \right]^2 + \beta^2 \sum_{i=1}^{n} \sin^2(\omega t_i + \theta). \tag{3}$$

The second term in the above equation is data independent and can be omitted from the minimization operation. Hence, the criterion (2) is simplified to

$$\left(\widehat{\beta}, \widehat{\theta}\right) = \arg\min_{\beta,\theta} \sum_{i=1}^{n} \left[y_h(t_i) - \beta \cos(\omega t_i + \theta) \right]^2. \tag{4}$$

We further apply $a = \beta \cos(\theta)$ and $b = -\beta \sin(\theta)$ and derive an equivalent of (4) as follows:

$$\left(\widehat{a}, \widehat{b}\right) = \arg\min_{a,b} \sum_{i=1}^{n} \left[y_h(t_i) - a \cos(\omega t_i) - b \sin(\omega t_i) \right]^2. \tag{5}$$

The target of interest to the fitting problem now becomes $\widehat{a}$ and $\widehat{b}$ (instead of $\alpha(\omega)$), and the solution is well known to be

$$\begin{bmatrix} \widehat{a} \\ \widehat{b} \end{bmatrix} = \mathbf{R}^{-1}\mathbf{r}, \tag{6}$$

where

$$\mathbf{R} = \sum_{i=1}^{n} \begin{bmatrix} \cos(\omega t_i)^2 & \cos(\omega t_i)\sin(\omega t_i) \\ \sin(\omega t_i)\cos(\omega t_i) & \sin(\omega t_i)^2 \end{bmatrix}, \quad \mathbf{r} = \sum_{i=1}^{n} \begin{bmatrix} \cos(\omega t_i) \\ \sin(\omega t_i) \end{bmatrix} y_h(t_i). \tag{7}$$

After $\widehat{a}$ and $\widehat{b}$ are estimated, the least-squares periodogram can be derived.

2.2. Observation Interval and Resolution. Prior to implementation of RIAA for periodogram estimation, the observation interval $[0, \omega_{\max}]$ and the resolution in terms of grid size have to be selected. To this end, the maximum frequency $\omega_{\max}$ in the observation interval without aliasing errors for sampling instances $t_1, \ldots, t_n$, can be evaluated by

$$\omega_{\max} = \frac{\omega_0}{2}, \tag{8}$$

where ω_0 is given by

$$\omega_0 = \frac{2(n-1)\pi}{\sum_{i=1}^{n-1}(t_{i+1}-t_i)}. \quad (9)$$

The observation interval $[0, \omega_{\max}]$ is hence chosen after $\omega_{\max}$ is obtained.

To ensure that the smallest frequency separation in time-course expression data with regular or irregular sampling can be adequately detected, the grid size $\Delta\omega$ is chosen to be

$$\Delta\omega = \frac{2\pi}{t_n - t_1}, \quad (10)$$

which, in fact, is the resolution limit of the least-squares periodogram. As a result, the frequency grids ω_g considered in periodogram are

$$\omega_g = g\Delta\omega, \quad g = 1, \ldots, G, \quad (11)$$

where the number of grids G is given by

$$G = \left\lfloor \frac{\omega_{\max}}{\Delta\omega} \right\rfloor. \quad (12)$$

2.3. Implementation. The following notations are introduced for the implementation of RIAA at a specific frequency ω_g:

$$\begin{aligned} \mathbf{Y} &= \left[y_h(t_1) \quad \cdots \quad y_h(t_n) \right]^T, \\ \rho_g &= \left[a(\omega_g) \quad b(\omega_g) \right]^T, \\ \mathbf{A}_g &= \left[\mathbf{c}_g \quad \mathbf{s}_g \right], \end{aligned} \quad (13)$$

where

$$\begin{aligned} \mathbf{c}_g &= \left[\cos(\omega_g t_1) \quad \cdots \quad \cos(\omega_g t_n) \right]^T, \\ \mathbf{s}_g &= \left[\sin(\omega_g t_1) \quad \cdots \quad \sin(\omega_g t_n) \right]^T, \end{aligned} \quad (14)$$

and $a(\omega_g)$ and $b(\omega_g)$ denote variables a and b at frequency ω_g, respectively.

RIAA's salient feature is the addition of a weighted matrix $\mathbf{Q}_g$ to the least-squares fitting criterion. The weighted matrix $\mathbf{Q}_g$ can be viewed as a covariance matrix encapsulating the contributions of noise and other sinusoidal components in $\mathbf{Y}$ other than ω_g to the spectrum; it is defined as

$$\mathbf{Q}_g = \mathbf{\Sigma} + \sum_{m=1, m\neq g}^{G} \mathbf{A}_m \mathbf{D}_m \mathbf{A}_m^T, \quad (15)$$

where

$$\mathbf{D}_m = \frac{a^2(\omega_g) + b^2(\omega_g)}{2} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad (16)$$

and $\mathbf{\Sigma}$ denotes the covariance matrix of noise in expression data $\mathbf{Y}$, given by

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2 \end{bmatrix}. \quad (17)$$

Assuming that $\mathbf{Q}_g$ is invertible, in RIAA, a weighted least-squares fitting problem is formulated and considered for finding $\widehat{a}$ and $\widehat{b}$ (instead of using (5)), and it is written in the form of matrices using (13) as follows:

$$\widehat{\rho}_g = \arg\min_{\rho_g} \left[\mathbf{Y} - \mathbf{A}_g \rho_g \right]^T \mathbf{Q}_g^{-1} \left[\mathbf{Y} - \mathbf{A}_g \rho_g \right]. \quad (18)$$

In Stoica et al. [12], the solution to (18) has been shown to be

$$\widehat{\rho}_g = \frac{\mathbf{A}_g^T \mathbf{Q}_g^{-1} \mathbf{Y}}{\mathbf{A}_g^T \mathbf{Q}_g^{-1} \mathbf{A}_g}, \quad (19)$$

and the RIAA periodogram at $\omega = \omega_g$ can be derived by

$$\Phi_{\text{riaa}}(\omega_g) = \frac{1}{n} \widehat{\rho}_g^T \left(\mathbf{A}_g^T \mathbf{A}_g \right) \widehat{\rho}_g. \quad (20)$$

From (15) and (19), it is obvious that $\mathbf{Q}_g$ and $\widehat{\rho}_g$ are dependent on each other. An iterative approach (i.e., RIAA) is hence a feasible solution to get the estimate $\widehat{\rho}_g$ and the weighted matrix $\mathbf{Q}_g$.

The iteration for estimating spectrum starts with initial estimates $\widehat{\rho}_g^0$, in which the elements $\widehat{a}$ and $\widehat{b}$ are given by (6) with $\omega = \omega_g$, $g = 1, \ldots, G$. After initialization, the first iteration begins. First, the elements $\widehat{a}$ and $\widehat{b}$ of $\widehat{\rho}_g^0$ are applied to obtain $\widehat{\mathbf{D}}_m^1$ using (16). Secondly, to get a good estimate of $\widehat{\sigma}^1$, the frequency ω_p at which the largest value-p is located in the temporary periodogram $\Phi^0(\omega_g)$, $g = 1, \ldots, G$, derived using (20) with $\widehat{\rho}_g = \widehat{\rho}_g^0$, is applied for obtaining a reversed engineered signal $\widehat{\mathbf{Y}}^0$. The elements $\widehat{y}_h(t_i)$, $i = 1, \ldots, n$, in $\widehat{\mathbf{Y}}^0$ are given by

$$\widehat{y}_h(t_i) = \sqrt{2P} \cos(\omega_p t_i + s), \quad i = 1, \ldots, n. \quad (21)$$

The phase of the cosine function s is unknown; however, $\widehat{\sigma}^1$ is estimable using

$$\widehat{\sigma}^1 = \min_{s\in[0,2\pi]} \frac{\left\| \mathbf{Y} - \widehat{\mathbf{Y}}^0 \right\|^2}{n}, \quad (22)$$

where $\|\cdot\|$ is the Euclidean norm. With estimates $\widehat{\mathbf{D}}_m^1$ and $\widehat{\sigma}^1$, the estimates $\widehat{\mathbf{Q}}_g^1$, $g = 1, \ldots, G$, in the first iteration are hence given by (15). After this, $\widehat{\mathbf{Q}}_g^1$ are inserted into the right-hand side of (19) and updated estimates $\widehat{\rho}_g^1$, $g = 1, \ldots, G$, are derived. The algorithm consists of repeating these steps and updating $\widehat{\mathbf{Q}}_g^k$ and $\widehat{\rho}_g^k$ iteratively, where k denotes the number of iterations, until a termination criterion is reached. If the process stops at the Kth iteration, then the final RIAA periodogram is given by (20) using $\widehat{\rho}_g^K$. The pseudocode in Algorithm 1 represents a concise description of the iterative RIAA process.

3. Methods

Figure 1 demonstrates our scheme for periodicity detection and algorithm comparison. The first step involves a periodogram estimation, which converts the time-course gene

Algorithm RIAA

Initialization
Use (6) to obtain the initial estimates $\hat{a}$ and $\hat{b}$ in $\hat{\rho}_g^0$.
The First Iteration
Obtain $\widehat{\mathbf{D}}_m^1$ using (16) with parameters $\hat{a}$ and $\hat{b}$ given by $\hat{\rho}_g^0$. Obtain $\hat{\sigma}^1$ using (22). Using $\widehat{\mathbf{D}}_m^1$ and $\hat{\sigma}^1$ to drive the first weighted matrix $\widehat{\mathbf{Q}}_g^1$ by (15). Update estimate $\hat{\rho}_g^1$ by (19) with $\mathbf{Q}_g = \widehat{\mathbf{Q}}_g^1$.
Updating Iteration
At the kth iteration, $k = 1, 2, \ldots$, estimates $\widehat{\mathbf{Q}}_g^k$ and $\hat{\rho}_g^k$ are iteratively updated in the same way as the first iteration.
Termination
Terminate simply after 15 iterations ($K = 15$), or when the total changes in $d_g^k = ||\hat{\rho}_g^k||$ for $g = 1, \ldots, G$, is extremely small, say, $\sqrt{\sum_{g=1}^{G}(d_g^k - d_g^{k-1})^2} < 0.005\sqrt{\sum_{g=1}^{G}(d_g^{k-1})^2}$, then $K = k$.

Algorithm 1: The pseudocode of the iterative process in RIAA.

expression ratios into the frequency domain. Three methods are considered for comparison: RIAA, LS, and a detrend LS (termed DLS), which uses an additional detrend function (developed in LSPR) before regular LS periodogram estimation is applied. The derived spectra are then analyzed using hypothesis testing. This study is conducted using a Fisher's test, with the null hypothesis that there are no periodic signals in the time domain and hence no significantly large peak in the derived spectra. The algorithm performance is evaluated and compared via simulations and receiver operating characteristic (ROC) curves. In real microarray data analysis, three published benchmark sets are utilized as standards of cell cycle genes for performance comparison.

3.1. Fisher's Test. After the spectrum of time-course expression data is obtained via periodogram estimation, a Fisher's statistic f for gene h with the null hypothesis H_0 that the peak of the spectral density is insignificant against the alternative hypothesis H_1 that the peak of the spectral density is significant is applied as

$$f_h = \frac{\max_{1\le g\le G}\left(\Phi\left(\omega_g\right)\right)}{G^{-1}\sum_{g=1}^{G}\Phi\left(\omega_g\right)}, \tag{23}$$

where Φ refers to the periodogram derived using RIAA, LS, or DLS. The null hypothesis H_0 is rejected, and the gene h is claimed as a periodic gene if its p-value, denoted as p_h, is less than or equal to a specific significance threshold. For simplicity, p_h is approximated from the asymptotic null distribution of f assuming Gaussian noise [13] as follows:

$$p_h = 1 - e^{-ne^{-f_h}}. \tag{24}$$

In real data analysis, deviation might be invoked for the estimation of p_h when the time-course data is short. This issue was carefully addressed by Liew et al. [14], and, as suggested, alternative methods such as random permutation may provide less deviation and better performance. However, permutation also has limitations such as tending to be conservative [15]. While finding the most robust method for the p-value evaluation remains an open question, it gets beyond the scope of this study since the algorithm comparison via ROC curves is threshold independent [16], and the results are unaffected by the deviation.

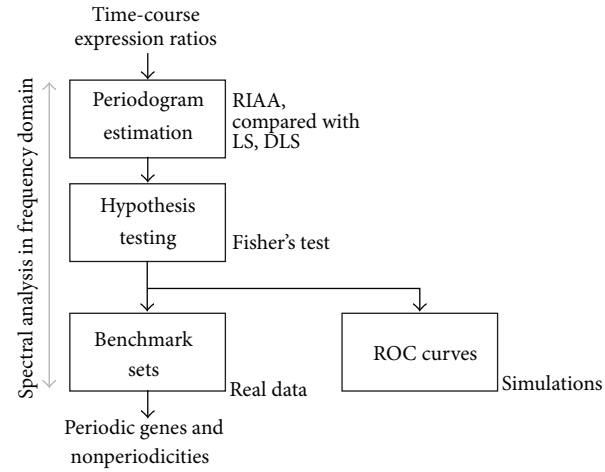

Figure 1: The scheme of the process for detecting periodicities in time-course expression data.

3.2. Simulations. Simulations are applied to evaluate the performance of RIAA. The simulation models and sampling strategies used for simulations are described in the following paragraphs.

3.2.1. Periodic and Nonperiodic Signals. Three models, one for periodic signals and two for nonperiodic signals, are considered as transcriptional signals. Since periodic genes are transcribed in an oscillatory manner, the expression levels y_s embedded with periodicities are assumed to be

$$y_s\left(t_i\right) = M\cos\left(\omega_s t_i\right) + \epsilon_{t_i}, \quad i = 1, \ldots, n, \tag{25}$$

where M denotes the sinusoidal amplitude; ω_s refers to the signal frequency; ϵ_{t_i} are Gaussian noise independent and

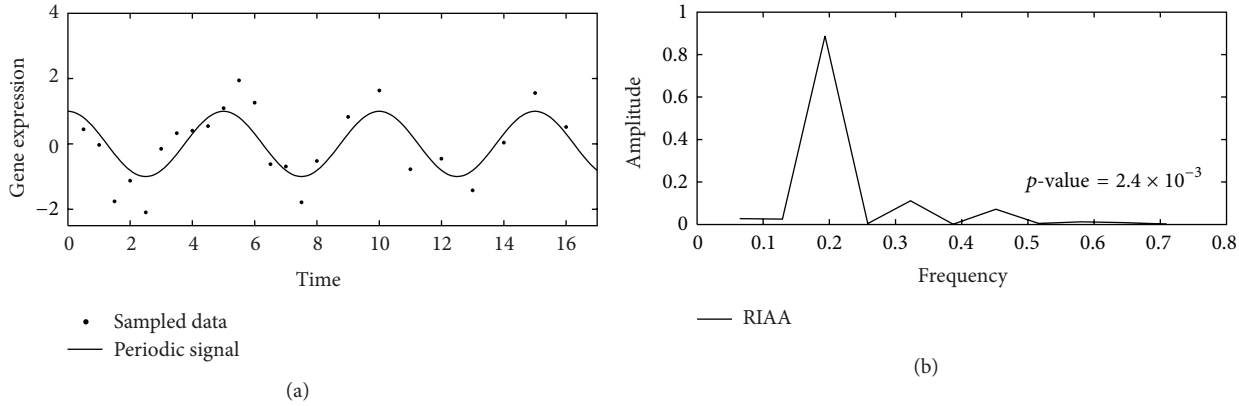

Figure 2: (a) A time-course periodic signal with frequency = 0.2 sampled by the bio-like sampling strategy; 16 time points are assigned to the interval (0,8], and 8 time points are assigned to the interval (8,16]. (b) The periodogram derived using RIAA. The maximum value (peak) in the periodogram locates at frequency = 0.195.

identically distributed (i.i.d.) with parameters μ and σ. For nonperiodic signals, the first model y_n is simply composed of Gaussian noise, given by

$$y_n(t_i) = \epsilon_{t_i}, \quad i = 1, \ldots, n. \tag{26}$$

Additionally, as visualized by Chubb et al., gene transcription can be nonperiodically activated with irregular intervals in a living eukaryotic cell, like pulses turning on and off rapidly and discontinuously [17]. Based on this, the second nonperiodic model y_n' incorporates one additional transcriptional burst and one additional sudden drop into the Gaussian noise, which can be written as

$$y_n'(t_i) = I_b(t_i) - I_d(t_i) + \epsilon_{t_i}, \quad i = 1, \ldots, n, \tag{27}$$

where I_b and I_d are indicator functions, equal to 1 at the location of the burst and the drop, respectively, and 0 otherwise. The transcriptional burst assumes a positive pulse while the transcriptional drop assumes a negative pulse. Both of them may be located randomly among all time points and are assumed to last for two time points. In other words, the indicator functions are equal to 1 at two consecutive time points, say, $I_b = 1$ at t_i and t_{i+1}. The burst and the drop have no overlap.

3.2.2. Sampling Strategies. As for the choices of sampling time points t_i, $i = 1, \ldots, n$, four different sampling strategies, one with regular sampling and three with irregular sampling, are considered. First, regular sampling is applied in which all time intervals are set to be $1/c$, where c is a constant. Secondly, a bio-like sampling strategy is invoked. This strategy tends to have more time points at the beginning of time-course experiments and less time points after we set the first 2/3 time intervals as $1/c$ and set the next 1/3 time intervals as $2/c$. Third, time intervals are randomly chosen between $1/c$ and $2/c$. The last sampling strategy, in which all time intervals are exponentially distributed with parameter c, is less realistic than the others but it is helpful for us to evaluate the performance of RIAA under pathological conditions.

ROC curves are applied for performance comparison. To this end, 10,000 periodic signals were generated using (25) and 10,000 nonperiodic signals were generated using either (26) or (27). Sensitivity measures the proportion of successful detection among the 10,000 periodic signals and specificity measures the proportion of correct claims on the 10,000 nonperiodic simulation datasets. Sampling time points are decided by one of the four sampling strategies and the number of time points n is chosen arbitrarily. For all ROC curves in Section 4, $c = 2$ and $n = 24$.

3.3. Real Data Analysis. Two yeast cell cycle experiments synchronized using an alpha-factor, one conducted by Spellman et al. [2] and one conducted by Pramila et al. [18], are considered for a real data analysis. The first time-course microarray data, termed dataset alpha and downloaded from the Yeast Cell Cycle Analysis Project website (http://genome-www.stanford.edu/cellcycle/), harbors 6,178 gene expression levels and 18 sampling time points with a 7-minute interval. The second time-course data, termed dataset alpha 38, is downloaded from the online portal for Fred Hutchinson Cancer Research Center's scientific laboratories (http://labs.fhcrc.org/breeden/cellcycle/). This dataset contains 4,774 gene expression levels and 25 sampling time points with a 5-minute interval. Three benchmark sets of genes that have been utilized in Lichtenberg et al. [19] and Liew et al. [20] as standards of cell cycle genes are also applied herein for performance comparison. These benchmark sets, involving 113, 352, and 518 genes, respectively, include candidates of cycle cell regulated genes in yeast proposed by Spellman et al. [2], Johansson et al. [21], Simon et al. [22], Lee et al. [23], and Mewes et al. [24] and are accessible in a laboratory website (http://www.cbs.dtu.dk/cellcycle/).

4. Results

RIAA performed well in the conducted simulations. As shown in Figure 2(a), a periodic signal (solid line) with amplitude $M = 1$ and frequency $\omega_s = 0.4\pi$ is sampled

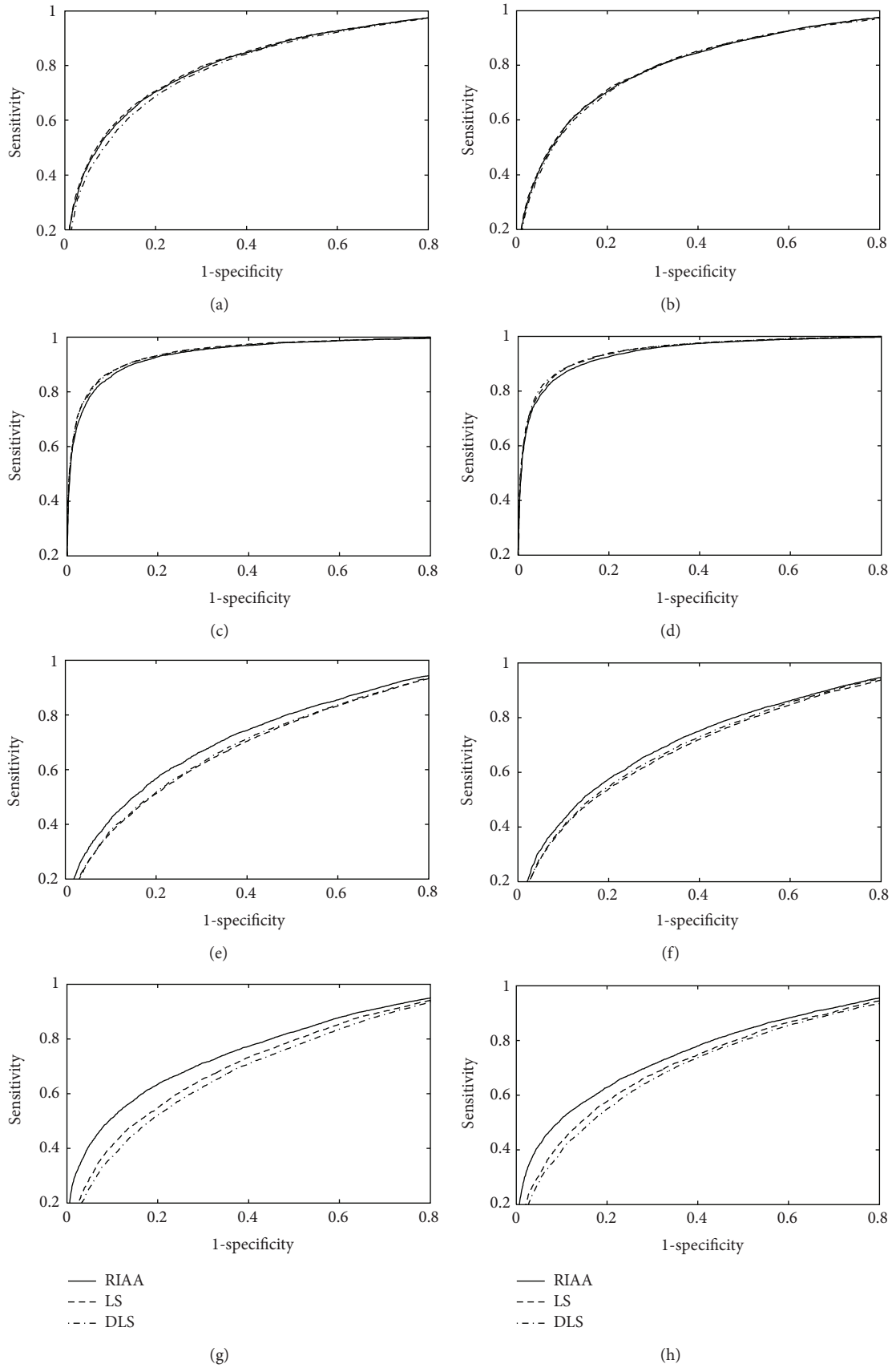

Figure 3: The ROC curves derived from simulations with 24 sampling time points, signal amplitude $M = 1$, $\omega_s = 0.4\pi$, and Gaussian noise $\mu = 0$ and $\sigma = 0.5$. Description of subplots is provided in Section 4.

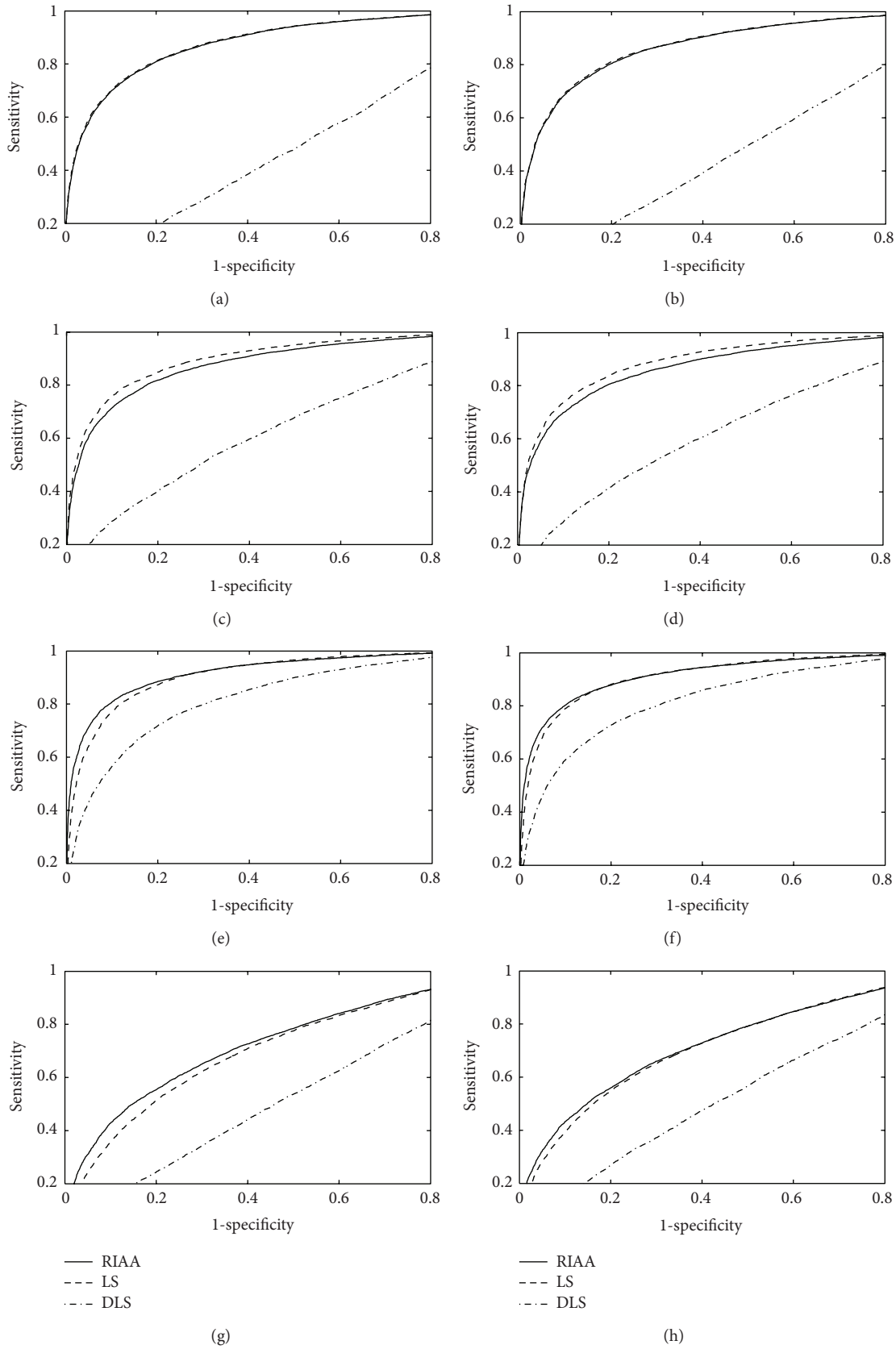

FIGURE 4: The ROC Curves derived from simulations with 24 sampling time points, signal amplitude $M = 1$, $\omega_s = 0.1\pi$, and Gaussian noise $\mu = 0$ and $\sigma = 0.5$. Description of subplots is provided in Section 4.

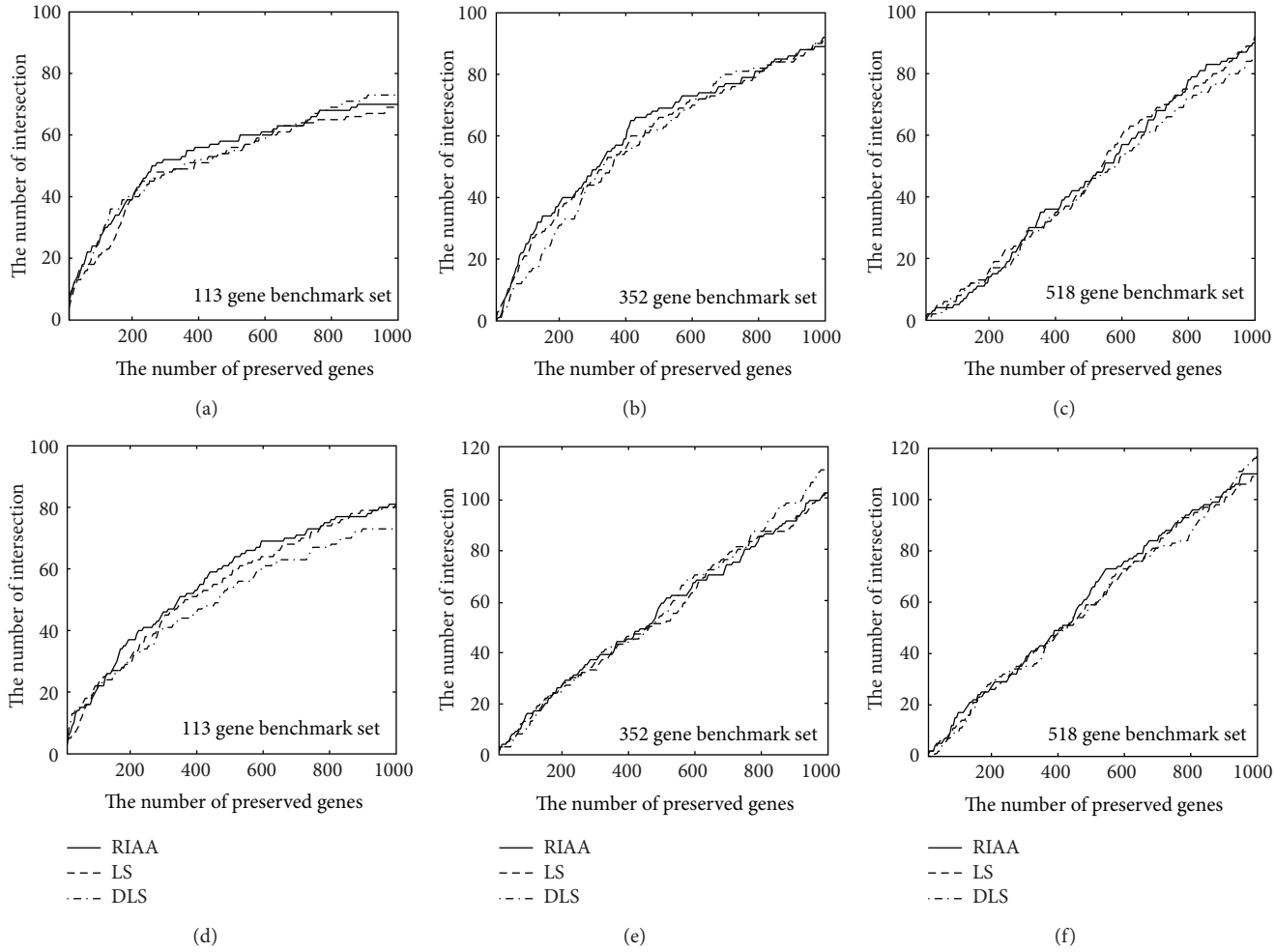

Figure 5: The intersection of preserved genes and the benchmark sets using RIAA, LS, and DLS algorithms. (a), (b), and (c) reveal the analysis results when dataset alpha was applied. (d), (e), and (f) reveal the analysis results when dataset alpha 38 was applied.

using the bio-like sampling strategy, which applies 16 time points in (0,8] and 8 more time points in (8,16]. Gaussian noise with parameters $\mu = 0$ and $\sigma = 0.5$ is assumed during microarray experiments. The resulting time-course expression levels (dots), at a total of 24 time points and the sampling time information were treated as inputs to the RIAA algorithm. Figure 2(b) demonstrates the result of periodogram estimation. In this example, the grid size $\Delta\omega$ was chosen to be 0.065 and a total of 11 amplitudes corresponding to different frequencies were obtained and shown in the spectrum. Using Fisher's test, the peak at the third grid (frequency = 0.195) was found to be significantly large (p-value = 2.4×10^{-3}), and hence a periodic gene was claimed.

ROC curves strongly illustrate the performance of RIAA. In Figures 3 and 4, subplots (a)-(b), (c)-(d), (e)-(f), and (g)-(h) refer to the simulations with regular, bio-like, binomially random, and exponentially random sampling strategies, respectively. Additionally, in the left-hand side subplots (a), (c), (e), and (g), nonperiodic signals were simply Gaussian noise with parameters $\mu = 0$ and $\sigma = 0.5$, while in the right-hand side subplots (b), (d), (f), and (h), nonperiodic signals involve not only the Gaussian noise but also a transcriptional burst and a sudden drop (27). Periodic signals were generated using (25) with amplitude $M = 1$, $c = 2$, and $n = 24$. The only difference in simulation settings between Figures 3 and 4 is the frequency of periodic signals; they are $\omega_s = 0.4\pi$ and 0.1π, respectively. As shown in these figures, LS and DLS can perform well as RIAA when the time-course data are regularly sampled, or mildly irregularly sampled; however, when data are highly irregularly sampled, RIAA outperforms the others. The superiority of RIAA over DLS is particularly clear when the signal frequency is small.

Figure 5 illustrates the results of the real data analysis when these three algorithms, namely, the RIAA, LS, and DLS, were applied. On the x-axis, the numbers indicate the thresholds η that we preserved and classified as periodicities among all yeast genes; on the y-axis, the numbers refer to the intersection of η preserved genes and the proposed periodic candidates listed in the benchmark sets. Figures 5(a)–5(c) demonstrate the results derived from dataset alpha when the 113-gene benchmark set, 352-gene benchmark

set, and 518-gene benchmark set were applied, respectively. Similarly, Figures 5(d)–5(f) demonstrate the results derived from dataset alpha 38. The RIAA does not result in significant differences in the numbers of intersections when compared to those corresponding to LS and DLS in most of these cases. However, RIAA shows slightly better coverage when the dataset alpha 38 and the 113-gene benchmark set was utilized (Figure 5(d)).

5. Conclusions

In this study, the rigorous simulations specifically designed to comfort with real experiments reveal that the RIAA can outperform the classical LS and modified DLS algorithms when the sampling time points are highly irregular, and when the number of cycles covered by sampling times is very limited. These characteristics, as also claimed in the original study by Stoica et al. [12], suggest that the RIAA can be generally applied to detect periodicities in time-course gene expression data with good potential to yield better results. A supplementary simulation further shows the superiority of RIAA over LS and DLS when multiple periodic signals are considered (see Supplementary Figure s1 available online at http://dx.doi.org/10.1155/2013/171530). From the simulations, we also learned that the addition of a transcriptional burst and a sudden drop to nonperiodic signals (the negatives) does not affect the power of RIAA in terms of periodicity detection. Moreover, the detrend function in DLS, designed to improve LS by removing the linearity in time-course data, may fail to provide improved accuracy and makes the algorithm unable to detect periodicities when transcription oscillates with a very low frequency.

The intersection of detected candidates and proposed periodic genes in the real data analysis (Figure 5) does not reveal much differences among RIAA, LS, and DLS. One possible reason is that the sampling time points conducted in the yeast experiment are not highly irregular (not many missing values are included), since, as demonstrated in Figures 3(a)–3(d), the RIAA just performs equally well as the LS and DLS algorithms when the time-course data are regularly or mildly irregularly sampled. Also, the very limited time points contained in the dataset may deviate the estimation of p-values [14] and thus hinder the RIAA from exhibiting its excellence. Besides, the number of true cell cycle genes included in the benchmark sets remains uncertain. We expect that the superiority of RIAA in real data analysis would be clearer in the future when more studies and more datasets become available.

Besides the comparison of these algorithms, it is interesting to note that the bio-like sampling strategy could lead to better detection of periodicities than the regular sampling strategy (as shown in Figures 3(c) and 3(d)). It might be beneficial to apply loose sampling time intervals at posterior periods to prolong the experimental time coverage when the number of time points is limited.

Acknowledgments

The authors would like to thank the members in the Genomic Signal Processing Laboratory, Texas A&M University, for the helpful discussions and valuable feedback. This work was supported by the National Science Foundation under Grant no. 0915444. The RIAA MATLAB code is available at http://gsp.tamu.edu/Publications/supplementary/agyepong12a/.

References

[1] W. Zhao, K. Agyepong, E. Serpedin, and E. R. Dougherty, "Detecting periodic genes from irregularly sampled gene expressions: a comparison study," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2008, Article ID 769293, 2008.

[2] P. T. Spellman, G. Sherlock, M. Q. Zhang et al., "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization," *Molecular Biology of the Cell*, vol. 9, no. 12, pp. 3273–3297, 1998.

[3] G. Rustici, J. Mata, K. Kivinen et al., "Periodic gene expression program of the fission yeast cell cycle," *Nature Genetics*, vol. 36, no. 8, pp. 809–817, 2004.

[4] M. Menges, L. Hennig, W. Gruissem, and J. A. H. Murray, "Cell cycle-regulated gene expression in Arabidopsis," *Journal of Biological Chemistry*, vol. 277, no. 44, pp. 41987–42002, 2002.

[5] M. Ahdesmäki, H. Lähdesmäki, R. Pearson, H. Huttunen, and O. Yli-Harja, "Robust detection of periodic time series measured from biological systems," *BMC Bioinformatics*, vol. 6, article 117, 2005.

[6] M. Ahdesmäki, H. Lähdesmäki, A. Gracey et al., "Robust regression for periodicity detection in non-uniformly sampled time-course gene expression data," *BMC Bioinformatics*, vol. 8, article 233, 2007.

[7] E. F. Glynn, J. Chen, and A. R. Mushegian, "Detecting periodic patterns in unevenly spaced gene expression time series using Lomb-Scargle periodograms," *Bioinformatics*, vol. 22, no. 3, pp. 310–316, 2006.

[8] R. Yang, C. Zhang, and Z. Su, "LSPR: an integrated periodicity detection algorithm for unevenly sampled temporal microarray data," *Bioinformatics*, vol. 27, no. 7, pp. 1023–1025, 2011.

[9] E. R. Dougherty, "Small sample issues for microarray-based classification," *Comparative and Functional Genomics*, vol. 2, no. 1, pp. 28–34, 2001.

[10] Y. Tu, G. Stolovitzky, and U. Klein, "Quantitative noise analysis for gene expression microarray experiments," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 22, pp. 14031–14036, 2002.

[11] Z. Bar-Joseph, "Analyzing time series gene expression data," *Bioinformatics*, vol. 20, no. 16, pp. 2493–2503, 2004.

[12] P. Stoica, J. Li, and H. He, "Spectral analysis of nonuniformly sampled data: a new approach versus the periodogram," *IEEE Transactions on Signal Processing*, vol. 57, no. 3, pp. 843–858, 2009.

[13] J. Fan and Q. Yao, *Nonlinear Time Series: Nonparametric and Parametric Methods*, Springer, New York, NY, USA, 2003.

[14] 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, no. 4, pp. 549–556, 2009.

[15] V. Berger, "Pros and cons of permutation tests in clinical trials," *Statistics in Medicine*, vol. 19, no. 10, pp. 1319–1328, 2000.

[16] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[17] J. R. Chubb, T. Trcek, S. M. Shenoy, and R. H. Singer, "Transcriptional pulsing of a developmental gene," *Current Biology*, vol. 16, no. 10, pp. 1018–1025, 2006.

[18] T. Pramila, W. Wu, W. Noble, and L. Breeden, "Periodic genes of the yeast Saccharomyces cerevisiae: a combined analysis of five cell cycle data sets," 2007.

[19] U. Lichtenberg, L. J. Jensen, A. Fausbøll, T. S. Jensen, P. Bork, and S. Brunak, "Comparison of computational methods for the identification of cell cycle-regulated genes," *Bioinformatics*, vol. 21, no. 7, pp. 1164–1171, 2005.

[20] 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, article 137, 2007.

[21] D. Johansson, P. Lindgren, and A. Berglund, "A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription," *Bioinformatics*, vol. 19, no. 4, pp. 467–473, 2003.

[22] I. Simon, J. Barnett, N. Hannett et al., "Serial regulation of transcriptional regulators in the yeast cell cycle," *Cell*, vol. 106, no. 6, pp. 697–708, 2001.

[23] T. I. Lee, N. J. Rinaldi, F. Robert et al., "Transcriptional regulatory networks in *Saccharomyces cerevisiae*," *Science*, vol. 298, no. 5594, pp. 799–804, 2002.

[24] H. W. Mewes, D. Frishman, U. Güldener et al., "MIPS: a database for genomes and protein sequences," *Nucleic Acids Research*, vol. 30, no. 1, pp. 31–34, 2002.