COMPARATIVE PERFORMANCE ANALYSIS OF THE ALGORITHMS FOR DETECTING PERIODICALLY EXPRESSED GENES

A Thesis

by

KWADWO SEFA AGYEPONG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

August 2012

Major Subject: Electrical Engineering

Comparative Performance Analysis of the Algorithms

for Detecting Periodically Expressed Genes

COMPARATIVE PERFORMANCE ANALYSIS OF THE ALGORITHMS FOR DETECTING PERIODICALLY EXPRESSED GENES

A Thesis

by

KWADWO SEFA AGYEPONG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Approved by:

Co-Chairs of Committee,	Erchin Serpedin
	Edward R. Dougherty
Committee Members,	Ulisses Braga-Neto
	Samiran Sinha
Head of Department,	Costas Georghiades

August 2012

Major Subject: Electrical Engineering

ABSTRACT

Comparative Performance Analysis of the Algorithms
for Detecting Periodically Expressed Genes. (August 2012)

Kwadwo Sefa Agyepong, B.S., Prairie View A&M University

Co–Chairs of Advisory Committee: Erchin Serpedin
Edward R. Dougherty

Thus far, a plethora of analysis on genome-wide gene expression microarray experiments on the cell cycle have been reported. Time series data from these experiments include gene expression profiles that might be periodically expressed. However, the numbers and actual genes that are periodically expressed have not been reported with consistency, analysis on similar experiments reports disparate numbers of genes that are periodically expressed with scant overlap. This work ultimately compares the performance of five spectral estimation schemes in their ability to recover periodically expressed genes profiles. Lomb-Scargle (LS), Capon, Missing-Data Amplitude and Phase Estimation (MAPES), Real Value Iterative Adaptive Approach (RIAA) and Lomb-Scargle Periodogram Regression (LSPR) are rigourously studied and pitted against each other in various simulated testing conditions. Results obtained using synthetic and microarray data reveals that RIAA is an efficient and robust method for the detection of periodically expressed genes in short time series data that might be characterized with noisy and irregularly sampled data points.

To Leona

ACKNOWLEDGMENTS

I would like to thank my committee co-chairs, Dr. Serpedin and Dr. Dougherty, and my committee members, Dr. Bragga-Neto, and Dr. Sinha, for their support throughout the course of my studies in Texas A&M University and for introducing me to the rigors of research work. I would also like to thank the department faculty and staff, especially Ms. Tammy Carda, for all the help and guidance I received throughout my studies. Thanks to my friends and colleagues in the GSP Lab who helped in creating a conducive atmosphere for research and learning. Lastly, I would also like to thank my parents for their encouragement and understanding and to my wife for her unflinching support.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION[1]

Biological processes undergo rhythms that are dictated by various cell activities such as the cell cycle process. This phenomenon recur at regular intervals and may be annual, seasonal, circadian or even ultradian. These rhythms are controlled by endogenous biological clocks and understanding their molecular basis is of fundamental interest in biology. Knowledge of these rhythms leads to insights into diagnosis and treatment of illness. The rhythmic signals help a living organism to organize its behavior and physiology. Understanding these rhythmic activities has been an important problem in systems biology for many years. Advances in microarray technology equipped us with a means to directly measure and quantify the expressional concentration levels of mRNA, the basic unit structure that encodes chemical instructions for a protein product. These measurements provide a tangible means to characterize regulations in the cell.

Microarray experiments exploit high-throughput gene chips to measure gene expressions at various sampling time points per the suitability of the experimenter and experimental constraints. Experimental constraints [1] lead to scarce sample size, as large sample sizes are not economically feasible due to the cost of gene chips and the maintenance of a conducive ambience for cell cultures over time. The limited data set generally present missing values at random time points. This is due to defective slides

The journal model is *IEEE Transactions on Automatic Control.*

[1] Part of this chapter is reprinted with permission from "Detecting periodic genes from irregularly sampled gene expressions: a comparison study" by Zhao, W. and Agyepong, K. and Serpedin, E. and Dougherty, E.R., vol.2008, *EURASIP Journal on Bioinformatics and Systems Biology*, 2008, Copyright 2008 by Zhao and Agyepong and Serpedin and Dougherty.

and the inability of microarrays to cipher non-ideal spots. Experimental noise also corrupts the limited samples, leading to uncertainty that must be addressed within a stochastic framework [1].

The mechanisms of the underlying process is well understood, but the analyses of the datasets led to inconclusive reports on the numbers of periodically expressed genes for many organisms. Work on *Saccharomyces cerevisiae* [2] [3] has so far reported about 400 to 800 genes that are cell cycle regulated, meaning that they are periodically expressed. For *Schizosaccharomyces pombe*, about 400 to 700 genes [4][5] have been found to be periodically expressed. In *Aradidiopsis thaliana*, about 500 to 600 are reported to be cell cycle regulated [6]. The need for an analysis tool that overcomes the innate undesirable characteristics of the microarray data is evident. On experiments that are available to the general public, it is interesting that one cannot get an overlap of more than 400 genes between two different analyses based on similar experimental synchronization designs. There have been many microarray experiments conducted on the budding yeast [2] [4]. The budding yeast in [2] is the most used data source in many analytical experiments for the detection of periodically expressed genes. This is because of its grounding breaking results and the relatively large sample size it provided to literature citeKwadwo08. Spellman [2] analyzed the data on the budding yeast via a scoring criterion where a combination of a correlation score and a Fourier based score were used to rank 800 genes believed to be periodically expressed.

There are basically two main approaches used in the literature to evaluate schemes and models. The norm is to search for hits from a set of 104 genes that are known to be cell cycle regulated [7]. These 104 genes were found from traditional methods where expression profile were visually inspected [2]. The other way of putting a measure of performance on a scheme or statistical test is to combine the results of similar works, by taking a heuristic threshold overlap of results publicly available and counting the overlap of genes between the results of ones model and the overlap of results from other methods.

Our earlier work looked at three spectral analysis tools which could overcome the undesirable characteristics of the microarray experimental data set. The performance of Lomb-Scargle periodogram (LS) [8] , Capon (Robust Capon) [9] and Missing data Amplitude and Phase Estimation (MAPES) [10] were compared. Each scheme possesses the ability to detect periodically expressed genes from the expression measurements of mRNA provided that some conditions are met. Lomb-Scargle proved to be the most efficient method when all three schemes were applied on cdc15 dataset from Spellman's experiments [1]. The previous three schemes are included in the present comprehensive study for detailed analysis on a myriad simulated conditions that are semblant to microarray dataset. Stoica's [11] new method called Real Value Iterative Adaptive Approach (RIAA) and a scheme employed by Yang [12] called LSPR have been added to this study. LSPR is a new periodicity detection algorithm that has its foundation built on Lomb-Scargle periodogram and harmonic regression. There have been many methods proposed to detect periodicity in the cell cycle of organisms. Yang [13] used an algorithm which combined time domain and frequency domain analysis to obtain and identify rhythmic expression profiles. It utilizes spectral estimation technique to obtain periodically expressed profile candidates and model these candidates with a time-series model. Giurcaneanu [14] used generalized Gaussian distributions to investigate stochastic complexity inherent in the detection mechanism of genes that are periodically expressed. Ahdesmaki [15] employed a robust periodicity testing procedure that used a non-Gaussian noise assumption and considered a regression method to aide in simulating irregular sampling. Luan [16] used a selection of 'guide' genes and constructed cubic B-spline based periodic functions as a model [1]. The statistical approach by Luan[16] allowed for the identification of thresholds for false discovery rate. Lu [17] proposed a Bayesian approach to estimate a periodic-normal mixture model from five different experiments. Several additional power spectral density estimation schemes have been used in the literature. Wichert [18] applied the traditional periodogram where any missing data present for all genes were imputed via interpolation. Bowles used synthetic data to compare Capon method and Robust Capon approach[19]. Lichtenberg [20] compared [2], [16] and [17] using a a score obtained via the combination of periodicity and regulation magnitude. Most of the works cited above employed their methods on evenly sampled data. Missing data points were interpolated and in cases where the missing data set were more than 30%, the genes were discarded [1].

Microarray experiments are generally characterized by having datasets that are irregularly sampled. To address the issue of unequally spaced measurements, Lomb [21] and Scargle [22] discovered that a phase shift restores the orthogonality lost by Fourier analysis, due to the unevenness of the data, in the sine and cosine terms. Glynn [8] used the Lomb-Scargle scheme to analyze *Plasmodium falciparum* data set. Stoica [23] modified the Capon method to adapt to irregular sampled data in the field of signal processing. Wang et al. [10] proposed a new approach called missing-data amplitude and phase estimation (MAPES). MAPES estimates any missing data and computes the spectral density estimate iteratively via the Expectation Maximization (EM) algorithm. Real Value Iterative Adaptive Approach (RIAA) [11] induced the present interest to revisit our prior work given the fact that preliminary results show that it presents much promise in being robust to deficiencies in microarray data set. The rest of this work will illustrate the capability of each method while providing a complete review of the work in [1]. The following nested questions are posed and answered in this study: Which scheme performs best in the presence of (1) Noise, (2) Small sample size, (3) Clusters of missing data or irregular sampling? Both synthetic and experimental data are used in this work. The aim of this work is to nominate a scheme that will address the problem of scant overlap in in the existing results assessing periodically expressed genes in the same organism. Results shows that RIAA outperforms the schemes considered in this work on both synthetic and the Cdc 15 yeast data in Spellman's dataset. RIAA is also applied to two different data set, Spellman [2] and Pramila [24] alpha synchronized datasets, to obtain a consistent overlap of results for periodically expressed genes. Full results are provided in the Appendices including Matlab codes, the list of 104 plus 9 new genes provided by Johansson [7] are also included.

CHAPTER II

METHODS[1]

This section begins by examining RIAA and proceeds with a recapitulation of the existing methods for a proper perspective of the subject. The material of this chapter relies on our previous paper [1]. RIAA belongs to the class of power spectral density estimators that employ least-squares to estimate the spectral density for a sequential data with discrete spectra. Lomb [21] used phase-shift of the sine and cosine functions to restore orthogonality that is lost, due to unevenly sampling, between the cosine and sine harmonics. Scargle [22] extensively reanalyzed Lomb's periodogram to provide derivation of a null hypothesis distribution for the periodogram. The Lomb-Scarlge periodogram has been cited numerous times in many fields and applications including genomics see e.g.,[8], [12].. Capon approach represents a filter bank approach for power spectrum density estimation, where a finite-length data spectrum estimator is constructed by estimating the spectral power's distribution over narrow spectral bands. MAPES was developed for regular sampling times with missing data but as mentioned in [10], it belongs to the family of non-parametric spectral estimation techniques. It exploits the expectation maximization (EM) algorithm to estimate missing samples. LSPR is based on Lomb-Scargle periodogram, where inferences made from LS are used as inputs into a harmonic regression model whose output acts as inputs in Akaike's information criterion [25] to obtain a $p-$value.

[1] Part of this chapter is reprinted with permission from "Detecting periodic genes from irregularly sampled gene expressions: a comparison study" by Zhao, W. and Agyepong, K. and Serpedin, E. and Dougherty, E.R., vol.2008, *EURASIP Journal on Bioinformatics and Systems Biology*, 2008, Copyright 2008 by Zhao and Agyepong and Serpedin and Dougherty.

A. Real Value Iterative Adaptive Approach - RIAA

Real Value Iterative Adaptive Approach (RIAA) is a spectral estimator (periodogram), designed to alleviate undesirable characteristics that arise in the spectral density estimation of non-uniformly sampled data. This method can be thought of as an iterative weighted least-squares method which utilizes an adaptive weighting matrix obtained from the most recent spectral density estimate [11]. Let $(t_l, y_l), l = 0, \ldots, N-1$, denote N time-series observations where t_l are the observational times or time lag and y_l is the expression measurement of a gene or time series. RIAA is formulated within the framework of least-squares periodogram and so to explain RIAA, it is prudent to expound on the ordinary least-squares periodogram. The Fourier transform periodogram of the data set will normally be expressed as:

$$\Phi_{FT}(\omega) = \frac{1}{N^2} \left| \sum_{l=0}^{N-1} y_l e^{-j\omega t_l} \right|^2, \tag{2.1}$$

where ω is the angular frequency variable. An equivalent expression for $\Phi_{FT}(\omega)$ can be obtained via least-squares theory [26] as,

$$\begin{aligned} \Phi_{FT}(\omega) &= |\hat{\alpha}(\omega)|^2, \\ \hat{\alpha}(\omega) &= \arg\min_{\alpha(\omega)} \sum_{l=0}^{N-1} |y_l - \alpha(\omega) e^{j\omega t_l}|^2. \end{aligned} \tag{2.2}$$

Employing real value signals, Equation(2.2) can be re-written as:

$$\min_{\substack{\Theta \geq 0 \\ \phi \in [0, 2\pi]}} \sum_{l=0}^{N-1} [y_l - \Theta \cos(\omega t_l + \phi)]^2, \tag{2.3}$$

where Θ and ϕ depend on ω. Set $a = \Theta \cos(\phi)$ and $b = -\Theta \sin(\phi)$ to obtain:

$$\min_{a,b} \sum_{l=0}^{N-1} [y_l - a\cos(\omega t_l) - b\sin(\omega t_l)]^2. \tag{2.4}$$

The solution to Equation (2.4) is given by:

$$\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = \mathbf{R^{-1}r}. \tag{2.5}$$

Where,

$$\mathbf{R} = \sum_{l=0}^{N-1} \begin{bmatrix} \cos(\omega t_l)^2 & \cos(\omega t_l)\sin(\omega t_l) \\ \sin(\omega t_l)\cos(\omega t_l) & \sin(\omega t_l)^2 \end{bmatrix}. \tag{2.6}$$

and

$$\mathbf{r} = \sum_{l=0}^{N-1} \begin{bmatrix} \cos(\omega t_l) \\ \sin(\omega t_l) \end{bmatrix} y_l. \tag{2.7}$$

The ordinary least squares periodogram can then be defined as:

$$\Phi_{LSP}(\omega) = \frac{1}{N}\mathbf{r^T R^{-1} r}. \tag{2.8}$$

1. Frequency Window and Grid Size

A spectral window that can resolve spectral tendencies without aliasing from the sampling times is presented in this section. Proceeding with the premise that other sinusoidal components are present in the data, an error term is introduced into the spectral density estimate and taking the spectral norm of this error term as in [11], a solution is obtained that depends on the sampling pattern. From this solution, the spectral window can be derived as a function of ω. Stoica [11] approximated this window as $W(\omega) = |\sum_{l=0}^{N-1} e^{j\omega t_l}|^2$. This window is used to find the smallest frequency ω_o for which the spectral window function is at its next maximum, different from the global maximum obtained at $\omega = 0$. If there are no frequencies that have a maximum nearest the peak of N^2, set $\omega_o = \infty$ or a value representative for the data under study. Using ω_o, the maximum frequency is defined as

$$\omega_{max} = \frac{\omega_o}{2} \tag{2.9}$$

which provides the interval $[0, \omega_{max}]$. In this window, care must be taken to ensure that the smallest frequency separation can be adequately detected in choosing a frequency search grid $\Delta\omega$. There are many grid size approximations used in the literature [11][27]. However, Equation(2.10) can be used since it is a widely used approximation for irregular sampling:

$$\Delta\omega = \frac{2\pi}{t_{N-1} - t_0}. \tag{2.10}$$

The number of grid points is then given by:

$$J = \frac{\lfloor \omega_{max} \rfloor}{\Delta\omega}. \tag{2.11}$$

And this leads to a uniform frequency grid as in [28] given by

$$\omega_j = \Delta\omega_j, \quad j = 1, \ldots, J. \tag{2.12}$$

Thus far, the ordinary least-squares spectral estimation method has been described. To continue formulating RIAA, there is a need to introduce the following parameters,

$$\mathbf{y} = \begin{bmatrix} y_0 \\ \vdots \\ y_{N-1} \end{bmatrix}, \qquad \mathbf{A_j} = \begin{bmatrix} \mathbf{c_j} & \mathbf{s_j} \end{bmatrix}, \; \mathbf{\Theta_j} = \begin{bmatrix} a(\omega_j) \\ b(\omega_j) \end{bmatrix},$$

$$\mathbf{c_j} = \begin{bmatrix} \cos(\omega_j t_0) \\ \vdots \\ \cos(\omega_j t_{N-1}) \end{bmatrix}, \; \mathbf{s_j} = \begin{bmatrix} \sin(\omega_j t_0) \\ \vdots \\ \sin(\omega_j t_{N-1}) \end{bmatrix}. \tag{2.13}$$

Re-parametrization of Equation (2.2) presents the following solution,

$$\min_{\mathbf{\Theta_j}} \|\mathbf{y} - \mathbf{A_j}\mathbf{\Theta_j}\|^{\mathbf{2}}$$

$$\hat{\Theta}_j = (\mathbf{A}_j^T\mathbf{A}_j)^{-1}\mathbf{A}_j^T\mathbf{y}. \tag{2.14}$$

The covariance matrix of other possible components in the data other than the component with ω_j is defined:

$$\mathbf{Q}_j = \sum_{m=1, m\neq j}^{J} (a_m^2 + b_m^2)\mathbf{A}_m\mathbf{A}_m^T. \tag{2.15}$$

At ω_j, all other frequency components are considered to be noise and Equation(2.15) carries their contribution. Using Eq.(2.15) if available, the following weighted least squares approach is employed because it is known to be more accurate under general conditions than the ordinary least squares [29].

$$\min_{\alpha_j} \|\mathbf{y} - \mathbf{A}_j\alpha_j\|^2{}_{\mathbf{Q}_j^{-1}} \tag{2.16}$$

The solution to the problem above is given as:

$$\hat{\Theta}_j = \frac{\mathbf{A}_j^T\mathbf{Q}_j^{-1}\mathbf{y}}{\mathbf{A}_j^T\mathbf{Q}_j^{-1}\mathbf{A}_j}. \tag{2.17}$$

Then RIAA also known as the weighted least square periodogram (WLSP) is defined as:

$$\Phi_{WLSP}(\omega_j) = \frac{1}{N}\hat{\Theta}_j^T(\mathbf{A}_j^T\mathbf{A}_j)\hat{\Theta}_j.$$

$$\Phi_{WLSP}(\omega_j) = |\alpha_j|^2. \tag{2.18}$$

Initialization Use the ordinary least squares to obtain the initial value of α_j^0.

Iteration At the kth iteration, the estimate of $\hat{\alpha}_j$ i.e., at ω_j is $\alpha_j^k = \frac{\mathbf{A}_j^T(\mathbf{Q}_j^k)^{-1}\mathbf{y}}{\mathbf{A}_j^T(\mathbf{Q}_j^k)^{-1}\mathbf{A}_j}$ for $k = 1, \ldots, K$ where $\mathbf{Q}^k{}_j = \sum_{m=1, m\neq j^J} |\alpha_j^{k-1}|^2 \mathbf{A}_m \mathbf{A}_m^T$.

End Iteration is terminated after 15 iterations or when $|\alpha_j^{k+1} - \alpha_j^k|^2 < 10^{-4}$.

After the last iterative step, $\{\hat{\Theta}_j^K\}$ is used to compute the power spectral density for RIAA:

$$\Phi_{RIAA}(\omega_j) = \frac{1}{N}(\hat{\boldsymbol{\Theta}}_j^K)^T(\mathbf{A}_j^T\mathbf{A}_j)(\hat{\Theta}_j^K), \quad j = 0, \ldots, J. \tag{2.19}$$

RIAA does not suffer from the global and local leakage that are characteristic for the other methods . Therefore, peaks detected by RIAA have a high probability of denoting cyclicity and simulation results show that RIAA does not suffer from the spurious peaks problem of LS, which leads to false positives.

B. Lomb-Scargle Periodogram

In cases where evenly sampled data cannot be obtained, Lomb-Scargle periodogram has been the method of choice when estimating spectral components in the data. Lomb-Scargle periodogram ignores the unevenness of the data by imputing a phase-shift to the sine and cosine harmonic functions. This restores the orthogonality which, otherwise, is lost due to the nature of the data. Given N time-series observations $(t_l, y_l), l = 0, \ldots, N-1$, where t stands for the time tag and y stands for the value of a time series point or sampled expression of a specific gene, the normalized Lomb-

Scargle periodogram at an angular frequency ω is defined as in [1]

$$\Phi_{LS}(\omega_j)=\frac{1}{2\hat{\sigma}^2}\left(\frac{\left(\sum_{l=0}^{N-1}[y_l-\bar{y}]cos[\omega_j(t_l-\tau)]\right)^2}{\sum_{l=0}^{N-1}cos^2[\omega_j(t_l-\tau)]}+\frac{\left(\sum_{l=0}^{N-1}[y_l-\bar{y}]sin[\omega_j(t_l-\tau)]\right)^2}{\sum_{l=0}^{N-1}sin^2[\omega_j(t_l-\tau)]}\right), \tag{2.20}$$

for $j=1,\dots,J$ as defined in Equation(2.12) where $\bar{y}$ and $\hat{\sigma}^2$ stand for the mean and variance of the sampled data, respectively, and τ is defined as:

$$\tau=\frac{1}{2\omega_j}atan\left(\frac{\sum_{l=0}^{N-1}sin(2\omega_j t_l)}{\sum_{l=0}^{N-1}cos(2\omega_j t_l)}\right). \tag{2.21}$$

The frequency grid defined under RIAA is also applied to the Lomb-Scargle periodogram. Lomb-Scargle periodogram is an efficient solution in estimating the spectra of unevenly sampled data sets especially when the underlying noise assumption is Gaussian.

C. Robust Capon Method

The general framework for the Capon method is reproduced from our earlier work [1] Capon method is a filter-bank approach that is based on a data-dependent bandpass filter [9]. It was originally designed for evenly sampled data. It estimates the spectral density of a time series input signal by first passing it through a bank of bandpass filters with varying center frequencies, called the steering frequencies. It then measures and uses the output power of the filter's passband. By dividing the measured power by the passband bandwidth, an estimate of the power spectrum density is obtained. The filter is designed in such a way that it minimizes all the contribution of other frequencies in the input signal except the frequency components at ω. In other words, the Capon method seeks to solve the following optimization problem:

$$\mathbf{h} = \arg\min_{\mathbf{h}} \mathbf{h}^H \mathbf{R} \mathbf{h} \quad \text{subject to} \quad \mathbf{h}^H \mathbf{a}(\omega_j) = 1, \tag{2.22}$$

whose solution provides the spectrum estimate at frequency ω_j as

$$\Phi_C(\omega_j) = \frac{1}{\mathbf{a}^H(\omega_j \Delta)\mathbf{R}^{-1}\mathbf{a}(\omega_j \Delta)}, \tag{2.23}$$

where matrix $\mathbf{R}$ stands for the data covariance matrix with a dimension N_0, which is inversely proportional to the bandwidth of the Capon filter. The steering vector is defined as follows

$$\mathbf{a}(\omega_j) = \left(1 \; e^{j\omega_j} \cdots e^{j\omega_j(N_0-1)}\right)^T. \tag{2.24}$$

To guarantee the existence of inverse $\mathbf{R}^{-1}$, the bandwidth parameter N_0 need not exceed $\lfloor (N-1)/2 \rfloor$. However, a smaller N_0, will adversely affect the resolution of the spectral estimates while the accuracy of the estimate of the covariance matrix will increase. Hence, N_0 should be set as a tradeoff between resolution and accuracy of the Capon method [23].

It has been proven that given an adequate number of samples, the Capon method yields a better spectral resolution compared with traditional periodogram [9]. The Capon method has been updated to cope with the presence of irregular samples [23]. The same frequency grid denoted in Equation (2.12) is employed. In order to take advantage of the best resolution, N_0 is set to be equal to $\lfloor (J-1)/2 \rfloor$, where J is defined in Equation (2.12). In simulation, an estimate of the autocorrelation matrix $\hat{\mathbf{R}}$ can is obtained from the Lomb-Scargle periodogram, which is represented by

$$\hat{\mathbf{R}} = \frac{1}{J\delta} \sum_{j=1}^{J} \mathbf{a}(\omega_j \delta)\mathbf{a}^H(\omega_j \delta)\Phi_{LS}(\omega_j). \tag{2.25}$$

The Capon method is slightly more computationally complex than LS and RIAA.

In simulated data, its resolution was better than LS and could rival RIAA if the sample sizes is increased to be greater than 40 samples, but on limited sample size and corrupted biological data, its performance was below a notch compared to LS and RIAA.

D. MAPES Method

The general framework for MAPES is also reproduced from our earlier work [1]. Given P time-series observations $(t_l, y_l), l = 0, \ldots, P-1$, the data are assumed to be sampled uniformly. However, only N data points are available and there are $P - N$ missing data points. The time-series signal with frequency ω is modeled as

$$y_l = \alpha(\omega)e^{j\omega l} + \varepsilon_l(\omega), \quad l = 0, \ldots, P-1, \quad \omega \in [0, 2\pi], \tag{2.26}$$

where $\alpha(\omega)$ represents the complex amplitude of the sinusoidal component and $\varepsilon_l(w)$ denotes a residual term. The same frequency grid as in Equation (2.12) is used. Using the expectation-maximization algorithm, MAPES iteratively estimates the missing data, and while updating the estimates of the spectra by minimizing the mean square error between consecutive estimates.

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

$$\tilde{\mathbf{y}} = \begin{pmatrix} \tilde{\mathbf{y}}_0 \\ \vdots \\ \tilde{\mathbf{y}}_{L-1} \end{pmatrix} = \mathbf{U}\boldsymbol{\gamma} + \mathbf{V}\boldsymbol{\mu}, \tag{2.27}$$

where $\boldsymbol{\gamma}$ $(N \times 1)$ and $\boldsymbol{\mu}$ $((P - N) \times 1)$ represent the available and missing data, respectively. $\mathbf{U}$ $(LM \times N)$ and $\mathbf{V}$ $(LM \times (P - N))$ denote binary selection matrices for the available and missing data, respectively. The selection matrices are orthogonal to each other: $U_N^T V_{P-N} = \mathbf{0}_{Nx(P-N)}$. In other words, 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

$$\boldsymbol{\gamma} = (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\tilde{\mathbf{y}} = \tilde{\mathbf{U}}^\dagger\tilde{\mathbf{y}}, \quad \text{where} \quad \tilde{\mathbf{U}}^\dagger = (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T, \tag{2.28}$$

$$\boldsymbol{\mu} = (\mathbf{V}^T\mathbf{V})^{-1}\mathbf{V}^T\tilde{\mathbf{y}} = \tilde{\mathbf{V}}^\dagger\tilde{\mathbf{y}}, \quad \text{where} \quad \tilde{\mathbf{V}}^\dagger = (\mathbf{V}^T\mathbf{V})^{-1}\mathbf{V}^T. \tag{2.29}$$

The residual vector and its covariance matrix are next defined

$$\mathbf{e}_l(\omega) = \left(\varepsilon_l(\omega)\ \varepsilon_{l+1}(\omega)\cdots\varepsilon_{l+M-1}(\omega)\right)^T, \tag{2.30}$$

$$\mathbf{Q}(\omega) = E\left(\mathbf{e}_l(\omega)\mathbf{e}_l^H(\omega)\right), \tag{2.31}$$

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

$$\boldsymbol{\rho}(\omega) = \begin{pmatrix} e^{j\omega 0}\mathbf{a}(\omega) \\ \vdots \\ e^{j\omega(L-1)}\mathbf{a}(\omega) \end{pmatrix}, \tag{2.32}$$

$$\mathbf{D}(\omega) = \begin{pmatrix} \mathbf{Q}(\omega) & & 0 \\ & \ddots & \\ 0 & & \mathbf{Q}(\omega) \end{pmatrix}. \tag{2.33}$$

Where $\mathbf{a}(\omega)$ represents the complex amplitude of the sinusoidal component and $\mathbf{Q}(\omega)$ is defined as in Equation(2.31). 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 Gaussian with mean and variance denoted by

$(\mathbf{b}, \mathbf{K})$ as follows

$$\begin{aligned}
\mathbf{b}_i(\omega) &= \tilde{\mathbf{U}}^T \boldsymbol{\rho}(\omega)\alpha_i(\omega) + \tilde{\mathbf{U}}^T \mathbf{D}_i(\omega)\tilde{\mathbf{V}}\Big(\tilde{\mathbf{V}}^T \mathbf{D}_i(\omega)\tilde{\mathbf{V}}\Big)^{-1}\Big(\boldsymbol{\gamma} - \tilde{\mathbf{V}}^T \boldsymbol{\rho}(w)\alpha_i(w)\Big), && (2.34)\\
\mathbf{K}_i(\omega) &= \tilde{\mathbf{U}}^T \mathbf{D}_i(\omega)\tilde{\mathbf{U}} - \tilde{\mathbf{U}}^T \mathbf{D}_i(\omega)\tilde{\mathbf{V}}\left(\tilde{\mathbf{V}}^T \mathbf{D}_i(\omega)\tilde{\mathbf{V}}\right)^{-1}\tilde{\mathbf{V}}^T \mathbf{D}_i(\omega)\tilde{\mathbf{U}}. && (2.35)
\end{aligned}$$

Where $\tilde{\mathbf{U}}$ and $\tilde{\mathbf{V}}$ are estimates of the selection matrices at the ith EM iteration and $\mathbf{D}_i(\omega)$ is the estimate of $\mathbf{D}(\omega)$,Equation(2.23) at the ith EM iteration. Then the estimates for spectral magnitude $\alpha(\omega)$ and residual matrix $\mathbf{Q}$ are updated in terms of equations

$$\begin{aligned}
\alpha_{i+1}(\omega) &= \frac{\mathbf{a}^H(\omega)\mathbf{S}^{-1}(\omega)\mathbf{Z}(\omega)}{\mathbf{a}^H(\omega)\mathbf{S}^{-1}(\omega)\mathbf{a}(\omega)}, && (2.36)\\
\mathbf{Q}_{i+1}(\omega) &= \mathbf{S}(\omega) + \left(\alpha_{i+1}(\omega)\mathbf{a}(\omega) - \mathbf{Z}(\omega)\right)\left(\alpha_{i+1}(\omega)\mathbf{a}(\omega) - \mathbf{Z}(\omega)\right)^H, && (2.37)
\end{aligned}$$

where the auxiliary matrices are defined as follows

$$\begin{aligned}
\begin{pmatrix} \mathbf{z}_0 \\ \vdots \\ \mathbf{z}_{L-1} \end{pmatrix} &= \mathbf{U}\boldsymbol{\gamma} + \mathbf{V}\mathbf{b}(\omega), && (2.38)\\
\mathbf{Z}(\omega) &= \frac{1}{L}\sum_{l=0}^{L-1} \mathbf{z}_l e^{-j\omega l}, && (2.39)\\
\mathbf{S}(\omega) &= \frac{1}{L}\sum_{l=0}^{L-1} \boldsymbol{\Gamma}_l + \frac{1}{L}\sum_{l=0}^{L-1} \mathbf{z}_l \mathbf{z}_l^H - \mathbf{Z}(\omega)\mathbf{Z}^H(\omega). && (2.40)
\end{aligned}$$

In Equation(2.40), $\boldsymbol{\Gamma}_{\mathbf{0}}, \cdots, \boldsymbol{\Gamma}_{\mathbf{L-1}}$ are $M \times M$ sub-block matrices located on the main diagonal of matrix $\mathbf{UKU^T}$.

Finally, the MAPES power spectral density estimator is expressed as

$$\Phi_{MAPES}(\omega) = \frac{|\alpha(\omega)|^2}{J}. \tag{2.41}$$

E. LSPR Method

LSPR is not necessary an acronym, however, the LSP stands for Lomb-Scargle periodogram and R stand for regression. As mentioned in Chapter 1, its foundation is built on LS. It uses the output from LS as inputs to a harmonic regression. The algorithm is provided below as shown in [12].

LSPR algorithm

1. Detrend data and denote it as $\dot{y}$.
2. Smooth detrended data $\dot{y}$ with the fourth-order Savitzky-Golay algorithm and denote the resulting data as $\ddot{y}$.
3. Apply LS on both $\dot{y}$ and $\ddot{y}$ and select periods $\{\dot{T}_j\}$ and $\{\ddot{T}_j\}$.
4. Use $\{\dot{T}_j\}$ and $\{\ddot{T}_j\}$ as inputs into a harmonic regression for $\{\dot{y}\}$
5. Employ Akaike information criterion (AIC) to find the best harmonic regression model and $p-$value of $\{\dot{y}\}$ from Step 4.
6. Set FDR to be less than 0.05.

Harmonic regression is then used to fit the detrended data $\dot{y}$ with sinusoidal functions as:

$$\dot{y}_l = \mu + \sum_{j=1}^{J} \alpha_j \cos(\frac{2\pi}{T} t_l + \phi_j) + \varepsilon_l, t \tag{2.42}$$

where μ is the mean of $\{\dot{T}_j\}$, α_j are the amplitudes of the predictor trigonometric functions, ϕ_j are the phases of the peaks relative to the time zero, ε_l are uncorrelated noise, and T_j are the periods inferred from LS. The smoothing version of the detrended data produced worse results than the original dataset and hence simulations are limited to the detrended data but the poor performance of the smoothed data are also shown in some pertinent simulations. In Chapter 4, simulations of LSPR are only for $\dot{y}$, the detrended data. If it is necessary to compare the performance of the smoothed detrended data, it will be clearly stated. The advantage to this method only serves to reduce the number of false positives that Lomb-Scargle periodogram produces but does not improve on recovering misses that LS failed to observe. LSPR assumes that the trend in the data is known, and by removing it, a limitation of Lomb-Scargle is eliminated, but if the data contained outliers their effect will still be felt and LS will provide spurious peaks which will then propagate through the LSPR algorithm to come to a similar conclusion just like LS.

CHAPTER III

SIGNIFICANCE TESTING[1]

A. Periodicity Test

A time series data of length N is used as an input to each of the schemes to obtain power spectral density estimates. Based on peaks from the outputs of the schemes, the data is classified as cyclic or non-cyclic. The null hypothesis is taken to be that the measurements are originated from a Gaussian noise stochastic process [1]. There are a host of tests that can be employed to access the significance of peaks detected by the schemes. Akaike's information criterion (AIC) has been used by [30] to test for periodicity. Stoica employed the Bayesian information criterion (BIC) [9], Glynn used the Fisher test [8] to search for periodicity in *Plasmodium falciparum* microarray gene expression dataset. The likelihood ratio test has been used in [15], Fan [31] showed that χ^2 test can also be employed to determine the significance of the detected peaks. However, Stoica [11] implied that there was no satisfactory algorithm or approach for testing significance of detected peaks in the case of irregularly sampled, however one can use Fisher's test to determine the significance of peaks detected in a power spectral density estimator $\Phi(\omega)$ without any drop in performance when compared with other methods [1]. The Fisher's test statistic is defined as

$$T = \frac{\Phi(\omega_{k_{max}})}{N_0^{-1} \sum_{1 \leq k \leq N_0} \Phi(\omega_k)} \ , \tag{3.1}$$

[1] Part of this chapter is reprinted with permission from "Detecting periodic genes from irregularly sampled gene expressions: a comparison study" by Zhao, W. and Agyepong, K. and Serpedin, E. and Dougherty, E.R., vol.2008, *EURASIP Journal on Bioinformatics and Systems Biology*, 2008, Copyright 2008 by Zhao and Agyepong and Serpedin and Dougherty. Originally published by SpringerOpen

where $N_0 = \lfloor (J-1)/2 \rfloor$ for the defined symmetric frequency grid and the highest peak is $\Phi(\omega_{k_{max}})$.

Our synthetic data simulations also included testing for multiple peaks. This necessitated the use of Whittle's second peak detection formulation [32], since Fisher's test was only defined for the highest peak. Whittle's second peak detection statistics is defined as

$$T_2 = \frac{\Phi(\omega_{k_2})}{\sum_{k=1}^{N_0} \Phi(\omega_k) - \Phi(\omega_{k_{max}})}, \tag{3.2}$$

where $\Phi(\omega_{k_2})$ stands for the second highest peak. The p-value for detecting the largest peak is then given as [31]

$$P(T > t) = 1 - e^{-N_0 e^{-t}}. \tag{3.3}$$

The distribution for Fisher's test Equation(3.1) and Whittle's test Equation(3.2) is similar to that of Equation(3.3). The p-value measures the likelihood of obtaining such a peak if the series were generated by noise alone. Whereby a small p-value will give the indication that there is a small chance obtaining such a peak if the measurement were of noise alone. A p-value threshold serves as a threshold to decide if time series measurement contains any rhythms that are not due to chance. A rejection of the null hypothesis will imply that the magnitude of a frequency in the power spectral density is appreciably bigger than the mean and the time series data are samples from a periodic signal. For more details on the p-values, please see Fisher [33] or Brockwell [34].

Once the p-values are calculated for each time series or gene, they are ranked in ascending order and the threshold is employed to obtain significant results.

B. Multiple Testing Correction

For just one test, a fixed p-value is acceptable. For example, if the p-value is set to 0.05, the implication is that there is a 5% chance that the results obtained are not true positive. A 5% chance of false positives is high especially when considering over 6000 tests. To overcome the above problem, multiple testing approach must be used to control the results of the tests that were significant and not for all test. As proposed in [35] and [36], multiple testing correction is needed to control the false discovery rate (FDR). For each time series or measured gene, a p-value is calculated from the spectral density estimator or periodogram and used to test for periodicity. The p-values are ranked in an increasing order with the smallest ith p-value designated by $p_{(i)}$ [1]. For real biological data, the estimate for the number of non-cyclic genes among all n genes is taken to be $\widehat{n_0}$; it is acceptable to take $\widehat{n_0} = n$. The testing procedure make inference on the k genes with the lowest p-values, by using an adjusted p-value obtained from the FDR approach defined as

$$\widehat{FDR}_k = \frac{p_{(k)}\widehat{n_0}}{k}, \tag{3.4}$$

where $p_{(k)}\widehat{n_0}$ is an estimate of the number of false positives. Estimate of FDR, $\widehat{FDR}$, is not a monotonic function of k, the number inferred to be periodic. This makes it hard to choose a p-value threshold [1]. Storey [35] solved this problem by proposing an FDR adjusted p-value called q-value and is given by the following

$$q_k = \min_{k \leq j \leq n} \widehat{FDR}_j. \tag{3.5}$$

The q-value defined by Equation(3.5) is a monotonically increasing function of k. By specifying a q-value threshold as τ, the FDR can be controlled and through that the number of time series or genes to be inferred as periodic can then be derived as

$$k = \max_{1 \leq j \leq n} q_j \leq \tau. \tag{3.6}$$

CHAPTER IV

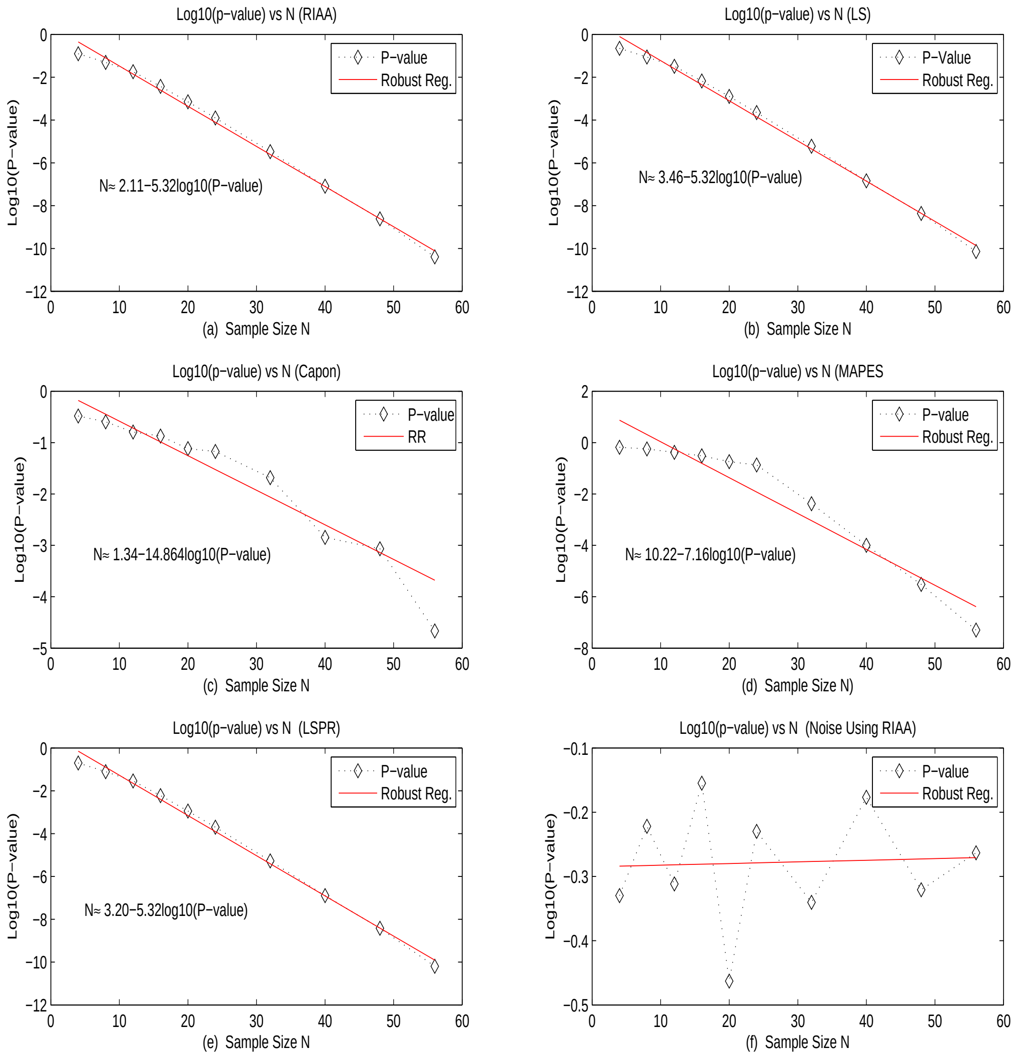
SIMULATION: ARTIFICIAL AND BIOLOGICAL DATA

A. Results

A natural query is the question of how to assess the performance of these schemes. The schemes are implemented to investigate the smallest number of samples that each requires to obtain significant results. A purely sinusoidal signal sampled irregularly with a Poisson sampling process was utilized. The schemes were then applied on artificial datasets, obtained from a periodic signal mixed with Gaussian noise and a non-periodic signal, to evaluate their ability to infer periodic signals in the presences of non-idealities. Performance was evaluated based on different p-value thresholds for a fixed sample size. The ability of the schemes under different signal to noise ratio (SNR) was also investigated for a fixed sample size and p-value. The computation time required by each scheme for different sample sizes was also analyzed. An ancillary aim classified the schemes under undesirable characteristics of the microarray dataset,i.e., missing values, sample size, and presence of noise. Finally, the best scheme is applied on two data sets to attempt bridging the gap of disparities in the reported results of periodically expressed genes for yeast, found in literature.

1. Simulation on Artificial Data

A purely sinusoidal signal was irregularly sampled to investigate the minimum number of samples each scheme needed to obtained significant results. For each N in Figure (1), the p-value was calculated as discussed Chapter 3 for each correctly inferred period in our signal, this technique is similar to that performed by Gylnn [8] for Lomb-Scargle periodogram. An approximation to the minimum number of samples that each



(a) Sample Size N

(b) Sample Size N

(c) Sample Size N

(d) Sample Size N)

(e) Sample Size N

(f) Sample Size N

Fig. 1: Determination of sample size.

scheme needs is illustrated in Figure (1) where Matlab's version of robust regression is used to obtain estimates of N . From Figures 1b and 1e, LS and LSPR needed approximately 12 samples and RIAA needed only 9 samples Figure (1a) to produce significant results for a p-value of 0.05. It can be seen from Figure (1c) that the Capon method needed the largest number of samples to obtained significant results. The reason is that the Capon method requires a tradeoff between resolution and statistical accuracy when it comes to the choice of the filter length. Our simulation revealed that choosing the filter length to be approximately equal to one half of the data length, a balance was established for both resolution and accuracy in the estimation of the covariance matrix for the Capon method. It is not surprising that LS and LSPR both needed the same number of samples, as much as LSPR attempts to obtain best fit models from its harmonic regression, as its core is based on LS. Table I shows the number of samples that each scheme needed to show significant results with $p-$value threshold set at 0.05 and 0.005, respectively. The choice of these p-values is explained later in the chapter.

Table I: Minimum number of samples needed based on p-value thresholds of 0.05 and 0.005

Method	$N_{0.05}$	$N_{0.005}$
RIAA	9	14
LS	12	19
MAPES	20	26
Capon	22	26
LSPR	12	19

RIAA required the smallest number of samples when the $p-$value was selected to be more stringent. However, Capon and MAPES approximately needed the same number (26) of samples to obtain significant results. It must again be highlighted that this only provides approximate values for the sample size needed.

2. Artificial Data Model

A modeled to generate artificial data set is given as follows:

$$y_l = \alpha \cos(\omega l + \phi) + \epsilon_l, \tag{4.1}$$

where $l = 0, \ldots, N-1$, $\phi \in (-\pi, \pi]$ and ϵ_l are i.i.d. noise sequence.

Two cases of non-idealities were considered: (1) Addition of Gaussian noise and (2) Addition of non-periodic data and Gaussian noise. Figure (2) shows a signal composed of non periodic pulses and Gaussian noise with zero mean and unit variance which was added to our data model. The pulses represents mRNA bursts that are characteristic for microarray data sets. An experiment similar to [15] was conducted where two thousand time series of length $N = 18, 48$, and 100 were generated. One hundred of the time series are generated from our data model in Equation (4.1) to be periodic and 1900 non-periodic. For each series, the p-value was evaluated and the testing methodology discussed in Chapter 3 was employed for FDR with $q-$values equal to 0.05, 0.01 and 0.005. The sampling was modeled as a Poisson process with parameter λ; this ensured that sampling was done on an average of every $\frac{1}{\lambda}$s. The Poisson process will inherently bring an irregular sampling format that will mimic microarray datasets characterized by uniform sampling, but with ample number of missing values.

Table II on page 27 shows the number of signals inferred to be periodic by each scheme when the number of samples time points N equal to 18 for q-values 0.05, 0.01

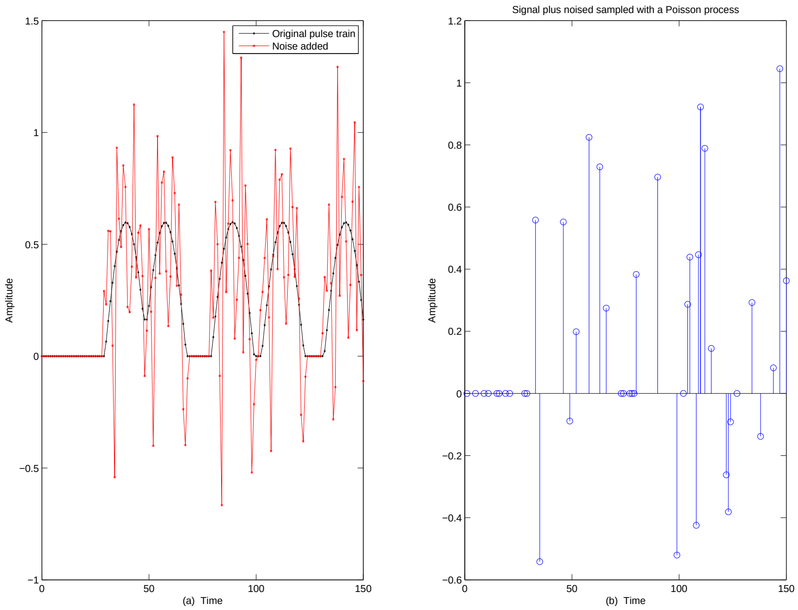


Fig. 2: A non-periodic signal of pulses.

and 0.005. With limited number of samples and SNR 5dB, RIAA was able to detect more periodic components in the data per q-value threshold than any other method. This is important because most microarray datasets have limited number of sample points and a scheme that can detect periodic components with limited resources is of premium. The number in parentheses are true positives. LSPR was employed for only the detrended data.

Table II: Inferred number of periodic time series: N=18

	$q-$value		
Method	0.05	0.01	0.005
RIAA	42(41)	29(29)	15(15)
LS	27(21)	11(9)	1(0)
MAPES	14(11)	10(6)	3(0)
Capon	9(9)	6(0)	1(0)
LSPR	23(21)	9(9)	0(0)

Table III: Inferred number of periodic time series: N=48

	$q-$value		
Method	0.05	0.01	0.005
RIAA	103(100)	76(72)	65(65)
LS	111(89)	68(59)	54(54)
MAPES	109(84)	72(64)	53(53)
Capon	105(86)	66(61)	54(53)
LSPR	111(89)	68(59)	54(54)

When the number of samples was increased to 48, Table III shows that RIAA still outperforms the other three schemes. When the number of samples was changed to 100 time points in Table IV, all the schemes were able to accurately preserve the periodic components in the dataset when the $q-$value was set as 0.05 and 0.01,respectively. However, the false positives in RIAA and Capon were less than all other schemes the $q-$value was set as 0.005.

Table IV: Inferred number of periodic time series: N=100

	$q-$value		
Method	0.05	0.01	0.005
RIAA	105(100)	101(100)	100(100)
LS	111(100)	107(100)	104(100)
MAPES	117(100)	113(100)	101(100)
Capon	113(100)	111(100)	100(100)
LSPR	105(100)	102(100)	101(100)

The schemes were also compared on their ability to infer closely embedded multiple frequencies in the data set. Gaussian noise was added to sinusoids with frequencies, $f_1 = 0.29Hz$, $f_2 = 0.32Hz$ and sampled irregularly using the same Poisson process as in Figure (1), the signal to noise ratio was set to 3dB.

With only 16 samples, only RIAA is able to detect the embedded frequencies consistently. LS and LSPR were able to detect the frequencies but based on Figures 3b and 3e, our testing methodology would have resulted in a miss for these frequencies. Capon and MAPES performed poorly for 16 sample points.

However, when the number of samples were increased to 24 points, but with SNR of 2dB, Figure (4) shows the performance of the five scheme with the same frequencies as Figure (3), $f_1 = 0.29Hz$, $f_2 = 0.32Hz$ and amplitudes 0.45, and 0.35

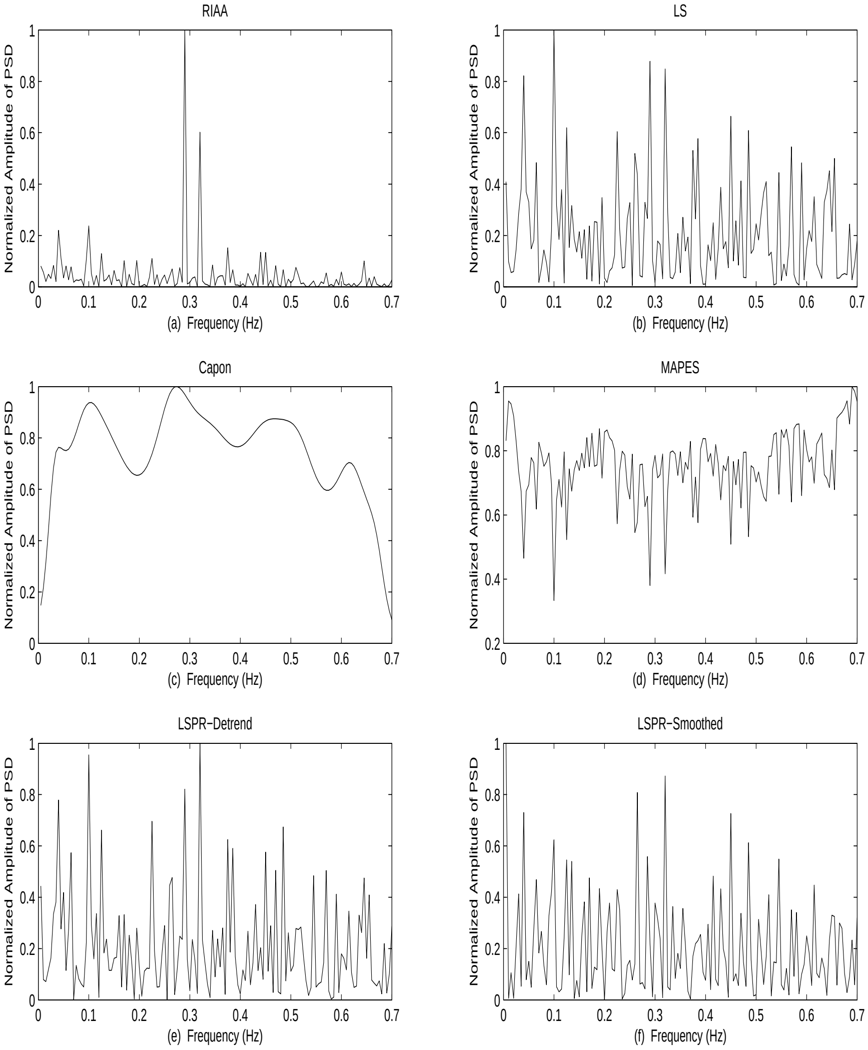


Fig. 3: Performance comparison based on an artificial data set (N=16) with sinusoids with frequencies $f_1 = 0.29Hz$, $f_2 = 0.32Hz$ (a) RIAA (b) LS (c) Capon (d) MAPES (e) LSPR Detrend (f) LSPR Smoothed

respectively. From the graph, it is obvious to see that RIAA does not suffer from detrimental sidelobes nor mainlobe leakages that LS and LSPR appear to exhibit. Still, both Capon and MAPES are lacking behind in detecting the frequencies. They are able to detect the frequency at 0.29Hz but not the frequency at 0.32Hz.

When the sample size is increased to 100 points, SNR still at 2dB, Figure (5) shows that Capon and MAPES improved dramatically. However, the smoothed version of LSPR could still not detect the two frequencies consistently. The number of samples had to be increased to over 200 samples points before it detected the two frequencies. Such a method is not ideal for microarray data sets where sample size is of premium. An auxiliary interest was to investigate the computational time required by each scheme.

From simulations, Figure (6) shows the disparities in computation time between MAPES and the other schemes. Due to the expectation maximization step in MAPES, it was the only scheme that required noticeable time in computing the power spectral estimates.

The ability of the schemes to detect a periodic signal, sampled with a Poisson process was investigated. With sample times points just 18 and SNR increased from 0 to 3dB, 200 simulations were run for each SNR value and Figure (7) shows the number of times the periodic signal was detected at the exact frequency. In Figure(7), RIAA at SNR=2.7dB was able to detect the embedded frequency out of the 200 simulations runs. It was not after 3dB that the other four schemes were able to detect the frequency for all 200 simulations runs with 18 time points sampled irregularly.

3. Simulation on Spellman's Yeast Data

The schemes were then evaluated on a real biological data set from Spellman's experiment [2]. Performance was judged based on their ability to recover genes from

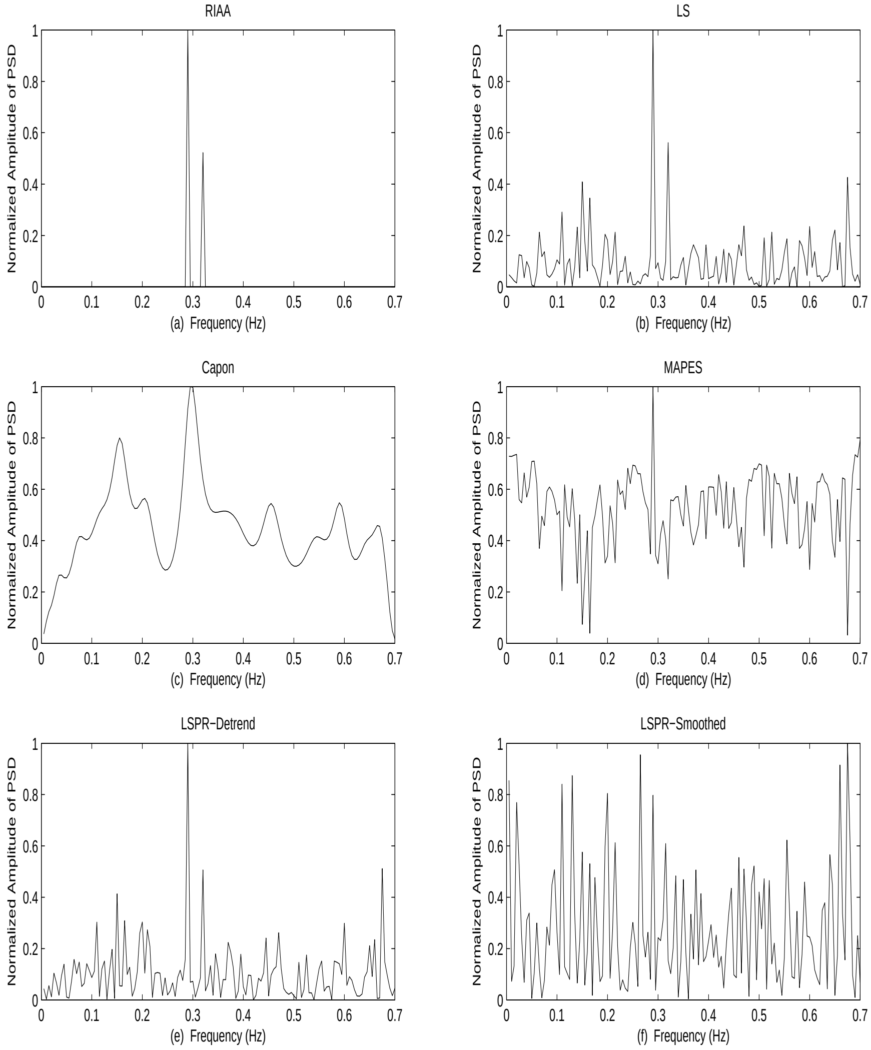


Fig. 4: Performance comparison based on artificial data set (N=24) with sinusoids with frequencies $f_1 = 0.29Hz$, $f_2 = 0.32Hz$. (a) RIAA (b) LS (c) Capon (d) MAPES (e) LSPR Detrend (f) LSPR Smoothed

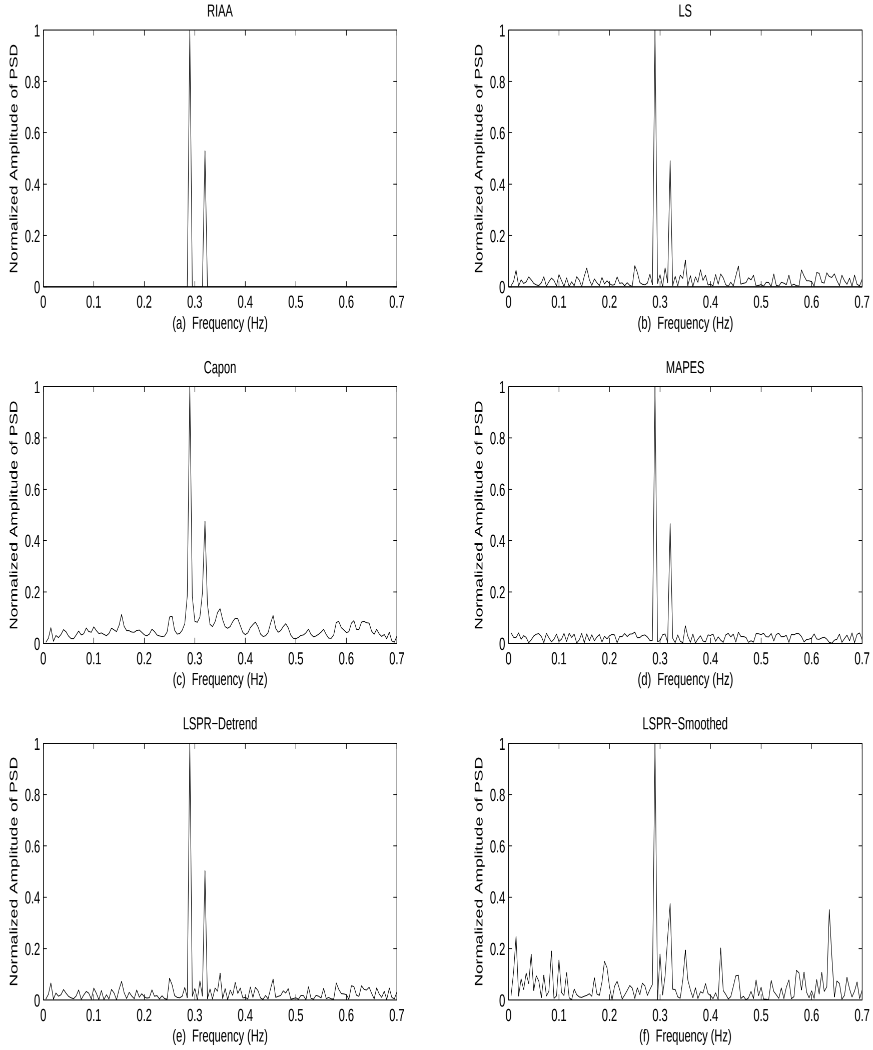


Fig. 5: Performance comparison based on an artificial data set (N=100) with sinusoids with frequencies $f_1 = 0.29Hz$, $f_2 = 0.32Hz$. (a) RIAA (b) LS (c) Capon (d) MAPES (e) LSPR Detrend (f) LSPR Smoothed

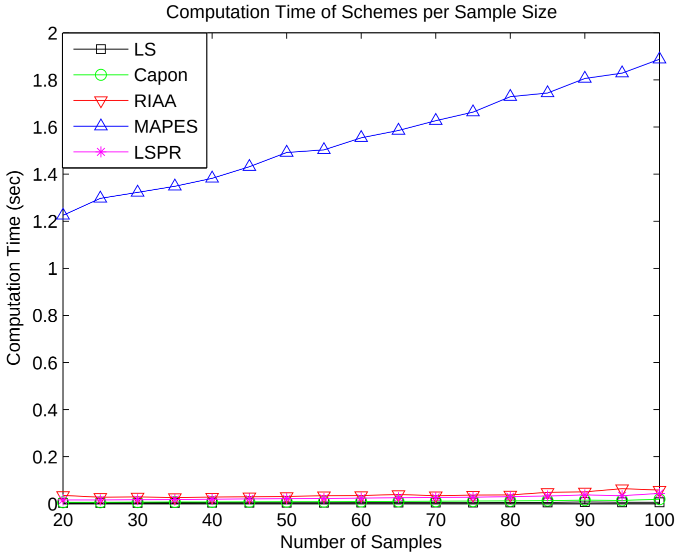


Fig. 6: Computational time for all schemes based on number of sample time points available.

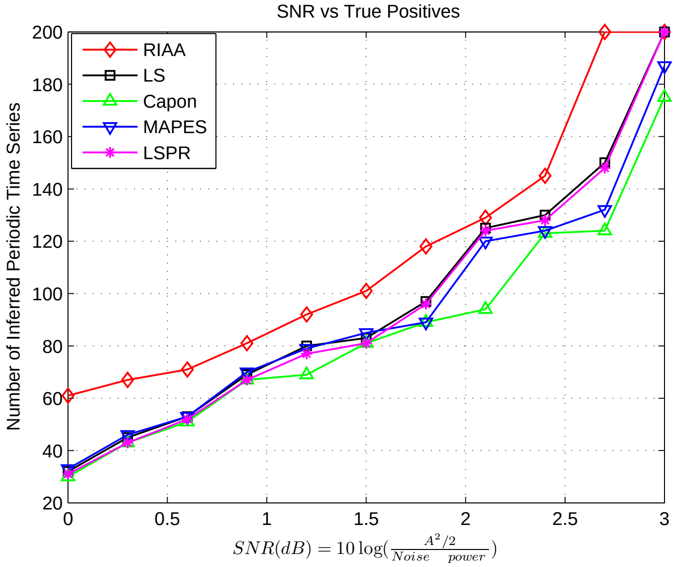


Fig. 7: Performance comparison as SNR is increased from 0 to 3dB.

a set of known periodic genes that were obtained from a small scale experiment. At the time of Spellman's work, there were 104 known periodic genes for the yeast, later in 2003, Johansson [7] added nine genes to provide researchers with 113 cell cycle regulated reference genes. From here on, the 113 cell cycle regulated genes will be referred to as Benchmark set A.

As mentioned in Chapter 1, the standard method in evaluating the performance of schemes that seek to detect periodically expressed genes is to determine the percentage of the reference genes the scheme was able to infer as periodic. The best schemes are expected to have a high number of the reference genes present in the fewest number of inferred genes. For example, Spellman was unable to obtain 92% of the 104 reference genes until 800 genes were inferred or judged to be periodic.

For comparison, the dataset for Cdc15 arrest and Alpha arrest synchronization from the experiment of Spellman [2] were used. Cdc15 data set had 24 sample time points and Alpha data set had 18 time points, there were too few samples for cdc28 and elutriation synchronization data and thus not ideal for Capon as has been demonstrated via artificial data simulation.

The comparison procedure was done as follows, based on the given dataset, each schemes infer a pre-specified number of genes. The inferred genes are designated as periodically expressed genes and are crossed with Benchmark set A. A percentage is obtained from the number of the referenced gene set that are present in the pre-specified number of genes inferred. This is illustrated in Figure (8) where the superiority of RIAA is clearly demonstrated in identifying more known periodically expressed genes than any other scheme for the Cdc15 experiment.

Capon method however performed much better on biological data set than the MAPES based on the criteria used to measure performance. Since only one frequency was believed to be present, resolution for the Capon method was sacrificed in favor of

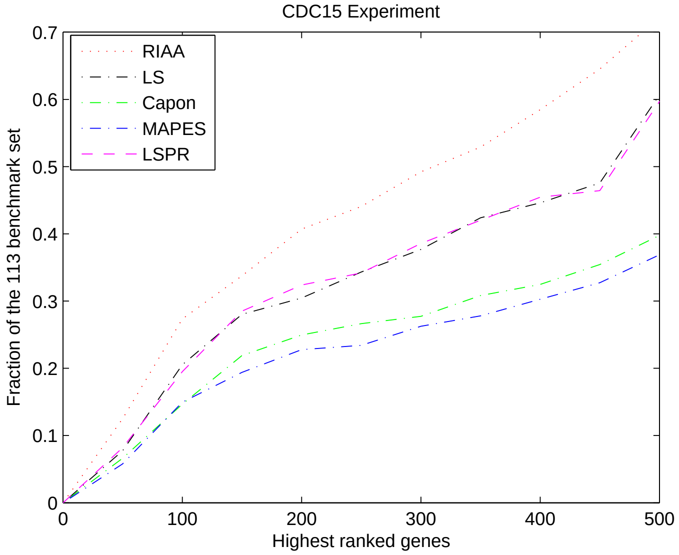


Fig. 8: Performance comparison based on cdc15 arrest data set.

accuracy and this gave the Capon more samples to use within the confines of its methodology. As mentioned previously, the Capon method needs to decide on a tradeoff between resolution and accuracy and the filter length plays a central role in this tradeoff. A small filter length affects the resolution especially in the case when there is a need to differentiate between two closely embedded frequencies. As expected with LS and LSPR, there was no appreciable performance separation between the two.

Applying the schemes on the Alpha data set, RIAA continued to demonstrate its efficacy in matching the referenced genes set per pre-specified inferred genes. With only 18 time points available with some genes having missing data as well, MAPES outperformed the Capon method on this dataset. Again the performance of LS and LSPR were almost identical. As can be seen from Figure (9), there was a slight drop off in the percentage of referenced genes that RIAA and all the other schemes were able to pick, this was expected and understandable with the limited time points available for the Alpha data set. From these figures, it is easy to see that RIAA outperforms the other schemes and should be the analysis tool of choice when the goal of an analysis on a microarray experiment data set is to seek periodically expressed genes.

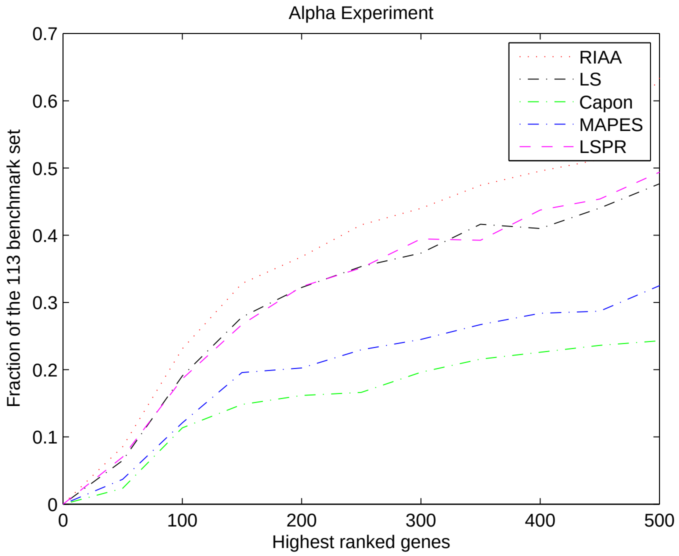


Fig. 9: Performance comparison based on Alpha arrest data set.

B. Discussion and Conclusions

The datasets of Spellman [2] and Pramila [24] were analyzed using RIAA. Pramila's alpha arrest experiment data set has 25 samples and Spellman's cdc15 experiment has 24 time points. There were numerous missing data points rendering the data set as irregularly sampled. With a q value set to be not more than 0.05, 609 genes were adjudged to be periodic in Spellman's dataset and 596 in Pramila's dataset, the results are shown in Table B1 and B2 respectively in Appendix B. An overlap of 543 genes was obtained between the two data sets. Using RIAA, the results obtained establishes a better level of clarity in the overlap of periodically expressed genes between two different datasets. The overlap of 543 genes is appreciably more than any two different results reported on the yeast which can be found on [37].

Compared with Spellman [2], there was only an overlap of 357 genes. RIAA using only 550 genes detected 97% out of 104 genes that were known to be cell cycled regulated at the time of Spellman's work while Spellman's method only got 92% out of 800 genes. It must be added that expression profiles of the genes may be more complex than simple sinusoidal curves, however, the visual inspection of the time series profile reveals that the genes inferred to be periodic appeared as sinusoidal in nature and made the assumption sinusoids sound. The earlier work in [1], concluded that LS was effective and an accurate tool to use, but through artificial simulations, it has been seen that LS can be sensitive to large outliers that could be present due to perturbation in the measurement environment. RIAA does not suffer from such sensitivity and is innately designed to limit false discoveries. The Capon filter is a powerful tool, also robust to the presence of noise, but desires a bigger sample size than a typical microarray data sets provides. MAPES is computational expensive and also requires much more data points to be effective in inferring periodically expressed gene from microarray experiments. LSPR turned out not to outperform LS in terms of the actual number of true positives but reduced the number of false positives that LS picks. It is recommended to future researchers seeking to find periodically expressed genes in a microarray experiment to employ RIAA as it has been proven to be an effective tool in identifying periodic gene expression profiles. It is robust to small sample sizes, missing data or clusters of missing data and irregularly sampled data.

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, 2008.

[2] 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, no. 12, pp. 3273–3297, 1998.

[3] R.J. Cho, M.J. Campbell, E.A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T.G. Wolfsberg, A.E. Gabrielian, D. Landsman, D.J. Lockhart, et al., "A genome-wide transcriptional analysis of the mitotic cell cycle," *Molecular Cell*, vol. 2, no. 1, pp. 65–73, 1998.

[4] S. Marguerat, T.S. Jensen, U. de Lichtenberg, B.T. Wilhelm, L.J. Jensen, and J. Bähler, "The more the merrier: comparative analysis of microarray studies on cell cycle-regulated genes in fission yeast," *Yeast*, vol. 23, no. 4, pp. 261–277, 2006.

[5] 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, no. 8, pp. 809–817, 2004.

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

[7] 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.

[8] 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.

[9] P. Stoica and R.L. Moses, *Spectral analysis of signals*, Upper Saddle River:Prentice Hall, 2005.

[10] Y. Wang, P. Stoica, J. Li, and T.L. Marzetta, “Nonparametric spectral analysis with missing data via the em algorithm,” *Digital Signal Processing*, vol. 15, no. 2, pp. 191–206, 2005.

[11] H. He, J. Li, and P. Stoica, “Spectral analysis of non-uniformly sampled data: A new approach versus the periodogram,” in *Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop, 2009. DSP/SPE 2009. IEEE 13th*. IEEE, 2009, pp. 375–380.

[12] 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.

[13] R. Yang and Z. Su, “Analyzing circadian expression data by harmonic regression based on autoregressive spectral estimation,” *Bioinformatics*, vol. 26, no. 12, pp. i168–i174, 2010.

[14] C.D. Giurcaneanu, “Stochastic complexity for the detection of periodically expressed genes,” in *Genomic Signal Processing and Statistics, 2007. GENSIPS*

2007. IEEE International Workshop on. IEEE, 2007, pp. 1–4.

[15] 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, no. 1, pp. 117, 2005.

[16] Y. Luan and H. Li, "Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data," *Bioinformatics*, vol. 20, no. 3, pp. 332–339, 2004.

[17] X. Lu, W. Zhang, Z.S. Qin, K.E. Kwast, and J.S. Liu, "Statistical resynchronization and bayesian detection of periodically expressed genes," *Nucleic Acids Research*, vol. 32, no. 2, pp. 447–455, 2004.

[18] S. Wichert, K. Fokianos, and K. Strimmer, "Identifying periodically expressed transcripts in microarray time series data," *Bioinformatics*, vol. 20, no. 1, pp. 5–20, 2004.

[19] T. Bowles, A. Jakobsson, and J. Chambers, "Detection of cell-cyclic elements in mis-sampled gene expression data using a robust capon estimator," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*. IEEE, 2004, vol. 5, pp. V–417.

[20] U. De 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.

[21] N.R. Lomb, "Least-squares frequency analysis of unequally spaced data," *Astrophysics and Space Science*, vol. 39, no. 2, pp. 447–462, 1976.

[22] J.D. Scargle, “Studies in astronomical time series analysis. ii-statistical aspects of spectral analysis of unevenly spaced data,” *The Astrophysical Journal*, vol. 263, pp. 835–853, 1982.

[23] P. Stoica and N. Sandgren, “Spectral analysis of irregularly-sampled data: Paralleling the regularly-sampled data approaches,” *Digital Signal Processing*, vol. 16, no. 6, pp. 712–734, 2006.

[24] T. Pramila, W. Wu, S. Miles, W.S. Noble, and L.L. Breeden, “The forkhead transcription factor hcm1 regulates chromosome segregation genes and fills the s-phase gap in the transcriptional circuitry of the cell cycle,” *Genes & Development*, vol. 20, no. 16, pp. 2266–2278, 2006.

[25] H. Akaike, “A new look at the statistical model identification,” *Automatic Control, IEEE Transactions on*, vol. 19, no. 6, pp. 716–723, 1974.

[26] J. Taylor and S. Hamilton, “Some tests of the vaníček method of spectral analysis,” *Astrophysics and Space Science*, vol. 17, no. 2, pp. 357–367, 1972.

[27] L. Eyer and P. Bartholdi, “Variable stars: which nyquist frequency?,” *Arxiv Preprint Astro-Ph/9808176*, 1998.

[28] P. Babu and P. Stoica, “Spectral analysis of nonuniformly sampled data–a review,” *Digital Signal Processing*, vol. 20, no. 2, pp. 359–378, 2010.

[29] S.M. Kay, *Fundamentals of statistical signal processing: Estimation theory*, Englewood Cliffs: Prentice-Hall, 1993.

[30] U. de Lichtenberg, R. Wernersson, T.S. Jensen, H.B. Nielsen, A. Fausbøll, P. Schmidt, F.B. Hansen, S. Knudsen, and S. Brunak, “New weakly expressed cell cycle-regulated genes in yeast,” *Yeast*, vol. 22, no. 15, pp. 1191–1201, 2005.

[31] J. Fan and Q. Yao, *Nonlinear time series: Nonparametric and parametric methods*, New York: Springer Verlag, 2003.

[32] P. Whittle, "Tests of fit in time series," *Biometrika*, vol. 39, no. 3/4, pp. 309–318, 1952.

[33] R.A. Fisher, "Tests of significance in harmonic analysis," *Proceedings of the Royal Society of London. Series A*, vol. 125, no. 796, pp. 54–59, 1929.

[34] P.J. Brockwell and R.A. Davis, *Time series: theory and methods*, New york: Springer Verlag, 2009.

[35] J.D. Storey, "A direct approach to false discovery rates," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 64, no. 3, pp. 479–498, 2002.

[36] J.D. Storey, "The positive false discovery rate: A bayesian interpretation and the q-value," *Annals of Statistics*, pp. 2013–2035, 2003.

[37] N.P. Gauthier, M.E. Larsen, R. Wernersson, U. De Lichtenberg, L.J. Jensen, S. Brunak, and T.S. Jensen, "Cyclebase. orga comprehensive multi-organism online database of cell-cycle experiments," *Nucleic Acids Research*, vol. 36, no. suppl 1, pp. D854–D859, 2008.

APPENDIX A

CODES

1. Matlab codes

Lomb-Scargle function

```
    function psd = LombScargle(T,X,W)
% this function is to use loom-scargle
% inputs:
% T - time points
% X - sampled data
% W - frequencies
% outputs:
% psd - power spectral density corresponds to the frequencies
std_X = std(X);
mean_X = mean(X);

for k = 1:length(W)
tau = 1/2/W(k) * atan(sum(sin(2*W(k)*T))/sum(cos(2*W(k)*T)));
psd(k) = 1/2/std_X^2 * ( sum((X-mean_X).*cos(W(k)*(T-tau)))^2
/sum(cos(W(k)*(T-tau)).^2) ...
+ sum((X-mean_X).*sin(W(k)*(T-tau)))^2/sum(sin(W(k)*(T-tau)).^2)  );
end
```

MAPES Function

```
    function PSD = pmapes(X,T,W)

% T has to be integers

%W = 0.05:0.05:pi;

n = T(end)-T(1);


if size(X,1) == 1   % row vector

    X = X.';    % change it to column vector

end

if size(T,1) == 1   % row vector

    T = T.';    % change it to column vector

end


% ---initilization--------------------------

% set 0 to missing data

XX = [];

avail = []; % availability

for k=1:length(T)

    XX = [XX,X(k)];

    avail = [avail,1];

    if k~=length(T) && T(k+1)-T(k)>1     % not the tail,

    therefore k+1 is valid

        XX = [XX,zeros(1,T(k+1)-T(k)-1)];  % set zeros

        to missing positions

        avail = [avail,zeros(1,T(k+1)-T(k)-1)];

    end

end
```

```
N = length(XX);
miss = ones(1,N) - avail;
g = length(X);  % # data available
M = ceil(N/2);
L = N-M+1;
% initilize Q, Sg, Sm
Sg = zeros(L*M,g);
Sm = zeros(L*M,N-g);
for l = 0:(L-1)
    for k = 1:M
        if avail(l+k) == 1  % there is a datum here
            Sg(l*M+k,sum(avail(1:l+k))) = 1;
        else    % there is a miss here
            Sm(l*M+k,sum(miss(1:l+k))) = 1;
        end
    end
end
Sg_tilde = (inv(Sg.'*Sg)*Sg.').';
Sm_tilde = (inv(Sm.'*Sm)*Sm.').';

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

for n = 1:length(W)
    w = W(n);

    % ---initilize alpha--------------------------
    alpha = sqrt(LombScargle(X,T,w));
    a = exp((0:M-1)*j*w).’;
    % initilize Q
    Q = zeros(M,M);
    for l = 0:(L-1)
        yl = XX((1+l):(l+M)).’;
        Q = Q + (yl-alpha*a*exp(j*w*l))*(yl-alpha*a*exp(j*w*l))’;
    end
    Q = Q/L;

% ---start iterations of EM--------------------
e = 0;    % arbitrarily set error to a large value
alpha_old = inf;
rho = [];
for l=0:L-1
    rho = [rho;exp(j*w*l)*a];
end

loops = 0;
while abs(alpha-alpha_old)/abs(alpha) > 0.1 && loops<100
    loops = loops+1;
    alpha_old = alpha;

D = Q;
for l=1:L-1
    D = [D,zeros(size(D,2),M);zeros(M,size(D,2)),Q];
```

```
end
b = Sm_tilde.’*rho*alpha + Sm_tilde.’*D*Sg_tilde*
inv(Sg_tilde.’*D*Sg_tilde)*(X-Sg_tilde.’*rho*alpha);
K = Sm_tilde.’*D*Sm_tilde + Sm_tilde.’*D*Sg_tilde*
inv(Sg_tilde.’*D*Sg_tilde)*Sg_tilde.’*D*Sm_tilde;

S_tilde = zeros(M,M);
Z = zeros(M,1);
SmKSm = Sm*K*(Sm.’);
SgrSmb = Sg*X+Sm*b;
for l=0:L-1
Gammal = SmKSm((l*M+1):(l*M+1),(l*M+M):(l*M+M));
zl = SgrSmb((l*M+1):(l*M+M));
S_tilde = S_tilde + Gammal + zl*zl’;
Z = Z+ zl*exp(-j*w*l);
end
Z = Z/L;
S_tilde = S_tilde/L - Z*Z’;
S_tilde = S_tilde + 0.01*diag(diag(S_tilde));   %diagnol loading
invS_tilde = inv(S_tilde);
alpha = (a’*invS_tilde*Z)/(a’*invS_tilde*a);
Q = S_tilde + (alpha*a-Z)*((alpha*a-Z)’);
end

    PSD(n)=alpha;
end
```

```
PSD = abs(PSD).^2;
PSD=circshift(PSD,[0,1]);
```

RIAA function

```
    function [P_RIAA ] = PSD_RIAA(X, W, T, N, K, s_no,)

P_RIAA = zeros(K, 1);
Theta = zeros(2, K);
set_A = zeros(N, 2, K);
for j = 1:K
    omega = W(j);
    set_A(:,:,j) = [cos(omega*t_n) sin(omega*t_n)];
end

% Initialization with Least Squares Periodogram
for k = 1:K
    A = set_A(:,:,k);
    Theta(:,k) = inv(A'*A) * A'*y;
end

% Power in signal estimation and initiation of iteration
num_o= 0;
flag = 1;
while flag
    Theta_tmp = Theta;
    Alpha_tmp = sqrt(Theta_tmp(1,:).^2 + Theta_tmp(2,:).^2);
   gam = zeros(N, N);
    y_esti = zeros(N, 1);
    for j = 1:K   % calculate gam
        A = set_A(:,:,j);
       gam = Gam + (Theta(1,j)^2 + Theta(2,j)^2) / 2 * A * A’;
        y_est = y_est+ A * Theta(:, j);
    end

    in_Gam = inv(Gam);
    for j = 1:K
        A = set_A(:,:,j);
        Theta(:,j) = inv(A’ * in_Gam * A) * (A’ * in_Gam * y);
    end
    num_o= num_o+ 1;
    if num_o>= stop_no
        flag = 0;
    end
    Alpha = sqrt(Theta(1,:).^2 + Theta(2,:).^2);
    sentinel = norm(Alpha - Alpha_tmp) / norm(Alpha_tmp);

    if (sentinel < 5e-3)
        flag = 0;
    end
end
```

```
for j = 1:K
    A = set_A(:,:,j);
    P_RIAA(j) = 1/N * Theta(:,j)’ * (A’ * A) * Theta(:,j);
end
```

Capon function

```
    function P = pcapon(X,T,m,W)
% function [P,W] = pcap(X,T)
% power spectral density estimation by using capon method
% irregular sampling
% P - power spectral
% W - frequency list
% X - input data sequence
% T - data sampling time points
% m - the order of the filter

X = (X-mean(X))/std(X);

per = LombScargle(T,X,W);
%per = pergram(t,X,W);

R = zeros(m+1,m+1);
wdelta = W(2)-W(1);
for k = 1:length(W)
```

```
    a = exp(-i*wdelta*k*(0:m)).';
    R = R + a*a'*per(k);
end
R = (R/length(W));
%R = R + diag(0.01*diag(R));     % diagnal loading

% from stoica's Forward-backward
J = zeros(m+1,m+1);
for k = 1:m+1
    J(k,m+2-k) = 1;
end
R = 0.5*(R+J*transpose(R)*J);
invR = inv(R);

for k = 1:length(W)
    w = W(k);
    a = exp(-i*w*(0:m)).';
    P(k) = 1/(a'*invR*a);
end
```

Data generator

```
function [Perodata sampltimes N noise]= datagenerator(f,lamda,N,A)

m=N;
```

```
times= exprnd(1/lamda,[m 1]);
times = cumsum(times);

num_sinu=length(A);
sampltimes=times;

phi = 2*pi * rand([num_sinu,1]);
variance=0.01;

noise = sqrt(variance) * randn(m, 1);

y=(A’*cos(2*pi*f*times’ + repmat(phi, [1,m])))’;

Perodata=y;
```

APPENDIX B

RESULTS

2. Spellman's dataset

Table V: RIAA Cyclic genes from Spellman's dataset

YMR215W	YIL132C	YCR040W	YBL009W	YNL044W	YIL052C
YBL002W	YBR038W	YLR170C	YBR189W	YCR084C	YML119W
YPL163C	YGR099W	YOL105C	YGR189C	YIL146C	YDR345C
YHR175W	YGL028C	YPR119W	YPL267W	YOR058C	YGL253W
YMR305C	YGL089C	YJR022W	YNL300W	YJL167W	YNL037C
YJL092W	YPL256C	YBR158W	YDR055W	YPL273W	YDL133W
YHR086W	YOR070C	YJL200C	YOR234C	YJL052W	YNL233W
YPL128C	YGL161C	YGL254W	YOR127W	YBR214W	YJL122W
YDL055C	YLR342W	YKL127W	YNL074C	YLR333C	YER088C
YLL028W	YBR071W	YPL090C	YDR450W	YJR009C	YDR089W
YER001W	YJL174W	YER041W	YDL082W	YHR052W	YFL037W
YBR243C	YBR092C	YGR086C	YNL145W	YOL143C	YCL040W
YJL159W	YKL175W	YLR121C	YPL168W	YFL026W	YEL026W
YNL283C	YDL164C	YLR390W-A	YOR025W	YER006W	YJL118W
YNL015W	YCL014W	YDR452W	YJL079C	YNR014W	YKL184W
YKR042W	YOR084W	YPL158C	YGL013C	YGR240C	YCR048W
YBR093C	YDL227C	YDL170W	YHL027W	YHR174W	YNL289W
YAR007C	YPL127C	YGR092W	YIL066C	YKR077W	YGR214W

YEL042W	YLR194C	YER026C	YNL162W	YLR183C	YPL187W
YGR044C	YMR011W	YJL157C	YDL191W	YOR312C	YJL181W
YDL224C	YDR097C	YLR325C	YDR224C	YKL148C	YLR287C-A
YGR108W	YML083C	YOR264W	YMR023C	YKR037C	YCR089W
YML052W	YKL045W	YNR001C	YCR069W	YML051W	YOR315W
YOL012C	YGR065C	YNL160W	YML085C	YKL185W	YML102W
YIL123W	YGL200C	YDR302W	YDR146C	YGR230W	YGR192C
YKL096W	YMR189W	YBL032W	YNL192W	YJR046W	YJR092W
YLR286C	YOR383C	YMR058W	YER056C-A	YDL018C	YNR019W
YKL066W	YHR005C	YCR067C	YLL041C	YPR181C	YOR378W
YAR071W	YGL008C	YNL327W	YJL051W	YDL010W	YGR041W
YKL165C	YMR163C	YMR179W	YBR083W	YML064C	YKL067W
YHR143W	YEL002C	YAR018C	YGL116W	YPL242C	YJL187C
YOL007C	YER070W	YJR010W	YAR050W	YOL091W	YKR024C
YDL003W	YHR211W	YGL154C	YPR032W	YDR042C	YML099C
YNL176C	YBL102W	YPR120C	YPL188W	YOR023C	YLR275W
YMR042W	YOR247W	YJL206C	YMR032W	YNL078W	YOR322C
YPR149W	YPL032C	YOL060C	YHR215W	YGL255W	YLR180W
YKL164C	YHR061C	YBR049C	YJL185C	YBR196C	YBR102C
YER095W	YKL001C	YJL137C	YOR009W	YLR448W	YHR165C
YML058W	YLR190W	YDR033W	YMR078C	YGL093W	YNL112W
YBR221C	YLR164W	YOR382W	YMR048W	YLR437C	YPL190C
YJL134W	YJR127C	YLR103C	YHR188C	YBR181C	YGL090W
YMR307W	YIL056W	YDR488C	YJL074C	YER065C	YNR009W

YOR144C	YLR373C	YKR013W	YNR068C	YNL096C	YPL081W
YPL208W	YDL101C	YDL155W	YOL123W	YBR200W	YLR182W
YNL058C	YIR018W	YGL055W	YPR019W	YJL078C	YLR176C
YNL197C	YLR372W	YOR176W	YMR019W	YBR142W	YER124C
YER003C	YAL022C	YOR019W	YOR044W	YKL180W	YMR271C
YHR016C	YDL066W	YBL103C	YJR021C	YDL105W	YDR297W
YKL081W	YJR098C	YOR358W	YPL202C	YJR150C	YDR216W
YHR178W	YBR295W	YGR143W	YDR534C	YKR046C	YDR379W
YOR004W	YEL009C	YGR029W	YBR203W	YOR371C	YJR051W
YOR355W	YER146W	YIR038C	YBL063W	YDR309C	YNL312W
YNL072W	YLR048W	YBR009C	YEL017W	YNL298W	YGL184C
YKL009W	YIL018W	YKL025C	YCR061W	YOR308C	YLR210W
YDL064W	YER093C-A	YFR053C	YBR067C	YLL022C	YPR030W
YGR152C	YAR073W	YDR115W	YNL030W	YLR005W	YIL016W
YHR021C	YKR071C	YPR132W	YHR149C	YIL011W	YDR463W
YOL019W	YMR141C	YCL067C	YDR028C	YGR166W	YFR016C
YPL253C	YDR436W	YBR073W	YNL248C	YFR034C	YKL104C
YOR256C	YAL043C	YIL131C	YGL030W	YGR272C	YDR381W
YKR039W	YML056C	YBR219C	YDR356W	YOR205C	YHR202W
YGR221C	YHR123W	YPR106W	YPL221W	YEL057C	YCR065W
YGR161C	YLR274W	YDR481C	YDR408C	YDL179W	YAR035W
YPL177C	YGR079W	YLR212C	YGL237C	YGL027C	YGL259W
YDR447C	YIR039C	YDR085C	YBR138C	YAL024C	YCR024C-A
YBR070C	YBR240C	YPL061W	YOR221C	YDR124W	YKL008C

YLR284C	YCL024W	YAR002W	YMR021C	YDR416W	YKL048C
YGL037C	YER178W	YML041C	YIL152W	YGR109C	YHR141C
YOR095C	YGR034W	YIL133C	YOR120W	YDR386W	YCL027W
YMR317W	YDL028C	YLR455W	YIL050W	YBR021W	YDR025W
YDR067C	YOR182C	YOR016C	YGR279C	YKL113C	YGR068C
YLR367W	YER129W	YHR006W	YLR378C	YLR426W	YMR164C
YDL220C	YLR290C	YKR010C	YOR198C	YDL239C	YDR446W
YDL142C	YOR288C	YDR047W	YNL002C	YKR019C	YOL036W
YMR016C	YLR409C	YER118C	YLR457C	YGR075C	YIR022W
YDR528W	YOR204W	YDR425W	YNR047W	YPR001W	YLR326W
YBR267W	YPR156C	YJL063C	YLR353W	YGR220C	YDR421W
YGR027C	YHR153C	YOR272W	YDR507C	YLR131C	YOL127W
YDL194W	YKR099W	YOL090W	YGL062W	YGL207W	YHL047C
YBR130C	YER111C	YEL032W	YKL011C	YGR282C	YMR261C
YLR288C	YKR094C	YER075C	YOR153W	YJR137C	YBL003C
YAL032C	YDL012C	YLR213C	YOR178C	YPL089C	YHR158C
YOL070C	YPL014W	YFL033C	YMR199W	YLR049C	YDR255C
YHR094C	YGR288W	YKL020C	YMR184W	YEL050C	YBL054W
YOR142W	YDL087C	YJL056C	YPL153C	YOR313C	YER167W
YDR310C	YIL129C	YJR155W	YHR203C	YAR008W	YLR304C
YOR338W	YFL021W	YMR031C	YOR310C	YLR394W	YGR013W
YNL103W	YIL122W	YMR145C	YOR122C	YNL069C	YMR070W
YLR313C	YCL061C	YER089C	YPL146C		

3. Pramila's dataset

Table VI: RIAA Cyclic genes from Pramila's dataset

YDR225W	YMR076C	YDL018C	YPR019W	YBR202W	YFL008W
YBL003C	YPL127C	YOL090W	YPL061W	YPL116W	YPL255W
YNL300W	YKL113C	YMR215W	YDR507C	YBL009W	YPL124W
YBR009C	YNL312W	YMR011W	YMR179W	YGR109C	YDL093W
YER070W	YDR055W	YLR274W	YOR273C	YOR114W	YAL040C
YNL030W	YGR189C	YJL115W	YLR254C	YCR065W	YLR342W
YBL002W	YAR007C	YML058W	YER003C	YML060W	YML033W
YPL163C	YKL101W	YJL074C	YDL101C	YDL156W	YOR073W
YDR224C	YBL035C	YIL026C	YMR003W	YJL157C	YGR014W
YNL289W	YER001W	YOR247W	YDL197C	YLL021W	YDL096C
YBR089W	YIL140W	YDR097C	YNL233W	YER032W	YOR229W
YDL003W	YHR152W	YNL058C	YKL045W	YOR373W	YBR067C
YJL159W	YBR070C	YNL126W	YGR152C	YNL057W	YKL008C
YBR010W	YLR103C	YLR194C	YCR042C	YNL088W	YLR383W
YPL256C	YGL021W	YEL032W	YMR307W	YLL022C	YGR221C
YFL026W	YNL145W	YMR078C	YIL131C	YCR024C-A	YNL166C
YOL007C	YHR154W	YLR045C	YBR088C	YKR042W	YNL192W
YNL031C	YPL267W	YGR092W	YNL262W	YOR083W	YBL111C
YLR183C	YDR222W	YPL153C	YLR121C	YJL073W	YJR030C
YNL102W	YMR031C	YIL106W	YDR297W	YJL019W	YKL104C
YOR074C	YGL116W	YER095W	YHR172W	YDL055C	YGR099W
YBR071W	YKL209C	YOR195W	YCL061C	YIL123W	YNL082W

YOR066W	YCL024W	YDL164C	YER111C	YBR139W	YDR545W
YHR005C	YDR113C	YFL067W	YPR135W	YML061C	YBR073W
YJL187C	YKR013W	YEL061C	YMR199W	YOR321W	YDR400W
YFL008W	YEL076C-A	YLR463C	YPL141C	YGR286C	YHR023W
YPL255W	YCL040W	YER037W	YGL225W	YOR127W	YDL155W
YPL124W	YML027W	YJL173C	YPR175W	YOL069W	YDR518W
YDL093W	YGR098C	YBL023C	YJR006W	YNL339C	YHR151C
YAL040C	YBR275C	YEL076C	YGR140W	YCR005C	YJL225C
YLR342W	YEL017W	YML085C	YPL221W	YGR296W	YDR481C
YML033W	YLR273C	YHL026C	YPL057C	YMR001C	YIL158W
YOR073W	YEL076W-C	YHR146W	YKL067W	YMR306W	YAR018C
YGR014W	YDL103C	YNL072W	YJR143C	YLL002W	YMR190C
YDL096C	YJL051W	YBR243C	YFL065C	YMR132C	YBR296C
YOR229W	YOR058C	YFL006W	YIL159W	YOL158C	YDR379W
YBR067C	YDR077W	YBL113C	YLR313C	YLR467W	YLR341W
YKL008C	YFL037W	YOR144C	YDR528W	YFL027C	YPR174C
YLR383W	YGR279C	YNR001C	YHR218W	YOR313C	YKL089W
YGR221C	YDR191W	YER118C	YOR246C	YNL338W	YMR006C
YNL166C	YNL273W	YKR010C	YHL021C	YNL150W	YOR288C
YNL192W	YPL032C	YKR098C	YLR032W	YDR146C	YKR037C
YBL111C	YML052W	YPR018W	YMR292W	YOR248W	YMR253C
YJR030C	YGL027C	YKL042W	YFL068W	YHL050C	YBR093C
YKL104C	YDR488C	YBL031W	YLR455W	YDL138W	YDR503C
YGR099W	YMR048W	YGR143W	YIR010W	YLR326W	YPL242C

YNL082W	YKL185W	YDL102W	YIL122W	YBR138C	YDR307W
YDR545W	YDR279W	YIL015W	YEL075C	YKL165C	YJR053W
YBR073W	YAL039C	YBL109W	YCL012W	YLR234W	YLR049C
YDR400W	YFL066C	YJR092W	YLR386W	YML125C	YPR076W
YKR090W	YNL111C	YHR086W	YDR501W	YBR140C	YOR111W
YGR188C	YOL138C	YIL066C	YOL025W	YOR363C	YDR460W
YPL209C	YER114C	YHR158C	YGL207W	YHR159W	YGR075C
YOR372C	YJL155C	YDL105W	YDR516C	YDL056W	YOR315W
YDR464W	YDR219C	YBR015C	YER190W	YBR028C	YPR139C
YGL061C	YEL042W	YOR016C	YPR004C	YDR457W	YOR026W
YBR086C	YBR072W	YOL124C	YOR307C	YOR033C	YNL165W
YPL283C	YML069W	YDL095W	YER016W	YNL309W	YEL031W
YML133C	YNL310C	YPR031W	YIR044C	YCR090C	YBR153W
YJL044C	YPR203W	YLR074C	YPL004C	YKL048C	YGR022C
YHL049C	YML119W	YAR008W	YGR142W	YML021C	YMR075W
YLR247C	YBR087W	YJR076C	YER189W	YDR436W	YDR544C
YBR038W	YDR261C	YGL037C	YDR245W	YLR210W	YKL129C
YNL334C	YBR042C	YJR054W	YLL067C	YMR117C	YLR457C
YLR462W	YBR092C	YFL060C	YBR242W	YAL007C	YJL185C
YPR149W	YLR182W	YLL032C	YDR277C	YIL155C	YML034W
YLR464W	YAL024C	YHR217C	YLR151C	YLR465C	YER053C
YOR176W	YLR380W	YDL248W	YDR147W	YNL335W	YKL052C
YEL077C	YOR233W	YBR187W	YPR035W	YKL225W	YIL177C
YPL208W	YOL019W	YBL034C	YDL163W	YKL210W	YOL147C

YGR292W	YJR043C	YPR202W	YJR112W	YHR136C	YOR188W
YBR161W	YDR052C	YGR108W	YGL253W	YLR372W	YFL064C
YLR382C	YLR373C	YLR466W	YDR089W	YDL211C	YNL176C
YLL066C	YMR160W	YNL263C	YGL065C	YDL011C	YPL253C
YJL092W	YLL031C	YEL040W	YMR251W-A	YBL112C	YBR103W
YOR084W	YGL200C	YPR034W	YMR030W	YCR023C	YNL095C
YIL127C	YJR003C	YOR017W	YBR289W	YAL034W-A	YAL033W
YDR440W	YNL062C	YHR215W	YNL180C	YPL128C	YNL291C
YKL004W	YCL001W	YOR162C	YAL023C	YJL151C	YGR089W
YOR326W	YOR095C	YNL333W	YHR169W	YGL163C	YPL144W
YGR153W	YMR247C	YLR250W	YDR537C	YNL160W	YIL156W
YGL216W	YGL013C	YDR227W	YJL034W	YJL186W	YDL028C
YML065W	YMR032W	YHR127W	YBL071C	YER105C	YLR025W
YMR144W	YPR052C	YNL149C	YBR276C	YFL044C	YGL050W
YJL137C	YDR489W	YPL007C	YLR063W	YMR197C	YGR026W
YKL160W	YOL030W	YKL049C	YIL144W	YOR025W	YMR127C
YGR012W	YLR034C	YDR302W	YNL197C	YKL161C	YLR190W
YLL028W	YHR219W	YML012W	YJL176C	YGL241W	YCR072C
YBR098W	YLR335W	YDL219W	YDL166C	YER170W	YMR163C
YML124C	YPL227C	YNL238W	YDR212W	YBR133C	YBR302C
YLR212C	YER044C	YML020W	YDL115C	YJL029C	YBR012W-A
YOL017W	YER014W	YAL059W	YBR198C	YPR104C	YAR003W
YJL080C	YIL007C	YIL047C	YAR071W	YGR113W	YBR041W
YFR038W	YKL066W	YCL064C	YDR343C	YMR258C	YJL075C

YGL083W	YBR203W	YCR037C	YKL010C	YKL046C	YLR088W
YDR179C	YNL141W	YOR228C	YGR159C	YNL181W	YIL027C
YPL066W	YGL012W	YOL142W	YKR060W	YPL247C	YJL116C
YGL101W	YLR188W	YDL030W	YKL136W	YPL212C	YDR085C
YNL056W	YOR320C	YLR189C	YJL084C	YIL115C	YLR438W
YNR009W	YDR276C	YOR256C	YIL103W	YNL134C	YBL085W
YGR250C	YKL151C	YNL296W	YFR028C	YDR189W	YLL004W
YGR245C	YJR086W	YDR177W	YJR124C	YDL119C	YER122C
YMR274C	YFL017C	YHR170W	YPL058C	YPR075C	YKR050W
YDR325W	YGL006W	YJL072C			