Supplementary Information

LSPR: an integrated periodicity detection algorithm for unevenly sampled temporal microarray data

Rendong Yang[1,†], Chen Zhang[2,†] and Zhen Su[1,*]

[1]Division of Bioinformatics, State Key Laboratory of Plant Physiology and Biochemistry, College of Biological Sciences, China Agricultural University, Beijing 100193, China

[2]Department of Applied Mathematics, College of Science, China Agricultural University, Beijing 100083, China

1 INTRODUCTION[1]

LSPR is a new periodicity identification algorithm based on the Lomb-Scargle periodogram and harmonic regression method for unevenly sampled time series. It combines the spectral estimation and curve fitting techniques. For a given irregularly sampled time series, our method first estimates the period by the Lomb-Scargle periodogram in the frequency domain, and then models the periodic signals by a harmonic regression method in time domain. Such a joint strategy overcomes the limitations of the Lomb-Scargle periodogram, and gives better descriptions for periodic patterns. In this study, we compared the performances of LSPR with ARSER (Yang and Su, 2010), the Lomb-Scargle periodogram (Lomb, 1976; Scargle, 1982) and COSOPT (Straume, 2004). Results obtained using synthetic and experimental data showed that LSPR was an efficient method of identifying sinusoidal and non-sinusoidal periodic patterns in short, noisy and unevenly sampled time-series.

2 METHODS

2.1 Data preprocessing

Detrending and smoothing are used before rhythm analysis by LSPR. The detrending procedure involves the least square fitting of an increasing or decreasing straight line to the data. If the total variance in the data is significantly less about the trend line than about the mean, then there is a trend in the data. A significant trend will be subtracted from raw time-series $\{x_i\}$, and the detrended time-series $\{\dot{x}_i\}$ will be used in subsequent spectral analysis and harmonic regression in the LSPR algorithm.

Microarray data are commonly perceived as being extremely noisy, and the noise may cause incorrect period identification during spectral analysis. To accurately determine the periods of the cyclical components present in $\{\dot{x}_i\}$, besides directly calculating the spectrum for $\{\dot{x}_i\}$, LSPR also uses the 4th degree Savitzky-Golay filter to smooth $\{\dot{x}_i\}$ and calculate the spectrum for the smoothed time-series $\{\ddot{x}_i\}$. The smoothing procedure can efficiently remove pseudo-peaks caused by noise in a spectrum.

However, smoothing may change the shape of gene expression pattern, making non-periodic signals more likely to be periodic signals, thus yielding artificially significant *p*-values if fitting the smoothed time-series $\{\ddot{x}_i\}$ by a harmonic regression model. To address this problem, LSPR only fits the detrended time-series $\{\dot{x}_i\}$ for harmonic analysis, while the periods used in harmonic regression fitting are derived from spectral analysis of both the smoothed time-series $\{\ddot{x}_i\}$ and detrended time-series without smoothing $\{\dot{x}_i\}$.

2.2 Spectral analysis of unevenly sampled data

LSPR follows the same procedure as ARSER for detecting and analyzing periodicity: data preprocessing, spectral analysis and harmonic regression modeling. The only difference between LSPR and ARSER is the method used in the spectral analysis. ARSER employs autoregressive spectral estimation for the spectral analysis. This method can only be applied to analyze evenly sampled data. For unevenly sampled data, a different spectral analysis method must be used.

LSPR employs a well-established spectral analysis method, the Lomb-Scargle periodogram, to estimate the spectrum for unevenly spaced time series. This method was developed by Lomb and additionally elaborated by Scargle, and makes use of each time point rather than computing based on even intervals. For time series $x_i \equiv x(t_i)(i = 1,2,...,N)$, where t_i is the observation number, the Lomb-Scargle periodogram estimates the spectrum by

$$P_x(\omega) = \frac{1}{2}\left(\frac{[\sum_i x_i \cos\omega(t_i-\tau)]^2}{\sum_i \cos^2\omega(t_i-\tau)} + \frac{[\sum_i x_i \sin\omega(t_i-\tau)]^2}{\sum_i \sin^2\omega(t_i-\tau)}\right) \quad (1)$$

where ω is the corresponding frequency in radians ($\omega = 2\pi f$) and the constant τ is defined by the relation

$$\tan(2\omega\tau) = \frac{\sum_i \sin 2\omega t_i}{\sum_i \cos 2\omega t_i} \quad (2)$$

The generated periodogram for $\{x_i\}$ will exhibit one or several significant peaks accounting for the periodic components. M periods $\{T_j\}(T_j = 1/f_j, j = 1,2,...,M)$ corresponding to these peaks are selected as the cycle length of $\{x_i\}$ for the further harmonic analysis if T_i belongs to the user defined period interval $[T_{start}, T_{end}]$.

2.3 Harmonic regression

Lomb-Scargle spectral analysis predicts the wavelengths for $\{x_i\}$. If we want to determine the presence or absence of such a periodic signal, we must provide a quantitative description of the significance of the periodicity identified from $P_x(\omega)$. The statistical significance of periodicity in $P_x(\omega)$ can be obtained according to the null hypothesis of $P_x(\omega)$, which follows an exponential probability distribution as shown in Scargle, (1982). However this hypothesis testing may fail to detect the significant periodic patterns for short time-series (see Equation 8).

LSPR uses a different method to obtain the statistical significance of periodicity, which is based on the harmonic regression fitting for a detrended time-series $\{\dot{x}_i\}$ according to the periods obtained in the spectral analysis.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First authors.

*To whom correspondence should be addressed. zhensu@cau.edu.cn

Harmonic regression models are used to represent the cyclic trends by fitting $\{\dot{x}_i\}$ with sinusoidal functions as follows:

$$\dot{x}_i = \mu + \sum_{j=1}^{M} \beta_j \cos(\frac{2\pi}{T_j} t_i + \varphi_j) + \varepsilon_i \quad (3)$$

where μ is the mean level of $\{\dot{x}_i\}$, β_j are the amplitudes of different sinusoidal waveforms; φ_j are the phases or the locations of peaks relative to time zero; ε_i is an uncorrelated random variable; t_i are sampling time-points, and T_j are estimated periods from Lomb-Scargle periodogram.

Equation 3 can be reduced to a simple linear equation

$$\dot{x}_i = \mu + \sum_{j=1}^{M} \left\{ p_j \cos(\frac{2\pi}{T_j} t_i) + q_j \sin(\frac{2\pi}{T_j} t_i) \right\} + \varepsilon_i \quad (4)$$

where $p_j = \beta_j \cos\varphi_j$, $q_j = -\beta_j \sin\varphi_j$, and an ordinary least squares (OLS) regression procedure is used to estimate parameters p_j, q_j, μ. Amplitude $\beta_j = \sqrt{p_j^2 + q_j^2}$ and phase φ_j can be obtained from $\tan\varphi_j = -q_j / p_j$.

By applying the harmonic regression, periodicity can be fully described by four parameters: period, phase, amplitude and mean level. An F-test for β_j is employed to determine the statistical significance of periodicity with the associated p-value.

2.4 Multiple testing correction

To analyze large-scale temporal expression profiles, LSPR employed two approaches to making multiple testing corrections. The first, the q-value proposed by Storey and Tibshirani, (2003) is to examine the distribution of p-values from the given data set so that an estimate of the proportion that is truly non-rhythmic can be derived. The p-value for each transcript is converted to a more stringent q-value, which represents the false discovery rate. Another method was proposed by Benjamini and Hochberg, (1995), which controls the FDR, the rate of expected proportion of errors among the rejected hypotheses and is more stringent than the q-value method. As suggested by Glynn, *et al.* (2006), the Lomb-Scargle periodogram uses this method to get FDR values. In our study, we consider genes with an FDR value < 0.05 to be rhythmically expressed.

2.5 Simulation data

In our numerical experiment, we prepared periodic and non-periodic data sets to test the effectiveness of LSPR. To generate the periodic time-series, we applied two models to simulate the expression data regulated by periodic genes. The stationary model representing an "ideal" gene expression is defined as

$$x_i = \mathrm{SNR} \cdot 2\cos(\frac{2\pi}{T} t_i - \varphi) + \varepsilon_i \quad (5)$$

where SNR is signal-to-noise ratio; T is period; φ is phase; and ε_i is a normally distributed $(\mu = 0, \delta = 50)$ noise term.

The second is non-stationary model which considers the dampening effect of free-running rhythms. It is defined as

$$x_i = 500 \cdot e^{-0.01 t_i} + \mathrm{SNR} \cdot 100 \cdot e^{-0.01 t_i} \cdot \cos(\frac{2\pi}{T} t_i - \varphi) + \varepsilon_i \quad (6)$$

where ε_i is normally distributed $(\mu = 0, \delta = 50)$ noise; and the mean level and amplitude exponentially decay over time. t_i are unevenly sampled at 0 h, 1 h, 2 h, 4 h, 8 h, 12 h, 13 h, 14 h, 16 h, 20 h with one replicate for each time point, respectively. By varying SNR, T and φ in the above two models, multiple periodic time-series are generated. Non-periodic signals are generated from two random processes: the standard normally distributed white noise and the AR(1) model defined by

$$x_i = c + \alpha x_{i-1} + \varepsilon_i \quad (7)$$

where α is the AR coefficient, c is a constant and ε_i is white noise.

3 RESULTS

3.1 Statistical detection of periodicity for short time-series

In this study, we analyzed the temporal microarray data to search genes that are periodically expressed. We developed the LSPR algorithm based on the Lomb-Scargle periodogram. Glynn, *et al.* (2006) applied the Lomb-Scargle periodogram to analyze unevenly sampled gene expression data. The authors generated a set of simulated time-series and gave the estimated relationship between the number of time-points, N, and the statistical significance of periodicity, p-value, as follows:

$$N \approx 5[1 - \log_{10}(p\text{-value})] \quad (8)$$

According to Equation (8), for a statistically significant rhythmic gene under the p-value threshold 0.05, the number of time-points of the expression profiles will need to be more than 20. The diurnal and circadian time course studies are usually designed to collect data every 4 hours over a course of 48 hours, generating expression profiles with 12 or 13 time-points (Yamada and Ueda, 2007). Such short time-series make Lomb-Scargle periodogram fail to detect the periodic gene expression pattern.

LSPR overcomes the limitation of Lomb-Scargle method by evaluating the p-value using harmonic regression fitting. Similar estimation between N and p-value was conducted for LSPR, and the relationship between these quantities is as follows:

$$N \approx 5.6 - \log_{10}(p\text{-value}) \quad (9)$$

According to Equation (9), LSPR can identify a periodically expressed gene as significant under a p-value less than 0.05 from a time-series with only 7 time-points.

Moreover, we applied LSPR, COSOPT, Lomb-Scargle periodogram and ARSER to analyze a set of short periodic time-series (Table 1). This data set was generated by Michael, *et al.* (2008) with five periodic patterns based on studies available in the literature. The data set contains 120 time series in total, with 24 samples for each periodic pattern, respectively. We found that Lomb-Scargle identified none of the periodic signals under an FDR threshold of 0.05. The test time-series contained 12 time-points, which led to the p-value of the Lomb-Scargle periodogram being not significant for such short time-series. COSOPT detected the periodicity by curve fitting, which measures the goodness-of-fit between experimental data and a series of cosine curves with varying phases and period lengths. p-values are then calculated by scrambling the experimental data and re-fitting it to cosine curves in order to determine the probability that the observed data matches a cosine curve by chance. However, COSOPT is a sinusoidal-based fitting algorithm, which limited its power in detecting periodicity for non-sinusoidal waveforms. Table 1 shows that COSOPT identified less periodic time-series with spike and box-like patterns than LSPR, while LSPR and ARSER gave the best prediction for this test data set.

Table 1. Number of inferred periodic samples from short time-series with multiple periodic patterns

Method	Waveforms					
	Spike	Box1	box2	rigid	sine	random
Lomb-Scargle	0	0	0	0	0	0
COSOPT*	7	24	7	24	24	1
ARSER*	16	24	16	24	24	1
LSPR	16	24	16	24	24	0

*note: The data generated by the COSOPT and ARSER algorithms are from Yang and Su, (2010)

Since LSPR and ARSER have a similar algorithm design, we directly compared them with our generated unevenly sampled data (see Method section for details), and the ROC curve is shown in Figure S1.

Cubic spline interpolation was used to generate 2-h evenly spaced data sets from the unevenly sampled data before input into ARSER. The results showed LSPR performed better on unevenly sampled data than ARSER with interpolation.

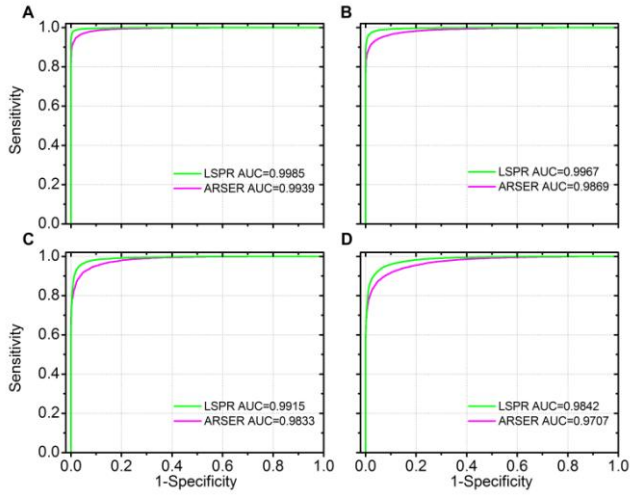

Fig. S1. ROC analysis of LSPR and ARSER on unevenly sampled datasets: (A) 10000 stationary periodic signals and 10000 white noise, (B) 10000 non-stationary periodic signals and 10000 white noise, (C) 10000 stationary periodic signals and 10000 AR(1)-based signals and (D) 10000 non-stationary periodic signals and 10000 AR(1)-based signals. We employed cubic splines to interpolate the testing sets into a two-hour sample interval before input into ARSER.

3.2 Analysis of cell cycle periodic patterns in yeast expression data

Generally, periodicity detection algorithms are applied to circadian rhythm and cell cycle expression data. We applied LSPR to analyze Yeast cell cycle expression data generated by Spellman, *et al.* (1998). In our benchmark, we included three computational methods: LSPR, the Lomb-Scargle periodogram and the spectral estimation method proposed by Liew, *et al.* (2007). As suggested by the study of Lichtenberg *et al.* (2005), we set a benchmark for these methods by applying them to three Yeast experimental data sets: alpha, cdc15 and cdc28. For Alpha and cdc15 experiments, the raw data were obtained from http://genome-www.stanford.edu/cellcycle/. For the cdc28 experiment, renormalized data were generated by Lichtenberg *et al.* (2005). The Alpha experiment harvested samples at 7 min intervals for 119 min with a total of 18 time-points. Cdc15 data were measured every 10 min for 290 min, lacking observations for 20, 40, 60, 260 and 280 min time-points, giving a total 24 time-points. For the cdc28 experiment, samples were taken every 10 min for 160 min with a total of 17 time-points. Alpha and cdc28 were treated as having evenly sampled data, while cdc15 was taken to have unevenly sampled data. In our analysis, we removed the genes with missing values for all sample points and obtained 4489 genes for alpha, 4381 genes for cdc15 and 6214 genes for cdc28.

To measure the performance of different algorithms, de Lichtenberg, *et al.* (2005) proposed three benchmark sets B1, B2 and B3. B1 contains a total of 113 genes previously identified as periodically expressed in small-scale experiments. B2 contains 352 genes whose promoters were bound by at least one of the 9 known cell cycle transcription factors in two Chromatin IP studies, and therefore many of the genes in this benchmark set should be expected to be cell cycle regulated. B3 contains 518 genes annotated in MIPS (Mewes, *et al.*, 2002) as "cell cycle and DNA processing". However, since a large number of genes involved in the cell cycle are not subjected to transcriptional regulation (not periodic) and genes found in B1 were explicitly removed, only a small fraction of the genes in B3 are expected to be periodically expressed.

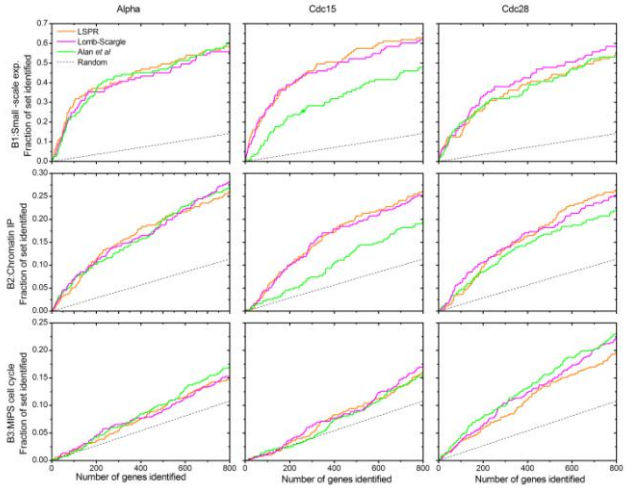

Fig. S2. Comparison of periodicity detection algorithms in Yeast cell cycle data. The fraction of benchmark genes contained in the top 800 ranked list is plotted for each algorithm, benchmark set (B1, B2, B3) and experiment (alpha, cdc15, cdc28), respectively. The methods are coloured as follows: LSPR (our method, orange), Lomb-Scargle (magenta) and Alan *et al.* (green). Each ranked list is formed by sorting the *p*-values of gene profiles in ascending order. A random performance could be observed in the black dot.

Figure S2 shows the performance of each method for the three Yeast cell cycle data sets. There is no single method that outperforms the others across all benchmark sets in all experiments. Ap-

parently, all methods perform significantly better than random. The LSPR and Lomb-Scargle methods give similar performances for cell cycle data, which is in contrast to the results of analyzing circadian expression data. This demonstrates that the Lomb-Scargle periodogram is more efficient to analyze cell cycle expression data other than circadian expression data, since the cell cycle experiments have more sampling time points. According to the comparisons of the three methods used to analyze cdc15 and cdc28, both the LSPR and Lomb-Scargle methods show better performance than the method proposed by Alan and his co-workers.

4 CONCLUSIONS

We have designed the LSPR algorithm for periodicity detection, which is a three-step algorithm based on the Lomb-Scargle periodogram and harmonic regression. LSPR is designed to identify periodically expressed genes from unevenly sampled time-course expression profiles with short lengths. When comparing LSPR with well-established algorithms for well-defined synthetic data and for microarray data for diurnal and cell cycle experiments, the results illustrated that LSPR was superior in identifying periodic patterns as well as in its robustness to short time-series.

REFERENCES

Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society Series B-Methodological*, **57**, 289-300.

de Lichtenberg, U., *et al.* (2005) Comparison of computational methods for the identification of cell cycle-regulated genes, *Bioinformatics*, **21**, 1164-1171.

Glynn, E.F., Chen, J. and Mushegian, A.R. (2006) Detecting periodic patterns in unevenly spaced gene expression time series using Lomb-Scargle periodograms, *Bioinformatics*, **22**, 310-316.

Liew, A.W., *et al.* (2007) Spectral estimation in unevenly sampled space of periodically expressed microarray time series data, *BMC Bioinformatics*, **8**, 137.

Lomb, N.R. (1976) Least-Squares Frequency-Analysis of Unequally Spaced Data, *Astrophysics and Space Science*, **39**, 447-462.

Mewes, H.W., *et al.* (2002) MIPS: a database for genomes and protein sequences, *Nucleic Acids Res*, **30**, 31-34.

Michael, T.P., *et al.* (2008) Network discovery pipeline elucidates conserved time-of-day-specific cis-regulatory modules, *PLoS Genet*, **4**, e14.

Scargle, J.D. (1982) Studies in Astronomical Time-Series Analysis .2. Statistical Aspects of Spectral-Analysis of Unevenly Spaced Data, *Astrophysical Journal*, **263**, 835-853.

Spellman, P.T., *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, *Mol Biol Cell*, **9**, 3273-3297.

Storey, J.D. and Tibshirani, R. (2003) Statistical significance for genomewide studies, *Proc Natl Acad Sci U S A*, **100**, 9440-9445.

Straume, M. (2004) DNA microarray time series analysis: automated statistical assessment of circadian rhythms in gene expression patterning, *Methods Enzymol*, **383**, 149-166.

Yamada, R. and Ueda, H.R. (2007) Microarrays: statistical methods for circadian rhythms, *Methods Mol Biol*, **362**, 245-264.

Yang, R. and Su, Z. (2010) Analyzing circadian expression data by harmonic regression based on autoregressive spectral estimation, *Bioinformatics*, **26**, i168-i174.