BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 7 2011, pages 1023–1025
doi:10.1093/bioinformatics/btr041

Gene expression

Advance Access publication February 3, 2011

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

Associate Editor: Trey Ideker

ABSTRACT

Summary: We propose a three-step periodicity detection algorithm named LSPR. Our method first preprocesses the raw time-series by removing the linear trend and filtering noise. In the second step, LSPR employs a Lomb–Scargle periodogram to estimate the periodicity in the time-series. Finally, harmonic regression is applied to model the cyclic components. Inferred periodic transcripts are selected by a false discovery rate procedure. We have applied LSPR to unevenly sampled synthetic data and two *Arabidopsis* diurnal expression datasets, and compared its performance with the existing well-established algorithms. Results show that LSPR is capable of identifying periodic transcripts more accurately than existing algorithms.

Availability: LSPR algorithm is implemented as MATLAB software and is available at http://bioinformatics.cau.edu.cn/LSPR

Contact: zhensu@cau.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 5, 2010; revised on December 25, 2010; accepted on January 15, 2011

1 INTRODUCTION

Most algorithms for analyzing cell cycle or circadian rhythm temporal expression data require that the input signals are equally sampled. For irregularly sampled time-series, the Lomb–Scargle periodogram has been proposed to analyze time-course gene expression data (Glynn *et al.*, 2006). A comparison study of periodicity detection methods for irregularly sampled data concluded that the Lomb–Scargle method performed better than most existing methods (Zhao *et al.*, 2008). However, this method is subject to noise and is not powerful for short time-series.

In this study, we propose a new periodicity identification algorithm based on the Lomb–Scargle periodogram and harmonic regression method for unevenly sampled time-series. Our algorithm, named LSPR, has a similar procedure to that of the ARSER algorithm, which was proposed to analyze evenly sampled temporal expression profiles (Yang and Su, 2010). The main difference is that LSPR uses spectrum analysis for unevenly sampled data introduced by Lomb and additionally elaborated by Scargle, while ARSER employs autoregressive spectral estimation, which can only analyze evenly sampled data.

For a given irregularly sampled time-series, LSPR first estimates the period by the Lomb–Scargle periodogram in the frequency domain, and then models the periodic signals by the harmonic regression method in the time domain. Such a joint strategy overcomes the limitations of the Lomb–Scargle periodogram and gives better descriptions of periodic patterns. By applying LSPR to the analysis of synthetic data and *Arabidopsis* diurnal expression data, we found our method was more powerful in detecting periodic transcripts compared with two well-established periodicity detection algorithms.

2 METHODS

2.1 Overview

Broadly, LSPR is a three-step integrated algorithm, including data preprocessing, spectral analysis and harmonic regression. Data preprocessing entails detrending and smoothing procedures. Detrending removes the linear trend in the raw time-series using ordinary least squares (OLS) regression. This process can remove the effects of accumulating data from a trend, to show only the absolute changes in values and to allow potential cyclical patterns to be identified. The detrended time-series are input into a spectral analysis and harmonic regression procedure. Including noise may influence the ability of the spectral analysis to predict the exact periods, so a fourth degree Savitzky–Golay filter (Savitzky and Golay, 1964) is used to smooth the detrended time-series, before the spectral analysis is used as another input to evaluate periods. In the spectral analysis step, LSPR estimates the spectrum by the Lomb–Scargle periodogram (Lomb, 1976; Scargle, 1982), which enables the extraction of periodic components from unevenly sampled time-series. Finally, to validate the statistical significance of identified periodicity from the spectral analysis, LSPR employs harmonic regression to model the cyclical components. The harmonic regression model fits the detrended time-series with sinusoids and gives predictions for amplitude, phase, mean value and *P*-value. The *P*-values calculated for each time-series are adjusted for multiple testing corrections under two alternative false discovery rate (FDR) procedures (Benjamini and Hochberg, 1995; Storey and Tibshirani, 2003) (see Supplementary Material for details).

2.2 Procedure of the LSPR algorithm

LSPR identifies periodicity using the following step-by-step procedures for an input time-series $\{x_i\}$:

(1) Remove the linear trend in the time-series $\{x_i\}$, denoting the detrended time-series as $\{\dot{x}_i\}$.

*To whom correspondence should be addressed.

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

(2) Smooth $\{\dot{x}_i\}$ by a fourth-order Savitzky–Golay algorithm. The smoothed time-series is denoted as $\{\ddot{x}_i\}$.

(3) Calculate the Lomb–Scargle periodogram of $\{\ddot{x}_i\}$ (Supplementary Equation 1). Then select all periods $\{\ddot{T}_j\} \in [T_{\text{start}}, T_{\text{end}}]$ that show peaks in the spectrum. T_{start} and T_{end} are the start and end of the selected period range.

(4) Calculate the Lomb–Scargle periodogram of detrended time-series $\{\dot{x}_i\}$ (Supplementary Equation 1), and select all periods $\{\dot{T}_j\} \in [T_{\text{start}}, T_{\text{end}}]$ that show peaks in the spectrum.

(5) The periods $\{\dot{T}_j\}$ and $\{\ddot{T}_j\}$ are chosen as input for the harmonic regression for $\{\dot{x}_i\}$ (Supplementary Equation 3).

(6) Use Akaike's information criterion (Akaike, 1974) to determine the best harmonic regression model among the models generated in Step (5) and give outputs of period, amplitude, phase and P-value for $\{x_i\}$. Users can use P-values to identify periodic patterns.

(7) For large-scale time-course data, FDR values are used to determine periodicity. In this study, genes with FDR values <0.05 are considered as periodic genes.

3 RESULTS

3.1 Detecting periodicity in synthetic data

To test the LSPR algorithm and compare its performance with prior methods, we prepared comprehensive datasets containing periodic and non-periodic samples (see Supplementary Material). The periodic samples were generated by two models. One is the stationary periodic model defined by a cosine curve with constant amplitude and mean value. The stationary model was widely used to generate synthetic data in previous studies (Liew *et al*., 2007; Ptitsyn *et al*., 2006; Wichert *et al*., 2004). Considering the dampening effect of the circadian rhythm (Westermark *et al*., 2009), the periodic data were also generated by another non-stationary model defined by a cosine curve with exponentially damped amplitude and mean value. Compared with the stationary model, the non-stationary model is more likely to approximate the natural biological rhythm (Refinetti, 2004).

The non-periodic data in the synthetic datasets were generated by two random processes. One was white noise following a standard normal distribution. Another widely accepted stochastic process in time-series is the autoregressive model considering the general correlation between successive measurements. Here, we used an autoregressive process of order one (AR(1)) to generate the random time-series.

In our synthetic datasets, the periodic signals include 10 000 stationary and 10 000 non-stationary samples with varied period, phase and signal-to-noise ratio (SNR). The non-periodic signals include 10 000 white noise signals and 10 000 AR(1) signals. Each time-series possessed 20 unevenly sampled time-points over the course of 2 days. This sampling design is consistent with the method adopted by Smith *et al*. (2004).

Four datasets were constructed by combining the periodic and non-periodic signals. We applied LSPR, the Lomb–Scargle periodogram and COSOPT to analyze these four datasets. Since the task is to separate periodic signals from non-periodic signals, we can treat it as a binary classification problem. This allows receiver operating characteristic (ROC) analysis to be conducted. Figure 1 shows the ROC analysis for the three algorithms according to their determination thresholds (FDR q-value for LSPR and Lomb–Scargle, pMMC-β for COSOPT). We found LSPR performed the best of the three algorithms in all the cases.

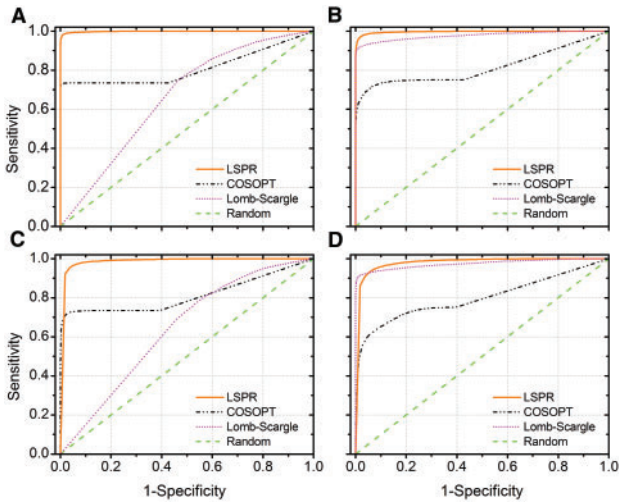

Fig. 1. ROC analysis of LSPR, COSOPT and the Lomb–Scargle periodogram. Testing datasets contains: (**A**) 10 000 stationary periodic signals and 10 000 white noise, (**B**) 10 000 non-stationary periodic signals and 10 000 white noise, (**C**) 10 000 stationary periodic signals and 10 000 AR(1)-based signals and (**D**) 10 000 non-stationary periodic signals and 10 000 AR(1)-based signals. A greater area under the ROC curve means more accuracy for prediction. LSPR gives the best performance of all the three algorithms.

3.2 Application of LSPR in *Arabidopsis* diurnal expression data

Here, we applied LSPR to analyze two independent datasets from the studies of the diurnal gene expression of a model plant, *Arabidopsis*. These datasets were named after their respective first author: Smith data (Smith *et al*., 2004) and Blasing data (Blasing *et al*., 2005). Smith data were unevenly sampled at 11 time-points with 2 biological replicates for each. Blasing data were evenly sampled at six time-points with 4 h intervals and three biological replicates.

We also applied the Lomb–Scargle periodogram and COSOPT, a widely used algorithm for analyzing circadian or diurnal expression data (Straume, 2004) to analyze the same datasets as those processed with LSPR. We found the Lomb–Scargle method identified zero transcripts as rhythmic in both Smith and Blasing data at an FDR cutoff of 0.05. Thus, we studied the results given by COSOPT and LSPR. Of all the 22 810 *Arabidopsis* genes, we found LSPR identified 7851 transcripts (35% of the complete *Arabidopsis* genome) as rhythmic genes in the Blasing data and 6709 transcripts (30% of the complete *Arabidopsis* genome) in the Smith data with an FDR cutoff of 0.05 (Fig. 2A). These fractions of clock-regulated genes were consistent with an estimate of between 31% and 41% of expressed genes being circadian regulated reported by a recent study (Covington *et al*., 2008). The overlap of identified rhythmically expressed genes between the Smith and Blasing data is 19% of all expressed genes, which is larger than the 13% obtained by COSOPT (Fig. 2A). Moreover, we found the set of 3602 genes present in Blasing but absent in Smith, according to LSPR, has significant overlap with the set of 2832 genes present in Blasing but absent in

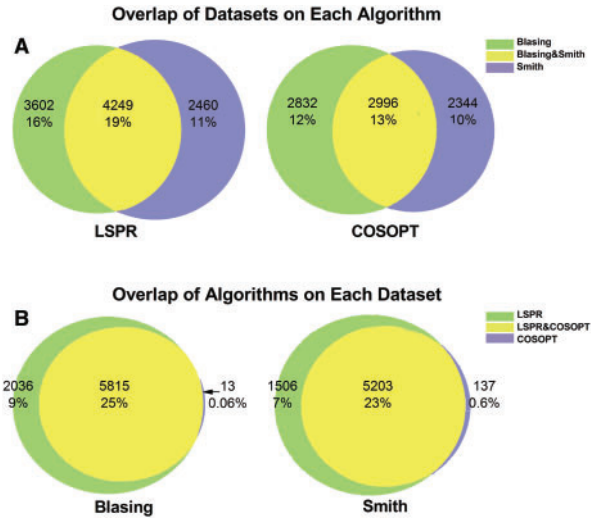

Fig. 2. Comparisons of periodicity detecting algorithms applied to two *Arabidopsis* datasets. (**A**) Overlap of periodic transcripts identified by LSPR (left) and COSOPT (right) for the Blasing and Smith data. (**B**) Overlap of periodic transcripts in the Blasing data (left) and Smith data (right), identified by LSPR and COSOPT.

Smith according to COSOPT (55% overlap, 2284 genes in common). Similarly, the set of identified genes present in Smith but absent in Blasing as found by LSPR also has significant overlap with the set of identified genes present in Smith but absent in Blasing as found by COSOPT (54% overlap, 1676 genes in common).

Figure 2B shows LSPR identified 99% and 97% of the rhythmic genes identified by COSOPT in the Blasing and Smith data, respectively. In addition, LSPR newly identified 9% and 7% of the complete genome transcripts as rhythmic genes in the Blasing and Smith data, respectively. Also, results can be compared with a benchmark set of 28 known clock-regulated genes reported in previous studies (Dodd *et al*., 2007; Pruneda-Paz, 2009). Two of these genes were found among the newly identified genes found by LSPR in the Blasing data. One is ZEITLUPE (ZTL), which encodes clock-associated PAS protein (Somers *et al*., 2000). The other is PHYTOCHROME B (PHYB), an element in the input of the cytokinin signal to the circadian phase (Hanano *et al*., 2006).

ACKNOWLEDGEMENTS

We thank Daofeng Li for web server assistance, and Wenying Xu and Yi Ling for helpful advice and discussions.

Funding: Ministry of Science and Technology of China (2006CB100105); College Student Research and Career-creation Program of Beijing (2010).

Conflict of Interest: none declared.

REFERENCES

Akaike,H. (1974) A new look at the statistical model identification. *IEEE Trans. Automat. Contr.*, **19**, 713–723.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.*, **57**, 289–300.

Blasing,O. *et al.* (2005) Sugars and circadian regulation make major contributions to the global regulation of diurnal gene expression in Arabidopsis. *Plant Cell*, **17**, 3257–3281.

Covington,M.F. *et al.* (2008) Global transcriptome analysis reveals circadian regulation of key pathways in plant growth and development. *Genome Biol.*, **9**, R130.

Dodd,A.N. *et al.* (2007) The Arabidopsis circadian clock incorporates a cADPR-based feedback loop. *Science*, **318**, 1789–1792.

Glynn,E.F. *et al.* (2006) Detecting periodic patterns in unevenly spaced gene expression time series using Lomb-Scargle periodograms. *Bioinformatics*, **22**, 310–316.

Hanano,S. *et al.* (2006) Multiple phytohormones influence distinct parameters of the plant circadian clock, *Genes Cells*, **11**, 1381–1392.

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 Space Sci.*, **39**, 447–462.

Pruneda-Paz,J.L. *et al.* (2009) A functional genomics approach reveals CHE as a component of the Arabidopsis circadian clock. *Science*, **323**, 1481–1485.

Ptitsyn,A.A. *et al.* (2006) Permutation test for periodicity in short time series data. *BMC Bioinformatics*, **7** (Suppl. 2), S10.

Refinetti,R. (2004) Non-stationary time series and the robustness of circadian rhythms. *J. Theor. Biol.*, **227**, 571–581.

Savitzky,A. and Golay,M. (1964) Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.*, **36**, 1627–1639.

Scargle,J.D. (1982) Studies in astronomical time-series analysis. 2. Statistical aspects of spectral-analysis of unevenly spaced data. *Astrophys. J.*, **263**, 835–853.

Smith,S.M. *et al.* (2004) Diurnal changes in the transcriptome encoding enzymes of starch metabolism provide evidence for both transcriptional and posttranscriptional regulation of starch metabolism in Arabidopsis leaves. *Plant Physiol.*, **136**, 2687–2699.

Somers,D.E. *et al.* (2000) ZEITLUPE encodes a novel clock-associated PAS protein from Arabidopsis. *Cell*, **101**, 319–329.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **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.

Westermark,P.O. *et al.* (2009) Quantification of circadian rhythms in single cells. *PLoS Comput. Biol.*, **5**, e1000580.

Wichert,S. *et al.* (2004) Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics*, **20**, 5–20.

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

Zhao,W. *et al.* (2008) Detecting periodic genes from irregularly sampled gene expressions: a comparison study. *EURASIP J. Bioinform. Syst. Biol.*, 769293.