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

Rendong Yang¹†, Chen Zhang²† and Zhen Su¹,*

¹Division of Bioinformatics, State Key Laboratory of Plant Physiology and Biochemistry, College of Biological Sciences, China Agricultural University, Beijing 100193 and ²Department 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 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.cauc.edu.cn/LSPR

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 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 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 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.

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 \( \{\hat{x}_i\} \).

Received on November 5, 2010; revised on December 25, 2010; accepted on January 15, 2011
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; Pitsyn 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 (Reffetti, 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 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 un evenly 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 p-value for COSOPT). We found LSPR performed the best of the three algorithms in all the cases.

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

(2) Smooth \( \{x_i\} \) by a fourth-order Savitzky–Golay algorithm. The smooth function is denoted as \( \{\hat{x}_i\} \).

(3) Calculate the Lomb–Scargle periodogram of \( \{\hat{x}_i\} \) (Supplementary Equation 1). Then select all periods \( \{T_i\} \in \{T_{\text{min}}, T_{\text{max}}\} \) that show peaks in the spectrum. \( T_{\text{min}} \) and \( T_{\text{max}} \) are the start and end of the selected period range.

(4) Calculate the Lomb–Scargle periodogram of detrended time-series \( \{x_i\} \) (Supplementary Equation 1), and select all periods \( \{T_i\} \in \{T_{\text{min}}, T_{\text{max}}\} \) that show peaks in the spectrum.

(5) The periods \( \{T_i\} \) and \( \{\hat{T}_i\} \) are chosen as input for the harmonic regression for \( \{\hat{x}_i\} \) (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 \( \{\hat{x}_i\} \). 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.

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

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**


