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

Rendong Yang, Chen Zhang and Zhen Su

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

ABSTRACT

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

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 evaluate 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 \( x(t) \):

1. Remove the linear trend in the time-series \( x(t) \), denoting the detrended time-series as \( x_d(t) \).
To test the LSPR algorithm and compare its performance with their determination thresholds (FDR receiver operating characteristic (ROC) analysis to be conducted. we can treat it as a binary classification problem. This allows the task is to separate periodic signals from non-periodic signals, periodogram and COSOPT to analyze these four datasets. Since adopted by Smith course of 2 days. This sampling design is consistent with the method time-series possessed 20 unevenly sampled time-points over the time-series. In our synthetic datasets, the periodic signals include 10 000 stationary and 10 000 non-stationary samples with varied period, time-series is the autoregressive model considering the general normal distribution. Another widely accepted stochastic process in time-series is the autoregressive model considering the general normal distribution. These datasets were named after their respective first author: Smith 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.

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 signals (see Supplementary Material). The periodic signals 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; Pittsyn 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 LSPR–Scargle, pMMC–β for COSOPT). We found LSPR performed the best of the three algorithms in all the cases.

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

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