BIOINFORMATICS ORIGINAL PAPER

Vol. 22 no. 3 2006, pages 310–316
doi:10.1093/bioinformatics/bti789

Gene expression

Detecting periodic patterns in unevenly spaced gene expression time series using Lomb–Scargle periodograms

Earl F. Glynn[1], Jie Chen[1,2,*] and Arcady R. Mushegian[1,3]

[1]Stowers Institute for Medical Research, 1000 East 50th Street, Kansas City, MO 64110, USA, [2]Department of Mathematics and Statistics, University of Missouri-Kansas City, 5100 Rockhill Road, Kansas City, MO 64110, USA and [3]Department of Microbiology, Immunology and Molecular Genetics, University of Kansas Medical Center, Kansas City, KS 66160, USA

Received on June 13, 2005; revised on October 26, 2005; accepted on November 15, 2005
Advance Access publication November 22, 2005
Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Periodic patterns in time series resulting from biological experiments are of great interest. The commonly used Fast Fourier Transform (FFT) algorithm is applicable only when data are evenly spaced and when no values are missing, which is not always the case in high-throughput measurements. The choice of statistic to evaluate the significance of the periodic patterns for unevenly spaced gene expression time series has not been well substantiated.

Methods: The Lomb–Scargle periodogram approach is used to search time series of gene expression to quantify the periodic behavior of every gene represented on the DNA array. The Lomb–Scargle periodogram analysis provides a direct method to treat missing values and unevenly spaced time points. We propose the combination of a Lomb–Scargle test statistic for periodicity and a multiple hypothesis testing procedure with controlled false discovery rate to detect significant periodic gene expression patterns.

Results: We analyzed the *Plasmodium falciparum* gene expression dataset. In the Quality Control Dataset of 5080 expression patterns, we found 4112 periodic probes. In addition, we identified 243 probes with periodic expression in the Complete Dataset, which could not be examined in the original study by the FFT analysis due to an excessive number of missing values. While most periodic genes had a period of 48 h, some had a period close to 24 h. Our approach should be applicable for detection and quantification of periodic patterns in any unevenly spaced gene expression time-series data.

Availability: The computations were performed in R. The R code is available from http://research.stowers-institute.org/efg/2005/LombScargle

Contact: chenj@umkc.edu

Supplementary information: The online supplement is available at http://research.stowers-institute.org/efg/2005/LombScargle

INTRODUCTION

Rhythmic processes occur at all levels of biological organization with periods ranging from less than a second to years (Goldbeter, 2002). Time-series experiments are a common way to study rhythmic processes, and inherent periodicity in such data is indicative of the underlying 'clocks'. Examples of biological rhythms include cell division (Mitchison, 2003), circadian rhythms (Crosthwaite, 2004; Prolo *et al*., 2005), morphogenesis of periodic structures, such as somites in vertebrates (Dale *et al*., 2003), complex life cycles of some microorganisms (Lakin-Thomas, 2004; Rovery *et al*., 2005) and many others.

With the help of gene expression technology, biologists can study the mechanisms that control a particular biological rhythm more closely. For instance, to study how the major oscillator in the suprachiasmatic nuclei (SCN) and in the liver regulates behavioral and physiological rhythms in the whole organism, Panda *et al*. (2002) used high-density oligonucleotide arrays to measure gene expression in the mouse tissue samples taken every 4 h during two complete circadian cycles and applied a cosine wave-fitting algorithm (Harmer *et al*., 2000) to identify clusters of circadian-regulated genes among more than 7000 genes. They found that about 650 cycling transcripts were under circadian regulation specific to either the SCN or the liver.

As with many other types of high-dimensional data, the choice of algorithm and statistic to identify significantly periodic patterns of gene expression is a challenge (Wichert *et al*., 2004). Zhao *et al*. (2001) used a single-pulse model (SPM) to identify periodic transcripts in the *Saccharomyces cerevisiae* yeast microarray data based on the assumption that the cell-cycle-regulated transcripts will peak only once per cycle. Bar-Joseph *et al*. (2003) developed an algorithm to represent time series of gene expression as continuous curves using a cubic spline method and used this algorithm to estimate missing values in time series. Langmead *et al*. (2003) proposed an algorithm that uses autocorrelation to perform linear-time search in frequency and phase, and then use the undirected Hausdroff distance as a similarity measure to cluster genes of similar cyclic patterns together. Johansson *et al*. (2003) used a multivariate partial least squares regression model to identify cell cycle-regulated genes in the *S.cerevisiae* yeast data. Luan and Li (2004) proposed to use the shape-invariant model of Lawton *et al*. (1972) and Wang and Brown (1996) combined with a B-spline estimation to model periodic gene expression profiles. Luan and Li (2004) also applied their approach to several publicly available microarray data including the *S.cerevisiae* yeast data.

A common computational technique to study periodic data is the Fast Fourier Transform (FFT) algorithm (Priestley, 1981), which finds periodicities by searching for sharp peaks in the ordinary periodograms calculated from the Fourier transform of the time

[*]To whom correspondence should be addressed.

series (Durbin, 1967). For example, Spellman *et al*. (1998) studied cell cycle-regulated genes in *S.cerevisiae* gene expression data using Fourier analysis and found about 800 periodic genes. Straume (2004) compared four algorithms used in circadian gene expression analysis and pointed out that FFT performs well when time points are evenly spaced. Wichert *et al*. (2004) defined an average periodogram based on the ordinary periodograms or classical periodograms (Priestley, 1981; Scargle, 1982) and used it as a graphical tool for finding possible periodic signals in yeast and human cells.

One limitation of the FFT algorithm and the ordinary periodograms is that it requires evenly spaced time series (Priestley, 1981). Time series produced by biological experiments are, however, often unevenly spaced, for a variety of reasons, many of which have to do with the limitations of the instruments and with inherent experimental constraints on biological samples. Another limitation of the FFT algorithm is that it does not tolerate missing values. When missing values in the time series are present, data must be imputed prior to the application of the FFT algorithm. The effects of data imputation and the optimal way to do so, however, are generally not known. Moreover, FFT scores do not directly address the significance of the observed periodicity and have to be validated by a heuristic cutoff score or permutation studies.

These limitations can be overcome using the statistical approach first introduced in astrophysics, the Lomb–Scargle periodogram. When studying variable stars in astronomy, Lomb (1976) sought a way to find periodicities in unevenly spaced data. Astronomers could not always control viewing times, telescope availability and the position of an object in the sky—all of which is reminiscent of similar experimental problems in biology. In an attempt to find an alternative to imputing pseudo-data in sinusoidal models, Lomb (1976) proposed to use least-squares fits to sinusoidal curves. Scargle (1982) extended Lomb's work by defining the Lomb–Scargle periodogram and by deriving the null distribution for it. The *p*-value of the Lomb–Scargle periodogram can be obtained using these results. Horne and Baliunas (1986) noted a particular standardization of the periodogram that resulted in known statistical properties. Press and Rybicki (1989) proposed a practical mathematical formulation, later implemented in C (Press *et al*., 2002).

Ruf (1999) was one of the first to use Lomb–Scargle periodograms to analyze biological data. Ruf applied the technique to telemetric temperature data of an alpine marmot over more than 12 days prior to hibernation and detected a circadian rhythm with a period of 24 h. Van Dongen *et al*. (1999) analyzed the data of human oral temperatures in search of a circadian rhythm. Periodicity of 24 h was successfully identified in the unevenly spaced time series using the Lomb–Scargle method. In their attempt to detect rhythmic components in the circadian cycle of the Crassulacean acid metabolism plants, Bohn *et al*. (2003) used Lomb–Scargle method for periodogram estimation.

In this paper, we propose to use the Lomb–Scargle periodogram to search for periodic patterns in unevenly spaced time series that represent gene expression profiles. As gene expression data have typically large dimensionality (hundreds to thousands of profiles), searching for all periodic genes under such conditions requires statistical multiple hypothesis testing, which can be achieved using the multiple comparison procedure that controls the false discovery rate (FDR) (Benjamini and Hochberg, 1995). Our main example is gene expression in the malaria parasite, recorded by Bozdech *et al*. (2003). The techniques and the software that we developed, nevertheless, are readily applicable to periodic pattern discovery in any time series.

METHODS

Lomb–Scargle periodogram

For a gene g and expression level observed at time t_i, we denote the time series by $Y_g(t_i)$ for $i = 1, \ldots, N$ and $g = 1, \ldots, \mathcal{G}$. To model $Y_g(t_i)$ for periodicity, we assume

$$Y_g(t_i) = \eta_g(t_i) + \varepsilon_g(t_i),$$

where $\eta_g(t_i)$ is a periodic function with a smallest positive period T_g for gene g, i.e. $\eta_g(t_i + T_g) = \eta_g(t_i)$ for all t_i; and $\varepsilon_g(t_i)$ is assumed to be a sequence of non-observable normal random errors with mean 0 and homogenous variance σ^2 for all g and t_i (Scargle, 1982). Let the expression of gene g at time t_i be $y_g(t_i)$, and the average gene expression for gene g be

$$\bar{y}_g = \frac{1}{N}\sum_{i=1}^{N} y_g(t_i),$$

then the error variance can be estimated by the sample variance as follows:

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}[y_g(t_i) - \bar{y}_g]^2.$$

Unlike Fourier analysis, in which the Fourier frequencies are used, we assume that there are M test frequencies, $f_1, f_2, \ldots, f_M$ and their corresponding angular frequencies are $\omega_j = 2\pi f_j$, for $j = 1, \ldots, M$.

The Lomb–Scargle periodogram is defined in Press and Rybicki (1989) as

$$P_g(\omega_j) = \frac{1}{2\hat{\sigma}^2}\left\{\frac{\left(\sum_{i=1}^{N}[y_g(t_i) - \bar{y}_g]\cos[\omega_j(t_i - \tau)]\right)^2}{\sum_{i=1}^{N}\cos^2[\omega_j(t_i - \tau)]} + \frac{\left(\sum_{i=1}^{N}[y_g(t_i) - \bar{y}_g]\sin[\omega_j(t_i - \tau)]\right)^2}{\sum_{i=1}^{N}\sin^2[\omega_j(t_i - \tau)]}\right\} \tag{1}$$

for $j = 1, \ldots, M$, where τ is defined by

$$\tan(2\omega_j\tau) = \frac{\sum_{i=1}^{N}\sin(2\omega_j t_i)}{\sum_{i=1}^{N}\cos(2\omega_j t_i)}.$$

The choice of M depends on the number of independent frequencies, N_0 (Press *et al*., 2002). Horne and Baliunas (1986) performed extensive Monte Carlo simulations to investigate the relationship between M and N_0. They gave a simple least squares formula to estimate the number of independent frequencies N_0 from the number of observations, N, in a time series:

$$N_0 \approx -6.362 + 1.193N + 0.00098N^2.$$

This empirical deterministic formula is adequate for most purposes of choosing M by actually taking M as N_0 (Press *et al*., 2002).

Statistical hypothesis testing for periodicity using Lomb–Scargle periodogram

Scargle (1982) showed that the null distribution of the Lomb–Scargle periodgram $Z_j = P_g(\omega_j)$ at a given frequency ω_j is exponentially distributed, i.e. the cumulative distribution function (CDF) of Z_j is

$$\begin{aligned} F(z) &= \Pr[Z_j \leq z] \\ &= 1 - e^{-z}. \end{aligned} \tag{2}$$

To search for periodic gene expression, we test the null hypothesis that gene g is non-periodic versus the alternative that it is periodic. We calculate how likely it is for an observed peak in the Lomb–Scargle periodogram of gene g to occur by chance. If the peak in the Lomb–Scargle periodogram of gene g is attained at frequency ω_k among M independent frequencies, we denote such a peak by $X_g = \max_j P_g(\omega_j) = P_g(\omega_k)$. Then, for independently normally distributed noise ε_g, if there were M independent frequencies to test, the probability that the peak Lomb–Scargle periodogram X_g is smaller than x_g is given by

$$\Pr[X_g \leq x_g] = \Pr[Z_j \leq x_g,\ j = 1,\ \ldots, M] = (1 - \mathrm{e}^{-x_g})^M,$$

in view from Equation (2). Thus, the observed statistical significance level, the p-value, of testing the null hypothesis that such a peak in Lomb–Scargle periodogram of gene g is due to chance is calculated by

$$p_g = p\text{-value} = 1 - (1 - \mathrm{e}^{-x_g})^M, \qquad (3)$$

for gene g with $g = 1, \ldots, \mathcal{G}$.

In genome-wide datasets, $\mathcal{G}$ ranges from 10^3 to 10^6, and a multiple testing approach must be employed to control the family-wise error rate when comparing $\mathcal{G}$ profiles simultaneously. Dudoit *et al.* (2003) surveyed several different approaches to multiple hypothesis testing for finding differentially expressed genes in microarray experimental data. The well-known Bonferroni adjustment, when applied to gene expression analysis is somewhat conservative in the sense that too many genes will be rejected for periodicity (Tsai *et al.*, 2003; Knudsen, 2004). Benjamini and Hochberg (1995) proposed another approach to multiple testing by controlling the FDR, the rate of expected proportion of errors among the rejected hypotheses. It is a step-down type of multiple testing procedures in combination with Bonferroni approach. The Benjamini and Hochberg FDR approach is less stringent than the family-wise error rate and more powerful than the methods proposed in Holm (1979), Hochberg (1988) and Hommel (1988). Storey and Tibshirani (2003) proposed an extension of the FDR, called q-value, to give each test its own individual measure of significance. For analyzing large-scale gene expression time-series data, we suggest to use the Benjamini and Hochberg FDR approach in combination with a periodicity-searching method.

The approach of searching for periodic genes proposed in this paper consists of the following steps:

(1) For each gene expression time series, calculate the normalized Lomb–Scargle periodogram, $P_g(\omega_j)$, given by Equation (1), for $j = 1, \ldots, M$, and $g = 1, \ldots, \mathcal{G}$; and find the peak Lomb–Scargle periodogram, $P_g(\omega_k)$, for each gene g.

(2) For each of the $P_g(\omega_k)$ obtained in step 1 above, calculate the p-value, p_g, according to Equation (3).

(3) For the p-values obtained in step 2, obtain the ordered p-values: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(\mathcal{G})}$. Then, find according to $\hat{k}$ according to

$$\hat{k} = \arg \max_{1 \leq k \leq G} \{k : p_{(k)} \leq qk/\mathcal{G}\} \qquad (4)$$

for a desired FDR level q.

(4) Identify the genes whose p-values correspond to $p_{(1)}, p_{(2)}, \ldots, p_{(\hat{k})}$; these genes can be claimed to show statistically significant periodic behavior for the given FDR level q according to Benjamini and Hochberg (1995).

NUMERICAL EXPERIMENTS

Several numerical experiments were performed to assess the effectiveness of the Lomb–Scargle periodogram procedure. This section summarizes the main results of applying the Lomb–Scargle periodogram to search for periodicity in simulated signals taken on unevenly spaced time points; the details and the case of evenly spaced time points data are included in the online supplement (http://research.stowers-institute.org/efg/2005/LombScargle).

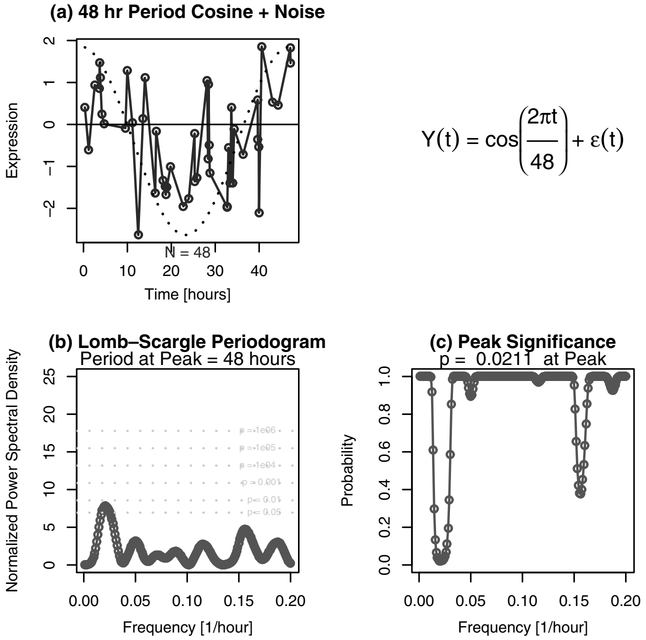

Fig. 1. Simulated cosine signal taken on unevenly spaced time points mixed with Gaussian noise.

Single periodicity detection with unevenly spaced time points

A cosine curve is often used to model an ‘ideal’ periodic gene expression (Ueda, 2002). Since the *Plasmodium falciparum* dataset used in this study contained data from 48 hourly samples, a 48-point cosine signal mixed with normal noise (mean 0 and variance 1) was simulated. Figure 1a shows such a simulated expression profile for a gene that has a 48 h period with data values taken randomly in the 48 h interval. Figure 1b shows a peak near a frequency of 0.0208 per hour ≈1/48 per hour, or a period of 48 h in the Lomb–Scargle periodogram. A p-value curve (Fig. 1c) is obtained for different frequencies; the lowest p-value (highest significance) of 0.0211 is achieved at peak frequency 1/48 per hour.

The highest frequency for which the unevenly spaced data may be evaluated (Van Dongen *et al.*, 1999) is called the Nyquist frequency. The Nyquist frequency for data spaced by an interval Δt, an approximation of the mean of an unevenly time interval, is

$$f_{\text{nyquist}} = 1/(2\Delta t).$$

Spurious peaks can be seen in a periodogram near or above the Nyquist limit. To avoid these problems near the Nyquist limit, and because longer-period biological signals may be of more interest, the frequency range for periodograms was restricted from just above 0 to 0.20 per hour.

A 48-point (unevenly spaced) cosine signal with the period of 24 h mixed with Gaussian noise was also simulated to study for a dominant frequency. The periodograms performed as expected with a single high peak and the p-value curve gave the lowest value of 0.000925 at frequency 1/23.4 per hour (Fig. 2).

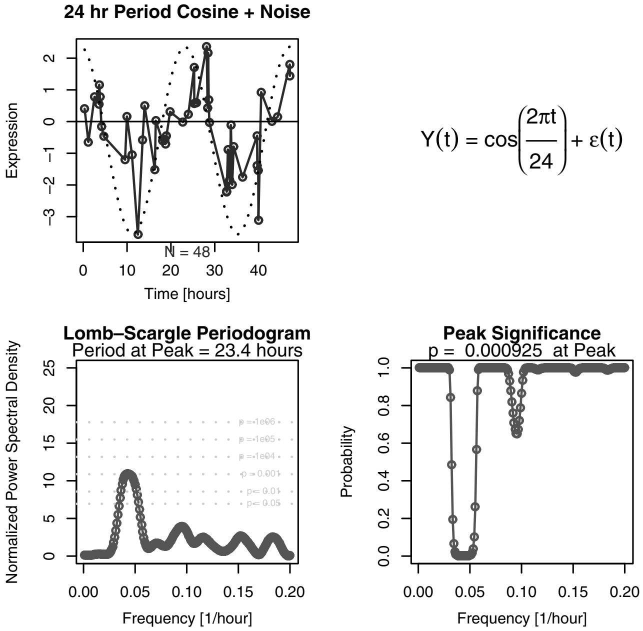

Fig. 2. Simulated cosine signal taken on unevenly spaced time points (mixed with Gaussian noise) with single dominant period.

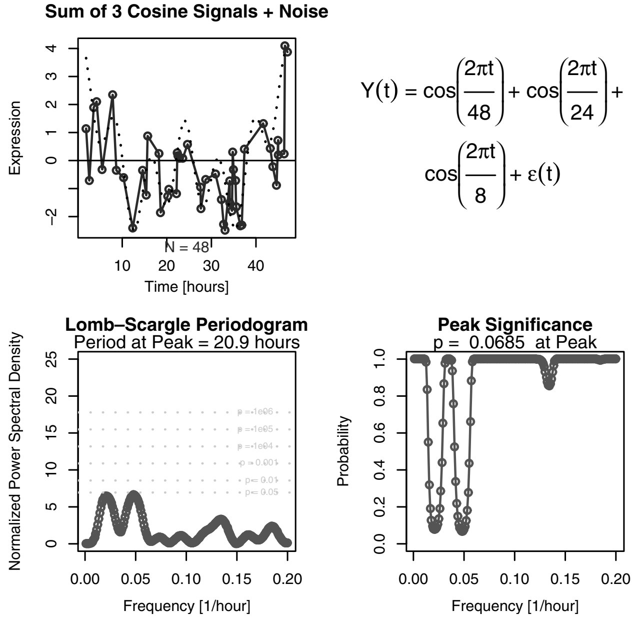

Fig. 3. Simulated cosine signal taken on unevenly spaced time points (mixed with Gaussian noise) with multiple periods.

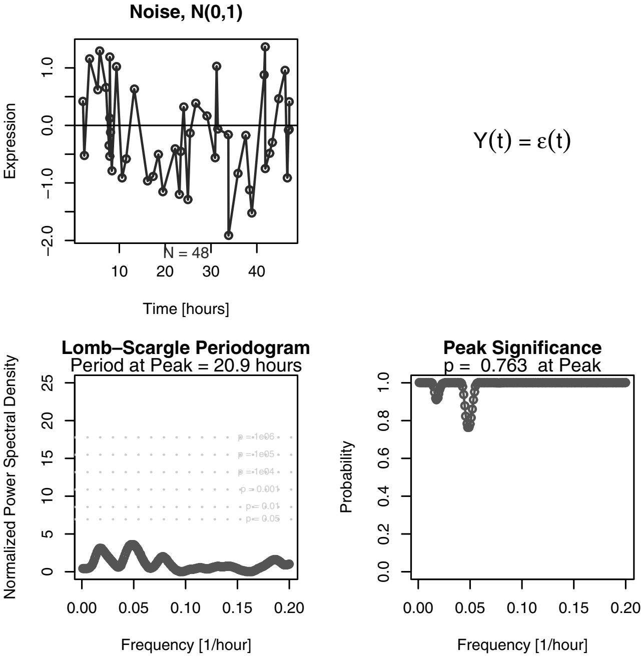

Fig. 4. Simulated Gaussian noise taken at unevenly spaced time points with no periodicity.

Multiple periodicities detection with unevenly spaced time points

Although the biology of the malaria parasite suggests one dominant frequency of periodic gene expression (Bozdech *et al*., 2003), we cannot exclude the possibility that different genes have different expression periodicity and that expression of some genes may reflect two or more periodic processes with different frequencies. Van Dongen *et al*. (1999) have proposed a procedure of multiple periods searching in an unequally spaced human oral temperature time-series data using Lomb–Scargle method.

A simulated profile (on unevenly spaced time points) with three periodicities of 8-, 24-, and 48 h (mixed with a Gaussian noise) was obtained (Fig. 3). All three periods had equal contributions in Figure 3 and the Lomb–Scargle periodogram p-value curve showed two low p-values that indicated at least two dominant frequencies in the signal. Additional results of the numerical experiments on multiple periodicities are in the online supplement.

When the simulated observations were purely Gaussian noise taken on unevenly spaced time points, the Lomb–Scargle periodogram showed no significant peak, and the corresponding p-value curve had a lowest p-value of 0.763 (Fig. 4) indicating no significant periodic signal.

ANALYSIS OF THE *PLASMODIUM FALCIPARUM* INTRAERYTHROCYTIC GENE EXPRESSION

While studying the transcriptional program of the asexual intraerythrocytic developmental cycle (IDC) of malaria parasite *P.falciparum*, Bozdech *et al*. (2003) obtained expression profiles using a DNA microarray. The profiles represent the expression of nearly every gene in that species. Since the symptomology and pathogenesis of malaria are strongly periodic, the study attempted to identify genes that were also strongly periodic, which might be useful for understanding the transcriptional program of *Plasmodium* IDC and for drug intervention. The data (available at http://malaria.ucsf.edu/SupplementalData.php) include three datasets: Complete, Quality Control, and Overview (see online supplement for more discussion). Gene expression was measured every hour throughout the 48 h IDC. The Complete dataset of 7091 probes had

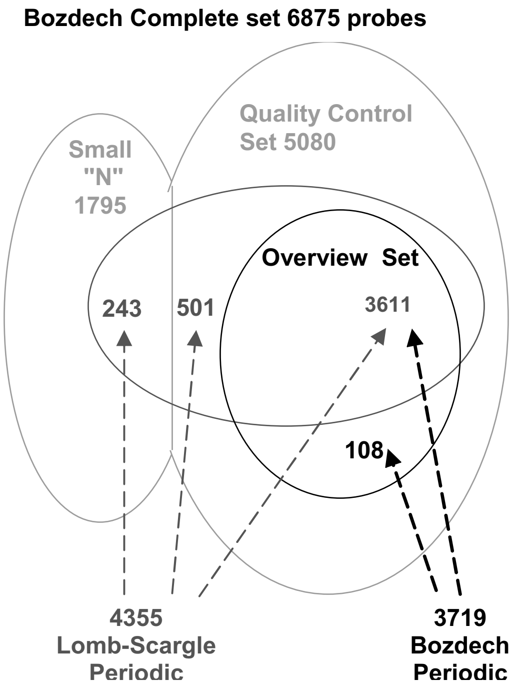

Fig. 5. Periodic gene sets identified by the two methods.

18.23% missing values, including all time points at hour 23 and hour 29. The large number of missing values motivates us to use the Lomb–Scargle periodogram analysis for searching for periodic patterns in this same dataset.

Instead of imputing the missing values, we treated any two points with a missing value in-between as two unevenly spaced time points. The Lomb–Scargle algorithm was implemented in ‘R’ (R Development Core Team, 2004, www.R-project.org), largely based on MatLab code by Glover (2000, http://w3eos.whoi.edu/12.747/notes/lect07/l07s05.html), with additional information from Horne and Baliunas (1986) and Press *et al*. (2002). Several extreme points in the Complete Bozdech dataset of 7091 profiles were identified and profiles with empty Oligo_ID were ignored (see supplement), resulting in a set of 6875 profiles. These 6875 profiles in the Complete dataset (with 15.66% missing values) were later used in our analysis of periodic patterns. More details about the data processing are provided in the online supplement.

Discoveries

Of the 6875 probes in the Complete dataset we found that 4355 probes, or 63%, were considered to be significantly periodic with FDR level $q = 1 \times 10^{-4}$, and 2520 were not considered as significantly periodic. Figure 5 summarizes the results from the Lomb–Scargle algorithm for the Bozdech’s Complete dataset of 6875 probes compared with the results of Bozdech *et al*. (2003) based on the Quality Control dataset.

Bozdech *et al*. (2003) did not analyze the Complete dataset for periodic genes, but rather a subset of the Complete dataset known as the Quality Control dataset. They found that 3719 profiles (the Overview dataset) in the Quality Control dataset of 5080 probes were periodic using a 70% FFT score cutoff and a 75% maximum frequency magnitude cutoff. Using the Lomb–Scargle method, we found that of the 5080 probes in the Quality Control dataset, 4112 were periodic.

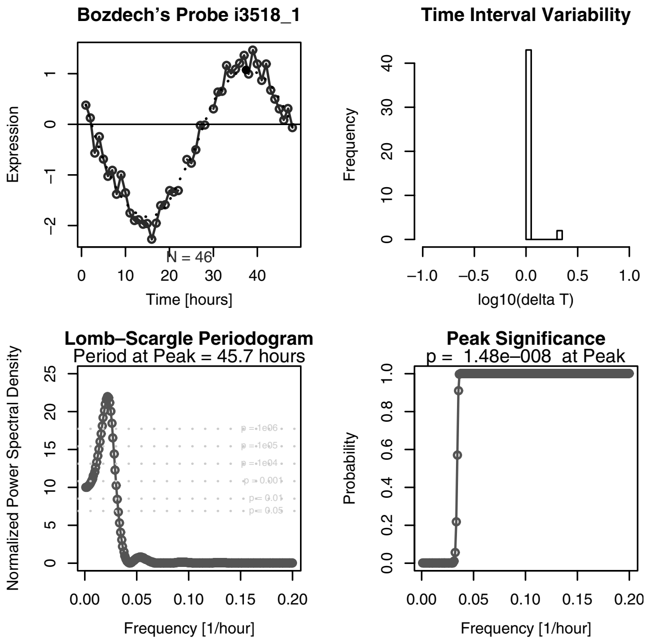

Fig. 6. Periodic gene expression pattern of *i*3518_1 in the *P. falciparum* gene expression dataset.

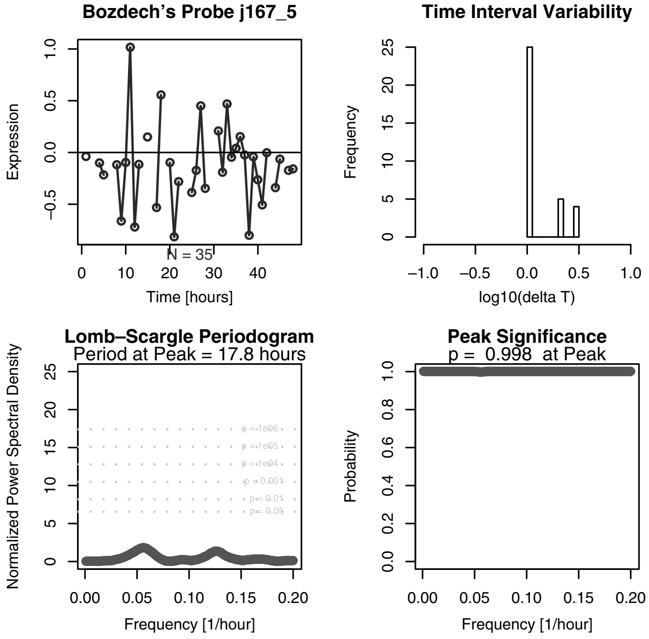

Fig. 7. Non-periodic gene expression pattern of *j*167_5 in the *P. falciparum* gene expression dataset.

One typical periodic gene (probe ID *i*3518_1) with its corresponding Lomb–Scargle periodogram and *p*-value curve is given in Figure 6. A non-periodic gene (probe ID *j*167_5) along with its corresponding Lomb–Scargle periodogram and *p*-value curve is given in Figure 7.

Of the 4355 probes that we selected as periodic from the Complete dataset, 744 now identified as periodic were not considered periodic by Bozdech *et al*. (2003). See the supplement for the gene list and the representative periodograms in this selected set. Among these 744 probes, 243 were processed by the Lomb–Scargle algorithm with a sample size less than 43. These 243 probes resolved to 151 distinct genes. The analysis of these 151 distinct genes presents a picture of sequence conservation generally similar to *Plasmodium* genome as a whole (Aravind *et al*., 2003), with 26% of proteins apparently unique to plasmodia, 36% shared only with other apicomplexan species, and the rest broadly conserved across the phylogeny. Predicted molecular functions of the latter group of genes, however, show uneven distribution of different functional classes of transcripts: the single largest category (13 genes, accounting for 9% of all newly identified periodic genes and for more than one-quarter of genes with predicted molecular function) consists of genes involved in ribosome assembly and RNA maturation (supplementary table). There were no other overrepresented functional categories in the set of 243. Some late trophozoite or early schizont-specific functions highlighted by Bozdech *et al*. (2003), e.g. proteasome subunits and TCA cycle enzymes, were missing from the newly recovered periodic gene set altogether. This suggests that the newly recovered 243 periodic genes perhaps mostly come from the ring or early trophozoite stage (compare Fig. 2 in Bozdech *et al*., 2003). The reasons for that are not clear at present, but most likely they include experimental artifacts and early phase shift for some genes (see supplement).

A histogram by period (see supplement) of the 4355 selected probes shows a dominant frequency at 48 h, as expected. A histogram of *p*-value shows two peaks and suggests a mixture of two distributions (see supplement).

Since the Overview dataset was the most restrictive Bozdech dataset, the 108 probes identified as non-periodic by the Lomb–Scargle algorithm were studied (see supplement). Most of these probes may be claimed as periodic by the Lomb–Scargle test with slightly more lenient selection criteria.

Fourier analysis directly provides an approximate value for the phase of a periodic curve. In the Lomb-Scargle computation, such a phase value is not provided directly, but it can be approximated by the location of the peak value in the expression profile after appropriate smoothing (see supplement).

Missing value tolerance

As pointed out earlier, the *P.falciparum* gene expression Complete dataset of 6875 probes contains 15.66% missing values. In Bozdech *et al*. (2003), missing values in the profiles were imputed using LOESS smoothing before the FFT algorithm was applied. Bozdech *et al*. (2003) concluded that 79.5% of the profiles in the Quality Control set of 5080 probes, or 54.10% of profiles in the Complete set of 6875 probes, were periodic, where periodicity was defined as the FFT score of the profile being higher than a heuristic 70% cutoff score and the profile being in the top 75% of the maximum frequency magnitudes. Heuristic rules have been used to reject certain profiles from analysis if they contained too many missing values and the LOESS smoothing failed to provide good imputation for those missing values. Therefore, Bozdech *et al*. (2003) only analyzed gene expression time series with at most five missing values. The Lomb–Scargle method does not need missing value imputation and just treats a series with missing value as unevenly spaced. Our above analysis illustrated the effectiveness of Lomb–Scargle method in treating missing values.

We further examined extreme cases when many missing values are consecutive in the time series (see supplement). As the original data had missing values for all profiles at hour 23 and hour 29, we deleted all 7 consecutive time points from hour 23 to hour 29 in the Complete dataset of 6875 probes (with missing value rate increased from 15.66 to 26.08% in the dataset), and found that at the FDR level of $q = 1 \times 10^{-4}$, the Lomb–Scargle method identified 3617 probes as periodic (53% periodic versus 54.10% in original analysis). Of the 3617 identified, 3609 were in the original set of 4355 identified periodic genes. When we deleted 11 consecutive time points (with missing value rate increased from 15.66 to 34.41%) from hour 21 to hour 31, the Lomb–Scargle method can still identify 2506 (or 36%) probes as periodic; and of the 2506 identified periodic probes, 2502 were in the original set of 4355 identified periodic genes. These studies indicated that the Lomb–Sargle method has good ‘tolerance’ toward missing values.

DISCUSSION

The Lomb–Scargle periodogram is a promising technique of searching for time series with periodic patterns. It requires no special treatment of missing values and can be used in data taken on unevenly spaced time points. Under the normal random noise assumption, the *p*-value for labeling each gene as periodic is easily calculated. There is no need for an *ad hoc* scoring system of power in peaks or for use of random permutations to assess significance of a peak. Weighting of data occurs on a ‘per point’ basis instead of on a ‘per frequency interval’ basis (Press and Rybicki, 1989). As all other methods in the search for periodic signals, the Lomb–Scargle periodogram requires the assumption of Gaussian noise on the error term. When the Gaussian noise assumption is invalid, the conclusion of Lomb–Scargle method might be misleading. Schimmel (2001a,b) also discussed the limitation of the Lomb–Scargle method when the signal is periodic with non-sinusoidal shapes or with outliers. These issues become topics of our further investigation. Our current studies indicate that for the *P.falciparum* dataset the Lomb–Scargle periodogram method is more appropriate, as it is better suited to handle unevenly spaced time series.

The question of how many time points should be planned to observe a particular periodical expression pattern might be raised at the planning stage of the experiment. We provide a simple guideline for estimating the sample size *N* (number of time points) of one profile for a given *p*-value as follows:

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

The derivation of Equation (5) is through a simple regression curve and is given in the online supplement. This equation only provides a guideline on how many data points should be planned if one given profile will be viewed as significant at the given *p*-value. After a large number of profiles are obtained, a given FDR level should be used to identify actually how many profiles are significantly periodic.

CONCLUSION

The Lomb–Scargle periodogram algorithm is an effective tool for finding periodic gene expression profiles in microarray data, especially when data may be collected at arbitrary time points or when a significant proportion of data is missing.

ACKNOWLEDGEMENTS

We thank two anonymous referees whose comments greatly improved the manuscript. We appreciate discussions with Galina Glazko, Chris Seidel, Mary-Lee Dequeant, Oliver Pourquie and Mike Melko. J.C. is supported by the NSF grant DMS-0426148 for this research and is a visiting scientist at the Bioinformatics Center of the Stowers Institute for Medical Research while on leave from the Department of Mathematics and Statistics of the University of Missouri-Kansas City.

Conflict of Interest: none declared.

REFERENCES

Aravind,L. *et al*. (2003) Plasmodium biology: genomic gleanings. *Cell*, **115**, 771–785.

Bar-Joseph,Z. *et al*. (2003) Continuous representations of time series gene expression data. *J. Comp. Biol*., **10**, 341–356.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc*., **B57**, 289–300.

Bohn,A. *et al*. (2003) Identification of rhythmic subsystems in the circadian cycle of crassulacean acid metabolism under thermoperiodic perturbations. *Biol. Chem*., **384**, 721–728.

Bozdech,Z. *et al*. (2003) The Transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS Biol*., **1**, 1–16.

Crosthwaite,S.K. (2004) Circadian clocks and natural antisense RNA. *FEBS Lett*., **567**, 49–54.

Dale,J.K. *et al*. (2003) Periodic notch inhibition by lunatic fringe underlies the chick segmentation clock. *Nature*, **421**, 275–278.

Dudoit,S. *et al*. (2003) Multiple hypothesis testing in microarray experiments. *Stat. sci*., **18**, 71–103.

Durbin,J. (1967) Tests of serial independence based on the cumulated periodogram. *Bull. Int. Stat. Inst*., **42**, 1039–1049.

Glover,D.M. (2000) Non-Uniform Time Series, Woods Hole Oceanographic Institute.

Goldbeter,A. (2002) Computational approaches to cellular rhythms. *Nature*, **420**, 238–245.

Harmer,S.L. *et al*. (2000) Orchestrated transcription of key pathways in *Arabidopsis* by the circadian clock. *Science*, **290**, 2110–2113.

Hochberg,Y. (1988) A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, **75**, 800–803.

Holm,S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Stat*., **6**, 65–70.

Hommel,G. (1988) A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, **75**, 383–386.

Horne,J.H and Baliunas,S.L (1986) A prescription for period analysis of unevenly sampled time series. *Astrophys. J*., **302**, 757–763.

Johansson,D. *et al*. (2003) A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription. *Bioinformatics*, **19**, 467–473.

Knudsen,S. (2004) *Guide to Analysis of DNA Microarray Data*. 2nd edn, Wiley Liss, Hoboken, New Jersey.

Lakin-Thomas,P.L. and Brody,S. (2004) Circadian rhythms in microorganisms: new complexities. *Annu. Rev. Microbiol*., **58**, 489–519.

Langmead,C.J. *et al*. (2003) Phase-independent rhythmic analysis of genome-wide expression patterns. *J. Comp. Biol*., **10**, 521–536.

Lawton,W.H. *et al*. (1972) Self-modeling nonlinear regression. *Technometrics*, **13**, 513–532.

Lomb,N.R. (1976) Least-squares frequency analysis of unequally spaced data. *Astrophys. Space Sci*., **39**, 447–462.

Luan,Y. and Li,H. (2004) Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data. *Bioinformatics*, **20**, 332–339.

Mitchison,J.M. (2003) Growth during the cell cycle.. *Int. Rev. Cytol*,, **226**, 165–258.

Panda,S. *et al*. (2002) Coordinated transcription of key pathways in the mouse by the circadian clock. *Cell*, **109**, 307–320.

Press,W.H. and Rybicki,G.B. (1989) Fast algorithm for spectral analysis of unevenly sampled data. *Astrophysical J*., **338**, 277–281.

Press,W.H., Teukolsky,S.A., Vetterling,W.T. and Flannery,B.P. (2002) *Numerical Recipes in C++*, 2nd edn, Cambridge University Press, Cambridge.

Priestley,M.B. (1981) *Spectral Analysis and Time Series*. Academic Press, San Diego.

Prolo,L.M. *et al*. (2005) Circadian rhythm generation and entrainment in astrocytes. *J. Neurosci*., **12**, 404–408.

R Development Core Team (2004) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rovery,C. *et al*. (2005) Transcriptional response of *Rickettsia conorii* exposed to temperature variation and stress starvation. *Res. Microbiol*., **156**, 211–218.

Ruf,T. (1999) The Lomb–Scargle Periodogram in biological rhythm research: analysis of incomplete and unequally spaced time-series. *Biol. Rhythm Res*., **30**, 178–201.

Scargle,J.D. (1982) Statistical aspects of spectral analysis of unevenly spaced data. *Astrophys. J*., **263**, 835–853.

Schimmel,M. (2001a) Emphasizing difficulties in the detection of rhythms with Lomb–Scargle periodograms. *Biol. Rhythm Res*., **32**, 341–345.

Schimmel,M. (2001b) The issue of significant features in random noise. *Biol. Rhythm Res*., **32**, 355–360.

Shedden,K. and Cooper,S. (2002) Analysis of cell-cycle gene expression in *Saccharomyces cerevisiae* using microarrays and multiple synchronization methods. *Nucleic Acids Res*., **30**, 2920–2929.

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 analysis. *Proc. Natl Acad. Sci*., **100**, 9440–9445.

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

Tsai,C. *et al*. (2003) Estimation of false discovery rates in multiple testing: application to gene microarray data. *Biometrics*, **59**, 1071–1081.

Ueda,H.R. *et al*. (2002) Genome-wide transcriptional orchestration of Circadian rhythms. *J. Biol. Chem*., **277**, 14048–14052.

Van Dongen,H.P.A. *et al*. (1999) A procedure of multiple periods searching in unequally spaced time-series with Lomb–Scargle Method. *Biol. Rhythm Res*., **30**, 149–177.

Van Dongen,H.P.A. *et al*. (2001) Letter to the Editor: analysis of problematic time series with the Lomb–Scargle method, a reply to ‘Emphasizing Difficulties in the Detection of Rhythms with Lomb-Scargle Periodograms’. *Biol. Rhythm Res*., **32**, 347–354.

Wang,Y. and Brown,M.M. (1996) A flexible model for human circadian rhythms. *Biometrics*, **52**, 588–596.

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

Zhao,L.P. *et al*. (2001) Statistical modeling of large microarray data sets to identify stimulus–response profiles. *Proc. Natl Acad. Sci. USA*, **98**, 5631–5636.