

# Searching for Periodic Gene Expression Patterns Using Lomb-Scargle Periodograms

Earl F. Glynn and Arcady R. Mushegian  
Stowers Institute for Medical Research  
1000 East 50<sup>th</sup> Street  
Kansas City, MO 64110  
816.926.4412

[efg@Stowers-Institute.org](mailto:efg@Stowers-Institute.org)

## ABSTRACT

The Lomb-Scargle periodogram approach was applied to the search for periodically expressed genes in a *Plasmodium falciparum* dataset. The Lomb-Scargle algorithm has several computational advantages over more common approaches, such as Fourier analysis, including direct treatment of missing values and a periodogram that has known statistical properties. Hierarchical clustering of periodograms shows explicit partitioning of multiple periodicities present in some gene expression patterns.

The Lomb-Scargle algorithm performance was compared to earlier analysis of CAMDA 2004 challenge dataset by Bozdech et al. [1], based on Fast Fourier Transforms (FFT). We identified an additional 265 genes with 48-hr periodic expression, which were not considered by the FFT because they had too many missing values. We also automatically detected, in the same analysis, expression patterns with periodicity close to 24 hr, and other interesting patterns.

## Keywords

Lomb-Scargle periodogram, periodicity, gene expression profile, time series, missing values, false discovery rate, Fourier analyses, FFT, microarray analysis, periodogram clustering

## 1. INTRODUCTION

### 1.1 Problems with Fourier Analysis

Time series often contain unknown periodicities, some of which may be of interest. A common technique used to study periodic data is Fourier analysis, and in particular, an algorithm called the Fast Fourier Transform (FFT). Periodicities are found by searching the “standard” periodogram from Fourier analysis and looking for sharp peaks. Usually these peaks correspond to intrinsic periodicity in the time series. However, evenly sampled data are required for applying the FFT algorithm, which can be difficult to achieve.

Missing values cause gaps in what otherwise would be evenly sampled data, and even sampling may sometimes be impossible in the first place for technical reasons. When Fourier analysis is used, and some data points are missing, they have to be provided in a more or less arbitrary way. Bozdech et al. [1] studied the patterns of gene expression in malaria parasite *Plasmodium falciparum*, and they used a common approach of imputing missing values by applying a weighted regression algorithm.

The *Complete* Bozdech dataset has many missing values. This dataset has expression profiles for 6875 “non-empty” probes. Figure 1 summarizes the number of missing values by time period. Points for time periods 23 and 29 are completely missing in all profiles. Most of the profiles had 46 of the 48 hourly time points, but many profiles had less.

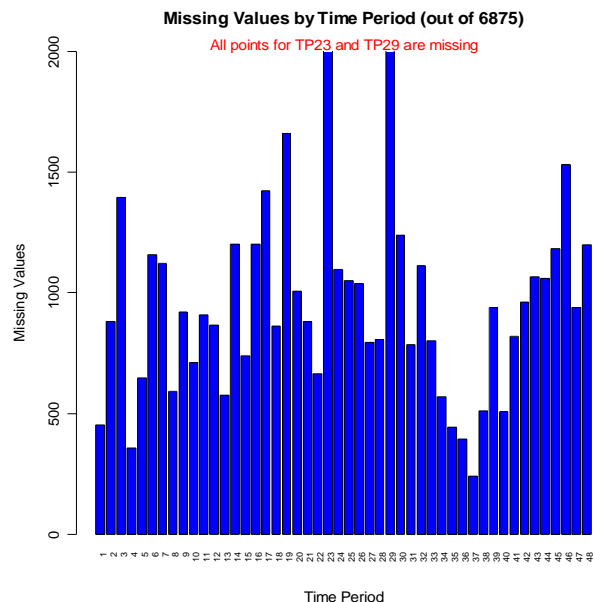


Figure 1. Missing Values by Time Period in *Complete* Bozdech Dataset

With missing values appearing throughout the dataset, time series have a variable sample size, unless these missing values are assigned somehow. In addition, to comply with their FFT algorithm requirement of a sample size that is a power of 2, Bozdech added a variable number of zeros to each series to create 64 time points.

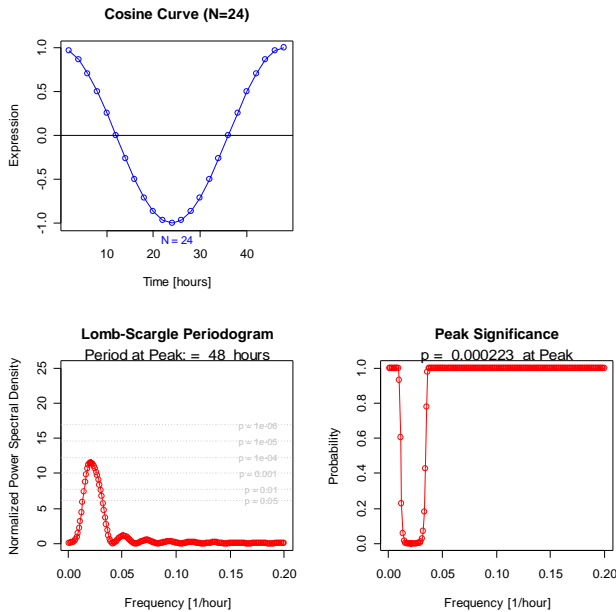
It is not known whether these heuristic data changes have an effect on the analysis. A better approach would be to handle missing data directly. It is also desirable to have statistics that can evaluate the significance of each periodogram peak.

### 1.2 Lomb-Scargle Periodogram Overview

Lomb, who studied variable stars in astronomy [4], 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 points, Lomb computed statistics on least-squares fits to sinusoidal data models. Lomb attempted to find the probability distribution of the height of the highest peak in a periodogram. Scargle [8] extended the work of Lomb to include a “false alarm probability”, which gives an estimate of the significance of calling a point a peak given its height. Scargle showed that peak heights in the Lomb-Scargle periodogram, when properly normalized, follow an exponential probability distribution. Home and Baliunas [3] showed that correct “normalization” for the Lomb-Scargle periodogram power is achieved by dividing by the total variance of the data. [Note: this “normalization” should not be confused with any kind of microarray data normalization.] Press and Rybicki [5] presented the algorithm in a way where normalization is always correctly performed.

Before taking a look at the mathematical details of the Lomb-Scargle periodogram, let’s study what its output will tell us. A generalized cosine curve has been used to represent the ideal expression of a gene that goes from “on” state, to an “off” state, and then back to “on” [10, 11]. The upper left corner of Figure 2 shows such an idealized expression profile for a gene that has a 48-hour period with data samples taken every two hours. The Lomb-Scargle Periodogram at the lower left of Figure 2 shows a peak near a frequency of 0.0208/hour  $\approx 1/48$  /hour. This should be the case for the 48-hour period curve, since frequency and period are inverses.



**Figure 2. Lomb-Scargle Periodogram for Idealized Gene Expression Profile (Cosine Curve)**

Press [6] explains how we can give a quantitative answer to the question, “How significant is a peak in this spectrum?” In this question, the null hypothesis is that each value in the Lomb-Scargle periodogram is an independent Gaussian random variable. Because of the Lomb-Scargle periodogram normalization, the peak values follow an exponential distribution with unit mean. A small value for this false alarm probability statistic indicates a significant periodic signal.

A p-value for every point in the periodogram can be computed as shown in the lower right of Figure 2. Since the cosine is a periodic curve, the significance is quite high and the p-value is quite small.

The Lomb-Scargle algorithm is not equivalent to the conventional periodogram analysis [6].

### 1.3 Mathematical Details

Assume we have N data points at specific times,  $t_1, t_2, t_3, \dots, t_N$ , which may or may not be evenly spaced. Assume the corresponding data points are  $h_1, h_2, h_3, \dots, h_N$ . When data are missing, the missing points can be excluded from the analysis and no other special treatment is necessary.

The first step in the algorithm is to compute the mean and variance of the data:

**Mean**

$$\bar{h} = \frac{h_1 + h_2 + h_3 + \dots + h_N}{N} = \frac{1}{N} \sum_{i=1}^N h_i \quad [1]$$

**Variance**

$$s^2 = \frac{1}{N-1} \sum_{i=1}^N (h_i - \bar{h})^2 \quad [2]$$

The Lomb-Scargle periodogram will be evaluated at M frequencies (discussed below),  $f_1, f_2, f_3, \dots, f_k, \dots, f_M$ . The corresponding angular frequencies,  $\omega_k$ , are computed as follows:

$$\omega_k = 2 \pi f_k \quad [3]$$

A value  $\tau$  is computed as follows:

$$\tan(2 \omega_k \tau) = \frac{\sum_{i=1}^N \sin(2 \omega_k t_i)}{\sum_{i=1}^N \cos(2 \omega_k t_i)} \quad [4]$$

See Press [6] for some interesting properties of  $\tau$ .

The formula for the normalized spectral power density is a bit overwhelming:

$$SPD(W_k) = \frac{1}{2 \cdot s^2} \left[ \frac{\left[ \sum_{i=1}^N (h_i - \bar{h}) \cos [W_k (t_i - t)] \right]^2}{\sum_{i=1}^N \cos^2 [W_k (t_i - t)]} + \frac{\left[ \sum_{i=1}^N (h_i - \bar{h}) \sin [W_k (t_i - t)] \right]^2}{\sum_{i=1}^N \sin^2 [W_k (t_i - t)]} \right]$$

### Normalized Spectral Power Density (SPD) [5]

The Lomb-Scargle periodogram is a graph of  $SPD(\omega_k)$  versus the frequencies,  $f_k$ . There is no consensus about the number of frequencies,  $M$ , to use in this graph. Unlike Fourier frequencies in the normal periodogram that are all independent, not all these  $M$  frequencies will be independent. We need just enough test frequencies to not miss any peak in the periodogram. We only care about the number of independent frequencies since this value is needed to estimate the statistical significance of any peak in the periodogram.

Horne and Baliunas [3] give a simple least squares formula to estimate the number of independent frequencies,  $N_{\text{independent}}$ , from the number of observations in a time series,  $N$ :

$$N_{\text{independent}} \approx -6.362 + 1.193N + 0.00098N^2 \quad [6]$$

Apparently if we scan  $M > N_{\text{independent}}$  frequencies, the additional frequencies are not independent.

If we scan some  $N_{\text{independent}}$  frequencies, and find a peak at  $z = SPD(\omega_k)$ , the probability that no other peak will give a larger value is given by  $[1 - \exp(-z)]^{N_{\text{independent}}}$ . So the false-alarm probability can be computed as follows:

$$p = P(>z) = 1 - [1 - \exp(-z)]^{N_{\text{independent}}} \quad [7]$$

If we solve equation [7] for  $z$ , we can compute the horizontal significance level lines to be plotted with the Lomb Scargle periodogram, as shown in the graph at the lower left in Figure 2:

$$z = -\ln[1 - (1 - p)^{1/N_{\text{independent}}}] \quad [8]$$

## 2. NUMERICAL EXPERIMENTS

Several numerical experiments were performed to assess the effectiveness of the Lomb-Scargle periodogram technique before it was applied to the Bozdech dataset. The details of these experiments are included in the online supplement.

### 2.1 Single Periodicity

A cosine curve is often used to model an "ideal" periodic gene [10, 11]. Since the Bozdech dataset contained data from 48 hourly samples, the behavior of a 48-point cosine curve was studied for various frequencies, corresponding to periods from 4 hours to 72 hours. The periodograms performed as expected with a single high peak on the periodogram and a single low peak on the p-value curve. With periods 48 hours or shorter, the correct period was computed, with a quite significant p-value of about 3E-9.

The Nyquist frequency for data spaced by an interval,  $\Delta t$ , is:

$$f_{\text{Nyquist}} = \frac{1}{2\Delta t} \quad [9]$$

A Nyquist limit can be estimated for unevenly sampled data by using the mean time interval. With hourly data,  $\Delta t = 1$ , and  $f_{\text{Nyquist}} = 0.5$ . Spurious spikes were seen in the periodogram when frequencies above the Nyquist limit were inspected, especially when viewing the data for the 4-hour period. To avoid any problems near the Nyquist limit, and because longer-period biological signals were the primary ones under consideration, the frequency range for periodograms was restricted from just above 0, to 0.20/hour in subsequent analyses.

### 2.2 Double and Triple Periodicities

It is not clear whether all periodically expressed genes have the same or different periodicities, or indeed whether a periodically expressed gene has a single dominant periodicity. A study of ideal profiles with two simultaneous periodicities, 24 hours and 48 hours, showed that two periodicities can only be observed with statistical significance if the ratio of the contributions was within a factor of approximately 2. When the factor was about 2:1 or 4:1, one dominant frequency is still quite statistically significant, and the lesser frequency can be seen in the periodogram, but with questionable statistical significance. With a ratio greater than about 4:1, only the dominant frequency can be observed.

Simulated profiles with three periodicities of 8-, 24-, and 48-hours were also studied. Again, if all three had nearly equal contributions, they all could be resolved with statistical significance. With a ratio of 1:2:1, only the dominant periodicity could be resolved with statistical significance even though the other frequencies were visible in the periodogram.

The frequency peaks for 8-, 24-, and 48-hours could all be simultaneously resolved on the periodogram. However, a period of 36-hours would have significant overlap with either 24- or 48-hour periodogram peaks, and may not be resolvable as part of a mixture.

### 2.3 Sample Size

The supplement shows the derivation of a simple regression curve for estimating the sample size for a given expected p-value:

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

If we are working with approximately 10,000 genes, and desire a false alarm probability of about 1E-4, which should result in few false positives, we need a sample size of about 25. Bozdech analyzed only profiles with more than 44 points. With  $N=25$ , we can analyze 920 more expression profiles than Bozdech considered.

### 2.4 "Noise" Experiments

The supplement shows a summary table from a number of numerical experiments of adding "noise" to a cosine "signal." For  $N=48$  points in a time series, even the addition of 50% noise didn't prevent the Lomb-Scargle technique from finding the signal with a reasonable p-value of approximately 2E-4. For 20%

noise added to a cosine signal of  $N=24$  points, the p-value only increased to  $9E-4$  from about  $2E-4$  with no noise.

Statistics of the  $\log_2$  expression values in the Bozdech dataset show that the mean was  $\sim 0.0$ , and the standard deviation was  $\sim 1.0$ . Random datasets with approximately the same statistics were created to determine experimentally how many “false positives” would be observed as “periodic” and how the Lomb-Scargle periodogram p-value statistic might be used to discriminate noise from signal. The supplement shows histograms of the p-value statistic for various sets of 5000 random expression profiles. A histogram of the  $\log_{10}(\text{p-value})$  shows a very narrow range, and is skewed toward 0, with a mean of about  $-0.76$ . In the 35,000 random expression profiles, only one  $\log_{10}(\text{p-value})$  was less than  $-3$ , with a value of  $-3.2$ , with the second smallest p-value being  $-2.7$ . A p-value of  $1E-4$  should be quite adequate to reject any noise coming from non-periodic expression.

### 3. METHODS

The *Complete* Bozdech dataset [1] with 7091 profiles was analyzed using the “R” software package. Several extreme points were identified (see supplement). The code in *CAMDA04.R* set these extremes to missing values and deleted 216 rows that were marked as “empty” in the *Oligo\_ID* field. The remaining 6,875 rows all had unique *Oligo\_IDs*.

The Lomb-Scargle algorithm was implemented in “R” [7] largely based on MatLab code by Glover [2] using additional information from references [3] and [6]. The *LombScargleLibrary.R* file provides a function, *ComputeLombScargle*, for processing a generic time series. A second function, *PlotLombScargle*, is used to plot the series, the periodogram, and a graph of the corresponding p-value for each inspected frequency.

Parameters to *ComputeLombScargle* include the time points and expression profile. Missing values for a row were excluded from the processed points and no other special processing of missing values was necessary. Based on numerical experiments described above, the test frequencies were within a range of 0 and 0.2/hour. 192 test frequencies were used for each profile, which was to correspond to  $4*N$  for a 48-point series, as suggested by Press [6].

The function, *ProcessExpressionData*, in *CAMDA04.R* can be used to process the whole Bozdech dataset. This function writes files to a target directory. The file *Dominant.CSV* includes summary information for the dominant frequency in each expression profile. Numeric values of the periodogram for each inspected frequency are written to a *Periodogram.CSV* file, with corresponding p-values written to a *pValue.CSV* file. In addition, a directory of low-resolution charts is written to disk so the results for any given probe can be viewed by *Oligo\_ID* if desired.

“R” code in *Analyze1.R* was used to filter the 6875 probes that passed various selection criteria, which included  $N \geq 25$  (6000 probes),  $\log_{10}(\text{p-value}) \leq -4$  (4454 probes), and  $8 \text{ hours} \leq \text{period} \leq 60 \text{ hours}$  (6425 probes).

Spotfire [9] was used to hierarchically cluster the periodograms using the *Periodogram.CSV* data file.

### 4. RESULTS

The mean expression profile for all probes (see supplement) shows a slight diurnal period (27.4 hours) with weak significance (p-value = 0.06).

Of the 6875 probes in the *Complete* Bozdech dataset, 4422 were considered to be periodic and 2453 were aperiodic, or noise-like. A histogram by period (see supplement) of the 4422 selected probes shows a dominant frequency at 48 hours as expected.

A histogram by p-value (see supplement) shows two peaks and suggests a mixture of two distributions. The distribution on the left represents probes that are periodic with a statistically significant p-value. The distribution on the right in the histogram is similar to the noise seen in the numerical experiments.

Table 1 shows the results from the Lomb-Scargle algorithm for each of the Bozdech’s datasets:

**Table 1. Probes Selected/Rejected as Periodic by Lomb-Scargle Technique**

Bozdech Dataset	Probes	Lomb-Scargle Results	
		Selected	Rejected
<i>Complete</i>	6875	4422	2453
<i>Quality Control</i>	5071	4157	914
<i>Overview</i>	3711	3631	80

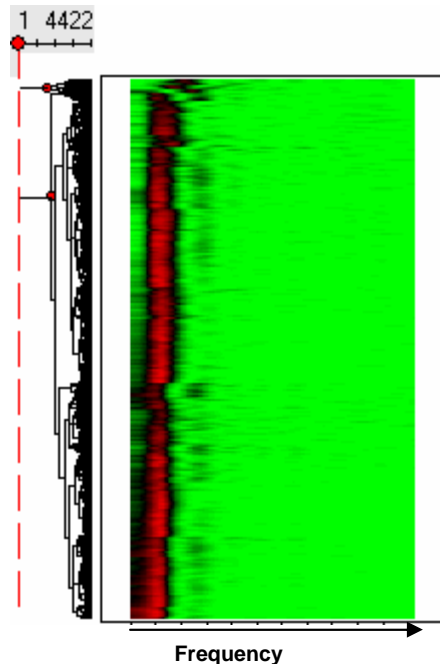
The Bozdech datasets were refinements in their selection process of periodic genes. Bozdech applied Fourier analysis to the *Quality Control* dataset and filtered those results to obtain the *Overview* dataset.

Of the 4422 probes that we selected as periodic from the *Complete* dataset, 265 now identified as periodic could not be processed, and/or considered periodic by Bozdech. See the supplement for a table and representative periodograms in this selected set. 258 of these 265 were processed by the Lomb-Scargle algorithm with a sample size less than 43. It’s unclear how Bozdech’s Fourier approach would have classified these profiles with smaller values of  $N$ . One probe, c230, with only 32 data points, was found to have a periodicity of 43.6 hours, with a p-value value of  $8E-5$ .

Annotations were available for only 211 of the 265 new probes identified as periodic. These 211 probes represented 198 unique ORFs. Of these 198 ORFs, 45 were selected by other probes and analyzed by Bozdech, which leaves 153 periodic ORFs to study further.

Since the *Overview* dataset was the most restrictive Bozdech dataset, the 80 probes rejected by the Lomb-Scargle algorithm were studied (see supplement). Most of these rejected probes would have been selected by Lomb-Scargle with slightly more lenient selection criteria.

Figure 3 shows the result of hierarchical clustering of the periodogram results. A clear 48-hour period dominant band is seen in red along the left side of this figure. A weaker diurnal band can often be seen to the right of this dominant band for many of the probes. Additional bands are rare.



**Figure 3. Hierarchical Clustering of Lomb-Scargle Periodograms for 4422 Selected Probes**

At the top of Figure 3, above the weak diurnal band, there are three small clusters of probes that do not have a dominant frequency of 48 hours. One of these small clusters of 40 probes has a dominant frequency of  $24.8 \pm 1.6$  hours with quite significant p-value from  $9.5E-5$  to  $4.8E-7$ . However, Bozdech categorized 39 of these 40 probes as having a periodicity of 48 hours. The biological relevance of these 40 probes (representing 26 genes) is under investigation.

Two other nearby clusters also have periodicities distinct from 48 hours: 28 probes with period  $30.5 \pm 2.4$  hours, and 46 probes with period  $23.7 \pm 6.4$  hours. This last set of 46 also shows a weak 48-hour periodicity, but a more dominant diurnal frequency.

Applying the Lomb-Scargle algorithm to the *Complete* dataset takes about 40 minutes on a 1.7 GHz Pentium M processor. Optimization of run time was not attempted.

## 5. DISCUSSION

Lomb-Scargle periodogram is a promising technique of searching for time series with periodic patterns. It requires no special treatment of missing values, and directly gives a p-value statistic to determine the significance of a periodicity peak in a periodogram. 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. Lomb-Scargle provides a more uniform way to treat all data. Weighting of data is on a "per point" basis instead of on a "per time interval" basis [5].

Wichert [11] proposed using "average periodograms" as a tool to finding periodic genes graphically. The hierarchical clustering of a series of periodograms should be more sensitive than an average periodogram in finding a small number of genes that may be related by their periodicities, but we did not attempt a direct comparison of the techniques yet.

## 6. CONCLUSIONS

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.

## 7. SUPPLEMENTARY MATERIALS

See this web page for supplementary information related to this paper: <http://research.stowers-institute.org/efg/2004/CAMDA/>

## 8. ACKNOWLEDGMENTS

Thanks to Galilna Glazko for help with sequences and discussions. We appreciate the feedback and discussions with Chris Seidel, Mary-Lee Dequeant, Oliver Pourquie, Mike Melko, Jie Chen, and Mark Frei.

## 9. REFERENCES

- [1] Bozdech, Zbynek, et al. The Transcriptome of the Intraerythrocytic Developmental Cycle of *Plasmodium falciparum*, *PLoS Biology*, 1: 1-16, 2003.
- [2] Glover, D.M. Non-Uniform Time Series, Woods Hole Oceanographic Institute, <http://w3eos.whoi.edu/12.747/notes/lect07/107s05.html>, 2000.
- [3] Horne, J.H and Baliunas, S.L. A Prescription for Period Analysis of Unevenly Sampled Time Series, *Astrophysical Journal*, 302: 757-763, 1986.
- [4] Lomb, N. R. Least-Squares Frequency Analysis of Unequally Spaced Data, *Astrophysics and Space Science*, 39: 447-462, 1976.
- [5] Press, W. H. and Rybicki, G.B. Fast Algorithm for Spectral Analysis of Unevenly Sampled Data. *Astrophysical Journal*, 338: 277-281, 1989.
- [6] Press, W. H., et al. Spectral Analysis of Unevenly Sampled Data. *Numerical Recipes in C++ (2<sup>nd</sup> edition)*, Cambridge University Press, 2002.
- [7] R Development Core Team (2004). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL [www.R-project.org](http://www.R-project.org)
- [8] Scargle, J.D. Statistical Aspects of Spectral Analysis of Unevenly Spaced Data. *Astrophysical Journal*, 263: 835-853, 1982.
- [9] Spotfire. URL [www.spotfire.com](http://www.spotfire.com).
- [10] Ueda, Hiroki R, et al. Genome-wide Transcriptional Orchestration of Circadian Rhythms, *Journal of Biological Chemistry*, 277: 14048-14052, 2002.
- [11] Wichert, S., et al. Identifying periodically expressed transcripts in microarray time series data, *Bioinformatics*, 20: 5-20, 2004.