Searching for Periodic Gene Expression Patterns Using Lomb-Scargle Periodograms

Earl F. Glynn
Stowers Institute

Arcady R. Mushegian
Stowers Institute & Univ. of Kansas Medical Center

Jie Chen
Stowers Institute & Univ. of Missouri Kansas City

http://research.stowers-institute.org/efg/2004/CAMDA

Critical Assessment of Microarray Data Analysis Conference
November 11, 2004
Searching for Periodic Gene Expression Patterns Using Lomb-Scargle Periodograms

- Periodic Patterns in Biology
- Introduction to Lomb-Scargle Periodogram
- Data Pipeline
- Application to Bozdech’s *Plasmodium* dataset
- Conclusions
Periodic Patterns in Biology

A vertebrate’s body plan: a segmented pattern. Segmentation is established during somitogenesis.

Photograph taken at Reptile Gardens, Rapid City, SD
www.reptile-gardens.com
Periodic Patterns in Biology

Intraerythrocytic Developmental Cycle of *Plasmodium falciparum*

Expression Ratio = \( \frac{\text{Cy5}}{\text{Cy3}} \) = \( \frac{\text{RNA from parasitized red blood cells}}{\text{RNA from all development cycles}} \)

Values for Log\(_2\)(Expression Ratio) are approximately normally distributed. Assume gene expression reflects observed biological periodicity.
Simple Periodic Gene Expression Model

Gene Expression = Constant × Cosine(2πf t)

“Periodic” if only observed over a single cycle?

Frequency = \[ \frac{1}{\text{period}} \]

\[ f = \frac{1}{T} \]

\[ \omega = \text{angular frequency} = 2\pi f \]
Introduction to Lomb-Scargle Periodogram

- What is a Periodogram?
- Why Lomb-Scargle Instead of Fourier?
- Example Using Cosine Expression Model
- Mathematical Details
- Mathematical Experiments
  - Single Dominant Frequency
  - Multiple Frequencies
  - Mixtures: Signal and Noise
What is a Periodogram?

- A graph showing frequency “power” for a spectrum of frequencies
- “Peak” in periodogram indicates a frequency with significant periodicity
Why Lomb-Scargle Instead of Fourier?

- Missing data handled naturally
- No data imputation needed
- Any number of points can be used
- No need for $2^N$ data points like with FFT
- Lomb-Scargle periodogram has known statistical properties

Note: The Lomb-Scargle algorithm is NOT equivalent to the conventional periodogram analysis based Fourier analysis.
Lomb-Scargle Periodogram
Example Using Cosine Expression Model

A small value for the false-alarm probability indicates a highly significant periodic signal.

\[ T = \frac{1}{f} \]

Evenly-spaced time points
Lomb-Scargle Periodogram
Example Using Noisy Cosine Expression Model

Unevenly-spaced time points
Lomb-Scargle Periodogram
Example Using Noise

- Noise (N=48)

- Time Interval Variability

- Lomb-Scargle Periodogram
  Period at Peak = 7.4 hours

- Peak Significance
  p = 0.973 at Peak
Lomb-Scargle Periodogram Mathematical Details

\[ h_i \equiv h(t_i), \quad i = 1, \ldots, N. \]

\[
\bar{h} = \frac{1}{N} \sum_{i=1}^{N} h_i, \quad \sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (h_i - \bar{h})^2 \tag{13.8.3}
\]

Now, the Lomb normalized periodogram (spectral power as a function of angular frequency \( \omega \equiv 2\pi f > 0 \)) is defined by

\[
P_N(\omega) \equiv \frac{1}{2\sigma^2} \left\{ \frac{\left[ \sum_j (h_j - \bar{h}) \cos\omega(t_j - \tau) \right]^2 + \left[ \sum_j (h_j - \bar{h}) \sin\omega(t_j - \tau) \right]^2}{\sum_j \cos^2\omega(t_j - \tau)} \right\} \tag{13.8.4}
\]

Here \( \tau \) is defined by the relation

\[
\tan(2\omega\tau) = \frac{\sum_j \sin 2\omega t_j}{\sum_j \cos 2\omega t_j} \tag{13.8.5}
\]

\( P_N(\omega) \) has an exponential probability distribution with unit mean.

\[
P(>z) = 1 - (1 - e^{-z})^M \tag{13.8.7}
\]

is the false-alarm probability of the null hypothesis, that is, the significance level of any peak in \( P_N(\omega) \).

Source: *Numerical Recipes in C* (2nd Ed), p. 577
Mathematical Experiment: Single Dominant Frequency

Expression = Cosine(2πt/24)

Mathematical Experiment: Multiple Frequencies

Expression = 
Cosine(2\pi t/48) + 
Cosine(2\pi t/24) + 
Cosine(2\pi t/8)

Multiple peaks in periodogram. Corresponding valleys in significance curve.
Mathematical Experiment: Multiple Frequencies

Expression = 3*Cosine(2\pi t/48) + Cosine(2\pi t/24) + Cosine(2\pi t/8)

"Weaker" periodicities cannot always be resolved statistically.
Mathematical Experiment:
Multiple Frequencies: “Duty Cycle”

50% 66.6% (e.g., human sleep cycle)

One peak with symmetric “duty cycle”.
Multiple peaks with asymmetric cycle.
Mathematical Experiment:
Mixtures: Periodic Signal Vs. Noise

“p” histogram

100% periodic genes

50% periodic
50% noise

100% noise
Mathematical Experiment: Mixtures: Periodic Signal Vs. Noise

Multiple-Hypothesis Testing

More False Negatives

Bonferroni

Holm

Hochberg

Benjamini & Hochberg FDR

None

More False Positives

50% periodic, 50% noise
Data Pipeline to Apply to Bozdech’s Data

1. Apply quality control checks to data
2. Apply Lomb-Scargle algorithm to all expression profiles
3. Apply multiple hypothesis testing to define “significant” genes
4. Analyze biological significance of significant genes
Bozdech’s *Plasmodium* dataset:
1. Apply Quality Control Checks

Global views of experiment.
Remove certain outliers.
Bozdech’s *Plasmodium* dataset:

1. Apply Quality Control Checks

Many missing data points require imputation for Fourier analysis.
Bozdech’s *Plasmodium* dataset:

2. Apply Lomb-Scargle Algorithm

A weak diurnal period is visible in “mean” data profile.
Bozdech’s *Plasmodium* dataset:

2. Apply Lomb-Scargle Algorithm

Periodic Expression Patterns

Examples of highly-significant periodic expression profiles.
Bozdech’s *Plasmodium* dataset:

2. Apply Lomb-Scargle Algorithm

Aperiodic/Noise Expression Patterns
Bozdech’s *Plasmodium* dataset:

2. Apply Lomb-Scargle Algorithm

Small “N”
Bozdech’s *Plasmodium* dataset:

2. Apply Lomb-Scargle Algorithm

Signal and Noise Mixture

![Histogram of p values](histogram-log10p.pdf)
Bozdech’s *Plasmodium* dataset:
3. Apply Multiple-Hypothesis Testing

- More False Negatives
- Bonferroni
- Holm
- Hochberg
- Benjamini & Hochberg FDR
- None

More False Positives

---

Significance $\alpha = 1E-4$

Log10(p) vs. Rank Order of Sorted p Values

- bonferroni
- holm
- hochberg
- fdr
- none

(Using R’s `p.adjust` methods)
Bozdech’s *Plasmodium* dataset:

3. Apply Multiple-Hypothesis Testing

<table>
<thead>
<tr>
<th>p Adjustment Method</th>
<th>α Significance Level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.05</td>
</tr>
<tr>
<td>Bonferroni</td>
<td>3707</td>
</tr>
<tr>
<td>Holm</td>
<td>3995</td>
</tr>
<tr>
<td>Hochberg</td>
<td>4009</td>
</tr>
<tr>
<td>Benjamini &amp; Hochberg FDR</td>
<td>5618</td>
</tr>
<tr>
<td>None</td>
<td>5648</td>
</tr>
</tbody>
</table>

A priori plan: Use Benjamini & Hochberg FDR level of 0.0001.

Observed number of periodic probes consistent with biological observation of ~60% of *Plasmodium* genome being transcriptionally active during the intraerythrocytic developmental cycle.
Bozdech’s *Plasmodium* dataset:

### 4. Analyze Biological Significance

Lomb-Scargle: 4358 Probes, $\alpha = 1E-4$ significance

Comparison with Bozdech’s Results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>N time series points</th>
<th>Probes</th>
<th>Lomb-Scargle Periodic</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bozdech Complete</td>
<td>43 .. 46</td>
<td>5080</td>
<td>4115</td>
<td>81.0%</td>
</tr>
<tr>
<td></td>
<td>(Bozdech Quality Control Dataset)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>32 .. 42</td>
<td>1795</td>
<td>243</td>
<td>13.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6875</td>
<td>4358</td>
<td>63.4</td>
</tr>
</tbody>
</table>

While Lomb-Scargle identified 243 new low “N” periodic probes, the low percentage in that group may indicate some other problem.
Bozdech’s *Plasmodium* dataset:

4. Analyze Biological Significance

Lomb-Scargle: 4358 Probes, $\alpha = 1\text{E-4}$ significance

Comparison with Bozdech’s Results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Probes</th>
<th>Lomb-Scargle Periodic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bozdech Overview</td>
<td>3711</td>
<td>3611</td>
</tr>
</tbody>
</table>

Unclear how to apply Bozdech’s ad hoc “Overview” criteria for use with Lomb-Scargle method: “70% power in max frequency with top 75% of max frequency magnitude.”

The best 3711 Lomb-Scargle “p” values contained 3449 (92.9%) of the Overview probes.
Bozdech’s *Plasmodium* dataset:

4. Analyze Biological Significance

“Phaseograms”

Lomb-Scargle Results
4358 Probes

Bozdech: “Overview” Dataset
2714 genes, 3395 probes
Bozdech’s *Plasmodium* dataset:

4. Analyze Biological Significance

Lomb-Scargle: 4358 Probes, $\alpha = 1E-4$ significance

Periodogram Map

- Shows periodograms, not expression profiles
- Shows frequency space, not time
- Dominant frequency band corresponds to 48-hr period
- Are “weak” bands indicative of complex expression, perhaps a diurnal component, or an asymmetric “duty cycle”?
## Summary

<table>
<thead>
<tr>
<th>Lomb-Scargle Method</th>
<th>Fourier Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weights data points</td>
<td>Weights frequency intervals</td>
</tr>
<tr>
<td>No special requirement</td>
<td>Requires uniform spacing</td>
</tr>
<tr>
<td>No special processing</td>
<td>Missing data imputed</td>
</tr>
<tr>
<td>No special requirement</td>
<td>$2^N$ points for FFT; 0 padding</td>
</tr>
<tr>
<td>Known statistical properties</td>
<td>Permutation tests needed to assess statistical properties</td>
</tr>
<tr>
<td>Use “p” values</td>
<td>Ad hoc scoring rules</td>
</tr>
<tr>
<td>Need estimate of number of “independent frequencies” but explore using continuum</td>
<td>Usually only look at “independent” Fourier frequencies</td>
</tr>
</tbody>
</table>
Conclusions

• Lomb-Scargle periodogram is effective tool to identify periodic gene expression profiles
• Results comparable with Fourier analysis
• Lomb-Scargle can help when data are missing or not evenly spaced

We wanted to validate the Lomb-Scargle method before applying to our somitogenesis problem, since the Fourier technique would be difficult to use. Scargle (1982): “surprising result is that the … spectrum of a process can be estimated … [with] only the order of the samples …”
Conclusions

• Conclusions should not be drawn using the individual p-value calculated for each profile. A multiple comparison procedure False Discovery Rate (FDR) must be used to control the error rate.
• Expression profiles may be more complex than simple cosine curves
• Power spectra of non-sinusoid rhythms are more difficult to interpret
Supplementary Information

http://research.stowers-institute.org/efg/2004/CAMDA
Acknowledgements

Stowers Institute for Medical Research

Pourquie Lab
Olivier Pourquie
Mary-Lee Dequeant