Using Lomb-Scargle Periodograms to Identify Periodic Genes in Somitogenesis

Stowers Institute for Medical Research

Earl F. Glynn
Scientific Programmer
29 March 2006
Using Lomb-Scargle Periodograms to Identify Periodic Genes in Somitogenesis

- Periodic Patterns in Biology
- Simple Periodic Gene Expression Model
- Introduction to Lomb-Scargle Periodogram
- Data Pipeline
- Methodology Validation Study: Bozdech’s *Plasmodium* dataset
- Application to Mary-Lee’s somitogenesis 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
Simple Periodic Gene Expression Model

Gene Expression = Amplitude × Cosine(2πf t + PhaseShift)

“Periodic” if only observed over a single cycle?
Introduction to Lomb-Scargle Periodogram

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

<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>Data can be unevenly sampled</td>
<td>Requires uniform spacing</td>
</tr>
<tr>
<td>No data imputation</td>
<td>Missing data imputed</td>
</tr>
<tr>
<td>Any number of data points</td>
<td>$2^N$ points for FFT; 0 padding</td>
</tr>
<tr>
<td><strong>Known statistical properties</strong></td>
<td>Permutation tests needed to assess statistical properties</td>
</tr>
<tr>
<td>“p” value</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>
Lomb-Scargle Periodogram
Example Using Cosine Expression Model

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

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

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

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

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

(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

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

(13.8.4)

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

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

(13.8.5)

\[ P(> z) \equiv 1 - (1 - e^{-z})^M \]  

(13.8.7)

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

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

Source: Numerical Recipes in C (2nd Ed), p. 577
Lomb-Scargle Periodogram Experiment: Single Dominant Frequency

Expression = \cos(\frac{2\pi t}{24})

Lomb-Scargle Periodogram Experiment: Multiple Frequencies

Expression =
\[ \text{Cosine}(2\pi t/48) + \text{Cosine}(2\pi t/24) + \text{Cosine}(2\pi t/8) \]

Multiple peaks in periodogram. Corresponding valleys in significance curve.
Lomb-Scargle Periodogram Experiment: Multiple Frequencies

Expression = 
3*Cosine(2πt/48) + 
Cosine(2πt/24) + 
Cosine(2πt/8)

“Weaker” periodicities cannot always be resolved statistically.
Lomb-Scargle Periodogram Experiment: Multiple Frequencies: “Duty Cycle”

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

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

log(p) histogram

100% periodic genes  50% periodic  50% noise  100% noise
Lomb-Scargle Experiment: Mixtures: Periodic Signal Vs. Noise
Multiple-Hypothesis Testing

Bonferroni
Holm
Hochberg
Benjamini & Hochberg FDR
None

More False Negatives

More False Positives

50% periodic, 50% noise
Data Pipeline to Apply to Microarray Dataset

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


Intraerythrocytic Developmental Cycle of *Plasmodium falciparum*  
Critical Assessment of Microarray Data Analysis Conference (CAMDA), Nov 2004

http://research.stowers-institute.org/efg/2005/LombScargle/
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

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 showing periodic and aperiodic probes or noise](histogram-log10p.pdf)

- Missing Noise Spike?
Bozdech’s *Plasmodium* dataset:

3. Apply Multiple-Hypothesis Testing

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

Multiple Testing Correction Methods

(Using R’s `p.adjust` methods)

Significance

\[ \alpha = 1E-4 \]
Bozdech’s *Plasmodium* dataset:

4. Analyze Biological Significance

![Diagram showing sets and relationships between Bozdech's Plasmodium dataset subsets](image)
Bozdech’s *Plasmodium* dataset:

4. Analyze Biological Significance

Lomb-Scargle Results
4355 Probes

“Phaseograms”

Probes Ordered by Phase

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

**powerMAX vs Power at Peak Frequency**

Periodic Genes
Bozdech’s *Plasmodium* dataset: Bozec’s Ad Hoc Scoring Vs Lomb-Scargle p values
Bozdech’s *Plasmodium* dataset:
Bozdech’s “Phase” Vs. Peak of Smoothed Time Series
Mary-Lee’s Somitogenisis Dataset

We wanted to validate the Lomb-Scargle method with Bozdech’s dataset before applying to our somitogenesis problem, since the Fourier technique could not be used:

Scargle (1982):
“surprising result is that the … spectrum of a process can be estimated … [with] only the order of the samples …”
Somitogenesis Dataset

<table>
<thead>
<tr>
<th>N</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.1</td>
</tr>
<tr>
<td>17</td>
<td>0.006</td>
</tr>
<tr>
<td>20</td>
<td>0.002</td>
</tr>
<tr>
<td>48</td>
<td>3E-9</td>
</tr>
</tbody>
</table>
Somitogenesis Dataset

20 embryos = 20 time points

- At different phases of the Clock cycle, covering 5 cycles
- Reordering based on *in situ* hybridization

Mary-Lee, Science Club, 2004
Somitogenesis Dataset

Mary-Lee, Science Club, 2004
Somitogenesis Dataset

Do we care about time order, or only periodic genes?

22,690 Affy Probesets
Somitogenesis Dataset

Do we care about time order, or only periodic genes?

Plan: Use only known periodic genes

TimeOrderingExperiment1.doc, Nov 2004
Somitogenesis Dataset

Perturbation ("Jitter") and Permutation Tests

Perturbation (times):
27.75443 33.08618 42.6607 11.04462 -0.6288767 18.55415 12.93994 44.90501 59.98529 67.2538 46.87758 55.91625 74.72591 78.9921 83.98062 104.9910 78.12926 77.38472 111.5297 109.2331

Permutation (order for fixed times):
7 5 6 8 4 3 2 1 | 11 12 9 10 | 13 | 17 15 16 18 14 | 20 19
Somitogenesis Dataset

Perturbation ("Jitter") and Permutation Tests

Log10(p) Variability by Gene for 10,000 Perturbations
'p' values from Lomb-Scargle analysis

Best Rank Product Worst
JitterSpreadHistograms.pdf, Feb 2005

Log10(p) Variability by Gene for 10,000 Permutations
'p' values from Lomb-Scargle analysis

Best Rank Product Worst
PermutedSpreadHistograms.pdf
## Somitogenesis Dataset

### Perturbation ("Jitter") and Permutation Tests

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>O</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>All ProbeSET</td>
<td>Permuted Rank</td>
<td>Permuted GeneSymbol</td>
<td>Permuted logranks</td>
<td>Permuted top500</td>
<td>Permuted top100</td>
<td>Permuted top250</td>
<td>Permuted top500</td>
<td>Permuted Rank</td>
<td>Permuted GeneSymbol</td>
<td>Permuted logranks</td>
<td>Permuted top500</td>
<td>Permuted top100</td>
<td>Permuted top250</td>
<td>Permuted top500</td>
<td>Permuted RankProduct</td>
</tr>
<tr>
<td>2</td>
<td>1420360_at</td>
<td>1 Dkk1</td>
<td>2542.1</td>
<td>10000</td>
<td>10000</td>
<td>10000</td>
<td>10000</td>
<td>1 Dkk1</td>
<td>2371.1</td>
<td>10000</td>
<td>10000</td>
<td>10000</td>
<td>10000</td>
<td>10000</td>
<td>10000</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1418102_at</td>
<td>2 Hes1</td>
<td>5216.9</td>
<td>10000</td>
<td>10000</td>
<td>10000</td>
<td>10000</td>
<td>2 Hes1</td>
<td>4345.4</td>
<td>9996</td>
<td>10000</td>
<td>10000</td>
<td>10000</td>
<td>10000</td>
<td>10000</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>1427600_at</td>
<td>3 ---</td>
<td>8449.9</td>
<td>9836</td>
<td>981</td>
<td>10000</td>
<td>10000</td>
<td>3 ---</td>
<td>7684.7</td>
<td>9910</td>
<td>991</td>
<td>10000</td>
<td>10000</td>
<td>10000</td>
<td>10000</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>1425896_a_at</td>
<td>4 Ild1</td>
<td>13803.4</td>
<td>9282</td>
<td>991</td>
<td>10000</td>
<td>10000</td>
<td>4 Ild1</td>
<td>12392.8</td>
<td>9632</td>
<td>991</td>
<td>10000</td>
<td>10000</td>
<td>10000</td>
<td>10000</td>
<td>16</td>
</tr>
<tr>
<td>6</td>
<td>1426845_at</td>
<td>5 Asin2</td>
<td>17411.4</td>
<td>3924</td>
<td>6442</td>
<td>9387</td>
<td>9937</td>
<td>6 Asin2</td>
<td>16266.2</td>
<td>5139</td>
<td>7521</td>
<td>9577</td>
<td>9959</td>
<td>10000</td>
<td>10000</td>
<td>30</td>
</tr>
<tr>
<td>7</td>
<td>1422914_at</td>
<td>6 Sp5</td>
<td>17514.0</td>
<td>3864</td>
<td>5287</td>
<td>9541</td>
<td>9974</td>
<td>5 Sp5</td>
<td>16103.0</td>
<td>5504</td>
<td>7301</td>
<td>9967</td>
<td>9924</td>
<td>10000</td>
<td>10000</td>
<td>30</td>
</tr>
<tr>
<td>8</td>
<td>1426001_at</td>
<td>7 Eomes</td>
<td>20822.9</td>
<td>18.29</td>
<td>3791</td>
<td>7311</td>
<td>9386</td>
<td>7 Eomes</td>
<td>19885.0</td>
<td>2417</td>
<td>4771</td>
<td>8212</td>
<td>9673</td>
<td>49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1416039_x_at</td>
<td>8 Cys1</td>
<td>21436.2</td>
<td>1473</td>
<td>3230</td>
<td>6962</td>
<td>9176</td>
<td>8 Cys1</td>
<td>20580.4</td>
<td>1936</td>
<td>4069</td>
<td>7752</td>
<td>9477</td>
<td>64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1424942_a_at</td>
<td>9 Myc</td>
<td>22459.1</td>
<td>940</td>
<td>2546</td>
<td>8140</td>
<td>8744</td>
<td>10 Myc</td>
<td>21707.3</td>
<td>1170</td>
<td>3033</td>
<td>7053</td>
<td>9235</td>
<td>90</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

PermutedPerturbComparison.xls, Jan 2005
Somitogenesis Dataset

Perturbation ("Jitter") and Permutation Tests

Top 50: Somites Vs Random

StatComparison1000.xls: Top50, Jan 2005
Somitogenesis Dataset

Perturbation (“Jitter”) and Permutation Tests

StatComparison1000.xls: TopLogRankSum

Somites Vs Random
Somitogenesis Dataset

How many periodic genes are in the dataset?
Somitogenesis Dataset

What periodicities are present?
Periodogram Clustering

Hierarchical Clustering

119 min  59 min
Somitogenesis Dataset

Unconstrained Permutations to Estimate False Hit Rate
(10,000 Permutations, 17-point time series, 7544 Affy probes)

Use p-value cutoffs from “baseline” with permutations

<table>
<thead>
<tr>
<th>Top</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>50</th>
<th>75</th>
<th>100</th>
<th>200</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td>20</td>
<td>25</td>
<td>50</td>
<td>75</td>
<td>100</td>
<td>200</td>
<td>500</td>
</tr>
<tr>
<td>Permutated</td>
<td>0.2</td>
<td>2.7</td>
<td>5.2</td>
<td>9.8</td>
<td>12.2</td>
<td>25.5</td>
<td>42.9</td>
<td>60.9</td>
<td>164.5</td>
<td>462.7</td>
</tr>
<tr>
<td>&quot;Good&quot; Hits</td>
<td>4.8</td>
<td>7.3</td>
<td>9.8</td>
<td>10.2</td>
<td>12.8</td>
<td>24.5</td>
<td>32.1</td>
<td>39.1</td>
<td>35.5</td>
<td>37.3</td>
</tr>
</tbody>
</table>

About 35-40 cyclic genes detected?
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
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 may be difficult to interpret
Acknowledgements

Pourquie Lab
Olivier Pourquié
Mary-Lee Dequéant

Bioinformatics
Arcady Mushegian
Galina Glasko
Jie Chen