Are we overestimating the number of cell-cycling genes? The impact of background models for time series data

Matthias E. Futschik & Hanspeter Herzel
Institute for Theoretical Biology, Charité, Humboldt-University, Berlin, Germany

German Conference in Bioinformatics, Potsdam, 2007
Cell cycle

- Fundamental process underlying growth and reproduction
- Replication of DNA and cytokinesis
- Comprises 4 phases: G1, S, G2 & M
- Core machinery is well-studied (Cyclins and CDKs)
However:

- microarray measurements comprise high noise and experimental bias
- little overlap in periodic genes between independent studies
Identification of periodic gene expression patterns

**Challenging**
- Detection of periodic patterns occurring by chance due to large number of genes measured

**Various statistical approaches proposed**
- Recent comparison showed that Fourier-based scoring performs superior (de Lichtenberg et al., Bioinformatics, 2005):

\[
F[g] = \left( \sum_i \cos(2 \cdot \pi \cdot t_i / T) \cdot g_i \right)^2 + \left( \sum_i \sin(2 \cdot \pi \cdot t_i / T) \cdot g_i \right)^2
\]

\(g\): gene expression; \(t_i\) sampling time, \(T\): cell cycle period
Assessment of Significance

Significance is usually assessed by comparison to randomly permuted or Gaussian patterns

\[ FDR(t) = \frac{\sum \delta_i(F_b \geq t)/n}{\sum \delta_j(F_o \geq t)} \]
Autocorrelation in microarray time series data

Random or Gaussian background models assume that non-periodic patterns in microarray data are random i.e. do not show correlation between sequential measurements.

However, that is usually not the case.

Futschik & Carlisle, Journal of Computational Biology and Bioinformatics, 2006
Correlation structures can be captured by autoregressive processes:

$$X_t = \alpha_1 \cdot X_{t-1} + \alpha_2 \cdot X_{t-2} + \ldots + \alpha_p \cdot X_t + Z_t$$

$X$: time dependent variable; $Z$: independent random variable

**AR(1) process:** $X_t = \alpha_1 \cdot X_{t-1} + Z_t$

$\alpha_1$: autocorrelation for time lag of one; variance: $\sigma_z^2 = \sigma_x^2 (1 - \alpha_1^2)$
Examples of generated background time series
Influence on detected periodicity in background

Background signals with smallest Fourier scores

Background signals with median Fourier scores

Background signals with largest Fourier scores
Yeast cell cycle experiments re-analysed

- **CDC 28 by Chu *et al.*:**
  - Synchronization by temperature sensitive yeast cells (CDC28)
  - Sampling over 160 minutes in 10 minute intervals comprising two cell cycles
  - Affymetrix GeneChips (6000 genes) : 400 periodic genes

- **CDC 15 by Spellman *et al.*:**
  - Synchronization using CDC15 strain
  - Sampling over 270 minutes comprising three cell cycles
  - cDNA arrays (6000 genes) : 800 periodic genes
Data processing

- Log2-transformation
- Missing values by the knn-method
- Standardization: $\mu = 0$ and $\sigma = 1$
- For each experiment, 100 independent random datasets of the same size were generated for each background models
Autocorrelation

- Successive measurements show correlation in original data sets
- Random and Gaussian background models do not capture correlation structures
- AR(1) reflects short-time correlation
Distribution of Fourier scores

Calculation of Fourier Scores for observed data and background models

- Random and Gaussian background models led to very similar distributions with small number of high scores
- AR(1)-based background model yielded a larger number of high-scoring expression vectors
Power spectrum

Power spectrum $I$ represents the strength of periodic components in a signal with respect to their frequency

$$I[f_p] = \left[ \left( \sum_i \cos(2 \cdot \pi \cdot f_p) \cdot g_i \right)^2 + \left( \sum_i \sin(2 \cdot \pi \cdot f_p) \cdot g_i \right)^2 \right] / N\pi$$
- AR(1) processes can lead to a higher spectral density depending on cell cycle period and autocorrelation coefficient.

- Observed strong enrichment of autocorrelation coefficients in the critical range, where correlation leads to higher spectral density.

- Larger Fourier scores for the observed cell cycle frequencies compared to random processes.
Significance is strongly dependent on chosen background model

- Random and Gaussian background result in very similar number of significant genes
- AR(1) background leads to a considerable reduction of the number of significant genes independent of the chosen FDR.
<table>
<thead>
<tr>
<th>Background Model</th>
<th>CDC15</th>
<th>CDC28</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>258</td>
<td>192</td>
<td>0.01</td>
</tr>
<tr>
<td>Gauss</td>
<td>257</td>
<td>152</td>
<td>0.05</td>
</tr>
<tr>
<td>AR(1)</td>
<td>119</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td>420</td>
<td>448</td>
<td></td>
</tr>
<tr>
<td>Gauss</td>
<td>413</td>
<td>419</td>
<td></td>
</tr>
<tr>
<td>AR(1)</td>
<td>257</td>
<td>52</td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td>551</td>
<td>649</td>
<td>0.10</td>
</tr>
<tr>
<td>Gauss</td>
<td>527</td>
<td>614</td>
<td></td>
</tr>
<tr>
<td>AR(1)</td>
<td>326</td>
<td>126</td>
<td></td>
</tr>
</tbody>
</table>

Number of genes detected as significantly periodically expressed
Assessment of detected genes

Benchmark sets for yeast cell-cycle genes

- **Small scale experiment:** comprises a total of 113 genes identified as periodically expressed in small scale experiments
- **Chromatin IP:** consists of 352 genes which underlie the control of known cell cycle transcription factors
- **MIPS:** comprises 518 genes annotated in MIPS as “cell cycle and DNA processing” after the exclusion of genes included in the two other benchmark sets.
<table>
<thead>
<tr>
<th>Benchmark Set</th>
<th>CDC15 Random</th>
<th>CDC15 AR(1)</th>
<th>CDC28 Random</th>
<th>CDC28 AR(1)</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small scale experiments</td>
<td>0.21</td>
<td>0.31</td>
<td>0.19</td>
<td>0.66</td>
<td>0.01</td>
</tr>
<tr>
<td>Chromatin IP</td>
<td>0.17</td>
<td>0.18</td>
<td>0.18</td>
<td>0.33</td>
<td>0.01</td>
</tr>
<tr>
<td>MIPS</td>
<td>0.10</td>
<td>0.06</td>
<td>0.20</td>
<td>-</td>
<td>0.05</td>
</tr>
<tr>
<td>Small scale experiments</td>
<td>0.15</td>
<td>0.23</td>
<td>0.11</td>
<td>0.32</td>
<td>0.05</td>
</tr>
<tr>
<td>Chromatin IP</td>
<td>0.16</td>
<td>0.17</td>
<td>0.12</td>
<td>0.19</td>
<td>0.05</td>
</tr>
<tr>
<td>MIPS</td>
<td>0.11</td>
<td>0.10</td>
<td>0.18</td>
<td>0.19</td>
<td>0.05</td>
</tr>
<tr>
<td>Small scale experiments</td>
<td>0.13</td>
<td>0.18</td>
<td>0.09</td>
<td>0.21</td>
<td>0.10</td>
</tr>
<tr>
<td>Chromatin IP</td>
<td>0.14</td>
<td>0.18</td>
<td>0.10</td>
<td>0.19</td>
<td>0.10</td>
</tr>
<tr>
<td>MIPS</td>
<td>0.10</td>
<td>0.11</td>
<td>0.16</td>
<td>0.25</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Positive predicted value
Conclusions

- Autocorrelation in microarray data interferes with the detection of periodically expressed genes
- Choice of background model has drastic effect on the number of genes detected as significantly periodically expressed
- Random and Gaussian background models neglect the dependency structure within the observed data and lead to overestimation of significance
- AR(1)-based background models give a more adequate representation of correlations between measurements and reduced, but more reliable number of significantly periodically expressed genes
Thank you!

Questions?