Identifying Periodically Expressed Transcripts in Microarray Time Series Data

Korbinian Strimmer
Department of Statistics, University of Munich, Germany.

Munich, 19 February 2004

Motivation and Questions

- Many time series data sets monitoring gene expression during the cell-cycle are available. .
- We want to learn about cell cycle regulated genes. How can we identify the subset of genes with a clear periodic signature?
- Controversy about data quality
- Problem of synchronization methods
- Statistical significance of results (chance fluctuations)?

Program of Work

Guideline: avoid reinventing the wheel, use standard statistical time series methods (no ad-hoc approaches please..)

- Visual assessment via average periodogram
- Test of hidden periodicities (i.e. unknown periodically expressed genes) using Fisher's exact g test
- Multiple testing with False Discovery Rate
- Reassessment of most popular benchmark data sets.

Periodogram

Consider gene expression time series data $Y_1, ..., Y_N$:

Then the *periodogram* is the corresponding power spectrum:

$$I(\omega) = \frac{1}{N} \left| \sum_{t=1}^{N} Y_t \exp(-i\omega t) \right|^2, \quad \omega \text{ discrete } \in [0, \pi] \tag{1}$$

A simple graphical device is to search for *significant peaks* in $I(\omega)$.

Average Periodogram

When Y_{it} denotes the ith observed time series at time t where $i = 1, \ldots, G$ and $t = 1, \ldots, N$.

The *average periodogram* can then be defined as:

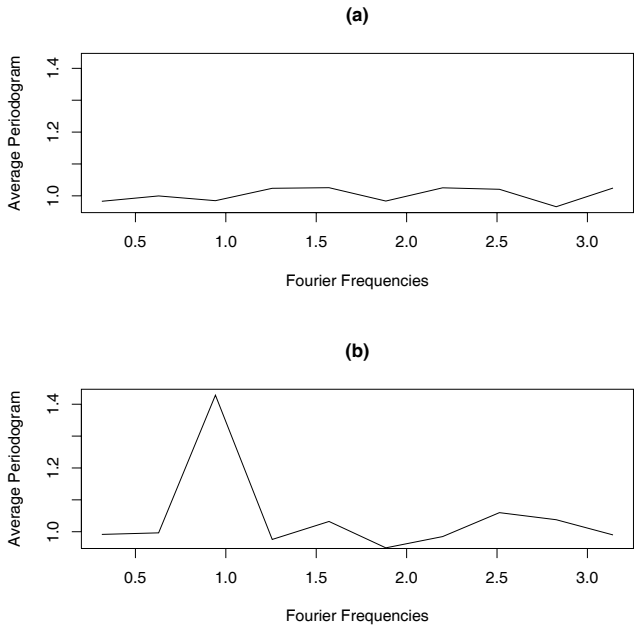
$$AI(\omega) = \frac{1}{G} \sum_{i=1}^{G} I_i(\omega), \tag{2}$$

where $I_i(\omega)$ is the periodogram of the ith time series.

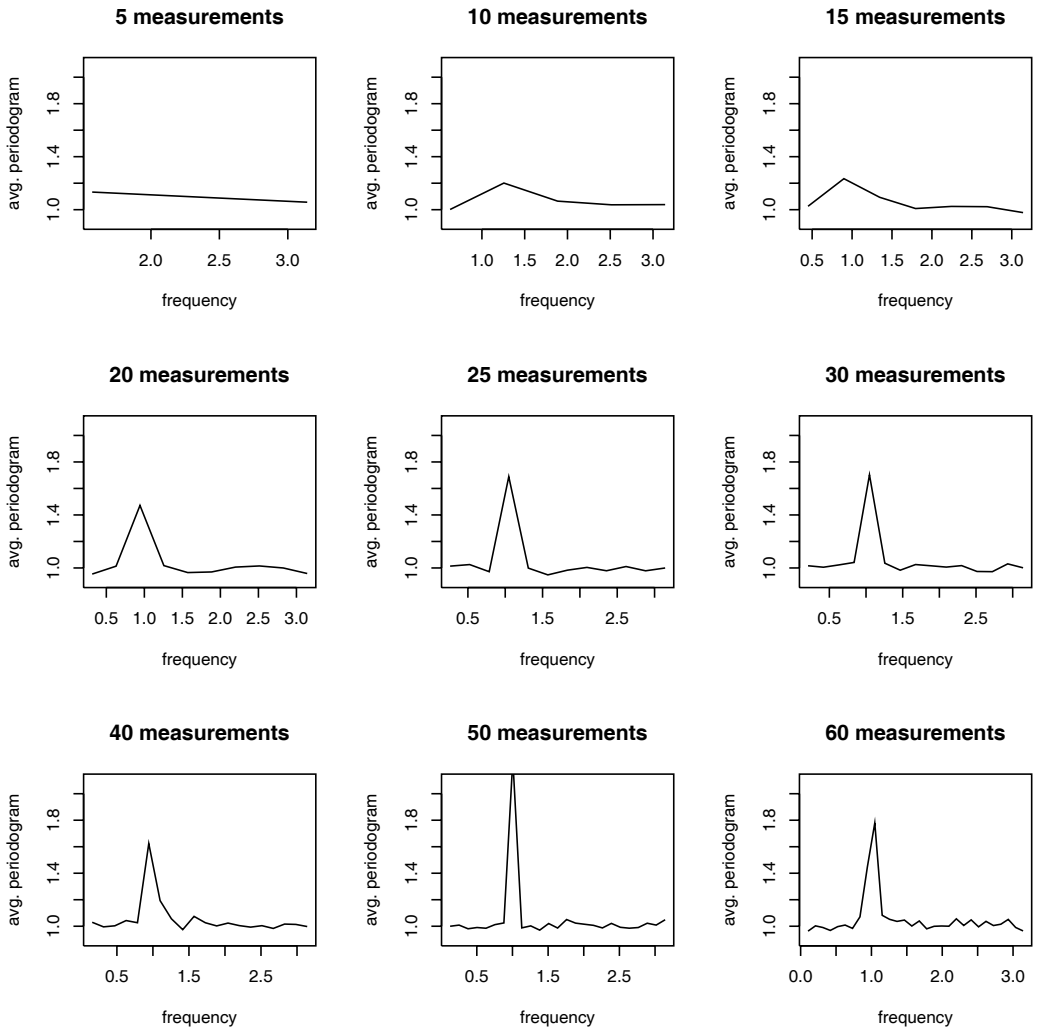
Graphical Interpretation

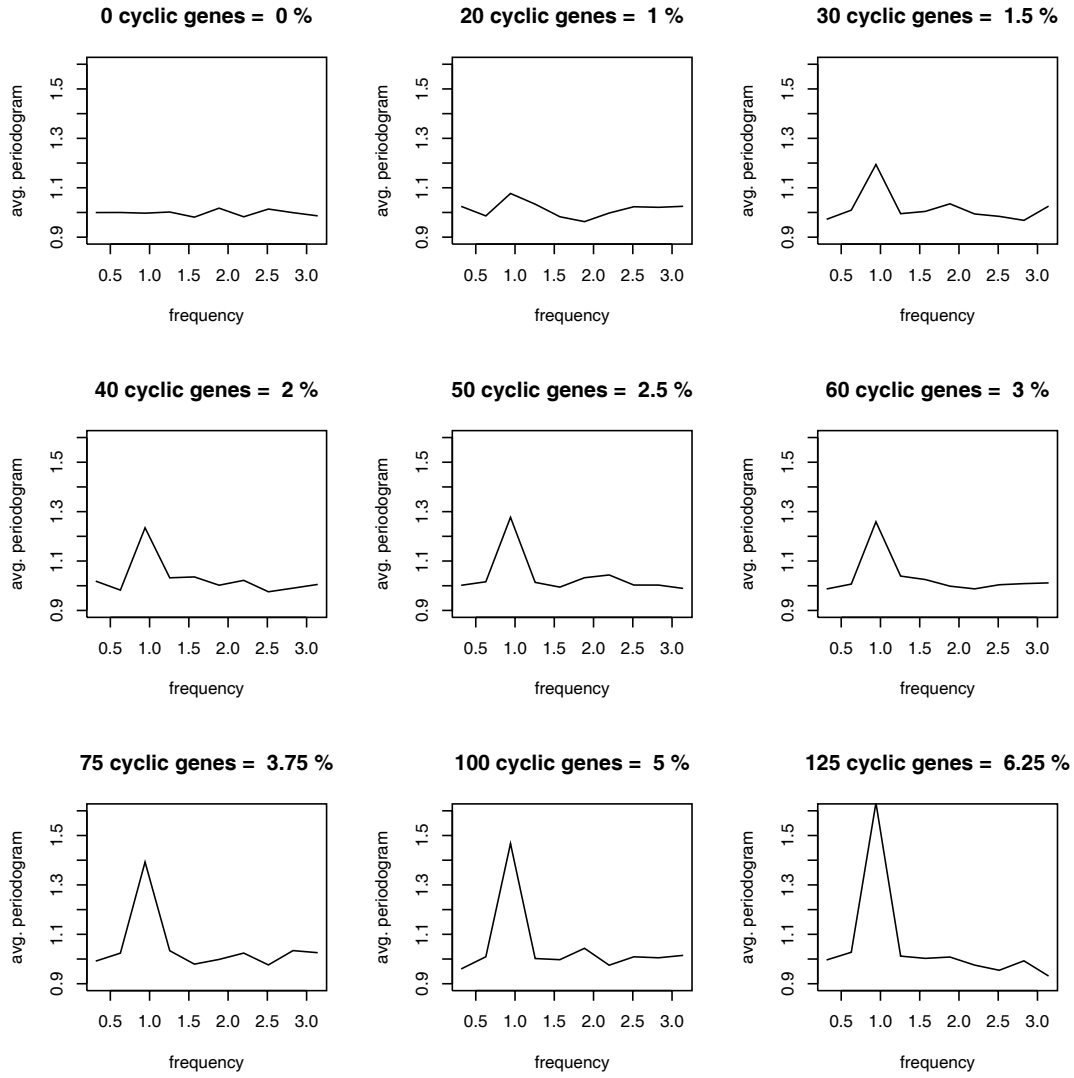
- If a time series contains a sinusoidal component with frequency $\omega_0 \in [0, \pi]$, then the periodogram exhibits a peak at that frequency.
- If a time series does not contain a periodic component, then the periodogram reduces to a straight line.
- Hence, if there is no periodic component in the data then the average spectral density of all time should will also reduce to a straight line.
- If there are a few time series exhibiting a strong cycle then their corresponding periodogram ordinates contribute a large amount to the average periodogram and consequently any visible peak should indicate the presence of a periodic component.

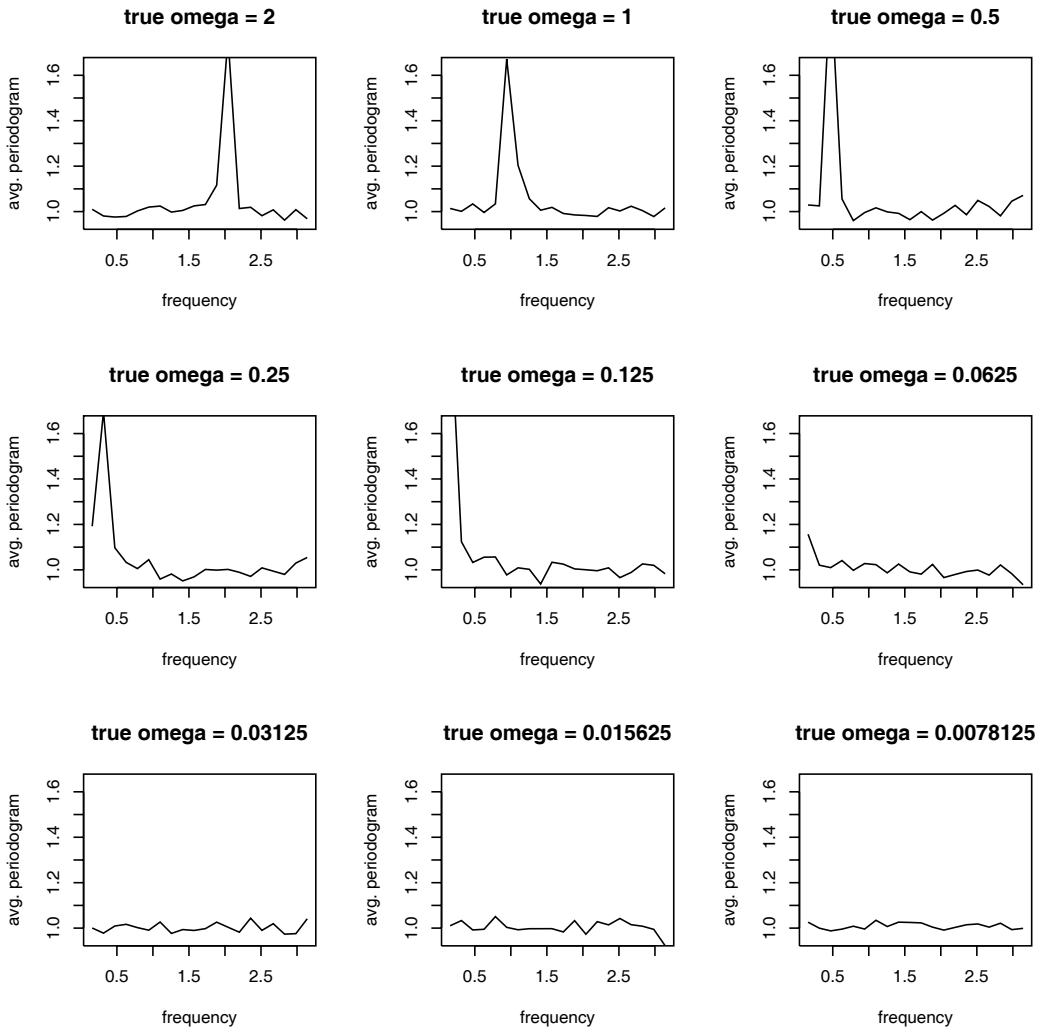
Example of the Average Periodogram



Average Periodogram for simulated data with 2000 time series (genes) of length 20. (a) corresponds to a white noise process. (b) includes 100 time series with frequency 1.


5 measurements
10 measurements
15 measurements
20 measurements
25 measurements
30 measurements
40 measurements
50 measurements
60 measurements
avg. periodogram
frequency


0 cyclic genes = 0 %
20 cyclic genes = 1 %
30 cyclic genes = 1.5 %
40 cyclic genes = 2 %
50 cyclic genes = 2.5 %
60 cyclic genes = 3 %
75 cyclic genes = 3.75 %
100 cyclic genes = 5 %
125 cyclic genes = 6.25 %
avg. periodogram
0.9
1.1
1.3
1.5
0.5
1.0
1.5
2.0
2.5
3.0
frequency


true omega = 2
true omega = 1
true omega = 0.5
true omega = 0.25
true omega = 0.125
true omega = 0.0625
true omega = 0.03125
true omega = 0.015625
true omega = 0.0078125
avg. periodogram
frequency
1.0
1.2
1.4
1.6
0.5
1.5
2.5

Fisher's Exact g Test

Fisher (1929!) proposed a test of hidden periodicities in a time series using the maximum periodogram coordinate by introducing the *g statistic*:

$$g = \frac{\max_k I(\omega_k)}{\sum_{k=1}^{[N/2]} I(\omega_k)}. \tag{3}$$

- The distribution of g can be computed exactly (important for small sample size!) under a Gaussian process.
- Large values of g lead to the rejection of the null hypothesis.

Gene Selection and Multiple Testing

- For each gene calculate the g-statistic and corresponding p-value

- Use FDR multiple comparison procedure on ordered p-values $p_{(1)}, p_{(2)}, \ldots, p_{(G)}$:

  1. Let i_q be the largest i for which $p_{(i)} \leq \frac{i}{G} q$,
  2. then reject the null hypothesis for all genes $g_{(1)}, g_{(2)}, \ldots, g_{(i_q)}$.

  It can be shown that this procedure controls the FDR at level q (Benjamini and Hochberg 1995).

FDR Simulations

$N =$	10	20	40	45	50	100	200
$q = 0.15$	3	21	65	103	118	121	117
$q = 0.10$	1	13	41	97	114	111	111
$q = 0.05$	1	3	30	90	107	104	104
$q = 0.01$	0	2	1	78	99	99	100
$q = 0.001$	0	0	0	45	88	99	99
Z	10	52	64	93	98	100	100

The simulations were carried out with 1900 random genes and 100 periodic genes. N is the sample size, q the desired FDR level (expected type I error), and Z the number of correctly identified periodic genes among the first 100 genes ranked according to their p-values.

Summary and Recipe for Analysis

1. Check graphically using the average periodogram whether or not there are periodic components in the data.

2. For each time series calculate Fisher's g statistic.

3. For each of the test statistic calculate the corresponding p-value.

4. Identify the genes that show strong cyclic behavior under the desired FDR level (e.g. $q = 0.05$).

Investigated Molecular Data Sets

The datasets we have used are all available on the web or in public data bases.

- Yeast *Saccharomyces cerevisiae* (4, Spellman *et al.*,1998)
- *Caulobacter crescentus* bacterial cell cycle (1, Laub *et al.*,2000)
- Human fibroblasts (2, Cho *et al.*,2001)
- Human cancer cell line (5, Whitfield *et al.*,2002)

Results from FRD g Test

Cell type	Experiment	N	G	C	C/G
Yeast	cdc15	24	4289	766	17.9%
Yeast	cdc28	17	1365	105	7.7%
Yeast	alpha	18	4415	468	10.6%
Yeast	elution	14	5695	193	3.4%
Caulobacter crescentus	bacteria	11	1444	44	3.0%
Human Fribroblasts	N2	13	4574	0	0%)
Human Fribroblasts	N3	12	5079	0	0%
Human HeLa	score1	12	14728	0	0%
Human HeLa	score2	26	15472	134	0.9%
Human HeLa	score3	48	39724	6043	15.2%
Human HeLa	score4	19	39192	56	0.1%
Human HeLa	score5	9	34890	0	0%

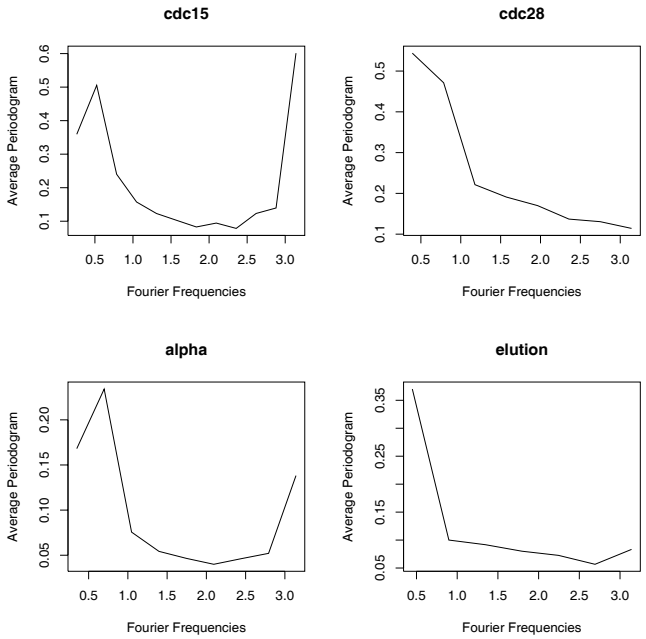
Notation: N is the sample size, G the total number of genes, C the number of periodic genes that are statistically significant for a FRD level of $q = 0.05$).

Yeast Cell Cycle

- four gene expression experiments
- three cell cycle synchronization techniques:
  - temperature arrest (cdc15, cdc28)
  - alpha factor arrest (alpha)
  - elutriation synchronization (elution)

Controversy: synchronization method may affect the results.

Average periodograms for Yeast



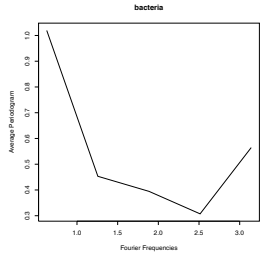
Clear signal of periodicity

Results obtained for Yeast

- the number of cyclic genes for cdc15 is similar to the one found in previous studies, whereas we detected a smaller number of cyclic genes for the other synchronization methods
- elution data provides little statistically significant information with regard to cell cycle regulation
- difficult to distinguish cell cycle specific genes from an artifact of the method used to synchronize the cells
- Only few periodic genes are identical across experiments

Bacterial Cell Cycle

- shortest time series considered (1444 genes measured over 11 time points)



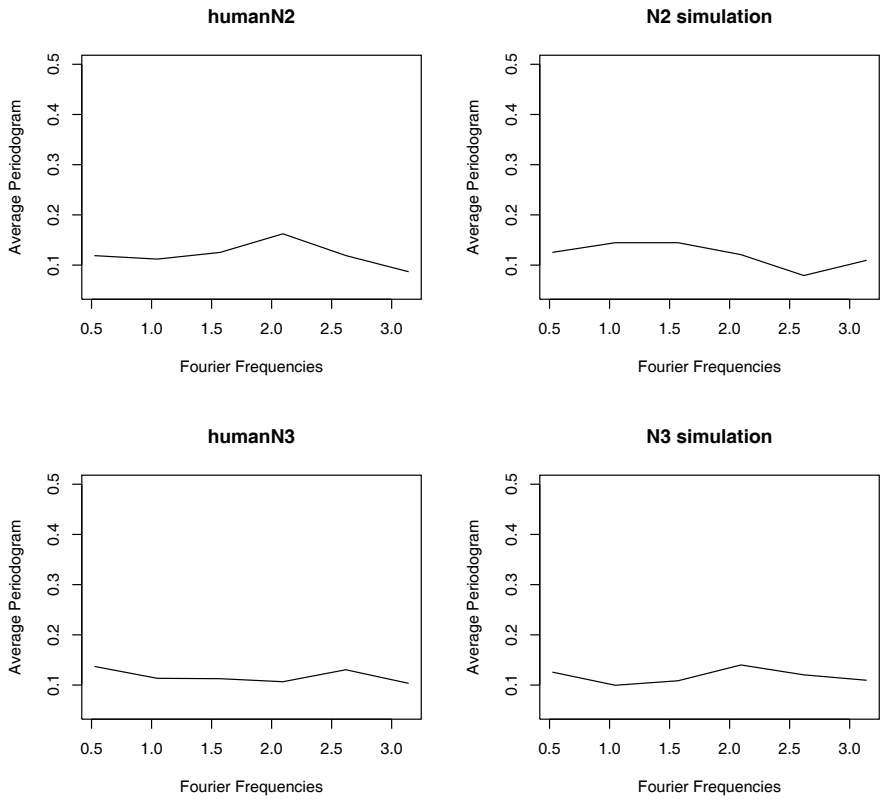
- the average periodogram indicates the presence of cell cycle specific genes

- the g-test identified 44 significant periodically expressed genes (as compared with Laub *et al.* (2000) who have found 553).

Human Fibroblasts Cell Cycle

- consist of two microarray experiments
  - `N2` with 13 time points per gene
  - `N3` with 12 time points per gene

Average Periodogram for Human Fibroblasts



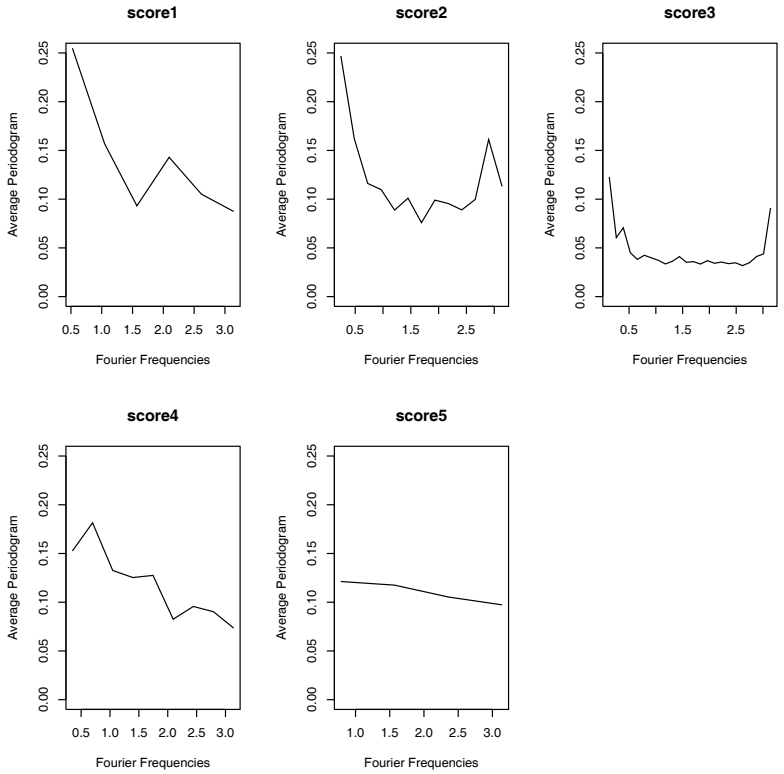
Results obtained from the Human Fibroblasts data set

- the average periodogram provides no evidence of periodic gene expression in the overall data
- the formal g-test statistic did also not detect any periodicity
- these results cannot be due to the short sample size. E.g. the $elution$ (from the yeast experiment) and the $bacteria$ data sets are also very short
- previous authors have raised doubts, on biological grounds, of whether this data is suited for statistical analysis (Shedden and Cooper, 2002)

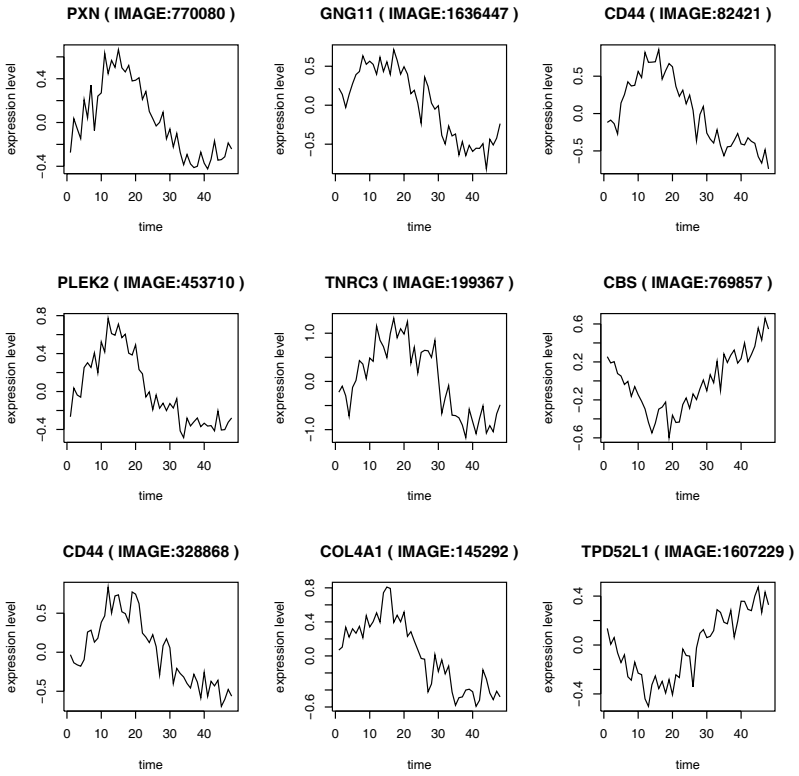
Human Cancer Cell Line

- consists of five experiments (`score1`, `score2`, `score3`, `score4`, `score5`) measuring the expression levels of approximately 30,000 genes
- three cell cycle synchronization methods were used
  - double thymidine block (`score1`, `score2`, `score3`)
  - thymidine followed by arrest in mitosis with Nocodazole (`score4`)
  - mitotic shake-off using an automated cell shake (`score5`)
- measurements were taken for up to 48 time points (`score3`), which makes this study one of the most extensive microarray time series experiment so far

Average Periodogram of Human Cancer Cell Line



Most Significant Periodic Genes (Hela/score3)



Results obtained from the Human Cancer Cell Line

- the average periodogram showed evidence of periodicity in all data sets except `score5` (which shows a flat line)
- the g-test statistic detected the occurrence of periodic genes in all scores apart from `score1` and `score5`
- we were able to find a substantial amount of additional periodic genes in `score3` compared with other previous studies (e.g. Whitfield *et al.*, 2002)
- but results may also be due to cell perturbations rather than generic cell cycle effects.

Summary

- we have suggested two statistical tools for microarray time series analysis:
  - the average periodogram as an exploratory device to assess the presence of periodically expressed transcripts
  - a formal statistical test for gene selection based on Fisher's g-statistic and the FDR multiple testing that allows screening for individual periodic genes
- Application to real microarray data shows that many data sets do not contain many statistically significant periodic genes (and often these are due to shock-response rather than generic activity).

Advantages

- our approach is applicable to data sets with a small number of measurements per gene
- average periodogram is essentially a non-parametric approach that takes advantage of the parallel structure of the data
- the g statistic allows to detect periodically expressed genes even with small amplitudes and is well defined for finite samples
- FDR multiple testing instead of arbitrary cut-off values

Cons

- g statistic assumes a null-model a purely Gaussian process
  - it is unclear whether this is a valid assumption for microarray time series data (maybe preprocessing steps help).

- investigated genes are correlated - this may have an adverse impact on the analysis (though non-independence is not critical for FDR testing).

- Open problem of interpreting statistically significant genes: how can one distinguish perturbation artifacts from generic cell cycle activity

Further conclusions

- there seems to be a remarkable gradient of signal quality in the data sets publicly available

- for a reliable detection of cell-cycle-regulated genes at least 40 time points per gene should be sampled.

Reference and Software

Reference:

Wichert, S., K. Fokianos, and K. Strimmer. 2004. Indentifying periodically expressed transcripts in microarray time series data. Bioinformatics **20**:5–20.

Software:

The methods presented are implemented in the R package "GeneTS" written by our group.

GeneTS is available from the R package archive CRAN and from the Bioconductor web page.