

Detection of hidden periods in microarray data

Jörg Aßmus, Hans Arnfinn Karlsen, Dag Tjøstheim

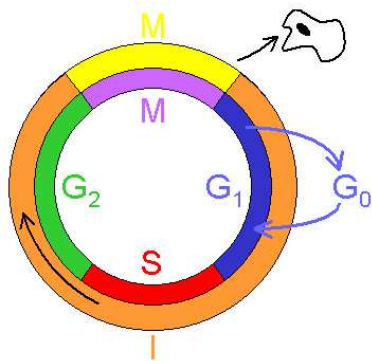


Outline

1. Introduction - Cell cycles
2. Method - Priodogram based tests
3. A simulation study
4. Data Analysis
5. Conclusion



Introduction I: Cell cycles - CV of the cell



Life cycle of a cell

- controlled by proteins
- information carrier: DNA, RNA

Reasonable idea:

Cycle dependent or periodic effects visible in the protein activity or RNA-transmission

Questions:

1. How do we handle different states of the cells in a culture?
2. How do we observe the activity?
3. How do we detect periodicities?

Introduction II: Synchronization

Approaches:

1. Selection:

Selection of cells assumed to be in the same state according to exterior criteria:

- *Elutriation*: Size

2. "Stop and Go":

(a) Stopping the cell cycle at a predefined state

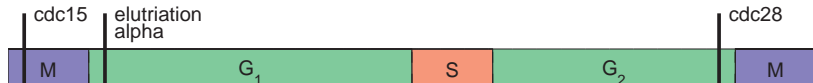
(b) Waiting until all cells have reached this point

(c) Restart at the same time

- α -factor: Stopping by adding an α -factor

- *cdc15*: Stopping by influence proteine *cdc15* (heat)

- *cdc28*: Stopping by influence proteine *cdc28* (heat)



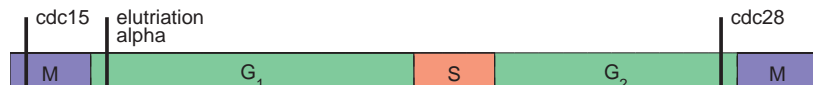
Introduction III: Microarray Data

A well-known benchmark data set

Gene expressions of yeast (*Saccharomyces cerevisiae*) monitored over one cell cycle with length T using different synchronizing techniques (Spellman [1998], Cho [1998])



Method	N	M	Cycle start	Reference
elutriation	5981	14	early G_1	Spellman [1998]
alpha	4489	18	early G_1	Spellman [1998]
cdc15	4381	24	M	Spellman [1998]
cdc28	6214	17	late G_2	Cho [1998]



Methods I: Statistical Approach

Model: Consider a panel $\mathbf{Y} = \{Y_{ij}\}_{i=1,\dots,N;j=1,\dots,M}$ with

$$Y_{ij} = A_i \cos[\omega_i(t_j + \phi_i)] + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathbf{N}[0, \sigma^2]$$

Test problem: For each gene i we test

$H_{i,0} : A_i = 0$ i.e. there is **no** periodicity

$H_{i,1} : A_i \neq 0$ i.e. there is periodicity

Methods:

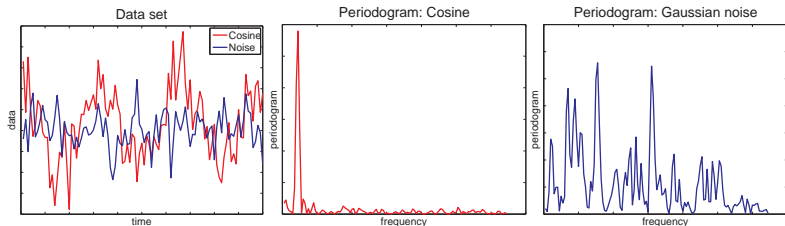
- Periodogram based tests (Fisher, PKF)
- Multiple testing methods (>4000 tests!)

Previous investigations:

- Monitoring (Spellman, 1998; Cho, 1998)
- Fisher test, average histogram (Wichert et al., 2004)
- Fisher, reduced PFK, LR-test (Aßmus, 2006, PhD)



Methods II: Periodogram based tests



Fisher-test (Fisher [1929])

Idea: Is there any dominating peak in the histogram?

$$T_{FS} = \frac{\max_k P_{yy,i}(\omega_k)}{\sum_{k=1}^{M/2} P_{yy,i}(\omega_k)}$$

Exact distribution

PFK-test (Brockwell-Davis [1986])

Idea: Is there at a predefined ω_0 a dominating peak?

$$T_{PFK} = \frac{(M-3)I_i(\omega_0)}{\sum_{j=1}^{M/2} Y_{ij}^2 - I_i(0) - 2I_i(\omega_0)}$$

Asymptotically $F_{2,M-3}$

Methods III: Distribution of the p -values

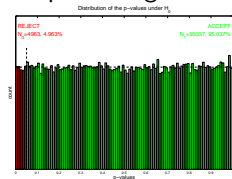
Multiple testing

Testing for each gene

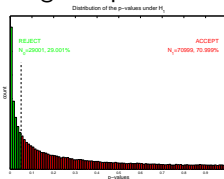
\Rightarrow > 4000 p -values

2 ideal cases \longrightarrow

No periodic genes



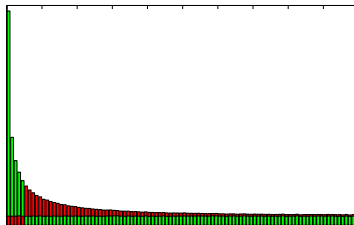
All genes periodic



Mixture of periodic and non-periodic genes

Methods to handle:

- Bonferroni
- False discovery rate

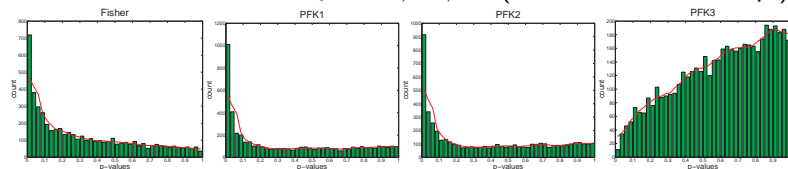


Simulation Experiment

Data set: Generation of 5000 gene expressions at 15 time points.

- 2500 genes with one full oscillation
- 2500 genes with two full oscillations

Tests: Fisher, PFK for $\omega_0 = 2\pi, 4\pi, 6\pi$ (1,2,3 oscillations resp.)



Conclusions:

- Incorrectly formulated alternatives. ($H_{i,0} : A_i = 0$)

Fisher $H_{i,1} : A_i \neq 0, \omega \in \{\omega_1, \dots, \omega_{M/2}\}$

PFK $H_{i,1} : A_i \neq 0, \omega = \omega_k$

- Wrong distribution of p -values in certain cases
⇒ Usual test criteria do not work properly
- PFK is more powerful

Yeast data I: Analysis

Tests:

1. Fisher-test
2. PFK-test, $\omega_0 = 2\pi/T$ - 1 oscillation per cycle (PFK 1)
3. PFK-test, $\omega_0 = 4\pi/T$ - 2 oscillations per cycle (PFK 2)
4. PFK-test, $\omega_0 = 6\pi/T$ - 3 oscillations per cycle (PFK 3)

Test results:

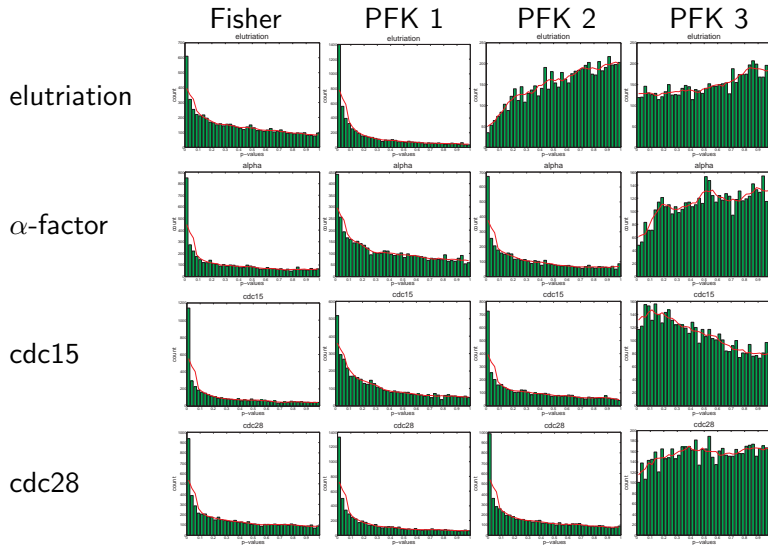
Rejections at 5%-level (ignoring multiple effects!)

n_1 : number of the detected periodically expressed genes

data set	Fisher		PFK 1		PFK 2		PFK 3	
	n_1	%	n_1	%	n_1	%	n_1	%
elutriation	931	15.57	1242	20.77	33	0.55	89	1.49
alpha	1124	25.04	696	15.50	927	20.65	101	2.25
cdc15	1433	32.71	812	18.53	979	22.35	239	5.46
cdc28	1324	21.31	1835	29.53	1344	21.63	239	3.85

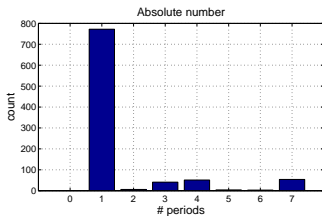


Yeast data II: p-value distributions

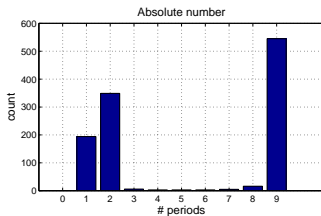


Yeast data V: Frequency maxima

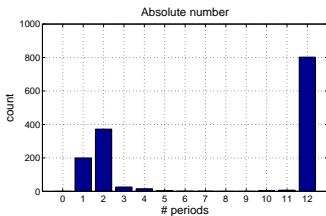
Elutriation



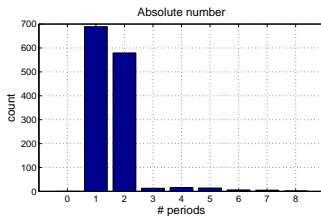
α -factor



cdc15



cdc28



Conclusions

1. Method

- Fisher-test: · Reliable but low power
 - We don't know, what we see
- PFK-test: · Better power than Fisher-test
 - Specified test for interesting frequencies
 - Hypotheses do not cover all possible effects
 - ⇒ multiple methods not applicable (No FDR!)
- Cosine Model very restrictive (⇒ other models? LR-test?)

2. Data analysis

- A large amount of periodically expressed genes detected
 - Mainly one and two oscillations per cell cycle
- Very different results for the considered synchronizations
- Many artefacts in α -factor, cdc15-data
- Newer data should be investigated

