

Clustering periodically-expressed genes using microarray data: a statistical analysis of the yeast cell cycle data

Jim Booth

Department of Statistics

University of Florida

<http://www.stat.ufl.edu/~jbooth>

jbooth@stat.ufl.edu

Background

Joint work with **George Casella**, **Janice Cooke** and **John Davis** at the University of Florida.

Arose out of a genomics discussion group at the University of Florida.
Two key papers are

- **Spellman et al.** (1998), *Molecular Biology of the Cell*
- **Alter et al.** (2000), *Proc. Nat. Acad. Sci.*

The papers concern statistical techniques for identifying and classifying cell cycle-regulated genes in the yeast genome; specifically

- **Fourier Analysis** (Spellman et al.)
- **Singular Value Decomposition** (Alter et al.)

Goals

- To explain and compare the statistical techniques used in the Spellman and Alter papers.
- Relate them to standard statistical techniques.
- Develop new statistical tools for the analysis of this and similar data.

Yeast Cell Cycle Data

- 10 million yeast cells required to harvest enough RNA to produce a microarray
- Synchronized population of cells produced by
 - elutriation (size based)
 - alpha-pheromone arrest
 - temperature based arrest

- 2-channel competitive hybridization
 - **Treatment** RNA (synchronized cells) used to to synthesize a cDNA-Cy5 labelled probe (**red**)
 - **Control** RNA (unsynchronized cells) used to to synthesize a cDNA-Cy3 labelled probe (**green**)
 - Expression or intensity level measures the amount of cDNA “hybridized” to chip
- Measurement is ratio of Cy5 to Cy3 expression levels

$$y = \log(\text{expression ratio})$$

Why take logarithms? Symmetry: $\log(1/2) = -\log(2)$

Data Matrix

- $y_{ij} = j$ th measurement (i.e. log expression ratio) on i th gene.
- $Y = \{y_{ij}\}$, data matrix

gene	time/treatment			
	t_1	t_2	\cdots	t_m
1	y_{11}	y_{12}	\cdots	y_{1m}
	\vdots			\vdots
i	y_{i1}	y_{i2}	\cdots	y_{im}
	\vdots			\vdots
n	y_{n1}	y_{n2}	\cdots	y_{nm}

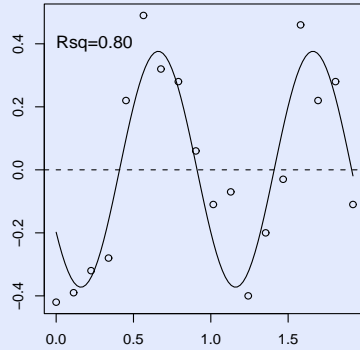
- Preprocessing/background noise: intensity measurements for each spot are summaries based on grid of pixels.

Yeast data from **Spellman et al.** (1998) and **Cho et al.** (1998)
($n = 6178$ genes)

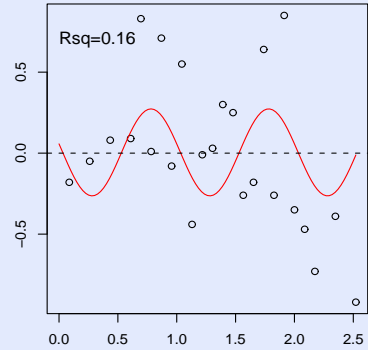
- **Elutriation:** $m = 14$ at 30 minute intervals
- **Alpha-factor:** $m = 18$ at 7 minute intervals
- **Cdc15-arrest:** $m = 24$ at 10 or 20 minute intervals
- **Cdc28-arrest:** $m = 17$ at 10 minute intervals
(Oligonucleotide array data)

<http://genome-www.stanford.edu/cellcycle/data/rawdata/>

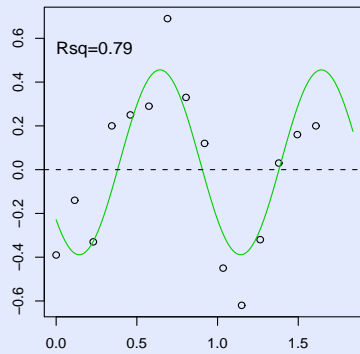
alpha-factor



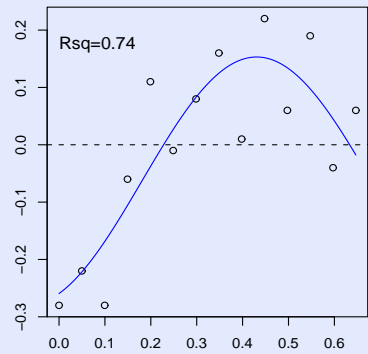
cdc15



cdc28



elutriation



Spellman et al. modeled the variation in log expression ratios over the course of the cell cycle for each gene using a linear combination of cosine and sine waves:

$$y(t) = \frac{a_0}{2} + a_1 \cos(2\pi t/T + \theta) + b_1 \sin(2\pi t/T + \theta)$$

- T is the length or period of the cell cycle
- θ is the initial phase
- Times of peak expression above and below the mean are two solutions ($T/2$ apart) of the equation:

$$\tan(2\pi t/T + \theta) = b_1/a_1$$

Corresponding angles $\phi = 2\pi t/T$ determine opposite points on the unit circle.

Model Fitting

- Estimates of the Fourier coefficients can be obtained by a **least squares fit** of the log expression profiles to the linear model

$$y_{ij} = \frac{a_{0i}}{2} + a_{1i} \cos(2\pi t_j/T) + b_{2i} \sin(2\pi t_j/T) + e_{ij}$$

- **Goodness-of-fit** of Fourier model to each gene's expression profile measured by R^2 .
- Rank genes by their P-values. Assuming **Gaussian errors**:

$$P = P(R^2 \geq R_{obs}^2) = P(F > F_{obs})$$

- The period T is **estimated separately**. However, R^2 value independent of initial phase!

Alpha-factor experiment

$m = 18$ arrays

$n = 5917$ genes + 103 known with fewer than 4 missing values

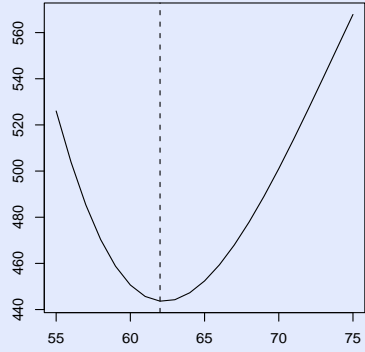
Numbers of genes with P-values lower than a threshold

P-value	Expected	Actual	Known
0.05	296	1434	78
0.01	59	652	62
0.005	30	482	57
0.001	6	259	47

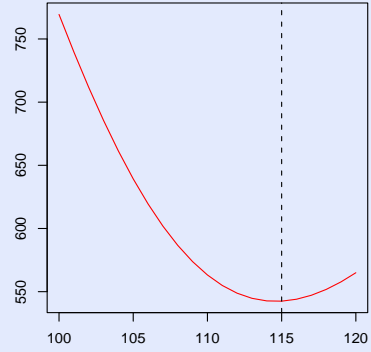
Estimation of Cell Cycle Period

- Spellman et al. found coefficient estimates to be “unstable to small variations” in T .
- Their analysis involved averaging coefficient estimates obtained for a range of values of T .
e.g. for the alpha-factor experiment they used 40 equally spaced values over the range 66 ± 11 .
- In our analysis we estimated T for each experiment by the **minimizer of the total sum of squared errors** over the 104 known genes.
(e.g. $\hat{T} = 62$ for the alpha-factor data.)

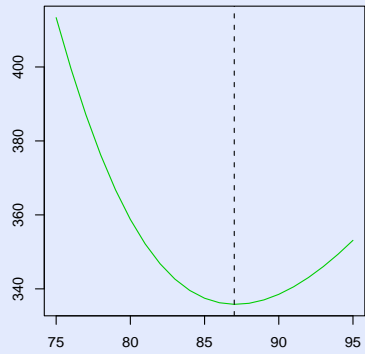
alpha-factor



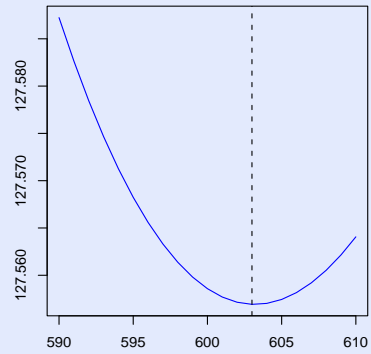
cdc15



cdc28



elutriation



Combining P-values

- P-values from the different experiments can be combined using Fisher's method (See Goutis, Wells and Casella, 1996, JASA):
- For a given gene, let P_{af} , P_{15} and P_{28} be P-values obtained from the alpha-factor, cdc15 and cdc28. Then

$$C^2 = -2(\ln P_{af} + \ln P_{15} + \ln P_{28})$$

has a chisquard(6) distribution if the gene's expression profile is constant over time.

Numbers (out of 6178) genes with P-values lower than a threshold

P-value	Expected	Actual	Known
0.05	296	2289	100
0.01	59	1348	94
0.005	30	1100	90
0.001	6	737	84
0.005	30	482	57

- Red line based on alpha-factor experiment only

Gaussian Errors?

- F-test tests the hypothesis of **no time effect**;
i.e. the mean $\log(\text{expression ratio})$ is constant over time.
- Based an assumption of **i.i.d. normal errors**!
- **Idea!**: Calibrate P-values using “randomization” distribution of F-statistic.

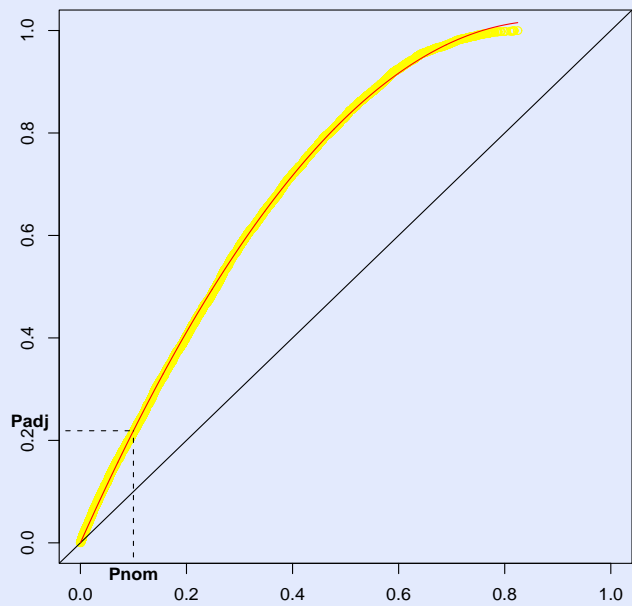
Bootstrap null distribution

- Let \hat{e}_{ij} denote the residual at time j for gene i , based on Fourier model fit.
- Construct **resampled profiles** for each gene:

$$y_{ij}^* = \bar{y}_i + \hat{e}_{ij}^*, \quad j = 1, \dots, m,$$

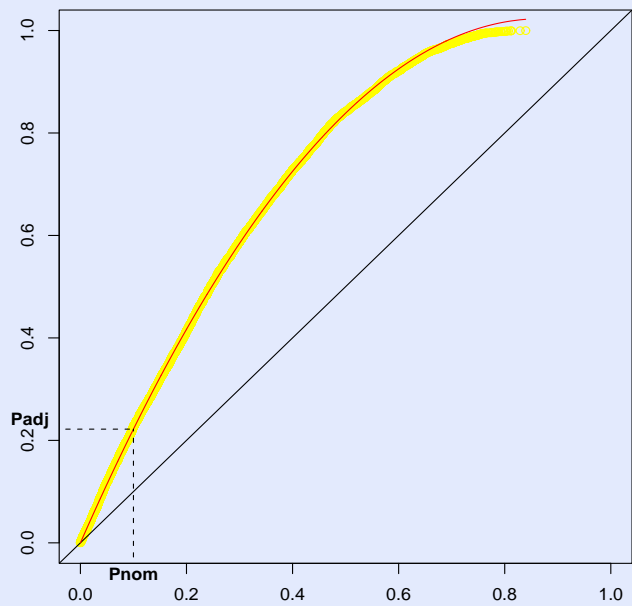
where \hat{e}_{ij}^* is drawn at random (with replacement) from $(\hat{e}_{i1}, \dots, \hat{e}_{im})$.

- Fit the Fourier model and calculate F-statistics, and P-values, for each resampled profile.
- Compare P-value distribution to uniform!



Permutation-based null distribution

- Permute the times in each gene profile in each experiment
- Fit the Fourier model and calculate F-statistics, and P-values, for each permuted profile.
- Compare P-values distribution to uniform!



P-value calibration

- See previous slide.
- Plot suggests quadratic model

$$F(p) = ap^2 + bp + c.$$

With constraints $F(0) = 0$ and $F(1) = 1$, reduces to

$$F(p) = bp(1 - p) + p^2.$$

- Almost perfect fit: $R > .9999$.
- P-value adjustment:

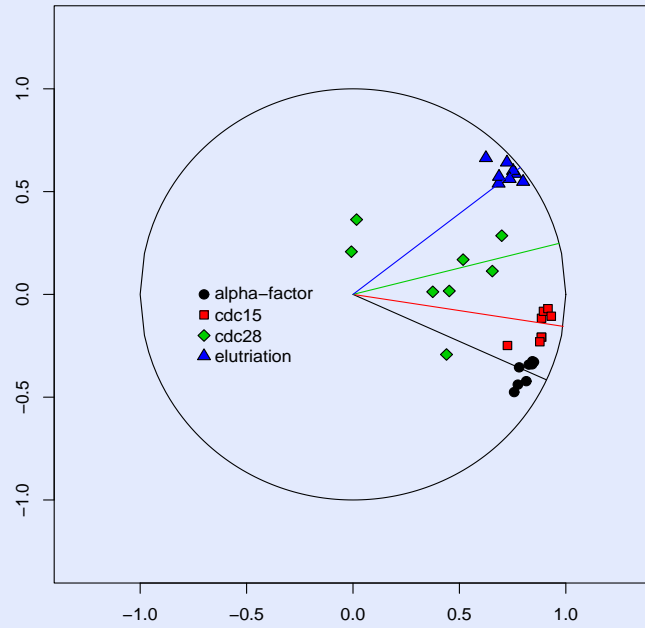
$$p_{\text{adj}} = F(p_{\text{nom}})$$

e.g. $F(.005) = .0117$.

- Corresponding bootstrap adjustment is almost identical!

Synchronization of Initial Phase

- Plot genes in a unit circle with angle determined by estimated phase of peak expression, and distance from the origin proportional to the **multiple correlation coefficient** R . (Alter et al.)
- Tight grouping of 8 known synthesis genes.
- Synchronize initial phases of Cdc15 and Cdc28 experiments by rotating plots so that **(weighted) mean phases** of synthesis genes line up with alpha-factor experiment.
- Alternatively, maximize the average length of the 104 resultant vectors formed by combining data on each gene from the alpha-factor and cdc15 (or cdc28) experiments. (Spellman et al.)
- **Combined estimate of phase of peak expression.**



Comparison with Spellman et al.

- Spellman et al. ranked the genes using an **aggregate CDC score** of the form:

$$\text{CDC} = \sqrt{B^2 + A^2},$$

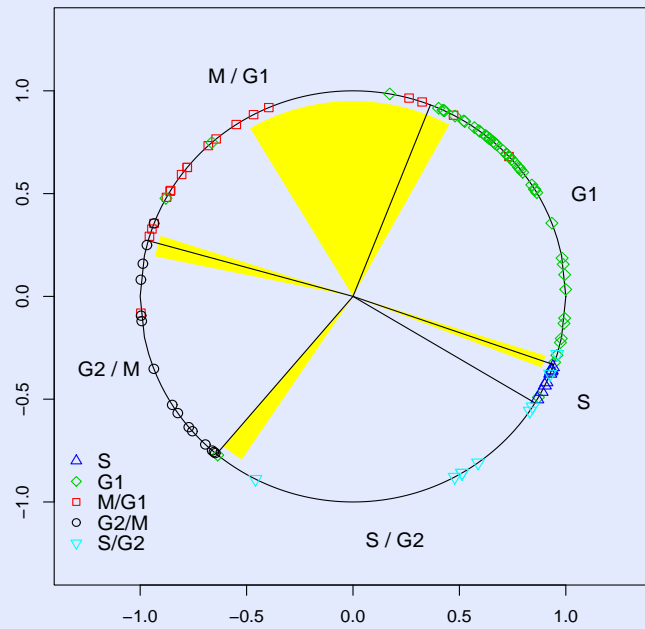
where (B, A) is a weighted average of the estimated coefficients (b_1, a_1) over the alpha-factor, cdc15 and cdc28 experiments.

- The **threshold** for declaring a gene to be cell cycle-regulated was chosen to capture 90% (91 out of 104) of the known genes. A total of **800 genes exceeded the threshold**.
- 75 recalculated CDC scores exceeded this threshold after **permuting** the times in each sample profile, suggesting a **significance level of 0.012** ($=75/6178$).

Classification

The cell cycle phase grouping is known for 104 genes. These can be used as **training sample** to produce a gene classifier using a **stochastic search algorithm**

- Find the “best” boundaries between 5 cell cycle phases: S, S/G2, G2/M, M/G1, G1. This is equivalent to placing 5 radii on the circleplot, with radii falling midway between two adjacent genes.
- Select the radii that maximize the proportion of the training sample that is correctly classified.
- Number of ways of choosing 5 radii with 104 genes is approximately 13 billion.



Stochastic Search Algorithm

1. Fix 4 radii at current values
2. Move the remaining radius (j) to new position with probabilities

$$p_i = \frac{(c_i + \lambda)/(d_i + \lambda)}{\sum_k (c_k + \lambda)/(d_k + \lambda)}$$

where c_i/d_i are the numbers of genes correctly/incorrectly classified between radii $j - 1$ and $j + 1$.

3. One iteration consists of a move for all 5 radii.
4. Repeat for M iterations (say $M = 20,000$)
5. Sort iterations according to number of genes correctly classified.

Summary

- Fourier analysis of Spellman et al. can be explained using simple and standard statistical methods.
- The combined P-value approach is a more powerful way of identifying cell cycle-regulated genes.
- Stochastic search algorithm provides data-driven method for classifying genes.
- Not clear that SVD analysis adds anything to Spellman approach. Fourier and SVD sorts are similar.
(Actual analysis of Alter et al. involved extensive data processing and manipulation.)

Singular Value Decomposition

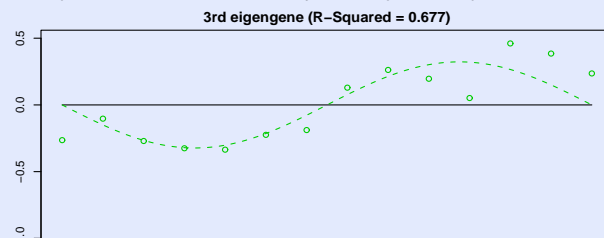
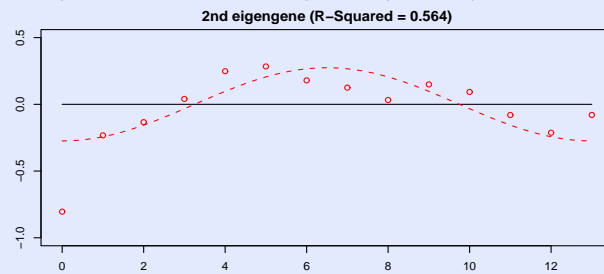
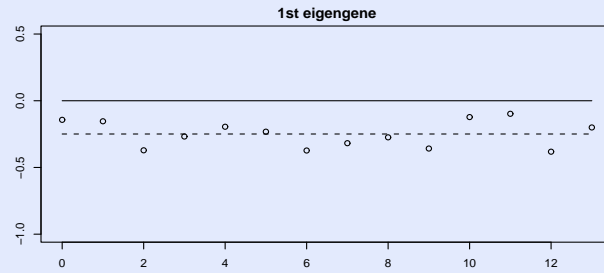
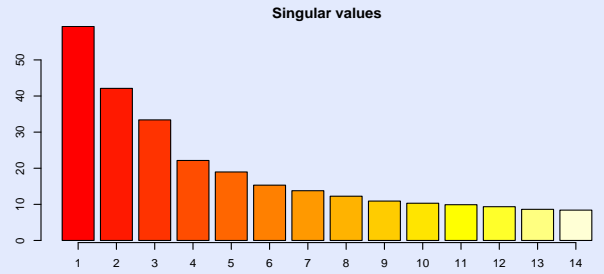
Microarray matrix, Y , can be decomposed into a product involving two matrices with orthonormal columns and a diagonal matrix; i.e.

$$Y = USV^T = \sum_{k=1}^m s_k u_k v_k^T$$

- Alter et al (2000)

$u_k = k$ th eigenvector of $Y'Y$ or k th “eigenarray”

$v_k = k$ th eigenvector of YY' or k th “eigengene”



- Approximation using three components gives

$$y_{ij} = (s_1 u_{i1}) v_{1j} + (s_2 u_{i2}) v_{2j} + (s_3 u_{i3}) v_{3j} + \hat{e}_{ij}$$

- $s_k u_{ik}$, $k = 1, 2, 3$ are precisely the least squares estimates obtained by regressing the i th gene's profile on the first three eigengenes.
- **By analogy** with Fourier model, estimate the phase of peak expression for i th gene as solution to

$$\tan(\phi_i) = \frac{s_3 u_{i3}}{s_2 u_{i2}}$$

Circular Correlation

Agreement between two sorts can be measured using **circular correlation** coefficients: Fisher (1995) “Statistical Analysis of Circular Data”

Circular correlations between Fourier, log-SVD and Alter et al. sorts of 100 known cell cycle-regulated genes using the elutriation data.

	Fourier	log-SVD	Alter
Fourier	1	0.863	0.946
log-SVD		1	0.912
Alter			1