Statistical Issues in Microarray Experiments

George Casella
University of Florida
casella@stat.ufl.edu

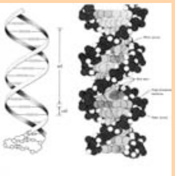
Work done with Jim Booth and Sam Wu –Statistics

John Davis and Janice Cooke – Forest Genetics

Outline

- Brief Introduction to DNA Microarrays
- Example of up- and down-regulated genes
- Cell Cycle Analysis
- Fourier vs. SVD analysis
- Classification of Genes
- Final thoughts

Deoxyribonucleic Acid (DNA)



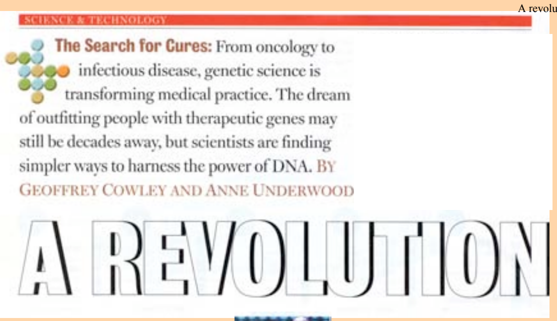
•Double helix – storing information necessary to direct the production of proteins.

•Four types of base: A, T, C, G

•Two base pairs: (A, T), (C, G)

•Complementary strands, capable of precise self-replication.

A revolution



SCIENCE & TECHNOLOGY

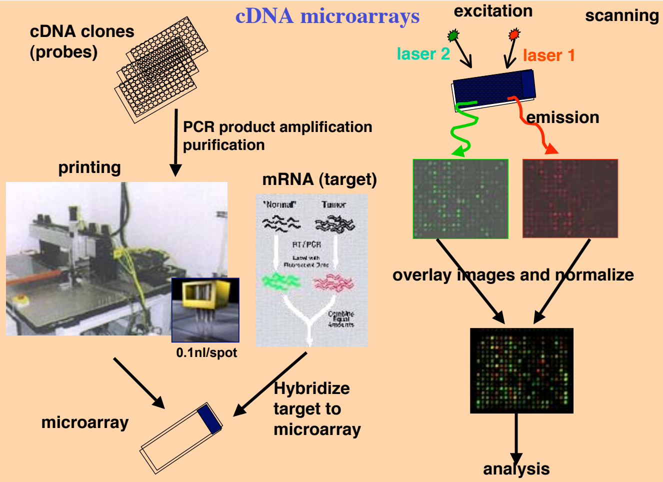
The Search for Cures: From oncology to infectious disease, genetic science is transforming medical practice. The dream of outfitting people with therapeutic genes may still be decades away, but scientists are finding simpler ways to harness the power of DNA. BY GEOFFREY COWLEY AND ANNE UNDERWOOD

A REVOLUTION

DNA microarrays also known as "DNA Chips"

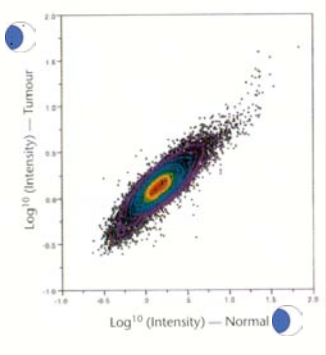
Newsweek April 10, 2000

cDNA microarrays

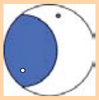


•Example of gene expression data

Comparison of Prostate Epithelial Cells and Tumor Cells from an Individual Patient



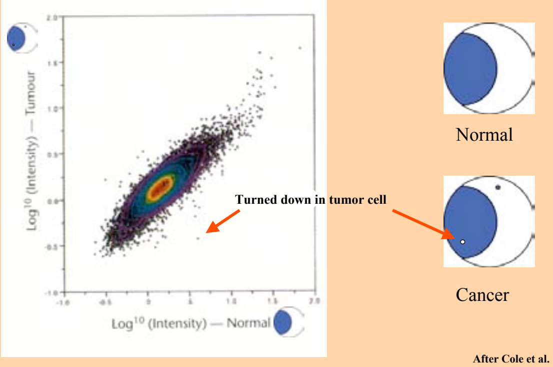
Normal



Cancer

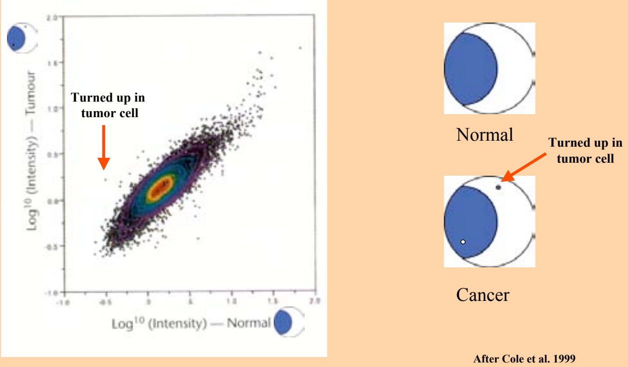
After Cole et al. 1999

Comparison of Prostate Epithelial Cells and Tumor Cells from an Individual Patient



After Cole et al. 1999

Comparison of Prostate Epithelial Cells and Tumor Cells from an Individual Patient



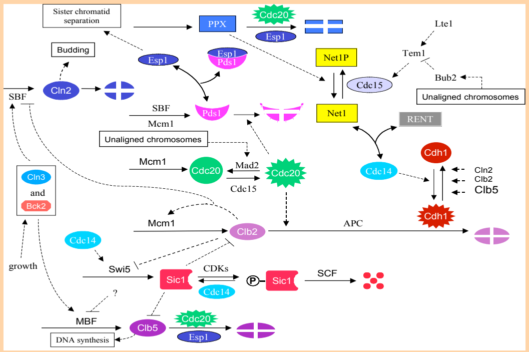
After Cole et al. 1999

- Today’s topic: Analysis of yeast genome
- Determine which genes are “cell-cycle regulated”
- Technical report available at http://web.stat.ufl.edu/~jbooth

Introduction

- 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.
  - Provide simpler, standard statistical techniques as alternatives.
  - Develop new statistical tools for the analysis of this and similar data.

Network Diagram of the Yeast Cell Cycle



From: Kohn, 1999

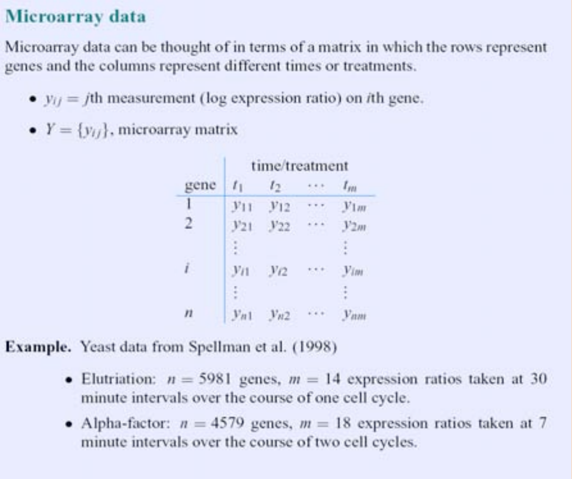
Yeast Genome Data

- Several 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: $\log(1/2) = -\log(2)$

Microarray Data



Microarray data

Microarray data can be thought of in terms of a matrix in which the rows represent genes and the columns represent different times or treatments.

- y_{ij} = jth measurement (log expression ratio) on ith gene.
- $Y = \{y_{ij}\}$, microarray matrix

$$\begin{array}{c|cccc}
 & \multicolumn{4}{c}{\text{time/treatment}} \\
\text{gene} & t_1 & t_2 & \cdots & t_m \\
\hline
1 & y_{11} & y_{12} & \cdots & y_{1m} \\
2 & y_{21} & y_{22} & \cdots & y_{2m} \\
 & \vdots & & & \vdots \\
i & y_{i1} & y_{i2} & \cdots & y_{im} \\
 & \vdots & & & \vdots \\
n & y_{n1} & y_{n2} & \cdots & y_{nm}
\end{array}$$

Example. Yeast data from Spellman et al. (1998)

- Elutriation: $n = 5981$ genes, $m = 14$ expression ratios taken at 30 minute intervals over the course of one cell cycle.
- Alpha-factor: $n = 4579$ genes, $m = 18$ expression ratios taken at 7 minute intervals over the course of two cell cycles.

Image issues

- # Image Issues

1. **Spot parameters**

   Layout, Distance between spot, Spot size

2. **Spot Information**

   Spot intensity, Background, Quality measure

3. **Where is the spot?**

From: Yang et al. 2000

Normalization Issues

- # Normalization Issues

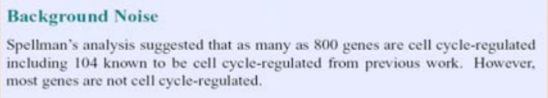
Normalization is the process of removing systematic variation in microarray experiments so that DNA expression levels can be compared across slides.

The observed intensity for each spot is determined by:

- ***Amount of target DNA on the microarray***
- ***Amount of probes available for hybridization***
- ***Experiment conditions of hybridization***
- ***Location on the array***
- ***Dye bias***

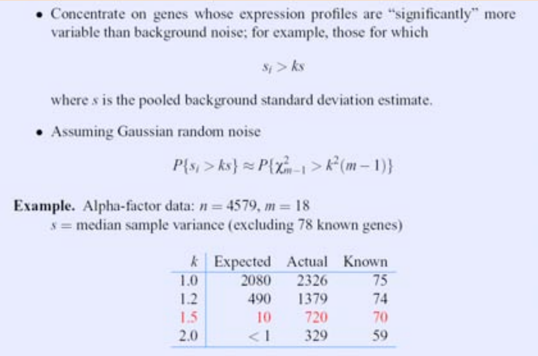
From: Yang et al. 2001

Background noise



Background Noise

Spellman's analysis suggested that as many as 800 genes are cell cycle-regulated including 104 known to be cell cycle-regulated from previous work. However, most genes are not cell cycle-regulated.



- Concentrate on genes whose expression profiles are "significantly" more variable than background noise; for example, those for which

$$s_i > ks$$

where s is the pooled background standard deviation estimate.

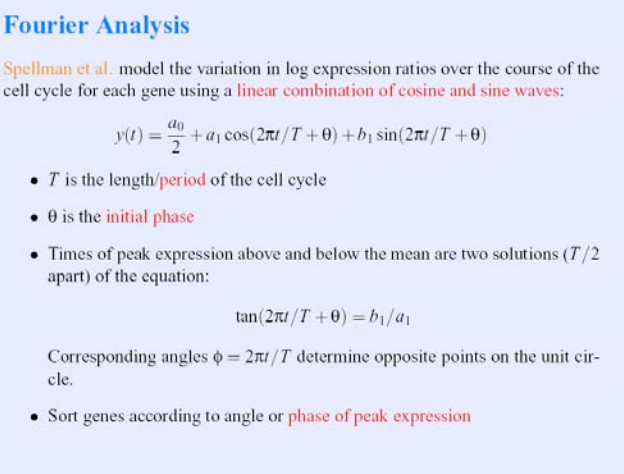
- Assuming Gaussian random noise

$$P\{s_i > ks\} \approx P\{\chi^2_{m-1} > k^2(m-1)\}$$

Example. Alpha-factor data: $n = 4579$, $m = 18$
s = median sample variance (excluding 78 known genes)

k	Expected	Actual	Known
1.0	2080	2326	75
1.2	490	1379	74
1.5	10	720	70
2.0	< 1	329	59

Fourier analysis



Fourier Analysis

Spellman et al. model 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:

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

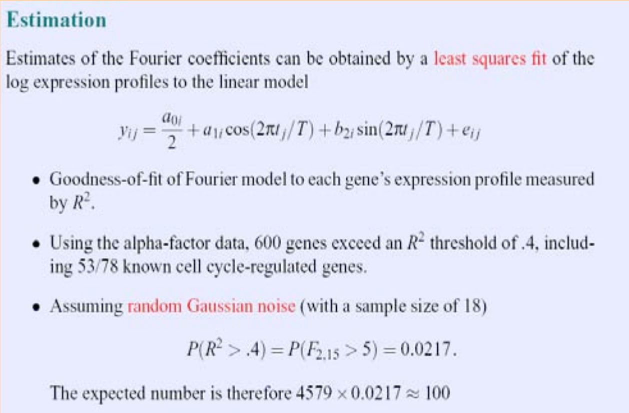
- T is the length/period of the cell cycle
- θ is the initial phase
- Times of peak expression above and below the mean are two solutions ($T/2$ apart) of the equation:

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

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

- Sort genes according to angle or phase of peak expression

Estimation



Estimation

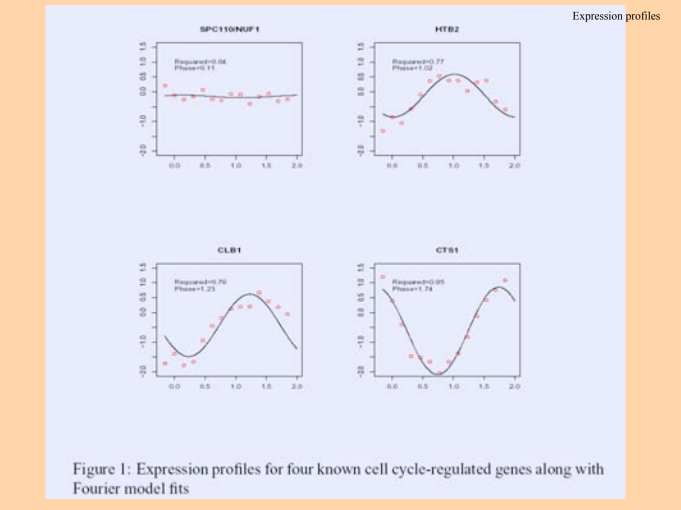
Estimates of the Fourier coefficients can be obtained by a least squares fit of the log expression profiles to the linear model

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

- Goodness-of-fit of Fourier model to each gene's expression profile measured by R^2.
- Using the alpha-factor data, 600 genes exceed an R^2 threshold of .4, including 53/78 known cell cycle-regulated genes.
- Assuming random Gaussian noise (with a sample size of 18)

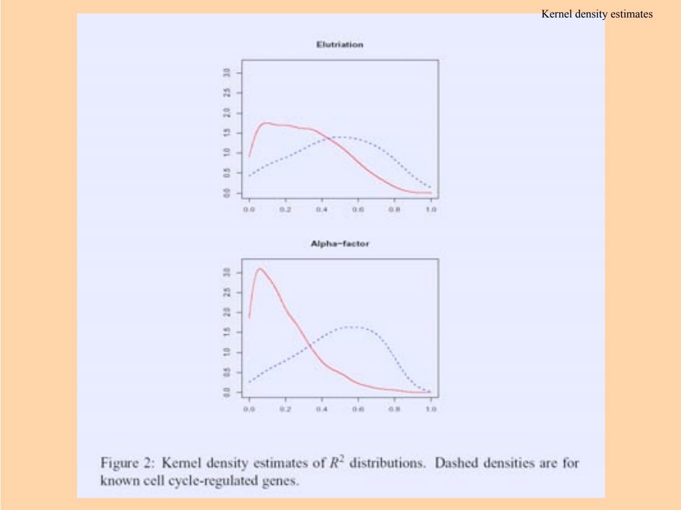
$$P(R^2 > .4) = P(F_{2,15} > 5) = 0.0217.$$

  The expected number is therefore $4579 \times 0.0217 \approx 100$



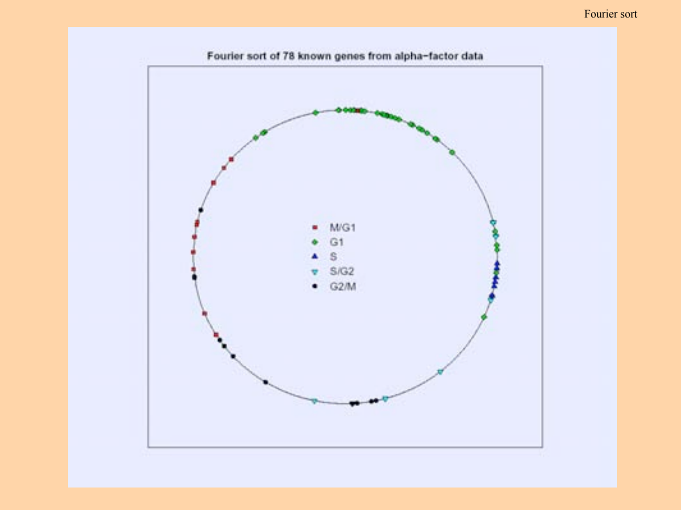
Expression profiles

Figure 1: Expression profiles for four known cell cycle-regulated genes along with Fourier model fits

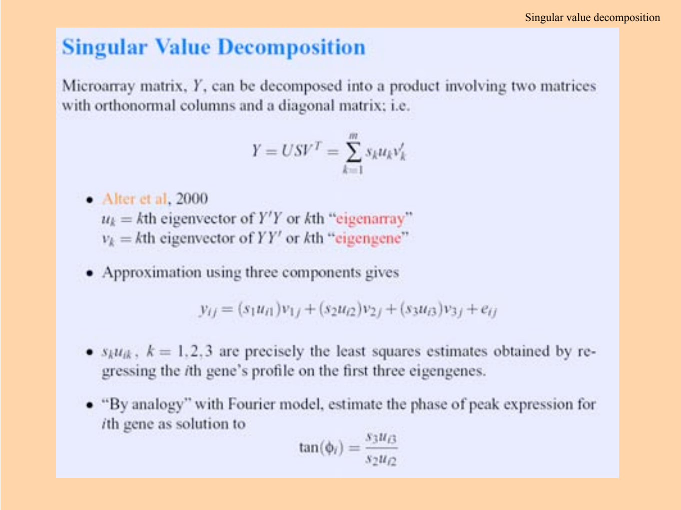


Kernel density estimates

Figure 2: Kernel density estimates of R^2 distributions. Dashed densities are for known cell cycle-regulated genes.



Fourier sort



Singular value decomposition

Singular Value Decomposition

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

$$Y = USV^T = \sum_{k=1}^{m} s_k u_k v_k'$$

- Alter et al, 2000
  $u_k = k$th eigenvector of $Y'Y$ or kth "eigenarray"
  $v_k = k$th eigenvector of YY' or kth "eigengene"

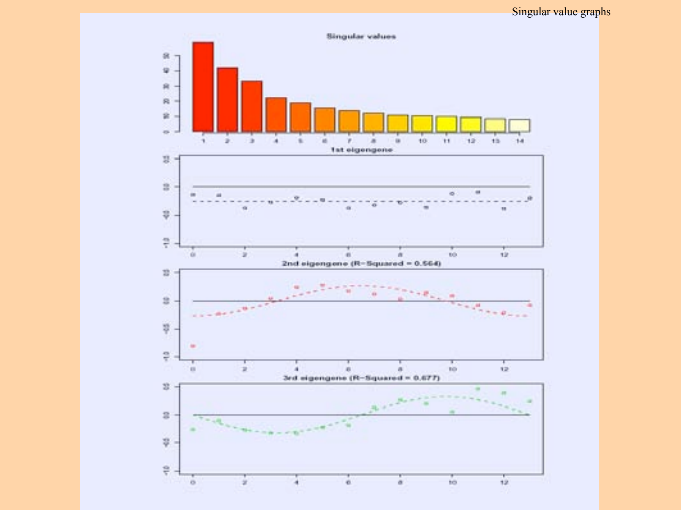
- Approximation using three components gives

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

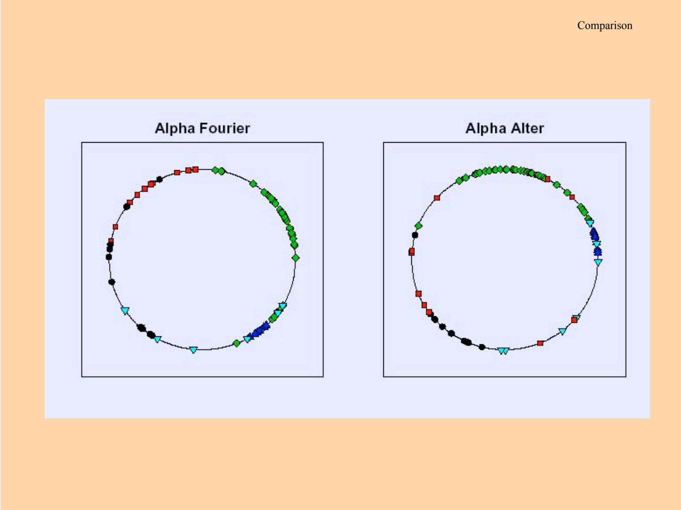
- $s_k u_{ik}$, $k = 1,2,3$ are precisely the least squares estimates obtained by regressing the ith gene's profile on the first three eigengenes.

- "By analogy" with Fourier model, estimate the phase of peak expression for ith gene as solution to

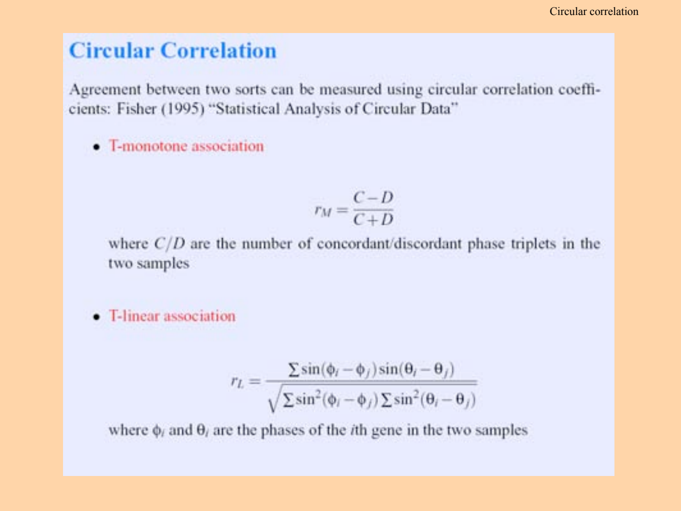
$$\tan(\phi_i) = \frac{s_3 u_{i3}}{s_2 u_{i2}}$$



Singular value graphs



Comparison



Circular correlation

Circular Correlation

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

- T-monotone association

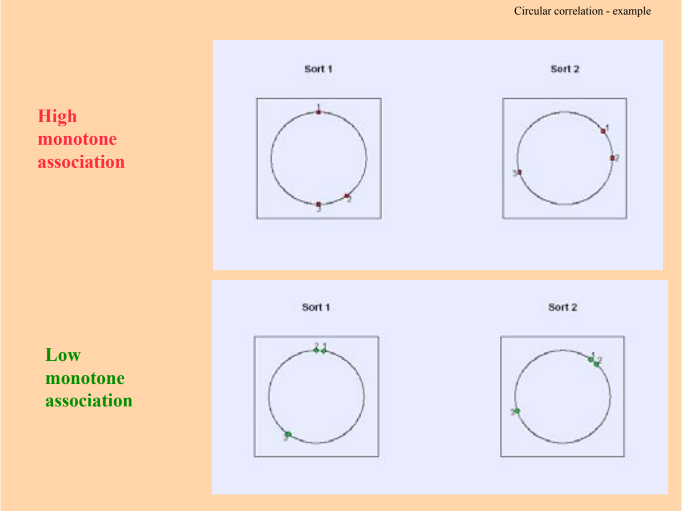
$$r_M = \frac{C-D}{C+D}$$

where C/D are the number of concordant/discordant phase triplets in the two samples

- T-linear association

$$r_L = \frac{\sum \sin(\phi_i - \phi_j)\sin(\theta_i - \theta_j)}{\sqrt{\sum \sin^2(\phi_i - \phi_j)\sum \sin^2(\theta_i - \theta_j)}}$$

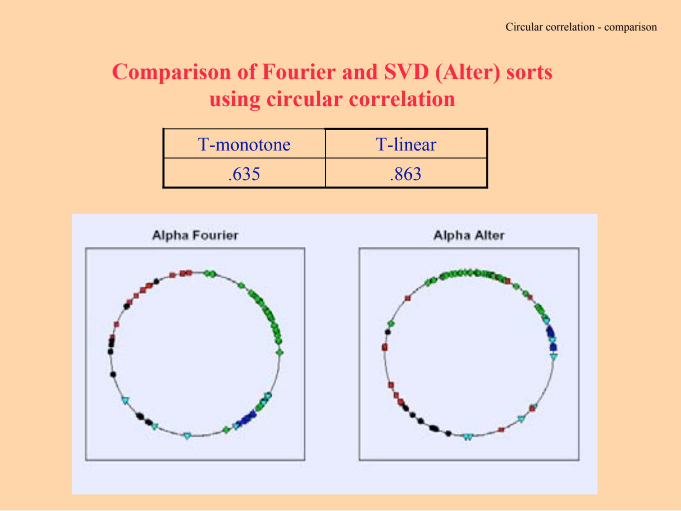
where ϕ_i and θ_i are the phases of the ith gene in the two samples



Circular correlation - example

High monotone association

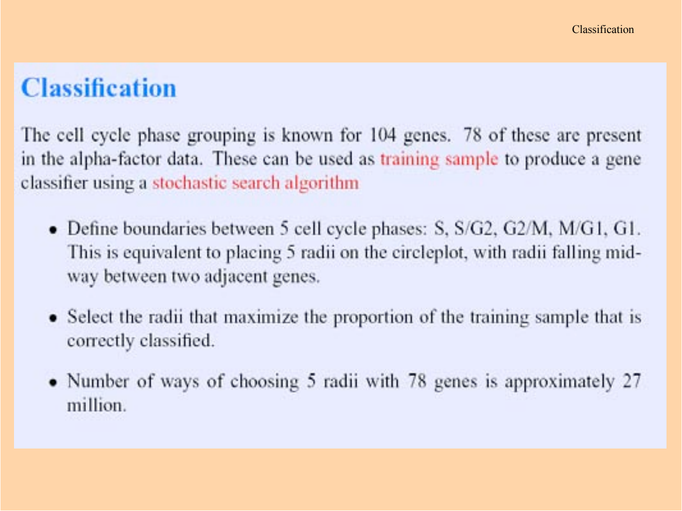
Low monotone association



Circular correlation - comparison

Comparison of Fourier and SVD (Alter) sorts using circular correlation

T-monotone	T-linear
.635	.863

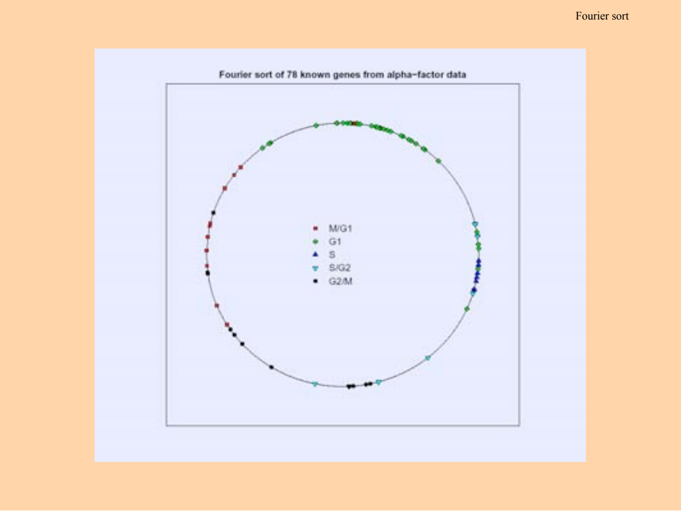


Classification

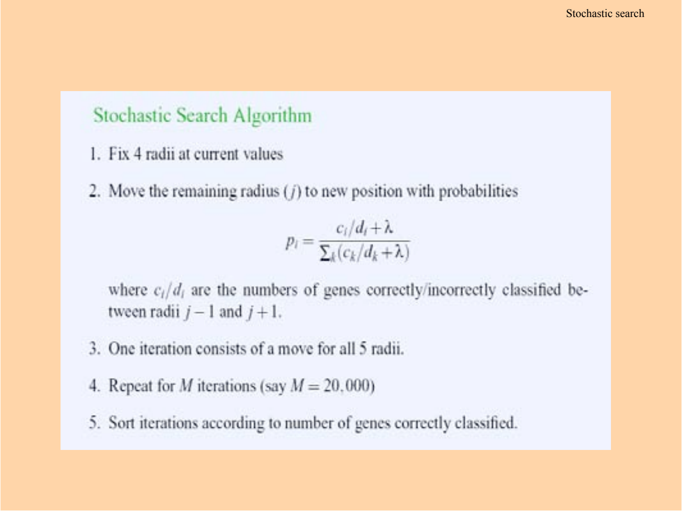
Classification

The cell cycle phase grouping is known for 104 genes. 78 of these are present in the alpha-factor data. These can be used as training sample to produce a gene classifier using a stochastic search algorithm

- Define 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 78 genes is approximately 27 million.



Fourier sort



Stochastic search

Stochastic Search Algorithm

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

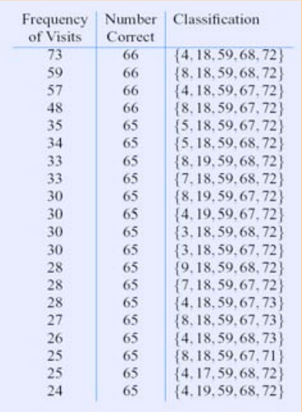
$$p_i = \frac{c_i/d_i + \lambda}{\sum_k (c_k/d_k + \lambda)}$$

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.

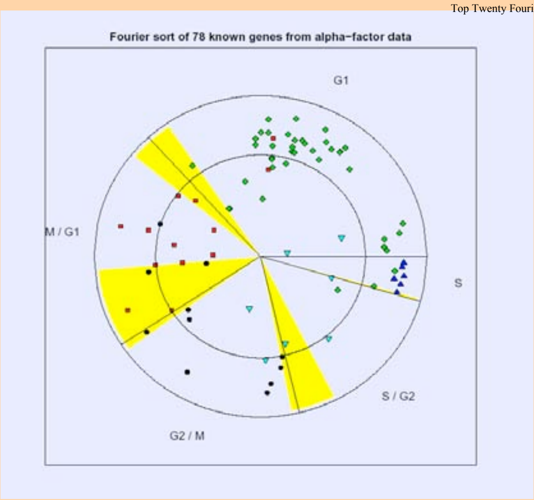
Top Twenty Classifications

Top Twenty Classifications



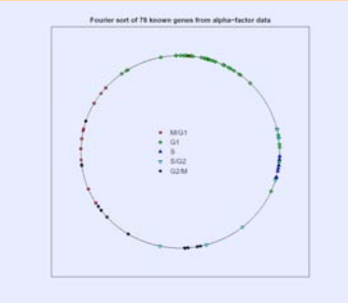
Frequency of Visits	Number Correct	Classification
73	66	$\{4,18,59,68,72\}$
59	66	$\{8,18,59,68,72\}$
57	66	$\{4,18,59,67,72\}$
48	66	$\{8,18,59,67,72\}$
35	65	$\{5,18,59,67,72\}$
34	65	$\{5,18,59,68,72\}$
33	65	$\{8,19,59,68,72\}$
33	65	$\{7,18,59,68,72\}$
30	65	$\{8,19,59,67,72\}$
30	65	$\{4,19,59,67,72\}$
30	65	$\{3,18,59,68,72\}$
30	65	$\{3,18,59,67,72\}$
28	65	$\{9,18,59,68,72\}$
28	65	$\{7,18,59,67,72\}$
28	65	$\{4,18,59,67,73\}$
27	65	$\{8,18,59,67,73\}$
26	65	$\{4,18,59,68,73\}$
25	65	$\{8,18,59,67,71\}$
25	65	$\{4,17,59,68,72\}$
24	65	$\{4,19,59,68,72\}$

Top Twenty Fourier sort

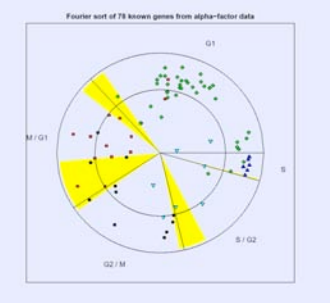


Both Sorts

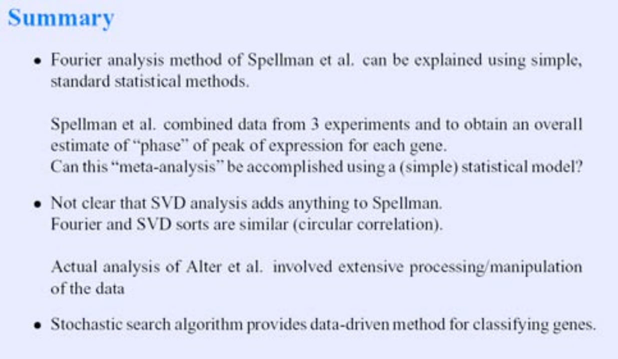
Radius = 1



Radius = r^2



Summary



Summary

- Fourier analysis method of Spellman et al. can be explained using simple, standard statistical methods.

  Spellman et al. combined data from 3 experiments and to obtain an overall estimate of “phase” of peak of expression for each gene.
  Can this “meta-analysis” be accomplished using a (simple) statistical model?
- Not clear that SVD analysis adds anything to Spellman.
  Fourier and SVD sorts are similar (circular correlation).

  Actual analysis of Alter et al. involved extensive processing/manipulation of the data
- Stochastic search algorithm provides data-driven method for classifying genes.

- Simultaneous Inferences

Last Thoughts: Simultaneous Inferences

To identify differentially expressed genes, we have to control error rates in thousands of simultaneous hypotheses tests.

1. **Multiple comparison techniques were explored in Dudoit et al. (2000)**
2. **False discovery rates approaches were studied in Tusher et al. (2001)**
3. **Empirical Bayes analysis was developed in Efron et al (2001)**