OR 779  Functional Data Analysis
Course Project

# Analyzing Microarray Time–course Genome–wide Data

Presented by

**Xin Zhao**

April 29, 2002
Cornell University

# Overview

1. Introduction

    - Biological Background
    - Biological Questions
    - Data Description
    - Our Approach

2. Data Analysis
    - Functional Data Analysis Approach

3. Conclusions

4. Possible Future Ideas

5. Acknowledgements

6. References

# Biological Background

Refer to Jing Qiu's talk.

## Cell Cycle

- Includes G1, S, G2, M, M/G1 five phases in literature.

## Cell–cycle–specific Gene

- Gene expresses periodically over cell cycle.
- Intuitively, call it "periodic gene".

# Biological Background (cont.)

## Yeast Genome

- Include over 6,000 genes
- By 1998, 104 “known” periodic genes
- Spellman (1998)
  - Identified 800 periodic genes
  - 94 “known” genes were included
  - Considered as “standard”
- Main data source for studying periodic genes

# Biological Background (cont.)

Gene expression level

- Measured by cDNA-microarray experiment.
- Expression ratio between test gene and reference gene.

Figure: yeast genome-wide gene expressions over 2 cell cycles

- Data source:
  genome–www.Stanford.edu/cellcycle.
- Referenced by over 293 published papers so far.

# Biological Background (cont.)

Figure: yeast genome-wide gene expressions over 2 cell cycles

- Each curve represents the time series for each gene over 18 sampling points
- Sampling at 7-min intervals for 120 minutes.
- The time interval covers approximately 2 cell cycles
- 4,489 time series in the population

x–axis: sampling points over time.

y–axis: $\log_2$(gene expression level)

# Biological Background (cont.)

## Periodic gene classification

- Peak expression at a specific phase during a cell cycle.
- Have five groups (G1, S, G2, M, M/G1)

  e.g., G1 group peak expression at G1 phase

Figure: yeast cell genome-wide periodic gene classification (200 genes)

# Biological Questions

- Identification of  Periodic Genes

Figure: yeast genome-wide periodic genes identification

- Classification of Periodic Genes

Figure:
Yeast cell genome-wide periodic gene classification (200 genes)

# Data Description

- Yeast cells synchronized by $\alpha$-factor arrest

- Raw data

$$\begin{pmatrix} x_{1,1} \\ \vdots \\ x_{d,1} \end{pmatrix}, \cdots, \begin{pmatrix} x_{1,n} \\ \vdots \\ x_{d,n} \end{pmatrix} \qquad \text{Where} \quad d = 18, \qquad n = 4{,}489$$

$x_{i,j}$ : $\log_2$ (expression level) for j$^{th}$ gene at i$^{th}$ sampling point.

# Our Approach

- Project Goals:

  Goal 1: Understand "population structure".

  Goal 2: Explore identification and classification for periodic genes.

- This is an exploratory analysis

- Fit in the framework of "Functional Data Analysis".

  Object Space $\leftrightarrow$ Feature Space
  (curves $\leftrightarrow$ data vectors)

# Our Approach (cont.)

Object Space → Feature Space (i.e., curves → data vectors)

- Approach to Goal 1:

  - PCA for data

  - PCA for projections onto an appropriate Fourier subspace

- Approach to Goal 2:

  Project data onto a 2-dim Fourier subspace

  - Identify periodic genes

  - Classify periodic genes

# Our Approach (cont.)

Feature Space → Object Space (i.e., data vectors → curves)

- Visualization of “population structures”

Figure: yeast cell genome-wide periodic gene classification (200 genes)

# Functional Data Analysis

Object Space View:

- Overlay plots of curves

Recall:

Figure: yeast genome-wide gene expressions over 2 cell cycles

# Functional Data Analysis (Cont.)

Feature Space Data Representation

- Data vectors

Recall:

$$\begin{pmatrix} x_{1,1} \\ \vdots \\ x_{d,1} \end{pmatrix}, \cdots, \begin{pmatrix} x_{1,n} \\ \vdots \\ x_{d,n} \end{pmatrix}$$ Where $d = 18,$ $n = 4{,}489$

$x_{i,j}$ : $\log_2$ (expression level) for j^th gene at i^th sampling point.

# Functional Data Analysis (Cont.)

## Feature Space Data Representation

Center data vector over time → centered data

$$\begin{pmatrix} x_{1,1} - \bar{x}_1 \\ \vdots \\ x_{d,1} - \bar{x}_1 \end{pmatrix}, \cdots, \begin{pmatrix} x_{1,n} - \bar{x}_n \\ \vdots \\ x_{d,n} - \bar{x}_n \end{pmatrix} \qquad \bar{x}_j = \frac{1}{d} \sum_{i=1}^{d} x_{i,j}$$

data matrix = $\begin{pmatrix} x_{1,1} - \bar{x}_1 & \cdots & x_{1,n} - \bar{x}_n \\ \vdots & \ddots & \vdots \\ x_{d,1} - \bar{x}_1 & \cdots & x_{d,n} - \bar{x}_n \end{pmatrix}$ 18×4,489

# Understand Population Structure

1. PCA for data

Curve View Graphic: a nice approach to view PCA

Figure: spellman_alpha_complete_pca.eps

- No dominant eigenvalue, as shown clearly in the power plot.
- 1st PC only explains about 25% of total energy.
- 2nd PC only explains about 16% of total energy.
- Periodic direction is not the PC directions, but might be a rotation of the PC directions.

# Understand Population Structure (Cont.)

2. PCA for projections onto a Fourier subspace

Fourier basis B = $\{\sin(i\omega t), \cos(i\omega t),\ i = 2, 4, 6, 8,\ t = 1, \ldots, 18\}_{18\times 8}$

$$\text{Where}\quad \omega = \frac{2\pi}{T},\quad T = 18$$

Reasons:

- The time interval covers two cell cycles.
- The period of periodic genes is consistent with that of a cell cycle.
- 18 (equally spaced) sampling points available.

# Understand Population Structure (Cont.)

2. PCA for projections onto a Fourier subspace

Projection matrix = $B(B^TB)^{-1}B^T$(data matrix)
18 × 4,489                                   18 × 4,489

Perform PCA on the projected data

Figure: spellman_alpha_complete_proj_pca.eps

# Understand Population Structure (Cont.)

2. PCA for projections onto a Fourier subspace

Figure: spellman_alpha_complete_proj_pca.eps

- The first two PCs are dominant (explain about 65% of total energy)
- 1st PC explains 37.31% of total energy.
- 1st PC direction similar to a sine wave over two periods.
- 2nd PC explains 27.54% of total energy.
- 2nd PC direction similar to a cosine wave over two periods.

# Understand Population Structure (Cont.)

2. PCA for projections onto a Fourier subspace

- Projections onto the 2-dim Fourier subspace spanned by $\{\sin(2\omega t), \cos(2\omega t)\}$ captures the main feature of the periodicity in the data.

- Time shift issue is captured in the subspace:

  Reason: $\sin(2\omega t + \phi) = \cos(\phi)\sin(2\omega t) + \sin(\phi)\cos(2\omega t)$

  $= a_1\sin(2\omega t) + a_2\cos(2\omega t)$,

  where $a_1$ and $a_2$ are constants

  similar to $\cos(2\omega t + \phi)$

# Identification and classification for periodic genes

1. Identification of periodic genes

Recall: Subspace spanned by {sin(2$\omega$t), cos(2$\omega$t)} captures periodicity.

Idea: Periodic genes have large distance to the origin in the subspace.

Q: Which distance is considered as large?

# Identification and classification for periodic genes (cont.)

1. Identification of periodic genes

Q: Which distance is considered as large?

We consider a range of thresholding criteria:

- Rank all genes in decreasing order according to the distance to the origin in the subspace.

- Choose a range of thresholding values:

  First 200, 400, 600, 800, and 1,000 genes

Figure: Periodic gene identification scatterplots

# Identification and classification for periodic genes (cont.)

Figure: Periodic gene identification scatterplots

- G1 group is in red
- S group is in green
- G2 group is in blue
- M group is in yellow
- M/G1 group is in cyan
- Non-periodic genes are in black
- Purple lines are boundaries (explained later)

# Identification and classification for periodic genes (Cont.)

2. Classification of periodic genes

Idea: set the boundaries for the angles in the 2-dim subspace.

Q: Do we know the timing for G1, S, G2, M, M/G1 phases?

A: No.

Solution:

- Initial guess from previous result: Spellman’s classification.
- Modification using Sizer plot of angles and scatterplot in the subspace.

# Identification and classification for periodic genes (Cont.)

2. Classification of periodic genes

I. Results by Spellman (1998):

Figure: Spellman’s classification

II. Modification by SiZer plot of first 200 gene angles:

Figure: SiZer plot of first 200 gene angles

III. Modification by scatterplot in the subspace

Figure: periodic genes scatterplot by sizer (200, 400 genes)

# Identification and classification for periodic genes (Cont.)

## 2. Classification of periodic genes

Our set of angle boundaries for the five periodic gene groups:

| | |
|---|---|
| M/G1 phase: | [0.83, 2.04] |
| G1 phase: | [2.04, 3.74] |
| S phase: | [3.74, 4.58] |
| G2 phase: | [4.58, 5.72] |
| M phase: | [5.72, 0.83] |

Note:

- This selection came from previous “standard” results and eyeball examination of the Sizer plot and scatterplot.
- It may not be statistically robust.

# Identification and classification for periodic genes (Cont.)

2. Classification of periodic genes

Figure: compare classification results for 2 thresholds (200, 800)

- Sizer plot for first 200 genes shows 2 significant bumps, G1 and S groups.

- The Sizer plot for the first 800 genes shows 2 significant bumps: G1 and G2 groups.

- Suggests that S group has more highly periodic genes, and G2 group has more low periodic genes.

- Might have biological interpretation.

# Identification and classification for periodic genes (cont.)

## 2. Classification of periodic genes

Kernel density estimator of periodic genes for different thresholds:

Figure: Kde plot of periodic genes for 2 thresholds (200, 800)

- Kde plot of first 200 genes: four bumps for G1, S, G2, and M/G1 groups.
- Kde plot for first 800 genes: G2 group became most significant.
- G1 group is significant for each threshold.

# Identification and classification for periodic genes (cont.)

2. Classification of periodic genes

Figure: Compare percentage of genes in each group for different thresholds

- G1 group is the largest group for each threshold.
- S and M groups are the small groups for each threshold.
- As the threshold increases, percentage of periodic genes
  - *decreases* in G1 group
  - *increases* in G2 group
  - relatively stable in S, M, and M/G1 groups

There might exist some meaningful biological interpretation.

# Visualization of "population structures"

In object space, compare five groups of periodic genes for different thresholds:

- Plot periodic gene curves for each threshold in each group.

Figure:

Plot of periodic gene curves for each threshold in G1 group

# Conclusions

- Fourier subspace spanned by {sin(2ωt), cos(2ωt)} captures periodicity.
- G1 and S groups has more highly periodic genes
- G2 group has more low periodic genes
- Periodicity in M, and M/G1 groups is relatively uniform-distributed.

# Possible Future Ideas (cont.)

I. Apply our approach to different microarray experiments

Current microarray experiments:

| | Sampling Interval (minutes) | Total Experiment Time (minutes) | Cell Cycle Time (minutes) | Number of Sampling Points |
|---|---|---|---|---|
| Alpha-factor [a] | 7 | 120 | 66 | 18 |
| CDC 15 [a] | 10 | 290 | 110 | 24 |
| CDC28 [b] | 10 | 160 | 85 | 17 |
| Elutriation [a] | 30 | 390 | 390 | 14 |

[a]: data is from Spellman, et al

[b]: data is from Cho, et al

# Possible Future Ideas (cont.)

II. Patterns should be experimentally reproducible and statistically significant.

Q: How reproducible are the patterns in current microarray experiments?

“Genes that are periodic under one synchronization procedure are not necessarily periodic under a different synchronization procedure.”

- Shedden, Kerby and Cooper, Stephen (2002)

# Acknowledgements

- My deep appreciation to Prof. Steve Marron, under whose guidance this project was conducted.

- Thank Prof. Martin Wells for his initiation on the project.

# References

Some publications in this area:

## Clustering Method:

Eisen, M.B, Spellman, P.T., Brown, P.O., and Botstein (1998), “Cluster Analysis and display of genome-wide expression patterns”, *Proc. Natl. Acad. Sci. USA*, 95, 14863-14868.

Whitfield, M.L., Sherlock, G, Saldanha, A., Murray, JI, Ball, C.A., Alexande K.E., Matese J.C., Perou, C.M., Hurt M.M., Brown, P.O., and Botstein, D. (2002), “Identification of genes periodically expressed in the human cell cycle and their expression in tumors”, *Mol. Biol. Cell*, 10, ???-???

## Corresponse Analysis:

Fellenberg, K., Hauser, N.C., Brors, B., Neutzner, A, Hoheisel, J.D., and Vingron, M. (2001), “Correspondence analysis applied to microarray data”, *Proc. Natl. Acad. Sci. USA*, 98(19), 10781-10786.

## Singular Value Decomposition Method:

Holter, N., Mitra, M., Maritan, A., Cieplak, M., Banavar, JR, and Fedoroff, NV. (2000), “Fundamental patters underlying gene expression profiles: simplicity from complexity”, *Proc. Natl. Acad. Sci. USA*, 97, 8409-8414.

Alter, O., Brown, P.O., and Botstein, D. (2000) “Singular value decomposition for genome-wide expression data processing and modeling”, *Proc. Natl. Acad. Sci. USA*, 97(18), 10101-10106

## Others:

Spellman, P. T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botsein, D. and Futcher, B. (1998), “Comprehensive Identification of Cell Cycle–regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization”, *Molecular Biology of the Cell*, 9, 3273–3297.

Cho R.J., et al. (1998), “A Genome-wide Transcriptional Analysis of the Mitotic Cell Cycle”, *Molecular Cell*, 2, 65-73

Cooper, S. “Cell Cycle Analysis and Microarrays”, *Trends in Genetics*, accepted for publication.

Shedden, K. and Cooper, S. (2002), “Analysis of cell-cycle-specific gene expression in human cells as determined by microarrays and double-thymidine block synchronization”, PNAS, 99(7), 4379-4384