Microarray Data Analysis –
Bayesian Techniques, Clustering, Imputation

Ming Ouyang*
William J. Welsh† and Panos G. Georgopoulos*
*Environmental and Occupational Health Sciences Institute,
Computational Chemodynamics Laboratory (CCL)
†Informatics Institute
Rutgers University & UMDNJ – R. W. Johnson Medical School,
Piscataway, NJ
Overview

• Significance of differential expression
• Cluster analysis
• Imputation: missing value estimation
Microarray Experiment

Spot (DNA probe):
- known cDNA or
- Oligo

RT and label with fluor dyes

Mix and hybridize target to microarray

Treated
Control
Significance of Differential Expression

• The data: Freeman et al. (2001) Neuroscience 108(3):371-80
  – rat hippocampus
  – cocaine vs. saline injections
  – two replicates (four arrays)

• The question:
  – Which genes’ expression are significantly changed?
Significance of Differential Expression 2/3

• The fold approach: significant if
  – $\text{mean}(\log_2(\text{treated/control})) > 1$
  – $\text{mean}(\log_2(\text{treated/control})) < -1$
  – doubtful when there are only two replicates

• The Bayesian approach:
  – Incorporate background knowledge
  – Take the variance into account
Significance of Differential Expression 3/3

• The Bayesian technique:
• Data: \( D = \{ x_1, x_2 \} \)
• Prior: \( \Pr(\mu, \sigma^2) \)
• Likelihood: \( \Pr(D|\mu, \sigma^2) = \prod N(x_i; \mu, \sigma^2) \)
• Posterior: \( \Pr(\mu, \sigma^2|D) = \frac{[\Pr(D|\mu, \sigma^2) \cdot \Pr(\mu, \sigma^2)]}{\Pr(D)} \)

(Thomas Bayes, 1764)
Collections of Experiments

- Time course after a treatment
- Different treatments
- Disease cell lines
- Data are represented in a matrix
Cluster Analysis

• Grouping of genes with “similar” expression profiles
• Grouping of disease cell lines/toxicants with “similar” effects on gene expression
• Clustering algorithms
  – Hierarchical clustering
  – Self-organizing maps
  – K-means clustering
Measuring Similarity of Gene Expression

- **Euclidean (L_2) distance**
- **Manhattan (L_1) distance**
- **L_m**: \( (|x_1-x_2|^m + |y_1-y_2|^m)^{1/m} \)
- **L_\infty**: \( \max(|x_1-x_2|,|y_1-y_2|) \)
- **Inner product**: \( x_1x_2 + y_1y_2 \)
- **Correlation coefficient**
- **Spearman rank correlation coefficient**
Linkage in Hierarchical Clustering

- **Single linkage:**
  \[ S(A,B) = \min_a \min_b d(a,b) \]

- **Average linkage:**
  \[ A(A,B) = \frac{\sum_a \sum_b d(a,b)}{|A| \cdot |B|} \]

- **Complete linkage:**
  \[ C(A,B) = \max_a \max_b d(a,b) \]

- **Centroid linkage:**
  \[ M(A,B) = d(\text{mean}(A), \text{mean}(B)) \]

- **Hausdorff linkage:**
  \[ h(A,B) = \max_a \min_b d(a,b) \]
  \[ H(A,B) = \max(h(A,B), h(B,A)) \]

- **Ward linkage:**
  \[ W(A,B) = \frac{|A| \cdot |B| \cdot (M(A,B))^2}{|A|+|B|} \]
Hierarchical Clustering

3 clusters? 2 clusters?
K-means Clustering

The intended clusters are found.
K-means Clustering

Relocate a point

The intended clusters are not found.
Robust K-means Clustering via Sub-sampling
Bradley and Fayyad (1998)
Proc. 15th International Conf on Machine Learning
Computation Time and Memory Requirement
n genes and m experiments

Computation time:
• Hierarchical clustering
  – $O( m n^2 \log(n) )$
• K-means clustering
  – $t$: number of iterations
  – $O( k t m n )$

Memory requirement:
• Hierarchical clustering
  – $O( mn + n^2 )$
• K-means clustering
  – $t$: number of iterations
  – $O( mn + kn )$

- Yeast subject to environmental changes
- 6198 genes
- 5 to 10 time points
- 156 experiments

- Heat shock
- $H_2O_2$
- Menadione
- Dithiothrietol
- Diamide
- Hyper-osmotic
- Hypo-osmotic
- Amino acid starvation
- Nitrogen depletion
- Different carbon sources
- Stationary phase
- Steady-state growth
Yeast Environmental Stress Response, Gasch et al. (1998)

hierarchical clustering, correlation coefficient, average linkage
Yeast Environmental Stress Response, Gasch et al. (1998)
two clusters: ~600 repressed and ~300 induced genes
Correlations among Time Points in Amino Acid Starvation

<table>
<thead>
<tr>
<th></th>
<th>0.5 hr</th>
<th>1 hr</th>
<th>2 hr</th>
<th>4 hr</th>
<th>6 hr</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5 hr</td>
<td>1.000</td>
<td>0.647</td>
<td>0.353</td>
<td>0.342</td>
<td>0.413</td>
</tr>
<tr>
<td>1 hr</td>
<td>0.647</td>
<td>1.000</td>
<td>0.575</td>
<td>0.408</td>
<td>0.445</td>
</tr>
<tr>
<td>2 hr</td>
<td>0.353</td>
<td>0.575</td>
<td>1.000</td>
<td>0.497</td>
<td>0.435</td>
</tr>
<tr>
<td>4 hr</td>
<td>0.342</td>
<td>0.408</td>
<td>0.497</td>
<td>1.000</td>
<td>0.694</td>
</tr>
<tr>
<td>6 hr</td>
<td>0.413</td>
<td>0.445</td>
<td>0.435</td>
<td>0.694</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Transient versus stationary responses
Issues in Cluster Analysis

• A lot of clustering algorithms
• A lot of distance/similarity metrics
• Which clustering algorithm runs faster and uses less memory?
• How many clusters after all?
• Are the clusters stable?
• Are the clusters meaningful?
Determine the Number of Clusters

- **Dunn’s validity index:**
  - \((\text{min inter-cluster distance}) / (\text{max intra-cluster distance})\)

- **The gap statistic:**
  - \(W_k\): sum of all intra-cluster distances
  - \(\text{Gap}(k) = E(\log(W_k)) - \log(W_k)\)

- **The silhouette statistic:** for a point \(x_i\)
  - \(a(i)\): average distance to points in its own cluster
  - \(b(i)\): average distance to points in the nearest cluster
  - The silhouette statistic: \((b(i) - a(i)) / \max(a(i), b(i))\)


- **Traditional statistics:** Calinski and Harabsz (1974), Hartigan (1975), Krzanowski and Lai (1985)
Our Approach

• Employ the statistic B:
  – Non-parametric correlations between clusters

• Analysis method:
  1. Compute two random K-means
  2. Compute the statistic B
  3. Repeat Steps 1, 2 for 100 times
  4. Repeat Steps 1, 2, 3 for $K = 2, 3, \ldots, 10$
Cluster Stability of Yeast Environmental Stress Data

Correlation coefficients
Cluster Stability of Yeast Cell Cycle Data

Correlation coefficients
The Severity of Missing Values

- Yeast cell cycle: Eisen et al. (1998) PNAS 95:14863-14868
  - 6221 genes, 80 experiments
- 3222 genes with all 80 values
- 1583 genes with 1 missing value
- 478 genes with 2 missing values
- 178 genes with 3 missing values
- What to do:
  - Discard all genes with missing values, or
  - Estimate the missing values
Missing Value Estimations

- Zeros
- Column mean
- Row mean
- Singular Value Decomposition (SVD) based imputation
- K Nearest Neighbor (KNN) imputation
- Multivariate normal distribution (MVN) based imputation
SVD Based Imputation

- $A = U \cdot S \cdot V^T$
- $S$: singular values
- $V^T$: orthogonal vectors
- $S \cdot V^T$: principal components
- Regress a point with missing values against $V$
KNN and MVN Imputation
Imputation: Yeast Cell Cycle
3222 genes, 80 experiments, 3 clusters

Number of genes with missing values

Number of mis-clustered genes

Number of genes with missing values

- MVN
- KNN
- SVD
- Zero
- Col mean
- Row mean
Imputation: Yeast Cell Cycle
3222 genes, 80 experiments, 3 clusters

Root mean squared error

Number of genes with missing values
Acknowledgement

• U.S. EPA funded Center for Exposure and Risk Modeling (CERM) at EOHSI (EPAR-827033)