CAP 5510: Introduction to Bioinformatics
CGS 5166: Bioinformatics Tools

Giri Narasimhan
ECS 254; Phone: x3748
giri@cis.fiu.edu
www.cis.fiu.edu/~giri/teach/BioinfF18.html
Gene Expression

- Process of transcription and/or translation of a gene is called gene expression.
- Every cell of an organism has the same genetic material, but different genes are expressed at different times.
- Patterns of gene expression in a cell is indicative of its state.
Hybridization

- If two complementary strands of DNA or mRNA are brought together under the right experimental conditions they will hybridize.

- **A hybridizes to B** ⇒
  - A is reverse complementary to B, or
  - A is reverse complementary to a subsequence of B.

- It is possible to experimentally verify whether A hybridizes to B, by labeling A or B with a radioactive or fluorescent tag, followed by excitation by laser.
Measuring gene expression

- Gene expression for a single gene can be measured by extracting mRNA from the cell and doing a simple hybridization experiment.

- Given a sample of cells, gene expression for every gene can be measured using a single microarray experiment.
Microarray/DNA chip technology

- High-throughput method to study gene expression of thousands of genes simultaneously.

- Many applications:
  - Genetic disorders & Mutation/polymorphism detection
  - Study of disease subtypes
  - Drug discovery & toxicology studies
  - Pathogen analysis
  - Differing expressions over time, between tissues, between drugs, across disease states
# Microarray Data

<table>
<thead>
<tr>
<th>Gene</th>
<th>Expression Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene1</td>
<td></td>
</tr>
<tr>
<td>Gene2</td>
<td></td>
</tr>
<tr>
<td>Gene3</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
Gene Chips
DNA Chips & Images
Gene g

Probe 1

Probe 2

...
Microarray/DNA chips (Simplified)

- **Construct probes** corresponding to reverse complements of genes of interest.
- Microscopic quantities of probes placed on solid surfaces at defined spots on the chip.
- Extract mRNA from sample cells and **label** them.
- Apply labeled sample (mRNA extracted from cells) to every spot, and allow hybridization.
- Wash off unhybridized material.
- Use optical detector to measure amount of fluorescence from each spot.
Affymetrix DNA chip schematic

www.affymetrix.com
What’s on the slide?
Affymetrix & Agilent

Differ in:

- Method to place DNA: Spotting vs. photolithography
- Length of probe
- Complete sequence vs. series of fragments
Study effect of treatment over time
2-color DNA microarray

Treated mRNA → Cy5 Probe
Control mRNA → Cy3 Probe

Normalization → Data extraction → Scanning

Simultaneous hybridization
How to compare 2 cell samples with Two-Color Microarrays?

- mRNA from sample 1 is extracted and labeled with a red fluorescent dye.
- mRNA from sample 2 is extracted and labeled with a green fluorescent dye.
- Mix the samples and apply it to every spot on the microarray. Hybridize sample mixture to probes.
- Use optical detector to measure the amount of green and red fluorescence at each spot.
Sources of Variations & Experimental Errors

- Variations in cells/individuals
- Variations in mRNA extraction, isolation, introduction of dye, variation in dye incorporation, dye interference
- Variations in probe concentration, probe amounts, substrate surface characteristics
- Variations in hybridization conditions and kinetics
- Variations in optical measurements, spot misalignments, discretization effects, noise due to scanner lens and laser irregularities
- Cross-hybridization of sequences with high sequence identity
- Limit of factor 2 in precision of results
- Variation changes with intensity: larger variation at low or high expression levels

Need to Normalize data
Clustering is a general method to study patterns in gene expressions.

Several known methods:

- Hierarchical Clustering (Bottom-Up Approach)
- K-means Clustering (Top-Down Approach)
- Self-Organizing Maps (SOM)
Hierarchical Clustering: Example
A Dendrogram
Hierarchical Clustering [Johnson, SC, 1967]

- Given \( n \) points in \( \mathbb{R}^d \), compute the distance between every pair of points.

- While (not done)
  - Pick closest pair of points \( s_i \) and \( s_j \) and make them part of the same cluster.
  - Replace the pair by an average of the two \( s_{ij} \).

**Try the applet at:** [http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletH.html](http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletH.html)
For clustering, define a distance function:

- **Euclidean distance metrics**
  \[ D_k(X, Y) = \left[ \sum_{i=1}^{d} (X_i - Y_i)^k \right]^{1/k} \]
  
  \( k=2: \) Euclidean Distance

- **Pearson correlation coefficient**
  \[ \rho_{xy} = \frac{1}{d} \sum_{i=1}^{d} \left( \frac{X_i - \bar{X}}{\sigma_x} \right) \left( \frac{Y_i - \bar{Y}}{\sigma_y} \right) \]
  \(-1 \leq \rho_{xy} \leq 1\)
### EXHIBIT 3.4  Joint Probability Model for the Ratings of Two People

(a) $\rho_{xy} = 0$

<table>
<thead>
<tr>
<th></th>
<th>y 1</th>
<th>y 2</th>
<th>y 3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>x 3</td>
<td>1/9</td>
<td>1/9</td>
<td>1/9</td>
<td>1/3</td>
</tr>
<tr>
<td>x 2</td>
<td>1/9</td>
<td>1/9</td>
<td>1/9</td>
<td>1/3</td>
</tr>
<tr>
<td>x 1</td>
<td>1/9</td>
<td>1/9</td>
<td>1/9</td>
<td>1/3</td>
</tr>
<tr>
<td>Total</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
<td>1</td>
</tr>
</tbody>
</table>

(b) $\rho_{xy} = -1$

<table>
<thead>
<tr>
<th></th>
<th>y 1</th>
<th>y 2</th>
<th>y 3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>x 3</td>
<td>1/18</td>
<td>1/18</td>
<td>4/18</td>
<td>1/3</td>
</tr>
<tr>
<td>x 2</td>
<td>1/18</td>
<td>4/18</td>
<td>1/18</td>
<td>1/3</td>
</tr>
<tr>
<td>x 1</td>
<td>4/18</td>
<td>1/18</td>
<td>1/18</td>
<td>1/3</td>
</tr>
<tr>
<td>Total</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
<td>1</td>
</tr>
</tbody>
</table>

(c) $\rho_{xy} = -1/2$

<table>
<thead>
<tr>
<th></th>
<th>y 1</th>
<th>y 2</th>
<th>y 3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>x 3</td>
<td>4/18</td>
<td>1/18</td>
<td>1/18</td>
<td>1/3</td>
</tr>
<tr>
<td>x 2</td>
<td>1/18</td>
<td>4/18</td>
<td>1/18</td>
<td>1/3</td>
</tr>
<tr>
<td>x 1</td>
<td>1/18</td>
<td>4/18</td>
<td>1/18</td>
<td>1/3</td>
</tr>
<tr>
<td>Total</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
<td>1</td>
</tr>
</tbody>
</table>

(d) $\rho_{xy} = 1$

<table>
<thead>
<tr>
<th></th>
<th>y 1</th>
<th>y 2</th>
<th>y 3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>x 3</td>
<td>1/27</td>
<td>2/27</td>
<td>6/27</td>
<td>1/3</td>
</tr>
<tr>
<td>x 2</td>
<td>2/27</td>
<td>5/27</td>
<td>2/27</td>
<td>1/3</td>
</tr>
<tr>
<td>x 1</td>
<td>6/27</td>
<td>2/27</td>
<td>1/27</td>
<td>1/3</td>
</tr>
<tr>
<td>Total</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
<td>1</td>
</tr>
</tbody>
</table>

(e) $\rho_{xy} = -8$

<table>
<thead>
<tr>
<th></th>
<th>y 1</th>
<th>y 2</th>
<th>y 3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>x 3</td>
<td>6/27</td>
<td>3/27</td>
<td>1/27</td>
<td>1/3</td>
</tr>
<tr>
<td>x 2</td>
<td>2/27</td>
<td>5/27</td>
<td>2/27</td>
<td>1/3</td>
</tr>
<tr>
<td>x 1</td>
<td>1/27</td>
<td>2/27</td>
<td>6/27</td>
<td>1/3</td>
</tr>
<tr>
<td>Total</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
<td>1</td>
</tr>
</tbody>
</table>

(f) $\rho_{xy} = 1/2$

<table>
<thead>
<tr>
<th></th>
<th>y 1</th>
<th>y 2</th>
<th>y 3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>x 3</td>
<td>1/36</td>
<td>2/36</td>
<td>9/36</td>
<td>1/3</td>
</tr>
<tr>
<td>x 2</td>
<td>2/36</td>
<td>8/36</td>
<td>2/36</td>
<td>1/3</td>
</tr>
<tr>
<td>x 1</td>
<td>9/36</td>
<td>2/36</td>
<td>9/36</td>
<td>1/3</td>
</tr>
<tr>
<td>Total</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
<td>1</td>
</tr>
</tbody>
</table>

(g) $\rho_{xy} = -1$

<table>
<thead>
<tr>
<th></th>
<th>y 1</th>
<th>y 2</th>
<th>y 3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>x 1</td>
<td>9/36</td>
<td>2/36</td>
<td>9/36</td>
<td>1/3</td>
</tr>
<tr>
<td>x 2</td>
<td>2/36</td>
<td>8/36</td>
<td>2/36</td>
<td>1/3</td>
</tr>
<tr>
<td>x 1</td>
<td>1/36</td>
<td>2/36</td>
<td>9/36</td>
<td>1/3</td>
</tr>
<tr>
<td>Total</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
<td>1</td>
</tr>
</tbody>
</table>
Clusterings of gene expressions

- Represent each gene as a vector or a point in $d$-space where $d$ is the number of arrays or experiments being analyzed.
Clustering Random vs. Biological Data

Example from Andrew Moore’s tutorial on Clustering.
K-means
1. Decide on how many clusters you desire. (e.g., 3)
2. Randomly guess k initial center locations
3. Least square fit each center’s Euclidean distances
   4. Each data point finds out which center it is closest to
   5. Each center becomes a set of representatives

K-means continues

K-means terminates

Start

End
K-Means Clustering [McQueen ’67]

Repeat
- Start with randomly chosen cluster centers
- Assign points to give greatest increase in score
- Recompute cluster centers
- Reassign points

until (no changes)

Try the applet at: http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletH.html
Comparisons

- **Hierarchical clustering**
  - Number of clusters not preset.
  - Complete hierarchy of clusters
  - Not very robust, not very efficient.

- **K-Means**
  - Need definition of a mean. Categorical data?
  - More efficient and often finds optimum clustering.
Functionally related genes behave similarly across experiments.
Self-Organizing Maps [Kohonen]

- Kind of neural network.
- Clusters data and find complex relationships between clusters.
- Helps reduce the dimensionality of the data.
- Map of 1 or 2 dimensions produced.
- Unsupervised Clustering
- Like K-Means, except for visualization
SOM Architectures

- 2-D Grid
- 3-D Grid
- Hexagonal Grid