Microarray data analysis

Russ B. Altman
BMI 214
CS 274

Microarrays: DNA Base Pairing

Microarrays: Experimental Protocol

Typical DNA array for Yeast

Reproducibility of data sets

- Preparation of mRNA and experimental design
- Hybridization of RNA to DNA
  - Sequence-specific effects
  - Length-related effects
- Quality of spotted genes on array
  - Proper sequence spotted evenly
- Finding and digitizing spot intensities
- Comparability of experiments on same chip, experiments on different chips

What are they good for?

- Follow population of (synchronized) cells over time, to see how expression changes (vs. baseline).
- Expose cells to different external stimuli and measure their response (vs. baseline).
- Take cancer cells (or other pathology) and compare to normal cells.
Informatics Approaches

- Low level image processing of spots to assess amount of fluorescence.
- Need to deal with missing values (due to experimental artifacts, etc…)
- Need to decide how much of a change is significant (e.g. “2-fold increase” in expression).
- Creation of databases with the info (SMD)

Data Analysis

- DATA: Thousands of genes by hundreds of cases.
- Grouping gene or cases by clustering expression profiles (unsupervised machine learning)
- Comparing genes or cases to known profiles using pattern recognition (supervised machine learning)

(Raychaudhuri & Altman, Trends in Biotechnology, 2001)

Review of Microarray Analysis

- Basic Signal Processing
- Clustering methods
  - Manage the amount of data (1D clustering)
  - Look for modules (2D clustering)
- Classification methods
  - Use to associate external labels with genes
- Combining with other sources of data
  - Sequence data, Natural language, gene fusions, comparative genomics
- Reconstructing genetic networks

http://classify.stanford.edu/

Can build trees from cluster analysis, groups genes by common patterns of expression.
Average of clustered wave forms

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>Wave 1</td>
</tr>
<tr>
<td><img src="image1.png" alt="Wave forms" /></td>
<td><img src="image2.png" alt="Wave forms" /></td>
</tr>
</tbody>
</table>

Typical “wave forms” observed (note: not lots of bumps)

Cluster Analysis Result

Clustering Lymphomas

Works well if we use the appropriate 143 GC specific genes

Classifying Lymphomas

Estimate Missing Values?

- Complete data set
- Data set with 30% entries missing (missing values appear black)
- Data set with missing values estimated by KNNimpute algorithm
Independent Components Analysis

- Find projection where distribution maximizes a measure of non-normality.
- Use kurtosis as measure of non-normality.

PCA--first two components

Principle Component Analysis of Conditions in Microarray Experiment

Look for interesting genes

w_i * After + w_j * Before

w_i * After + w_j * Before
"Interesting" Projections

- max variance = PCA
- max kurtosis
- max non-Gaussianity
  - Normal = not interesting (PJ Huber ’85, Jone & Sibson ‘87)
- max information

Discovered MSE Promoter

<table>
<thead>
<tr>
<th>Gene</th>
<th>ORF</th>
<th>Site</th>
<th>Consensus</th>
</tr>
</thead>
<tbody>
<tr>
<td>YLL004W</td>
<td>ORC3</td>
<td>1</td>
<td>ATTTGTGTCAT</td>
</tr>
<tr>
<td>YML066C</td>
<td>x</td>
<td>1</td>
<td>TTTTGTGTCAT</td>
</tr>
<tr>
<td>YPR022C</td>
<td>CDC14</td>
<td>1</td>
<td>TTTTGTGACCT</td>
</tr>
<tr>
<td>YNL018C</td>
<td>x</td>
<td>2</td>
<td>TTTTGTGACAC</td>
</tr>
<tr>
<td>YIL212C</td>
<td>x</td>
<td>1</td>
<td>TTTTGTGACCT</td>
</tr>
<tr>
<td>YNL174W</td>
<td>x</td>
<td>1</td>
<td>ATTTGTGACCT</td>
</tr>
<tr>
<td>YBR069C</td>
<td>VAP1</td>
<td>1</td>
<td>TTTTGTGACAT</td>
</tr>
<tr>
<td>YDR191W</td>
<td>HST4</td>
<td>1</td>
<td>TTTTGTGACAT</td>
</tr>
<tr>
<td>YBR124W</td>
<td>HDBO</td>
<td>2</td>
<td>TTTTGTGACAT</td>
</tr>
<tr>
<td>YLR228C</td>
<td>x</td>
<td>1</td>
<td>TTTTGTGACAT</td>
</tr>
</tbody>
</table>

Consensus: TTTTGTGACAT

the regulatory element that recognizes this promoter...

Imagine other array technologies

- Protein chips to assess interaction of proteins (lay down proteins, and then label others, and look for binding events).