Effective Methods for Gene Expression Analysis and Validation

Shin-Mu Tseng 曾新穆
tsengsm@mail.ncku.edu.tw

Dept. Computer Science and Information Engineering
National Cheng Kung University
Taiwan, R.O.C.
Outline

- Gene Expression Analysis
- Microarray techniques
- Goal & steps of gene expression analysis
- A new approach for efficient microarray data analysis
- Future Directions
Current Status

- Human genome project is at finishing stage, revealing the sequences of most functional genes in a human cell

- For more than 90% of the genes, we know little about their real functions
Gene Expression Analysis

- Gene expression analysis can help understand the functions and relations of genes
- Paradigm of gene expression analysis
  - Gene expression grouping
    - Clustering methods and validation techniques
  - Reasoning
    - Classification-based methods with gene properties
    - Frequently used methods: Decision Tree, SVM, Neural Network, etc.
Microarray Techniques

- **Main Advantage of Microarray Techniques**
  - allow simultaneous studies of the expression of thousands of genes in a single experiment

- **Microarray Process**
  - Arrayer
  - Experiments: Hybridization
  - Image Capturing of Results
  - Analysis
Microarray Process

1. Normal RNA RT-PCR
   2. Experimental RNA RT-PCR
   3. Hybridize Wash Scan
   4. Est Array
   5. Yellow = Equal Expression
   6. Green = Decreased Expression
   7. Red = Increased Expression

---

Microarray Scanner

---

TreeView

---

Cluster

---

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>1</td>
<td>YCHR</td>
<td>NAME</td>
<td>GWEIGHT</td>
<td>opc0</td>
<td>opc1</td>
<td>opc2</td>
</tr>
<tr>
<td>2</td>
<td>EWEIGHT</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>YA1005WIEB1</td>
<td>1</td>
<td>0.23</td>
<td>-1.79</td>
<td>-2.29</td>
<td>-2.06</td>
</tr>
<tr>
<td>4</td>
<td>YA1004WYA1004W</td>
<td>1</td>
<td>0.41</td>
<td>-0.39</td>
<td>-0.89</td>
<td>-1.06</td>
</tr>
<tr>
<td>5</td>
<td>YA1005CSEX</td>
<td>1</td>
<td>0.61</td>
<td>-0.07</td>
<td>-1.29</td>
<td>-1.29</td>
</tr>
<tr>
<td>6</td>
<td>YA1005CMDA10</td>
<td>1</td>
<td>0.16</td>
<td>-0.15</td>
<td>-0.06</td>
<td>-1.25</td>
</tr>
</tbody>
</table>
**Goal of Microarray Data Analysis**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Test</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>...</th>
<th>...</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6</td>
<td>0.4</td>
<td>0.2</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
<td>0.9</td>
<td>0.8</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0.3</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.7</td>
<td>0.5</td>
<td>0.2</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0.3</td>
<td>0.8</td>
<td>0.7</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Gene Expression Profile Analysis*
# Goal of Microarray Data Analysis

Gene Expression Profile Analysis

<table>
<thead>
<tr>
<th>gene</th>
<th>test 1</th>
<th>test 2</th>
<th>test 3</th>
<th>test 4</th>
<th>...</th>
<th>...</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6</td>
<td>0.4</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
<td>0.9</td>
<td>0.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.7</td>
<td>0.5</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0.3</td>
<td>0.8</td>
<td>0.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Sample Clustering Results
# Time Course Data

<table>
<thead>
<tr>
<th>YORF</th>
<th>0 minutes</th>
<th>30 minutes</th>
<th>1 hour</th>
<th>2 hours</th>
<th>4 hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>YAL001C</td>
<td>1</td>
<td>1.3</td>
<td>2.4</td>
<td>5.8</td>
<td>2.4</td>
</tr>
<tr>
<td>YAL002W</td>
<td>0.9</td>
<td>0.8</td>
<td>0.7</td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td>YAL003W</td>
<td>0.8</td>
<td>2.1</td>
<td>4.2</td>
<td>10.1</td>
<td>10.1</td>
</tr>
<tr>
<td>YAL005C</td>
<td>1.1</td>
<td>1.3</td>
<td>0.8</td>
<td></td>
<td>0.4</td>
</tr>
<tr>
<td>YAL010C</td>
<td>1.2</td>
<td>1</td>
<td>1.1</td>
<td>4.5</td>
<td>8.3</td>
</tr>
</tbody>
</table>
Time Course Data (cont.)

Absolute offset

Scaling

Global similarity
Gene Expression Analysis Steps

1. Data quantification
   - Data quality problems
2. Data filtering & adjustment
3. Data Mining
   - Adopted mining methods?
   - How good is the mining result?
4. Related information retrieval
   - Web access
5. Reporting
Data Mining Methods

- **Clustering Methods**
  - Partitioning: K-Means, K-Medoids, PAM, CLARA, ...
  - Hierarchical: HAC, BIRCH, CURE, ROCK
  - Density-based: CAST, DBSCAN, OPTICS, CLIQUE, ...
  - Grid-based: STING, CLIQUE, WaveCluster, ...
  - Model-based: SOM, COBWEB, CLASSIT, AutoClass, ...
  - Two-way Clustering
  - Block clustering

- **Attributes & Sample Selection Methods**
  - Principle component analysis (PCA)
  - Factor Analysis (FA)
  - Sampling
Clustering Methods (cont.)

Partitioning
K-means Clustering (cont.)
K-means Clustering (cont.)
Clustering Methods (cont.)

Hierarchical

0 1 2 3 4 agglomerative

0 1 2 3 4 divisive

Hierarchical

a, b
c
d, e
ea, b, c, d, e
c, d, e
da, e
d, e
e

17
Clustering Methods (cont.)
Clustering Methods (cont.)

Density-based
CAST Clustering

- **Input**
  - $S$: a symmetric $n \times n$ Similarity Matrix, $S(i,j) \in [0, 1]$
  - $t$: Affinity Threshold (0 < $t$ < 1)

- **Method**
  1. Choose a seed for generating a new cluster
  2. ADD: add qualified items to the cluster
  3. REMOVE: remove unqualified items from the stable cluster
  4. Repeat Steps 1-3 till no more clusters can be generated
Similarity Measurements

- Types of similarity measurements
  - distance measurements
  - correlation coefficients
  - association coefficients
  - probabilistic similarity coefficients
Similarity Measurements: Correlation Coefficients

- The most popular correlation coefficient is **Pearson** correlation coefficient. (1892)
- Correlation between \( X=\{X_1, X_2, \ldots, X_n\} \) and \( Y=\{Y_1, Y_2, \ldots, Y_n\} \):
  \[
  r = \frac{1}{n} \sum_{k=1}^{n} \left( \frac{X_k - \bar{X}}{\sigma_X} \right) \left( \frac{Y_k - \bar{Y}}{\sigma_Y} \right)
  \]
- Where
  \[
  \sigma_G = \sqrt{\frac{\sum_{k=1}^{n} (G_k - \bar{G})^2}{n}}
  \]
Similarity Measurements
Correlation Coefficients (cont.)

- It captures similarity of the “shapes” of two expression profiles, and ignores differences between their magnitudes.

\[ r = 1.0 \]
Problems in Microarray Mining

- How to cluster microarray data with the following requirements met simultaneously?
  - Efficiency
  - Accuracy
  - Automation
Problems in Microarray Mining (cont.)

- How to cluster microarray data with the following requirements met simultaneously?
  - Efficiency
  - Accuracy
  - Automation

Efficient Clustering Methods + Validation Techniques
Validation Techniques: Hubert’s Γ Statistics

- \( X = [X(i, j)] \) and \( Y = [Y(i, j)] \) are two \( n \times n \) matrix
  - \( X(i, j) \): similarity of gene \( i \) and gene \( j \)
  - \( Y(i, j) = \begin{cases} 1 & \text{if genes } i \text{ and } j \text{ are in same cluster,} \\ 0 & \text{otherwise} \end{cases} \)
- Hubert’s \( \Gamma \) statistic represents the point serial correlation:
  \[
  \Gamma = \frac{1}{M} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left( \frac{X(i, j) - \overline{X}}{\sigma_X} \right) \left( \frac{Y(i, j) - \overline{Y}}{\sigma_Y} \right)
  \]
  - where \( M = n(n - 1)/2 \)
- A higher value of \( \Gamma \) represents the better clustering quality.
Efficient Microarray Mining [Tseng 02]

- CAST algorithm for clustering efficiency
- Hubert’s $\Gamma$ statistic for validating clustering results
- Iterative computation for automatic mining

![Graph showing Hubert's $\Gamma$ statistic over $t$ (%)]

$\Gamma$ (Hubert's statistic) vs. $t$ (percentage)
Iterative Approach

1. Narrow down the threshold range
2. Split and Conquer: find “nearly-best” result

LM: Left Margin
RM: Right Margin

Affinity Threshold
Experimental Evaluation

Biological Dataset

- Original dataset
  - Data source: Lawrence Berkeley National Lab (LBNL) (http://rana.lbl.gov/EisenData.htm)
  - microarray expression data of yeast saccharomyces cerevisiae
  - contain the expressions of 6221 genes under 80 experimental conditions

- Testing datasets
  - Dataset I: low similarity dataset (avg similarity: 0.137)
  - Dataset II: high similarity dataset (avg similarity: 0.696)
Experimental Evaluation
Low Similarity Dataset

Table 1. Experimental results (dataset I)

<table>
<thead>
<tr>
<th>Methods</th>
<th>time (sec)</th>
<th># cluster</th>
<th>$\Gamma$ statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Approach</td>
<td>27</td>
<td>57</td>
<td>0.514</td>
</tr>
<tr>
<td>CAST</td>
<td>246</td>
<td>57</td>
<td>0.514</td>
</tr>
<tr>
<td>k-means ($k=3 \sim 21$)</td>
<td>404</td>
<td>5</td>
<td>0.447</td>
</tr>
<tr>
<td>k-means ($k=3 \sim 39$)</td>
<td>1092</td>
<td>5</td>
<td>0.447</td>
</tr>
</tbody>
</table>

Table 2. Distribution of clusters (dataset I)

<table>
<thead>
<tr>
<th>Methods</th>
<th>Size</th>
<th>1~10</th>
<th>11~100</th>
<th>101~400</th>
<th>401~600</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Approach</td>
<td>38</td>
<td>15</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>CAST</td>
<td>38</td>
<td>15</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>k-means ($k=3 \sim 21$)</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>k-means ($k=3 \sim 39$)</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Experimental Evaluation—
High Similarity Dataset

Table 3. Experimental results (dataset II)

<table>
<thead>
<tr>
<th>Methods</th>
<th>time (sec)</th>
<th># cluster</th>
<th>Γ statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Approach</td>
<td>13</td>
<td>63</td>
<td>0.833</td>
</tr>
<tr>
<td>CAST</td>
<td>41</td>
<td>62</td>
<td>0.833</td>
</tr>
<tr>
<td>k-means ((k=2 \sim 20))</td>
<td>77</td>
<td>12</td>
<td>0.309</td>
</tr>
<tr>
<td>k-means ((k=2 \sim 38))</td>
<td>267</td>
<td>12</td>
<td>0.309</td>
</tr>
</tbody>
</table>

Table 4. Distribution of clusters (dataset II)

<table>
<thead>
<tr>
<th>Methods</th>
<th>Size</th>
<th>1~10</th>
<th>11~100</th>
<th>101~300</th>
<th>1901~2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Approach</td>
<td>62</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>CAST</td>
<td>61</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>k-means ((k=2 \sim 20))</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>k-means ((k=2 \sim 38))</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Experimental Evaluation

Artificial Dataset

- Dataset
  - Dataset III (2000 items, 15 dimensions)
    - Cluster 1, 2 and 3 contain 800, 500 and 300 items mixed with 400 noises.
  - Dataset IV (1640 items, 15 dimensions)
    - Cluster 1, 2 and 3 contain 800, 500 and 300 items mixed with 40 noises.

- Seed item of the three clusters: generated by SPSS
Experimental Evaluation - Artificial Dataset (cont.)

- Profile of each cluster
Experimental Evaluation - Artificial Dataset (cont.)
Experimental Evaluation—Dataset III

Table 5. Experimental results (Pearson’s correlation coefficient)

<table>
<thead>
<tr>
<th>methods</th>
<th>time(sec)</th>
<th>#cluster</th>
<th>$\Gamma$ statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Approach</td>
<td>3</td>
<td>167</td>
<td>0.947</td>
</tr>
<tr>
<td>CAST-FI</td>
<td>12</td>
<td>153</td>
<td>0.947</td>
</tr>
<tr>
<td>k-means ($k=2\sim20$)</td>
<td>176</td>
<td>5</td>
<td>0.827</td>
</tr>
<tr>
<td>k-means ($k=2\sim39$)</td>
<td>874</td>
<td>5</td>
<td>0.827</td>
</tr>
</tbody>
</table>

Table 6. Hit rate (Pearson’s correlation coefficient)

<table>
<thead>
<tr>
<th>methods</th>
<th>C1 (#800)</th>
<th>C2 (#500)</th>
<th>C3 (#300)</th>
<th>Noises (#400)</th>
<th>Hit rate(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Approach</td>
<td>804(+4)</td>
<td>505(+5)</td>
<td>304(+4)</td>
<td>387(-13)</td>
<td>99.35</td>
</tr>
<tr>
<td>CAST-FI</td>
<td>804(+4)</td>
<td>505(+5)</td>
<td>305(+5)</td>
<td>386(-14)</td>
<td>99.3</td>
</tr>
<tr>
<td>k-means ($k=2\sim20$)</td>
<td>874(+74)</td>
<td>584(+84)</td>
<td>215(+58)/202(+59)</td>
<td>125(-275)</td>
<td>86.25</td>
</tr>
<tr>
<td>k-means ($k=2\sim39$)</td>
<td>874(+74)</td>
<td>584(+84)</td>
<td>215(+58)/202(+59)</td>
<td>125(-275)</td>
<td>86.25</td>
</tr>
</tbody>
</table>
# Experimental Evaluation

## Dataset IV

### Table 9. Experimental results (*Pearson’s correlation coefficient*)

<table>
<thead>
<tr>
<th>methods</th>
<th>time(sec)</th>
<th>#cluster</th>
<th>$\Gamma$ statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Approach</td>
<td>3</td>
<td>36</td>
<td>0.97</td>
</tr>
<tr>
<td>CAST-FI</td>
<td>8</td>
<td>36</td>
<td>0.97</td>
</tr>
<tr>
<td>k-means ($k=2$~$20$)</td>
<td>129</td>
<td>3</td>
<td>0.939</td>
</tr>
<tr>
<td>k-means ($k=2$~$39$)</td>
<td>549</td>
<td>3</td>
<td>0.939</td>
</tr>
</tbody>
</table>

### Table 10. Hit rate (*Pearson’s correlation coefficient*)

<table>
<thead>
<tr>
<th>methods</th>
<th>C1 (#800)</th>
<th>C2 (#500)</th>
<th>C3 (#300)</th>
<th>Noises (#400)</th>
<th>Hit rate(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Approach</td>
<td>800(0)</td>
<td>500(0)</td>
<td>300(0)</td>
<td>40(0)</td>
<td>100</td>
</tr>
<tr>
<td>CAST-FI</td>
<td>800(0)</td>
<td>500(0)</td>
<td>300(0)</td>
<td>40(0)</td>
<td>100</td>
</tr>
<tr>
<td>k-means ($k=2$~$20$)</td>
<td>812(+12)</td>
<td>516(+16)</td>
<td>312(+12)</td>
<td>0(-40)</td>
<td>97.561</td>
</tr>
<tr>
<td>k-means ($k=2$~$39$)</td>
<td>812(+12)</td>
<td>516(+16)</td>
<td>312(+12)</td>
<td>0(-40)</td>
<td>97.561</td>
</tr>
</tbody>
</table>
Future Directions

- More effective clustering methods
  - Handling also time-course data
- Effective validation techniques
- Classification modeling based on gene expression clustering results
  - Understand gene functions
  - Building relation network of genes
Thanks