For written notes on this lecture, please read chapter 14 of *The Practical Bioinformatician.*

CS2220: Introduction to Computational Biology

Unit 4: Gene expression analysis

Li Xiaoli
Plan

- Microarray background
- Gene expression profile clustering
- Some standard clustering methods
Background on Microarrays
What is a Microarray?

- Gene expression is the process by which information from a gene is used in the synthesis of a functional gene products, e.g. functional RNA, proteins.

- Genes are expressed by being transcribed into RNA, and this transcript may then be translated into protein.

http://en.wikipedia.org/wiki/Gene_expression
What is a Microarray?

• Contain large number of DNA molecules spotted on glass slides, nylon membranes, or silicon wafers

• Detect what genes are being *expressed* in a cell of a tissue sample

• Measure expression of thousands of genes simultaneously
Good Videos on Microarray Introduction

- **Short Video (1-3 min each)**
  - [http://www.youtube.com/watch?v=_6ZMEZK-alM](http://www.youtube.com/watch?v=_6ZMEZK-alM)
  - [http://www.youtube.com/watch?v=VNsThMNjKhM](http://www.youtube.com/watch?v=VNsThMNjKhM)
  - [http://www.youtube.com/watch?v=SNbt--d14P4](http://www.youtube.com/watch?v=SNbt--d14P4)

- **Long Video (25 min)**
  - [http://www.youtube.com/watch?v=0Hj3f7vQFZU](http://www.youtube.com/watch?v=0Hj3f7vQFZU)
Perform Web Lab experiments

• Key Idea: If a gene is expressed, then it will generate mRNA. When we produce cDNA from mRNA, cDNA and DNA will be attracted to bind together.

According to base pairing rules (A with T and C with G), *hydrogen bonds* bind the bases of the two separate polynucleotide strands (DNA, cDNA) together

How to do Wet Lab experiments
http://www.bio.davidson.edu/Courses/genomics/chip/chip.html
A Sample Affymetrix GeneChip Data File (U95A)

<table>
<thead>
<tr>
<th>Positive</th>
<th>Negative</th>
<th>Pairs</th>
<th>lnA_W</th>
<th>Avg Diff</th>
<th>Abs Call</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFFX-Murl</td>
<td>5</td>
<td>2</td>
<td>19</td>
<td>297.5</td>
<td>A</td>
<td>M16762 Mouse interleukin 2 (IL-2) gene, exon 4</td>
</tr>
<tr>
<td>AFFX-Murl</td>
<td>3</td>
<td>2</td>
<td>19</td>
<td>554.2</td>
<td>A</td>
<td>M37897 Mouse interleukin 10 mRNA, complete cds</td>
</tr>
<tr>
<td>AFFX-Murl</td>
<td>4</td>
<td>2</td>
<td>19</td>
<td>308.6</td>
<td>A</td>
<td>M25892 Mus musculus interleukin 4 (II-4) mRNA, complete cds</td>
</tr>
<tr>
<td>AFFX-Murf</td>
<td>1</td>
<td>3</td>
<td>19</td>
<td>141</td>
<td>A</td>
<td>M83649 Mus musculus Fas antigen mRNA, complete cds</td>
</tr>
<tr>
<td>AFFX-BioE</td>
<td>13</td>
<td>1</td>
<td>19</td>
<td>9340.6</td>
<td>P</td>
<td>J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r)</td>
</tr>
<tr>
<td>AFFX-BioE</td>
<td>15</td>
<td>0</td>
<td>19</td>
<td>12862.4</td>
<td>P</td>
<td>J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r)</td>
</tr>
<tr>
<td>AFFX-BioE</td>
<td>12</td>
<td>0</td>
<td>19</td>
<td>8716.5</td>
<td>P</td>
<td>J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r)</td>
</tr>
<tr>
<td>AFFX-BioC</td>
<td>17</td>
<td>0</td>
<td>19</td>
<td>25942.5</td>
<td>P</td>
<td>J04423 E coli bioC protein (-5 and -3 represent transcript)</td>
</tr>
<tr>
<td>AFFX-BioC</td>
<td>16</td>
<td>0</td>
<td>20</td>
<td>28838.5</td>
<td>P</td>
<td>J04423 E coli bioC protein (-5 and -3 represent transcript)</td>
</tr>
<tr>
<td>AFFX-BioD</td>
<td>17</td>
<td>0</td>
<td>19</td>
<td>25765.2</td>
<td>P</td>
<td>J04423 E coli bioD gene dethiobiotin synthetase (-5 and -3 represent transcript)</td>
</tr>
<tr>
<td>AFFX-BioC</td>
<td>19</td>
<td>0</td>
<td>20</td>
<td>140113.2</td>
<td>P</td>
<td>J04423 E coli bioD gene dethiobiotin synthetase (-5 and -3 represent transcript)</td>
</tr>
<tr>
<td>AFFX-CreX</td>
<td>20</td>
<td>0</td>
<td>20</td>
<td>280036.6</td>
<td>P</td>
<td>X03453 Bacteriophage P1 cre recombinase protein (-5 and -3 represent transcript)</td>
</tr>
<tr>
<td>AFFX-CreX</td>
<td>20</td>
<td>0</td>
<td>20</td>
<td>401741.8</td>
<td>P</td>
<td>X03453 Bacteriophage P1 cre recombinase protein (-5 and -3 represent transcript)</td>
</tr>
<tr>
<td>AFFX-BioE</td>
<td>7</td>
<td>5</td>
<td>18</td>
<td>-483</td>
<td>A</td>
<td>J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r)</td>
</tr>
<tr>
<td>AFFX-BioE</td>
<td>5</td>
<td>4</td>
<td>18</td>
<td>313.7</td>
<td>A</td>
<td>J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r)</td>
</tr>
<tr>
<td>AFFX-BioE</td>
<td>7</td>
<td>6</td>
<td>20</td>
<td>-1016.2</td>
<td>A</td>
<td>J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r)</td>
</tr>
</tbody>
</table>

The imp’t field is “Avg Diff”, which gives the expression level of the gene. The “Abs Call” field is also imp’t, which tells whether the corresponding number in the “Avg Diff” field reliable or not. “P” means present and thus the number is reliable. “A” and “M” tell you the number is unreliable and should be ignored.

http://yfgdb.princeton.edu/Affymetrix_Empirical.txt

Copyright 2015 © Wong Limsoon, Li Xiaoli
Some additional biological knowledge on gene expression regulation

- Regulation of gene expression refers to the control of the *amount* and *timing* of appearance of the functional product of a gene.

- Control of expression is vital to allow a cell to produce the gene products it needs when it needs them; in turn this gives cells the flexibility to adapt to a variable environment, external signals, damage to the cell.

The patchy colours of a tortoiseshell cat are the result of different levels of expression of pigmentation genes in different areas of the skin.
Gene types depending on how they are regulated

• A **constitutive** gene continually transcribes to mRNA

• A **housekeeping** gene is typically a constitutive gene that is transcribed at a relatively **constant** level. The housekeeping gene's products are typically needed for maintenance of the cell.

• A **facultative/ inducible** gene is a gene only transcribed when needed as opposed to a constitutive gene. Its expression is either responsive to environmental change or dependent on the position in the cell cycle.
Example of Real Gene Expression Data

• [http://nemates.org/uky/520/Lab/lab10/yeastall_public.txt](http://nemates.org/uky/520/Lab/lab10/yeastall_public.txt)

• **Exercise:** store the whole gene expression data into a excel file to understand more
### Type of Gene Expression Datasets

- **Gene-Conditions or Gene-Sample** *(numeric or discretized)*

<table>
<thead>
<tr>
<th>Class</th>
<th>Gene1</th>
<th>Gene2</th>
<th>Gene3</th>
<th>Gene4</th>
<th>Gene5</th>
<th>Gene6</th>
<th>Gene7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample1</td>
<td>Cancer</td>
<td>0.12</td>
<td>-1.3</td>
<td>1.7</td>
<td>1.0</td>
<td>-3.2</td>
<td>0.78</td>
</tr>
<tr>
<td>Sample2</td>
<td>Cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.3</td>
</tr>
<tr>
<td>~Cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SampleN</td>
<td>~Cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Gene-Time** *(different genes)*

![Gene-Time Diagram](image)

- **Gene-Sample-Time**

![Gene-Sample-Time Diagram](image)
## Type of Gene Expression Datasets

- **Gene-Conditions or Gene-Sample** (numeric or discretized)

<table>
<thead>
<tr>
<th></th>
<th>Class</th>
<th>Gene1</th>
<th>Gene2</th>
<th>Gene3</th>
<th>Gene4</th>
<th>Gene5</th>
<th>Gene6</th>
<th>Gene7</th>
<th>.....</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample1</td>
<td>Cancer</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>....</td>
</tr>
<tr>
<td>Sample2</td>
<td>Cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>....</td>
</tr>
<tr>
<td>SampleN</td>
<td>~Cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>....</td>
</tr>
</tbody>
</table>

- **Gene-Time**

- **Gene-Sample-Time**

![Graph showing gene expression levels over time](image)

![Diagram illustrating gene-sample-time relationship](image)
Application: Disease Diagnosis

Gene Expression data to perform diagnostic task
Identify the biomarkers of people who will benefit from continued use of the drug. We can thus predict the treatment outcomes, e.g. working or not-working or should we give a patient the treatment?
Application: Drug Action Detection

Which group of genes are the drug affecting on?

With drugs, if the gene expression values have big changes?
Gene Expression Profile Clustering

Novel Disease Subtype Discovery
Childhood Acute lymphoblastic leukemia (ALL)

• Existing major subtypes:
  1. T-ALL,
  2. E2A-PBX,
  3. TEL-AML,
  4. BCR-ABL,
  5. MLL genome rearrangements,
  6. Hyperdiploid>50
## Type of Gene Expression Dataset

### Gene-Sample (numeric)

<table>
<thead>
<tr>
<th>Gene 1</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
<th>Sample 4</th>
<th>Sample 5</th>
<th>Sample 6</th>
<th>Sample 7</th>
<th>.....</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.12</td>
<td>0.34</td>
<td>-0.23</td>
<td>-0.34</td>
<td>0.28</td>
<td>0.11</td>
<td>0.23</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene 2</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene N</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
</table>
Is there a new subtype?

- Hierarchical clustering of gene expression profiles reveals a novel subtype of childhood ALL.
Clustering Methods

- K-means
- Hierarchical Clustering
What is Cluster Analysis?

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups.

Intra-cluster distances are minimized

Inter-cluster distances are maximized
Notion of a cluster can be ambiguous

How many clusters?
- Two Clusters
- Four Clusters
- Six Clusters

We use colors to represent the clustering results/groups.
We could also have
K-means Clustering

- Partitional clustering approach
- Each cluster is associated with a centroid (center point)
- Each point is assigned to the cluster with the closest centroid
- Number of clusters, $K$, must be specified
- The basic algorithm is very simple

1: Select $K$ points as the initial centroids.
2: repeat
3: Form $K$ clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: until The centroids don’t change
K-means Clustering Illustration

Iteration 6
K-means Clustering Illustration
Importance of Choosing Initial Centroids ...
Hierarchical Clustering

- **Two main types of hierarchical clustering**
  - Agglomerative:
    - Start with the points as individual clusters
    - At each step, merge *the closest pair of clusters* until only one cluster (or $k$ clusters) left
  - Divisive:
    - Start with one, all-inclusive cluster
    - At each step, split a cluster until each cluster contains a point (or there are $k$ clusters)

- **Traditional hierarchical algorithms use a similarity or distance matrix**
  - Merge or split one cluster at a time
Agglomerative Clustering Algorithm

- More popular hierarchical clustering technique
- Basic algorithm is straightforward
  1. Compute the proximity matrix
  2. Let each data point be a cluster
  3. Repeat
  4. Merge the two closest clusters
  5. Update the proximity matrix
  6. Until only a single cluster remains

- Key operation is the computation of the proximity of two clusters
  - Different approaches to defining the distance/similarity between clusters distinguish the different algorithms
Visualization of Agglomerative Hierarchical Clustering

Traditional Hierarchical Clustering

Traditional Dendrogram
Single Linkage defines distance between two clusters as the minimum distance between them.

Complete Linkage defines distance between two clusters as the maximum distance between them.

Exercise: Give definition of “average linkage”
Simulation: Starting Situation

• Start with clusters of individual points and a proximity matrix

<table>
<thead>
<tr>
<th></th>
<th>p1</th>
<th>p2</th>
<th>p3</th>
<th>p4</th>
<th>p5</th>
<th>\ldots</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Proximity Matrix

\[
\begin{array}{cccccc}
p1 & p2 & p3 & p4 & p5 & \ldots \\
p1 &   &   &   &   &   \\
p2 &   &   &   &   &   \\
p3 &   &   &   &   &   \\
p4 &   &   &   &   &   \\
p5 &   &   &   &   &   \\
\end{array}
\]
Intermediate Situation

- After some merging steps, we have some clusters

\[
\begin{array}{cccccc}
 & C1 & C2 & C3 & C4 & C5 \\
C1 &   &    &    &    &    \\
C2 &    &   &    &    &    \\
C3 &    &   &   &    &    \\
C4 &    &   & &   &    \\
C5 &    &   & &   &    \\
\end{array}
\]

Proximity Matrix
• We want to **merge** the two *closest* clusters (C2 and C5) and update the proximity matrix.

![Proximity Matrix Diagram]

[Diagram showing the merging of clusters C2 and C5 and updating the proximity matrix.]
After Merging

• The question is “How do we **update** the proximity matrix?”

![Proximity Matrix Diagram](image)

```
C2  
<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C5</th>
<th>C3</th>
<th>C4</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C2 U C5</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>C3</td>
<td>?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C4</td>
<td>?</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```
How to Define Inter-Cluster Similarity

- MIN
- MAX
- Group Average
- Distance Between Centroids

<table>
<thead>
<tr>
<th></th>
<th>p1</th>
<th>p2</th>
<th>p3</th>
<th>p4</th>
<th>p5</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Proximity Matrix
How to Define Inter-Cluster Similarity

- MIN
- MAX
- Group Average
- Distance/similarity Between Centroids

<table>
<thead>
<tr>
<th></th>
<th>p1</th>
<th>p2</th>
<th>p3</th>
<th>p4</th>
<th>p5</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Proximity Matrix
How to Define Inter-Cluster Similarity

- **MIN**
- **MAX**
- **Group Average**
- **Distance/similarity Between Centroids**

<table>
<thead>
<tr>
<th></th>
<th>p1</th>
<th>p2</th>
<th>p3</th>
<th>p4</th>
<th>p5</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Proximity Matrix
How to Define Inter-Cluster Similarity

- MIN
- MAX
- Group Average
- Distance/similarity Between Centroids

Proximity Matrix

<table>
<thead>
<tr>
<th></th>
<th>p1</th>
<th>p2</th>
<th>p3</th>
<th>p4</th>
<th>p5</th>
<th>…</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Proximity Matrix
How to Define Inter-Cluster Similarity

- MIN
- MAX
- Group Average
- Distance/similarity Between Centroids

Proximity Matrix

<table>
<thead>
<tr>
<th></th>
<th>p1</th>
<th>p2</th>
<th>p3</th>
<th>p4</th>
<th>p5</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

...
Cluster Similarity: MIN or Single Link

- Similarity of two clusters is based on the two most similar (closest) points in the different clusters
  - Determined by one pair of points, i.e., by one link in the proximity graph.

<table>
<thead>
<tr>
<th></th>
<th>p1</th>
<th>p2</th>
<th>p3</th>
<th>p4</th>
<th>p5</th>
<th>p6</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td>0.00</td>
<td>0.24</td>
<td>0.22</td>
<td>0.37</td>
<td>0.34</td>
<td>0.23</td>
</tr>
<tr>
<td>p2</td>
<td>0.24</td>
<td>0.00</td>
<td>0.15</td>
<td>0.20</td>
<td>0.14</td>
<td>0.25</td>
</tr>
<tr>
<td>p3</td>
<td>0.22</td>
<td>0.15</td>
<td>0.00</td>
<td>0.15</td>
<td>0.28</td>
<td>0.11</td>
</tr>
<tr>
<td>p4</td>
<td>0.37</td>
<td>0.20</td>
<td>0.15</td>
<td>0.00</td>
<td>0.29</td>
<td>0.22</td>
</tr>
<tr>
<td>p5</td>
<td>0.34</td>
<td>0.14</td>
<td>0.28</td>
<td>0.29</td>
<td>0.00</td>
<td>0.39</td>
</tr>
<tr>
<td>p6</td>
<td>0.23</td>
<td>0.25</td>
<td>0.11</td>
<td>0.22</td>
<td>0.39</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 8.4. Euclidean distance matrix for 6 points.
Hierarchical Clustering: MIN

Single Link Clustering

Single Link Dendrogram
Strength of MIN

• Can handle non-elliptical shapes

Original Points

Two Clusters

The algo likely to merge the points within same clusters if they are clearly separated
Limitations of MIN

• Sensitive to noise and outliers: cc

Original Points

Two Clusters
Cluster Similarity: MAX or Complete Linkage

- **Similarity of two clusters is based on the two least similar (most distant) points in the different clusters**
  - Determined by all pairs of points in the two clusters

<table>
<thead>
<tr>
<th></th>
<th>p1</th>
<th>p2</th>
<th>p3</th>
<th>p4</th>
<th>p5</th>
<th>p6</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td>0.00</td>
<td>0.24</td>
<td>0.22</td>
<td>0.37</td>
<td>0.34</td>
<td>0.23</td>
</tr>
<tr>
<td>p2</td>
<td>0.24</td>
<td>0.00</td>
<td>0.15</td>
<td>0.20</td>
<td>0.14</td>
<td>0.25</td>
</tr>
<tr>
<td>p3</td>
<td>0.22</td>
<td>0.15</td>
<td>0.00</td>
<td>0.15</td>
<td>0.28</td>
<td>0.11</td>
</tr>
<tr>
<td>p4</td>
<td>0.37</td>
<td>0.20</td>
<td>0.15</td>
<td>0.00</td>
<td>0.29</td>
<td>0.22</td>
</tr>
<tr>
<td>p5</td>
<td>0.34</td>
<td>0.14</td>
<td>0.28</td>
<td>0.29</td>
<td>0.00</td>
<td>0.39</td>
</tr>
<tr>
<td>p6</td>
<td>0.23</td>
<td>0.25</td>
<td>0.11</td>
<td>0.22</td>
<td>0.39</td>
<td>0.00</td>
</tr>
</tbody>
</table>

**Table 8.4.** Euclidean distance matrix for 6 points.
Hierarchical Clustering: MAX

Note we still want to merge two most similar clusters each time. However, we define the distance between clusters based on MAX.
Strength of MAX

• distance based on most distant points in the different clusters

Original Points

Two Clusters

• Less susceptible to noise and outliers
Limitations of MAX

Original Points
- Tends to break large clusters (two big, so they are far away)
- Biased towards globular clusters

Two Clusters
Cluster Similarity: Group Average

- Proximity of two clusters is the average of pairwise proximity between points in the two clusters.

\[
\text{proximity(Cluster}_i, \text{Cluster}_j) = \frac{\sum_{p_i \in \text{Cluster}_i, p_j \in \text{Cluster}_j} \text{proximity}(p_i, p_j)}{|\text{Cluster}_i| \times |\text{Cluster}_j|}
\]

- Need to use average connectivity for scalability since total proximity favors large clusters.

<table>
<thead>
<tr>
<th></th>
<th>p1</th>
<th>p2</th>
<th>p3</th>
<th>p4</th>
<th>p5</th>
<th>p6</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td>0.00</td>
<td>0.24</td>
<td>0.22</td>
<td>0.37</td>
<td>0.34</td>
<td>0.23</td>
</tr>
<tr>
<td>p2</td>
<td>0.24</td>
<td>0.00</td>
<td>0.15</td>
<td>0.20</td>
<td>0.14</td>
<td>0.25</td>
</tr>
<tr>
<td>p3</td>
<td>0.22</td>
<td>0.15</td>
<td>0.00</td>
<td>0.15</td>
<td>0.28</td>
<td>0.11</td>
</tr>
<tr>
<td>p4</td>
<td>0.37</td>
<td>0.20</td>
<td>0.15</td>
<td>0.00</td>
<td>0.29</td>
<td>0.22</td>
</tr>
<tr>
<td>p5</td>
<td>0.34</td>
<td>0.14</td>
<td>0.28</td>
<td>0.29</td>
<td>0.00</td>
<td>0.39</td>
</tr>
<tr>
<td>p6</td>
<td>0.23</td>
<td>0.25</td>
<td>0.11</td>
<td>0.22</td>
<td>0.39</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 8.4. Euclidean distance matrix for 6 points.
Hierarchical Clustering: Group Average

Group Average Clustering

Group Average Dendrogram
Hierarchical Clustering: Group Average

- **Compromise between Single and Complete Link**

- **Strengths**
  - Less susceptible to noise and outliers

- **Limitations**
  - Biased towards globular clusters
Hierarchical Clustering: Comparison

MIN

MAX

Group Average
Hierarchical Clustering: Time and Space requirements

• \(O(N^2)\) space since it uses the proximity matrix.
  – \(N\) is the number of points.

• \(O(N^3)\) time in many cases
  – There are \(N\) steps and at each step the size, \(N^2\), proximity matrix must be updated and searched
  – Complexity can be reduced to \(O(N^2 \log(N))\) time for some approaches
Bi-clustering in Gene Expression Datasets

• What happens if the similarity does not exist for all the attributes?
• More advanced clustering techniques: Bi-clustering, i.e. cluster both rows and columns simultaneously

• http://www.powershow.com/view/11b05a-ZTg4N/Biclustering_in_Gene_Expression_Datasets_powerpoint_ppt.presentation

• Slide 1 - 7
Thank You

Contact: xlli@i2r.a-star.edu.sg if you have questions
For written notes on this lecture, please read chapter 14 of *The Practical Bioinformatician*.

CS2220: Introduction to Computational Biology

Unit 4: Gene expression analysis (2)

Li Xiaoli
Plan

• Normalization

• Compute Similarities/Distances Between Two Gene’s Expression Profiles

• Gene Expression Profile Classification

• Gene Interaction Prediction

• Simple Introduction of Gene Ontology
Normalization
Sometimes, a gene expression study may involve batches of data collected over a long period of time…
In such a case, batch effect may be severe... to the extent that you can predict the batch that each sample comes!

⇒ Need normalization to correct for batch effect
Approaches to Normalization

- **Aim of normalization:** Reduce variance w/o increasing bias

- **Scaling method**
  - Intensities are scaled so that each array has same ave value
  - E.g., Affymetrix’s

- **Xform data so that distribution of probe intensities is same on all arrays**
  - E.g., \((x - \mu) / \sigma\)

- **Quantile normalization**
Quantitative Normalization

- Given \( n \) arrays of length \( p \), form \( X \) of size \( p \times n \) where each array is a column.
- Sort each column of \( X \) to give \( X_{\text{sort}} \).
- Take means across rows of \( X_{\text{sort}} \) and assign this mean to each element in the row to get \( X'_{\text{sort}} \).
- Get \( X_{\text{normalized}} \) by arranging each column of \( X'_{\text{sort}} \) to have same ordering as \( X \).

- Implemented in some microarray s/w, e.g., EXPANDER.
Can you perform Quantitive Normalization?

Array 1, 2, …, n

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>…</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.8</td>
<td>0.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>…</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sort each column to give $X_{\text{sort}}$

Take means across rows of $X_{\text{sort}}$ and assign this mean to each element in the row to get $X'_{\text{sort}}$

Get $X_{\text{normalized}}$ by arranging each column of $X'_{\text{sort}}$ to have same ordering as $X$. 

Gene 1, 2, …, p
Exercise


- Arrays 1 to 3, genes A to D

<table>
<thead>
<tr>
<th></th>
<th>Array 1</th>
<th>Array 2</th>
<th>Array 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>D</td>
<td>4</td>
<td>2</td>
<td>8</td>
</tr>
</tbody>
</table>

How to perform quantile normalization?
Rank->Average-> Replace (same order)
Sometimes, a gene expression study may involve batches of data collected over a long period of time...

After quantile normalization

In such a case, batch effect may be severe... to the extent that you can predict the batch that each sample comes!

⇒ Need normalization to correct for batch effect

Figure 3.6: GEPs after the batch effects removing.
References


Quantile Normalization in Statistics

- QN is a technique for making two distributions identical in statistical properties.
- To quantile normalize two or more distributions to each other, we sort, then set to the average of the distributions.
- The highest value in all cases becomes the mean of the highest values; the second highest value becomes the mean of the second highest values, and so on.
- Quantile normalization is frequently used in microarray data analysis.
Quantile Normalization (rank array)

- Arrays 1 to 3, genes A to D
  
<table>
<thead>
<tr>
<th></th>
<th>Array 1</th>
<th>Array 2</th>
<th>Array 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>D</td>
<td>4</td>
<td>2</td>
<td>8</td>
</tr>
</tbody>
</table>

- For each column determine a rank from lowest to highest and assign number i-iv
  
<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>iv</td>
<td>iii</td>
</tr>
<tr>
<td>B</td>
<td>i</td>
<td>i</td>
</tr>
<tr>
<td>C</td>
<td>ii</td>
<td>iii</td>
</tr>
<tr>
<td>D</td>
<td>iii</td>
<td>ii</td>
</tr>
</tbody>
</table>

These rank values are set aside to use later. We will convert the ranks into actual values.
Quantile Normalization (average gene’s rank values across array)

- Go back to the first set of data. Rearrange that first set of column values so each column is in order going lowest to highest value. The result is:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>4</td>
<td>2</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

- Now find the mean for each row to determine the ranks

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>5</td>
<td>4</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

- Smallest Values
- Largest Values
Quantile Normalization (explanation)

- Go back to the first set of data. Rearrange that first set of column values so each column is in order going lowest to highest value. The result is:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>5</th>
<th>4</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>4</td>
<td>2</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

- Now find the mean for each row to determine the ranks

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>2</th>
<th>1</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>5</td>
<td>4</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

| A (2 1 3 )/3 = 2.00 = rank i | Average of the smallest |
| B (3 2 4 )/3 = 3.00 = rank ii | Average of the second smallest |
| C (4 4 6 )/3 = 4.67 = rank iii | Average of the second largest |
| D (5 4 8 )/3 = 5.67 = rank iv | Average of the largest |
Quantile Normalization (Replace)

2.00 = rank i, 3.00 = rank ii, 4.67 = rank iii, 5.67 = rank iv

- Now take the ranking order and substitute in new values

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>iv</td>
<td>5.67</td>
<td>2.00</td>
<td>3.00</td>
<td>4.67</td>
</tr>
<tr>
<td>i</td>
<td>4.67</td>
<td>2.00</td>
<td>4.67</td>
<td>3.00</td>
</tr>
<tr>
<td>iii</td>
<td>2.00</td>
<td>3.00</td>
<td>5.67</td>
<td>4.67</td>
</tr>
<tr>
<td>i</td>
<td>3.00</td>
<td>4.67</td>
<td>5.67</td>
<td>4.67</td>
</tr>
<tr>
<td>ii</td>
<td>5.67</td>
<td>4.67</td>
<td>5.67</td>
<td>4.67</td>
</tr>
<tr>
<td>iii</td>
<td>4.67</td>
<td>3.00</td>
<td>5.67</td>
<td>5.67</td>
</tr>
<tr>
<td>iv</td>
<td>5.67</td>
<td>4.67</td>
<td>5.67</td>
<td>5.67</td>
</tr>
</tbody>
</table>

Original Data

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4</td>
<td>8</td>
<td>3</td>
</tr>
</tbody>
</table>
Compute Similarities/Distances Between Two Gene’s Expression Profiles
Cosine Similarity

• If \( g_1 \) and \( g_2 \) are two gene profile vectors, then

\[
\cos( g_1, g_2 ) = \frac{(g_1 \cdot g_2)}{||g_1|| \cdot ||g_2||},
\]

where \( \cdot \) indicates vector dot product and \( || \ g || \) is the length of vector \( g \).

• It is a measure of the cosine of the angle between the two vectors.

• Example:

\[
g_1 = 3 \ 2 \ 0 \ 5 \ 0 \ 0 \ 0 \ 2 \ 0 \ 0 \\
g_2 = 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 2
\]

\[
g_1 \cdot g_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5
\]

\[
||g_1|| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.4807
\]

\[
||g_2|| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.4495
\]

\[
\cos( g_1, g_2 ) = \frac{5}{(6.4807*2.4495)} = 0.3150
\]
Pearson correlation coefficient

• In statistics, the Pearson correlation coefficient (typically denoted by $r$) is a measure of the correlation (linear dependence) between two variables $X$ and $Y$.

• The values of $r$ are between -1 and +1 inclusive.

• It is widely used in the sciences as a measure of the strength of linear dependence between two variables.

• In our case, variables are genes, we measure the correlation between their expression profiles.
Example

• $X = (X_1, X_2, X_3) = (0.03, 0.08, 1.83)$
• $Y = (Y_1, Y_2, Y_3) = (0.01, 0.09, 2.12)$
• $Z = (Z_1, Z_2, Z_3) = (2.51, 0.10, 0.01)$

• $r(X, Y) =$?
• $r(X, Z) =$?

X, Y, Z could be very high dimension vectors!!!
Formula - Pearson's correlation coefficient

- Pearson's correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations:

$$r = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y},$$

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

$$r = \frac{N \sum XY - (\sum X)(\sum Y)}{\sqrt{N \sum X^2 - (\sum X)^2} \sqrt{N \sum Y^2 - (\sum Y)^2}}.$$
Scatter plots showing the correlation from −1 to 1.

1. Scatter plots illustrating correlations from -1 to 1.
An example to compute Pearson's correlation coefficient

- I will show an example to compute Pearson's correlation coefficient using Excel in Tutorial

- You can replace the numbers in the excel file to check how the values affect the PCC results.
Euclidean Distance

- Euclidean Distance between two n-dimensional vectors (objects) \( p \) and \( q \)

\[
dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}
\]

- where \( p=\{p_1, p_2, p_k, \ldots, p_n\} \), \( q=\{q_1, q_2, q_k, \ldots, q_n\} \). 
  \( n \) is the number of dimensions (attributes) and \( p_k \) and \( q_k \) are the \( k^{th} \) attributes (components) of data objects \( p \) and \( q \), respectively.
Euclidean Distance in 2D

Example:

<table>
<thead>
<tr>
<th>point</th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>p2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>p3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>p4</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

Euclidean Distance Matrix

<table>
<thead>
<tr>
<th></th>
<th>p1</th>
<th>p2</th>
<th>p3</th>
<th>p4</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td>0.0</td>
<td>2.828</td>
<td>3.162</td>
<td>5.099</td>
</tr>
<tr>
<td>p2</td>
<td>2.828</td>
<td>0.0</td>
<td>1.414</td>
<td>3.162</td>
</tr>
<tr>
<td>p3</td>
<td>3.162</td>
<td>1.414</td>
<td>0.0</td>
<td>2.0</td>
</tr>
<tr>
<td>p4</td>
<td>5.099</td>
<td>3.162</td>
<td>2.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Euclidean distance with feature importance

Given two vectors: $p = \{p_1, p_2, p_k, \ldots, p_n\}$, $q = \{q_1, q_2, q_k, \ldots, q_n\}$.

- May not want to treat all attributes the same.
- We could use weights $w_k$ to indicate the importance for each feature
- $w_k$ will be between 0 and 1 and

$$\sum_{k=1}^{n} w_k = 1$$

$$dist = \sqrt{\sum_{k=1}^{n} w_k (p_k - q_k)^2}$$
Gene Expression Profile Classification

• Diagnosis of Childhood Acute Lymphoblastic Leukemia and Optimization of Risk-Benefit Ratio of Therapy
Childhood ALL

- 6 Major subtypes: T-ALL, E2A-PBX, TEL-AML, BCR-ABL, MLL genome rearrangements, Hyperdiploid>50
- Diff subtypes respond differently to same Tx
- Over-intensive Tx
  - Development of secondary cancers
  - Reduction of IQ
- Under-intensive Tx
  - Relapse: suffer deterioration after a period of improvement.
- The subtypes look similar
- Conventional diagnosis
  - Immunophenotyping
  - Cytogenetics
  - Molecular diagnostics
- Unavailable in most ASEAN countries
Mission

• Conventional risk assignment procedure requires difficult expensive tests and collective judgement of *multiple specialists*

• Generally available only in major advanced hospitals

⇒ Can we have a single-test easy-to-use platform instead?
Single-Test Platform of Microarray & Machine Learning
For each subtype, select genes to develop classification model for diagnosing that subtype.
Subtype Diagnosis by PCL

- Gene expression data collection
- Classifier training by emerging pattern
- Apply classifier for diagnosis of future cases by PCL
Childhood ALL Subtype Diagnosis Workflow

A tree-structured diagnostic workflow was recommended by Prof Limsoon’s doctor collaborator.
### Training and Testing Sets

<table>
<thead>
<tr>
<th>Paired datasets</th>
<th>Ingredients</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-ALL vs OTHERS1</td>
<td>$\text{OTHERS1} = { \text{E2A-PBX1, TEL-AML1, BCR-ABL, Hyperdip&gt;50, MLL, OTHERS} }$</td>
<td>28 vs 187</td>
<td>15 vs 97</td>
</tr>
<tr>
<td>E2A-PBX1 vs OTHERS2</td>
<td>$\text{OTHERS2} = { \text{TEL-AML1, BCR-ABL, Hyperdip&gt;50, MLL, OTHERS} }$</td>
<td>18 vs 169</td>
<td>9 vs 88</td>
</tr>
<tr>
<td>TEL-AML1 vs OTHERS3</td>
<td>$\text{OTHERS3} = { \text{BCR-ABL, Hyperdip&gt;50, MLL, OTHERS} }$</td>
<td>52 vs 117</td>
<td>27 vs 61</td>
</tr>
<tr>
<td>BCR-ABL vs OTHERS4</td>
<td>$\text{OTHERS4} = { \text{Hyperdip&gt;50, MLL, OTHERS} }$</td>
<td>9 vs 108</td>
<td>6 vs 55</td>
</tr>
<tr>
<td>MLL vs OTHERS5</td>
<td>$\text{OTHERS5} = { \text{Hyperdip&gt;50, OTHERS} }$</td>
<td>14 vs 94</td>
<td>6 vs 49</td>
</tr>
<tr>
<td>Hyperdip&gt;50 vs OTHERS</td>
<td>$\text{OTHERS} = { \text{Hyperdip47-50, Pseudodip, Hypodip, Normo} }$</td>
<td>42 vs 52</td>
<td>22 vs 27</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Training Data</th>
<th>Type1</th>
<th>Type2</th>
<th>Type3</th>
<th>Type4</th>
<th>Type5</th>
<th>Type6</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td># Examples</td>
<td>28</td>
<td>18</td>
<td>52</td>
<td>9</td>
<td>14</td>
<td>42</td>
<td>52</td>
</tr>
<tr>
<td>Negatives</td>
<td>187</td>
<td>169</td>
<td>117</td>
<td>108</td>
<td>94</td>
<td>52</td>
<td></td>
</tr>
</tbody>
</table>
Emerging Patterns

• An emerging pattern is a set of conditions
  – usually involving several features
  – that *most* members of a class satisfy
  – but *none or few of the other class* satisfy

• A jumping emerging pattern (JEP) is an emerging pattern that
  – *some* members of a class satisfy
  – but *no* members of the other class satisfy

• We only study *jumping* emerging patterns
### Examples of JEP

<table>
<thead>
<tr>
<th>Patterns</th>
<th>Frequency (P)</th>
<th>Frequency (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>{9, 36}</td>
<td>38 instances</td>
<td>0</td>
</tr>
<tr>
<td>{9, 23}</td>
<td>38</td>
<td>0</td>
</tr>
<tr>
<td>{4, 9}</td>
<td>38</td>
<td>0</td>
</tr>
<tr>
<td>{9, 14}</td>
<td>38</td>
<td>0</td>
</tr>
<tr>
<td>{6, 9}</td>
<td>38</td>
<td>0</td>
</tr>
<tr>
<td>{7, 21}</td>
<td>0</td>
<td>36</td>
</tr>
<tr>
<td>{7, 11}</td>
<td>0</td>
<td>35</td>
</tr>
<tr>
<td>{7, 43}</td>
<td>0</td>
<td>35</td>
</tr>
<tr>
<td>{7, 39}</td>
<td>0</td>
<td>34</td>
</tr>
<tr>
<td>{24, 29}</td>
<td>0</td>
<td>34</td>
</tr>
</tbody>
</table>

Easy interpretation

Reference number 9: the expression of gene 37720_at > 215
Reference number 36: the expression of gene 38028_at ≤ 12
PCL: Prediction by Collective Likelihood

- Let $EP_1^P, \ldots, EP_i^P$ be the most general EPs of $D^P$ in descending order of support.

- Suppose the test sample $T$ contains these most general EPs of $D^P$ (in descending order of support):

  $$EP_{i_1}^P, EP_{i_2}^P, \ldots, EP_{i_x}^P$$

- Use $k$ top-ranked most general EPs of $D^P$ and $D^N$. Define the score of $T$ in the $D^P$ class as

  $$score(T, D^P) = \sum_{m=1}^{k} \frac{frequency(EP_{i_m}^P)}{frequency(EP_m^P)}$$

- Ditto for $score(T, D^N)$.

- If $score(T, D^P) > score(T, D^N)$, then $T$ is class $P$. Otherwise it is class $N$. 

T contains part of JEPs

Pos support score: example

Neg support score

Copyright 2015 © Wong Limsoon, Li Xiaoli
The idea of summarizing multiple top-ranked EPs is intended to avoid some rare tie cases.
Test example T ($k=3$)

**Top-Ranked EPs in Positive class**

- $\text{EP}_1^P$ (90%)
- $\text{EP}_2^P$ (86%)
- $\text{EP}_3^P$ (85%)
- $\text{EP}_4^P$ (83%)
- $\text{EP}_5^P$ (80%)
- $\text{EP}_6^P$ (79%)
- $\ldots$
- $\text{EP}_n^P$ (68%)

**Top-Ranked EPs in Negative class**

- $\text{EP}_1^N$ (100%)
- $\text{EP}_2^N$ (95%)
- $\text{EP}_3^N$ (92%)
- $\text{EP}_4^N$ (89%)
- $\text{EP}_5^N$ (85%)
- $\text{EP}_6^N$ (80%)
- $\ldots$
- $\text{EP}_n^N$ (80%)

The idea of summarizing multiple top-ranked EPs is intended to avoid some rare tie cases.
PCL Testing (classify a test sample, \(k=3\))

Most freq EP of pos class in the test sample

\[
\text{Score}^P = \frac{EP_1^P'}{EP_1^P} + \cdots + \frac{EP_k^P'}{EP_k^P} = 90/90 + 85/86 + 80/85
\]

Top-k ranked EP of pos class in the test sample

Similarly,

\[
\text{Score}^N = \frac{EP_1^N'}{EP_1^N} + \cdots + \frac{EP_k^N'}{EP_k^N}
\]

If \(\text{Score}^P > \text{Score}^N\), then positive class, Otherwise negative class

If test sample contains more freq positive JEPs and less negative JEPs, then it is a positive sample; otherwise it is a negative sample.
Accuracy of PCL (vs. other classifiers)

<table>
<thead>
<tr>
<th>Testing Data</th>
<th>Error rate of different models</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C4.5</td>
</tr>
<tr>
<td>T-ALL vs OTHERS1</td>
<td>0:1</td>
</tr>
<tr>
<td>E2A-PBX1 vs OTHERS2</td>
<td>0:0</td>
</tr>
<tr>
<td>TEL-AML1 vs OTHERS3</td>
<td>1:1</td>
</tr>
<tr>
<td>BCR-ABL vs OTHERS4</td>
<td>2:0</td>
</tr>
<tr>
<td>MLL vs OTHERS5</td>
<td>0:1</td>
</tr>
<tr>
<td>Hyperdiploid&gt;50 vs OTHERS</td>
<td>2:6</td>
</tr>
<tr>
<td><strong>Total Errors</strong></td>
<td>14</td>
</tr>
</tbody>
</table>

The classifiers are all applied to the 20 genes selected by $\chi^2$ at each level of the tree.

x:y: # errors in positive class vs # errors in negative class
Understandability of PCL

• E.g., for T-ALL vs. OTHERS1, one ideally discriminatory gene 38319_at was found, inducing these 2 EPs

\[ \text{EP}_1 \{ \mathit{gene}_{-38319\_at} @ (-\infty, 15,975.6) \} \]  
\[ \text{EP}_2 \{ \mathit{gene}_{-38319\_at} @ [15,975.6, +\infty) \}. \]

• These give us the diagnostic rule for test example

\[ \{ \mathit{gene}_{-38319\_at} @ (-\infty, 15,975.6) \} \]
\[ \{ \mathit{gene}_{-38319\_at} @ [15,975.6, +\infty) \}. \]

If the expression of 38319_at is less than 15,975.6, then this ALL sample must be a T-ALL. Otherwise it must be a subtype in OTHERS1.
Childhood ALL Cure Rates

- Conventional risk assignment procedure requires difficult expensive tests and collective judgement of multiple specialists

⇒ Not available in less advanced ASEAN countries
Childhood ALL Treatment Cost

- **Treatment for childhood ALL over 2 yrs**
  - Low intensity: US$36k
  - Intermediate intensity: US$60k
  - High intensity: US$72k

- **Treatment for relapse**: US$150k

- **Cost for side-effects**: Unquantified
Current Situation
(2000 new cases/yr in ASEAN)

- **Intermediate intensity conventionally applied in less advanced ASEAN countries**

  Childhood ALL Patients Profile
  - High (10%)
  - Low (50%)
  - Intermediate (40%)

- **Over intensive for 50% of patients, thus more side effects**
  (50% patients are supposed to use Low, but now we use intermediate intensity -> over)

- **Under intensive for 10% of patients, thus more relapse**
  (should use high but use intermediate > under)

Current Cost for these 2000 cases

- **US$120m (US$60k * 2000)** for intermediate intensity tx
- **US$30m (US$150k * 2000 * 10%)** for relapse tx (should use high)
- **Total US$150m/yr plus un-quantified costs for dealing with side effects**

Using Prof Limsoon’s Platform

• Low intensity applied to 50% of patients
• Intermediate intensity to 40% of patients
• High intensity to 10% of patients

⇒ Reduced side effects
⇒ Reduced relapse
⇒ 75-80% cure rates

Total cost for new solution
• US$36m (US$36k * 2000 * 50%) for low intensity
• US$48m (US$60k * 2000 * 40%) for intermediate intensity
• US$14.4m (US$72k * 2000 * 10%) for high intensity

• Total US$98.4m/yr
⇒ Save US$51.6m/yr

Low: US$36k, Intermediate: US$60k,
High: US$72k, relapse: US$150k
A Nice Ending…

- Asian Innovation Gold Award 2003
Gene Interaction Prediction
Beyond Classification of Gene Expression Profiles

- After identifying the candidate genes by feature selection, do we know which ones are *causal* genes and which ones are *surrogates*?
Gene Regulatory Circuits

- Genes are “connected” in “circuit” or network
- Expression of a gene in a network depends on expression of some other genes in the network
- Can we reconstruct the gene network from gene expression data?
Key Questions

• For each gene in the network:
  – Which genes affect it?
  – How they affect it?
Some Techniques

• **Bayesian Networks**
  – Friedman et al., *JCB* 7:601--620, 2000

• **Boolean Networks**
  – Akutsu et al., *PSB* 2000, pages 293--304

• **Differential equations**
  – Chen et al., *PSB* 1999, pages 29--40

• **Classification-based method**
A Classification-Based Technique

Soinov et al., *Genome Biology* 4:R6.1-9, Jan 2003

- Given a gene expression matrix $X$
  - each row is a gene
  - each column is a sample
  - each element $x_{ij}$ is expression of gene $i$ in sample $j$

- Find the average value $a_i$ of each gene $i$

- Denote $s_{ij}$ as state of gene $i$ in sample $j$,
  - $s_{ij} = \text{up}$ if $x_{ij} > a_i$
  - $s_{ij} = \text{down}$ if $x_{ij} \leq a_i$
To see whether the state of gene $g$ is determined by the state of other genes $i$

- see whether $\langle s_{ij} | i \neq g \rangle$ can predict $s_{gj}$ (use other gene’s same sample values to predict current gene’s sample value)
- if can predict with high accuracy, then “yes”
- Any classifier can be used, such as C4.5, PCL, SVM, etc.

To see how the state of gene $g$ is determined by the state of other genes

- apply C4.5 (or PCL or other “rule-based” classifiers) to predict $s_{gj}$ from $\langle s_{ij} | i \neq g \rangle$ (Rules are easy to understand)
- and extract the decision tree or rules used
Simple Introduction of Gene Ontology
Gene Ontology
(GO terms/concepts and relationships)

• **URL:** [http://www.geneontology.org/](http://www.geneontology.org/)
• **Download Ontology**
  – 10/31/2014 06:05PM 3,917,025
    [gene_ontology_edit.obo.2014-11-01.gz](gene_ontology_edit.obo.2014-11-01.gz) (consist of the following three parts; always updated one)
  – component.ontology (namespace: cellular_component)
  – function.ontology (namespace: molecular_function)
  – process.ontology (namespace: biological_process)
Associate Genes with Functions

• How to get a gene/gene product’s function information:
  – 1. Download whole file (for large scale analysis)
    • http://geneontology.org/page/download-annotations

• Saccharomyces cerevisiae

| Saccharomyces cerevisiae | Stanford University | 6381 | 94556 (48665 non-IEA) | 11/1/2014 | README | gene_association.sgd.gz (1 mb) |

1: DB, database contributing the file (always "SGD" for this file). 2: DB_Object_ID, SGDID (SGD's unique identifier for genes and features). 3: DB_Object_Symbol, see below 4: Qualifier (optional), one or more of 'NOT', 'contributes_to', 'colocalizes_with' as qualifier(s) for a GO annotation, when needed, multiples separated by pipe (|) 5: GO ID, unique numeric identifier for the GO term 6: DB:Reference(DB:Reference), the reference associated with the GO annotation 7: Evidence, the evidence code for the GO annotation 8: With (or) From (optional), any With or From qualifier for the GO annotation 9: Aspect, which ontology the GO term belongs (Function, Process or Component) 10: DB_Object_Name(Name) (optional), a name for the gene product in words, e.g. 'acid phosphatase' 11: DB_Object_Synonym(Synonym) (optional), see below 12: DB_Object_Type, type of object annotated, e.g. gene, protein, etc. 13: taxon(taxon), taxonomic identifier of species encoding gene product 14: Date, date GO annotation was defined in the format YYYYMMDD 15: Assigned_by, source of the annotation (always "SGD" for this file)
More detailed description of GO

The Gene Ontology provides a way to capture and represent biological knowledge in a computable form.

GO Slides from
Jennifer Clark, gene ontology consortium editorial office
How does the Gene Ontology work?

• GO isn’t just a flat list of biological terms
• Terms are related within a hierarchy
Relationships between GO terms
Genes’ Function

- all : all (166775)
  - GO:0008150 : biological_process (118690)
  - GO:0009987 : cellular_process (71171)
  - GO:0050875 : cellular_physiological_process (65087)
  - GO:0044237 : cellular_metabolism (41108)
  - GO:0006139 : nucleobase, nucleoside, nucleotide and nucleic_acid_metabolism (16561)
  - GO:0006259 : DNA_metabolism (4671)
  - GO:0006260 : DNA_replication (1115)
Ontology Structure

• Terms are linked by two relationships
  – 1. is-a
  – 2. part-of
Ontology Structure

cell

membrane

mitochondrial membrane

chloroplast

membrane

chloroplast membrane

is-a part-of
Ontology Structure

- Ontologies are structured as a hierarchical directed acyclic graph (DAG) [NO LOOP]
- Terms can have more than one parent and zero, one or more children
Ontology Structure

- cell
  - membrane
    - mitochondrial membrane
  - chloroplast membrane
- Directed Acyclic Graph (DAG) - multiple parentage allowed
How does GO work?

What information might we want to capture about a gene product?

- What does the gene product do?
- Where and when does it act?
- Why does it perform these activities?
GO structure

• **GO terms divided into three parts:**
  – cellular component
  – molecular function
  – biological process

• **What each of the three parts tell us??**
Cellular Component

- Where a gene product acts
Molecular Function

- Activities or “jobs” of a gene product

Glucose-6-phosphate isomerase activity
Molecular Function

insulin binding

insulin receptor activity
Molecular Function

- A gene product may have several functions; a function term refers to a reaction or activity
- Sets of functions make up a biological process
Biological Process
A commonly recognized series of events: cell division

**Preprophase**
- Centriole

**Prophase**
- Mitotic spindle

**Metaphase**
- Chromosomes arranged in equatorial plane, spindle completed, disappearance of nuclear envelope and nucleolus

**Telophase**
- Nuclear restitution, nuclear envelope and nucleolar formation, end of cell division

**Late anaphase**
- Aggregation of chromosomes at the poles, beginning of cell division, initiation of cleavage furrow

**Early anaphase**
- Longitudinal splitting of chromosomes and migration to poles
Biological Process: limb development
Annotation for Genes

Mitochondrial P450

This is a gene product that has already been annotated to all three gene ontologies. It is the Mitochondrial P450 gene product.
Where is it?

GO cellular component term: mitochondrial inner membrane ; GO:0005743
What does it do?

substrate + O₂ = CO₂ + H₂O product

GO molecular function term: monooxygenase activity ; GO:0004497
Which process is this?

**GO biological process term:**

electron transport ; GO:0006118

http://ntri.tamuk.edu/cell/mitochondrion/krebpic.html
References on gene expression data classification

Thank You

Contact: xlli@i2r.a-star.edu.sg if you have questions