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

CS2220: Introduction to Computational Biology
Unit 3: Gene Expression Analysis

Wong Limsoon
Plan

- Microarray background
- Gene expression profile classification
- Gene expression profile clustering
- Normalization
- Extreme sample selection
- Gene regulatory network inference
Background on microarrays
What is a microarray?

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

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

• Measure expression of thousands of genes simultaneously
Affymetrix GeneChip®
quartz is washed to ensure uniform hydroxylation across its surface and to attach linker molecules.

Exercise: What is the other commonly used type of microarray? How is that one different from Affymetrix’s?
Gene expression measurement by Affymetrix GeneChip®

Click to watch an interesting movie explaining the working of microarray
Sample Affymetrix GeneChip® data file (U95A)

<table>
<thead>
<tr>
<th>AFFX-Mural</th>
<th>AFFX-Mural</th>
<th>AFFX-Mural</th>
<th>AFFX-Mural</th>
<th>AFFX-Mural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Negative</td>
<td>Pairs</td>
<td>Avg</td>
<td>Avg Diff</td>
</tr>
<tr>
<td>AFFX-Mural</td>
<td>5 2</td>
<td>19</td>
<td>297.5</td>
<td>A</td>
</tr>
<tr>
<td>AFFX-Mural</td>
<td>3 2</td>
<td>19</td>
<td>554.2</td>
<td>A</td>
</tr>
<tr>
<td>AFFX-Mural</td>
<td>4 2</td>
<td>19</td>
<td>308.6</td>
<td>A</td>
</tr>
<tr>
<td>AFFX-Murf</td>
<td>1 3</td>
<td>19</td>
<td>141</td>
<td>A</td>
</tr>
<tr>
<td>AFFX-BioE</td>
<td>13 1</td>
<td>19</td>
<td>9340.6</td>
<td>P</td>
</tr>
<tr>
<td>AFFX-BioE</td>
<td>15 0</td>
<td>19</td>
<td>12862.4</td>
<td>P</td>
</tr>
<tr>
<td>AFFX-BioE</td>
<td>12 0</td>
<td>19</td>
<td>8716.5</td>
<td>P</td>
</tr>
<tr>
<td>AFFX-BioC</td>
<td>17 0</td>
<td>19</td>
<td>25942.5</td>
<td>P</td>
</tr>
<tr>
<td>AFFX-BioC</td>
<td>16 0</td>
<td>20</td>
<td>28838.5</td>
<td>P</td>
</tr>
<tr>
<td>AFFX-BioD</td>
<td>17 0</td>
<td>19</td>
<td>25765.2</td>
<td>P</td>
</tr>
<tr>
<td>AFFX-BioD</td>
<td>19 0</td>
<td>20</td>
<td>140113.2</td>
<td>P</td>
</tr>
<tr>
<td>AFFX-CreX</td>
<td>20 0</td>
<td>20</td>
<td>280036.6</td>
<td>P</td>
</tr>
<tr>
<td>AFFX-CreX</td>
<td>20 0</td>
<td>20</td>
<td>401741.8</td>
<td>P</td>
</tr>
<tr>
<td>AFFX-BioE</td>
<td>7 5</td>
<td>18</td>
<td>-483</td>
<td>A</td>
</tr>
<tr>
<td>AFFX-BioE</td>
<td>5 4</td>
<td>18</td>
<td>313.7</td>
<td>A</td>
</tr>
<tr>
<td>AFFX-BioE</td>
<td>7 6</td>
<td>20</td>
<td>-1016.2</td>
<td>A</td>
</tr>
</tbody>
</table>
Some advice on processing Affymetrix GeneChip® data

• Ignore AFFX genes
  – These genes are control genes

• Ignore genes with “Abs Call” equal to “A” or “M”
  – Measurement quality is suspect

• Upperbound 40000, lowerbound 100
  – Saturation of laser scanner

• Deal with missing values

Exercise: Suggest 2 ways to deal with missing value
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>
<th>.....</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>
<td>-0.12</td>
</tr>
<tr>
<td>Sample2</td>
<td>Cancer</td>
<td>1.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<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>
<td></td>
</tr>
</tbody>
</table>

100-500 rows

1000 - 100,000 columns

- **Gene-Time**

- **Gene-Sample-Time**
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>
<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>
</tr>
<tr>
<td>Sample2</td>
<td>Cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>SampleN</td>
<td>~Cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Gene-Time**

- **Gene-Sample-Time**
Application: Disease subtype diagnosis

samples

genes

benign
benign
benign
malign
malign
malign
malign

???
Application: Treatment prognosis

genes

samples

R
R
R
R
NR
NR
NR
NR

???
Type of gene expression datasets

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

<table>
<thead>
<tr>
<th>Gene1</th>
<th>Gene2</th>
<th>Gene3</th>
<th>Gene 4</th>
<th>Gene5</th>
<th>Gene6</th>
<th>Gene7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cond1</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>Cond2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.3</td>
</tr>
<tr>
<td>CondN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

100-500 rows

- **Gene-Time**

- **Gene-Sample-Time**

Application: Drug-action detection

• Which group of genes does the drug affect? Why?

Exercise #1
Gene expression profile classification

Childhood acute lymphoblastic leukemia subtype diagnosis
Childhood ALL

- 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

- 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

<table>
<thead>
<tr>
<th>Description</th>
<th>Positive</th>
<th>Negative</th>
<th>Pairs</th>
<th>InAver</th>
<th>Diff</th>
<th>Abs Call</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFFX-MurF</td>
<td>5</td>
<td>2</td>
<td>19</td>
<td>244.1</td>
<td>A</td>
<td>M16702</td>
</tr>
<tr>
<td>AFFX-MurF</td>
<td>3</td>
<td>2</td>
<td>19</td>
<td>62.4</td>
<td>A</td>
<td>M37367</td>
</tr>
<tr>
<td>AFFX-MurF</td>
<td>4</td>
<td>2</td>
<td>19</td>
<td>268.6</td>
<td>A</td>
<td>M38582</td>
</tr>
<tr>
<td>AFFX-MurF</td>
<td>1</td>
<td>3</td>
<td>19</td>
<td>14.1</td>
<td>A</td>
<td>M83669</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</td>
</tr>
<tr>
<td>AFFX-BioE</td>
<td>15</td>
<td>0</td>
<td>19</td>
<td>12662.4</td>
<td>P</td>
<td>J04423</td>
</tr>
<tr>
<td>AFFX-BioE</td>
<td>12</td>
<td>0</td>
<td>19</td>
<td>2716.5</td>
<td>P</td>
<td>J04423</td>
</tr>
<tr>
<td>AFFX-BioC</td>
<td>17</td>
<td>0</td>
<td>19</td>
<td>2545.5</td>
<td>P</td>
<td>J04423</td>
</tr>
<tr>
<td>AFFX-BioC</td>
<td>16</td>
<td>0</td>
<td>20</td>
<td>2535.5</td>
<td>P</td>
<td>J04423</td>
</tr>
<tr>
<td>AFFX-BioD</td>
<td>17</td>
<td>0</td>
<td>19</td>
<td>2576.2</td>
<td>P</td>
<td>J04423</td>
</tr>
<tr>
<td>AFFX-BioE</td>
<td>19</td>
<td>0</td>
<td>20</td>
<td>140113.2</td>
<td>P</td>
<td>J04423</td>
</tr>
<tr>
<td>AFFX-CreX</td>
<td>20</td>
<td>0</td>
<td>20</td>
<td>230036.6</td>
<td>P</td>
<td>X03452</td>
</tr>
<tr>
<td>AFFX-CreX</td>
<td>20</td>
<td>0</td>
<td>20</td>
<td>201741.3</td>
<td>P</td>
<td>X03452</td>
</tr>
<tr>
<td>AFFX-BioE</td>
<td>5</td>
<td>4</td>
<td>18</td>
<td>513.7</td>
<td>A</td>
<td>J04423</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</td>
</tr>
</tbody>
</table>
Overall strategy

- For each subtype, select genes to develop classification model for diagnosing that subtype
- For each subtype, select genes to develop prediction model for prognosis of that subtype
- For subtype-dependent prognosis
- Risk-stratified treatment intensity
Subtype diagnosis by PCL

- Gene expression data collection
- Gene selection by $\chi^2$
- Classifier training by emerging pattern
- Classifier tuning (optional for some machine learning methods)
- Apply classifier for diagnosis of future cases by PCL
Childhood ALL subtype diagnosis workflow

A tree-structured diagnostic workflow was recommended by our 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>OTHERS1 = {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>OTHERS2 = {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>OTHERS3 = {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>OTHERS4 = {Hyperdip&gt;50, MLL, OTHERS}</td>
<td>9 vs 108</td>
<td>6 vs 55</td>
</tr>
<tr>
<td>MLL vs OTHERS5</td>
<td>OTHERS5 = {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>OTHERS = {Hyperdip47-50, Pseudodip, Hypodip, Normo}</td>
<td>42 vs 52</td>
<td>22 vs 27</td>
</tr>
</tbody>
</table>
Signal selection basic idea

- Choose a signal with low intra-class distance
- Choose a signal with high inter-class distance
Signal selection by $\chi^2$

The $\chi^2$ value of a signal is defined as:

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}},$$

where $m$ is the number of intervals, $k$ the number of classes, $A_{ij}$ the number of samples in the $i$th interval, $j$th class, $R_i$ the number of samples in the $i$th interval, $C_j$ the number of samples in the $j$th class, $N$ the total number of samples, and $E_{ij}$ the expected frequency of $A_{ij}$ ($E_{ij} = R_i \times C_j / N$).
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 is an emerging pattern that
  – some members of a class satisfy
  – but no members of the other class satisfy

• We use only jumping emerging patterns
# Examples

<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>

Reference number 9: the expression of gene 37720_at > 215
Reference number 36: the expression of gene 38028_at ≤ 12

Easy interpretation
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 $N$. 
PCL learning

Top-Ranked EPs in Positive class

EP_1^P (90%)
EP_2^P (86%)
. .
EP_n^P (68%)

Top-Ranked EPs in Negative class

EP_1^N (100%)
EP_2^N (95%)
. .
EP_n^N (80%)

The idea of summarizing multiple top-ranked EPs is intended to avoid some rare tie cases.
PCL testing

Most freq EP of pos class in the test sample

\[ \text{Score}^P = \frac{EP_1^P'}{EP_1^P} + \ldots + \frac{EP_k^P'}{EP_k^P} \]

Most freq EP of pos class

Similarly,

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

If \( \text{Score}^P > \text{Score}^N \), then positive class,
Otherwise negative class
### Accuracy of PCL (vs. other classifiers)

<table>
<thead>
<tr>
<th>Testing Data</th>
<th>C4.5</th>
<th>SVM</th>
<th>NB</th>
<th>PCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-ALL vs OTHERS1</td>
<td>0:1</td>
<td>0:0</td>
<td>0:0</td>
<td>0:0</td>
</tr>
<tr>
<td>E2A-PBX1 vs OTHERS2</td>
<td>0:0</td>
<td>0:0</td>
<td>0:0</td>
<td>0:0</td>
</tr>
<tr>
<td>TEL-AML1 vs OTHERS3</td>
<td>1:1</td>
<td>0:1</td>
<td>0:1</td>
<td>1:0</td>
</tr>
<tr>
<td>BCR-ABL vs OTHERS4</td>
<td>2:0</td>
<td>3:0</td>
<td>1:4</td>
<td>2:0</td>
</tr>
<tr>
<td>MLL vs OTHERS5</td>
<td>0:1</td>
<td>0:0</td>
<td>0:0</td>
<td>0:0</td>
</tr>
<tr>
<td>Hyperdiploid&gt;50 vs OTHERS</td>
<td>2:6</td>
<td>0:2</td>
<td>0:2</td>
<td>0:1</td>
</tr>
<tr>
<td><strong>Total Errors</strong></td>
<td>14</td>
<td>6</td>
<td>8</td>
<td>4</td>
</tr>
</tbody>
</table>

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

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

\[
\{gene_{-(38319\_at)} @ (-\infty, 15975.6)\} \text{ and } \{gene_{-(38319\_at)} @ [15975.6, +\infty)\}.\]

- These give us the diagnostic rule

If the expression of 38319_at is less than 15975.6, then this ALL sample must be a T-ALL. Otherwise it must be a subtype in OTHERS1.
Multidimensional scaling plot for subtype diagnosis

Obtained by performing PCA on the 20 genes chosen for each level
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
  – Intermediate intensity: US$60k
  – Low intensity: US$36k
  – 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

- Over intensive for 50% of patients, thus more side effects
- Under intensive for 10% of patients, thus more relapse
- US$120m (US$60k * 2000) for intermediate intensity tx
- US$30m (US$150k * 2000 * 10%) for relapse tx
- Total US$150m/yr plus un-quantified costs for dealing with side effects
Using our 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

- 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
A nice ending…

- Asian Innovation Gold Award 2003
Gene expression profile clustering

Novel disease subtype discovery
Is there a new subtype?

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

Exercise: Name and describe one bi-clustering method.
Hierarchical clustering

- Assign each item to its own cluster
  - If there are N items initially, we get N clusters, each containing just one item

- Find the “most similar” pair of clusters, merge them into a single cluster, so we now have one less cluster

- Repeat previous step until all items are clustered into a single cluster of size N

More about this in a moment
Gene expression profile clustering

Diagnosis via guilt-by-association
Some patient samples

Mr. A: "???"

• Does Mr. A have cancer?
Let’s rearrange the rows…

<table>
<thead>
<tr>
<th>genes</th>
<th>samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>benign</td>
<td>benign</td>
</tr>
<tr>
<td>benign</td>
<td>malign</td>
</tr>
<tr>
<td>malign</td>
<td>malign</td>
</tr>
<tr>
<td>malign</td>
<td>malign</td>
</tr>
<tr>
<td>malign</td>
<td>malign</td>
</tr>
</tbody>
</table>

Mr. A:  

• Does Mr. A have cancer?
and the columns too…

Mr. A:

• Does Mr. A have cancer?
Introduction to simple clustering methods
What is cluster analysis?

- Finding groups of objects such that objects in a group are similar to one another and different from objects in other groups.

Intra-cluster distances are minimized

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

How many clusters?

Six Clusters

Two Clusters

Four Clusters
We can also have
K-means clustering

- Partitional clustering approach
- Each cluster is associated with a centroid
- Each point is assigned to the cluster with the closest centroid
- # of clusters, K, must be specified

---

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
K-means clustering illustration

Iteration 1

Iteration 2

Iteration 3

Iteration 4

Iteration 5

Iteration 6
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 hierarchical clustering

- More popular hierarchical clustering technique

- Basic algorithm
  
  Compute the proximity matrix
  Let each data point be a cluster
  Repeat
  Merge the two closest clusters
  Update the proximity matrix
  Until only a single cluster remains

- Key is computation of proximity of two clusters
  - Different approaches to defining the distance / similarity between clusters
Visualization of agglomerative hierarchical clustering

Traditional Hierarchical Clustering

Traditional Dendrogram
Single, complete, & average Linkage

**Single linkage** defines distance between two clusters as the minimum distance between them.

$$d(r, s) = \min \{ \text{dist}(x_{rj}, x_{sj}) \}$$

**Complete linkage** defines distance between two clusters as the maximum distance between them.

$$d(r, s) = \max \{ \text{dist}(x_{rj}, x_{sj}) \}$$

**Exercise:** Give definition of “average linkage”

Image source: UCL Microcore Website

Exercise #2
Simulation: Starting situation

- Start with clusters of individual points and a proximity matrix
Intermediate situation

- After some merging steps, we have some clusters

Proximity Matrix

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Intermediate situation

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

```
   C1   C2  C3  C4  C5
C1
C2
C3
C4
C5
```

Proximity Matrix
After merging

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

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C2</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>

Proximity Matrix
How to define inter-cluster similarity

- Min
- Max
- Group average
- Distance between centroids

Similarity?

Proximity Matrix
How to define inter-cluster similarity

- Min
- Max
- Group average
- Distance 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 between centroids

Proximity Matrix

\[
\begin{array}{ccccc}
\text{p1} & \text{p2} & \text{p3} & \text{p4} & \text{p5} & \ldots \\
p1 & & & & & \\
p2 & & & & & \\
p3 & & & & & \\
p4 & & & & & \\
p5 & & & & & \\
\vdots & & & & & \\
\end{array}
\]
How to define inter-cluster similarity

- Min
- Max
- Group average
- Distance 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>


How to define inter-cluster similarity:

- Min
- Max
- Group average
- Distance 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>
<tr>
<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>
</tr>
</tbody>
</table>

Proximity Matrix
Cluster similarity: Min / single linkage

- 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-linkage clustering

Single-linkage dendrogram
Food for thought

• What are the key strengths of single-linkage clustering?

• What are the key weaknesses of single-linkage clustering?
Cluster similarity: Max / 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

We still want to merge two most similar clusters each time. But we define the distance between clusters based on MAX.
Food for thought

• What are the key strengths of complete-linkage clustering?

• What are the key weaknesses of complete-linkage clustering?
Cluster similarity: Group average

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

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

<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 linkage

• Strengths
  – Less susceptible to noise and outliers

• Limitations
  – Biased towards globular clusters
Hierarchical clustering: Comparison

Min

Max

Group average
Food for thought

• What are the space and time complexity of hierarchical clustering?
Normalization
Sometimes, a gene expression study may involve batches of data collected over a long period of time…

Time Span of Gene Expression Profiles

Image credit: Dong Difeng
In such a case, batch effect may be severe… to the extent that you can predict the batch that each sample comes!

\[ \Rightarrow \text{Need normalization to correct for batch effect} \]
Normalization approaches

- **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., \( Z = \frac{x - \mu}{\sigma} \)

- **Quantile normalization**
Quantile 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 elem 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
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!

\[\text{Need normalization to correct for batch effect}\]

After quantile normalization

Figure 3.6: GEPs after the batch effects removing.
Food for thought

• Given a cancer vs normal dataset

• Should you apply quantile normalization to the dataset as a whole or should you apply quantile normalization to the cancer and the normal part separately? Why?
Food for thought

• Given a cancer vs normal dataset

• Should you apply Z-normalization to each phenotype separately or to the whole dataset in one go?

• Should you apply Z-normalization in a patient-wise or gene-wise manner? Why?
Selection of patient samples and genes for disease prognosis
Gene expression profile + clinical data ⇒ outcome prediction

- Univariate & multivariate Cox survival analysis (Beer et al 2002, Rosenwald et al 2002)
- Fuzzy neural network (Ando et al 2002)
- Partial least squares regression (Park et al 2002)
- Weighted voting algorithm (Shipp et al 2002)
- Gene index and “reference gene” (LeBlanc et al 2003)
- ......

Our approach

1. **Step 1: select training samples**
   - Training samples: long-term and short-term survivors
   - Test and evaluate
     - Assign risk score and risk group to each sample
     - Draw Kaplan-Meier curves

2. **Step 2: identify genes**
   - Genes related to survival

3. **Step 3: build SVM scoring function and form risk groups**

**“extreme” sample selection**

ERCOF
Extreme sample selection

Short-term Survivors v.s. Long-term Survivors

**Short-term survivors** who died within a *short* period

\[ F(T) < c_1 \text{ and } E(T) = 1 \]

**Long-term survivors** who were alive after a *long* follow-up time

\[ F(T) > c_2 \]

\[ T: \text{ sample} \]
\[ F(T): \text{ follow-up time} \]
\[ E(T): \text{ status (} 1: \text{ unfavorable}; \ 0: \text{ favorable}) \]
\[ c_1 \text{ and } c_2: \text{ thresholds of survival time} \]
Remove genes with expression values w/o cut point found (can’t be discretized)

Calculate Wilcoxon rank sum $w(x)$ for gene $x$. Remove gene $x$ if $w(x) \in [c_{lower}, c_{upper}]$

Group features by Pearson Correlation. For each group, retain the top 50% wrt class entropy.

**ERCOF**

Entropy-Based Rank Sum Test & Correlation Filtering
Risk score construction

Linear Kernel SVM regression function

\[ G(T) = \sum_i a_i y_i K(T, x(i)) + b \]

\( T \): test sample, \( x(i) \): support vector,
\( y_i \): class label (1: short-term survivors; -1: long-term survivors)

Transformation function (posterior probability)

\[ S(T) = \frac{1}{1 + e^{-G(T)}} \quad (S(T) \in (0,1)) \]

\( S(T) \): risk score of sample \( T \)
Diffuse large B-cell lymphoma

• DLBC lymphoma is the most common type of lymphoma in adults

• Can be cured by anthracycline-based chemotherapy in 35 to 40 percent of patients

⇒ DLBC lymphoma comprises several diseases that differ in responsiveness to chemotherapy

• Intl Prognostic Index (IPI)
  – age, “Eastern Cooperative Oncology Group” Performance status, tumor stage, lactate dehydrogenase level, sites of extranodal disease, ...

• Not very good for stratifying DLBC lymphoma patients for therapeutic trials

⇒ Use gene-expression profiles to predict outcome of chemotherapy?
Rosenwald et al., *NEJM* 2002

- **240 data samples**
  - 160 in preliminary group
  - 80 in validation group
  - each sample described by 7399 microarray features

- **Rosenwald et al.’s approach**
  - identify gene: Cox proportional-hazards model
  - cluster identified genes into four gene signatures
  - calculate for each sample an outcome-predictor score
  - divide patients into quartiles according to score
Knowledge discovery from gene expression of “extreme” samples

“extreme” sample selection: < 1 yr vs > 8 yrs

knowledge discovery from gene expression

T is long-term if $S(T) < 0.3$
T is short-term if $S(T) > 0.7$
## Discussions: Sample selection

<table>
<thead>
<tr>
<th>Application</th>
<th>Data set</th>
<th>Status</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Dead</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Alive</td>
<td></td>
</tr>
<tr>
<td>DLBCL</td>
<td>Original</td>
<td>88</td>
<td>160</td>
</tr>
<tr>
<td></td>
<td>Informative</td>
<td>47+1(*)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>72</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>25</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>73</td>
<td></td>
</tr>
</tbody>
</table>

Number of samples in original data and selected informative training set. 
(*): Number of samples whose corresponding patient was dead at the end of follow-up time, but selected as a long-term survivor.
## Discussions: Gene identification

<table>
<thead>
<tr>
<th>Gene selection</th>
<th>DLBCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>4937(*)</td>
</tr>
<tr>
<td>Phase I</td>
<td>132(2.7%)</td>
</tr>
<tr>
<td>Phase II</td>
<td>84(1.7%)</td>
</tr>
</tbody>
</table>

Number of genes left after feature filtering for each phase. (*): number of genes after removing those genes who were absent in more than 10% of the experiments.
Kaplan-Meier plot for 80 test cases

\[ p\text{-value of log-rank test: } < 0.0001 \]

Risk score thresholds: 0.7, 0.3
Improvement over IPI

(A) IPI low,  
\[ p-value = 0.0063 \]

(B) IPI intermediate,  
\[ p-value = 0.0003 \]
Merit of “extreme” samples

(A) W/o sample selection ($p = 0.38$)

(B) With sample selection ($p=0.009$)

No clear difference on the overall survival of the 80 samples in the validation group of DLBCL study, if no training sample selection conducted
About the inventor: Huiqing Liu

- **Huiqing Liu**
  - PhD, NUS, 2004
  - Currently PI at Incyte
  - Asian Innovation Gold Award 2003
  - New Jersey Cancer Research Award for Scientific Excellence 2008
  - Gallo Prize 2008
Beyond disease diagnosis & prognosis
Beyond classification of gene expression profiles

- After identifying the candidate genes by feature selection, do we know which ones are causal genes, which ones are surrogates, and which are noise?
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 and other data?

Source: Miltenyi Biotec
Key questions

For each gene in the network:

• Which genes affect it?

• How they affect it?
  – Positively?
  – Negatively?
  – More complicated ways?
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, 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 $x_{ij} \leq a_i$
A classification-based technique
Soinov et al., *Genome Biology* 4:R6.1-9, Jan 2003

- **To see whether the state of gene g is determined by the state of other genes**
  - See whether $\langle s_{ij} \mid i \neq g \rangle$ can predict $s_{gj}$
  - 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 $\langle s_{ij} \mid i \neq g \rangle$
  - Extract the decision tree or rules used
Advantages of this method

- Can identify genes affecting a target gene
- Don’t need discretization thresholds?
- Each data sample is treated as an example
- Explicit rules can be extracted from the classifier (assuming C4.5 or PCL)
- Generalizable to time series

- Discuss the point “Don’t need discretization thresholds”. Is it true?
Concluding remarks
Bcr-Abl

• Targeted drug dev
  – Know what molecular effect you want to achieve
    • E.g., inhibit a mutated form of a protein
  – Engineer a compound that directly binds and causes the desired effect

• Gleevec (imatinib)
  – 1st success for real drug
  – Targets Bcr-Abl fusion protein (i.e., Philadelphia chromosome, Ph)
  – NCI summary of clinical trial of imatinib for ALL at
    http://www.cancer.gov/clincaltrials/results/ALLimatinib1109/print
What have we learned?

• Technologies
  – Microarray
  – PCL, ERCOF

• Microarray applications
  – Disease diagnosis by supervised learning
  – Subtype discovery by unsupervised learning
  – Disease diagnosis via guilt-by-association
  – Gene network reconstruction

• Important tactic
  – Extreme sample selection
Useful packages

• **EXPANDER (EXPression Analyser & DisplayER)**
  - [http://acgt.cs.tau.ac.il/expander](http://acgt.cs.tau.ac.il/expander)

• **BRB-Array Tools**

• **NetProt**
  - [http://rpubs.com/gohwils/204259](http://rpubs.com/gohwils/204259)
  - [https://github.com/gohwils/NetProt/releases/](https://github.com/gohwils/NetProt/releases/)

Any question?
References