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

National University of Singapore
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>Description</th>
<th>Positive</th>
<th>Negative</th>
<th>Pairs In Avg</th>
<th>Avg Diff</th>
<th>Abs Call</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFFX-Murl Mouse interleukin 2 (IL-2) gene, exon 4</td>
<td>5</td>
<td>2</td>
<td>19</td>
<td>297.5</td>
<td>A</td>
</tr>
<tr>
<td>AFFX-Murl Mouse interleukin 10 mRNA, complete cds</td>
<td>3</td>
<td>2</td>
<td>19</td>
<td>554.2</td>
<td>A</td>
</tr>
<tr>
<td>AFFX-Murl Mus musculus interleukin 4 (II-4) mRNA, complete cds</td>
<td>4</td>
<td>2</td>
<td>19</td>
<td>308.6</td>
<td>A</td>
</tr>
<tr>
<td>AFFX-Murf Mus musculus Fas antigen mRNA, complete cds</td>
<td>1</td>
<td>3</td>
<td>19</td>
<td>141</td>
<td>A</td>
</tr>
<tr>
<td>AFFX-BioE E coli bioB gene biotin synthetase (-5, -M, -3 r)</td>
<td>13</td>
<td>1</td>
<td>19</td>
<td>9340.6</td>
<td>P</td>
</tr>
<tr>
<td>AFFX-BioE E coli bioB gene biotin synthetase (-5, -M, -3 r)</td>
<td>15</td>
<td>0</td>
<td>19</td>
<td>12862.4</td>
<td>P</td>
</tr>
<tr>
<td>AFFX-BioE E coli bioB gene biotin synthetase (-5, -M, -3 r)</td>
<td>12</td>
<td>0</td>
<td>19</td>
<td>8716.5</td>
<td>P</td>
</tr>
<tr>
<td>AFFX-BioC E coli bioC protein (-5 and -3 represent transcr</td>
<td>17</td>
<td>0</td>
<td>19</td>
<td>25942.5</td>
<td>P</td>
</tr>
<tr>
<td>AFFX-BioC E coli bioC protein (-5 and -3 represent transcr</td>
<td>16</td>
<td>0</td>
<td>20</td>
<td>28838.5</td>
<td>P</td>
</tr>
<tr>
<td>AFFX-BioD E coli bioD gene dethiobiotin synthetase (-5 ar</td>
<td>17</td>
<td>0</td>
<td>19</td>
<td>25765.2</td>
<td>P</td>
</tr>
<tr>
<td>AFFX-BioD E coli bioD gene dethiobiotin synthetase (-5 ar</td>
<td>19</td>
<td>0</td>
<td>20</td>
<td>140113.2</td>
<td>P</td>
</tr>
<tr>
<td>AFFX-CreX Bacteriophage P1 cre recombinase protein (-5</td>
<td>20</td>
<td>0</td>
<td>20</td>
<td>280036.6</td>
<td>P</td>
</tr>
<tr>
<td>AFFX-CreX Bacteriophage P1 cre recombinase protein (-5</td>
<td>20</td>
<td>0</td>
<td>20</td>
<td>401741.8</td>
<td>P</td>
</tr>
<tr>
<td>AFFX-BioE E coli bioB gene biotin synthetase (-5, -M, -3 r</td>
<td>7</td>
<td>5</td>
<td>18</td>
<td>-483</td>
<td>A</td>
</tr>
<tr>
<td>AFFX-BioE E coli bioB gene biotin synthetase (-5, -M, -3 r</td>
<td>5</td>
<td>4</td>
<td>18</td>
<td>313.7</td>
<td>A</td>
</tr>
<tr>
<td>AFFX-BioE E coli bioB gene biotin synthetase (-5, -M, -3 r</td>
<td>7</td>
<td>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></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>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>
<td></td>
</tr>
<tr>
<td>Sample2</td>
<td>Cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.3</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**
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>~Cancer</td>
<td></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>

- **Gene-Time**

- **Gene-Sample-Time**

---

Copyright 2021 © Wong Limsoon
Application: Disease subtype diagnosis

**genes**

**samples**

- **benign**
- **malign**

**???**
Application: Treatment prognosis

genes

samples

R
R
R
NR
NR
NR
NR

???
Type of gene expression datasets

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

  1000 - 100,000 columns

<table>
<thead>
<tr>
<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>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>

- **Gene-Time**

- **Gene-Sample-Time**
Application: Drug-action detection

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

Diagnosis of subtype

Subtype-dependent prognosis

Risk-stratified treatment intensity

• 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
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>$\text{OTHERS}_1 = {\text{E2A-PBX1, TEL-AML1, BCR-ABL, Hyperdip}&gt;50, \text{MLL, OTHERS}}$</td>
<td>28 vs 187</td>
<td>15 vs 97</td>
</tr>
<tr>
<td>E2A-PBX1 vs OTHERS2</td>
<td>$\text{OTHERS}_2 = {\text{TEL-AML1, BCR-ABL Hyperdip}&gt;50, \text{MLL, OTHERS}}$</td>
<td>18 vs 169</td>
<td>9 vs 88</td>
</tr>
<tr>
<td>TEL-AML1 vs OTHERS3</td>
<td>$\text{OTHERS}_3 = {\text{BCR-ABL Hyperdip}&gt;50, \text{MLL, OTHERS}}$</td>
<td>52 vs 117</td>
<td>27 vs 61</td>
</tr>
<tr>
<td>BCR-ABL vs OTHERS4</td>
<td>$\text{OTHERS}_4 = {\text{Hyperdip}&gt;50, \text{MLL, OTHERS}}$</td>
<td>9 vs 108</td>
<td>6 vs 55</td>
</tr>
<tr>
<td>MLL vs OTHERS5</td>
<td>$\text{OTHERS}_5 = {\text{Hyperdip}&gt;50, \text{OTHERS}}$</td>
<td>14 vs 94</td>
<td>6 vs 49</td>
</tr>
<tr>
<td>Hyperdip&gt;50 vs OTHERS</td>
<td>$\text{OTHERS} = {\text{Hyperdip}&lt;47-50, \text{Pseudodip, Hypodip, Normo}}$</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 $A_{ij}$ ($E_{ij} = R_i * C_j / N$).
### Accuracy

<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><strong>T-ALL vs OTHERS1</strong></td>
<td>0:1</td>
</tr>
<tr>
<td><strong>E2A-PBX1 vs OTHERS2</strong></td>
<td>0:0</td>
</tr>
<tr>
<td><strong>TEL-AML1 vs OTHERS3</strong></td>
<td>1:1</td>
</tr>
<tr>
<td><strong>BCR-ABL vs OTHERS4</strong></td>
<td>2:0</td>
</tr>
<tr>
<td><strong>MLL vs OTHERS5</strong></td>
<td>0:1</td>
</tr>
<tr>
<td><strong>Hyperdiploid&gt;50 vs OTHERS</strong></td>
<td>2:6</td>
</tr>
</tbody>
</table>

| Total Errors            | 14    | 6     | 8     | 4     |

The classifiers are all applied to the 20 genes selected by χ² at each level of the tree.
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 \times 2000) for intermediate intensity tx
- US$30m (US$150k \times 2000 \times 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
PCA in disease diagnosis and biomarker selection
PCA, intuitively

Credit: Alessandro Giuliani
PCA, in brief

\( X_{\text{raw}} \) \( n \) samples \( \times \) \( p \) genes data matrix

\( \mu \) mean vector of each gene

\[ X = X_{\text{raw}} - \mu \] the centered matrix

\( V \) \( p \times k \) eigenvectors; i.e. the PCA

\[ Z = X V \] the \( k \) PC projections

---

**Definition of PCA**

- Given \( X = (X_1, \ldots, X_p)' \)
- We call \( a'X \) a standard linear combination (SLC) if \( \sum a_i^2 = 1 \)
- Find the SLC \( a'_1 = (a_{11}, \ldots, a_{p1}) \) so that \( Y_1 = a'_1X \) has maximal variance
- Find the SLC \( a'_2 = (a_{12}, \ldots, a_{p2}) \) so that \( Y_2 = a'_2X \) has maximal variance, subject to the constraint that \( Y_2 \) is uncorrelated to \( Y_1 \).
- Find the SLC \( a'_3 = (a_{13}, \ldots, a_{p3}) \) so that \( Y_3 = a'_3X \) has maximal variance, subject to the constraint that \( Y_3 \) is uncorrelated to \( Y_1 \) and \( Y_2 \).
- Etc...
PCA in disease diagnosis

- The PCs capture different biological pathways. The values of PCs capture different states of these pathways

- Hypothesis: If patient X has ALL subtype T, X’s biological pathways are in state $S_T$

- Observation: John’s biological pathways are in state $S_T$

- Abduction: John has ALL subtype T
PCA in ALL subtype diagnosis

Steps:
Identify genes with high variance
Perform PCA on them
Plot using PC1 to 3
PCA in biomarker selection

• PCA is applied e.g. on gene expression data,
  – PCs w/ large variance ≈ diff expressed pathways
  – Variables w/ large coefficient/loading in a PC ≈ key genes in the pathway associated with that PC

• PCA can be a useful biomarker-selection approach, e.g., biomarkers ≈ genes w/ high loading
  – Loading of gene $x = \Sigma_j \mid \alpha x_j * \sigma_j^2 \mid$, where $\alpha x_j$ is coefficient of $x$ in $PC_j$, and $\sigma_j^2$ is variance of $PC_j$
Example

• Quiescence cancer cells (QCC) are a common feature of solid tumors, and represent an obstacle to long-term success of cancer therapies

⇒ Look for quiescence signature in lung cancer

Data

• NSCLC specimens taken from patients undergoing surgical resection

• NSCLC spheroids stained w/ dye PKH26, which is rapidly lost by proliferating cells but retained by quiescent cells

• PKH26-stained spheroid cells inoculated into immunocompromised mice, and tumors allowed to grow for 4 weeks: PKH26+ cells reached ~5% of total tumour cells

• PKH26+ and PKH26- cells separated by FACS sorting and lysed for gene expression analysis
Analysis procedure

• Apply PCA to the patient x gene matrix

• Project patients to the top PCs, which account for 99% of variance

• Determine the PC that corresponds to quiescence

• Determine genes with high loading on this PC; i.e. genes whose expression is highly correlated to this PC

• Test these genes for statistically significant association w/ quiescence, etc. etc.
PC3 perfectly separates PKH26+ and PKH26-

So, determine genes highly correlated with PC3, see what pathways are enriched with these genes, etc.
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
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>benign</td>
<td>malign</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>
</tbody>
</table>

Mr. A: ???

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

- 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?
- Two Clusters
- Six 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

Assignment
Update
K-means clustering illustration
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 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
Single, complete, & average Linkage

**Single linkage** defines distance between two clusters as min distance between them.

**Complete linkage** defines distance between two clusters as max distance between them.

Exercise: Give definition of “average linkage”

Image source: UCL Microcore Website
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>...</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
Intermediate situation

- After some merging steps, we have some clusters

After some merging steps, we have some clusters.
Intermediate situation

- We want to **merge** the two closest clusters (C2 and C5) and update the 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>
```

Proximity Matrix
After merging

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

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

![Diagram showing similarity and Proximity Matrix](Image)

<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

|     | p1 | p2 | p3 | p4 | p5 | ...
|-----|----|----|----|----|----|------
| p1  |    |    |    |    |    |      
| p2  |    |    |    |    |    |      
| p3  |    |    |    |    |    |      
| p4  |    |    |    |    |    |      
| p5  |    |    |    |    |    |      
|     |    |    |    |    |    |      

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>
<tr>
<td>. . .</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

<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

|     | p1 | p2 | p3 | p4 | p5 | ...
|-----|----|----|----|----|----|-----
| p1  |    |    |    |    |    |     
| p2  |    |    |    |    |    |     
| p3  |    |    |    |    |    |     
| p4  |    |    |    |    |    |     
| p5  |    |    |    |    |    |     
|     |    |    |    |    |    |     

• 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

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

Table 8.4. Euclidean distance matrix for 6 points.
Hierarchical clustering: Group average
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?
Batch effects & 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

Image credit: Dong Difeng
What are batch effects?

• Batch effects are unwanted sources of variation caused by different processing date, handling personnel, reagent lots, equipment/machines, etc.

• Batch effects is a big challenge faced in biological research, especially towards translational research and precision medicine
Visualizing batch effects

- Rank variables / genes by variance
- Keep those with high variance (e.g. top 30-50%)
- Perform PCA on them
- Make scatter plot of the first 2-3 PCs
  - Do the subjects clusters by batch?
- Make paired boxplot of each PC wrt class and batch variables
  - Is PC more correlated with batch?
Sometimes, a gene expression study may involve batches of data collected over a long period of time...

**PCA scatter plot**

Samples from diff batches are grouped together, regardless of subtypes and treatment response

Image credit: Difeng Dong’s PhD dissertation, 2011
Paired boxplots of PCs

It is easier to see which PC is enriched in batch effects by showing, side by side, the distribution of values of each PC stratified by class and suspected batch variables.
Dealing with batch effects

• **Batch effect removal**
  – Dropping PCs laden with batch effects
  – Batch effect-correction methods: ComBat, batch mean centering, gene fuzzing scoring (GFS), …

• **Normalization**
  – Scaling
  – Z-score transformation
  – Quantile normalization
Determine PCs associated with batch using paired boxplots of PCs

Batch effects dominate PC1

Goh & Wong, “Protein complex-based analysis is resistant to the obfuscating consequences of batch effects”, BMC Genomics 18(Suppl2):142, 2017
Samples separate by class post PC1 removal, no batch subgrouping

A and B are different datasets with different batch effects inserted

Batch effects dominate

Class-effect discrimination recovered

(Notation: A/B_D/D*_1/2 refers to the dataset, class and batches respectively)
Exercise

This “Batch effect-mitigation by PC removal” approach works in “PCA space”

How to do this in the original gene space? I.e., how to produce a batch-corrected gene expression matrix?
Exercise

Suggest a modification to the formula below to avoid selecting genes laden with batch effects

PCA can be a useful biomarker-selection approach

- E.g., biomarkers \( \approx \) genes w/ high loading
  
  \[
  \text{Loading of gene } x = \sum_j | \alpha_{xj} \times \sigma_j^2 |, \text{ where } \alpha_{xj} \text{ is coefficient of } x \text{ in } PC_j, \text{ and } \sigma_j^2 \text{ is variance of } PC_j
  \]
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...

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.
Caution: It is difficult to eliminate batch effects effectively

Green and orange are normal samples differing in processing date

a: Before normalization

b: Post normalization

c: Checks on individual genes susceptible to batch effects

d: Clustering after normalization (samples still cluster by...
Caution: “Over normalized” signals in cancer samples

A gene normalized by quantile normalization (RMA) was detected as down-regulated DE gene, but the original probe intensities in cancer samples were not diff from those in normal samples.

A gene was detected as an up-regulated DE gene in the non-normalized data, but was not identified as a DE gene in the quantile-normalized data.

Wang et al. Molecular Biosystems, 8:818-827, 2012
Impact on biomarker selection


A Balanced

B Moderately unbalanced

C Severely unbalanced

Precision  Recall  F-score
Impact on disease diagnosis
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?

Exercise #6
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?
Concluding remarks
What have we learned?

- **Technologies**
  - Microarray

- **Microarray applications**
  - Disease diagnosis by supervised learning
  - Subtype discovery by unsupervised learning
  - Disease diagnosis via guilt-by-association
  - Biomarker selection by PCA

- **Important issue**
  - Batch effect
  - Normalization
Any question?
References


• L. Zhou et al., “Examining the practical limits of batch effect-correction algorithms: When should you care about batch effects?”, *Journal of Genetics and Genomics*, 46:433—443, 2019