After this lecture, you can...

... think of “omics” data as being multidimensional vectors
... visualize multidimensional datasets and interpret PCA
... quantify distance between vectors using Euclidean, Manhattan, Pearson, and Spearman
... convert similarity into distance and *vice versa*
... cluster datasets with single, complete, and average linkage
... explain chaining and clumping behavior in clustering
... interpret clusterings and bi-clustering cladograms in terms of a biological question
• Gene expression of two housekeeping genes in 28 samples (18 lung tumors, 10 normal lung tissues)
  – Y-axis: glyceraldehyde-3-phosphate dehydrogenase expression
  – X-axis: hypoxanthin phosphoribosyltransferase 1 expression
  

• Expression of three proteins in many tissues
  – Housekeeping protein (GAPDH)
  – Signalling protein (EGFR)
  – Tumor-associated protein (CTNNB1)

Wilhelm et al. Nature 2014
Abundance profiles

• Many different biological questions can be answered with abundance profiles
• Sometimes you just want to know “What is in my sample?”
• Most of the time you will be interested in comparing samples

Heatmaps help visualize values

• Data values can be visualized using heat maps

Copyright © 2006 Nature Publishing Group
Nature Reviews Genetics

Data points are vectors

- A list of measurements is mathematically equal to a ...
  - Multidimensional vector
  - Point in multidimensional space

Mathematical toolbox

- By representing data as vectors, a mathematical toolbox unfolds
  - Visualizing data with PCA
  - Quantifying similarity between samples
  - Clustering

- These tools can be very valuable to understand your data
Visualizing highly-dimensional data

• Plotted in the graph is a large set of three-dimensional vectors (data points with X, Y, and Z coordinates, in gray)

• The two-dimensional projections are plotted in blue (in the X-direction), green (Y), and red (Z)

• Together, the three projections contain all the information in this graph

• **Rule:** *N* perpendicular projections always contain all information of a *N*-dimensional dataset

---

Principal components

• Projection into fewer dimensions is a commonly used approach to visualize highly-dimensional data

• Which direction(s) should we choose for the projection?

  ➔ The direction(s) that contain the most variation! (…in decreasing order)

• These directions are called the principal components of the data

• Directions do not have to correspond to original axes

• Principal components are always perpendicular to each other (90° angle)
Visualizing highly-dimensional data

- Most information in a $N$-dimensional dataset is contained in the first few principal components
  - Other dimensions predictable (no new info)
- This allows us to visualize most information on e.g. a sheet of paper (2D)

Human microbiome

- Microbiome information is very highly dimensional
  - Which bacterial phyla are present in different human body sites?
  - Which metabolic functions do they encode?
Principal Component Analysis (PCA)

- Lists of microbes are high dimensional vectors of numbers
- Visualizing this data by PCA quickly shows you:
  - What is the most important signal in the data: PC1
  - Whether there are clusters in the data

![PCA Diagram]

HMP, Nature 2012

Forensic analysis

- Unknown sample (X) most likely comes from Lansing

![Forensic Analysis Diagram]
Metadata vectors in a PCA

- Metadata (like physico-chemical composition) is also a vector
  - Metadata vectors can also be plotted on a PCA
  - This shows you which metadata vectors correlate with data points

![PCA Diagram](image)

**Figure 5.** Ordination diagram of the sampling sites by Principal Component Analysis (PCA), considering the physical and chemical variables of water from Streams 1, 2, 3 and 4, Uthelandia - 3OG, 2007. S - Stream (1, 2, 3 and 4), w - wet season, d - dry season.

The cell cycle

- The cell cycle is an important process in all cellular life forms
  - Reproduction
  - Growth and development
  - Tissue renewal

![Cell Cycle Diagram](image)

**Figure 6.** The cell cycle: 1. Cell division; 2. G1; 3. S; 4. G2; 5. M (division). Each cycle is divided into two phases: DNA synthesis (S) and mitosis (M). DNA synthesis occurs in the S phase and mitosis in the M phase.
Cell cycle genes

- Genes involved in the cell cycle are expected to be expressed during specific stages of the cycle.

- Expression of all genes in the genome of *Saccharomyces cerevisiae* was measured during several cell cycles.
  - 800 genes oscillate, indicating that they are involved in the cell cycle.

Predicting functions for unknown genes

- Unknown oscillating genes could be clustered with other genes that are known to function in specific cell cycle stages.
Quantifying distances between vectors

The levels are most similar between • and ■. The patterns are most similar between • and ▲.

Euclidean distance (levels)

The levels are most similar between • and ■. The patterns are most similar between • and ▲.

Example:

\[ d_{AB} = \sqrt{(X_A - X_B)^2 + (Y_A - Y_B)^2} \]

\[ d_{AB} = \sum (D_A - D_B)^2 \]

\[ d_{AB} = 0.103 \]
\[ d_{AB} = 0.253 \]
\[ d_{AB} = 0.178 \]
**Manhattan distance (levels)**

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>0.265</th>
<th>0.799</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0.265</td>
<td>0.799</td>
</tr>
<tr>
<td>0.265</td>
<td>0</td>
<td>0.534</td>
<td></td>
</tr>
<tr>
<td>0.799</td>
<td>0.534</td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

- Example:
  \[d = |0.20 - 0.15| + |0.17 - 0.15| + |0.16 - 0.16| + |0.20 - 0.15| + |0.20 - 0.16| + |0.17 - 0.16| + |0.16 - 0.15| + |0.20 - 0.15| + |0.18 - 0.16| + |0.16 - 0.15| = 0.265\]
  \[d = 0.799\]
  \[d = 0.534\]

**Distance matrices**

- We store distances in a **distance matrix**
  - Distance matrices are the input for clustering algorithms
  - Most distance measures are symmetrical: \[d_{AB} = d_{BA}\]
- Some methods calculate similarity rather than distance
  - Distance = 1 - similarity

**Similarity matrix**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>1 - x</th>
<th>1 - y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1 - x</td>
<td>1 - y</td>
</tr>
<tr>
<td>1 - x</td>
<td>1</td>
<td>1 - z</td>
<td></td>
</tr>
<tr>
<td>1 - y</td>
<td>1 - z</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

**Distance matrix**

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>x</td>
<td>y</td>
</tr>
<tr>
<td>x</td>
<td>y</td>
<td>z</td>
<td>0</td>
</tr>
</tbody>
</table>

The levels are most similar between ● and ▲. The patterns are most similar between ● and ▲.
Correlation (patterns)

- Pearson correlation, sensitive to outliers
- Spearman rank correlation, robust to outliers but less sensitive

A distance matrix allows us to create clusters

- In this example distance matrix:
  - and have the most similar vectors
  - and are the second most similar
  - and are the least similar

- Every time we merge the two most similar vectors/data points in the matrix
  - Most similar = lowest value in a distance matrix

- These relationships are hierarchical
- We can draw them as a cladogram
Hierarchical clustering

What is the distance between clusters ● & ●?

- Single linkage
  - Distance between clusters is the distance between closest vectors
- Complete linkage
  - Distance between clusters is distance between most distant vectors
- Average linkage (a.k.a. UPGMA)
  - Distance between clusters is the average distance between all vectors
  - Unweighted Pair Group Method with Arithmetic mean (UPGMA)
Clustering algorithm

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>8</td>
<td>7</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td>0</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

Always join most similar data points first [lowest distance]

Average linkage: the distance for a cluster is the average distance between its components

Single linkage: smallest distance
Complete linkage: largest distance

Chaining versus clumping behavior

- Single linkage cladograms typically show chaining behavior
  - Often, a single close data point (vector) is added to an existing cluster

- Conversely, complete linkage cladograms are more clumped

- Average linkage (UPGMA) is intermediate
Bracket-notation

- Cladograms can also be written as a **bracket-notation**
- The bracket-notation uses nested brackets to represent hierarchical clusters
- Every cluster is delimited by a pair of brackets
- Clusters that are part of a larger cluster are separated by commas
- A bracket-notation always ends with a semi-colon; so the computer knows to stop reading

\[
((A,B),((C,D),E));
\]

Branch lengths

\[
((A:0.5,B:0.5):2.67,((C:1.5,D:1.5):1,E:2.5):0.67));
\]
Distance measure affects clustering

Distance measures can significantly affect the outcome of clustering algorithms. Commonly used distance measures include Manhattan and Euclidean distances.

**Manhattan** and **Euclidian**

Gene 1
Gene 2
Gene 3

**Correlation**

Gene 1
Gene 2
Gene 3

**Biological interpretation of clusters**

- Clustering allows us to identify categories in our samples
  - Long branches show you the biggest differences in the data
  - Short branches show you the most similar data points
- Clusters might represent:
  - Genes involved in the same metabolic pathway
  - Micro-organisms responding to the same environmental signals
- Are gene expression profiles clustered for patients with similar disease outcome?

**Cladogram**

**Heat-map**

**Metadata**

**Abundance/Expression**

Time/environments/samples...

0
0.05
0.10
0.15
0.20
0.25

Gene 1
Gene 2
Gene 3
Bi-clustering: changing perspective

Gene/protein/microbe/etc

Time/Environments/Patients/Leaves

Distance matrix between categories

Distance matrix between samples

Bi-clustered heatmap

Infected

Uninfected

Genes

Samples

Increased

Decreased
How good is my clustering?
• Does the picture or cladogram fit your expectation?
• Are the samples/genes/microbes/patients/etc that “should” cluster together, indeed clustered together?
  – Genes involved in the same pathway
  – Symbiotic microbes
  – Patients with the same disease phenotype
  – Etc.

→ If so, information can be transferred to new/unknown samples/genes/microbes/patients/etc