• Duration: 50 minutes.

• Exam:
  – This is a closed book, closed notes exam. The use of any reference material is strictly forbidden.
  – No attempts of cheating will be tolerated. In case such attempts are observed, the students who took part in the act will be prosecuted. The legal code states that students who are found guilty of cheating shall be expelled from the university for a minimum of one semester!

• About the exam questions:
  – The points assigned for each question are shown in parenthesis next to the question.

• This exam consists of 6 pages including this page. Check that you have them all!

• GOOD LUCK!

<table>
<thead>
<tr>
<th>Question</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question 1</td>
<td>40</td>
</tr>
<tr>
<td>Question 2</td>
<td>40</td>
</tr>
<tr>
<td>Question 3</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
</tr>
</tbody>
</table>
(a) (20 pts) Given the following distance matrix $D$ for the four species $sp_1$, $sp_2$, $sp_3$, and $sp_4$, construct a phylogenetic tree using the UPGMA method. Show intermediate matrices and subtrees in your construction.

<table>
<thead>
<tr>
<th></th>
<th>$sp_1$</th>
<th>$sp_2$</th>
<th>$sp_3$</th>
<th>$sp_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$sp_1$</td>
<td>0</td>
<td>12</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>$sp_2$</td>
<td>12</td>
<td>0</td>
<td>6</td>
<td>14</td>
</tr>
<tr>
<td>$sp_3$</td>
<td>10</td>
<td>6</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>$sp_4$</td>
<td>4</td>
<td>14</td>
<td>12</td>
<td>0</td>
</tr>
</tbody>
</table>

Step (1) $sp_1$ and $sp_4$ are closest, merge them.

Step (2) $sp_2$ and $sp_3$ are closest, merge them.

Step (3) Only two nodes left to merge, merge them.
(b) (10 pts) Based on the tree you constructed in part a fill-in the following path
distance matrix d using the path distances between the four species.

\[
\begin{array}{cccc}
sp_1 & sp_2 & sp_3 & sp_4 \\
sp_1 & 0 & 12 & 12 & 4 \\
sp_2 & 12 & 0 & 6 & 12 \\
sp_3 & 12 & 6 & 0 & 12 \\
sp_4 & 4 & 12 & 12 & 0 \\
\end{array}
\]

(c) (10 pts) How close is the tree you generated in part a to the actual (i.e., true)
phylogenetic tree which preserves the original distances between the four species? In
other words, what is the error between matrix D and matrix d using LSQ (least
squares) error measure? (You may use upper or lower triangle of the matrix when
computing the LSQ error.)

\[
\text{LSQ error} = \frac{4}{4} + \frac{4}{4} + \frac{4}{4} + \frac{4}{4} + \frac{4}{4} = 4 + 4 = 8
\]

\[
\text{LSQ} = \sum \sum (d_{ij} - D_{ij})^2
\]
Consider the following gene expression values from 2 microarray experiments for 8 genes.

<table>
<thead>
<tr>
<th></th>
<th>exp1</th>
<th>exp2</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene1</td>
<td>-4</td>
<td>-3</td>
</tr>
<tr>
<td>gene2</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>gene3</td>
<td>1</td>
<td>-7</td>
</tr>
<tr>
<td>gene4</td>
<td>-4</td>
<td>-6</td>
</tr>
<tr>
<td>gene5</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>gene6</td>
<td>-1</td>
<td>5</td>
</tr>
<tr>
<td>gene7</td>
<td>-3</td>
<td>0</td>
</tr>
<tr>
<td>gene8</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

Negative expression values in the table mean that the gene is downregulated (i.e., expressed less) in the experimented cell, and positive values mean that the gene is upregulated (i.e., overexpressed). A biologist is trying to find out whether these 8 genes can be separated into two groups based on their behavior in the experimented conditions. In order to visualize the relationships, she sketched a 2-dimensional plot of the genes shown below.

Use k-means clustering to cluster these 8 genes into 2 clusters. Use gene7 (light grey data point on the 2-d plot) as the initial cluster center for cluster 1 and use genes (dark grey data point) as the initial cluster center for cluster 2. Indicate which data point belongs to which cluster and also give the coordinates of the centroids at each iteration. Iterate until convergence, i.e., until no change in clusters. When computing distances to centroids, you may use the squared distance (no need to take the square-root).
Compute new cluster centroids:

\[ \text{cluster 1: gene}_1, \text{gene}_4, \text{gene}_6, \text{gene}_7 \]
\[ (x, y) = \left( \frac{-4+4-3}{4}, \frac{-3-6+5+0}{4} \right) = (-3, -1) \]

\[ \text{cluster 2: gene}_2, \text{gene}_3, \text{gene}_5, \text{gene}_8 \]
\[ (x, y) = \left( \frac{6+1+4+3}{4}, \frac{5-7+6+0}{4} \right) = (3.5, 1) \]

Iteration 2: calculate new squared distances to cluster centroids:

\[ \text{cluster 1:} \]
\[ \text{gene}_1 = (4+3)^2 + (-3+1)^2 = 5 \]
\[ \text{gene}_2 = (6+3)^2 + (5+1)^2 = 117 \]
\[ \text{gene}_3 = (-1+3)^2 + (-7+1)^2 = 52 \]
\[ \text{gene}_4 = (-4+3)^2 + (-6+1)^2 = 26 \]
\[ \text{gene}_5 = (-4+3)^2 + (6+1)^2 = 98 \]
\[ \text{gene}_6 = (-1+3)^2 + (5+1)^2 = 40 \]
\[ \text{gene}_7 = (3+3)^2 + (0+1)^2 = 1 \]
\[ \text{gene}_8 = (3+3)^2 + (0+1)^2 = 10 \]

\[ \text{cluster 2:} \]
\[ \text{gene}_1 = (-4-3.5)^2 + (3-1)^2 = 5 \]
\[ \text{gene}_2 = (6-3.5)^2 + (5-1)^2 = 117 \]
\[ \text{gene}_3 = (-3.5)^2 + (-7-1)^2 = 52 \]
\[ \text{gene}_4 = (-4-3.5)^2 + (6-1)^2 = 26 \]
\[ \text{gene}_5 = (4-3.5)^2 + (6-1)^2 = 98 \]
\[ \text{gene}_6 = (-1-3.5)^2 + (5-1)^2 = 40 \]
\[ \text{gene}_7 = (-3.5)^2 + (0-1)^2 = 1 \]
\[ \text{gene}_8 = (3-3.5)^2 + (0-1)^2 = 10 \]

Cluster 1 and cluster 2 distances are compared and cluster assignments are shown in boxes.

Compute new cluster centroids:

\[ \text{cluster 1:} \]
\[ \text{gene}_1, \text{gene}_3, \text{gene}_4, \text{gene}_7 \]
\[ (x, y) = \left( \frac{-4+1+4-3}{4}, \frac{-3-7+6+0}{4} \right) = (-2.5, -4) \]

\[ \text{cluster 2:} \]
\[ \text{gene}_2, \text{gene}_5, \text{gene}_6, \text{gene}_8 \]
\[ (x, y) = \left( \frac{6+4-1+3}{4}, \frac{5-7+6+5+0}{4} \right) = (3, 4) \]

Iteration 3: calculate new squared distances to cluster centroids:

\[ \text{cluster 1:} \]
\[ \text{gene}_1 = (-4+2.5)^2 + (-3+4)^2 < 98 \]
\[ \text{gene}_2 = (4+2.5)^2 + (5+4)^2 > 10 \]
\[ \text{gene}_3 = (1+2.5)^2 + (-7+4)^2 < 19.5 \]
\[ \text{gene}_4 = (4+2.5)^2 + (-6+4)^2 < 14.9 \]
\[ \text{gene}_5 = (4+2.5)^2 + (6+4)^2 > 5 \]
\[ \text{gene}_6 = (-1+2.5)^2 + (5+4)^2 > 17 \]
\[ \text{gene}_7 = (-3+2.5)^2 + (0+4)^2 < 32 \]
\[ \text{gene}_8 = (3+2.5)^2 + (0+4)^2 > 16 \]

\[ \text{cluster 2:} \]
\[ \text{gene}_1 = (-4-3)^2 + (-3-4)^2 = 98 \]
\[ \text{gene}_2 = (6-3)^2 + (5-4)^2 = 10 \]
\[ \text{gene}_3 = (-3)^2 + (-7-4)^2 = 12.5 \]
\[ \text{gene}_4 = (-4-3)^2 + (-6-4)^2 = 14.9 \]
\[ \text{gene}_5 = (4-3)^2 + (6-4)^2 = 5 \]
\[ \text{gene}_6 = (-1-3)^2 + (5-4)^2 = 17 \]
\[ \text{gene}_7 = (-3-3)^2 + (0-4)^2 = 52 \]
\[ \text{gene}_8 = (3-3)^2 + (0-4)^2 = 16 \]

\[ \Rightarrow \text{clusters don't change} \Rightarrow \text{STOP}. \]
Short answer questions:

(a) (10 pts) When examining the quality of a structural alignment, is it enough to inspect RMSD alone? Why/Why not?

It is **NOT** enough to inspect RMSD alone. Because we also need the **length** of the alignment. It is possible to get very good RMSD values for very short alignments. However, this does not imply a high quality alignment. Therefore, **length of alignment** is a critical information that should be inspected along with RMSD.

(b) (10 pts) What is the difference between single-linkage and complete-linkage clustering techniques?

In single-linkage clustering, the distance between two clusters is defined as "the distance between the closest members of respective clusters." However, in complete-linkage clustering, the distance is defined as "the distance between the farthest members" of the respective clusters.