1. Distance-Based Phylogeny

1) Given the following pairwise distances, use a distance-based method to infer a phylogenetic tree for a, b, c, d, and e.

<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>4</td>
<td>16</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>b</td>
<td>0</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>0</td>
<td>10</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>0</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The data is ultrametric. Therefore, we should use UPGMA because we know it will construct a rooted tree that is consistent with the data.
2. Parsimony with Fitch’s Algorithm

5) Given the following three sequences, draw the possible rooted trees for the sequences, and use Fitch’s algorithm (traditional parsimony) to determine the hypothetical sequences for each internal node in the trees. Which tree would the parsimony principle prefer?

CGC
TAT
CAT

2. Parsimony

• three possible trees
• three sequence positions to consider
2. Parsimony (Continued)

```
C A T  \{C\}{A\}T
\{CT\}{GA\}{CT}   \{C\}{A\}T
\{CT\}{CT\}         \{CT\}{A\}T
-------------
CGC  TAT
2 1
3 total changes
```
2. Parsimony (Continued)

- so parsimony does not have a preference among these trees

![](image)
3. EM Clustering

Consider a one-dimensional clustering problem in which the data given are:

\[ x_1 = -4 \]
\[ x_2 = -3 \]
\[ x_3 = -1 \]
\[ x_4 = 3 \]
\[ x_5 = 5 \]

Show the first two steps in the EM-based clustering procedure in which we model the clusters using two Gaussians. The initial mean of the first Gaussian is 0 and the initial mean of the second is 2. The Gaussians have fixed width; you should use EM to determine only the means. The density function to use for the Gaussians is:

\[ f(x, \mu) = \frac{1}{\sqrt{8\pi}} e^{-\frac{1}{2} \left( \frac{x-\mu}{2} \right)^2} \]

where \( \mu \) denotes the mean (center) of the Gaussian.

**NOTE:** Here I show just the first step

\[
\begin{align*}
  f(-4, \mu) &= \frac{1}{\sqrt{8\pi}} e^{-\frac{1}{2} \left( \frac{-4-0}{2} \right)^2} = .0269 \\
  f(-3, \mu) &= .0646 \\
  f(-1, \mu) &= .176 \\
  f(3, \mu) &= .0646 \\
  f(5, \mu) &= .00874
\end{align*}
\]
3. EM Clustering: E Step

\[ h_1 = \frac{f(x_i, \mu_1)}{f(x_i, \mu_1) + f(x_i, \mu_2)} = \frac{.0269}{.0269 + .0022} = .0269 + .0022 \\
\]

\[ h_2 = \frac{f(x_i, \mu_2)}{f(x_i, \mu_1) + f(x_i, \mu_2)} = \frac{.0646}{.0646 + .00874} = .0646 + .00874 \\
\]

\[ h_3 = \frac{.176}{.176 + .0646} = .176 + .0646 \\
\]

\[ h_4 = \frac{.0646}{.0646 + .176} = .0646 + .176 \\
\]

\[ h_5 = \frac{.00874}{.00874 + .0646} = .00874 + .0646 \\
\]

3. EM Clustering: M-step

\[ \mu_1 = \frac{\sum h_{i1} \times x_i}{\sum h_{i1}} = \frac{-4 \times .924 + -3 \times .881 + -1 \times .732 + 3 \times .268 + 5 \times .119}{.924 + .881 + .732 + .268 + .119} = -1.94 \\
\]

\[ \mu_2 = \frac{\sum h_{i2} \times x_i}{\sum h_{i2}} = \frac{-4 \times .076 + -3 \times .119 + -1 \times .268 + 3 \times .732 + 5 \times .881}{.076 + .119 + .268 + .732 + .881} = 3.39 \\
\]
4. Threading

3) Consider a simple threading problem in which we have a template with three segments (i, j, k), each of which includes three amino acids. For a given sequence there are three possible starting positions for each segment. Suppose that you are given the following values for the scores of the individual segments and the scores for segment interactions.

\[
g_1(i, 2) = 5 \quad g_1(j, 8) = 9 \quad g_1(k, 13) = 3 \\
g_1(i, 3) = 2 \quad g_1(j, 9) = 7 \quad g_1(k, 14) = 4 \\
g_1(i, 4) = 8 \quad g_1(j, 10) = 6 \quad g_1(k, 15) = 1 \\
g_2(i, j, 2, 8) = 1 \quad g_2(i, j, 4, 8) = 7 \quad g_2(i, j, 3, 10) = 5 \\
g_2(i, j, 3, 8) = 5 \quad g_2(i, k, 9, 13) = 1 \quad g_2(i, k, 9, 14) = 6 \\
g_2(i, k, 3, 10) = 6 \quad g_2(i, k, 9, 15) = 8 \quad g_2(i, k, 10, 13) = 11 \\
g_2(j, k, 9, 13) = 7 \quad g_2(j, k, 8, 13) = 7 \quad g_2(j, k, 10, 14) = 12 \\
g_2(j, k, 4, 10) = 4 \quad g_2(j, k, 10, 15) = 13 \quad g_2(j, k, 13) = 19 \\
g_2(j, k, 14) = 19 \quad g_2(j, k, 15) = 28 \\
\]

Show how the branch and bound method would find the optimal threading. Show the threading sets considered in the search process and the lower bound calculated for each. Use the "simple lower bound" presented in class. When splitting a threading, split the segment having the minimal \( g_1 \) value (i.e. split on \( k \) first in the example above). To split the selected segment, divide it into three intervals of length one (this will work in the example because each segment has only three possible start positions).
5. Parsimony with Weighted Fitch

5) Consider the two simple phylogenetic trees shown below, and the symmetric cost matrix for assessing nucleotide changes. The tree on the right has a cost of 0.8.

\[
\begin{array}{c|cccc}
 & a & c & g & t \\
\hline
a & 0 & 0.8 & 0.2 & 0.9 \\
c & 0.8 & 0 & 0.7 & 0.5 \\
g & 0.2 & 0.7 & 0 & 0.1 \\
t & 0.9 & 0.5 & 0.1 & 0 \\
\end{array}
\]

5a) Show how the weighted version of Fitch’s algorithm would determine the cost of the tree on the left?

5b) What are the minimal cost characters for the internal nodes in the tree on the left?

5c) Which of the two trees would the maximum parsimony approach prefer?

The minimal cost characters for node 1 are either g or t. The minimal cost character for node 3 is g. The maximum parsimony approach would prefer the other tree.
6. Hierarchical Clustering

6) Given the following similarity matrix for 5 genes, show how complete-link clustering would cluster the genes.

<table>
<thead>
<tr>
<th></th>
<th>x_1</th>
<th>x_2</th>
<th>x_3</th>
<th>x_4</th>
<th>x_5</th>
</tr>
</thead>
<tbody>
<tr>
<td>x_1</td>
<td>4</td>
<td>8</td>
<td>5</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>x_2</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x_3</td>
<td>7</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x_4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>x_5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


6. Hierarchical Clustering

<table>
<thead>
<tr>
<th></th>
<th>x_1</th>
<th>x_2</th>
<th>x_3</th>
<th>x_4</th>
<th>x_5</th>
</tr>
</thead>
<tbody>
<tr>
<td>x_1</td>
<td>4</td>
<td>8</td>
<td>5</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>x_2</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x_3</td>
<td>7</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x_4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>x_5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6. Hierarchical Clustering

<table>
<thead>
<tr>
<th></th>
<th>x_2</th>
<th>x_4</th>
<th>x_5</th>
<th>c_1</th>
</tr>
</thead>
<tbody>
<tr>
<td>x_2</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>x_4</td>
<td>6</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x_5</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>c_1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
6. Hierarchical Clustering

<table>
<thead>
<tr>
<th>x_2</th>
<th>c_1</th>
<th>c_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>x_2</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>c_1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>c_2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Diagram showing hierarchical clustering with three clusters: c_1, c_2, and c_3, and five data points x_1, x_2, x_3, x_4, and x_5.