Problem Set 7: Clustering and Evolution

Handed out Wednesday, April 15. Due at the start of class Wednesday, April 22.

Homework Information: Please upload a PDF of your solutions to Moodle by **2pm central time**. If you write your solutions by hand, use an app like Adobe Scan to take a picture of it and turn it into a PDF.

1. (6 pts) Consider the following distance matrix.

	A	B	C	D	E
A	0	22	16	12	16
B	22	0	18	26	10
C	16	18	0	20	12
D	12	26	20	0	20
E	16	10	12	20	0

a. Is this distance matrix additive? Justify your answer.
b. Apply the AdditivePhylogeny algorithm described on the Large Additive Phylogeny Worksheet, page 2. Draw the final tree and the step-by-step results in a manner similar to the Additive Phylogeny Worksheets on the Course Website.
c. Describe an efficient algorithm for determining the value of δ in the AdditivePhylogeny algorithm.

2. (8 pts) Consider the following four sequences.

Chimp GTTG
Gorilla GTCA
Human ACCA
Orangutan ATTA

Assume the following scoring matrix.

	A	T	G	C
A	0	2	3	8
T	2	0	1	3
G	3	1	0	3
C	8	3	3	0

a. For these four species, draw any four binary tree topologies. Make at least two different tree structures.
b. Apply Sankoff's algorithm to the topologies created in (a) to determine the weighted parsimony score. What is the minimum parsimony score? What is the most plausible topology of the four you created? Why?
c. Apply Fitch's algorithm to the same topologies. Find the sequences for the roots for all. Which is the topology with the lowest parsimony score?
d. Do b) and c) give the same results? Why or Why not?

3. (10 pts) Consider the following set of points in two dimensional space.

$$(2.2, 3.3), (1.6, 4.5), (6.7, 3.3), (4.7, 1.2), (8.6, 7.5), (5.3, 5.6), (7.9, 6.5), (2.5, 4.6), (5.5, 4.5), (1.1, 3.8)$$

a. Assume that the Euclidean distance is used. Compute the distance matrix for these points.

b. Assume that the distance between two clusters is defined by the minimal distance of any pair of their elements. Draw the cluster tree constructed by the hierarchical clustering algorithm given below.

c. Assume that the distance between two clusters is defined as the average distance between their elements. Draw the hierarchical cluster tree.

d. We want to run the k-means algorithm with $k = 2$. What would be the clustering result if we take the first two points (i.e. (2.2, 3.3) and (1.6, 4.5)) as the initial cluster representatives.

e. Would the result be different if we used the last two points (5.5, 4.5) and (1.1, 3.8) as the initial cluster representatives instead? Explain your observations.

```
Hierarchical Clustering (d, n)
  Form n clusters each with one element
  Construct a graph T by assigning one vertex to each cluster
  while there is more than one cluster
    Find the two closest clusters C1 and C2
    Merge C1 and C2 into new cluster C with |C1| + |C2| elements
    Compute distance from C to all other clusters
    Add a new vertex C to T and connect to vertices C1 and C2
    Remove rows and columns of d corresponding to C1 and C2
    Add a row and column to d corresponding to the new cluster C
  return T
```

4. Optional (up to 6 points extra credit) Phylogenetic trees represent hypothetical relationships between different species or taxa. Generally, we do not directly observe these relationships, instead we infer them from data. As such, we must always be careful to understand the complexities of the data we are using. In recent years, molecular data has supplemented pre-existing trait based phylogenies (think Darwin and his observations of things like bird's beak size). This new information has provided new insights, but we must still be careful when constructing phylogenies based on this new data.

a. (3 points) Phylogenetic trees build based solely upon molecular data may not always accurately represent the evolutionary history of the underlying species. What are two reasons why this might happen? (Feel free to use any and all resources to answer this question. Please make sure to cite the resources you use.).

b. (3 points) We may use either DNA or protein sequences to reconstruct phylogenetic trees. Which type of data do you think would be more accurate in reconstructing evolutionary histories from the distant past? Why? (Feel free to use any and all resources to answer this question. Please make sure to cite the resources you use.)

To receive all 3 points on each part your answer should include all requested details in a clear and concise format, as well as appropriate references to outside sources.