HW 3: due Thursday, November 10th in class

For this assignment, all answers are to be submitted in hardcopy in class.

This problem compares different algorithms for tree inference. Note that even though in this case, you will know the correct tree, in the case of real data you would have only the distance matrix and not the tree available to you.

1. Compute the distance matrix for the taxa V, W, X, Y, and Z from the tree below:

2. Is the distance matrix additive? Is it an ultrametric? How do you know?

3. Suppose you do not know the underlying tree. Working only from the computed distance matrix in part 1, which taxa would the UPGMA algorithm first select as neighbors? Show the first subtree connecting these two taxa, with branch lengths, according to the UPGMA algorithm.

4. Suppose again that you do not know the underlying tree. Working only from the computed distance matrix in part 1, show the corrected distance matrix computed for the first step of the NJ algorithm. Which taxa would the neighbor joining (NJ) algorithm first select as neighbors? Show the subtree connecting these two taxa, with the branch lengths as determined by the NJ algorithm.

5. Which algorithm is a better one for reconstructing this tree, and why?
This problem was inspired by a problem assigned in Shamir’s 2001 Algorithms for Molecular Biology class.

Background: You are a biologist studying a rare human disease called Homework. After years of work, you understand that it is associated with specific malfunctioning cells. You harvest mRNA from such a cell, and get a cDNA, which you sequence. You reverse and complement the result to obtain the coding strand, and get:

```
GTGGCGCTTCTGCCAGCAGATGGTTCTTCAATTGCTGCTGGACA
ACGCGGCTGGCAAGCATATTACTGCCTAGAGAGGTGGGCTGGGCTGGCTGTGA
GCTGTGGCGCCTGGTCAAACTGGGGGGACGACCTCCTGACGCTGCTACCCCG
CAGTGAAGTCTCCTCTGTGGAGGCCCCTGGAAGCCGATGGAGAAAGCGCA
GTCACCTGAAACGAGACACAGAAGACCCAAGAGACCCAAGAGACACAGGTAGATCCGGCT
CATTGAT
```

1. Search this sequence for open reading frames that make sense, and translate them into a protein. Are the following statements correct/incorrect/possible for the sequence/a subsequence of it (explain your answers):

   (a) It is a coding region
   (b) It is an exon
   (c) It is an intron
   (d) It is a 5' untranslated region
   (e) It is a 3' untranslated region

2. Search the databases for your full length molecule. Which database should you use? Which of the following is correct/incorrect/possibly correct (justify)?

   (a) You have found a new gene
   (b) You have found a mutated version of a known gene (what are the mutations?)
   (c) Your sequencing machine made sequencing errors? (what are they?)
   (d) The gene is not human, but rather contamination of a bacteria or fungus or virus in the testube?
(e) The database contains multiple records of your gene? (why?)

(f) Your gene has a homologous counterpart in mouse.

(g) Your gene has a homologous counterpart in the fruit fly.

3. Find your gene in the human genome. Where does your gene lie? What can you say about its function? Can you find more genes with related functions? Are they located anywhere near the Homework-related gene along the genome??