Homework #3

Due Tuesday Feb. 5 at the beginning of class. Assignments turned in more than 5 minutes after the beginning of class will be penalized 10 points, with an additional 10 points every 24 hours thereafter. You may discuss the homework assignment with other students, but do not share your work.

All Python programs should be run before being turned in. Even experienced programmers can seldom write a program perfectly on the first try.

The first 4 questions use the following data:

- Red Panda: ATCCGTATA
- Giant Panda: ATCTGTAAA
- Raccoon: ATTTGCAAA
- Dog: CTCTGCACA

1. (9 points) Draw the three possible unrooted trees of these four species.

2. (9 points) Mark the mutations in these data on your three trees. Give the final parsimony score of each tree.

3. (10 points) If we can assume that dogs are an outgroup (in other words, the root of this tree is on the branch leading to dogs), are giant pandas more closely related to red pandas or to raccoons?

4. (17 points) Make a distance matrix of the raw distances among these four species.

5. (10 points) Calculate the corrected Jukes-Cantor distance for a raw distance of 7 substitutions in 100 bp. Hint:

   \[ D = \frac{3}{4} \ln(1 - \frac{4}{3}D_s) \]

6. (15 points) Write a Python program that takes a raw distance as command line input and returns the Jukes-Cantor corrected distance. Print an error message if the raw distance is outside the range (0,0.75). Hint: ln is the natural log, available as the \texttt{log} function in the math module.

7. (15 points) The web site provides a sample file \texttt{dna.txt}. The first line of this file gives the number of sequences and the number of bases in each sequence. In subsequent lines, the first 10 characters are the name of the species, and the rest is DNA. Write a Python program to read this type of file and check that the numbers on the first line are a correct summary of the contents. Print ”OK” if the file is correct and an informative message if it is not.

8. (15 points) Write a Python program to compute the raw distance between the first two sequences in a file like \texttt{dna.txt}. 
