

Name: _____

This matrix displays pairwise distances among three species.

	A	B	C
A	0	4	6
B	4	0	8
C	6	8	0

1. (10 points) Find the UPGMA tree and neighbor-joining trees associated with the distance matrix. Draw each tree to scale and indicate branch lengths with numbers. Indicate in your drawing if each tree is rooted or not.

2. (4 points) Many studies have shown that in a wide variety of settings for generating simulated molecular sequences, distance methods are less accurate (have smaller probabilities of estimating the correct tree) than alternatives including maximum parsimony and maximum likelihood. Provide one possible explanation for this.

3. (4 points) Distance methods do have one substantial practical advantage over maximum parsimony and maximum likelihood when the number of taxa in the estimation problem is very large. What is this advantage?

The following table summarizes the maximum log-likelihood values for each of three possible tree topologies in a four-taxon molecular data set for an alignment of 1200 nucleotide bases under a variety of likelihood models. The parameters include those associated with branch lengths of the tree and those associated with the substitution model of molecular evolution. The modifier Γ_4 means that the model included the 4-category discrete Gamma distribution for rate variation among sites and the modifier C means that the model partitioned sites by codon position.

Model	Tree Topology			# of Parameters
	1	2	3	
JC69	-2400	-2420	-2403	5
K80	-2350	-2365	-2348	6
HKY95	-2160	-2177	-2163	9
HKY95+ Γ_4	-2092	-2105	-2094	10
HKY95+ Γ_4 +C	-1880	-1897	-1881	22
GTR	-2140	-2159	-2142	14
GTR+ Γ_4	-2070	-2093	-2074	15
GTR+ Γ_4 +C	-1865	-1885	-1867	37

4. (8 points) For each model, find the tree topology that is the maximum likelihood estimate.

5. (4 points) On the basis of AIC, which of these models would be the best to use? Support your answer with a brief numerical calculation.

-
6. (4 points) Briefly explain in concept how one would use the bootstrap to assess the confidence in the maximum likelihood estimate tree topology using the best model from the previous problem by the AIC criterion.
7. (2 points) The bootstrap and Bayesian inference use different methods to produce collections of trees, which are then summarized. In one case, each tree is associated with a different data set. In the other, each tree is associated with a different set of parameter values for the same data set. Which is which?
8. (10 points) Bayesian phylogenetic inference is typically implemented by sampling trees using Markov chain Monte Carlo, but an alternative possibility would be to calculate the probability of the data for each tree topology (averaging over other parameters) and compute the posterior probability for each tree topology using Bayes' Theorem. If one assumed a uniform probability distribution over the three possible tree topologies and the average log-likelihoods (natural log) using the GTR+ Γ_4 +C model above were -1900 , -1918 , and -1903 , what would be the posterior probability of the first tree topology?