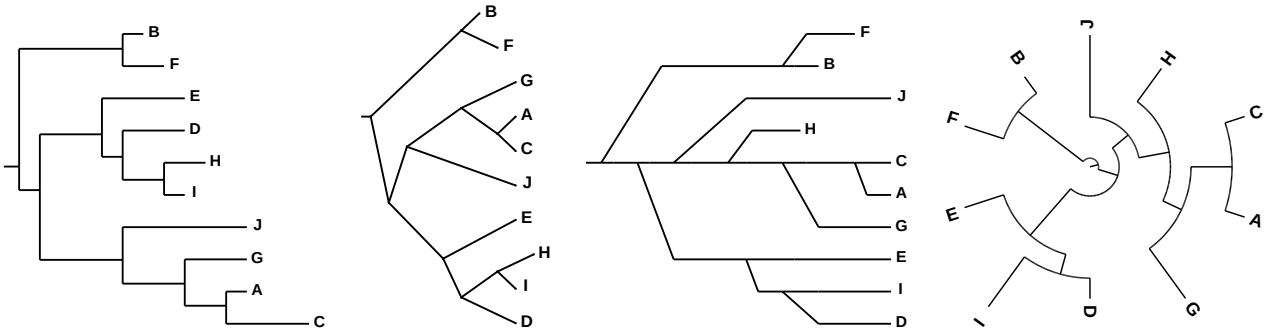
Genome 570
Phylogenetic Inference
Winter, 2016
J. Felsenstein

Name:_____
Address:_____
Address:_____

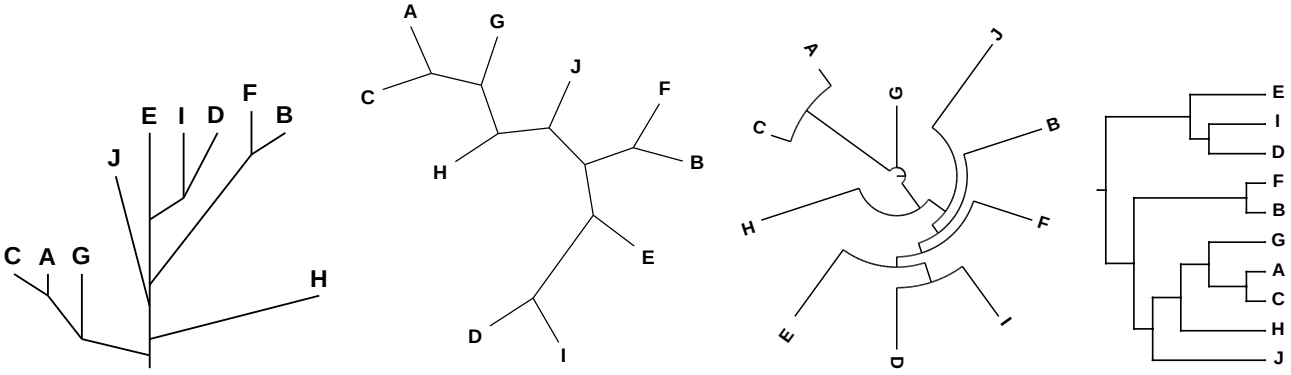
This exam is closed-book and totals 200 points. You can use a calculator or phone/calculator if you need to; if you don't have one you can instead leave the calculations as expressions such as $(3.67 \times 234)/1243 + 4.5$. Make sure to put your name on each page. Showing your work may help you get partial credit if the final answer is wrong. **To allow me to return your exam (with grade and course points) please fill in an address to mail it to** (preferably a campus box number, if not, an off-campus address).

1. (30 points) (**Basic tree literacy**)

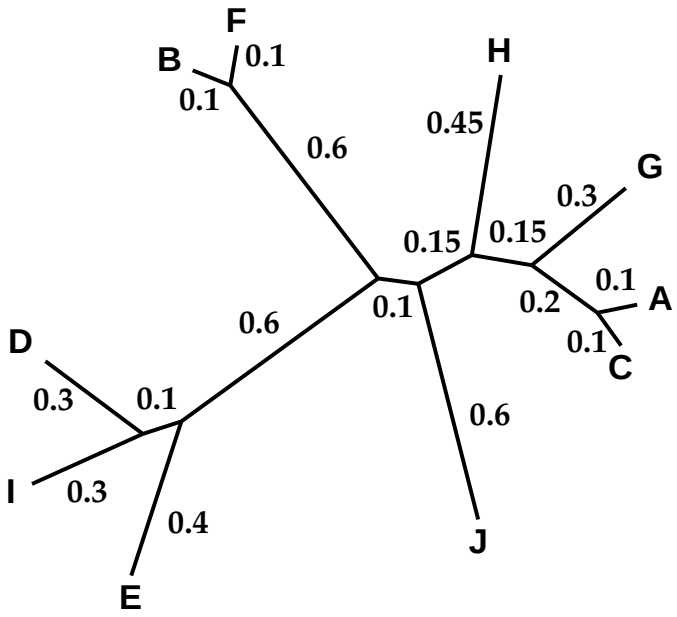
   (a) These four trees, drawn variously, are all rooted. Considered as rooted trees, there are only two different trees here. Indicate which ones are the same tree topology as each other. That means that you will divide them into two sets of trees.

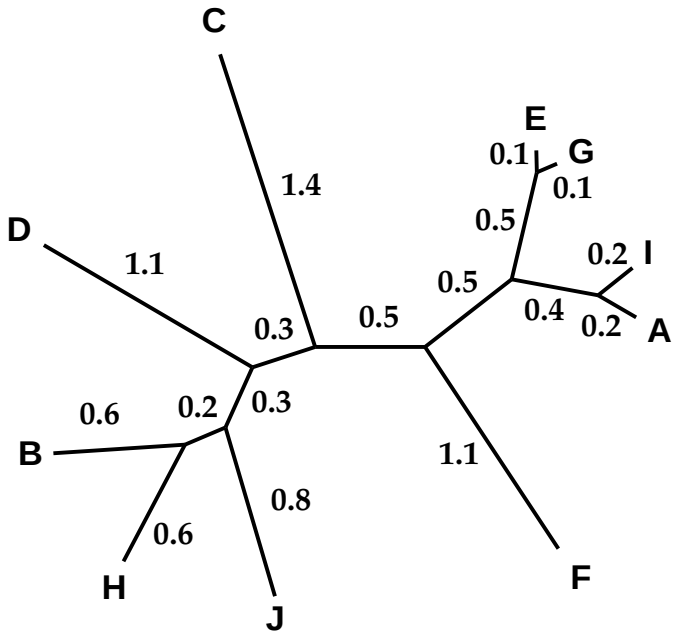


   (b) Consider these four trees as unrooted trees (even if some of them are currently rooted). Which ones are the same unrooted tree topologies?



(c) Here are two unrooted trees with branch lengths (drawn in various styles). Find a place on each tree that, if the root is there, the tree is a clocklike, ultrametric tree – one in which all tips are equidistant from the root. Each of these trees can be rooted so as to be clocklike. If this involves designating a point in the interior of a branch, make sure to note how much branch length there will be in each of the two resulting branches, one of them on each side of the root.





2. (35 points) Suppose that we are considering lineages of cancer cells in an individual, that have a branching tree as their genealogy (they do not mate or have recombination). We have genomic information on 9 cancer cells, which allows us to ask whether particular regions of their genome have deleted stretches of DNA. We find that in each of 7 regions there is one deletion, each present in some of the cells. By comparison with the individual's normal, noncancer DNA, we can see that the deletions arose in the cancer lineages, so presence is ancestral. We code presence as 1 and absence as 0 for each of these 7 regions. The names of the cells are arbitrary and do not indicate any order on their genealogy. The resulting data, including one noncancer cell, is:

normal	1	1	1	1	1	1	1
A	1	0	1	1	1	1	1
B	1	0	1	1	1	1	1
C	0	1	1	0	1	1	1
D	0	1	0	0	1	1	1
E	1	1	1	0	1	1	0
F	0	0	0	0	1	0	1
G	1	1	1	1	0	1	1
H	0	1	0	0	0	0	1
I	1	1	1	1	1	1	0

For these data, do the following. It will be best to work on the back of a page, so as to be able to write the results neatly on this page.

(a) Find a pair of characters that are not compatible, in the sense that they could not both evolve on the same tree with each changing only once. Tell me which characters they are and how you decided that they were not compatible.

(b) Make a compatibility matrix for all of the characters. Make clear to me how you are indicating that two characters are compatible (which symbol you are using).

(c) Make a graph whose points are the 7 characters, with points being connected when those two characters are compatible. Number the points so I know which character corresponds to each point.

(d) What is the largest clique?

(e) What tree is indicated by the characters in that clique?

(f) If we use that tree and have all characters change on it, how many changes are needed in each character?

(g) In effect, what parsimony method are you using when you count the changes of state?

3. (35 points) In distance methods on DNA data, one often sees people using the "p-distance" which is simply the fraction of sites that differ between the two sequences.

   (a) In effect, what are they assuming about how the sequences change?

   (b) Are they assuming that sites change independently? Why or why not?

   (c) How do their assumptions differ from those of the Jukes-Cantor model?

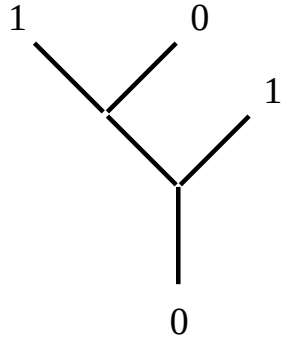
   (d) If you use p-distances in practice, you will be severely chastised by me. What (justified) criticisms do you think I could make, and why?

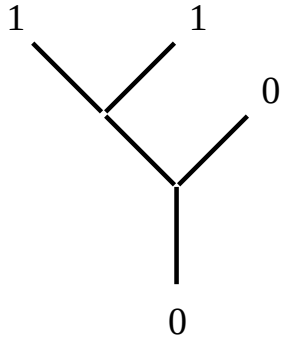
   (e) In what evolutionary situation might it not matter much whether you used p-distances or Jukes-Cantor distance? Explain.

4. (30 points) Suppose that we have a character with two states, 0 and 1. The state 0 is the ancestral state. State 0 can change to state 1, but state 1 cannot change back to state 0.

   Suppose that we consider two trees in which all branches are the same length, in the sense that in each branch the probability of a 0 changing to a 1 is p, and the probability of it remaining 0 is $1 - p$.

   For each of these two characters, calculate the likelihood for that character shown on the tree, given the states shown at the tips and at the root. These will be two expressions in p. This is not hard, since there are a very limited number of possible combinations of character states at the internal nodes of the trees.





5. (35 points) Indicate T (true) or F (false) by circling one for each question. If you feel that the answer needs explanation, use the space below on the page.

(a) In Bayesian inference of phylogeny we calculate probabilities for different possible trees that the tree is the true one.	T	F
(b) In Bayesian inference of phylogenies all tree topologies should have the same prior probability.	T	F
(c) Most Bayesian phylogeny programs use Markov Chain Monte Carlo (MCMC) methods.	T	F
(d) If we start from different initial trees, Bayesian MCMC programs are expected to give very different posterior distributions of trees.	T	F
(e) Posterior distributions in Bayesian methods typically have more variability than the initial prior distribution does.	T	F
(f) To get information about the uncertainty of the presence on one branch in the tree in a Bayesian MCMC run, it is necessary to do a bootstrap analysis.	T	F
(g) If we take 1000 trees sampled from a Bayesian MCMC run, and make a Majority Rule consensus tree for them, a branch that is seen to occur 80% of the time is inferred to have a probability of 80% of being a branch on the true tree.	T	F

6. (35 points) Suppose that we sample 2 gene copies of a particular autosomal gene from the nuclear genome from population #1, and 3 copies from population #2. Each population consists of 1,000 individuals. Suppose that in each generation, about 2 individuals in each of these populations is a newly arrived migrant from the other population.

   (a) If there were no migration, about how many generations back would we need to go before the two lineages in population #1 coalesce? Why?

   (b) In the absence of migration, do we have a higher or a lower probability, as we go back one generation, that there will be a coalescence in population #1 as compared to population #2 ? Why?

   (c) With migration allowed, about how many generations back will we need to go before we see one of the lineages in population #1 be now be found in population #2 ?

   (d) Based on these numbers, which is more likely to be the first event that we see as we go back in time in population #1, a coalescence or a migration? Why?

   (e) Based on that, do we expect the present location of these gene copies to be an accurate indication of their genealogical relationship? Why or why not?