This exam is closed-book and totals 200 points. You can use a 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 as the pages may be separated for grading. If you do any work on the blank backs of the pages, it will be best to work on the back of the same page, so that this work does not become separated from the question during grading. Showing your work may help you get partial credit if the final answer is wrong.

1. (39 points) **Basic Tree Literacy** (Arboricity? Dendricity?). Answer the following multiple-choice questions by circling the correct tree:

(a) One of these trees is not the same tree as the others. Which?

(b) One or more of the unrooted trees on the right are the same unrooted tree as the rooted tree on the left. Which one or ones?

(continued on next page)
(c) The midpoint rooting of a tree is rooting it at the middle of the longest path in the tree, the sequence of branches between the two tips that are farthest from each other. Which of the four trees on the right is the midpoint rooting of the unrooted tree on the left?
2. (36 points) For these sequences (with two sites)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha</td>
<td>G</td>
<td>T</td>
</tr>
<tr>
<td>Beta</td>
<td>T</td>
<td>G</td>
</tr>
<tr>
<td>Gamma</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td>Delta</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

and these three trees:

Showing your work if possible, so that partial credit can be given,

(1) Compute for each of the three trees the total number of changes in a parsimony analysis, using the Fitch algorithm.

(2) Compute the number of changes of state in a transversion parsimony analysis. This is a version of parsimony in which a change is counted only when a purine (A or G) changes to a pyrimidine (C or T) or vice versa. While you could compute this using the full Sankoff algorithm, an easier way is to convert the sequence symbols to R (purine) or Y (pyrimidine) and then use the Fitch algorithm.

(3) In each case, what is the best estimate (from the point of view of parsimony, for the ancestral state)?
3. (30 points) “Most esteemed Professor Felsenstein Joe: I am humble student Mr. Nietsneslef Eoj, of the Institute of Applied Semantics in Ruritania. I have collected 55 restriction sites which have been scored 0 (absent) or 1 (present) in 48 agricultural strains of crabgrass. I would like to do distance matrix analysis of their relationship. Will it be good to compute the distance simply as the fraction of sites differing? If not, what method can I use? Is it a problem if these are not distinct species but are strains?”

Help me advise this poor fellow, giving reasons for your opinions.
4. (30 points) A lab has DNA sequences from their favorite gene, and also protein sequences for the same species for that protein. They want to analyze these in a way that does not lose the insights that the protein sequence might give into the deep splits in the tree, nor the insights that the DNA might give into the relationships of more recently diverged species. For each of these three suggestions, comment on its suitability, pointing out any potential problems:

1. Compute a DNA distance matrix from the DNA sequences, using an HKY model, and also compute a protein distance matrix from the protein sequences, using a distance based on a JTT model. Average the two distances for each pair of species, and do a Neighbor-Joining tree.

2. Place the sequences end to end and do a parsimony analysis, where the bases and the amino acids represent the states, with changes allowed between all states with one step.

3. Use a likelihood or Bayesian analysis codon model, with the DNA sequences used to specify at the tips which codon is used by each sequence (in the coding region of the gene).
5. (30 points) For the following 5 (rooted) trees, compute an unrooted Majority-Rule consensus tree, and draw it.

If we did this for 1,000 trees, derived from 1,000 bootstrap samples from a data set of morphological characters such as bone measurements, which of the following worries should we have in interpreting the $P$ values as support of that clade? Explain very briefly why or why not for each.

- Bias of the bootstrap $P$ value. Yes: [ ] No: [ ]

- Multiple-tests problem. Yes: [ ] No: [ ]

- Correlated evolution of different characters. Yes: [ ] No: [ ]

- Different rates of evolution of different characters. Yes: [ ] No: [ ]
6. (35 points) In a rooted tree of two species, consider the coalescent tree connecting four copies of a gene, two sampled from each species (call these $a_1$ and $a_2$, $b_1$ and $b_2$). The effective population size of each species is $N$, and this is also the effective population size of their common ancestor. If the species diverged $t$ generations ago, what is the probability of the coalescent tree of these four copies having the species be “reciprocally monophyletic”? That means showing the tree topology $((a_1,a_2),(b_1,b_2))$. 

Careful! They could show this by being down to two lineages when they get back to the speciation, or by having three then, or even by having all four get back there without coalescing (and then happening to coalesce in the right way). (Hint – once you are back in the ancestral species the coalescent times don’t matter in answering this question, all that matters is the probability of the coalescent events being the right ones). *Show your work and I’ll show mercy.*