Short problems:

1. For this problem please draw LARGE trees and label the tips H, B, C, G, O for human, bonobo, chimp, gorilla, orangutan.
   (a) Draw a rooted tree which expresses the statement "Chimp and bonobo (pygmy chimpanzee) are most closely related. Humans and the two chimpanzees have a common ancestor more recently than their common ancestor with gorillas. Orangutans are the outgroup."
   (b) Draw an unrooted tree which corresponds to your rooted tree (it expresses the same relationships).
   (c) Starting with your unrooted tree, draw a *different* rooted tree which would also correspond to it. (In other words, show the rooted tree that would be produced if you moved the root to a new location.) Be sure your second tree is genuinely different, not just a different drawing of the same tree.
2. We draw a maximum likelihood tree of monkeys based on DNA sequences of several dozen unrelated protein coding genes. In this tree, the siamang is a typical monkey, closely related to several others. We also draw a tree based on gene PLP alone. In the PLP tree, siamangs appear only distantly related to other primates, and the branch leading to siamang is much longer than any other branch in the tree.
   (a) Give at least three plausible reasons for this discrepancy.
   (b) For one of your suggested reasions, give an experiment or analysis that could test it, and explain what the results would mean. (This does not have to be a phylogeny experiment–you can use any techniques you find relevant.)
3. Primates and rodents both have two copies of gene family A, called A1 and A2. We sample A1 and A2 from humans, chimps, rats and mice.
   (a) Draw the rooted phylogenetic tree you would expect if A1 and A2 arose from an ancient gene duplication in the ancestor of all mammals and have evolved independently since. Don't worry about branch lengths. Be sure there are EIGHT tips in your tree; name each tip with species and gene copy (i.e. C-A1, R-A2, etc).
   (b) Draw the rooted phylogenetic tree you would expect if A1 and A2 arose in the ancestor of all mammals, but are next to each other on the chromosome and experience strong concerted evolution.
4. Two researchers study the history of a bird species using a coalescent approach. One uses DNA both from recently caught birds and from birds found in glacier ice 10K-20K years old. The other uses only modern samples, but has over twice as many total samples. The first researcher finds evidence for multiple population bottlenecks in the bird species, whereas the second finds evidence for only a single, recent bottleneck. Assume that both sequenced the same loci and used their computational tools correctly. Can you explain the discrepancy?

Long problem:

(This problem is inspired by my research but the details are fictional.)

Barrett's Esophagus (BE) is an abnormal state of the esophageal lining that sometimes develops into cancer. In this condition, the mutation rate in the affected tissue is much higher than normal. We perform whole-genome sequencing on multiple biopsies from individual patients with BE. Each biopsy contains about 1 million cells; roughly speaking, sequencing detects nearly all mutations that are present in a majority of the cells and a decreasing fraction of those that are present in fewer cells.

We are interested in drawing evolutionary trees of samples within a patient by comparing the BE samples with a blood sample, thought to fairly accurately represent the patient's germ line state. Our biopsies turn out to have thousands to tens of thousands of mutations each, which should be ample data.

1. What hazards do you see in using whole-biopsy sequencing? (At this time single-cell is not technically feasible for these data.)

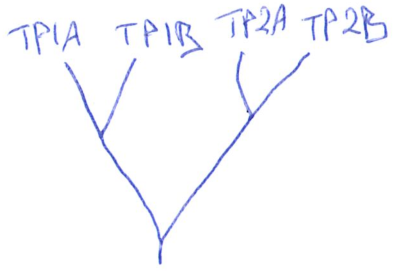
2. Mutations in $TP53$ are a strong predictor of cancer in this system. Some of the research team members argue that we should therefore draw trees based on $TP53$ mutations, using other genes only to break ties. Others argue that it would be better to use the whole genome, including non-coding regions and genes irrelevant to cancer. How would you argue?

One theory of cancer development predicts that the entire BE segment will experience selective sweeps based on favorable mutations. To test this, we use two pairs of biopsies, one collected six years ago, and one collected from the same patient today. Call the two old samples TP1A and TP1B, and the new ones TP2A and TP2B (TP stands for "time point").

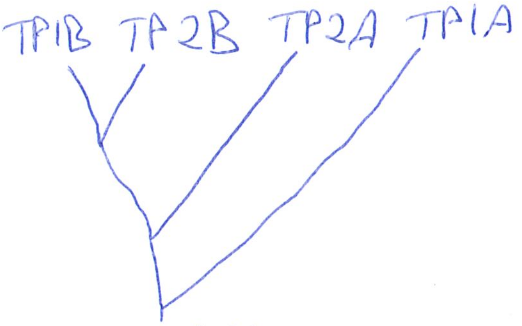
We count mutational differences between the four BE samples and the B sample. For Patient 1 this yields the following distance matrix:

	TP1A	TP1B	TP2A	TP2B
TP1A	–	418	435	223
TP1B	418	–	374	440
TP2A	435	374	–	457
TP2B	223	440	457	–

3. Construct a UPGMA tree using these data.
4. Is UPGMA an appropriate algorithm? Why or why not? Be specific.
5. Is the use of raw counts of differences appropriate? If not, how can we improve?
6. Is the tree you drew supportive of the hypothesis that there has been a complete selective sweep of the BE segment between TP1 and TP2? Explain briefly.
7. For Patient 2, we repeat the same experiment and run a bootstrap analysis on the results. In 1000 bootstrap replicates we see only 2 different trees, shown below:



612 times



388 times

How would you interpret this outcome? In particular, can you say anything about the probable nature of biopsy ~~T1B~~ TP1B in this patient?

8. Why are we assuming that genes on different chromosomes have the same phylogeny in this case?