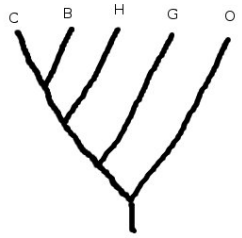
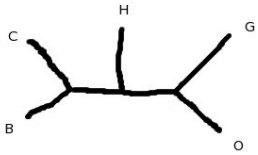
Short problems:

1. For this problem please draw LARGE trees and label the tips H, B, C, G, O for human, bonobo, chimp, gorilla, orangutan.

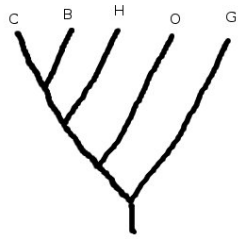
   (a) (1 pt) Draw a rooted tree which expresses the statement "Chimp and bonobo (pygmy chimpanzee) are most closely related. Humans and the two chimpanzees have a common ancestor more recently than their common ancestor with gorillas. Orangutans are the outgroup."



   (b) (1 pt) Draw an unrooted tree which corresponds to your rooted tree (it expresses the same relationships).



   (c) (2 pts) 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. *There are many solutions, including this:*



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) (3 pts) Give at least three plausible reasons for this discrepancy. *The siamang gene is a pseudogene, or has adapted to a new function, or is a duplicate, or has overdominance. This is a multi-gene family and we grabbed a paralog from siamang rather than a homolog. The siamang gene is hyper-mutable for some reason (hot spot? transposon integration?) The siamang gene has come in via horizontal gene transfer from a more distantly related species.*
   I took points off for any hypothesis that would support a shorter, rather than longer, branch to siamang, like the gene being highly conserved in siamangs or having a lower mutation rate.

   (b) (2 pts) 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.) *You could search the genome for other members of a putative multi-gene family; if they are present, you likely have the wrong one. You could use dN/dS or HKA to test if the siamang gene is likely to be under selection and if so, what kind. If it is neutral or under balancing selection that could explain your result. You could knock out the gene and see what happens: if nothing happens, it may be a pseudogene. You could sequence more of the genome and see if your multi-gene result really holds up. You could look for other species with a siamang-like PLP which might have participated in the putative horizontal gene transfer.*

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) (1 pt) 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). *The tree consists of two mirrored sides, one showing the expected human/chimp and mouse/rat clades for gene A1, and the other showing the same clades for gene A2.*

   (b) (1 pt) 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.*It looks like a single tree with human/chimp and mouse/rat clades, and then at the tip, the two gene copies from each species are nearest neighbors. I won't bother making a diagram as everyone got this question right.*

4. (2 pts) 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? *It is very hard to look back past a bottleneck with only contemporary DNA sequences, because very few lineages from before the bottleneck are still present–maybe only one, in which case nothing before the bottleneck can be seen. So I would expect that contemporary data would never show more than one bottleneck, and would tend to believe the ancient-DNA data as long as it had an adequate sample size. The ancient DNA data may include multiple sequences from before the most recent bottleneck and thus let us look back further in time.*

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. (2 pts) What hazards do you see in using whole-biopsy sequencing? (At this time single-cell is not technically feasible for these data.) *The biopsy is a mixture of cell lineages, not necessarily a single lineage. Worst case is that it was taken on the edge of two very different lineages and will contain two very different groups of cells. In this case, the data are not tree-like and any tree we draw will be questionable.*

   Also, it will be hard to phase mutations (are two mutations present in 20% of cells each present in the same cells, or different ones?)

2. (2 pts) 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? *Since TP53 is probably strongly selected in this system, it is about the worst gene for drawing a tree meant to represent the ancestry of the cell samples as a whole. We're much more likely to see convergent evolution than for an irrelevant gene. Also, using the whole genome is vastly more data.*

It is worthwhile to look at TP53 to see what story it is telling, but for phylogenetic analysis it is not a good choice. (Also, in the real data about half the patients don't have any mutations in TP53, so the tree could not be resolved.)

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. (3 pts) Construct a UPGMA tree using these data. *Samples TP1A and TP2B connect to each other, with branch lengths of 111.5 to their common ancestor. Then TP2A and TP1B connect to each other, with branch lengths of 187. Then–THIS IS THE PART SOME STUDENTS MISSED–the two pairs join to each other at the root. We have a total distance of 218.75 from root to tip. For the side leading to TP1A and TP2B, we have used up 111.5 of that, leaving a branch length of 107.25. For the side leading to TP1B and TP2A, we have used up 187 of that, leaving a branch length of 31.75. The common error was to make those interior branches both length 218.75, but if you look carefully at the resulting tree you will see it is far from clocklike, nor does it generate distances similar to the ones in the original data.*

4. (2 pts) Is UPGMA an appropriate algorithm? Why or why not? Be specific. *No. UPGMA assumes a clock and works badly if there is none. These samples were taken years apart so there is no reason to expect the same amount of evolution leading to the older and younger samples. Furthermore, these are pre-cancer cells with unstable genomes and the assumption that they all have the same mutation rate is highly questionable.*

5. (2 pts) Is the use of raw counts of differences appropriate? If not, how can we improve? *It would be better to use a mutational model that takes into account the chance of multiple hits; ideally, one fine-tuned to cancer data.*

6. (2 pts) 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. *No. If there was a complete sweep between TP1 and TP2, we would expect the two TP2 samples to be closely related as they both descend from the sweeping lineage. Instead, there were two distinct lineages present at TP1 and both are still present at TP2, which is not suggestive of a sweep.*

7. (2 pts) 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:

How would you interpret this outcome? In particular, can you say anything about the probable nature of biopsy T1B in this patient? *The trees look very well supported for everything but the position of T1B. I suspect this biopsy has cells from two different evolutionary lineages in it: some bootstrap trees group T1B with one of its components and some with the other. As it is clear that the data are not tree-like, I would avoid interpreting this tree further. (One of the patients in the real data looks exactly like this....)*

8. (2 pts) Why are we assuming that genes on different chromosomes have the same phylogeny in this case? *The cells don't reproduce sexually, so there is no reassortment of chromosomes: everything should be fully linked and reflect the same line of descent (with caveats about our samples not necessarily being just one lineage each).*