Solutions to Problems from Chapter 6

6.1(1). The CF tree T has $p(e) = 0.1$ for every edge e. The probability of $A = B = C = D$ is the sum over all ways of setting the internal nodes E and F of obtaining this outcome. We root the tree at A and let the internal node below A be E, and its children be B and F.

Then $pr(A = B = C = D = 0)$ is the sum of the following four terms:

- $pr(A = B = C = D = 0|E = F = 0)$.
- $pr(A = B = C = D = 0|E = 0, F = 1)$
- $pr(A = B = C = D = 0|E = 1, F = 0)$
- $pr(A = B = C = D = 0|E = F = 1)$

We analyze each one separately. Since every edge has the same probability of change, the only thing we need to know is the number of edges that have changes on the tree.

(1) $pr(A = B = C = D = 0|E = F = 0)$. Note that this has no change on any edge, and so is $(0.9)^5$, but multiplied by 0.5, the probability of $A = 0$.

(2) $pr(A = B = C = D = 0|E = 0, F = 1)$. This has change only on three edges, and no change on two edges. Therefore the probability is $(0.9)^2(0.1)^3$, multiplied by 0.5, the probability of $A = 0$.

(3) $pr(A = B = C = D = 0|E = 1, F = 0)$. This has change only on three edges, and so has the same probability as for case (2), i.e., $(0.5)(0.9)^2(0.1)^3$.

(4) $pr(A = B = C = D = 0|E = F = 1)$. This has change on four edges and no change on one edge. Therefore the probability is $(0.9)(0.1)^4$, but multiplied by 0.5, the probability of $A = 0$.

Therefore, the sum is $(0.5)[(0.9)^5 + 2(0.9)^2(0.1)^3 + (0.9)(0.1)^4]$.

Simplifying this we get $(0.5)[0.59049 + 0.00162 + 0.00009] = (0.5)(0.5922) = 0.2961$.

6.1(2). The CF model tree has the same topology as before (AB on one side and CD on the other), but has different substitution probabilities. Now the internal edge has $p(e) = 0.4$ and all other edges have probability 0.001.

1. We are asked to compute the probability of each of the parsimony informative patterns at the leaves. We do the calculation as before.

   - $pr(A = B = 0, C = D = 1)$. This is the sum of four possibilities, for the internal nodes E and F (defined as before).

     a) $E = F = 0$. Then there is change only on two external edges, and no change on the other three edges. The probability is $(0.5)(0.001)^2(0.6)(0.999)^2$. This is approximately 0.

     b) $E = 0, F = 1$. Then there is change on the internal edge but nowhere else. The probability of this is $(0.5)(0.999)^4(0.4)$, or approximately 0.2.

c) $E = 1, F = 0$. There is change on every edge. The probability of this is $(0.5)(0.1)^4(0.6)$, or approximately 0.

d) $E = F = 1$. There is change only on two external edges. The probability is $(0.5)(0.6)(0.001)^2(0.999)^2$, or approximately 0.

Summing these up, we obtain approximately 0.2.

- $pr(A = C = 0, B = D = 1)$. This is the sum of four possibilities, with internal nodes E and F as before.

  a) $E = F = 0$ or $E = F = 1$. Then there is change on two external edges but not on any other edges. The probability for each is $(0.5)(0.001)^2(0.999)^2$. Hence, the probability for these two events together is $(0.001)^2(0.999)^2$. Simplifying this is approximately 0.

  b) $E = 1, F = 0$ or $E = 0, F = 1$. For each of these cases there is change on the internal edge and on two of the four external edges, and no change on the other two external edges. Hence the probability for each is $(0.5)(0.4)(0.001)^2(0.999)^2$. Hence, the probability for these two events together is $(0.4)(0.001)^2(0.999)^2$. Simplifying this is approximately 0.

  Therefore, the total probability is approximately 0.

- $pr(A = D = 0, B = C = 1)$. It is clear that this is the same as for (b), hence approximately 0.

2. We are asked if maximum parsimony would be statistically consistent on the tree. The answer is *yes*, but the reasoning is slightly subtle. The point is that of the three parsimony informative patterns, only $A = B = 0, C = D = 1$ and $A = B = 1, C = D = 0$ have probabilities that are anything but extremely small. Since the parsimony uninformative patterns have no impact on maximum parsimony, this is all we care about. As the number of characters increases, with probability going to 1, the number of parsimony informative patterns that have $A = B = 0$ and $C = D = 1$ (or the reverse, $A = B = 1$ and $C = D = 0$) will be larger than any other single parsimony informative pattern. Therefore, the result from maximum parsimony will be the tree on which $A = B = 0, C = D = 1$ is compatible. Since that is the same tree as the model tree, maximum parsimony is statistically consistent for this model condition.

6.1(3). We have four model trees, each with the same tree topology but with different branch lengths.

1. To answer this problem, we consider the probability that two leaves have the same state for T_1. For leaves A and B, the probability that $A = B$ is approximately 0.5, since the two leaves are separated by a path that contains an edge with substitution probability 0.499 (and we round this value). More generally, the probability that A and any other leaf share the same state is about 0.5, and the same can be said for C and any other leaf. On the other hand, the probability that B and D have the same

state is close to 1. Thus, the parsimony informative pattern with $A = C$ and $B = D$ being two different states has probability approximately 0.25.

Now we consider the parsimony informative pattern $A = B$ and $C = D$, and calculate its probability. There are four possibilities, based upon the internal nodes. Suppose that $F \neq C$. Then there will be change on edges e_C and e_D, and no mattern how E is set, the probability of this is less than 0.0001. Similarly, if $E \neq A$ then there will be change on edges e_A and e_B, and the probability of this (no matter how F is set) will be less than 0.0001. Thus, the only settings for E and F that have the possibility of having a non-negligible probability have $E = A = B$ and $F = C = D$. But then $A \neq C$ implies that $E \neq F$, and so there is a change on edge e_I. This has probability 0.0001, and so this pattern also has a negligible probability of occuring. Therefore, the parsimony informative pattern $A = B \neq C = D$ has probability below 0.0001.

The same analysis can be performed for $A = D \neq B = C$, showing that this pattern has probability below 0.0001.

Thus the tree T_1 would be more likely to produce $A = C, B = D$ of all the parsimony informative patterns.

2. The same analysis as given above shows that for T_4, the only parsimony informative pattern with probability more than 0.01 is $A = B \neq C = D$. Furthermore, the probability of this pattern is approximately 0.5. Hence, the tree T_4 would be more likely to produce $A = B \neq C = D$ than any other parsimony informative pattern.

3. Consider tree T_1. Each site generated under this model has probability about 1/4 of having absolutely no change occuring at all on any edge of the tree, and a slightly larger larger probability of not exhibiting any change between the leaves (i.e., reversing changes that occur). Therefore, the probability of a random site having at least one change is fairly close to 3/4. It is therefore not very likely at all that there will be no change at all on a dataset with 100 sites.

   Similarly, consider T_2. Every edge has probability close to 1/2 of having a change, which essentially make all the states at the leaves random with respect to each other. The probability of having no change at all in any site is about 1/8. Therefore, the probability of having no change in 100 sites is extremely small.

   For tree T_4, the probability of all leaves having the same state is at most 1/2. Hence, the probability of having all four leaves have the same state for 100 sites is at most $1/2^{100}$, which is extremely small.

   Finally, for tree T_3, every edge has a very low probability of change. For this tree, the most likely scenario is that all leaves have the same state, and this has probability close to 1. The probability that this is still true for 100 sites is not all that close to 1, but it's still better than the others.

Therefore, if we had four long sequences that were identical, we'd pick T_3, since it has the highest probability of occuring.

4. B and D are the same and A and C are almost random with respect to each other (they differ in 5 of 10 positions). This looks like tree T_1.