CS 598AGB
Problem set for February 4, 2016

Some of these problems are similar to the ones in Homework #2, and so solving these will help you do the problems assigned for the homework. Others will be similar to problems you will do for later homeworks.

For these problems, MP = maximum parsimony, MC = maximum compatibility, and FPM = Four Point Method.

1. Suppose we have the following character matrix:

   - r = (0, 0, 0, 0)
   - a = (1, 1, 1, 0)
   - b = (1, 0, 0, 0)
   - c = (1, 0, 1, 1)

   Answer the following questions:

   (a) What characters are informative for the MP criterion? Solve MP for this input, and draw the unrooted four-leaf tree with the best MP score. Show the 4-tuples for the internal nodes of the tree, and calculate the MP score of the tree.

   (b) Answer the same questions for the MC criterion.

   (c) Compute the Hamming distance matrix for this set. What is the first sibling pair computed by UPGMA? What is the unrooted tree computed by UPGMA?

   (d) Does the Hamming distance matrix obey the four-point condition? What does FPM produce for this input?

   (e) Consider the characters described by the different sites (positions). Is there a tree on which they could evolve without any homoplasy? What does that tree look like?

   (f) What did you learn by computing these trees?

2. Consider the following matrix:

   - r = (0, 0, 0)
   - a = (1, 0, 1)
   - b = (1, 0, 2)
   - c = (1, 1, 1)
   - d = (1, 1, 3)

   Answer the following questions:

   (a) Solve MP for this dataset.

   (b) How many binary trees have the best possible MP score?

(c) What is the MP score of the star tree (one internal node adjacent to all the leaves)?

(d) Can you find any non-binary trees that also have the same best MP score?

(e) Find a tree that has a best possible MP score that is minimally resolved – i.e., if you collapse an edge it will no longer have the best possible MP score. What is that tree?

(f) Answer the same questions as above, but for MC.

3. Recall that a character matrix is compatible if there is a tree on which all the characters are compatible. Suppose my matrix is

   - a = (0, 1)
   - b = (0, ?)
   - c = (1, 1)
   - d = (1, 0)

   Can you set ? to 0 or 1 and have the matrix be compatible? If so, what value works?

4. Suppose you have the matrix

   - a = (0, 1)
   - b = (0, 2)
   - c = (1, 1)
   - d = (1, 3)
   - e = (1, 4)
   - f = (2, 5)

   Find all the solutions to MP. Do *not* solve this by examining all possible trees on $\{a, b, c, d, e, f\}$. Do the same problem for MC.

5. Solve MP for the following set of four sequences:

   - u = AAAAAAAAAA
   - v = ATTTTTTTTT
   - w = GAAAAATATT
   - x = GTATCTACAC

   Explain your work. (Think about parsimony-informative sites.)

6. Consider a model CFN tree $ab|cd$ and internal nodes x and y (with x adjacent to a and b, and y adjacent to c and d). Let the substitution probabilities be:

- $p(a, x) = p(c, y) = 0.49$
- $p(b, x) = p(x, y) = p(d, y) = 0.01$

Without doing any serious calculations, think about the parsimony informative sites under the CFN model, and make an educated guess about the most probable parsimony informative sites for this tree. What tree would they support? Now let the number of sites increase to infinity. What do you think MP (solved exactly) would return with probability converging to 1? What does that tell you about MP and statistical consistency on this tree? Now do the same thing for MC, UPGMA, and the Four Point Method.

7. Consider a model CFN tree $ab|cd$ and internal nodes x and y (with x adjacent to a and b, and y adjacent to c and d). Let the substitution probabilities be:

   - $p(a, x) = p(b, x) = (c, y) = p(d, y) = 0.01$
   - $p(x, y) = 0.49$

   Without doing any serious calculations, think about the parsimony informative sites under the CFN model, and make an educated guess about the most probable parsimony informative sites for this tree. What tree would they support? Now let the number of sites increase to infinity. What do you think MP (solved exactly) would return with probability converging to 1? What does that tell you about MP and statistical consistency on this tree? Now do the same thing for MC, UPGMA, and the Four Point Method.