COMP 571: Homework #4
Spring 2017

Assigned on April 6, 2017.

Due in class on April 18, 2017.

The Honor Code applies to all homework sets. Sign the pledge on your solutions.

1. Only two types of nucleotides, C and G, are present in sequences x^1 and x^2 of equal length. Given the ungapped alignment of x^1 and x^2, calculate the likelihood of a tree relating these sequences, assuming the Jukes-Cantor model of evolution. Show that the maximum likelihood edge lengths, t_1 and t_2, satisfy the following equation:

$$t_1 + t_2 = \frac{1}{4\alpha} \ln \frac{3(n_1 + n_2)}{3n_1 - n_2},$$

where n_1 is the number of alignment sites with identical residues, and n_2 is the number of sites with mismatches.

2. In class we described Fitch's algorithm for bifurcating trees and for equal costs for substitutions (one unit for change from x to y, when $x \neq y$, and 0 when $x = y$). Describe the modified algorithm so that it applies for trees in which nodes have arbitrary numbers of children, and for a given cost matrix W, where $W(\sigma, \sigma')$ is the cost of a change from state σ to state σ'.

3. Consider tree $T = ((A : 0.8, C : 0.1) : 0.05, (B : 0.8, D : 0.1) : 0.05)$, and base frequencies $f(A) = 0.1$, $f(C) = 0.2$, $f(G) = 0.3$, and $f(T) = 0.4$. Further, let the substitution rate matrix be the following:

	A	C	G	T
A	-	5	6	2
C	5	-	3	8
G	6	3	-	1
T	2	8	1	-

   (a) Using the Seq-gen tool under the REV model and the above settings, and for sequence lengths 100, 250, 500, 1000, 5000, and 10000, generate 20 sequence datasets for each of the sequence lengths.

   (b) For each dataset S_i, construct the MP tree T^i_{MP} and the ML tree T^i_{ML} (using publicly available programs; e.g., Phylip), and compare them to tree T using the Robinson-Foulds (RF) measure (the RF distance between two trees is the number of bipartition's in one, but not both of the trees, divided by 2).

   (c) Plot a graph with two curves that show the RF value of the two methods as a function of the sequence length (for each sequence length, plot the average RF of the 20 runs of both methods). Explain the behavior of the two methods.

In this problem, submit the plot from Part (c), as well as your explanation of the behavior and potential causes of the behavior (feel free to search the literature after you generate the plot, but clearly cite any sources you use).

4. The paper "Population genomic analysis of outcrossing and recombination in yeast," by Ruderfer *et al.* (Nature Genetics, 38(9): 1077-1081, 2006) describes an HMM-based method for estimating the rate of recombination from sequence data. Describe formally the method that they propose and discuss the assumptions underlying the method. In particular, you need to describe the HMM formally, and discuss its underlying assumptions.