

Homework no. 6  
Due Friday, May 13

Using the two species (Chimp and Human) of the mitochondrial DNA data set (the one you have been using), I want you to infer the branch length separating them (in effect, infer a two-species tree) by maximum likelihood, using the Hidden Markov Model machinery with two rates of evolution. The programming involved is not hard but the formulas have to be gotten right. We will use (for simplicity) a Jukes-Cantor model of base change. The search for optimal values of the parameters will be done by you by hand (unless you want to get ambitious and write hill-climbing optimization code).

Here are some detailed steps.

1. The program will read in both sequences, and some parameters (see below) These will set up rates of change  $r_1$  and  $r_2$ , a probability of being in state 1 ( $\pi_1$ ), an autocorrelation parameter ( $\lambda$ ), and the branch length  $t$ .
2. It will set up an array for the HMM machinery, with  $2 \times n$  entries. The  $(i, j)$  entry is for the  $i$ th site, with the  $j$ th rate. Compute in it the joint probability of the two bases at site  $i$  using rate  $j$ . If the two bases are different this will be

$$\frac{1}{16} \left(1 - e^{-\frac{4}{3}r_j t}\right)$$

and if they are the same it will be

$$\frac{1}{16} \left(1 + 3e^{-\frac{4}{3}r_j t}\right)$$

3. The Markov chain we are to use to assign rates going from one site to another has a probability  $\lambda\pi_2$  of changing to state 2 when it is in state 1, and a probability  $\lambda\pi_1$  of changing to state 1 when it is in state 2.  $\pi_1$  and  $\pi_2$  are then the equilibrium probabilities of the two rates (one of these can be read in as a parameter).  $\lambda$  is the autocorrelation parameter (it must be in the interval  $[0, 1]$ ). When it is large the blocks of similar rates will be large (the block lengths will be related to  $1/\lambda$ ).
4. Use rates  $r_1$  and  $r_2$  where the higher one is severalfold higher than the lower one. Choose also values of  $\pi_1$  and  $\pi_2$  where one is not many times the other. Any values you use are OK with me as long as you tell me what they are.
5. After you read in the rates, rescale them by dividing by their weighted average  $\pi_1 r_1 + \pi_2 r_2$  so that average rate is 1.

6. Use the Backward Algorithm to compute the likelihood for the given values of  $t$ ,  $\pi_1$ ,  $\lambda$ ,  $r_1$  and  $r_2$  that you read in (the  $r$ 's having been rescaled as above). Leave behind in an array as you do, numbers allowing you to backtrack and find the single combination of rates that makes the highest contribution to the likelihood.
7. For a given choice of  $r$ 's and  $\pi$ 's, try different values of  $\lambda$  and  $t$ . Choose the values of  $\lambda$  and  $t$  that yield the highest likelihood.
8. Report the values of  $r$ 's and  $\pi$ 's that you used, the best estimates you find of  $t$  and  $\lambda$ , and the resulting log-likelihood.
9. Also report the combination of rates that made the biggest contribution to the likelihood (as a string of 1's and 2's).
10. As usual, submit your source code as well as an attachment.

This looks hard but the programming part is not so bad.

If you want to, you could include any thoughts you had on how to find a region of low rate of change on this molecule, as what you are doing is equivalent to the widely-noticed "PhyloHMM" machinery of Siepel and Haussler, only with just two species in this case.