

MS2a, Week 3

Rune Lyngsø

November 2, 2011

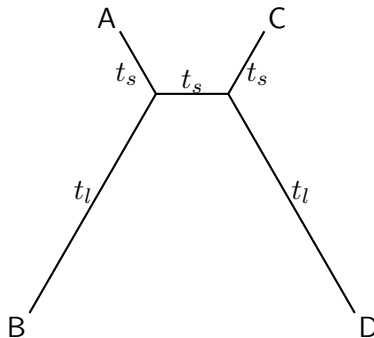
A Phylogeny Reconstruction

- a. Consider a binary character – for convenience denote the two possible states 0 and 1 – evolving according to the rate matrix

$$Q = \begin{bmatrix} -\alpha & \alpha \\ \alpha & -\alpha \end{bmatrix}$$

Determine $P(t) = e^{Qt}$.

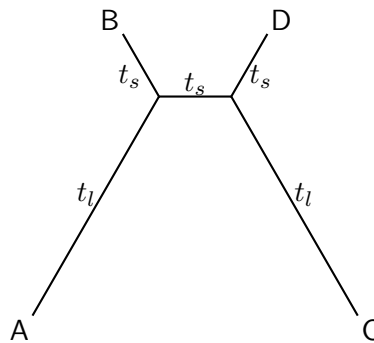
- b. Assume that we have a sequence of this binary character evolving on the following tree



with observed sequences A, B, C, and D. Let t_s be chosen such that $P(t)_{01} = 1/25$ and t_l be chosen such that $P(t)_{01} = 1/4$. What are the values of αt_s and αt_l meeting this requirement?

- c. What are the probabilities of observing each of the 16 possible combinations of the binary character at the four sequences, *i.e.* the probability of observing a 0 in all four sequences, a 0 in sequences A, B, and C and a 1 in sequence D, *etc.*?

- d. For sequence length $n \rightarrow \infty$ what tree would you expect to be preferred by the parsimony method, *i.e.* the tree requiring the fewest character changes when summed over all sequence positions?
- e. Write an expression in terms of the probabilities of the 16 possible character combinations (*i.e.* your variables will be $p_{0000}, \dots, p_{1111}$) that should be maximised to find the phylogeny the maximum likelihood method will converge to for $n \rightarrow \infty$?
- Without analytically solving for the MLE phylogeny, which phylogeny do you expect it to be?
- f. Assume now that half the positions of our sequence evolve on the tree above, and half the positions evolve on the following tree



that is the tree where A and C sit at the end of long branches instead of B and D. What is now the probability of observation of the 16 possible patterns of character states at the four sequences?

What is the tree expected to be preferred by the parsimony method?

- g. If you were told that the correct topology is in fact not the topology the maximum likelihood method will converge to, which of the alternate topologies would be your guess for the one the maximum likelihood method converges to instead?

B Recombination

- h. Can we find a tree for the data set

Pan	TTATCC
Gorilla	TTGTTC
Pongo	CCACCC
Hylobates	CCGTTC

such that only one substitution is required in each position? If yes, provide such a tree. If no, why not?

- i. Compute the minimum number of substitutions required for the above data set for each of the three possible unrooted tree topologies, e.g. by using Fitch's algorithm (or just eye-balling it if you feel confident about doing this).
- j. Assume that apart from substitutions, you are also allowed events arbitrarily changing the tree topology between consecutive sites (this is a simplification of recombination events – recombination events only allow certain changes to tree topology). What is the minimum number of events you need to explain the above data set.

Give an ancestral recombination graph explaining the data set with this number of events.

- k. How many recombination nodes are there in the ancestral recombination graph (ARG) you constructed in j?

Can you construct a data set by permuting the columns in the above data set that requires more recombination nodes for any ARG explaining it? If yes, give an example.

How many different marginal trees does the ARG you constructed in j have (a marginal tree is the tree relating the species at a particular position)?

Can you construct a data set by permuting the columns in the above data set that has more different marginal trees in any ARG explaining it? If yes, give an example.