

MS2a, Exercises Week 4

Rune Lyngsø

October 28, 2009

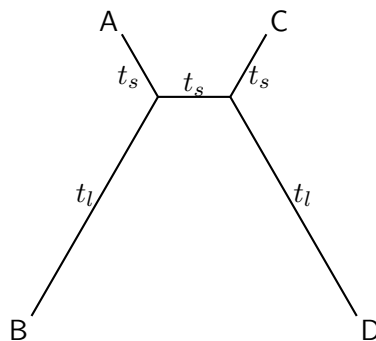
A Phylogeny Reconstruction

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

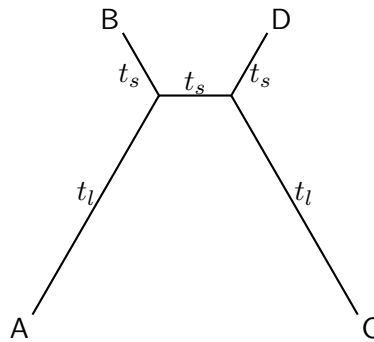
- b. Assume that we have a sequence of this binary character evolving on the following tree



with observed sequences A, B, C, and D. Let t_s be chosen such that $P(t)_{01} = 1/20$ 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 Alignment of Sequences

- a. Consider the three sequences

```

AGTCGGACAATGTC
GGCGAAAATGTA CTTC
GGCACAAGTCGTTCC

```

What is the longest sequence you can find that is a subsequence of *all* three sequences? For example, TTT is a subsequence of all three sequences, but CTTT is not (it is not a subsequence of the first sequence).

- b. What is the shortest sequence you can find such that all three sequences are a subsequence of it?
- c. Assume you are allowed three operations: changing one character (*e.g.* AGT \rightarrow ACT), deleting one character (*e.g.* ACT \rightarrow AC), and inserting one character (*e.g.* AC \rightarrow ATC). What is the best sequence you can find in terms of minimising the maximum number of changes required to obtain the three sequences above (also known as a median of the three sequences)?
- d. An alignment of a set of sequences is a matrix with one row for each sequence, where each entry contains either a character or a gap (usually depicted by $-$). When ignoring the gaps in a row, the remaining entries yields the corresponding sequence. Each column is required to contain at least one (non-gap) character. So an alignment of the above sequences could start as

```
A G T C G ...
G G - C G ...
G G - C - ...
```

How many different alignments are there of three sequences with three characters each?