Computational Biology HS 19 - Exercise 4
Answers to Theory Questions

Toby Law

20 November 2019

Question 1

Explain why there are two different probabilities at node 6 for A and G.

The child node of node 6 is tip node 3, whose identity is known to be A (i.e. likelihood of this site at node 3 being A is 1). So the likelihood of A at node 6 is higher due to this contribution from node 3. In contrast, none of the tip nodes that node 6 is ancestral to are G, and this is the main reason why likelihood of G at node 6 is lower, compared to that for A.

Question 2

After altering the tree with a nearest-neighbour interchange move, if you want to calculate the likelihood of the new tree using the same algorithm, can you reuse: none, some or all of the calculations you performed on the previous tree? Give a concise explanation.

Some of the calculations can be reused. If the node tips are interchanged, then some of the calculations of the directly connected parent nodes have to be redone. If the internal nodes are interchanged, the likelihood calculations done on them and their child nodes don't have to change as their connectivity is still preserved. We just have to redo the likelihood calculations for the nodes ancestral to them.

Question 3

List at least three key differences between the UPGMA tree reconstruction and the maximum-likelihood tree search.

UPGMA tree reconstruction is fast (polynomial time), distance-based, and assumes a strict molecular clock for evolution (evolution rate is constant).

Maximum-likelihood tree search is slow (exponential time) since we have to visit all trees in the tree space and experiment with branch lengths, assumes a probabilistic model of evolution, and allows for varying rates of evolution.

Question 4

How would the runtime of Felsenstein's algorithm change if you had 5 nucleotides instead of 4?

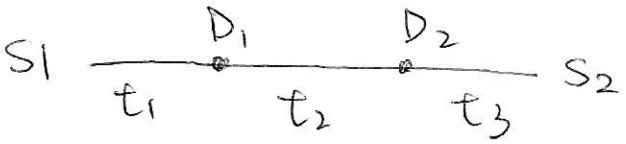
Runtime complexity of Felsenstein's algorithm is $O(nm)$, where n is the number of recursion steps/nodes, and m is the number of sites in the sequence alignment. If we had 5 nucleotides instead of 4, each recursion step/node will simply have the summation of one extra state, which does not change the runtime complexity of Felsenstein's algorithm.

Question 5

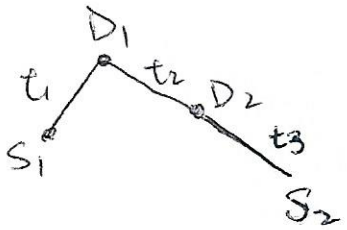
In lecture 4 we learn about the condition of "time reversibility". Could you place the root anywhere in your tree and still obtain the same likelihood if the substitution model was not time reversible?

No, the likelihood of the tree would differ depending on where the root is placed. If the substitution model is not time-reversible, it means that substitution rates between any two nucleotides amongst A,T,C,G will not be the same for two possible directions of substitution. eg. rate of A -> T substitution different from that for T -> A. We illustrate this with an example (Figure 1), where an unrooted tree has two possible roots, D_1 and D_2, and one site of the sequence at the two tips are s_1 and s_2. If we look at possibility 1, where D_1 is the root, the transition probabilities of going from internal state X to the observed nucleotide s_2 would be equivalent to first transitioning to another internal state at node D_2, then to s_2. When we look at possibility 2, where the tree is rooted at node D_2, now we have to consider the transition probabilities of the internal state X through another internal state at node D_1, then to s_1. Since if the substitution model is was not time-reversible, the substitution rates for any transition from node D_1 and D_2 would be different for that from node D_2 to D_1, so we would yield different likelihoods from these two possible trees.

Figure 1 Unrooted tree



Possibility (1)



Possibility (2)