Computational Biology
Assignment 4 - Report

Philip Hartout
phartout@student.ethz.ch

November 15, 2019

1. There is one reason why there are two different probabilities at node 6 for A and G, namely that A has occurred once in the sequence, hence increasing its likelihood.

2. The likelihoods of the subtree under the node that is interchanged is not going to be influenced by the calculations, because the sequence remains the same, as does the tree structure. Also, given K80 is time-reversible, the root of the tree will have the same likelihood, and the calculations leading to the root tree will not have to be changed.

3. Three differences are:

   (a) UPGMA have more stringent assumptions than Maximum likelihood methods, for instance, it assumes the presence of a strict molecular clock, which is often violated.

   (b) UPGMA is only valid for ultrametric trees, which means it can only be made from sequences sampled at one point in time.

   (c) UPGMA makes use of distance metrics between data points, whereas maximum likelihood-based trees make use of an evolutionary model.

4. The running time would still be in $\mathcal{O}(n)$, because summing over five states twice is still accomplished in constant time.

5. We cannot place the root anywhere in the tree and still obtain the same likelihood if the substitution model is not time reversible, because, suppose the likelihood of the tree starting in root D_1 is provided by:

$$P(D_1) = \sum_{X \in \mathcal{N}} \pi_X p_{X,s_1}(t_1)$$

   where $s_i \in \mathcal{N} = \{T, C, A, G\}$. We would then require the equality $p_{X,s_2}(t_2 + t_3)$ to equate $P(D_1)$ to $P(D_2)$, as provided below:

$$\begin{aligned} P(D_1) &= \sum_{X \in \mathcal{N}} \pi_X p_{X,s_1}(t_1) \sum_{Y \in \mathcal{N}} p_{X,Y}(t_2) p_{Y,s_2}(t_3) \\ &= \sum_{X \in \mathcal{N}} \sum_{Y \in \mathcal{N}} \pi_X p_{X,s_1}(t_1) p_{X,Y}(t_2) p_{Y,s_2}(t_3) \\ &= \sum_{X \in \mathcal{N}} \sum_{Y \in \mathcal{N}} \pi_X p_{X,Y}(t_2) p_{X,s1}(t_1) p_{Y,s_2}(t_3) \qquad \neq P(D_2) \end{aligned}$$