

# Computational Biology

## Assignment 3 - Report

Philip Hartout  
phartout@student.ethz.ch

November 10, 2019

1. If we take two very different sequences, we get an undefined distance metric, because in the indicated case,  $V = \frac{1}{3}$  and  $S = \frac{2}{3}$ . In turn,  $\frac{1}{2}\log(1 - 2 \cdot S - V) - \frac{1}{4} \cdot \log(1 - 2 \cdot V)$  becomes undefined because the log of negative values does not exist.
2. My implementation of the UPGMA algorithm will probably be influenced by the order in which the sequences are given, because my implementation can only merge one node at once, and in the case of multiple minimal values in the distance matrix, this will yield the first such smallest value.
3. There are several features that may contribute to such a discrepancy:
  - (a) It is likely that true evolution does not follow a strict molecular clock
  - (b) It can also be that the sequences have not been sampled at the same evolutionary time.
4. Non-ultrametric trees constructed by means of the neighbour-joining algorithm allow for rate changes in substitutions, which can remedy the fact that the strict molecular clock assumption is violated.