Computational Biology HW 3

Probst Jennifer, 16703423

7. November 2020

Question 1

With many differences in the sequences we get high S and V values. Calculating now the distance we might have to calculate the logrithm of negative values which is not defined.

Question 2

There can be two cases with equal minimal entries. Either you have two pairs of distinct sequences that have minimal distance and in this case the pairs will be two cherries in the tree; thus there is no influence which one is picked first. In the 2nd case for example seq a and seq b have the same seq distance as seq a and seq d. But then the distance from seq b to seq d has to be the same minimal distance aswell, meaning the distance from the two merged sequences to the remaining one is the same minimal one. This results in a polytomy (thus no influence on the order we attach the roots).

Question 3

UPGMA is a phenetic approach, which only uses pairwise differences and does not consider higher order correlations between sequences. Secondly, in the UPGMA algorithm it is assumed that our genetic data evolved according to a strict molecular clock, which means that substitution rates on each branch were always the same. Depending on the dataset these simplifications may result in inaccuracies.

Question 4

The first problem is a general problem of phenetic approaches. If we want to consider higherorder correlations between sequences, we have to use a different approach type e.g. a cladistic approach with the Fitch algorithm; which implicitly accounts for an evolutionary process. The neighbor-joining algorithm considers non constant evolutionary rates, which can be better suited for some datasets.