Sequence Bioinformatics
Assignment 04

Emil Paulitz, Raphael Olipitz

November 27, 2020

Task 2

We used maxiters = 16 to improve accuracy, and because the dataset is not too large, this is computationally not too costly.
Since our sequences are higly similar, we used diagonal optimization, and for tree-building we used the UPGMB algorithm because it is a mix of UPGMA and NJ, thus probably leading to a good result.

Task 3

The only option was whether gaps should be treated as missing or additional characters. We chose as additional character, since most of our sequences are complete and gaps rather point towards an indel than to missing data.

Task 4

Since this algorithm has many options and we do not have special insight in our data set except for its similarity and small size, we mostly chose the default.
Number of datasets is 15, datatype obviously amino acids, and we chose the default evolutionary model LG. The options are different models that result in different matrices to estimate the mutation probabilities between the amino acids. Since our data set is not too large, we used the model frequencies option.
As tree topology search algorithm, we chose option best, since it offers the best of the other two options and performance is not really an issue with our data set.
Before running the algorithm, we renamed the sequences since PhyML does not support ':'.

Task 5

See the tanglegram.png