Tutorial Assignment 1 - Phylogenetic Tree

Name: Poorva Pisal
Roll no.: 2020113001

Introduction

This assignment is used to create phylogenetic relationships between the given protein and nucleotide sequences. I have written the programs in jupyter notebook and have used the numpy, pandas and the biopython libraries. I have written the code corresponding to the UPGMA method which was taught in class.

Question 1

This question involves constructing a phylogenetic relationship for the given nucleotide sequences in Nucleotide.txt.

a) This part deals with writing a script (q1a.ipynb) to generate a distance matrix file (Ndistance.txt) for the given nucleotide sequences in Nucleotide.txt.

   We first extract the sequences and their names from nucleotide_aligned.txt. Then we compare each sequence with each other one in the list and calculate the distance between them and store them at the respective position in the distance matrix. The distance is calculated using the hamming distance formula for finding the number of mismatches and if the positions of dashes match then that is subtracted from the length while dividing for distances.

   Thus we get a distance matrix which is stored in Ndistance.txt

b) This part deals with writing a script (q1b.ipynb) that uses Ndistance.txt and generates a phylogenetic relationship between the sequences using the UPGMA method.

   We first extract the lower left triangle of the distance matrix in Ndistance.txt and the names of the sequences. Then while at least one sequence is available, we search for the smallest distance in the table and correspondingly then re-align the table and link the two names while deleting the second one. For each linked name we add the corresponding brackets and lengths of branches.

   Thus we get a phylogenetic relationship between all sequences this way.

Question 2

This question involves constructing a phylogenetic relationship for the given protein sequences in Protein.txt.

a) This part deals with writing a script (q2a.ipynb) to generate a distance matrix file (Pdistance.txt) for the given protein sequences in Protein.txt.

   We first store the BLOSUM62 matrix using MatrixInfo from the Bio.SubsMat library. Then we extract the sequences and their names from protein_aligned.txt. Then we compare each sequence with each other one in the list and calculate the distance between them and store them at the respective position in the distance matrix using the BLOSUM62 scoring matrix. If there are spaces at both places then the score is 1 and if it is only at one position it is -4.

   Thus using this we get a distance matrix which is stored in Pdistance.txt

b) This part deals with writing a script (q2b.ipynb) that uses Pdistance.txt and generates a phylogenetic relationship between the sequences using the UPGMA method.

   We first extract the lower left triangle of the distance matrix in Pdistance.txt and the names of the sequences. Then while at least one sequence is available, we search for the smallest distance in the table and correspondingly then re-align the table and link the two names while deleting the second one. For each linked name we add the corresponding brackets and lengths of branches.

   Thus we get a phylogenetic relationship between all sequences this way.

I have used the below link to align Nucleotide.txt to nucleotide_aligned.txt and Protein.txt to protein_aligned.txt:
https://www.ebi.ac.uk/Tools/msa/clustalo/

The phylogenetic relationships have been displayed at the end of q1b.ipynb and q2b.ipynb.