# Phylogenetic Tree

# Report

Siddharth Jain
2020113014

First, all the given sequences and proteins have been aligned by using the online software provided. The following is the brief description of the programs.

1. **q1a-** For generating a distance matrix of the given nucleotides we first convert the aligned .txt file into dictionary for easy usage. Then we make a function to calculate the distance between 2 sequences. Then we create a n*n distance matrix by iterating over the given n sequences and present them in form of a matrix using pandas. At the end we generate a csv file **Ndistance.txt** having the data about the distance matrix.

2. **q2a**- We convert the file into a dictionary just like before. Then we import the scoring scheme "BLOSUM62' to score the similarity between the various proteins . we create a function get_score to find the degree of difference between two given strings. And then we create the distance matrix by iterating over the given sequences. At the end we generate a csv file **Pdistance.txt** having the data about the distance matrix.

3. **q1b** and **q2b**- For generating the phylogenetic tree using UPGMA algo, we create a the following functions:

   a. lowest_value- to find the shortest pairwise distance
   b. join_labels- join the 2 sequences/groups with smallest distance
   c. join_table- joins the rows a/c to UPGMA algorithm

At the end we use the UPGMA function to create the tree in the Newick(or bracket) tree format.