ASSIGNMENT REPORT

Step 1: Using the global optimum alignment tool, align the nucleotide sequence (by selecting DNA (for q1) / Protein (for q2) from the drop-down and choosing the suitable output format).

Note: The code has been explained in detail using comments in the python notebook files.
Note: Please run every block of code in the python notebook to get the answer.

Step 2:
1a:

1. Open the the globally aligned nucleotide sequence in python and parse the information into a dictionary list with the Alphanumeric code mapping to the nucleotide sequence.
2. Write code to make the distance matrix by computing normalised values for each pair of nucleotide sequences.
3. Export to Ndistance.txt.
4. The code for 2a is mostly similar but the way to compute the scores for Pdistance.txt is using the BLOSUM62 convention/matrix.

1b:

1. Open Ndistance.txt and load it into a distance matrix by renaming the columns and rows from A through J for convenience.
2. Convert the symmetric distance matrix into a half list so that we don’t need to deal with the upper triangle of the symmetric matrix in every iteration of the calculation.
3. Write the code for UPGMA algorithm to find the required phylogenetic tree.
4. Note: I have created a Newick format without distance for the phylogenetic tree and have separately written the length (from the leaf node base of the tree – ie the base of the tree where all the names are present in an aligned manner) of the common ancestors for every iteration. Some additional book-keeping would yield the Newick format answer with distance.
5. The code for 2b is the same as the code for 1b (with minor changes).