

ANIKA ROY

BIOLOGY ASSIGNMENT 2 REPORT

Q1)

The Q1 folder contains final_1.py—for generating distance matrix using Nucleotide.txt. I have assigned 0 to mismatches, insertions and deletions and 1 to matches while minimising the sequence length. I calculate the pairwise distances using dp and keep printing to the output file Ndistance.txt.

upgma_f.py reads this output and implements UPGMA algorithm giving the following output

```
anika@anika-HP-ENVY-Laptop-14-eb0021tx: ~/Desktop/pyth...
(base) anika@anika-HP-ENVY-Laptop-14-eb0021tx:~/Desktop/python$ cd Bio-A2-anika/
(base) anika@anika-HP-ENVY-Laptop-14-eb0021tx:~/Desktop/python/Bio-A2-anika$ ls
AnikaRoy-final-ans-bio.pdf  Q1  Q2
(base) anika@anika-HP-ENVY-Laptop-14-eb0021tx:~/Desktop/python/Bio-A2-anika$ cd
Q1
(base) anika@anika-HP-ENVY-Laptop-14-eb0021tx:~/Desktop/python/Bio-A2-anika/Q1$
python3 final_1.py > Ndistance.py
^Z
[1]+  Stopped                  python3 final_1.py > Ndistance.py
(base) anika@anika-HP-ENVY-Laptop-14-eb0021tx:~/Desktop/python/Bio-A2-anika/Q1$
python3 final_1.py > Ndistance.txt
(base) anika@anika-HP-ENVY-Laptop-14-eb0021tx:~/Desktop/python/Bio-A2-anika/Q1$
python3 upgma_f.py
[[0], [1], [2], [3], [4], [5], [6], [7], [8], [9]] 0.17880794701986755
[[0], [1], [2], [3], [7], [4], [5], [6], [8], [9]] 0.21394230769230768
[[0], [1, 9], [2], [3], [7], [4], [5], [6], [8]] 0.2515592515592516
[[0], [1, 9], [2], [3], [7], [4, 5, 6], [8]] 0.2713130652718754
[[0], [1, 9, 4, 5, 6], [2], [3], [7], [8]] 0.2988911853050549
[[0], [1, 9, 4, 5, 6, 3, 7], [2], [8]] 0.3242846968876508
[[0], [1, 9, 4, 5, 6, 3, 7, 8], [2]] 0.34471277789265614
[[0, 2], [1, 9, 4, 5, 6, 3, 7, 8]] 0.4170692431561997
[[0, 2, 1, 9, 4, 5, 6, 3, 7, 8]] 0.45206421408393715
(base) anika@anika-HP-ENVY-Laptop-14-eb0021tx:~/Desktop/python/Bio-A2-anika/Q1$
```

at every stage, it shows what i have clustered together and the number at the end is the distance between the groups we just clustered together which i then divide equally in the graph. I have used this output to hand draw the graph.

Q2)

The Q2 folder contains final_2.py—for generating distance matrix using Protein.txt. Using the BLOSUM 62 table, I have assigned weights to mismatches and matches while minimising the sequence length. For all insertions and deletions, I have assigned a penalty of -1. I calculate the pairwise score using dp and keep printing the inverse(my distance parameter) to the output file Pdistance.txt.

upgma_f.py reads this output and implements UPGMA algorithm giving the following output.

```
anika@anika-HP-ENVY-Laptop-14-eb0021tx: ~/Desktop/pyth...
(base) anika@anika-HP-ENVY-Laptop-14-eb0021tx:~$ cd Desktop/python/Bio-A2-anika/
(base) anika@anika-HP-ENVY-Laptop-14-eb0021tx:~/Desktop/python/Bio-A2-anika$ cd
Q2
(base) anika@anika-HP-ENVY-Laptop-14-eb0021tx:~/Desktop/python/Bio-A2-anika/Q2$
python3 upgma_f.py
[[0], [1], [2], [3, 7], [4], [5], [6], [8], [9]] 0.001718213058419244
[[0], [1, 3, 7], [2], [4], [5], [6], [8], [9]] 0.001926810886667137
[[0], [1, 3, 7], [2], [4], [5, 6], [8], [9]] 0.0019455252918287938
[[0], [1, 3, 7, 5, 6], [2], [4], [8], [9]] 0.0020603589186455502
[[0], [1, 3, 7, 5, 6, 4], [2], [8], [9]] 0.0021211767836498566
[[0], [1, 3, 7, 5, 6, 4, 9], [2], [8]] 0.002157606241360992
[[0, 1, 3, 7, 5, 6, 4, 9], [2], [8]] 0.002592063173078429
[[0, 1, 3, 7, 5, 6, 4, 9, 8], [2]] 0.0027545541670585096
[[0, 1, 3, 7, 5, 6, 4, 9, 8, 2]] 0.003309668679021074
(base) anika@anika-HP-ENVY-Laptop-14-eb0021tx:~/Desktop/python/Bio-A2-anika/Q2$
```