Tutorial Assignment 1

Roll No. 2020101116

1) DOTplot Analysis

I have attached the graphs for every comparison to SARS-COV2 required.

I) SARS-CoV2 is clearly more similar to SARS-CoV (2003), as can be seen by looking at the DotPlot graphs obtained.

II) It is easier to see the similarity using protein, as the DNA matches for SARS CoV2 with the two are very distorted and noisy w.r.t the protein matches.

III) Parameters:

DOTTUP:

Word Size =10

DOTMATCHER:

Word Size = 10
Threshold = 50
Matrix = BLOSUM62 for protein comparison
We use DNAfull for DNA COMPARISON

2) Pairwise Alignment:

a) SARS CoV 2 and SARS CoV

(I) Identity and similarity:

Needle:

DNA:

Identity: 3338/3902 (85.5%)

```
      Similarity:  3338/3902 (85.5%)

Protein:

      Identity:     974/1277 (76.3%)

      Similarity:  1110/1277 (86.9%)

Water:

      DNA:

            Identity:    3339/3899 (85.6%)

            Similarity:  3339/3899 (85.6%)

      Protein:

            Identity:     974/1277 (76.3%)
```

```
                   Similarity:  1110/1277 (86.9%)
```

We observe that the identity and identity are same but for protein, the similarity is significantly larger than identity as protein sequences have more common ancestry and thus similar properties, however, the actual elements in the sequence might be different which leads to the lower identity percentage.

II) Identity is the number of characters that actually match in the two given sequences being compared without considering gaps in the sequence.
Similarity measures how much the two sequences resemble each other. Similar sequences will show similarities in their properties too.
III)The global and local alignments show no difference in the two sequences, both under DNA and Protein basis.
IV)Has been attached in the the tests folder under the “A” naming scheme

B)
I) Yes they are homologous
II) Two sequences are approximately homologous if their similarity is above 40%. Since the similarities obtained are more than this (protein similarity of 45.8% and DNA, 57.7% on needle). Thus, we can say that they are homologous.

3) Database Search:

I) Closest Homolog is Spike Glycoprotein- Bat Coronavirus HKU3

II)
- Percentage identity: 76%
- Percentage similarity: 96.35%
- Length of alignment: 1242
- e-value: 0

III) The SARS-CoV spike glycoprotein was one of the hits. Yes the percentages identity and similarity match the alignment obtained using water. The alignment tells us that the SARS-COV-2 might have originated from SARS-COV

IV)Yes I do, there is similarity in the SARS-CoV 2 and the virus found in bats.

- Score: 1959.1
- Percentage identity: 76%
- Percentage similarity: 96.35%
- Length of alignment: 1242
- e-value: 0

4)
- Uniprot: 230328648 sequence entries, comprising 80596791741 amino acids.
- Genbank: 236338284 sequences, 1173984081721 bases

(i) Matrix cells required to compute using DP for searching/comparing in protein database
= no. of (amino) acids in UniProtKB/TrEMBL * length of query sequence
= 80596791741 * 1000

= 80596791741000

Matrix cells required to compute using DP for searching in nucleotide database

= no. of bases in GenBank * length of query sequence

= 1173984081721 * 1000

= 1173984081721000

Carrying out search, for the protein database, number of iterations = no. of (amino) acids in UniProtKB/TrEMBL * length of query sequence = 80596791741 * 1000 = 80596791741000

Assuming the number of operations carried out in one second is 10^7, the time taken would be

= 80596791741000/(10^7)

= 8059679.1741 seconds

Also, For the nucleotide database, number of iterations = no. of bases in GenBank * length of query sequence

= 1173984081721 * 1000

= 1173984081721000

Therefore, assuming that the number of operations (iterations) carried out in one second is 10^7, the time taken would be = 1173984081721000/(10^7) = 117398408.1721 seconds

a. Uniprot: time taken = 80596791741 *1000 /10^7 = 8059679.1741
b. Genbank: time taken = 1173984081721 *1000 / 10^7 = 117398408.1721

(ii) Space complexity = m * n, where m is the length of gene/genome and n is the length of query sequence

For Human Chr 1 :

m = 249Mbp = 249 * 10^6 * 2 = 498 * 10^6

n = 1000

=> space complexity = m * n = 498 * 10^6 * 10^3 = 498 * 10^9

Also For Mouse Chr 1 :

m = 195Mbp = 195 * 10^6 * 2 = 390 * 10^6

n = 1000

space complexity = m * n = 390 * 10^6 * 10^3 = 390 * 10^9

a. Human: 249* 10^6*2*1000= 498 * 10^9
b. Mouse: 195 * 10^6 * 2 * 1000 = 390 * 10^9