Science II Tutorial 2 Assignment

Roll No: 2021101064

Name: Aman Raj

1)

(A) The following results were obtained by using the multiple sequence alignment. The similarity matrix derived comes out to be:

For DNA:

```
#
#
#   Percent Identity  Matrix - created by Clustal2.1
#
#


    1: MERS       100.00   45.98   46.26   46.08   46.54   46.66
    2: BATCOV      45.98  100.00   70.16   68.11   68.24   68.42
    3: COV         46.26   70.16  100.00   73.56   73.32   73.35
    4: PCOV        46.08   68.11   73.56  100.00   82.83   83.39
    5: BATRAG13    46.54   68.24   73.32   82.83  100.00   93.12
    6: COV2        46.66   68.42   73.35   83.39   93.12  100.00
```

For Protein:

```
#
#
#   Percent Identity  Matrix - created by Clustal2.1
#
#


    1: MERS-CoV   100.00   31.77   32.28   31.91   32.32   32.30
    2: BatCOV      31.77  100.00   75.85   71.61   72.28   72.26
    3: SARS-CoV    32.28   75.85  100.00   77.38   77.76   77.30
    4: PCOV        31.91   71.61   77.38  100.00   92.98   92.43
    5: RATG13      32.32   72.28   77.76   92.98  100.00   97.71
    6: SARS        32.30   72.26   77.30   92.43   97.71  100.00
```

As we can see from the Identity Matrix, the closest relative to SARS CoV 2 is Bat RatG13 since the percentage match is around 93.13 with SARS in DNA and 97.71 in
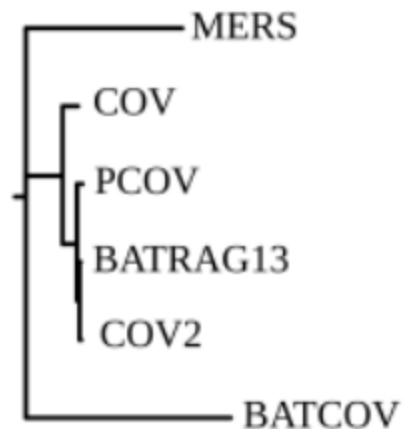
the Protein.

(B)

We can describe this from (A) that the source of SARS CoV 2 could be most likely be the Bat RatG13 due to their similarity.

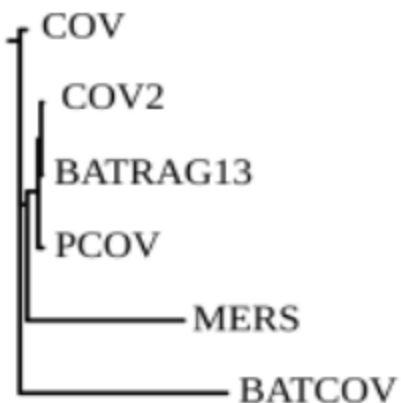For MERS, we cannot infer the from the data.
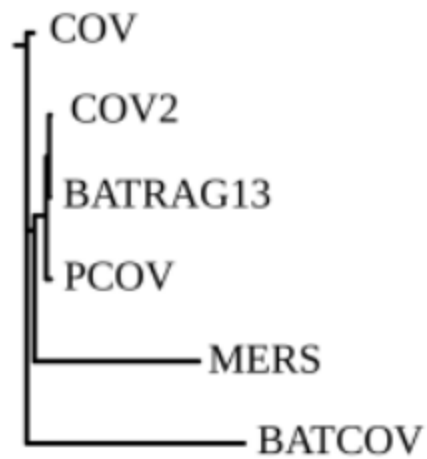
2)

The trees generated are as follows:
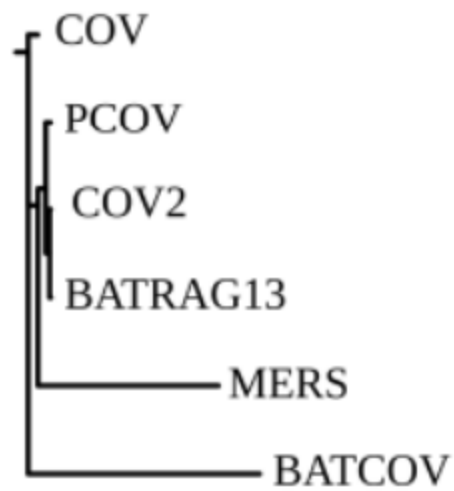
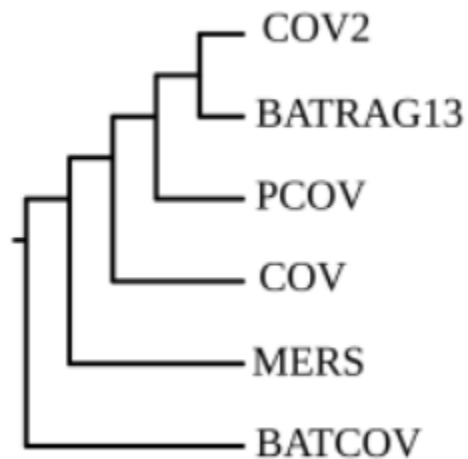Distance Without Bootstrap : -

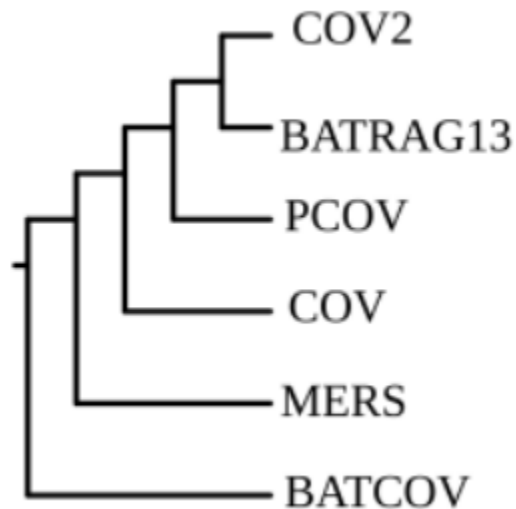Distance With Bootstrap : -

Likelihood Without Bootstrap : -

Likelihood With Bootstrap : -

Parsing Without Bootstrap : -

Parsing With Bootstrap : -

a) No, the trees that are obtained by the methods are not topology-wise.

b) There are essentially no differences between the bootstrapped and non bootstrapped data here. Bootstrapping is a statistical method that is used for resampling a large set to make a large number of stimulated samples.

c) Yes, the inferences derived from here are in really striking agreement to Q1(b). This is because here, pairing is similar for all the cases and for the one which we can't

determine, it is paired with the others.

Sequence 1 is closely related to "Arabidopsis thaliana genome" from mustard family.

Sequence 2 is closely related to "Tomato spotted wilt virus intergenic region, isolate TSWV-IT-CE, genomic RNA" from Arachis hypogea plant.