Science Tutorial Assignment 2

> Q1. SARS-CoV-2 is a zoonotic virus that jumped from animal species to humans. It is known that coronaviruses utilize one of its proteins, called Spike glycoprotein, to infect the host cells. The cellular entry of the coronavirus depends on binding between the viral Spike protein receptor-binding domain (RBD) and the angiotensin converting enzyme 2 (ACE2) target cell receptor. Hence it is important to understand how this protein has evolved for drug target.
> To understand how this protein has evolved to modify its host preference, perform multiple sequence alignment of the orthologs of these proteins.
> Further, this virus is reported to be closely related to bat and pangolin coronaviruses, and it is hypothesised that perhaps it is a blend of bat and pangolin viruses that emerged by a process, called recombination, in a bat, pangolin or another species. What do you infer from your analysis?

1. Download gene and protein sequences of Spike glycoprotein from six species: SARS-CoV-2, SARS-CoV, MERS-CoV, Bat coronavirus RaTG13, Bat-CoV, and P-CoV. [Hint: Go to NCBI Nucleotide search page, get the genomic sequences, from which search for spike protein and download CDS and protein sequences of spike glycoprotein in FASTA format. Generate a single multi-FASTA file for performing multiple-sequence alignment].

a. **Perform multiple sequence alignment and based on percentage identity, identify its closest relative. Compare the results with gene and protein sequences and give the % identity matrix in the two cases.**

→ **Identity Matrix Protein:**

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

     1: QHR63300.2               100.00   97.71   76.90   77.68   32.41   33.17
     2: QIG55955.1                97.71  100.00   76.68   77.22   32.38   33.20
     3: sp|Q0Q475.1|SPIKE_BC279   76.90   76.68  100.00   80.70   32.05   33.70
     4: YP_009825051.1            77.68   77.22   80.70  100.00   32.43   33.25
     5: QGW51920.1                32.41   32.38   32.05   32.43  100.00   66.32
     6: UJZ92542.1                33.17   33.20   33.70   33.25   66.32  100.00
```

SARS-CoV-2 here is the QIG55955.1 spike glycoprotein. Based on the identity matrix its closest relative is QHR63300.2 spike glycoprotein which is the **Bat coronavirus RaTG13.**

Identity Matrix DNA:

```
#
#
```

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

     1: KR559017.1               100.00   70.08   68.35   68.53   44.88   46.32
     2: NC_004718.3_21492-25259   70.08  100.00   73.46   73.48   45.78   46.12
     3: MN996532.2_21560-25369    68.35   73.46  100.00   93.12   45.26   46.08
     4: MT126746.1_21465-25286    68.53   73.48   93.12  100.00   45.04   46.31
     5: MN723544.1_21456-25517    44.88   45.78   45.26   45.04  100.00   65.68
     6: OM009282.1_21571-25647    46.32   46.12   46.08   46.31   65.68  100.00
```

MT126746.1:21465-25286 is SARS-CoV-2 and it's closest relative is MN996532.2:21560-25369 which is again the **Bat coronavirus RaTG13.**

b. From these results can you infer the possible source of origin of SARS-CoV-2 and that of MERS-CoV

For SARS-CoV-2 since the **Bat coronavirus RaTG13** is its closest relative we can guess that it was the most likely source of origin. For MERS-CoV, (MN723544.1:21456-25517 DNA and QGW51920.1 Spike glycoprotein) the closest relative is OM009282.1:21571-25647 Pangolin coronavirus and UJZ92542.1 spike glycoprotein [Pangolin coronavirus]. Hence we infer that its most likely source of origin is the **Pangolin coronavirus HKU4/P251T/pangolin/2018.**

2. Construct phylogenetic tree using the multiple sequence alignment files obtained in the previous Q. and analyse your results using Phylip. Use one of each parsimony, distance-based, and maximum likelihood methods and compare the three trees obtained (with and without bootstrapping). Submit the trees and summarize your observations.

a. Tool output DNA:

```
Using 8 threads
Read 6 sequences (type: DNA) from clustalo-I20220410-184421-0679-85626986-p2m.upfile
not more sequences (6) than cluster-size (100), turn off mBed
Calculating pairwise ktuple-distances...
Ktuple-distance calculation progress: 0 % (0 out of 21)
Ktuple-distance calculation progress: 4 % (1 out of 21)
Ktuple-distance calculation progress: 9 % (2 out of 21)
Ktuple-distance calculation progress: 14 % (3 out of 21)
Ktuple-distance calculation progress: 47 % (10 out of 21)
Ktuple-distance calculation progress: 61 % (13 out of 21)
Ktuple-distance calculation progress done. CPU time: 0.12u 0.00s 00:00:00.12 Elapsed: 00:00:00
Guide tree written to clustalo-I20220410-184421-0679-85626986-p2m.dnd
Guide-tree computation done.
Progressive alignment progress: 20 % (1 out of 5)
Progressive alignment progress: 40 % (2 out of 5)
Progressive alignment progress: 60 % (3 out of 5)
Progressive alignment progress: 80 % (4 out of 5)
Progressive alignment progress: 100 % (5 out of 5)
Progressive alignment progress done. CPU time: 11.16u 2.50s 00:00:13.66 Elapsed: 00:00:13
Alignment written to clustalo-I20220410-184421-0679-85626986-p2m.phylip
```

Tool output Protein:

```
Using 8 threads
Read 6 sequences (type: Protein) from clustalo-I20220410-184359-0427-4244495-p2m.upfile
not more sequences (6) than cluster-size (100), turn off mBed
Calculating pairwise ktuple-distances...
Ktuple-distance calculation progress: 0 % (0 out of 21)
Ktuple-distance calculation progress: 4 % (1 out of 21)
```

```
Ktuple-distance calculation progress: 9 % (2 out of 21)
Ktuple-distance calculation progress: 14 % (3 out of 21)
Ktuple-distance calculation progress: 61 % (13 out of 21)
Ktuple-distance calculation progress: 71 % (15 out of 21)
Ktuple-distance calculation progress done. CPU time: 0.02u 0.00s 00:00:00.02 Elapsed: 00:00:00
Guide tree written to clustalo-I20220410-184359-0427-4244495-p2m.dnd
Guide-tree computation done.
Progressive alignment progress: 20 % (1 out of 5)
Progressive alignment progress: 40 % (2 out of 5)
Progressive alignment progress: 60 % (3 out of 5)
Progressive alignment progress: 80 % (4 out of 5)
Progressive alignment progress: 100 % (5 out of 5)
Progressive alignment progress done. CPU time: 1.39u 0.27s 00:00:01.66 Elapsed: 00:00:02
Alignment written to clustalo-I20220410-184359-0427-4244495-p2m.phylip
```

The generated `.phylip` files can be found here:

Scince tut assignment 2

Instantly share code, notes, and snippets. You can't perform that action at this time. You signed in with another tab or window. You signed out in another tab or window. Reload to refresh your session. Reload to refresh your session.

https://gist.github.com/akcube/36b367f598b03f63ee2533bbdd330093

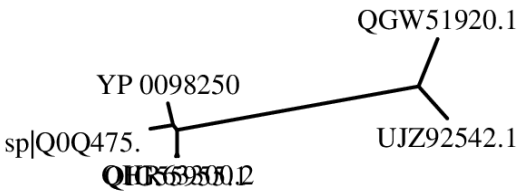
Non-Bootstrapped

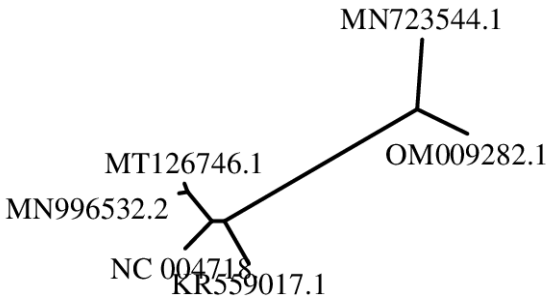
Protein `protdist`

```
+QIG55955.1
!
!     +---sp|Q0Q475.
!   +-3
!   ! +---YP_0098250
2---4
!   !                                  +-------QGW51920.1
!   +-----------------------------------1
!                                      +-----UJZ92542.1
!
+QHR63300.2
```

DNA `dnadist`

```
  +----NC_004718.
+-3
! !    +MN996532.2
! +----2
!      +MT126746.1
!
!                             +--------MN723544.1
4-----------------------------1
!                             +------OM009282.1
!
+-----KR559017.1
```





Protein `protpars`

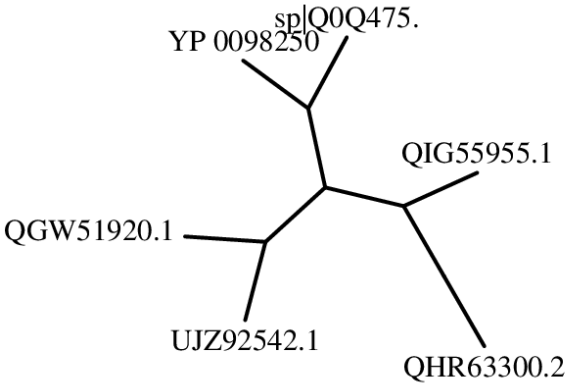
```
      +--UJZ92542.1
+-----5
!     +--QGW51920.1
```

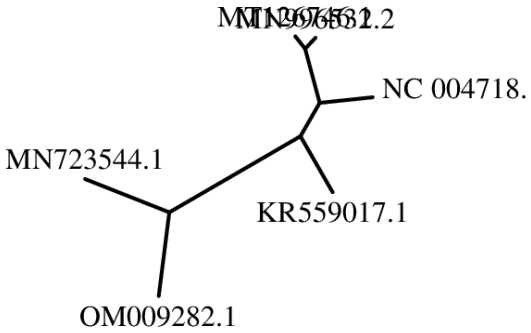
DNA `dnapars`

```
                    +----------OM009282.1
+-------------------4
|                   +-----------MN723544.1
```

```
   +--4
   !  !     +--YP_0098250
+--2  +-----3
!  !        +--sp|Q0Q475.
1  !
!  +-----------QIG55955.1
!
+--------------QHR63300.2
```

```
|
|           +--MT126746.1
|    +------3
1----2      +-MN996532.2
|    |
|    +------NC_004718.
|
+-------KR559017.1
```



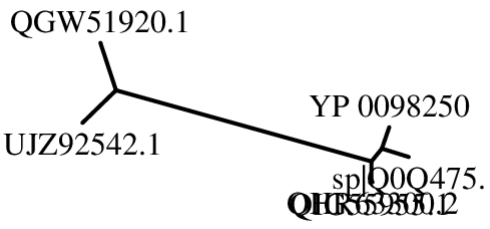


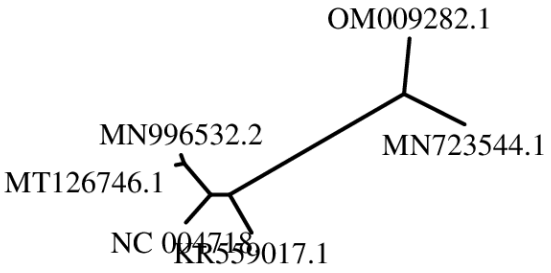
Protein `proml`

```
+QIG55955.1
|
|                                        +------UJZ92542.1
|  +-------------------------------------4
|  |                                     +------QGW51920.1
1--3
|  |  +---YP_0098250
|  +--2
|     +---sp|Q0Q475.
|
+QHR63300.2
```

DNA `dnaml`

```
   +----NC_004718.
+--1
|  |    +MT126746.1
|  +----2
|       +MN996532.2
|
|                          +------OM009282.1
3--------------------------4
|                          +--------MN723544.1
|
+-----KR559017.1
```





Bootstrapped

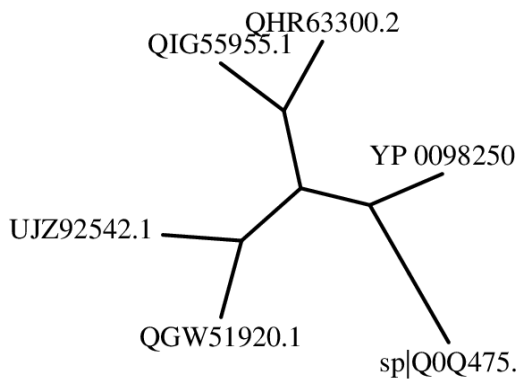
Protein `protdist`

```
                    +-------QGW51920.1
          +-100.0-|
          |         +-------UJZ92542.1
```

DNA `dnadist`

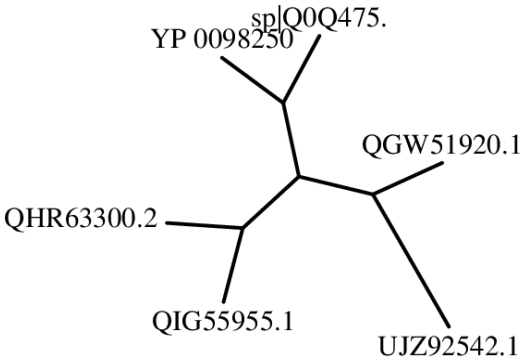
```
                              +-------MN996532.2
          +---------100.0-|
          |                 +-------MT126746.1
```

```
         +--90.0-|
         |       |       +-------QIG55955.1
+--------|       +-100.0-|
|        |               +-------QHR63300.2
|        |
|        +-----------------------YP 0098250
|
+--------------------------------sp|Q0Q475.
```

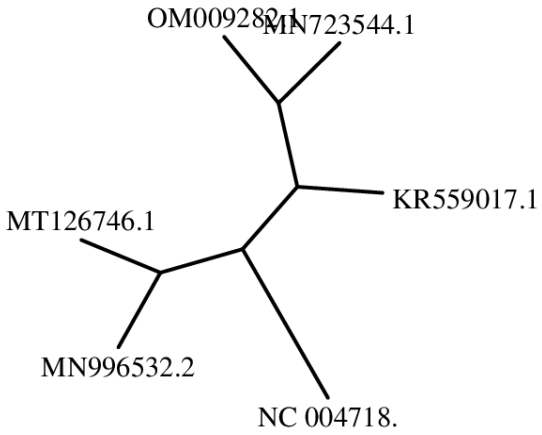


Protein `protpars`

```
                          +-------QIG55955.1
                +-100.0-|
                |       +-------QHR63300.2
        +-100.0-|
        |       |       +-------YP 0098250
+-------|       +--98.5-|
|       |               +-------sp|Q0Q475.
|       |
|       +-----------------------QGW51920.1
|
+--------------------------------UJZ92542.1
```

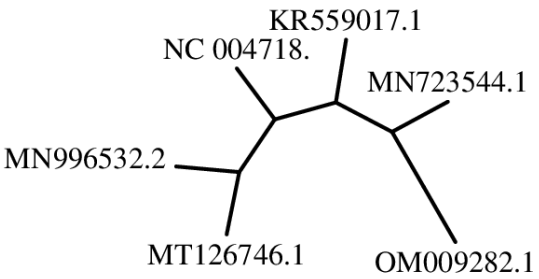



```
+-------|
|       |               +-------OM009282.1
|       |       +-100.0-|
|       +--97.0-|       +-------MN723544.1
|               |
|               +---------------KR559017.1
|
+--------------------------------NC 004718.
```



DNA `dnapars`

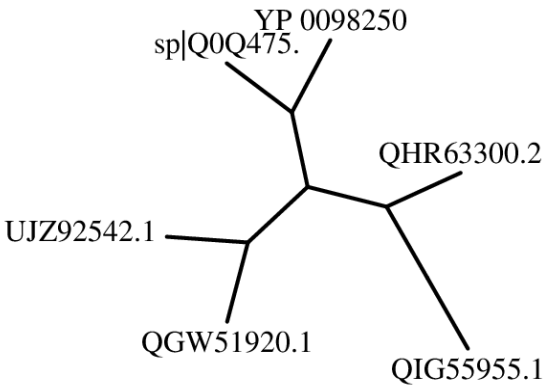
```
                                        +-------MT126746
                        +-100.0-|
                +-100.0-|       +-------MN996532.2
                |       |
        +-100.0-|       +---------------NC 004718.
        |       |
+-------|       +-----------------------KR559017.1
|       |
|       +--------------------------------MN723544.1
|
+---------------------------------------OM009282.1
```



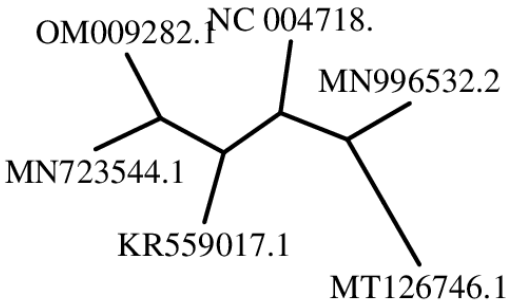
DNA `dnaml`

Protein `proml`

```
                              +-------QGW51920.1
                +-100.0-|
                |       +-------UJZ92542.1
        +-100.0-|
        |       |       +-------sp|Q0Q475.
+-------|       +--93.0-|
|       |               +-------YP 0098250
|       |
|       +----------------------QHR63300.2
|
+------------------------------QIG55955.1
```




```
                                  +---------------KR559017.1
                +--96.0-|
                |       |       +-------MN723544.1
        +-100.0-|       +-100.0-|
        |       |               +-------OM009282.1
+-------|       |
|       |       +----------------------NC 004718.
|       |
|       +------------------------------MN996532.2
|
+-------------------------------------MT126746.1
```



All the complete `outfiles` generated and the script containing all the commands run to generate the outfiles can be found here

a. **Are the trees obtained by different methods in agreement, topology-wise?**

   → Yes we can see that the trees are in agreement topology-wise, only the orientations of the trees change.

b. **Do you observe any difference with or without bootstrapping? What information does bootstrapping provide?**

   → No there is no observable difference from bootstrapping. Bootstrapping provides us with the information that despite mutations the results we get are similar.

c. **Are your inferences in agreement with those in Q.1(b) above.**

   → Yes my inferences are in agreement. No change from bootstrapping implies that the trees must be in agreement topology wise.