Bioinformatics

Assignment 6

Kushagra Agarwal
2018113012

1) Multiple Sequence Alignment of the spike protein

Gene sequences:

1) NC_048212.1:20814-24623 Bat coronavirus isolate CMR704-P12, complete cds
2) MN996532.1:21545-25354 Bat coronavirus RaTG13, complete genome
3) NC_019843.3:21456-25517 Middle East respiratory syndrome coronavirus, complete genome
4) MT799526.1 Pangolin coronavirus isolate cDNA31-S surface glycoprotein (S) gene, complete cds
5) NC_045512.2:21563-25384 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome
6) NC_004718.3:21492-25259 SARS coronavirus, complete genome

Protein:

1) YP_009824990.1 spike protein [Bat coronavirus]
2) QHR63300.2 spike glycoprotein [Bat coronavirus RaTG13]
3) YP_009047204.1 spike glycoprotein [Middle East respiratory syndrome-related coronavirus]
4) QLR06869.1 surface glycoprotein [Pangolin coronavirus]
5) YP_009724390.1 surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]
6) NP_828851.1 E2 glycoprotein precursor [Severe acute respiratory syndrome-related coronavirus]

MSA for Protein Sequences:

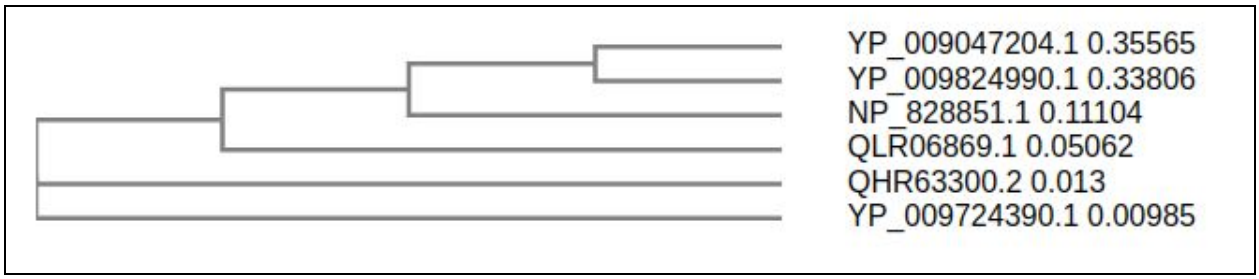
Link to Results:
https://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=clustalo-I20201001-095952-0100-85336523-p2m&analysis=alignments

Percentage Identity Matrix

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


  1: YP_009047204.1  100.00   30.63   32.53   32.32   32.29   32.10
  2: YP_009824990.1   30.63  100.00   34.68   34.00   33.89   33.72
  3: NP_828851.1      32.53   34.68  100.00   76.82   77.68   77.22
  4: QLR06869.1       32.32   34.00   76.82  100.00   89.64   90.59
  5: QHR63300.2       32.29   33.89   77.68   89.64  100.00   97.71
  6: YP_009724390.1   32.10   33.72   77.22   90.59   97.71  100.00
```

Phylogenetic Tree



Observations

Closest to SARS-COV-2 spike protein is the QHR63300.2 spike glycoprotein [Bat coronavirus RaTG13]. The percentage identity is 97.71% as can be seen in the 5th column of the 6th row in the matrix.

MSA for Gene Sequences:

Link to Results

https://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=clustalo-I20201001-100004-0762-94431923-p1m

Percentage Identity Matrix

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


  1: NC_019843.3_21456-25517  100.00   44.27   45.55   45.55   45.93   46.16
  2: NC_048212.1_20814-24623   44.27  100.00   47.17   46.92   47.11   47.68
  3: NC_004718.3_21492-25259   45.55   47.17  100.00   73.12   73.27   73.38
  4: MT799526.1                45.55   46.92   73.12  100.00   83.40   84.16
  5: MN996532.1_21545-25354    45.93   47.11   73.27   83.40  100.00   93.12
  6: NC_045512.2_21563-25384   46.16   47.68   73.38   84.16   93.12  100.00
```

Phylogenetic Tree

Observations

Closest to SARS-COV-2 spike protein is the MN996532.1:21545-25354 Bat coronavirus RaTG13, complete genome. The percentage identity is 93.12 % as can be seen in the 5th column of the 6th row in the matrix.

b) INFERENCES

Protein sequences are better to identify the closest relative (97.71 % compared to 93.12% in the case of gene sequences).

- For SARS-2 we can conclude that the origin was from the Bat coronavirus RaTG13. As the percentage identity is maximum for both protein and DNA alignments.
- For MERS, we see the closest relative in the case of DNA is SARS-COV-2 but all the values are in a very small range. (~45-46%). For protein MSA we find that the closest relative is E2 glycoprotein precursor [Severe acute respiratory syndrome-related coronavirus], but again all the values are in a very small range and therefore we cannot conclusively say which of them actually was the origin of MERS-COV.

---

2) Phylogenetic Trees using Phylip

Programs used

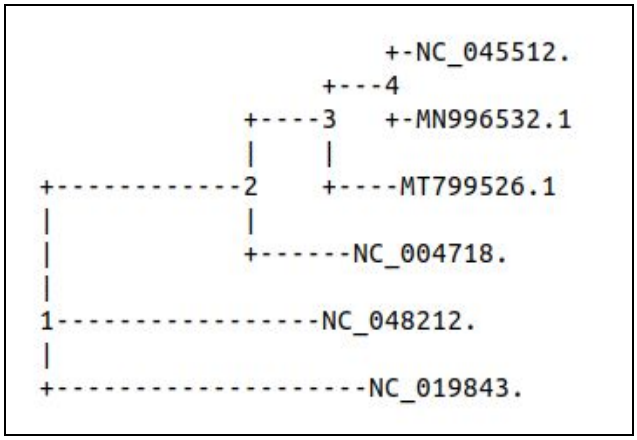
1) ./dnapars: for parsimony
2) ./dnadist: for distance-based (get dist matrix then seqboot)
3) ./dnaml: for maximum likelihood-based

http://sequenceconversion.bugaco.com/converter/biology/sequences/clustal_to_phylip.php was used to convert clustal alignment output to .phylip format.

For bootstrapping I first ran ./seqboot on original .phylip and used 11 as the random seed for my bootstrapper. The output file was used as input for all the 3 programs previously described.

PARSIMONY

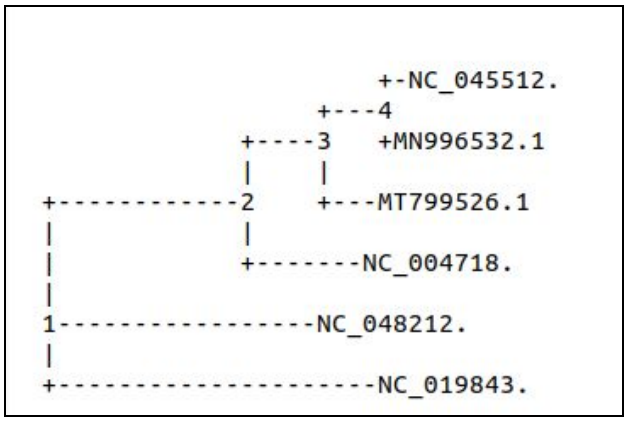
Without Bootstrap




```
requires a total of    5414.000

  between       and       length
  -------       ---       ------
     1            2       0.211524
     2            3       0.091261
     3            4       0.058909
     4       NC_045512.   0.031975
     4       MN996532.1   0.034713
     3       MT799526.1   0.074678
     2       NC_004718.   0.127036
     1       NC_048212.   0.305833
     1       NC_019843.   0.355887
```

With Bootstrap

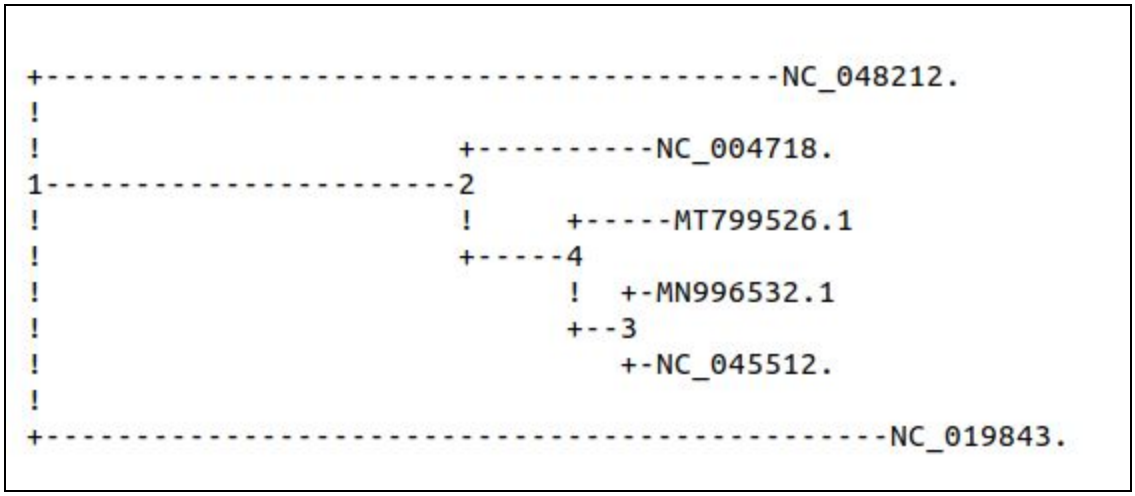


between	and	length
1	2	0.217661
2	3	0.086050
3	4	0.058220
4	NC_045512.	0.029789
4	MN996532.1	0.029664
3	MT799526.1	0.070182
2	NC_004718.	0.127169
1	NC_048212.	0.300652
1	NC_019843.	0.359782

DISTANCE-BASED

Without Bootstrap

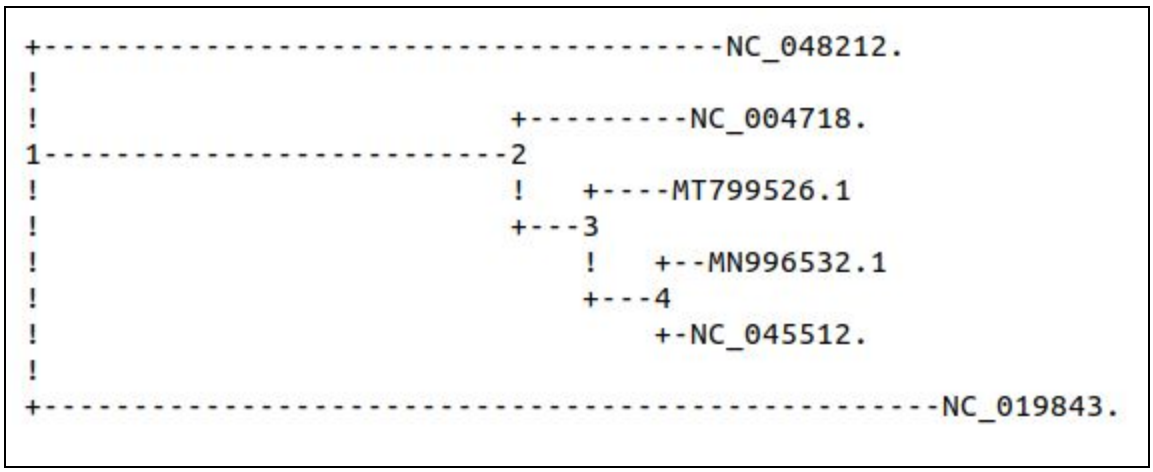
	6
NC_019843.	0.000000 1.493565 1.381526 1.393070 1.371912 1.369159
NC_048212.	1.493565 0.000000 1.270578 1.313813 1.300628 1.268629
NC_004718.	1.381526 1.270578 0.000000 0.357691 0.354678 0.353949
MT799526.1	1.393070 1.313813 0.357691 0.000000 0.194941 0.184380
MN996532.1	1.371912 1.300628 0.354678 0.194941 0.000000 0.072757
NC_045512.	1.369159 1.268629 0.353949 0.184380 0.072757 0.000000



Between	And	Length
1	NC_048212.	0.70153
1	2	0.40662
2	NC_004718.	0.17264
2	4	0.09040
4	MT799526.1	0.10214
4	3	0.05114
3	MN996532.1	0.04128
3	NC_045512.	0.03148
1	NC_019843.	0.79203

With Bootstrap

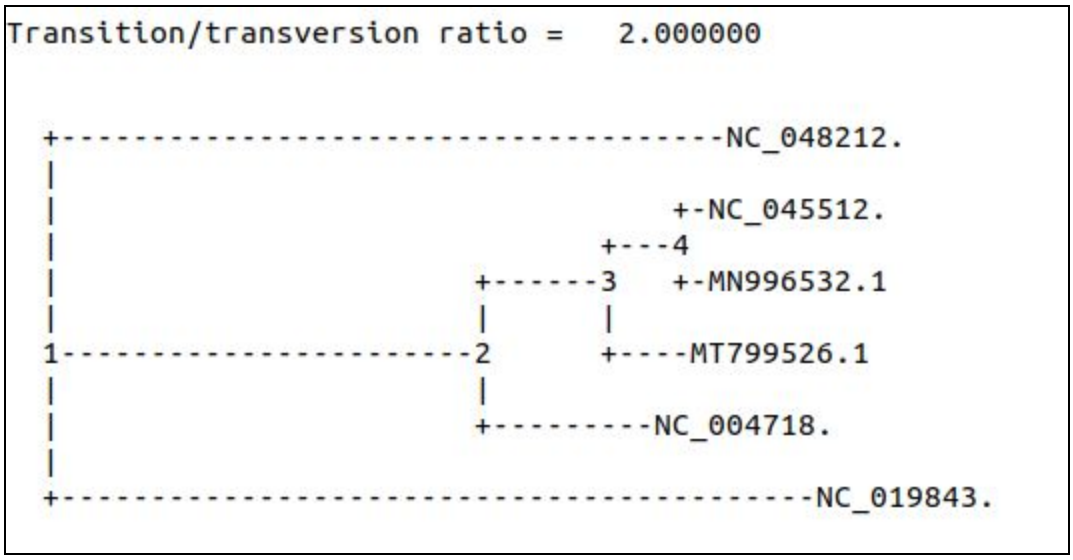
```
    6
NC_019843. 0.000000 1.503716 1.517145 1.417041 1.456515 1.464079
NC_048212. 1.503716 0.000000 1.235975 1.262552 1.326960 1.272894
NC_004718. 1.517145 1.235975 0.000000 0.340857 0.340344 0.336703
MT799526.1 1.417041 1.262552 0.340857 0.000000 0.182411 0.175982
MN996532.1 1.456515 1.326960 0.340344 0.182411 0.000000 0.063418
NC_045512. 1.464079 1.272894 0.336703 0.175982 0.063418 0.000000
```



Between	And	Length
1	NC_048212.	0.65731
1	2	0.45011
2	NC_004718.	0.17459
2	3	0.07194
3	MT799526.1	0.08010
3	4	0.06738
4	MN996532.1	0.03668
4	NC_045512.	0.02674
1	NC_019843.	0.84641

MAXIMUM-LIKELIHOOD-BASED

Without Bootstrap



```
Ln Likelihood = -21523.81020

 Between        And            Length      Approx. Confidence Limits
 -------        ---            ------      ------- ---------- ------

     1          NC_019843.        0.71156     (  0.64141,     0.78172) **
     1          NC_048212.        0.64060     (  0.57519,     0.70601) **
     1             2              0.39882     (  0.34488,     0.45277) **
     2             3              0.11376     (  0.09467,     0.13284) **
     3             4              0.06832     (  0.05722,     0.07941) **
     4          NC_045512.        0.03080     (  0.02404,     0.03757) **
     4          MN996532.1        0.04187     (  0.03432,     0.04942) **
     3          MT799526.1        0.09219     (  0.07972,     0.10466) **
     2          NC_004718.        0.16137     (  0.14112,     0.18160) **

     *  = significantly positive, P < 0.05
     ** = significantly positive, P < 0.01
```

With Bootstrap

```
Transition/transversion ratio =   2.000000


  +-----------------------------------NC_048212.
  |
  |                                      +-NC_045512.
  |                                 +---4
  |                           +-----3    +-MN996532.1
  |                           |     |
  1-------------------------2     +----MT799526.1
  |                           |
  |                           +---------NC_004718.
  |
  +--------------------------------------------NC_019843.
```

```
Ln Likelihood = -21228.29825

 Between        And            Length      Approx. Confidence Limits
 -------        ---            ------      ------- ---------- ------

     1          NC_019843.        0.76987     (  0.69404,     0.84571) **
     1          NC_048212.        0.60549     (  0.54139,     0.66959) **
     1             2              0.41641     (  0.36051,     0.47232) **
     2             3              0.10455     (  0.08602,     0.12310) **
     3             4              0.06732     (  0.05637,     0.07827) **
     4          NC_045512.        0.02757     (  0.02125,     0.03389) **
     4          MN996532.1        0.03568     (  0.02875,     0.04261) **
     3          MT799526.1        0.08702     (  0.07491,     0.09914) **
     2          NC_004718.        0.16035     (  0.14047,     0.18022) **

     *  = significantly positive, P < 0.05
     ** = significantly positive, P < 0.01
```

a) Even though all the trees are not identical, they are similar, as in the groupings are the same, even though the branch lengths are different in all the trees, which is also expected. Therefore all the trees have the same topology. MN996532.1:21545-25354 Bat coronavirus RaTG13, the complete genome is closest to NC_045512.2:21563-25384 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome in all of them.

b) In our case bootstrapping did not change any tree. It just caused a change in the distance matrices and not the topology of the trees.

Bootstrapping is a resampling analysis that involves taking columns of characters out of our analysis, rebuilding the tree, and testing if the same nodes are recovered. This is done through many (100 or 1000, quite often) iterations. If, for example, we recover the same node through 95 of 100 iterations of taking out one character and resampling our tree, then we have a good idea that the node is well supported (our bootstrap value, in that case, would be 0.95 or 95%).

A bootstrap value of 70 is often considered as the threshold for good confidence. See this paper by Hillis & Bull.
http://sysbio.oxfordjournals.org/content/42/2/182.short

In the statistical context, bootstrapping refers to using the data at hand to infer the uncertainty of said data. I.e. improve the statistic by pulling on its bootstraps. In practice, this is achieved by sampling or permuting the input data.

In terms of the phylogenetic trees, the bootstrapping values indicate how many times out of 100 (in your case) the same branch was observed when repeating the phylogenetic reconstruction on a re-sampled set of our data. If we get 100 out of 100 (and our data is sufficiently large to support this), we are pretty sure that the observed branch is not due to a single extreme data point. If we get 50 out of 100, we cannot be as certain.
https://www.sciencedirect.com/topics/medicine-and-dentistry/bootstrapping

c) Yes the results match with those of Q1)b). We again got Bat RaTG13 as the closest relative which is in agreement to Q1)b). MN996532.1:21545-25354 Bat coronavirus RaTG13, the complete genome is closest to NC_045512.2:21563-25384 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome in all of them.
