Project 3

Learning phylogenetic trees from multiple sequence alignment of the COVID data

Mahboobeh (Mariya) Golchinpour leili

Objective

The objective of this project is to investigate the evolutionary history of a virus strain by performing a phylogenetic analysis. To accomplish this, you will use the "MAFFT" and "RAxML-NG" tools to estimate the topology and parameters of the phylogenetic tree using the FASTA file data of Omicron, Delta, or Alpha strain genomes.

Data

The dataset is a portion of the GISAID dataset, comprising of SARS-CoV-2 strains collected from humans in various cities across Iran between 2021 and 2023. Each student will choose a subset of the data from the table below and study the evolution of a particular COVID-19 strain in Iran during a designated time period.

Data Nr.	1	2	3	4	5	6	7
strain	Omicron	Omicron	Omicron	Omicron	Delta	Delta	Alpha
Time interval	Jan 2022-Apr 2022	Jun 2022-Sep 2022	Aug 2022-Nov 2022	Sep 2022-now	Jan 2021-Nov 2022	Jan 2022-now	Oct 2021-now

Project steps

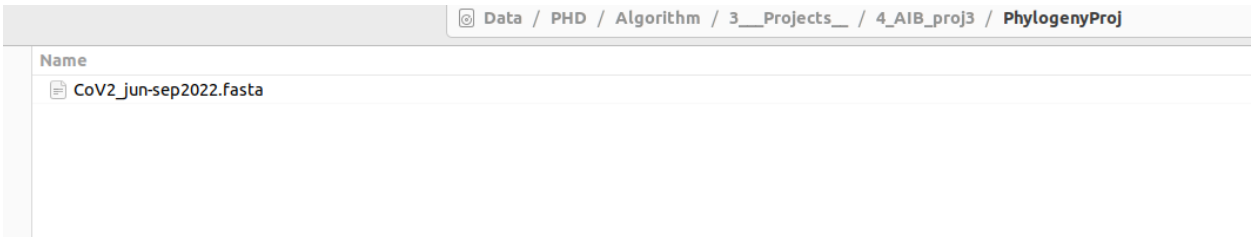
This project will involve the following steps:

1. **Data preparation**

I.Download the selected subset of the GISAID dataset.

My Data is number 2. Jun 2022-Sep2022 : **CoV2_jun-sep2022.fasta**

Directory:PhylogenyProj



i have a dataset of SARS-CoV-2 strains collected from humans in various cities across Iran between June 2022 and September 2022.

```
3376 >hCoV-19/Iran/NIC14010110MRQ1/2022|EPI_ISL_13369278|2022-03
3377 GAACAGGAAGAGAATCAGCAACTGTGTTGCTGATTATTCTGTCCTATATAATCTCGCACCATTTTTCACTTTTAAGTGTT
3378 ATGGAGTGTCTCCTACTAAATTAAATGATCTCTGCTTTACTAATGTCTATGCAGATTCATTTGTAATTAGAGGTGATGAA
3379 GTCAGACAAATCGCTCCAGGGCAAACTGGAAATATTGCTGATTATAATTATAAATTACCAGATGATTTTACAGGCTGCGT
3380 TATAGCTTGGAATTCTAACAAGCTTGATTCTAAGGTTAGTGGTAATTATAATTACCTGTATAGATTGTTTAGGAAGTCTA
3381 ATCTCAAACCTTTTGAGAGAGATATTTCAACTGAAATCTATCAGGCCGGTAACAAACCTTGTAATGGTGTTGCAGGTTTT
3382 AATTGTTACTTTCCTTTACGATCATATAGTTTCCGACCCACTTATGGTGTTGGTCACCAACCATACAGAGTAGTAGTACT
3383 TTCTTTTGAACTTCTACATGCACCAGCAACTGTTTGTGGACCTAAAAAGTCTACTAATTTGGTTAAAAACAAATGTGTC
3384 >hCoV-19/Iran/NIC14010110ALB3/2022|EPI_ISL_13369288|2022-03
3385 CTTGGAACAGGAAGAGAATCAGCAACTGTGTTGCTGATTATTCTGTCCTATATAATCTCGCACCATTTTTCACTTTTAAG
3386 TGTTATGGAGTGTCTCCTACTAAATTAAATGATCTCTGCTTTACTAATGTCTATGCAGATTCATTTGTAATTAGAGGTGA
3387 TGAAGTCAGACAAATCGCTCCAGGGCAAACTGGAAATATTGCTGATTATAATTATAAATTACCAGATGATTTTACAGGCT
3388 GCGTTATAGCTTGGAATTCTAACAAGCTTGATTCTAAGGTTAGTGGTAATTATAATTACCTGTATAGATTGTTTAGGAAG
3389 TCTAATCTCAAACCTTTTGAGAGAGATATTTCAACTGAAATCTATCAGGCCGGTAACAAACCTTGTAATGGTGTTGCAGG
3390 TTTTAATTGTTACTTTCCTTTACGATCATATAGTTTCCGACCCACTTATGGTGTTGGTCACCAACCATACAGAGTAGTAG
3391 TACTTTCTTTTGA
3392 >hCoV-19/Iran/NIC14010110SHH2/2022|EPI_ISL_13369292|2022-03
3393 TTTATGCTTGGAACAGGAAGAGAATCAGCAACTGTGTTGCTGATTATTCTGTCCTATATAATCTCGCACCATTTTTCACT
3394 TTTAAGTGTTATGGAGTGTCTCCTACTAAATTAAATGATCTCTGCTTTACTAATGTCTATGCAGATTCATTTGTAATTAG
3395 AGGTGATGAAGTCAGACAAATCGCTCCAGGGCAAACTGGAAATATTGCTGATTATAATTATAAATTACCAGATGATTTTA
3396 CAGGCTGCGTTATAGCTTGGAATTCTAACAAGCTTGATTCTAAGGTTAGTGGTAATTATAATTACCTGTATAGATTGTTT
3397 AGGAAGTCTAATCTCAAACCTTTTGAGAGAGATATTTCAACTGAAATCTATCAGGCCGGTAACAAACCTTGTAATGGTGT
3398 TGCAGGTTTTAATTGTTACTTTCCTTTACGATCATATAGTTTCCGACCCACTTATGGTGTTGGTCACCAACCATACAGAG
3399 TAGTAGTACTTTCTTTTGAACTTCTAC
3400 >hCoV-19/Iran/NIC14010110SHH3/2022|EPI_ISL_13369293|2022-03
3401 GAACAGGAAGAGAATCAGCAACTGTGTTGCTGATTATTCTGTCCTATATAATCTCGCACCATTTTTCACTTTTAAGTGTT
3402 ATGGAGTGTCTCCTACTAAATTAAATGATCTCTGCTTTACTAATGTCTATGCAGATTCATTTGTAATTAGAGGTGATGAA
3403 GTCAGACAAATCGCTCCAGGGCAAACTGGAAATATTGCTGATTATAATTATAAATTACCAGATGATTTTACAGGCTGCGT
3404 TATAGCTTGGAATTCTAACAAGCTTGATTCTAAGGTTAGTGGTAATTATAATTACCTGTATAGATTGTTTAGGAAGTCTA
3405 ATCTCAAACCTTTTGAGAGAGATATTTCAACTGAAATCTATCAGGCCGGTAACAAACCTTGTAATGGTGTTGCAGGTTTT
3406 AATTGTTACTTTCCTTTACGATCATATAGTTTCCGACCCACTTATGGTGTTGGTCACCAACCATACAGAGTAGTAGTACT
3407 TTCTTTTGAACTTCTACAT
3408 >hCoV-19/Iran/NIC14010110SMN2/2022|EPI_ISL_13369295|2022-03
3409 TTGGAACAGGAAGAGAATCAGCAACTGTGTTGCTGATTATTCTGTCCTATATAATCTCGCACCATTTTTCACTTTTAAGT
3410 GTTATGGAGTGTCTCCTACTAAATTAAATGATCTCTGCTTTACTAATGTCTATGCAGATTCATTTGTAATTAGAGGTGAT
3411 GAAGTCAGACAAATCGCTCCAGGGCAAACTGGAAATATTGCTGATTATAATTATAAATTACCAGATGATTTTACAGGCTG
3412 CGTTATAGCTTGGAATTCTAACAAGCTTGATTCTAAGGTTAGTGGTAATTATAATTACCTGTATAGATTGTTTAGGAAGT
3413 CTAATCTCAAACCTTTTGAGAGAGATATTTCAACTGAAATCTATCAGGCCGGTAACAAACCTTGTAATGGTGTTGCAGGT
3414 TTTAATTGTTACTTTCCTTTACGATCATATAGTTTCCGACCCACTTATGGTGTTGGTCACCAACCATACAGAGTAGTAGT
3415 ACTTTCTTTTGAACTTCTACATGCACCAGCAACTGTTTGTGGACCTAAAAAGTCTACTAATTTGGTTAAAAACAAATGTG
3416 TCAATT
3417 >hCoV-19/Iran/NIC14001116GLN1/2022|EPI_ISL_13370890|2022-02
3418 GAACAGGAAGAGAATCAGCAACTGTGTTGCTGATTATTCTGTCCTATATAATCTCGCACCATTTTTCACTTTTAAGTGTT
3419 ATGGAGTGTCTCCTACTAAATTAAATGATCTCTGCTTTACTAATGTCTATGCAGATTCATTTGTAATTAGAGGTGATGAA
3420 GTCAGACAAATCGCTCCAGGGCAAACTGGAAATATTGCTGATTATAATTATAAATTACCAGATGATTTTACAGGCTGCGT
3421 TATAGCTTGGAATTCTAACAAGCTTGATTCTAAGGTTAGTGGTAATTATAATTACCTGTATAGATTGTTTAGGAAGTCTA
3422 ATCTCAAACCTTTTGAGAGAGATATTTCAACTGAAATCTATCAGGCCGGTAACAAACCTTGTAATGGTGTTGCAGGTTTT
3423 AATTGTTACTTTCCTTTACGATCATATAGTTTCCGACCCACTTATGGTGTTGGTCACCAACCATACAGAGTAGTAGTACT
3424 TTCTTTTGAACTTCTACATGCACCAGCAACTGTTTGTGGACCTAAAAAGTCTACTAATTTGGTTAAAAACAAA
```

```
5311 >hCoV-19/Iran/Mashhad-NIC-25mo-23/2022|EPI_ISL_14806953|2022-08-11
5312 ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAA
5313 AATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAACTAATTACTGTCGTTGACAGG
5314 ACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTGTTGCAGCCGATCATCAGCACATCTAGGTTT
5315 TGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCCCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGC
5316 CTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACAT
5317 CTTAAAGATGGCACTTGTGGCTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAA
5318 ACGTTCGGATGCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTC
5319 GTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCTTCTTCGTAAG
5320 AACGGTAATAAAGGAGCTGGTGGCCATAGGTACGGCGCCGATCTAAAGTCATTTGACTTAGGCGACGAGCTTGGCACTGA
5321 TCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTGTTACCCGTGAACTCATGCGTGAGCTTAACG
5322 GAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGGCCCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTA
5323 GCACGTGCTGGTAAAGCTTCATGCACTTTGTCCGAACAACTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCG
5324 TGAACATGAGCATGAAATTGCTTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCAGACACCTTTTGAAATTAAAT
5325 TGGCAAAGAAATTTGACACCTTCAATGGGGAATGTCCAAATTTTGTATTTCCCTTAAATTCCATAATCAAGACTATTCAA
5326 CCAAGGGTTGAAAAGAAAAAGCTTGATGGCTTTATGGGTAGAATTCGATCTGTCTATCCAGTTGCGTCACCAAATGAATG
5327 CAACCAAATGTGCCTTTCAACTCTCATGAAGTGTGATCATTGTGGTGAAACTTCATGGCAGACGGGCGATTTTGTTAAAG

41060 AATGACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
41061 >hCoV-19/Iran/Gilan-NIC-6sh-4/2022|EPI_ISL_15013480|2022-08-21
41062 ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAA
41063 AATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAACTAATTACTGTCGTTGACAGG
41064 ACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTGTTGCAGCCGATCATCAGCACATCTAGGTTT
41065 TGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCCCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGC
41066 CTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACAT
41067 CTTAAAGATGGCACTTGTGGCTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAA
41068 ACGTTCGGATGCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTC
41069 GTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCTTCTTCGTAAG
41070 AACGGTAATAAAGGAGCTGGTGGCCATAGGTACGGCGCCGATCTAAAGTCATTTGACTTAGGCGACGAGCTTGGCACTGA
41071 TCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTGTTACCCGTGAACTCATGCGTGAGCTTAACG
41072 GAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGGCCCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTA
41073 GCACGTGCTGGTAAAGCTTCATGCACTTTGTCCGAACAACTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCG
41074 TGAACATGAGCATGAAATTGCTTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCAGACACCTTTTGAAATTAAAT
```

II. align the collected dataset using the MAFFT tool and check the MSA for any standard format issues, such as duplicate taxon names, invalid characters in taxon names, or duplicate sequences.

Install **MAFFT** by :

```
conda install-c bioconda mafft
```

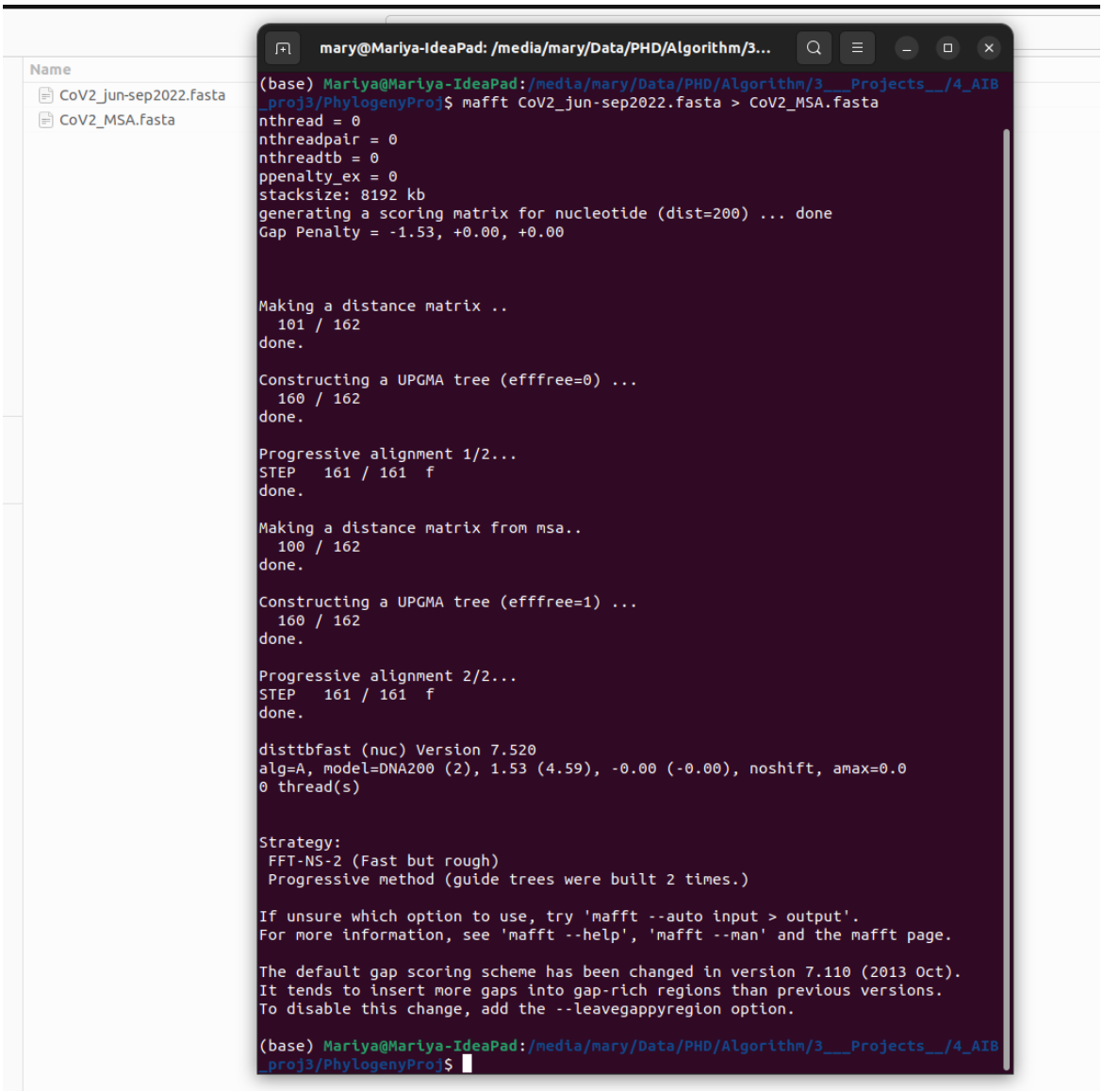
- Align the collected dataset using the MAFFT tool .

MAFFT is a popular **multiple sequence alignment tool** that is widely used for aligning DNA or protein sequences. aligning the dataset using the MAFFT tool is a common step in preparing the data for phylogenetic analysis.

This command will generate an output file called "**CoV2_MSA.fasta**" that contains the MSA.

```
PhylogenyProj$ mafft CoV2_jun-sep2022.fasta > CoV2_MSA.fasta
nthread = 0
nthreadpair = 0
nthreadtb = 0
ppenalty_ex = 0
stacksize: 8192 kb
generating a scoring matrix for nucleotide (dist=200) ... done
Gap Penalty = -1.53, +0.00, +0.00

Making a distance matrix ..
  101 / 162
done.

Constructing a UPGMA tree (efffree=0) ...
  160 / 162
done.
```

```
Progressive alignment 1/2...
STEP   161 / 161  f
done.

Making a distance matrix from msa..
  100 / 162
done.

Constructing a UPGMA tree (efffree=1) ...
  160 / 162
done.

Progressive alignment 2/2...
STEP   161 / 161  f
done.

disttbfast (nuc) Version 7.520
alg=A, model=DNA200 (2), 1.53 (4.59), -0.00 (-0.00), noshift, amax=0.0
0 thread(s)


Strategy:
 FFT-NS-2 (Fast but rough)
 Progressive method (guide trees were built 2 times.)

If unsure which option to use, try 'mafft --auto input > output'.
```



Visualizing multiple sequence alignments

Jalview is a free software package that is commonly used for visualizing and analyzing multiple sequence alignments. **CoV2_MSA.fasta** file are shown here.

- Check the MSA for any standard format issues, such as duplicate taxon names, invalid characters in taxon names, or duplicate sequences

for this part i used `RaxML-NG`. Installing `"RaxML-NG"` by conda.

```
conda install -c bioconda raxml-ng
```

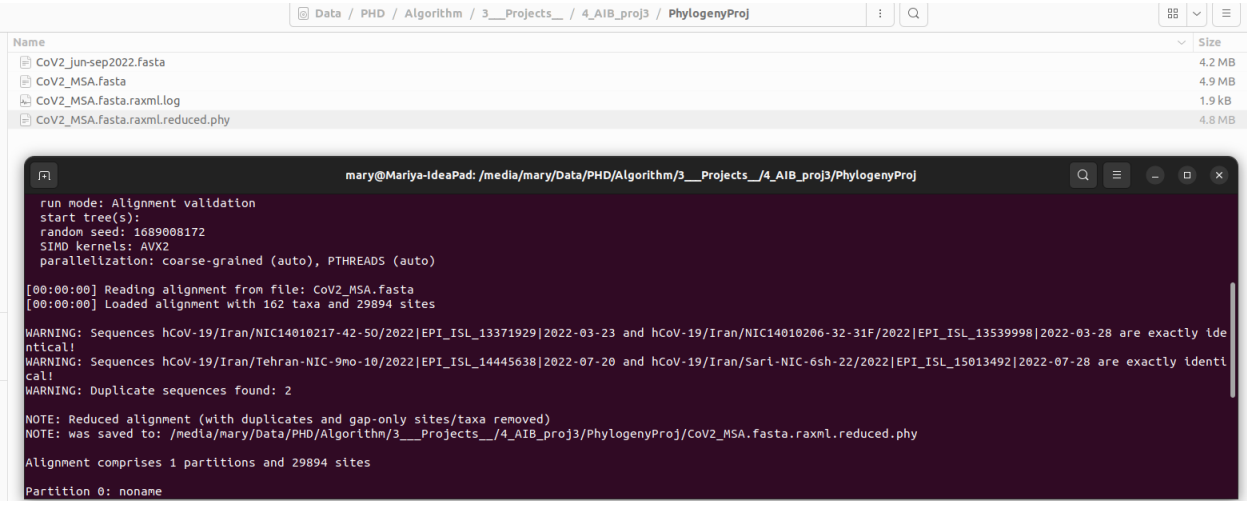
Before starting the actual analysis, performing a multiple sequence alignment (MSA) sanity check by calling RAxML-NG with the --check option is required to check duplicate taxon names, invalid characters in taxon names, or duplicate sequences.

MSA will check for several common format issues as well as data inconsistencies including:
-duplicate taxon names
-invalid characters in taxon names
-duplicate sequences
-fully undetermined (“gap-only”) sequences and columns
-incorrect or incompatible evolutionary models, partitioning scheme and starting trees

```
raxml-ng --check --msa CoV2_MSA.fasta --model GTR+G

###################################################
Analysis options:
  run mode: Alignment validation
  start tree(s):
  random seed: 1689008172
  SIMD kernels: AVX2
  parallelization: coarse-grained (auto), PTHREADS (auto)

[00:00:00] Reading alignment from file: CoV2_MSA.fasta
[00:00:00] Loaded alignment with 162 taxa and 29894 sites

WARNING: Sequences hCoV-19/Iran/NIC14010217-42-5O/2022|EPI_ISL_13371929
|2022-03-23 and hCoV-19/Iran/NIC14010206-32-31F/2022|EPI_ISL_13539998|
2022-03-28 are exactly identical!
WARNING: Sequences hCoV-19/Iran/Tehran-NIC-9mo-10/2022|EPI_ISL_14445638
|2022-07-20 and hCoV-19/Iran/Sari-NIC-6sh-22/2022|EPI_ISL_15013492|
2022-07-28 are exactly identical!
WARNING: Duplicate sequences found: 2
```

```
NOTE: Reduced alignment (with duplicates and gap-only sites/taxa removed)
NOTE: was saved to: /media/mary/Data/PHD/Algorithm/3___Projects__/4_AIB_proj3/
PhylogenyProj/CoV2_MSA.fasta.raxml.reduced.phy

Alignment comprises 1 partitions and 29894 sites

Partition 0: noname
Model: GTR+FO+G4m
Alignment sites: 29894
Gaps: 13.99 %
Invariant sites: 98.49 %

Alignment can be successfully read by RAxML-NG.
```

Result showed that 2 Duplicate sequences found and Reduced alignment saved as “**CoV2_MSA.fasta.raxml.reduced.phy**” file.



2. **Phylogenetic analysis**

   I. Estimate the optimal nucleotide substitution model using maximum likelihood or Bayesian methods by selecting the model with the lowest AIC or BIC score (you may use phangorn package in R).

To estimate the optimal nucleotide substitution model using maximum likelihood without a phylogenetic tree, modelTest() function in the phangorn package is used.

This function will estimate the optimal nucleotide substitution model based on the AIC or BIC criterion.

AIC (Akaike Information Criterion), AICc (corrected AIC), and BIC (Bayesian Information Criterion) are statistical measures used for model selection in various fields, including phylogenetics. These measures are used to compare the fit of different models to a given dataset, with the goal of selecting the model that best explains the data while avoiding overfitting. AIC and AICc are based on likelihood theory, which measures the probability of observing the data given the model parameters and the hypothesis being tested. BIC is similar to AIC, but it penalizes models more heavily for having more parameters. In phylogenetics, AIC, AICc, and BIC are commonly used to compare different models of nucleotide (or amino acid) substitution, which describe the process of evolution along a phylogenetic tree. The goal is to select the model that best explains the observed patterns of sequence variation, given the available data. AIC, AICc, and BIC can also be used to compare different phylogenetic trees inferred from the same data, to determine which tree is the best fit for the data under the selected model of evolution.

I used R for following steps.

`1-READ data` :**CoV2_MSA.fasta.raxml.reduced.phy**

`2-create substitution models` by modelTest()

```
library(ape)
library(phangorn)

# Load the nucleotide sequence data
sequences <- read.phyDat("CoV2_MSA.fasta.raxml.reduced.phy", format = "phylip")
sequences

# estimate the optimal nucleotide substitution model using maximum likelihood by mo
ml_model <- modelTest(sequences)
```

37:1 min of all models

Console | Terminal | Background Jobs

R 4.2.2 · /media/mary/Data/PHD/Algorithm/3___Projects__/4_AIB_proj3/2_Phylogeny_Analysis_R/proj/

```
> ml_model
          Model  df    logLik      AIC          AICw     AICc         AICcw      BIC
1            JC 317 -48126.54 96887.08  0.000000e+00 96893.90  0.000000e+00 99519.90
2          JC+I 318 -47704.12 96044.23  0.000000e+00 96051.09  0.000000e+00 98685.35
3       JC+G(4) 318 -47921.27 96478.54  0.000000e+00 96485.40  0.000000e+00 99119.66
4     JC+G(4)+I 319 -47675.11 95988.21  0.000000e+00 95995.11  0.000000e+00 98637.64
5           F81 320 -47169.96 94979.92  0.000000e+00 94986.87  0.000000e+00 97637.65
6         F81+I 321 -46760.36 94162.71 1.288284e-154 94169.70 1.470200e-154 96828.75
7      F81+G(4) 321 -46965.89 94573.78 7.053634e-244 94580.77 8.049663e-244 97239.82
8    F81+G(4)+I 322 -46731.23 94106.46 2.111506e-142 94113.50 2.357478e-142 96780.81
9           K80 318 -47859.62 96355.25  0.000000e+00 96362.11  0.000000e+00 98996.37
10        K80+I 319 -47436.58 95511.15  0.000000e+00 95518.06  0.000000e+00 98160.58
11     K80+G(4) 319 -47654.29 95946.59  0.000000e+00 95953.49  0.000000e+00 98596.02
12   K80+G(4)+I 320 -47407.39 95454.78  0.000000e+00 95461.72  0.000000e+00 98112.51
13          HKY 321 -46885.23 94412.46 7.575749e-209 94419.45 8.645504e-209 97078.49
14        HKY+I 322 -46491.05 93627.00  1.751324e-38 93634.03  1.955338e-38 96302.24
```

3-Select JC-K80-GTR

```
# Subset ml_model to extract rows with Model = "JC"
jc_model <- subset(ml_model, Model == "JC")
print(jc_model)

# Subset ml_model to extract rows with Model = "K80"
K80_model <- subset(ml_model, Model == "K80")
print(K80_model)

# Subset ml_model to extract rows with Model = "GTR"
GTR_model <- subset(ml_model, Model == "GTR")
print(GTR_model)
```

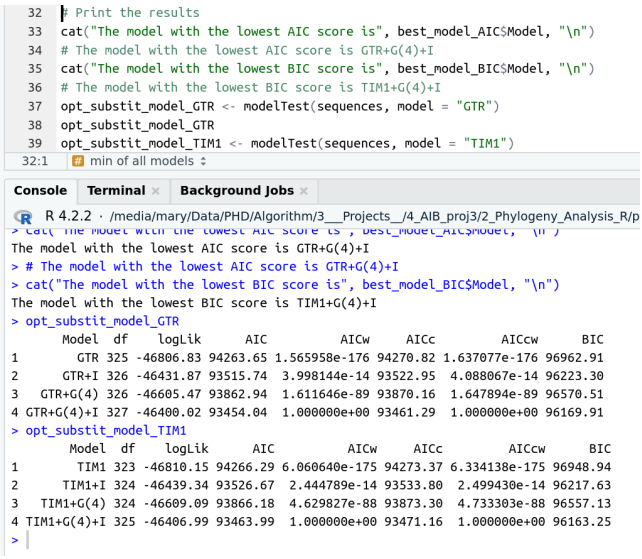
26:1 min of all models

Console | Terminal | Background Jobs

R 4.2.2 · /media/mary/Data/PHD/Algorithm/3___Projects__/4_AIB_proj3/2_Phylogeny_Analysis

```
> K80_model <- subset(ml_model, Model == "K80")
> # Subset ml_model to extract rows with Model = "GTR"
> GTR_model <- subset(ml_model, Model == "GTR")
> print(jc_model)
  Model  df    logLik      AIC AICw    AICc AICcw     BIC
1    JC 317 -48126.54 96887.08    0 96893.9     0 99519.9
> print(K80_model)
  Model  df    logLik      AIC AICw     AICc AICcw      BIC
9   K80 318 -47859.62 96355.25    0 96362.11     0 98996.37
> print(GTR_model)
   Model  df    logLik      AIC          AICw     AICc         AICcw      BIC
89   GTR 325 -46806.83 94263.65 1.554669e-176 94270.82 1.624744e-176 96962.91
>
```

4- Selecting the model with the lowest AIC or BIC score




```
library(ape)
library(phangorn)
# Load the nucleotide sequence data
sequences <- read.phyDat("CoV2_MSA.fasta.raxml.reduced.phy", format = "phylip")
sequences
# estimate the optimal nucleotide substitution model using maximum likelihood by modelTest()
ml_model <- modelTest(sequences)
ml_model
# Subset ml_model to extract rows with Model = "JC"
jc_model <- subset(ml_model, Model == "JC")
print(jc_model)
# Subset ml_model to extract rows with Model = "K80"
K80_model <- subset(ml_model, Model == "K80")
print(K80_model)

# Subset ml_model to extract rows with Model = "GTR"
GTR_model <- subset(ml_model, Model == "GTR")
print(GTR_model)

########################## min of all models ################################
# Find the model with the lowest AIC score
best_model_AIC <- ml_model[which.min(ml_model$AIC), ]

# Find the model with the lowest BIC score
best_model_BIC <- ml_model[which.min(ml_model$BIC), ]
# Print the results
cat("The model with the lowest AIC score is", best_model_AIC$Model, "\n")
# The model with the lowest AIC score is GTR+G(4)+I
cat("The model with the lowest BIC score is", best_model_BIC$Model, "\n")
# The model with the lowest BIC score is TIM1+G(4)+I
opt_substit_model_GTR <- modelTest(sequences, model = "GTR")
opt_substit_model_GTR
opt_substit_model_TIM1 <- modelTest(sequences, model = "TIM1")
opt_substit_model_TIM1
```

II.Calculate the pairwise distance between the sequences in the alignment.

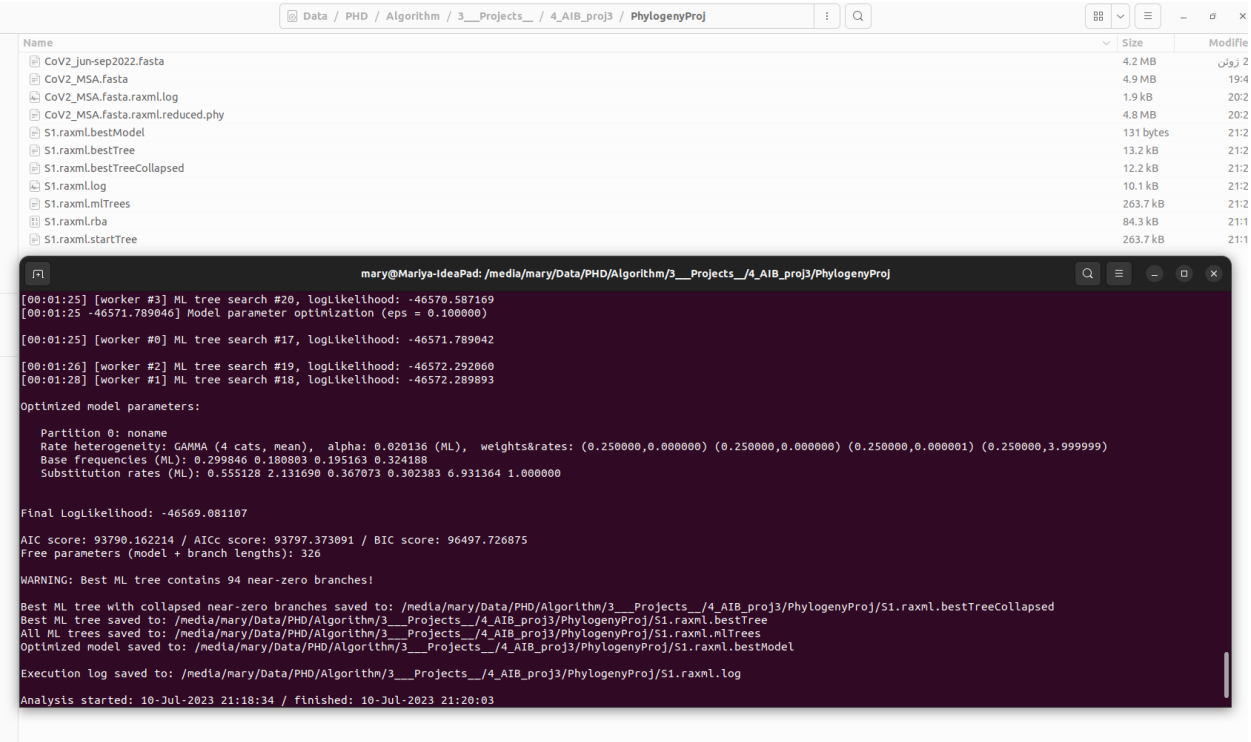
To calculate pairwise distances using raxml-ng, raxml-ng rfdist command calculates the Robinson-Foulds distance between two trees.The Robinson-Foulds distance is a measure of the topological distance between two trees and can be used as an estimate of the pairwise distance between sequences.

To calculate pairwise distances using `raxml-ng`, following command calculate `pairwise distance` file.

Tree search

this command will **perform 20 tree searches** using 10 random and 10 parsimony-based starting trees. **In the end it will pick the best-scoring topology**

```
 raxml-ng --msa CoV2_MSA.fasta.raxml.reduced.phy --model GTR+G --prefix S1

####################################################################
.................

[00:01:25] [worker #0] ML tree search #17, logLikelihood: -46571.789042

[00:01:26] [worker #2] ML tree search #19, logLikelihood: -46572.292060
[00:01:28] [worker #1] ML tree search #18, logLikelihood: -46572.289893

Optimized model parameters:

   Partition 0: noname
   Rate heterogeneity: GAMMA (4 cats, mean),  alpha: 0.020136 (ML),  weights&rates: (0.250000,0.000000) (0.250000,0.000000) (0.250000,0.000
   Base frequencies (ML): 0.299846 0.180803 0.195163 0.324188
   Substitution rates (ML): 0.555128 2.131690 0.367073 0.302383 6.931364 1.000000

Final LogLikelihood: -46569.081107

AIC score: 93790.162214 / AICc score: 93797.373091 / BIC score: 96497.726875
Free parameters (model + branch lengths): 326

WARNING: Best ML tree contains 94 near-zero branches!

Best ML tree with collapsed near-zero branches saved to:
S1.raxml.bestTreeCollapsed

Best ML tree saved to: S1.raxml.bestTree

All ML trees saved to: S1.raxml.mlTrees

Optimized model saved to: S1.raxml.bestModel

Execution log saved to: S1.raxml.log
```

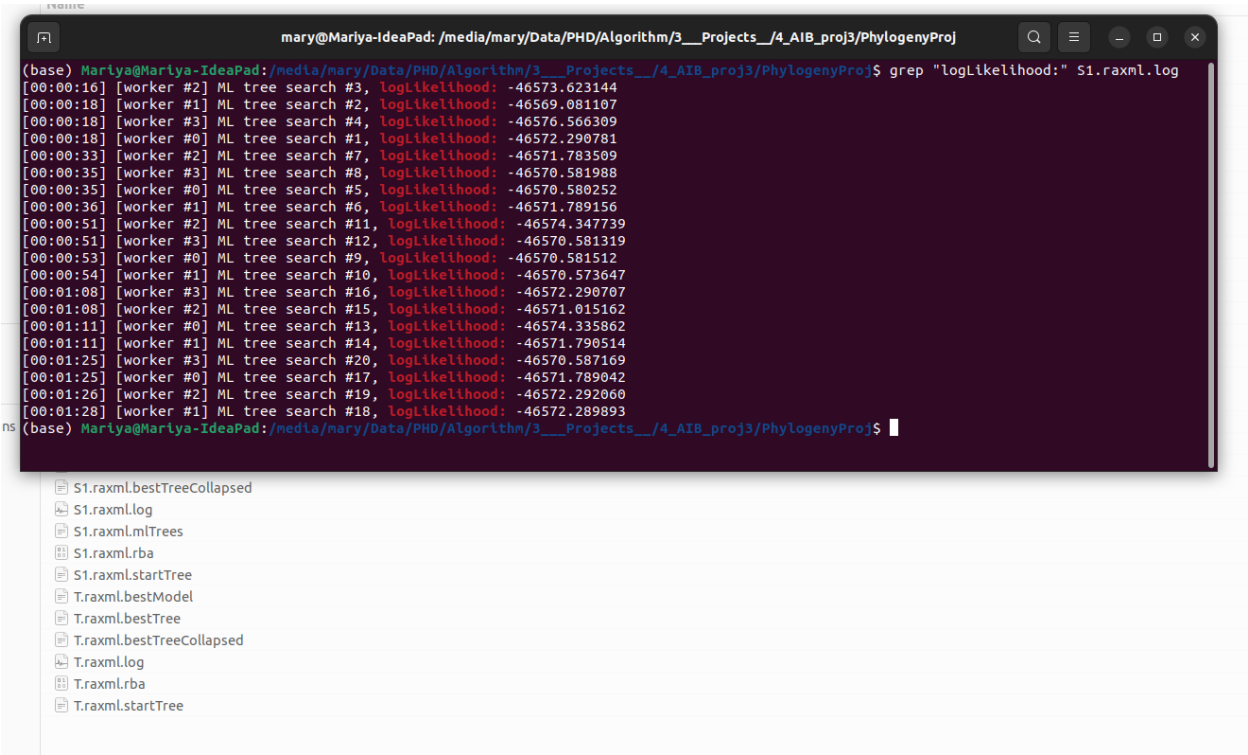


Best ML tree saved to: S1.raxml.bestTree
Optimized model saved to: S1.raxml.bestModel
Execution log saved to: S1.raxml.log

All ML trees saved to: S1.raxml.mlTrees

The log-likelihood value is a measure of the fit between a given tree and the multiple sequence alignment data. In general, trees with higher log-likelihood values are considered to be more likely to represent the true evolutionary history of the analyzed sequences.

loglikelihood of tree are:



the --rfdist command compute the topological Robinson-Foulds (RF) distance (Robinson and Foulds, 1981) between **all trees we have inferred from last step (S1.raxml.mlTrees):**

```
raxml-ng --rfdist --tree S1.raxml.mlTrees --prefix RF
```

`input file:` "S1.raxml.mlTrees"

`output:` "RF.raxml.rfDistances

```
(base) Mariya@Mariya-IdeaPad:/media/mary/Data/PHD/Algorithm/3___Projects__/4_AIB_proj3/PhylogenyProj$ ls
CoV2_jun-sep2022.fasta  CoV2_MSA.fasta.raxml.log          S1.raxml.bestModel  S1.raxml.bestTreeCollapsed  S1.raxml.mlTrees  S1.raxml.startTree
CoV2_MSA.fasta          CoV2_MSA.fasta.raxml.reduced.phy  S1.raxml.bestTree   S1.raxml.log                S1.raxml.rba
(base) Mariya@Mariya-IdeaPad:/media/mary/Data/PHD/Algorithm/3___Projects__/4_AIB_proj3/PhylogenyProj$ raxml-ng --rfdist --tree S1.raxml.mlTrees --prefix RF

RAxML-NG v. 1.2.0 released on 09.05.2023 by The Exelixis Lab.
Developed by: Alexey M. Kozlov and Alexandros Stamatakis.
Contributors: Diego Darriba, Tomas Flouri, Benoit Morel, Sarah Lutteropp, Ben Bettisworth, Julia Haag, Anastasis Togkousidis.
Latest version: https://github.com/amkozlov/raxml-ng
Questions/problems/suggestions? Please visit: https://groups.google.com/forum/#!forum/raxml

System: Intel(R) Core(TM) i5-10300H CPU @ 2.50GHz, 4 cores, 15 GB RAM

RAxML-NG was called at 10-Jul-2023 21:28:23 as follows:

raxml-ng --rfdist --tree S1.raxml.mlTrees --prefix RF

Analysis options:
  run mode: RF distance computation
  start tree(s): user
  random seed: 1689011903
  SIMD kernels: AVX2
  parallelization: coarse-grained (auto), PTHREADS (auto)

Reading input trees from file: S1.raxml.mlTrees
Loaded 20 trees with 160 taxa.

Average absolute RF distance in this tree set: 176.936842
Average relative RF distance in this tree set: 0.563493
Number of unique topologies in this tree set: 20

Pairwise RF distances saved to: /media/mary/Data/PHD/Algorithm/3___Projects__/4_AIB_proj3/PhylogenyProj/RF.raxml.rfDistances

Execution log saved to: /media/mary/Data/PHD/Algorithm/3___Projects__/4_AIB_proj3/PhylogenyProj/RF.raxml.log

Analysis started: 10-Jul-2023 21:28:23 / finished: 10-Jul-2023 21:28:23

Elapsed time: 0.013 seconds

(base) Mariya@Mariya-IdeaPad:/media/mary/Data/PHD/Algorithm/3___Projects__/4_AIB_proj3/PhylogenyProj$ ls
CoV2_jun-sep2022.fasta  CoV2_MSA.fasta.raxml.log          RF.raxml.log          S1.raxml.bestModel  S1.raxml.bestTreeCollapsed  S1.raxml.mlTrees  S1.raxml.startTree
CoV2_MSA.fasta          CoV2_MSA.fasta.raxml.reduced.phy  RF.raxml.rfDistances  S1.raxml.bestTree   S1.raxml.log                S1.raxml.rba
(base) Mariya@Mariya-IdeaPad:/media/mary/Data/PHD/Algorithm/3___Projects__/4_AIB_proj3/PhylogenyProj$
```

```
raxml-ng --rfdist --tree S1.raxml.mlTrees --prefix RF

Analysis options:
  run mode: RF distance computation
  start tree(s): user
  random seed: 1689011903
  SIMD kernels: AVX2
  parallelization: coarse-grained (auto), PTHREADS (auto)

Reading input trees from file: S1.raxml.mlTrees
Loaded 20 trees with 160 taxa.

Average absolute RF distance in this tree set: 176.936842
Average relative RF distance in this tree set: 0.563493
Number of unique topologies in this tree set: 20

Pairwise RF distances saved to: /media/mary/Data/PHD/Algorithm/3___Projects__/4_AIB_proj3/PhylogenyProj/RF.raxml.rfDistances

Execution log saved to: /media/mary/Data/PHD/Algorithm/3___Projects__/4_AIB_proj3/PhylogenyProj/RF.raxml.log
```

Average absolute RF distance in this tree set: 176.936842
Average relative RF distance in this tree set: 0.563493
Number of unique topologies in this tree set: 20

This tells us that, in fact, all 20 resulting topologies has RF dist 176.93 , So we have 20 distinct topologies in our set of 20 inferred trees, which correspond to distinct likelihood values we observed in the tree search output.

individual pairwise RF distances which are printed to the RF6.raxml.rfDistances file:

```
(base) Mariya@Mariya-IdeaPad:/media/mary/Data/PHD/Algorithm/3___Projects__/4_AIB_proj3/PhylogenyProj$ cat RF.raxml.
RF.raxml.log          RF.raxml.rfDistances
(base) Mariya@Mariya-IdeaPad:/media/mary/Data/PHD/Algorithm/3___Projects__/4_AIB_proj3/PhylogenyProj$ cat RF.raxml.rfDistances
0       1       186     0.592357
0       2       198     0.630573
0       3       172     0.547771
0       4       184     0.585987
0       5       172     0.547771
0       6       180     0.573248
0       7       182     0.579618
0       8       186     0.592357
0       9       178     0.566879
0       10      192     0.611465
0       11      180     0.573248
0       12      204     0.649682
0       13      182     0.579618
0       14      184     0.585987
0       15      162     0.515924
0       16      186     0.592357
0       17      160     0.509554
0       18      166     0.528662
0       19      186     0.592357
1       2       180     0.573248
1       3       190     0.605096
1       4       176     0.560510
1       5       170     0.541401
1       6       168     0.535032
1       7       184     0.585987
1       8       164     0.522293
```

Here, 1st and 2nd column contain tree indices in the NEWICK file, **3rd column shows the absolute RF distance between those two trees,** and 4th column shows the relative or normalized RF distance ranging from 0 to 1

III.Generate a first tree topology based on pairwise distances.

This command will use the pairwise distances calculated by `raxml-ng rfdist` to **generate a tree topology** based on the GTR model of nucleotide substitution.

The `--msa` option specifies the input alignment file,

The `--search` option specifies the search algorithm to use for tree inference.

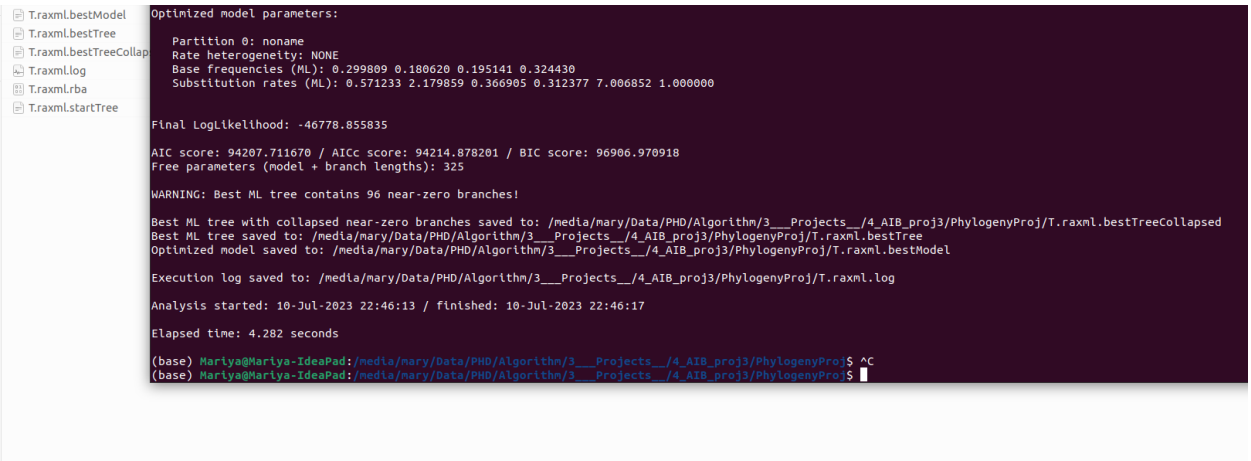
The `--model` option specifies the evolutionary model to use.

The `--tree` option specifies the starting tree for the tree search.

The `--prefix` option specifies the prefix to use for all output files.

After running this command, `raxml-ng rtree` will generate a tree topology based on the pairwise distances and save it to a file named "**T.raxml.bestTree**" in the current working directory.

```
raxml-ng rtree --msa CoV2_MSA.fasta.raxml.reduced.phy
 --search rfdist --model GTR --tree pars{1} --prefix T
```



IV.Optimize the phylogenetic tree topology, branch lengths, and nucleotide substitution model.

To optimize the phylogenetic tree topology, branch lengths, and nucleotide substitution model, **`raxml-ng`** perform a full **maximum likelihood analysis**. This will involve **searching for the best tree topology, branch lengths, and nucleotide substitution model that maximize the likelihood score based on the input alignment.**

```
raxml-ng --msa CoV2_MSA.fasta.raxml.reduced.phy --model
 GTR+G --tree pars{1} --prefix Cov_optim
```

This command will perform a **full maximum likelihood analysis on the input alignment file** "CoV2_MSA.fasta.raxml.reduced.phy", using the **GTR+G nucleotide substitution mode**l and the first parsimony tree as the starting tree for the tree search.

Generating 1 parsimony starting tree(s) with 160 taxa

```
[00:00:06] ML tree search #1, logLikelihood: -46570.586143


Optimized model parameters:

   Partition 0: noname
   Rate heterogeneity: GAMMA (4 cats, mean),  alpha: 0.020139 (ML),  weights&rates: (0.250000,0.000000) (0.250000,0.000000) (0.250000,0.000001) (0.250
000,3.999999)
   Base frequencies (ML): 0.299789 0.180794 0.195169 0.324249
   Substitution rates (ML): 0.555804 2.133123 0.367311 0.302659 6.955899 1.000000


Final LogLikelihood: -46570.586143

AIC score: 93793.172285 / AICc score: 93800.383162 / BIC score: 96500.736946
Free parameters (model + branch lengths): 326

WARNING: Best ML tree contains 96 near-zero branches!

Best ML tree with collapsed near-zero branches saved to: /media/mary/Data/PHD/Algorithm/3___Projects__/4_AIB_proj3/PhylogenyProj/Cov_optim.raxml.bestT
reeCollapsed
Best ML tree saved to: /media/mary/Data/PHD/Algorithm/3___Projects__/4_AIB_proj3/PhylogenyProj/Cov_optim.raxml.bestTree
Optimized model saved to: /media/mary/Data/PHD/Algorithm/3___Projects__/4_AIB_proj3/PhylogenyProj/Cov_optim.raxml.bestModel

Execution log saved to: /media/mary/Data/PHD/Algorithm/3___Projects__/4_AIB_proj3/PhylogenyProj/Cov_optim.raxml.log
```

Output files:

Best ML tree: Cov_optim.raxml.bestTreeCollapsed
Best ML tree Cov_optim.raxml.bestTree
Optimized model: Cov_optim.raxml.bestModel

Execution log: Cov_optim.raxml.log

-We can open the "**Cov_optim.raxml.bestTree**" file in R to visualize Tree.

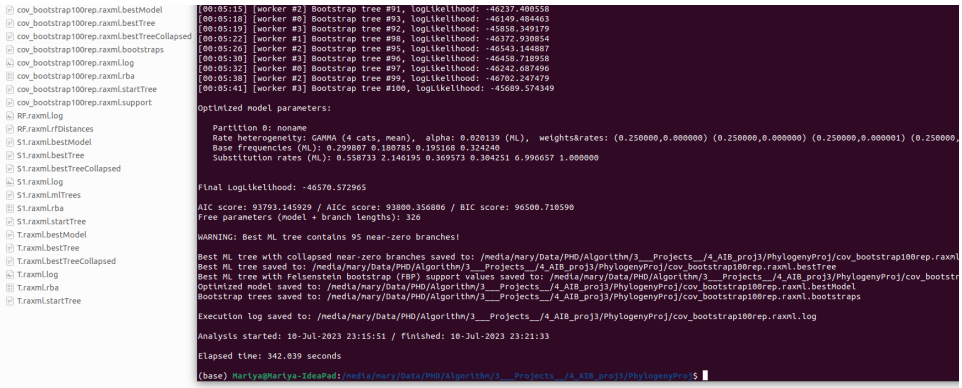
V. To assess the reliability of the phylogenetic tree, generate multiple bootstrap replicates of the dataset and estimate the phylogenetic tree for each replicate

Bootstrapping is a statistical resampling technique used in phylogenetic analysis to estimate the reliability of the inferred phylogenetic tree. The basic idea behind bootstrapping is to generate a large number of resampled datasets by randomly sampling columns (sites) from the original multiple sequence alignment (MSA) with replacement. Then, a phylogenetic tree is inferred from each resampled dataset using the same method and parameters as the original analysis. These trees are called bootstrap replicate trees.
The bootstrap method provides a measure of the support for each branch in the inferred tree, known as the bootstrap value. The bootstrap value represents the proportion of bootstrap replicate trees that contain the same branch. Generally, a bootstrap value of 70% or higher is considered to indicate strong support for a particular branch, while values between 50-70% indicate moderate support.
In the command you provided, the raxml-ng program is being used to infer bootstrap replicate trees from the multiple sequence alignment file "prim.phy" using the GTR+G model of nucleotide substitution. The "--bootstrap" option specifies that bootstrapping should be performed, and the "--prefix B1" option sets the prefix for the output files.

`Generate` `bootstrap replicates` `and estimate the phylogenetic tree for each replicate using raxml-ng, as described:` `100 replicate`

Bootstrap replicates were generated successfully, and the resulting trees are stored in the file **"cov_bootstrap100rep.raxml.bootstraps"** .

```
raxml-ng --all --bs-trees 100 --msa *.phy --model GTR+G --tree pars{1}
 --prefix cov_bootstrap100rep
```



Output files:

bootstrap tree: cov_bootstrap100rep.raxml.bootstraps
Best ML tree: cov_bootstrap100rep.raxml.bestTree
Optimized model: cov_bootstrap100rep.raxml.bestModel

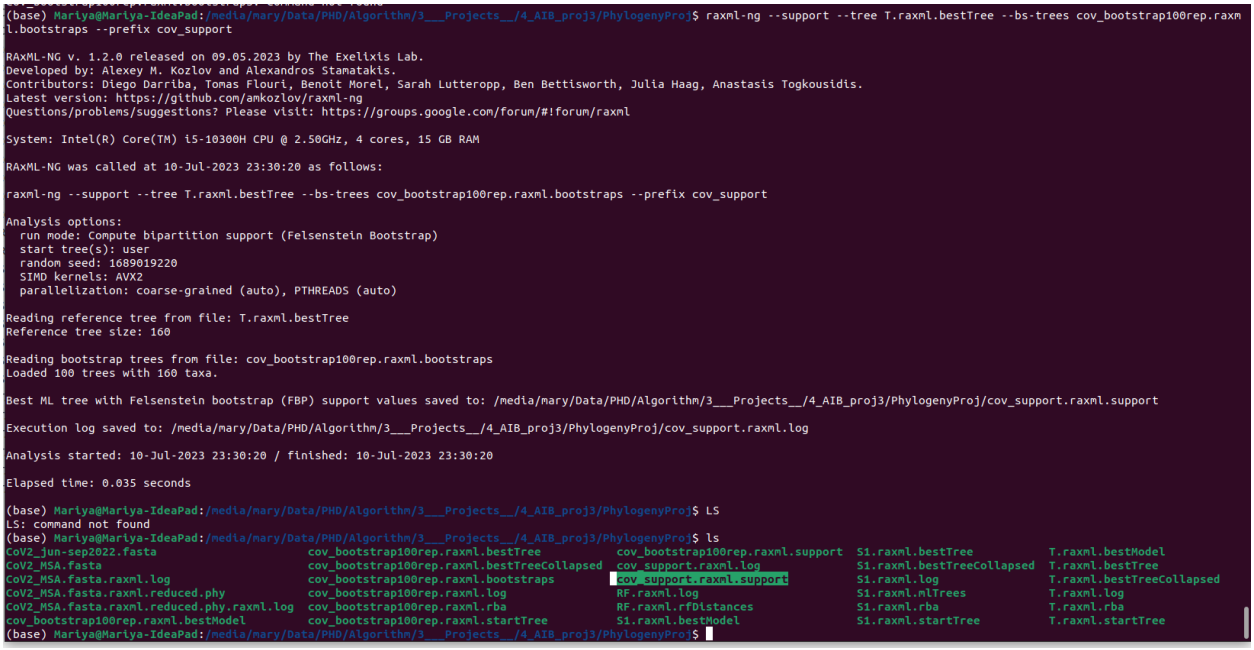
Execution log: cov_bootstrap100rep.raxml.log

VI.Infer bootstrap support for branches in the optimized tree.

Command to merge the bootstrap trees and calculate bootstrap support values for the optimized tree using raxml-ng:

```
raxml-ng --support --tree T.raxml.bestTree --bs-trees
cov_bootstrap100rep.raxml.bootstraps --prefix cov_support
```

Output file: **cov_support.raxml.support**



Map bootstrap support values to the best ML trees:

```
raxml-ng --support --tree T.raxml.bestTree --bs-trees
 cov_bootstrap100rep.raxml.bootstraps --prefix B2

###########################################################
Reading reference tree from file: T.raxml.bestTree
Reference tree size: 160

Reading bootstrap trees from file: cov_bootstrap100rep.raxml.bootstraps
Loaded 100 trees with 160 taxa.

Best ML tree with Felsenstein bootstrap (FBP) support values saved to:
TSUP.raxml.support

Execution log saved to: TSUP.raxml.log
```

Inputs:

T.raxml.bestTree --bs-trees
cov_bootstrap100rep.raxml.bootstraps (**100replicate bootstrap**)

outputs:

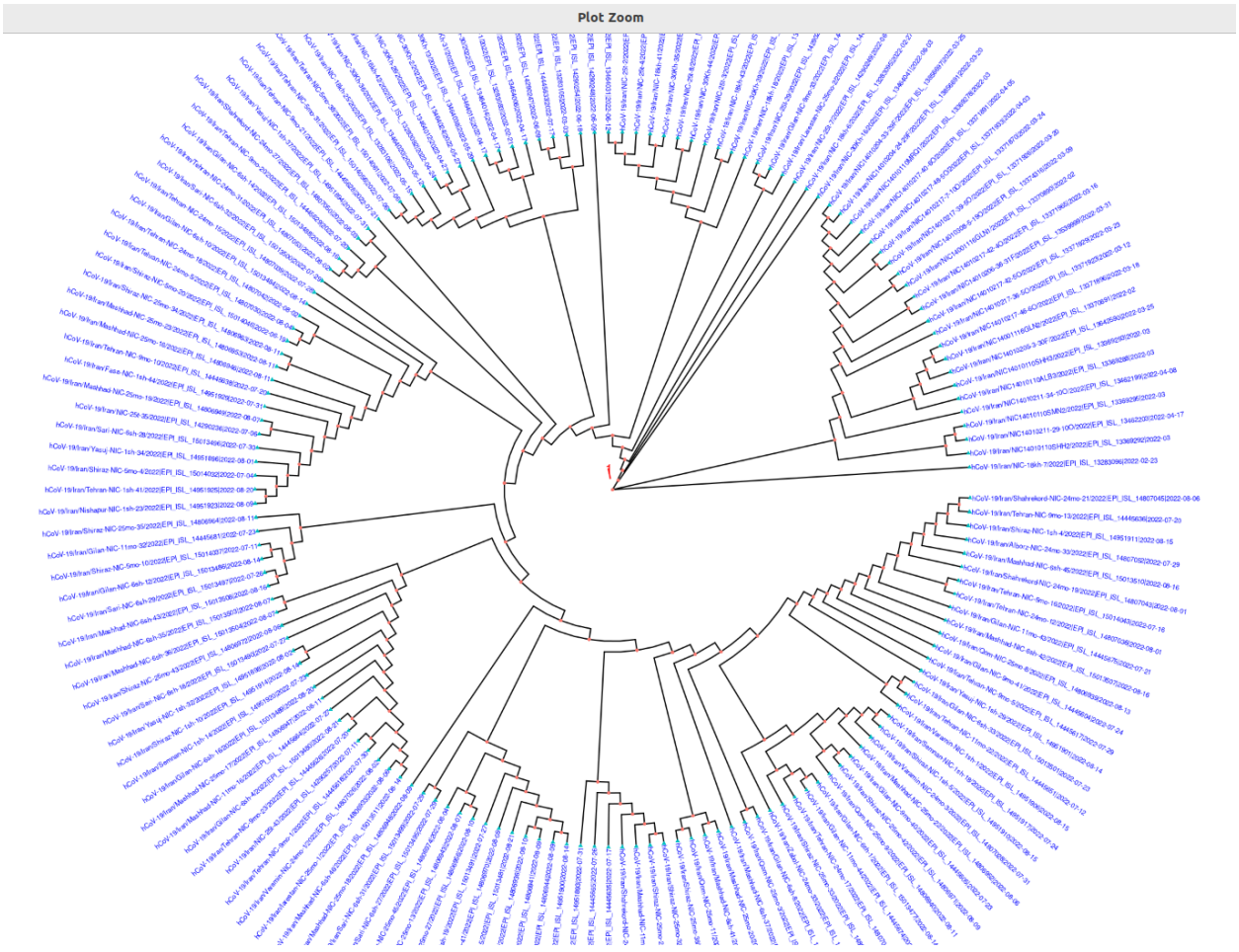
TSUP.raxml.support

```
(base) Mariya@Mariya-IdeaPad:/media/mary/Data/PHD/Algorithm/3___Projects__/4_AIB_proj3/PhylogenyProj$ raxml-ng --support --tree T.raxml.bestTree --bs-trees  cov_bootstrap100rep.raxml.bootstraps --prefix TSUP

RAxML-NG v. 1.2.0 released on 09.05.2023 by The Exelixis Lab.
Developed by: Alexey M. Kozlov and Alexandros Stamatakis.
Contributors: Diego Darriba, Tomas Flouri, Benoit Morel, Sarah Lutteropp, Ben Bettisworth, Julia Haag, Anastasis Togkousidis.
Latest version: https://github.com/amkozlov/raxml-ng
Questions/problems/suggestions? Please visit: https://groups.google.com/forum/#!forum/raxml

System: Intel(R) Core(TM) i5-10300H CPU @ 2.50GHz, 4 cores, 15 GB RAM

RAxML-NG was called at 11-Jul-2023 23:57:31 as follows:

raxml-ng --support --tree T.raxml.bestTree --bs-trees cov_bootstrap100rep.raxml.bootstraps --prefix TSUP

Analysis options:
  run mode: Compute bipartition support (Felsenstein Bootstrap)
  start tree(s): user
  random seed: 1689107251
  SIMD kernels: AVX2
  parallelization: coarse-grained (auto), PTHREADS (auto)

Reading reference tree from file: T.raxml.bestTree
Reference tree size: 160

Reading bootstrap trees from file: cov_bootstrap100rep.raxml.bootstraps
Loaded 100 trees with 160 taxa.

Best ML tree with Felsenstein bootstrap (FBP) support values saved to: /media/mary/Data/PHD/Algorithm/3___Projects__/4_AIB_proj3/PhylogenyProj/TSUP.raxml.support

Execution log saved to: /media/mary/Data/PHD/Algorithm/3___Projects__/4_AIB_proj3/PhylogenyProj/TSUP.raxml.log

Analysis started: 11-Jul-2023 23:57:31 / finished: 11-Jul-2023 23:57:31

Elapsed time: 0.132 seconds
```

3. **Data visualization**
   Use the phylogenetic tree and bootstrap support values to visualize the evolutionary history of the selected strain in Iran. (you may use ggtree package in R)

100replicate bootstrap

```
library(ggtree)

# Read Maping file of bootstrap support values to the best ML tree
tree <- read.tree("TSUP.raxml.support")

# Plot the tree with support values using ggtree
ggtree(tree, layout="circular", branch.length='none') +
  geom_treescale(x=1, y=45, width=1, color='red')+
  geom_tiplab(size=2, color='red',aes(angle=angle)) +
  # geom_tippoint(aes(size=label), shape=1, fill='gray') +
  theme(legend.position="none")+
 geom_point(aes(shape=isTip, color=isTip), size=1)

ggsave("TSUP.png", width = 30, height = 30, dpi = 300,limitsize = FALSE)
```


Plot Zoom

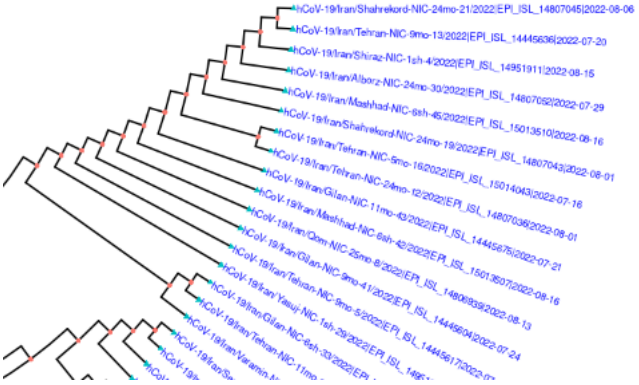
The phylogenetic tree is a rooted tree with 160 tips and 158 internal nodes, indicating it represents the evolutionary relationships among a set of 160 taxa (e.g., viral samples) based on genetic sequence data.

Each tip label corresponds to a specific sample and includes information about the sample name, the originating country (Iran), the laboratory (NIC), and the date of the sample collection. For example, "hCoV-19/Iran/NIC-18kh-7/2022|EPI_ISL_13283096|2022-02-23" indicates that the sample was collected on February 23, 2022, and was sequenced in the laboratory named NIC with the sample name hCoV-19/Iran/NIC-18kh-7.

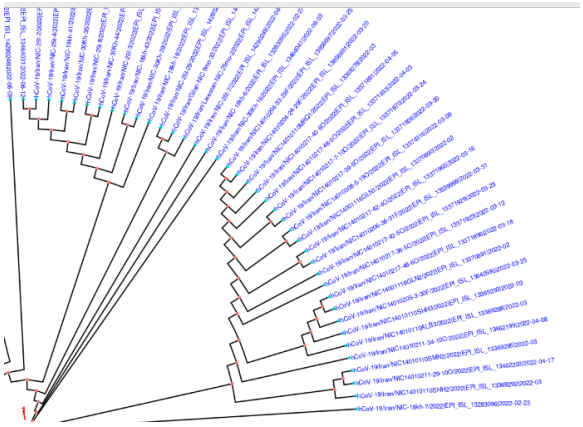
The branch lengths on the tree indicate the amount of genetic divergence between the sequences of the taxa connected by the branch. Longer branches generally indicate greater genetic distance and divergence between the sequences.

استرین مشاهده شده از ویروس در ماه ۸ ام ۲۰۲۲ در شاهرود با استرین مشاهده شده از ویروس درماه ۷ ام در تهران فاصله کمی داشته و دی یک کلاستر قرار دارند و درنتیجه به هم نزدیک بوده و جهش کمی داشته است .

ولی این استرین با استرین مشاهده شده از ویروس ازگیلان و یاسوج جهش ژنتیکی بیشتری داشته و در فاصله دورتری در یک کلاستر قرار دارند.

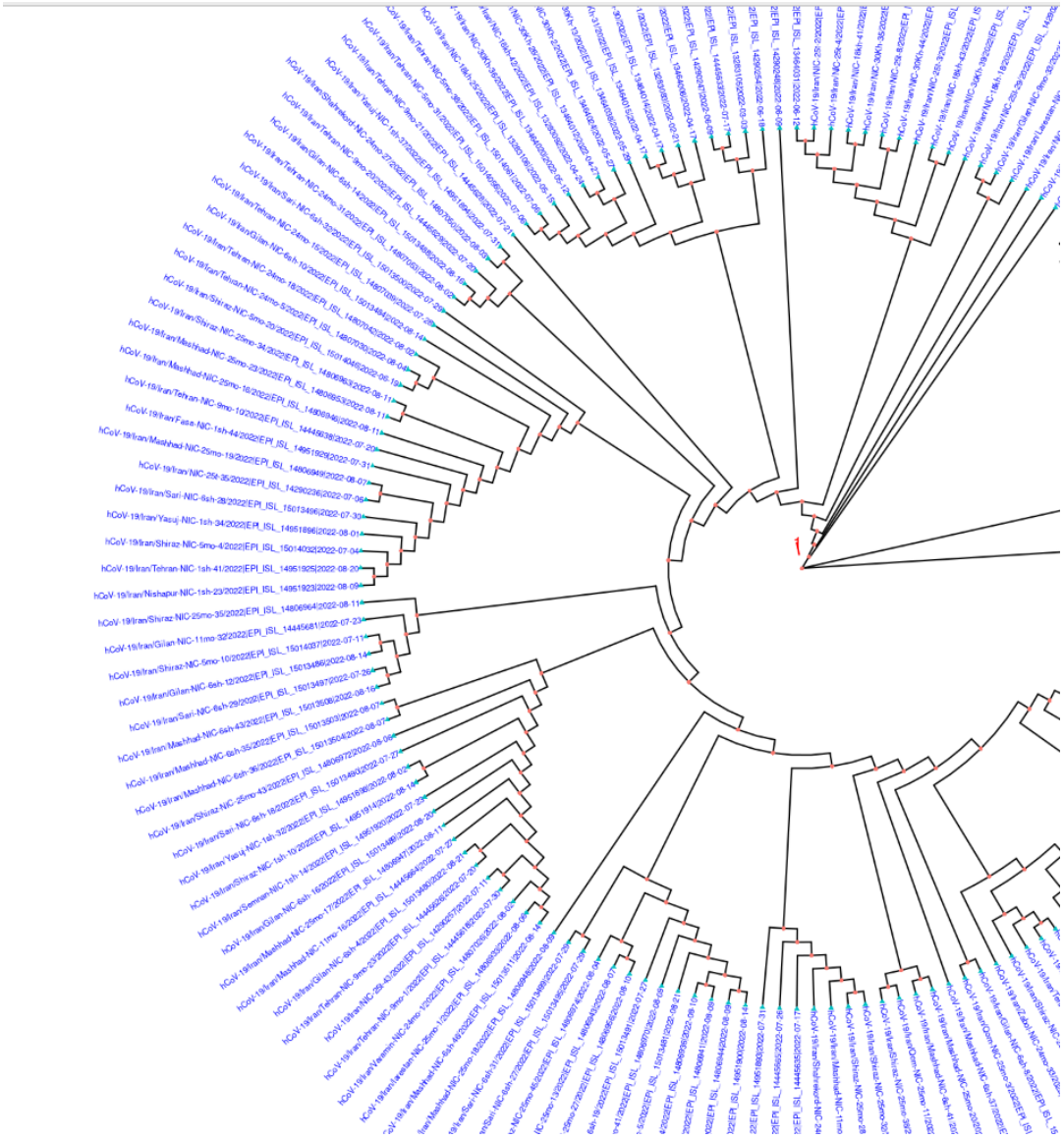


در سمت راست همه استرین های ویروس مربوط به ماه سوم و چهارم ۲۰۲۲ هستند که در یک کلاستر بزرگ نسبت به کلا کلاستر درخت قرار گرفته اند. این نشان میدهد در ماه سوم و چهارم جهش ژنتیکی ویروس تقریبا کم بوده است.

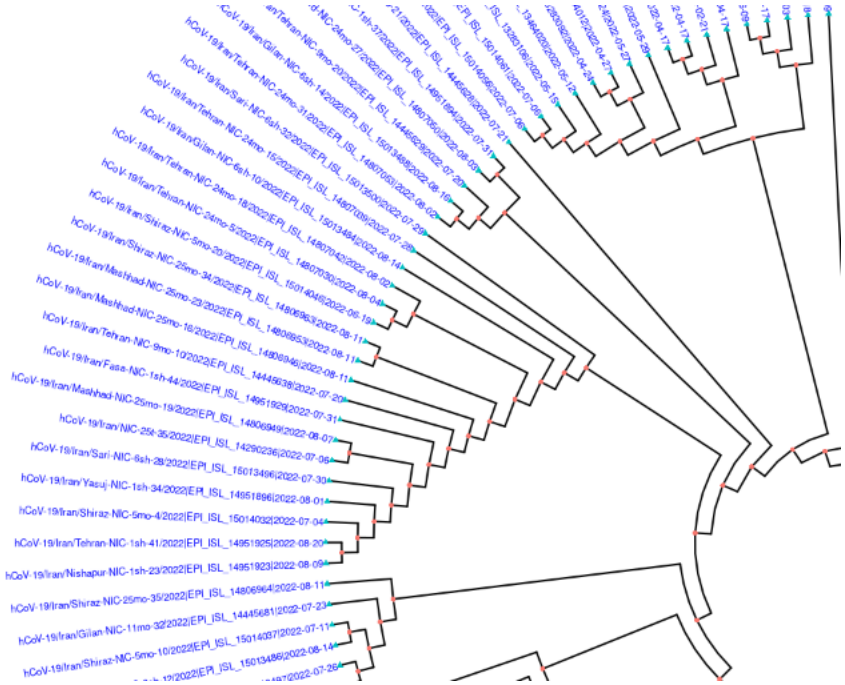


ولی در ماه های ۷ ام و ۸ ام ۲۰۲۲ ویروس جهش های زیادی داشته است . به طور کلی درخت فیلوژنی ما دو گروه نسبتا بزرگ دارد که یک گروه مربوط به ماه های ۳و ۴ است و گروه دیگر سایر ماه های به خصوص ماه ۷ ام و ۸ ام سال ۲۰۲۲

این نشان میدهد نرج جهش ویروس در ماه های ۷ ام و ۸ ام بسیار زیاد بوده نسبت به ماه سوم و چهارم



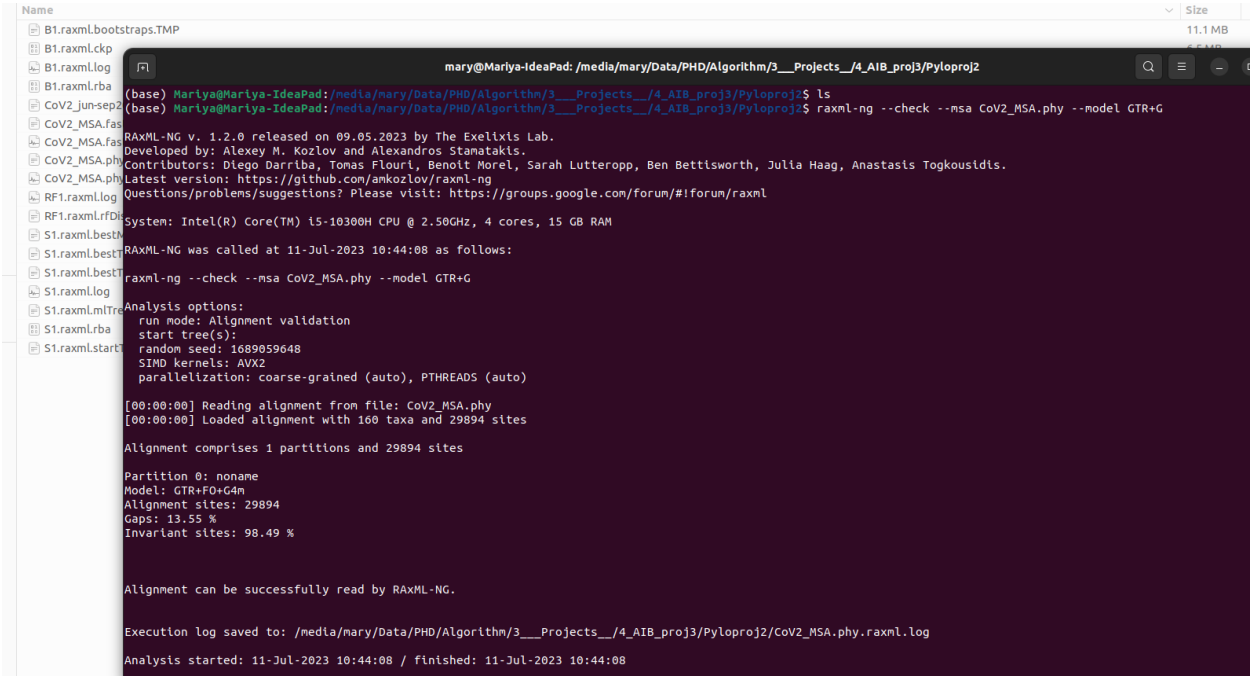
گونه مشاهده شده از ویروس در نیشابور در ماه هشتم با گونه مشاهده شده از تهرات درماه هشتم بسیار شبیه بود و نرخ جهش کم دارند نسبت بهم ولی این گونه با گونه مشاهده شده از گیلان در ماه هشتم جهش بیشتری دارد. یا کونه مشاهد شده از ساری در ماه هفتم نیز نرخ جهش بیشتری دارد



Run project again :)

Pyloproj2 Directory

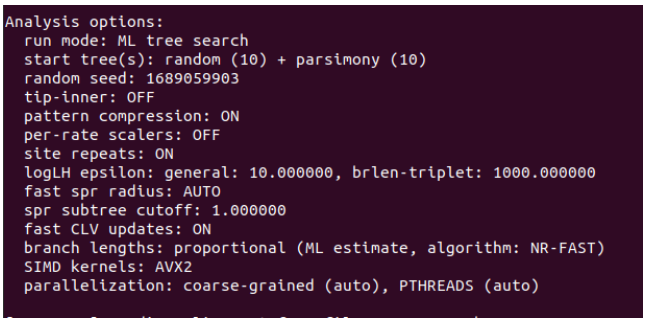
Check alignment for formatting errors



tree search

1. Run tree search
for prim.phy with default parameters

```
raxml-ng --msa CoV2_MSA.phy --model GTR+G --prefix S1

###################################################
Final LogLikelihood: -46569.081051

AIC score: 93790.162102 / AICc score: 93797.372979 / BIC score: 96497.726762
Free parameters (model + branch lengths): 326

WARNING: Best ML tree contains 94 near-zero branches!

Best ML tree with collapsed near-zero branches: S1.raxml.bestTreeCollapsed
Best ML tree saved to: Pyloproj2/S1.raxml.bestTree
All ML trees saved to: Pyloproj2/S1.raxml.mlTrees
Optimized model saved to:S1.raxml.bestModel

Execution log saved to: Pyloproj2/S1.raxml.log
```



Compare likelihoods of all 20 resulting trees

```
(base) Mariya@Mariya-IdeaPad:/media/mary/Data/PHD/Algorithm/3___Projects__/4_AIB_proj3/Pyloproj2$ grep "logLikelihood:" S1.raxml.log
[00:00:25] [worker #0] ML tree search #1, logLikelihood: -46571.136408
[00:00:25] [worker #1] ML tree search #2, logLikelihood: -46569.515582
[00:00:25] [worker #2] ML tree search #3, logLikelihood: -46570.579931
[00:00:27] [worker #3] ML tree search #4, logLikelihood: -46573.511112
[00:00:49] [worker #0] ML tree search #5, logLikelihood: -46572.877884
[00:00:51] [worker #1] ML tree search #6, logLikelihood: -46571.163949
[00:00:51] [worker #2] ML tree search #7, logLikelihood: -46580.070755
[00:00:54] [worker #3] ML tree search #8, logLikelihood: -46577.043517
[00:01:12] [worker #0] ML tree search #9, logLikelihood: -46572.301528
[00:01:14] [worker #2] ML tree search #11, logLikelihood: -46575.825116
[00:01:16] [worker #1] ML tree search #10, logLikelihood: -46572.348483
[00:01:16] [worker #3] ML tree search #12, logLikelihood: -46574.098703
[00:01:36] [worker #0] ML tree search #13, logLikelihood: -46572.290352
[00:01:38] [worker #2] ML tree search #15, logLikelihood: -46572.291077
[00:01:39] [worker #3] ML tree search #16, logLikelihood: -46575.833028
[00:01:40] [worker #1] ML tree search #14, logLikelihood: -46571.797730
[00:01:59] [worker #3] ML tree search #20, logLikelihood: -46573.504128
[00:02:01] [worker #0] ML tree search #17, logLikelihood: -46571.781543
[00:02:01] [worker #2] ML tree search #19, logLikelihood: -46569.081051
[00:02:02] [worker #1] ML tree search #18, logLikelihood: -46571.139682
```

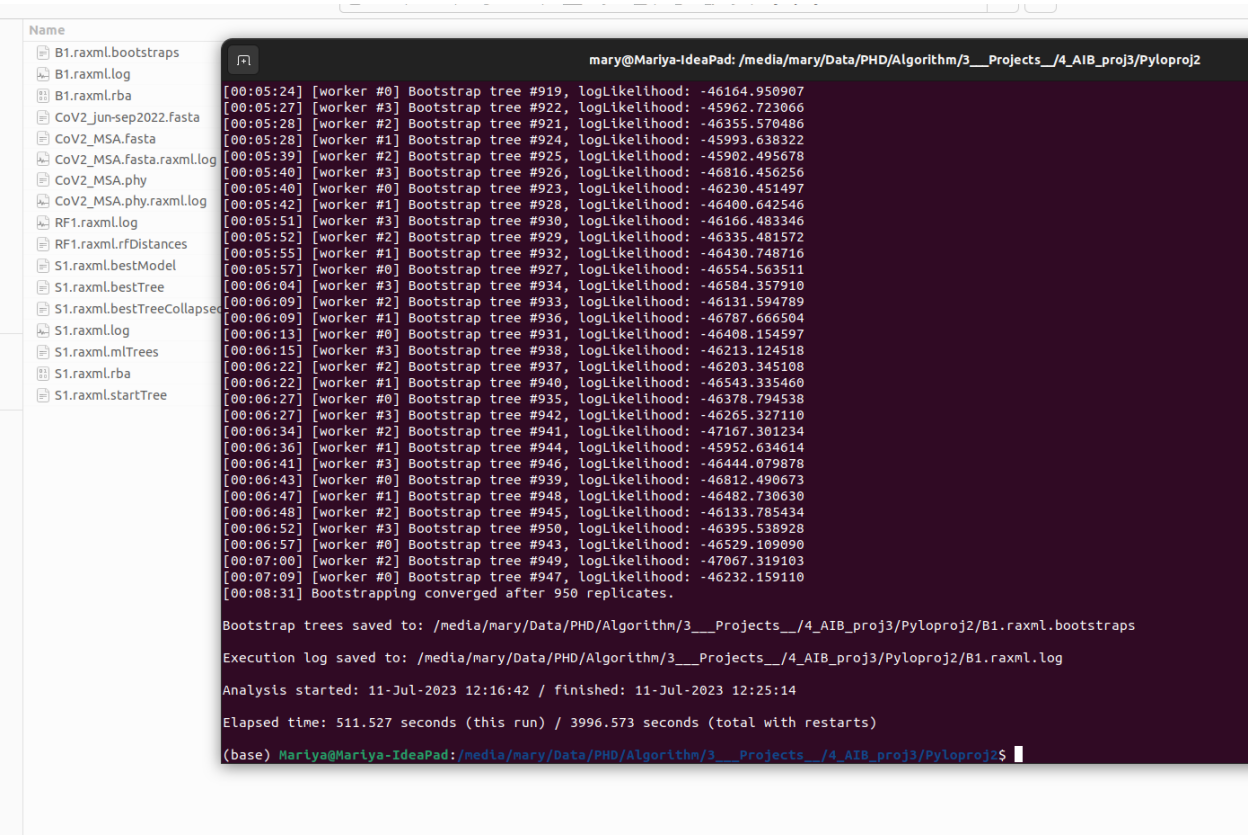
Check topological distances between all 20 trees(so-called Robinson-Foulds or RF distance)

Average topological (RF) distance

```
Reading input trees from file: S1.raxml.mlTrees
Loaded 20 trees with 160 taxa.

Average absolute RF distance in this tree set: 177.905263
Average relative RF distance in this tree set: 0.566577
Number of unique topologies in this tree set: 20

Pairwise RF distances saved to: /media/mary/Data/PHD/Algorithm/3___Projects__/4_AIB_proj3/Pyloproj2/RF1.raxml.rfDistances

Execution log saved to: /media/mary/Data/PHD/Algorithm/3___Projects__/4_AIB_proj3/Pyloproj2/RF1.raxml.log
```

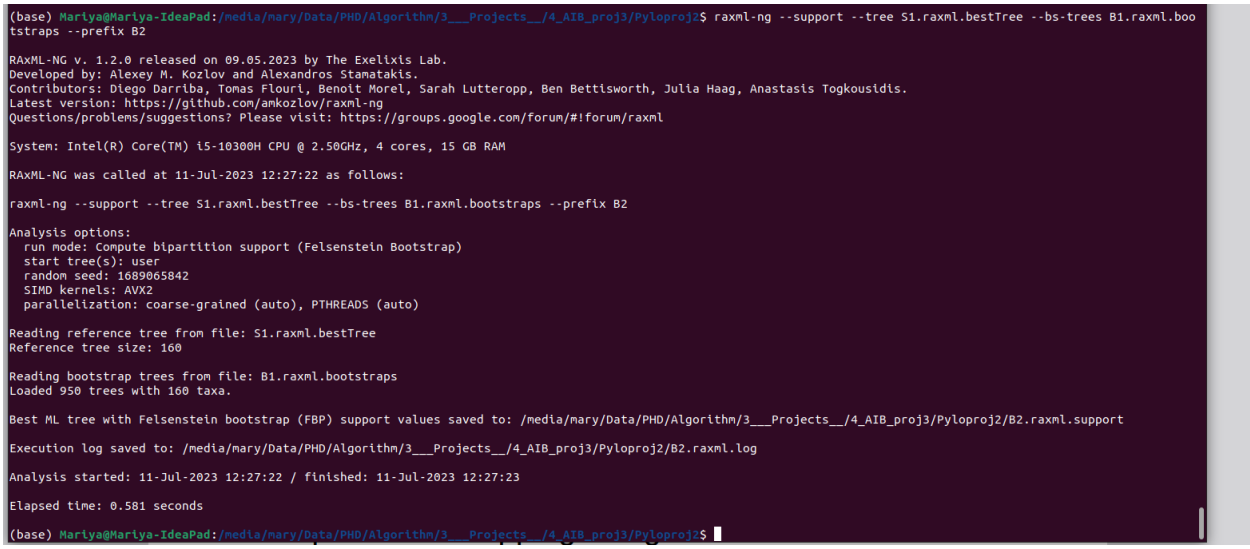
Run bootstrap tree inference with default parameters (**950 replicate**)

```
raxml-ng --bootstrap --msa CoV2_MSA.phy --model GTR+G --prefix B1
```



Map bootstrap support values to the best ML tree

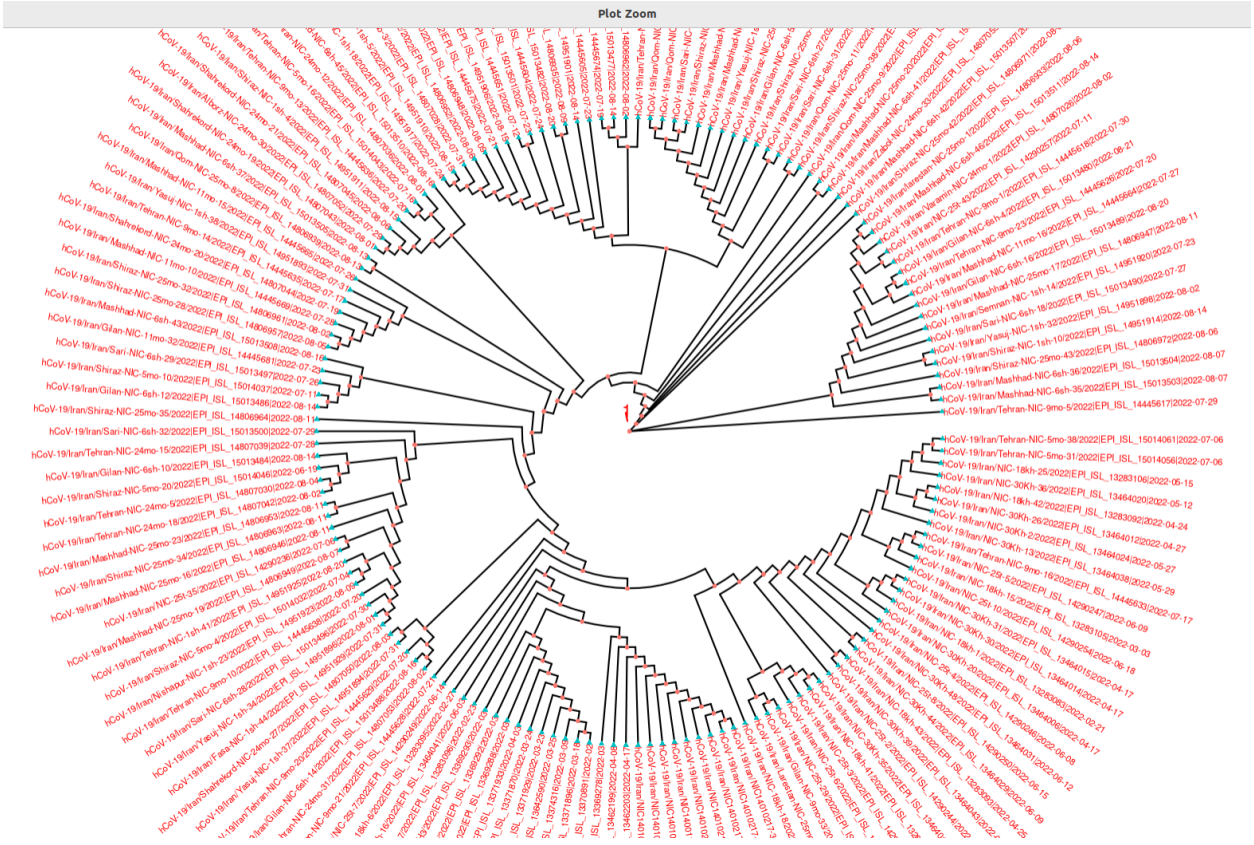
```
raxml-ng --support --tree S1.raxml.bestTree --bs-trees B1.raxml.bootstraps
 --prefix B2
```



Data visualization

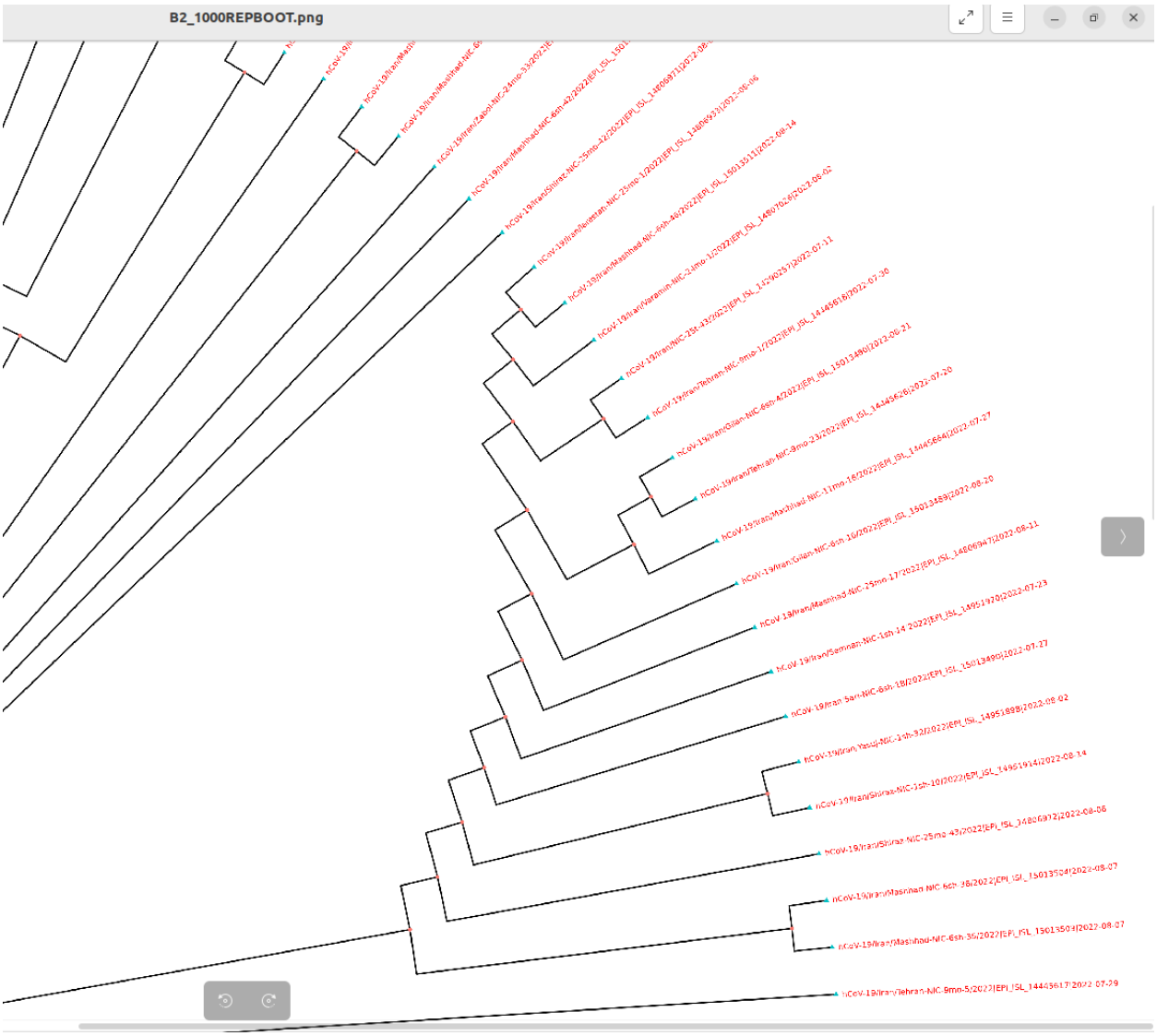
1000replicate bootstrap

```
library(ggtree)

# Read Maping file of bootstrap support values to the best ML tree
tree <- read.tree("B2_1000REPBOOT.raxml.support")

# Plot the tree with support values using ggtre
ggtree(tree, layout="circular", branch.length='none') +
  geom_treescale(x=1, y=45, width=1, color='red')+
  geom_tiplab(size=2, color='blue',aes(angle=angle)) +
  # geom_tippoint(aes(size=label), shape=1, fill='gray') +
  theme(legend.position="none")+
 geom_point(aes(shape=isTip, color=isTip), size=1)

ggsave("B2_1000REPBOOT.png", width = 30, height = 30, dpi = 300,limitsize = FALSE)
```



همانطور که مشاهده میشود در درخت گونه مشاهده شده از تهران در ماه ۷ ام با گونه ها یدیگر تفاوت ژنتیکی بیشتری دارد و شبیه هیچ گونه دیگری نیست. گونه های مشاهده شده از ویروس در شیراز مشهد و زابل نیز تفاوت ژنتیکی بالایی با بقیه گونه ها دارند.

ولی به طور مثال گونه تهران و گیلان در ماه ۸ ام تشابه ژنتیکی زیادی دارند و دریک کلاستر هستند همانطور که مشاهده میشود گونه های سمت راست تقریبا تشابه ژنتیکی دارند و در یک کلاستر کلی نسبت بهب قیه گونه های درخت قرار دارند.



سایر قسمت های درخت در یک کلاستر کلی قرار دارند و نرخ جهش زیادی در ویروس مشاهده میشود در این قسمت ها