Date:25/04/21

Phyloinformatics

Assignment-8

Q-1 Construct a phylogenetic tree using **PHYLIP software based o**n any **ten keratin protein sequences.**

-Design tree using **NJ and ML** methods (default parameters)

-Perform Bootstrapping of the results with default parameters and generate consensus tree accordingly.

-Compare your results.

Report contains 1. Input sequences 2. parameters 3. NJ and ML Tress 4. Bootstrap parameters and consensus tree of NJ and ML

Answer:

So for performing protein analysis we have two methods:-

1.Distance based methods :- It is made using prodist method and tree is designed using neighbour programs which has two methods of tree building:-

- UPGMA
- Neighbour Joining

2.Character based methods:-

- Maximum Likelihood: Where proml is the program used for proteins in ML.
- Parsimony: Which has protpars as protein programs used in Parsimony.

Step 1: Download PHYLIP software and MEGA software from chrome.

PHYLIP:

PHYLIP (the PHYLogeny Inference Package) is a package of programs for inferring phylogenies (evolutionary trees). It is available free over the Internet, and written to work on as many different kinds of computer systems as possible. The source code is distributed (in C), and executables are also distributed. In particular, already-compiled executables are available for Windows (95/98/NT/2000/me/xp/Vista), Mac OS X, and Linux systems. Older executables are also available for Mac OS 8 or 9 systems. Complete documentation is available on documentation files that come with the package.

Methods that are available in the package include **parsimony, distance matrix, and likelihood methods, including bootstrapping and consensus trees**. Data types that can be handled include molecular sequences, gene frequencies, restriction sites and fragments, distance matrices, and discrete characters.

The programs are controlled through a menu, which asks the users which options they want to set, and allows them to start the computation. The data are read into the program from a text file, which the user can prepare using any word processor or text editor (but it is

important that this text file *not* be in the special format of that word processor -- it should instead be in "flat ASCII" or "Text Only" format). Some sequence analysis programs such as the **ClustalW** alignment program can write data files in the PHYLIP format. Most of the programs look for the data in a file called "infile" -- if they do not find this file they then ask the user to type in the file name of the data file.

Output is written onto special files with names like **"outfile" and "outtree". Trees written onto "outtree" are in the Newick format**, an informal standard agreed to in 1986 by authors of a number of major phylogeny packages.

At this stage we do not have a mouse-windows interface for PHYLIP.

PHYLIP is probably the most widely-distributed phylogeny package. It is the sixth most frequently cited phylogeny package, after MrBayes, PAUP*, RAxML, Phyml, and MEGA. PHYLIP is also the oldest widely-distributed package. It has been in distribution since October, 1980, and has over 30,000 registered users. It is still being updated.

MEGA:

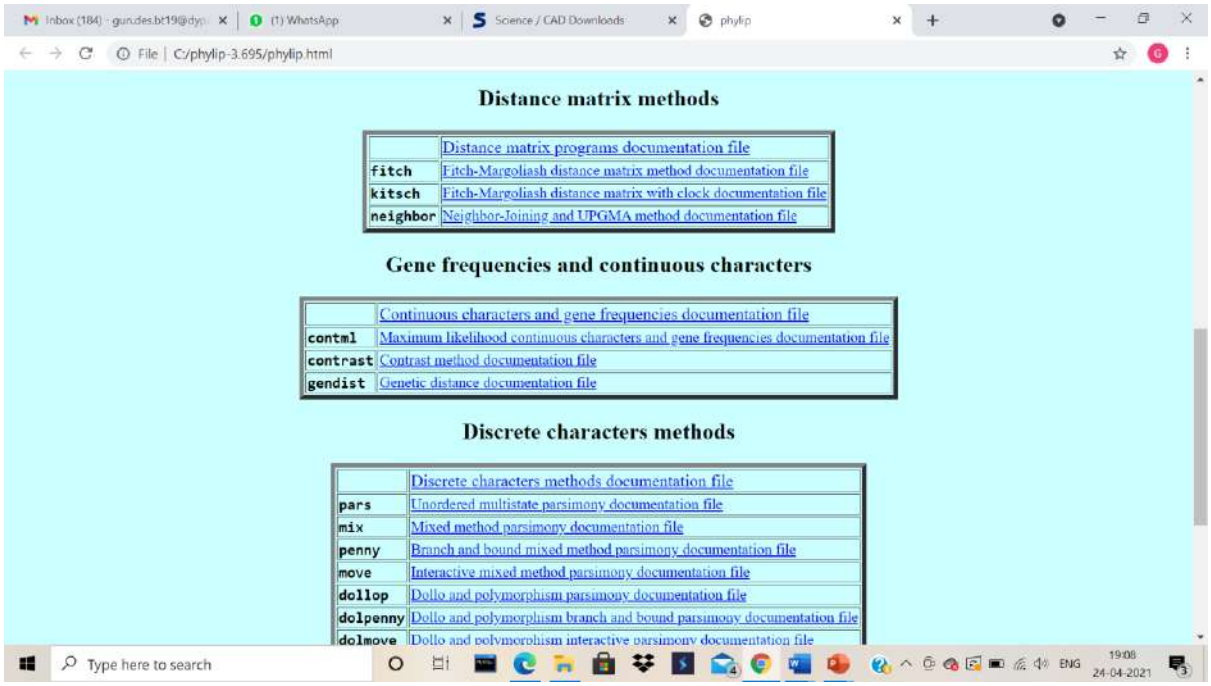
MEGA stands for Molecular Evolutionary Genetic Analysis and the X stands for version 10.

The objective of the MEGA software has been to provide tools for exploring, discovering, and analyzing DNA and protein sequences from an evolutionary perspective. The first version was developed for the limited computational resources that were available on the average personal computer in early 1990s. MEGA1 made many methods of evolutionary analysis easily accessible to the scientific community for research and education. MEGA2 was designed to harness the exponentially greater computing power and a graphical interface of the late 1990's, fulfilling the fast-growing need for more extensive biological sequence analysis and exploration software. It expanded the scope of its predecessor from single gene to genome wide analyses.

Step 2: Extract all the PHYLIP files to the C drive for convenient working.

Step 3: Go to the home page of PHYLIP.

The choice of molecular marker - **Protein analysis.**



To construct a phylogenetic tree using **PHYLIP software based o**n any **ten keratin protein sequences:**

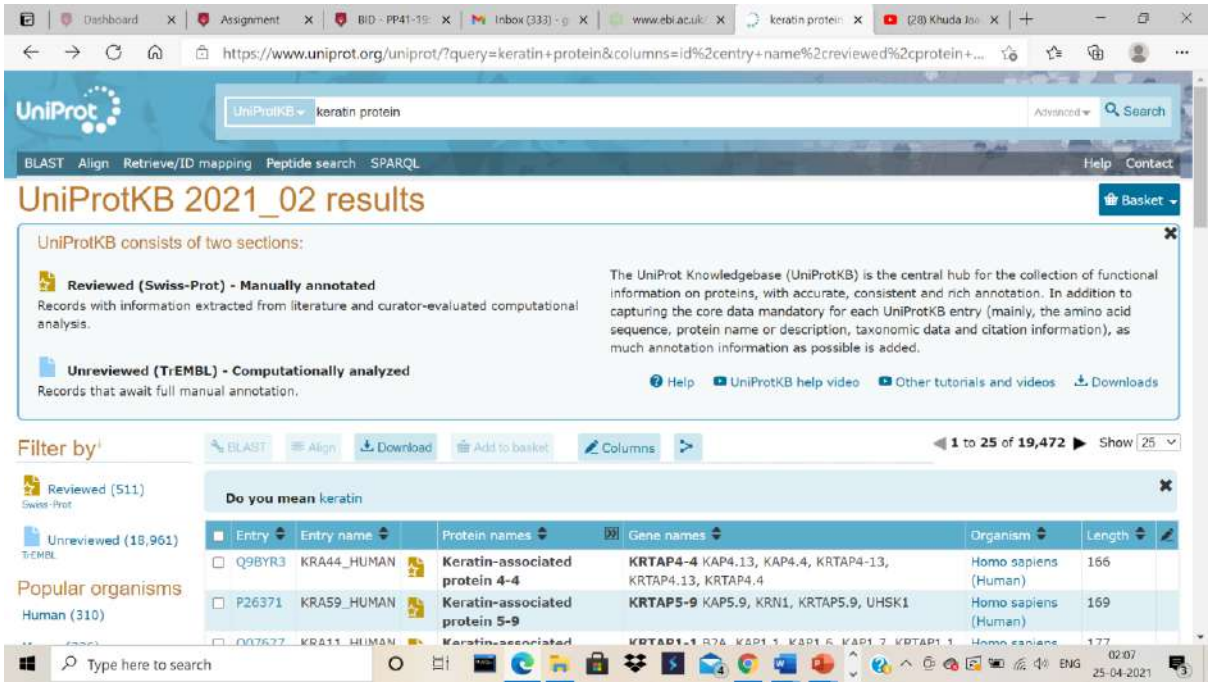
Step 1: Go to the home page of uniport

UniProtKB:

The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. In addition to capturing the core data mandatory for each UniProtKB entry (mainly, the amino acid sequence, protein name or description, taxonomic data and citation information), as much annotation information as possible is added. This includes widely accepted biological ontologies, classifications and cross-references, and clear indications of the quality of annotation in the form of evidence attribution of experimental and computational data.

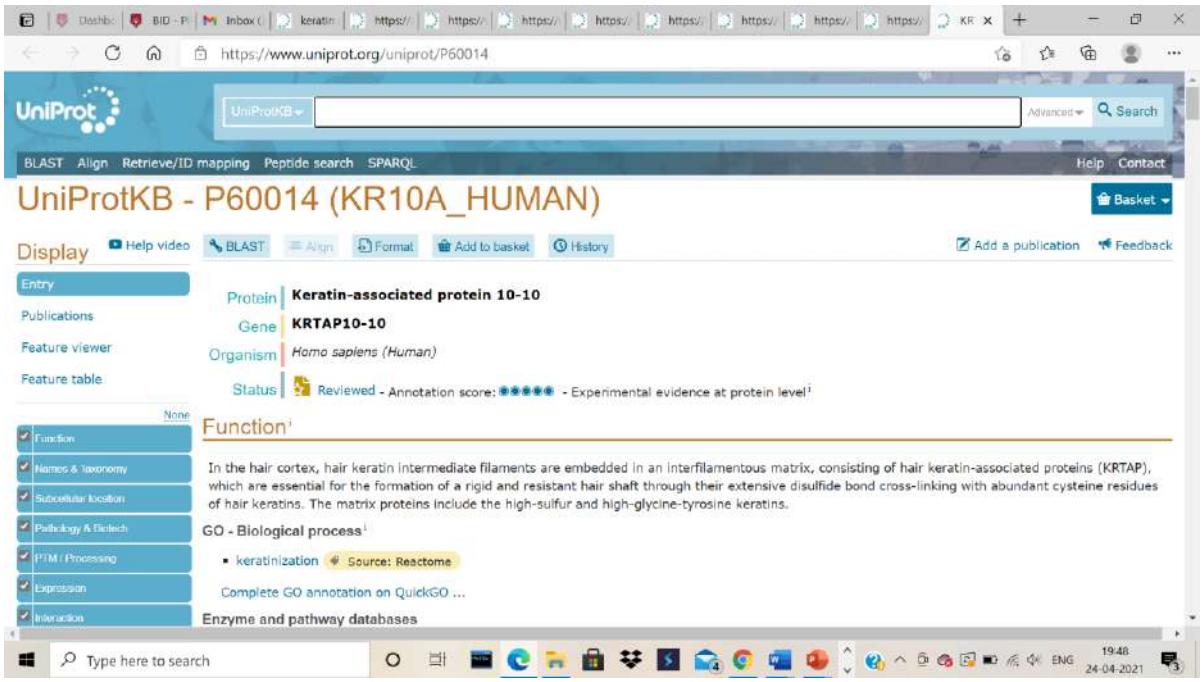
The UniProt Knowledgebase consists of two sections: a section containing manually-annotated records with information extracted from literature and curator-evaluated computational analysis, and a section with computationally analyzed records that await full manual annotation. For the sake of continuity and name recognition, the two sections are referred to as "UniProtKB/Swiss-Prot" (reviewed, manually annotated) and "UniProtKB/TrEMBL" (unreviewed, automatically annotated), respectively.

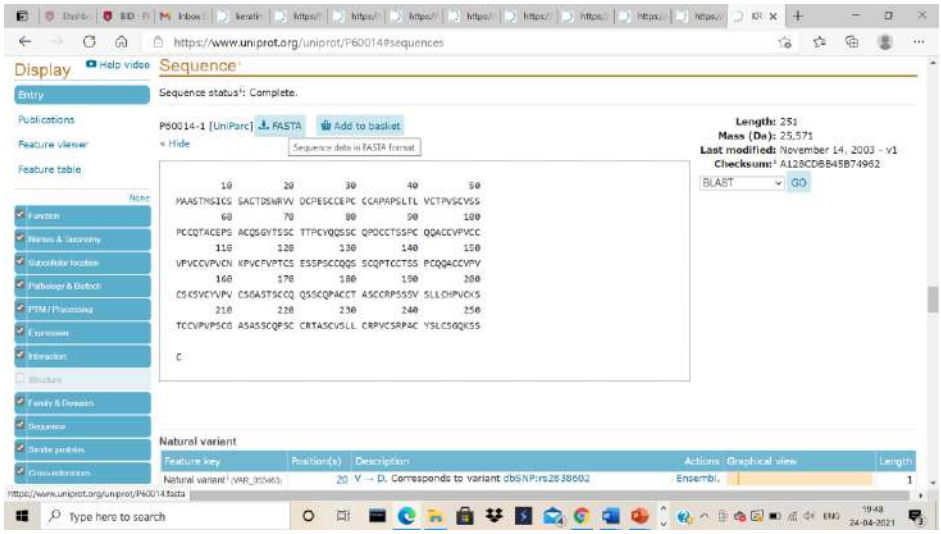
Step 2: In the search bar type **keratin protein**



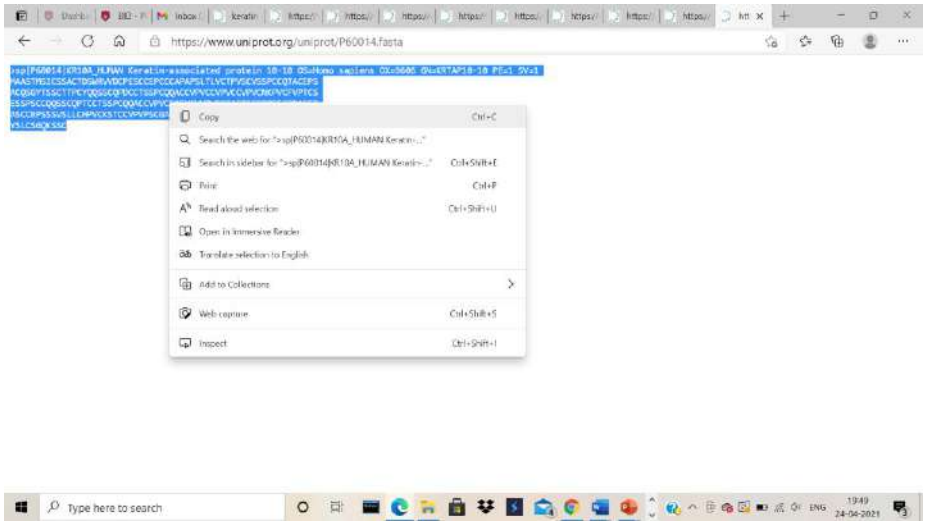
Step 3: Select any 10 keratin protein sequences.

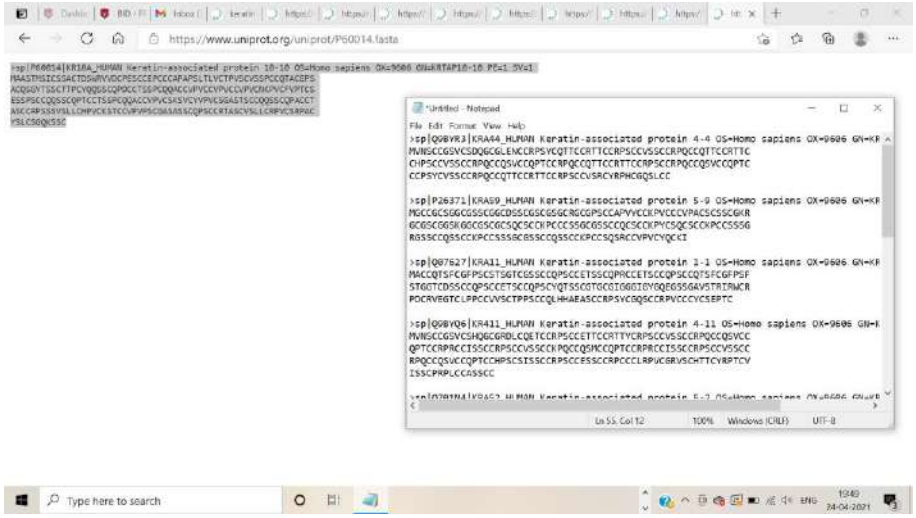
Step 4: Download all the fasta formats.





Step 5: Copy paste all fasta formats in notepad.

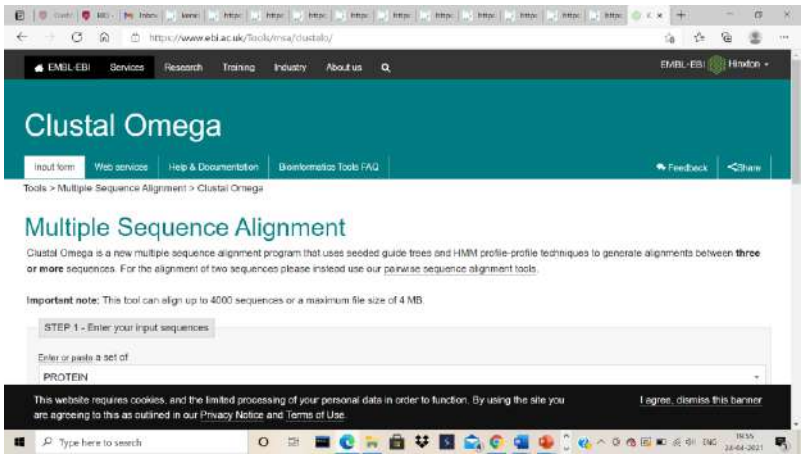




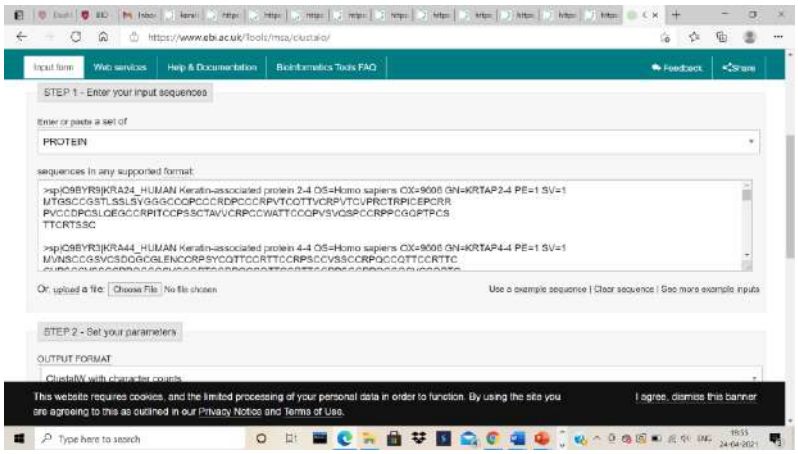
Step 6: Go to Clustal omega home page .

ClustalΩ (Omega): The current standard version.

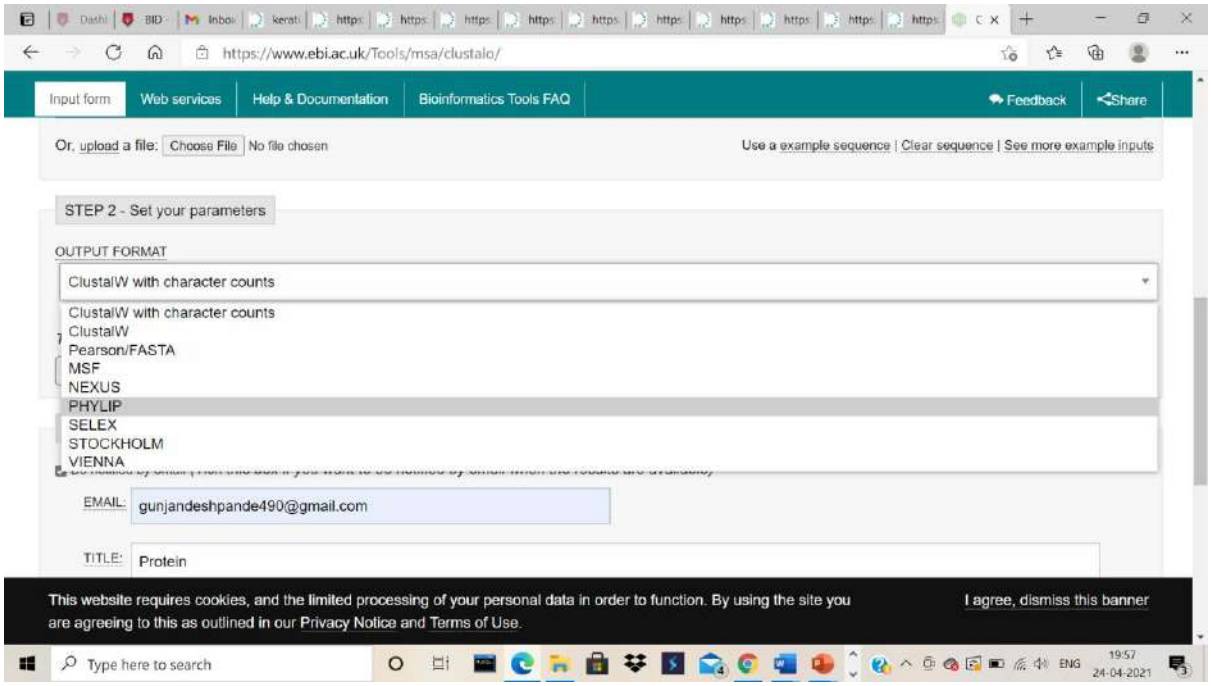
ClustalΩ (alternatively written as Clustal O and Clustal Omega) is a fast and scalable program written in C and C++ used for multiple sequence alignment. It uses seeded guide trees and a new HMM engine that focuses on two profiles to generate these alignments. The program requires three or more sequences in order to calculate the multiple sequence alignment, for two sequences use pairwise sequence alignment tools (EMBOSS, LALIGN). Clustal Omega is consistency-based and is widely viewed as one of the fastest online implementations of all multiple sequence alignment tools and still ranks high in accuracy, among both consistency-based and matrix-based algorithms.

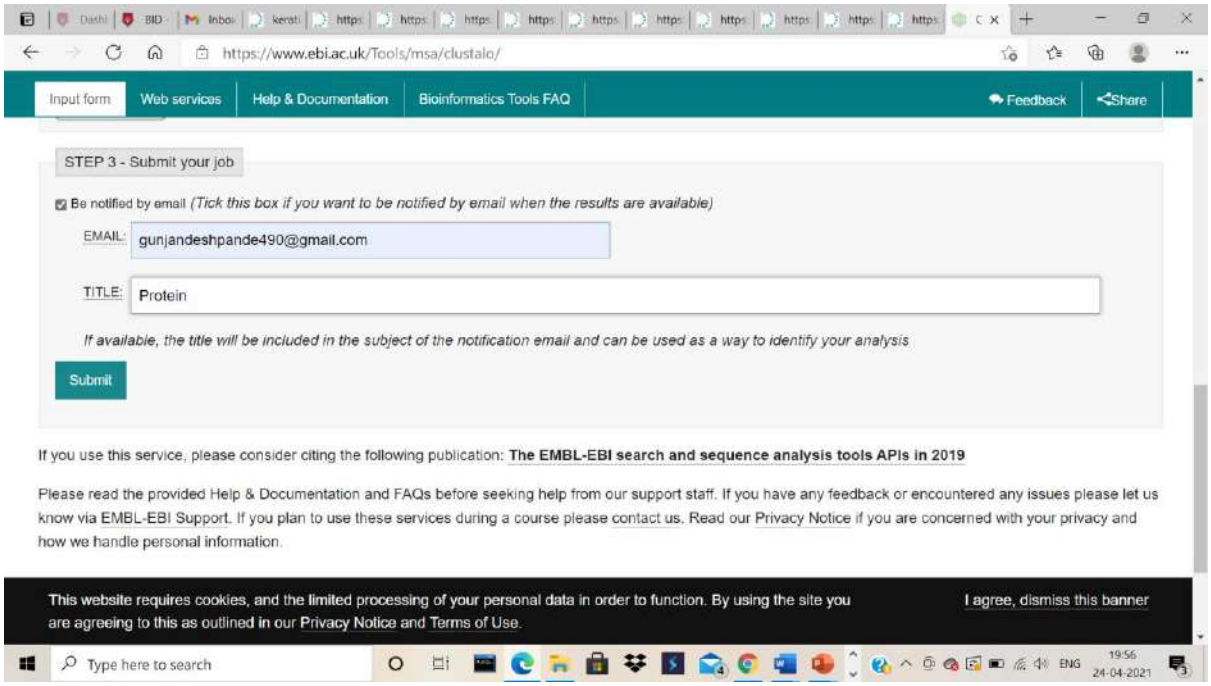


Step 7: Paste all 10 sequences in the box.

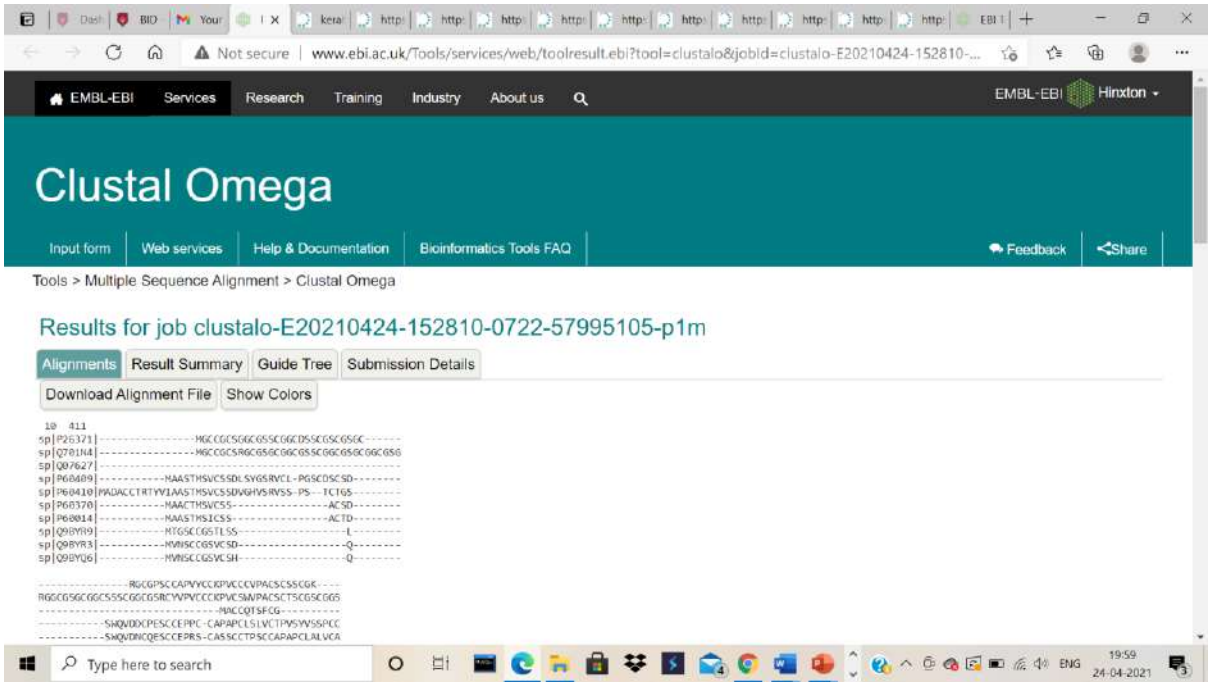


Step 8: Change the output format to PHYLIP.

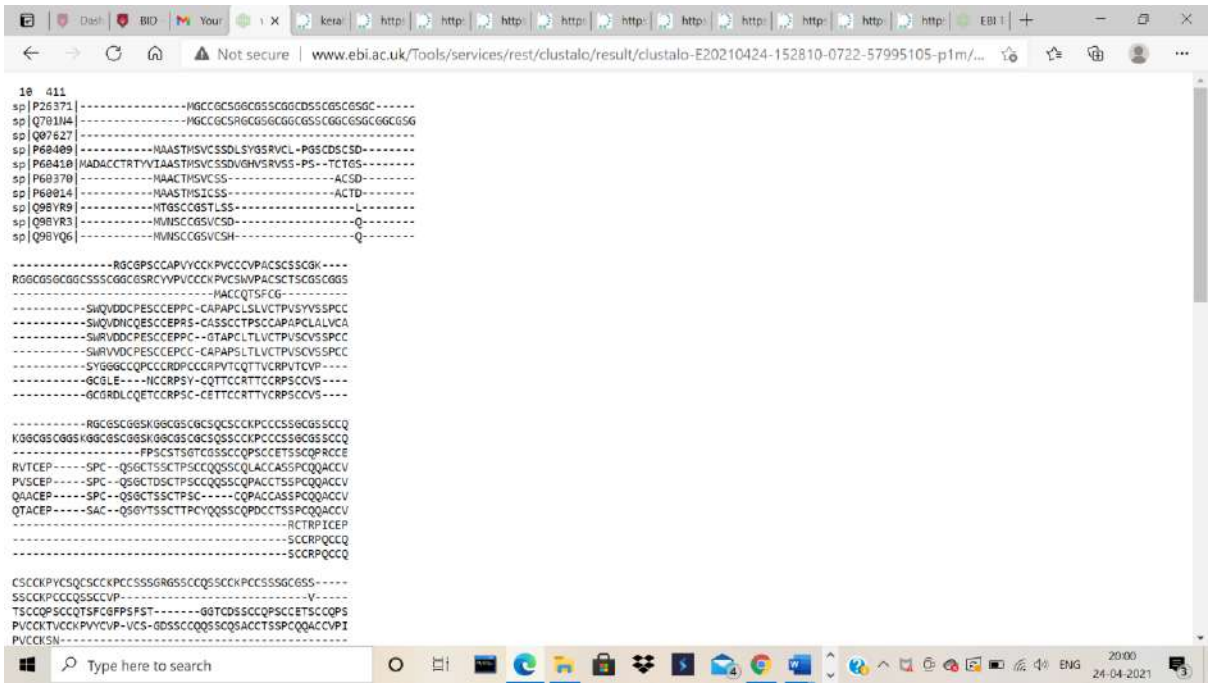




Result:



Step 9: Go to Download alignment file.

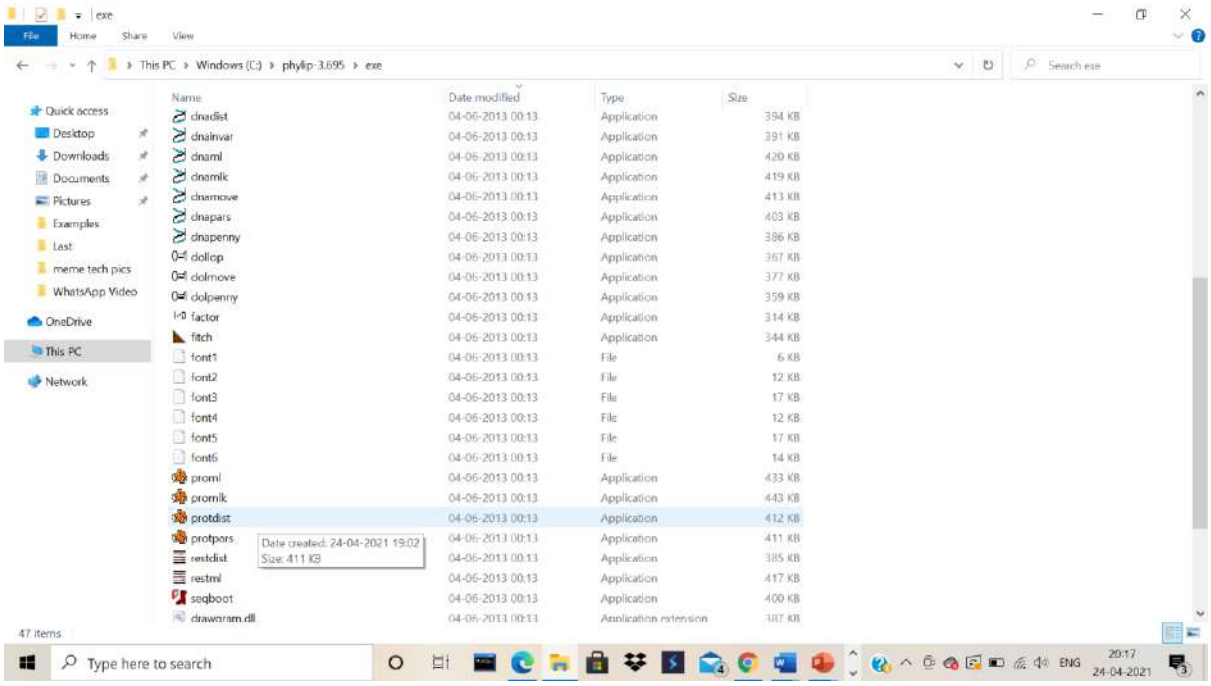


Step 10: Save this sequences in notepad and than save as in exe folder of PHYLIP file.

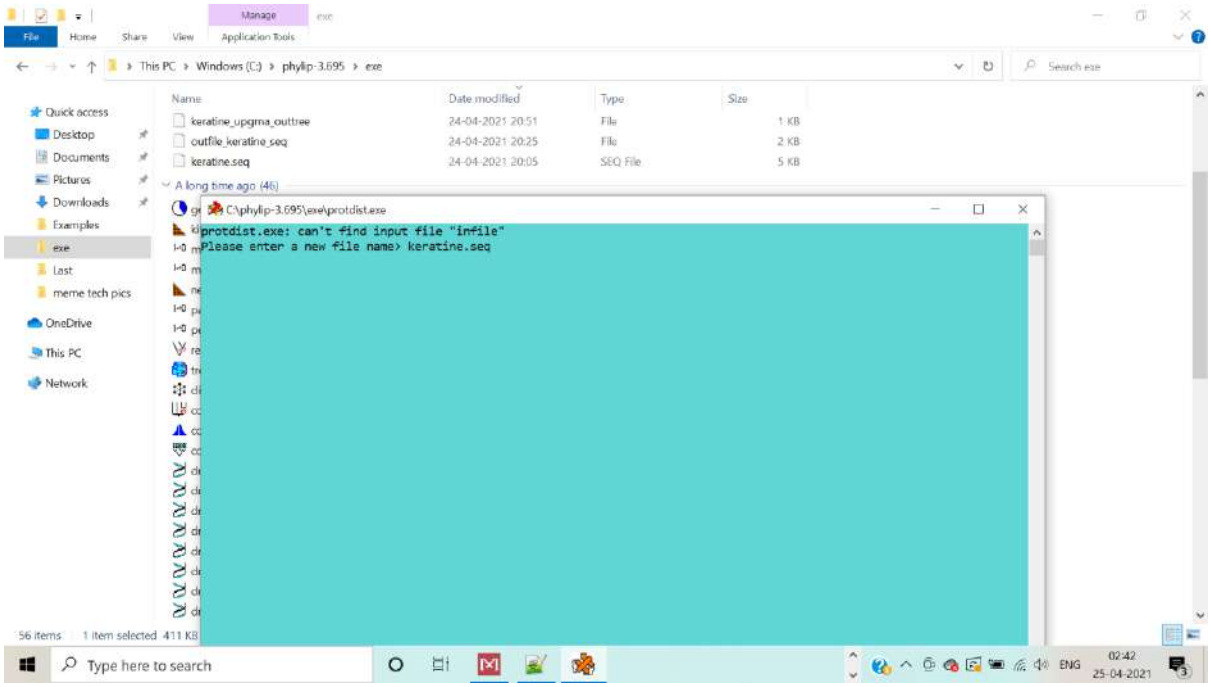
- ✓ **Performing protein analysis using Distance based methods**

Obtaining a distance matrix:

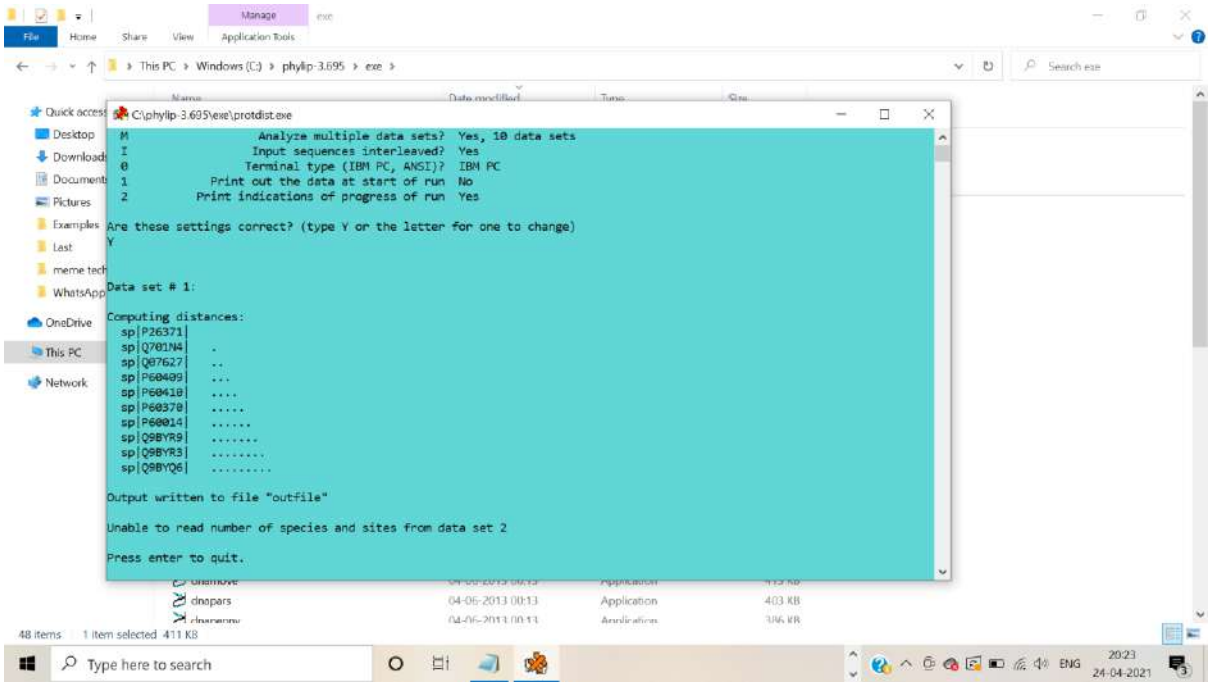
Step 1: Go to protdis



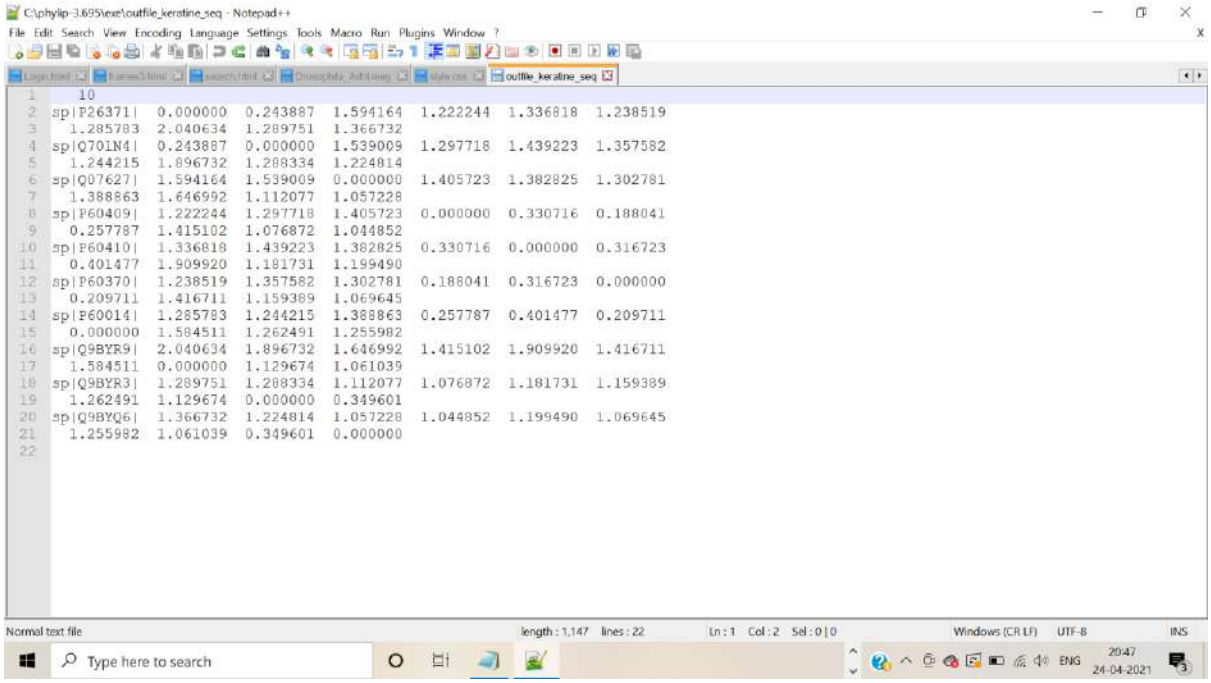
Step 2: Enter your new file name.



Step 3: Press Y and agree all the conditions and enter to exit.



Step 4: Open the output folder and you will get your **Distance matrix.**



- ✓ **To perform tree building using neighbour- joining method:**

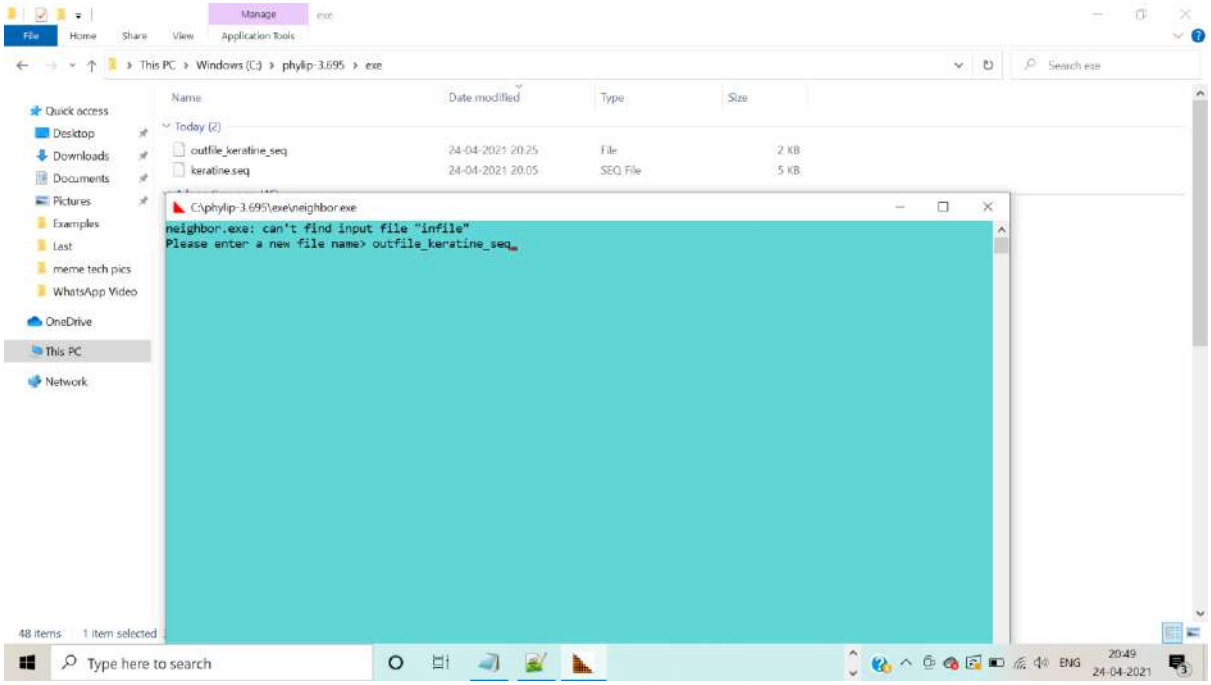
Neighbour- Joining Method:-

This method (Saitou and Nei 1987) is a simplified version of the minimum evolution (ME) method (Rzhetsky and Nei 1992). The ME method uses distance measures that correct for multiple hits at the same sites; it chooses a topology showing the smallest value of the sum of all branches (S) as an estimate of the correct tree. However, construction of an ME tree is time-consuming because, in principle, the S values for all topologies must be evaluated. Because the number of possible topologies (unrooted trees) rapidly increases with the number of taxa, it becomes very difficult to examine all topologies.

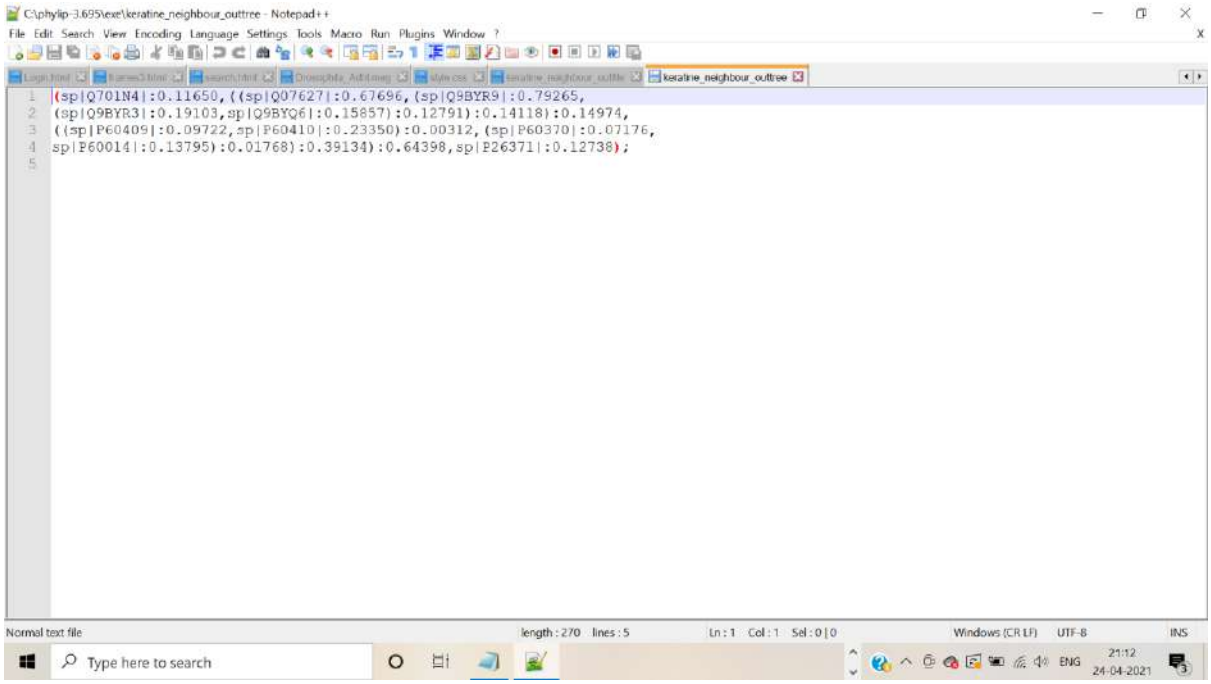
In the case of the NJ method, the S value is not computed for all or many topologies, but the examination of different topologies is embedded in the algorithm, so that only one final tree is produced.

The NJ method produces an unrooted tree because it does not require the assumption of a constant rate of evolution. Finding the root requires an outgroup taxon. In the absence of outgroup taxa, the root is sometimes given at the midpoint of the longest distance connecting two taxa in the tree, which is referred to as mid-point rooting.

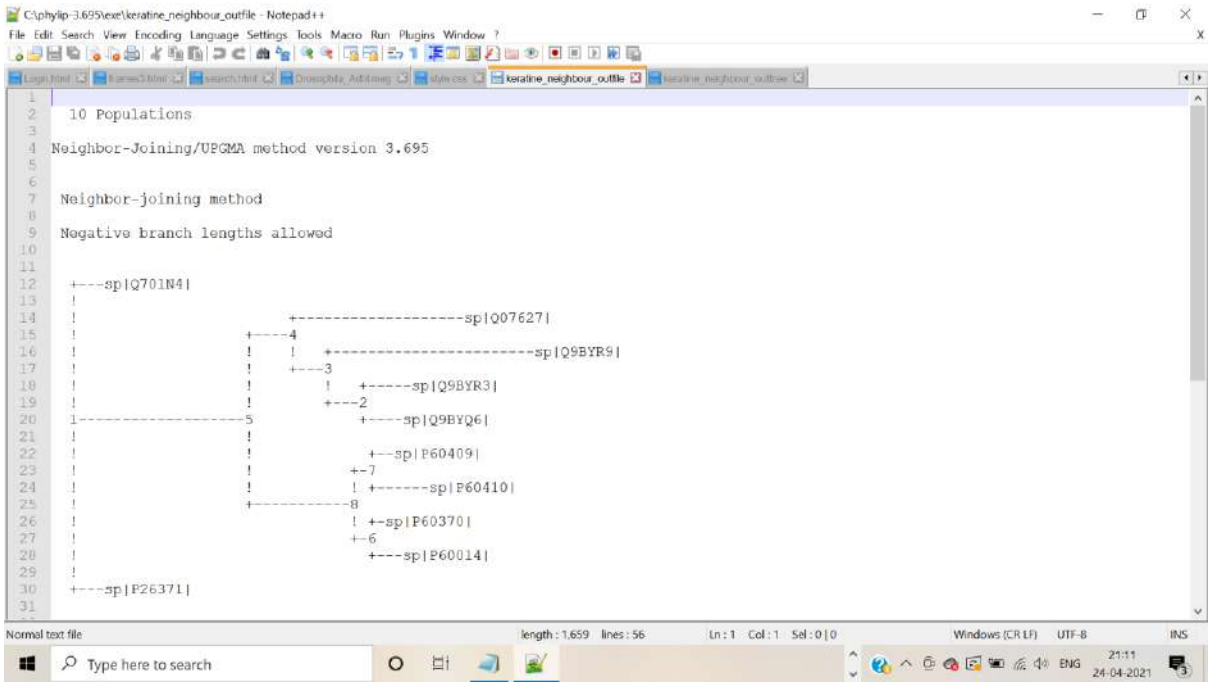
Step 1: Enter the file name:

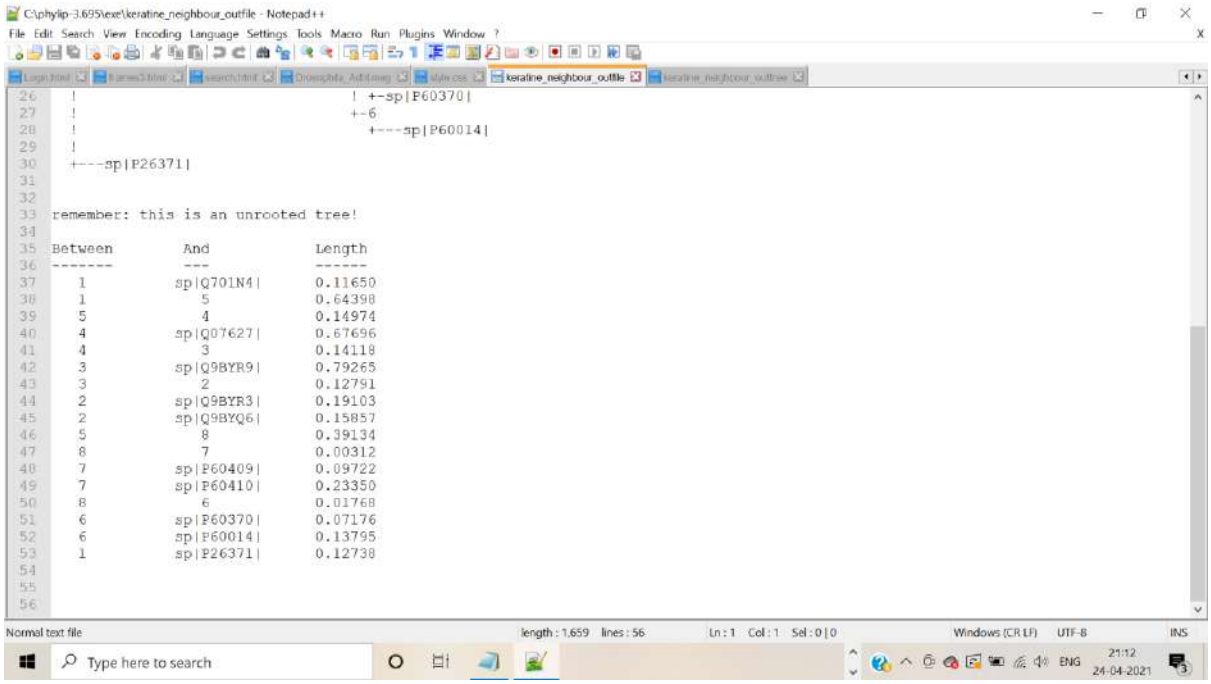


Step 2: Obtain the out tree.

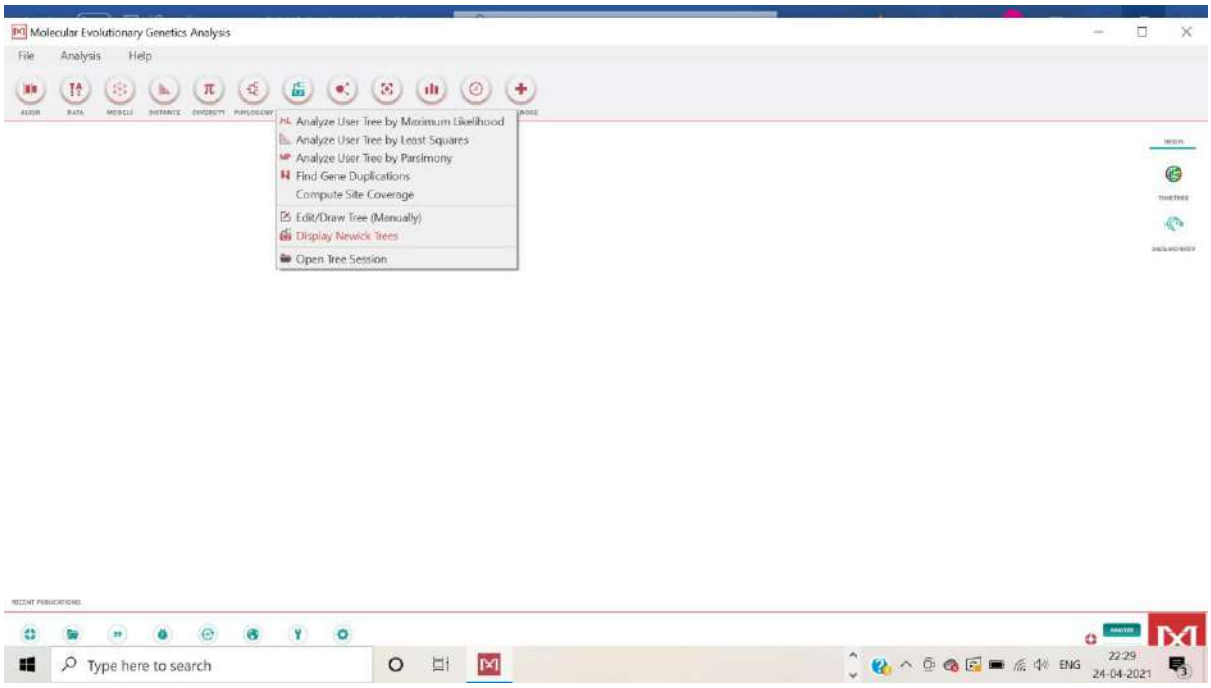


Step 3: Obtain a Un-rooted tree for Neighbour-joining method

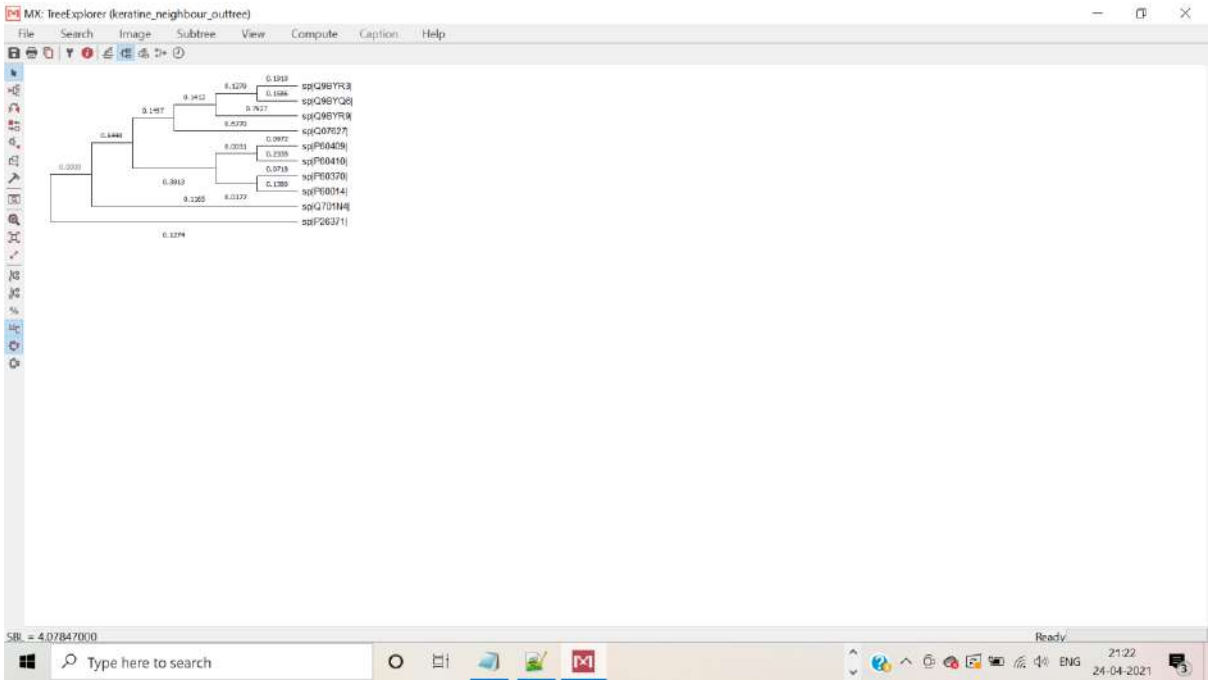




Step 4: View the tree in MEGA software, go to user tree and click on Display newick tree.



Step 5: Phylogenetic tree obtained for Neighbour-joining method.

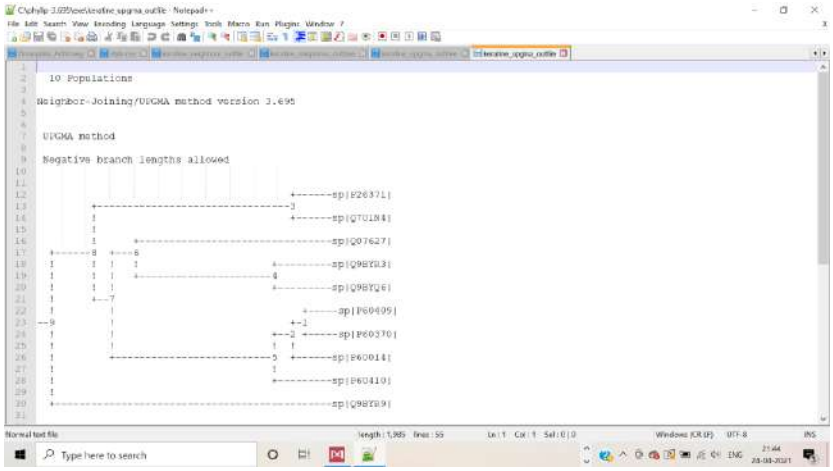


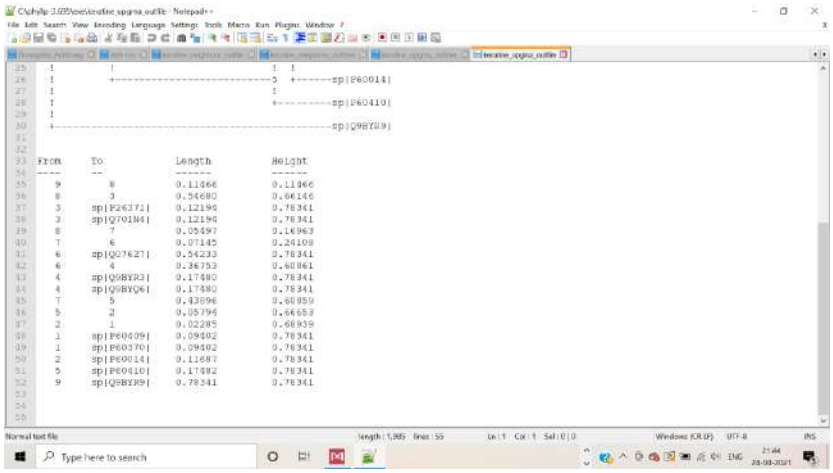
- ✓ **Performing Distance Based Tree Building**

UPGMA:-

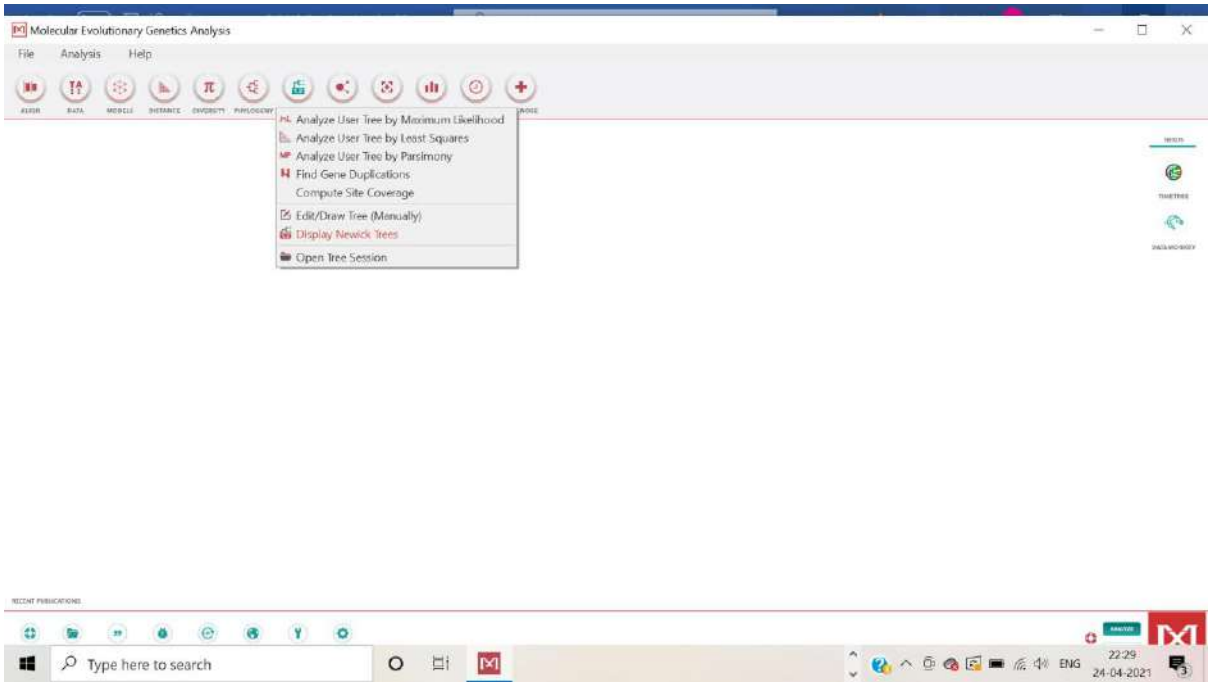
This method assumes that the rate of nucleotide or amino acid substitution is the same for all evolutionary lineages. An interesting aspect of this method is that it produces a tree that mimics a species tree, with the branch lengths for two OTUs being the same after their separation. Because of the assumption of a constant rate of evolution, this method produces a rooted tree, though it is possible to remove the root for certain purposes.

Step 1: Open the saved keratine_upgma_outfile in notepad ++ and you will see Negative branched tree in UPGMA method.

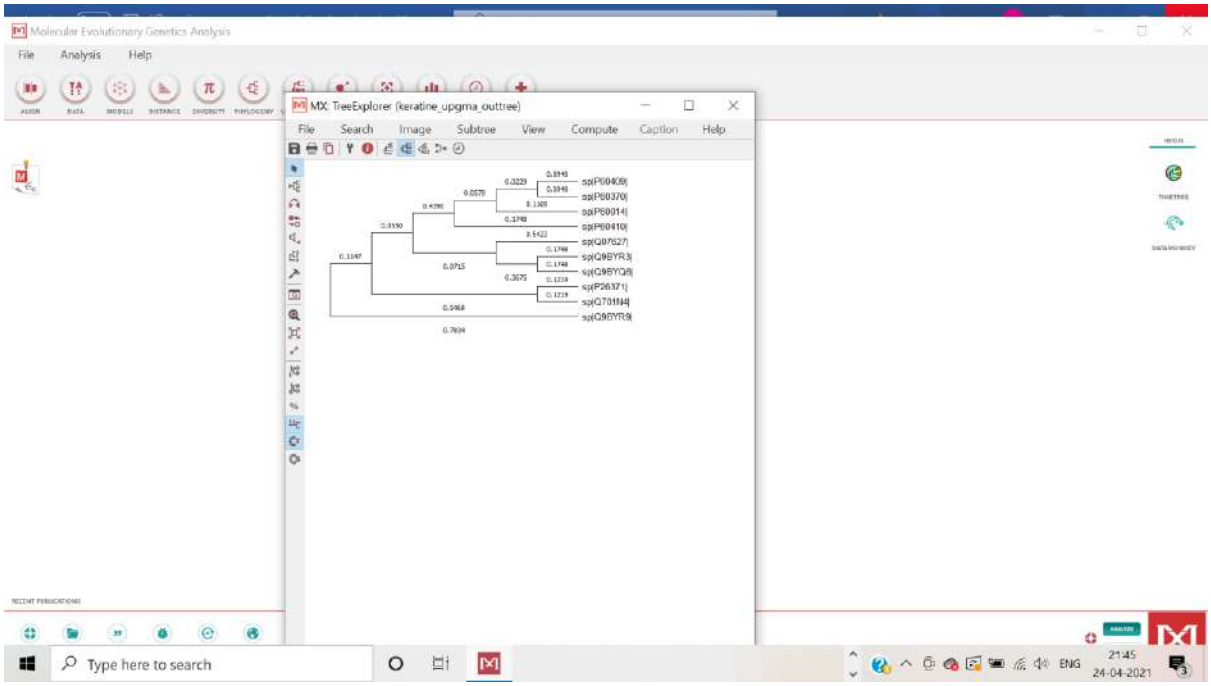




Step 2: Go to MEGA to obtain a phylogenetic tree. Click on user tree and select Display Newick tree.



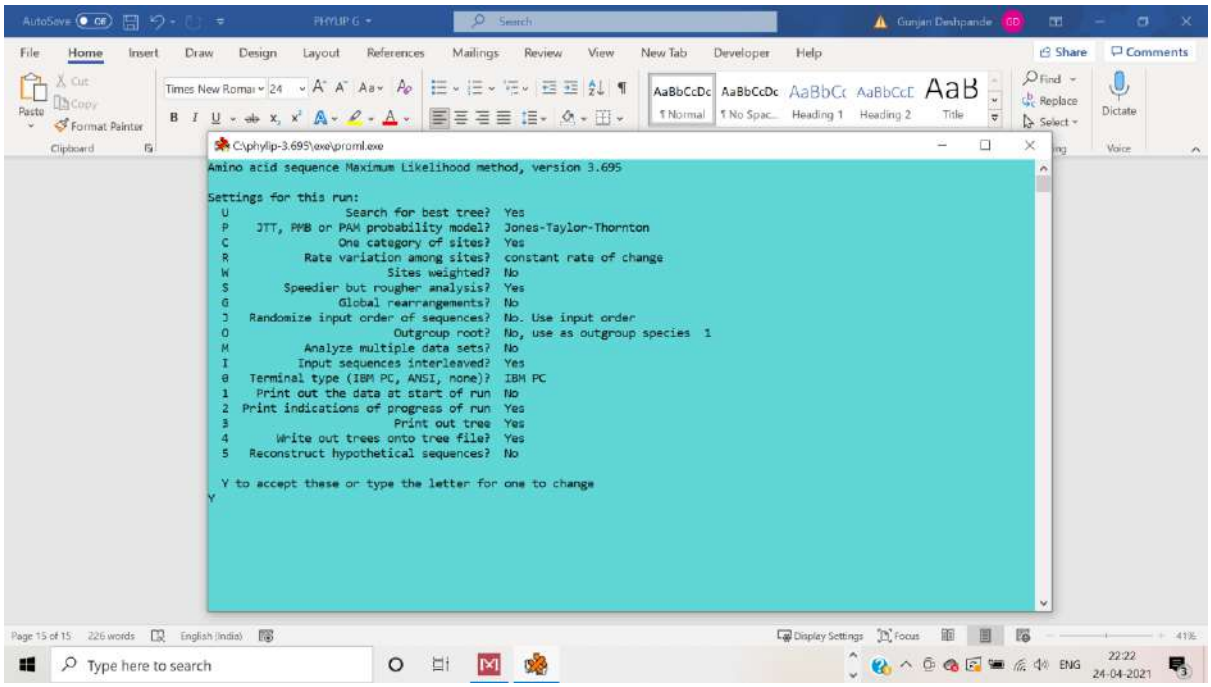
Step 3: UPGMA tree is obtained.



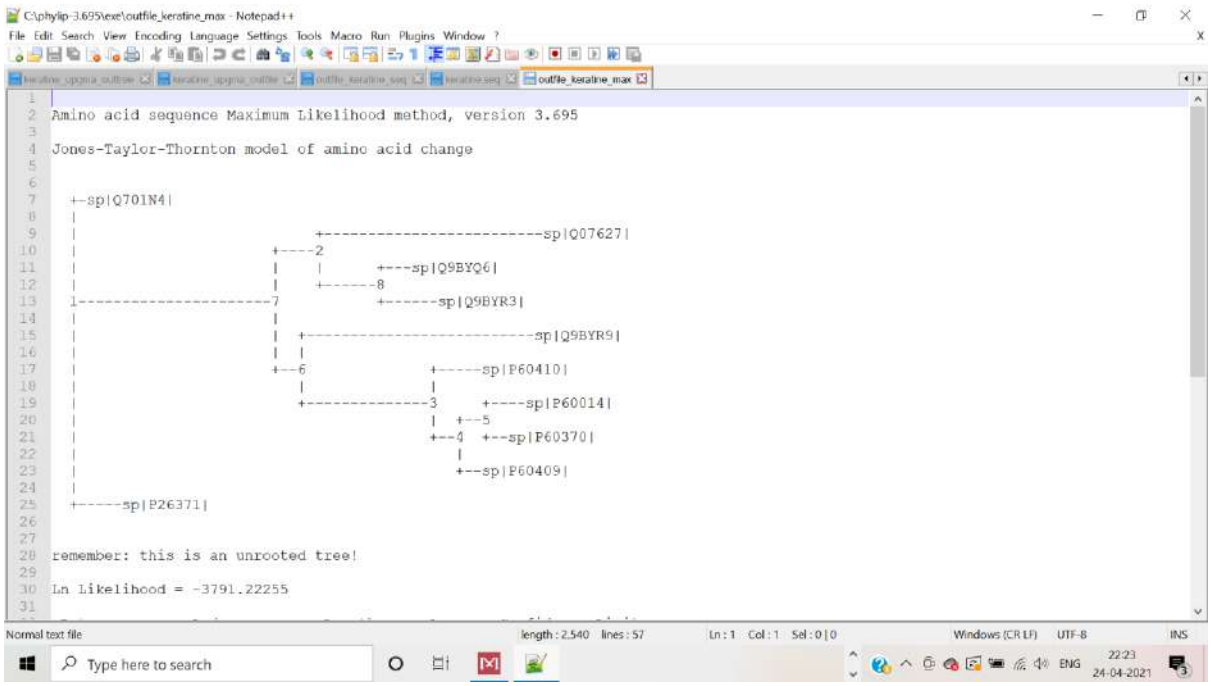
- ✓ **Performing protein analysis using Character based methods:-**

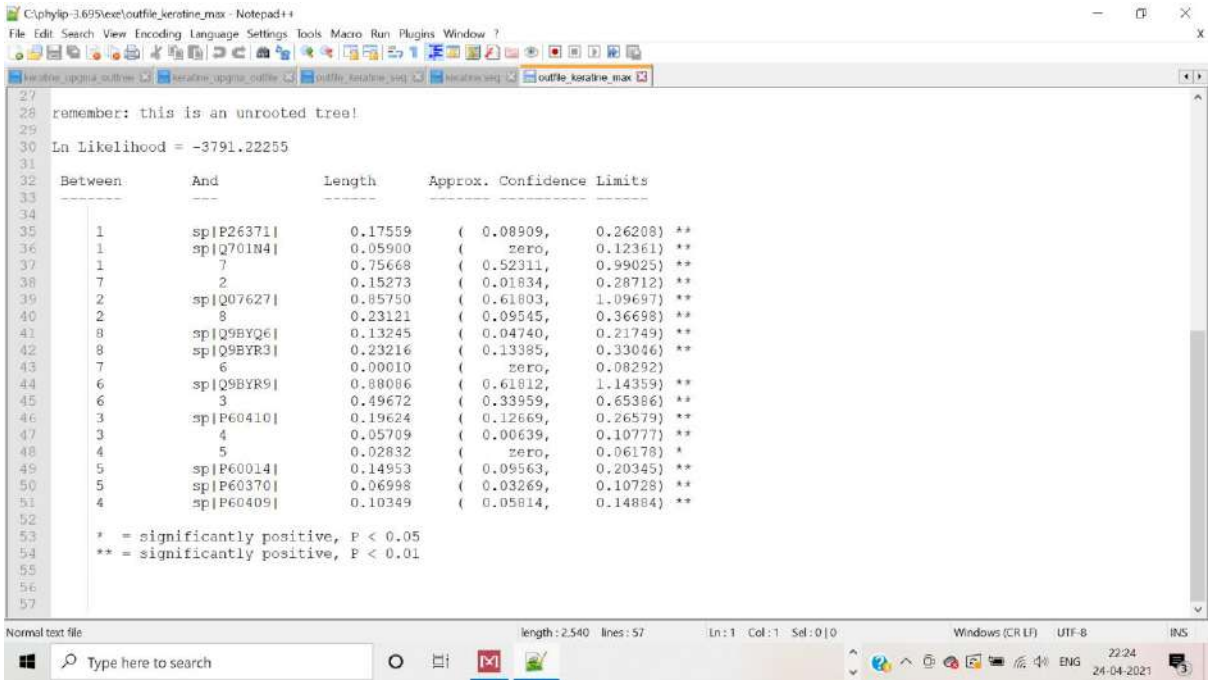
Maximum Likelihood: Where proml is the program used for proteins in ML.

Step 1: Open proml and type Y to agree and proceed with the conditions.

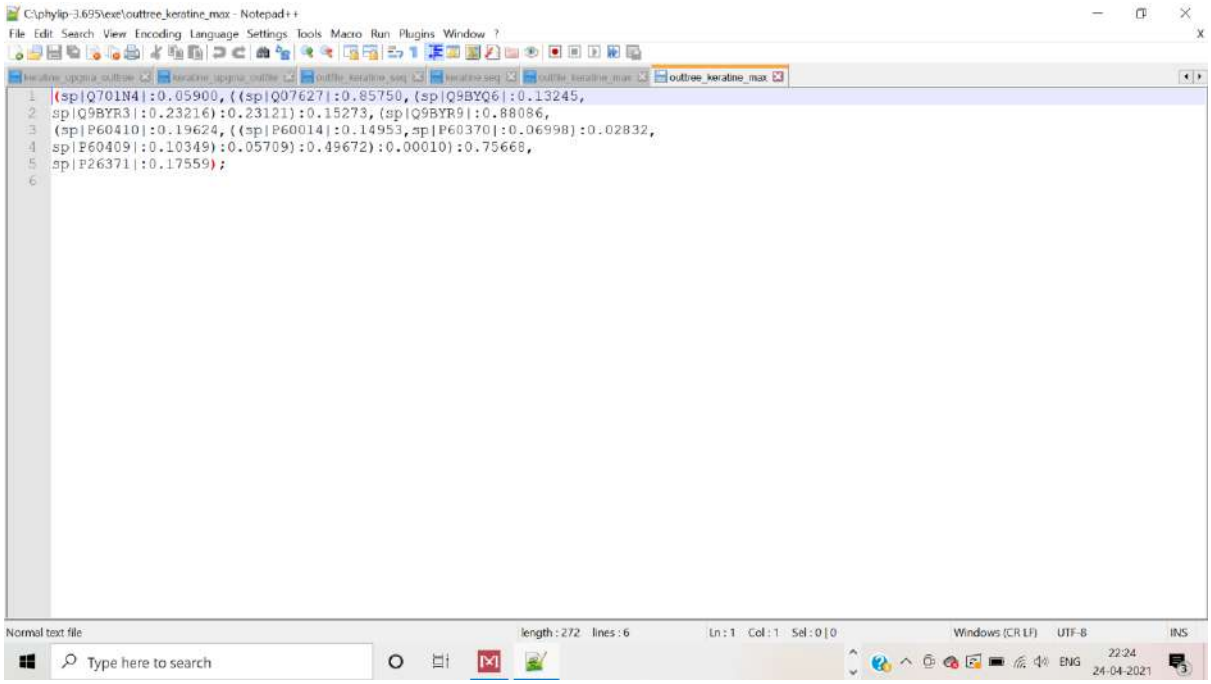


Step 2: Jones – Taylor – Thornton model is obtained for the sequence using maximum likelihood method.



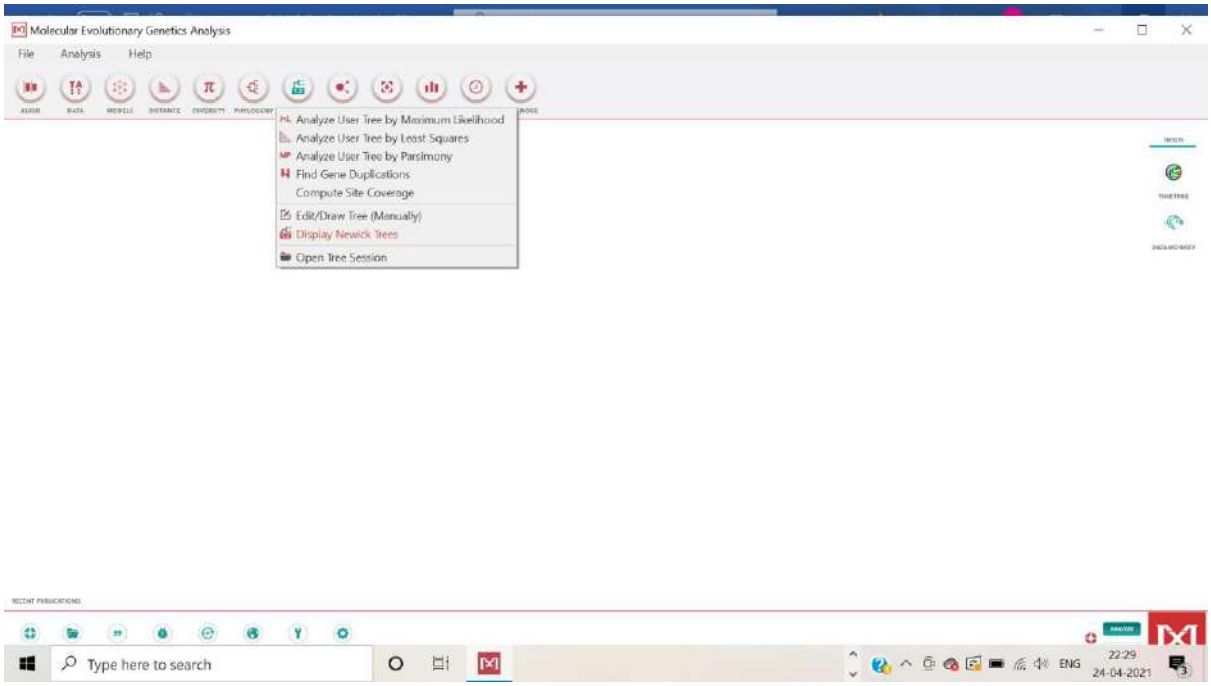


```
remember: this is an unrooted tree!

Ln Likelihood = -3791.22255

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

     1          sp|P26371|        0.17559     (  0.08909,     0.26208) **
     1          sp|Q701N4|        0.05900     (     zero,     0.12361) **
     1             7              0.75668     (  0.52311,     0.99025) **
     7             2              0.15273     (  0.01834,     0.28712) **
     2          sp|Q07627|        0.85750     (  0.61803,     1.09697) **
     2             8              0.23121     (  0.09545,     0.36698) **
     8          sp|Q9BYQ6|        0.13245     (  0.04740,     0.21749) **
     8          sp|Q9BYR3|        0.23216     (  0.13385,     0.33046) **
     7             6              0.00010     (     zero,     0.08292)
     6          sp|Q9BYR9|        0.88086     (  0.61812,     1.14359) **
     6             3              0.49672     (  0.33959,     0.65386) **
     3          sp|P60410|        0.19624     (  0.12669,     0.26579) **
     3             4              0.05709     (  0.00639,     0.10777) **
     4             5              0.02832     (     zero,     0.06178) *
     5          sp|P60014|        0.14953     (  0.09563,     0.20345) **
     5          sp|P60370|        0.06998     (  0.03269,     0.10728) **
     4          sp|P60409|        0.10349     (  0.05814,     0.14884) **

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

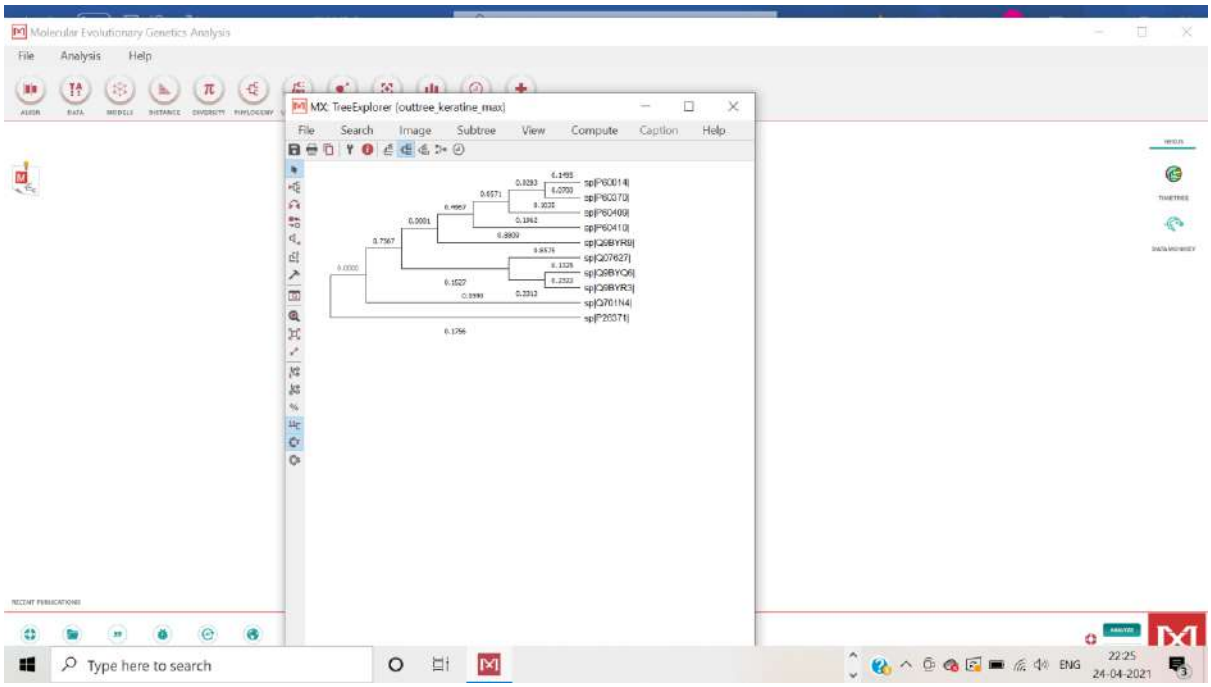


```
(sp|Q701N4|:0.05900,((sp|Q07627|:0.85750,(sp|Q9BYQ6|:0.13245,
sp|Q9BYR3|:0.23216):0.23121):0.15273,(sp|Q9BYR9|:0.88086,
(sp|P60410|:0.19624,((sp|P60014|:0.14953,sp|P60370|:0.06998):0.02832,
sp|P60409|:0.10349):0.05709):0.49672):0.00010):0.75668,
sp|P26371|:0.17559);
```

Step 3: Go to MEGA to obtain a phylogenetic tree. Click on user tree and select Display Newick tree.



Step 4: Phylogenetic tree is obtained using maximum likelihood method.



- ✓ Performing protein analysis using Character based methods:-

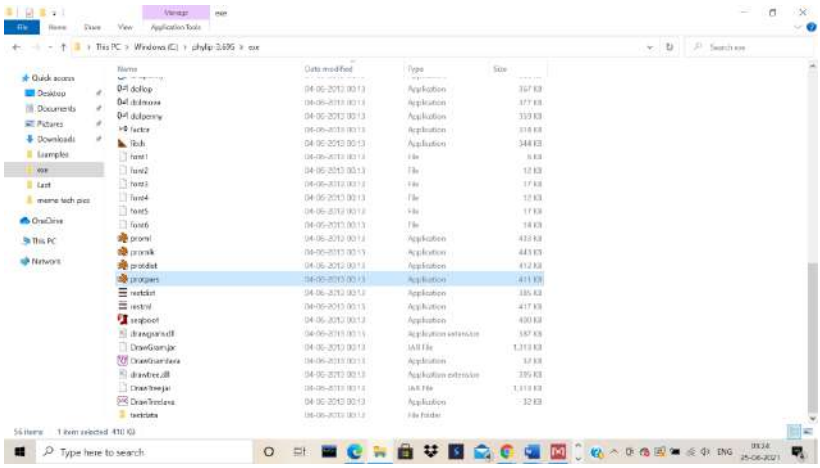
Parsimony: Which has protpars as protein programs used in Parsimony.

This will test the tree which you provide, and report on how accurate the tree is in relation to the data file you have open. The best tree with this method will be the one with the least evolutionary change required.

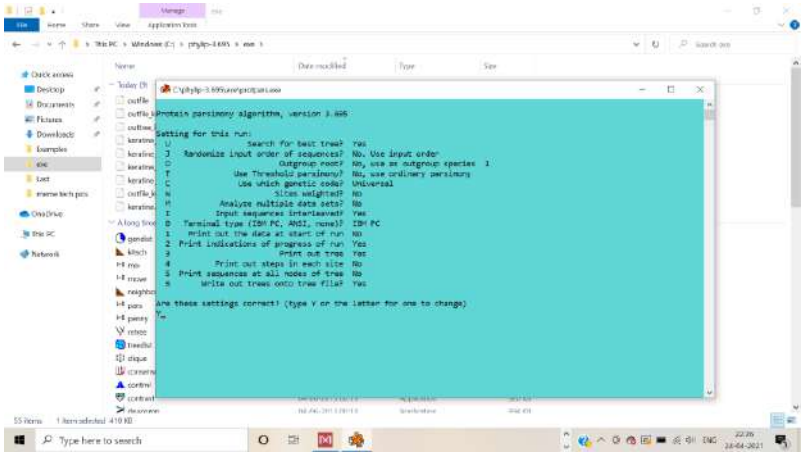
Reconstruction of the evolutionary history of genes and species is currently one of the most important subjects in molecular evolution. If reliable phylogenies are produced, they will shed light on the sequence of evolutionary events that generated the present day diversity of genes and species and help us to understand the mechanisms of evolution as well as the history of organisms.

Phylogenetic relationships of genes or organisms usually are presented in a treelike form with a root, which is called a *rooted tree*. It also is possible to draw a tree without a root, which is called an *unrooted tree*. The branching pattern of a tree is called a *topology*.

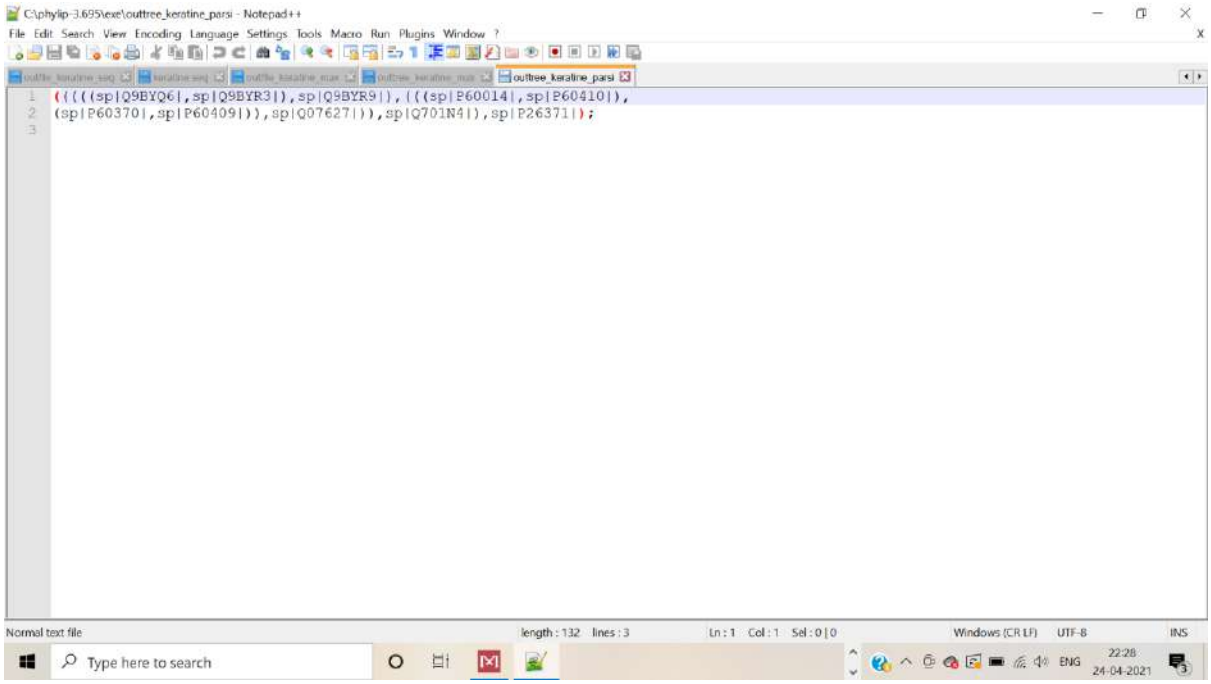
Step 1 : Open the protpars document.



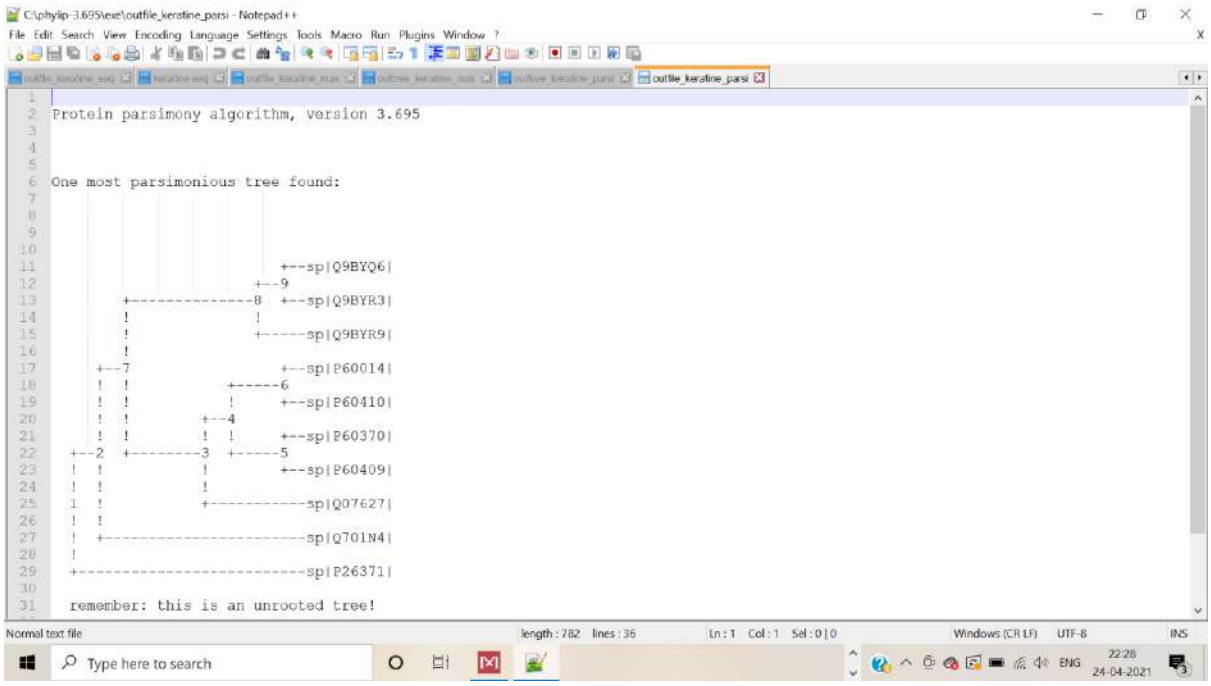
Step 2: Press Y and accept all the parameters.



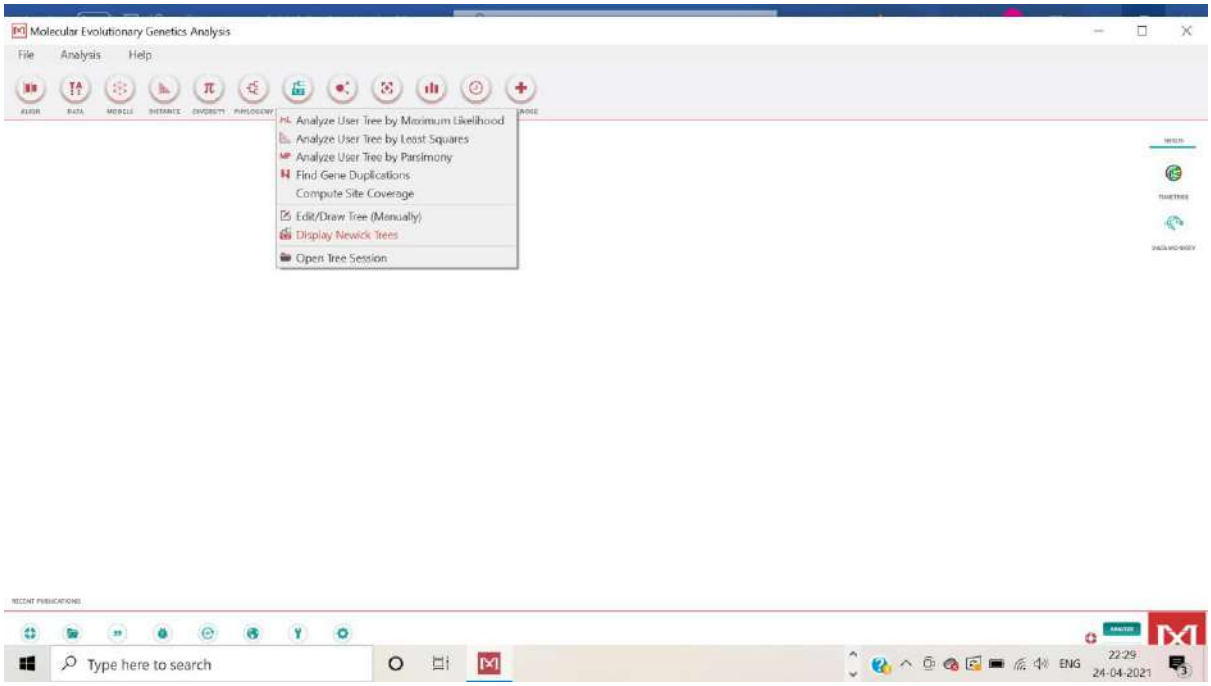
Step 3: Out tree for the respective result.



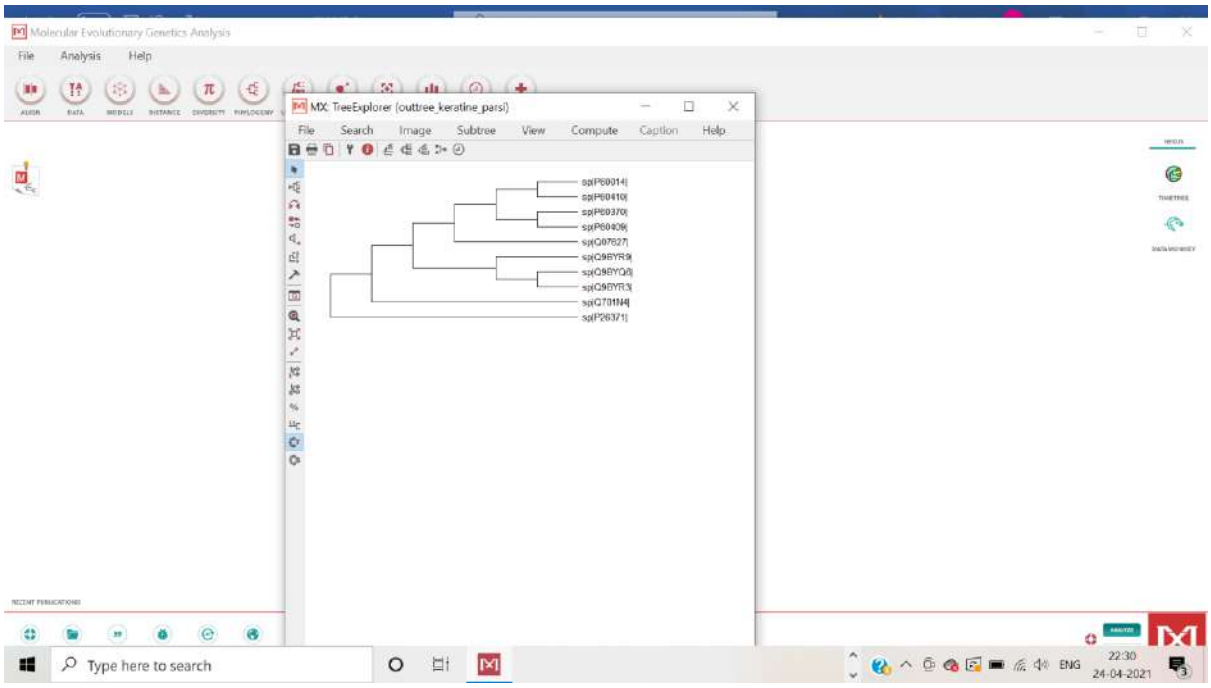
Step 4: You will find the most parsimonious tree.



Step 5: Go to MEGA and click on user tree and select Display Newick Tree.



Step 6: You will get the phylogenetic tree for parsimony method.

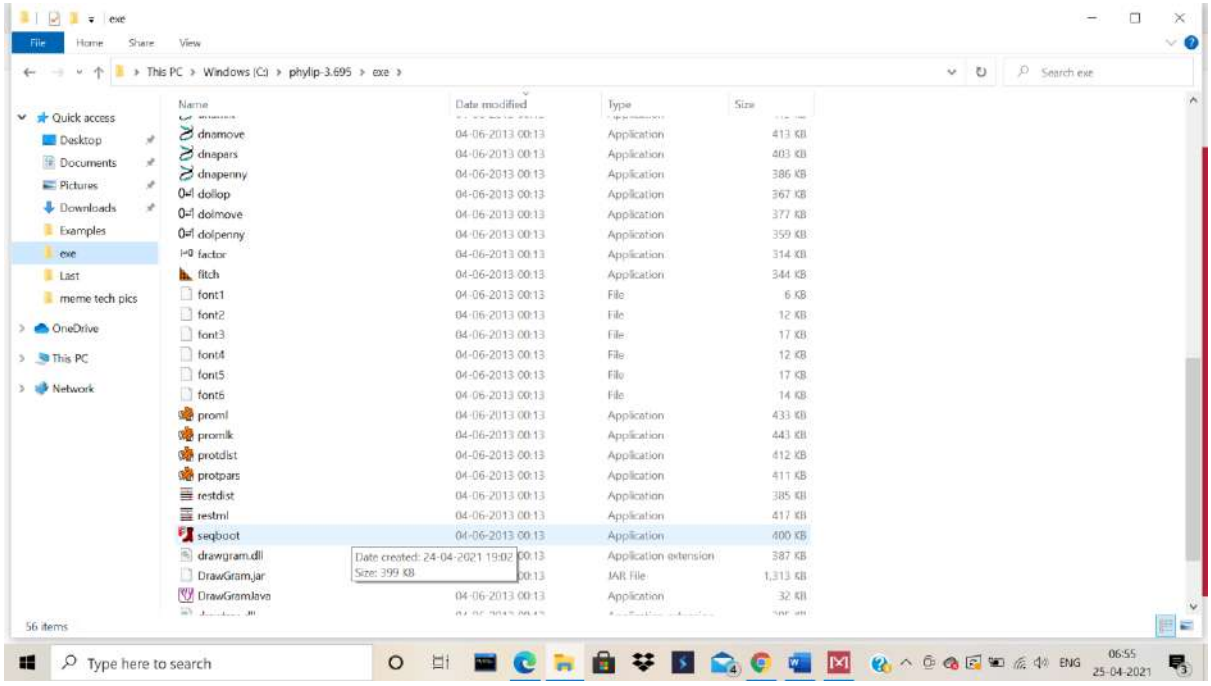


- ✓ **Performing Bootstrapping from the results of NJ and ML methods:**

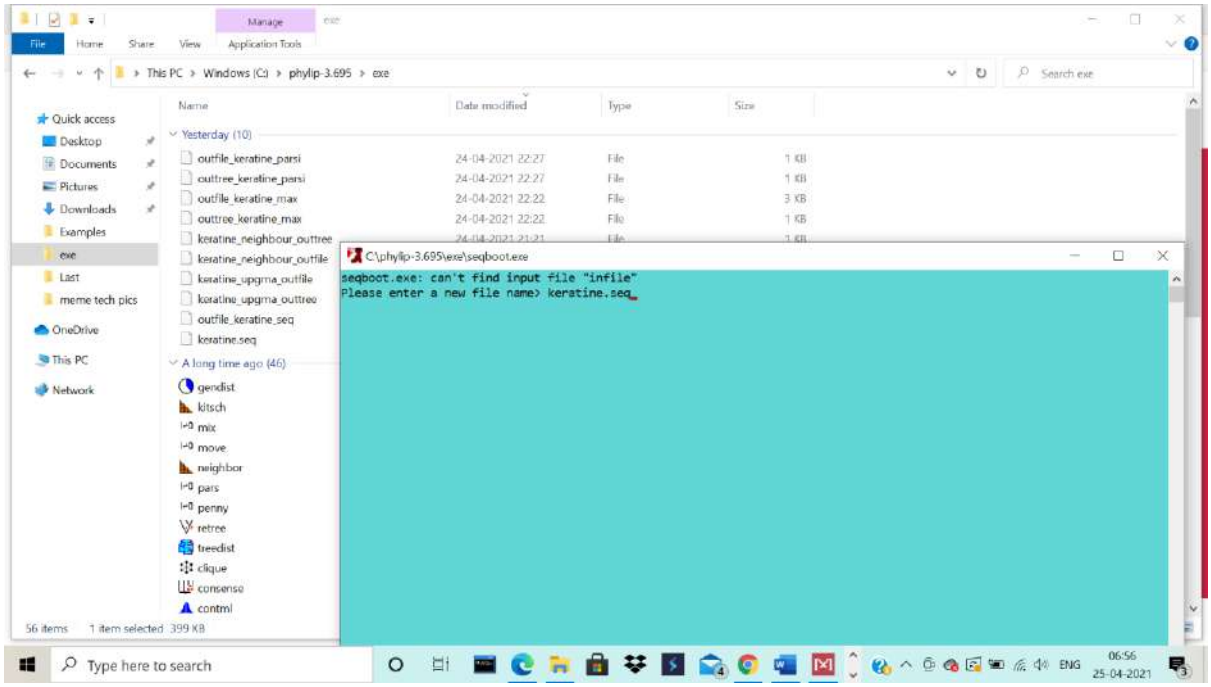
Bootstrapping:- It is basically a type of alignment which is already done , but we are actually reshuffling the positions of the sequences .i.e. certain positions are repeated, certain positions are deleted in short the shuffling of alignment. Bootstrap basically is used for building up the confidence values of the tree.

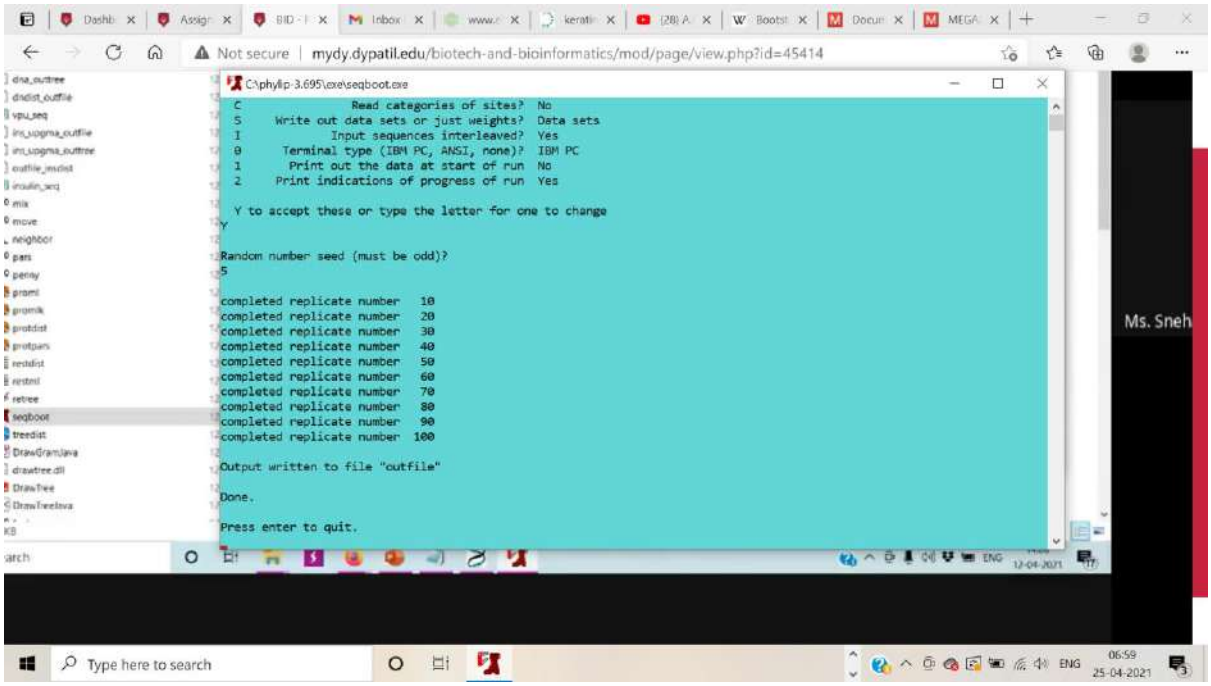
NJ:

Step 1: Select seqboot from exe folder.

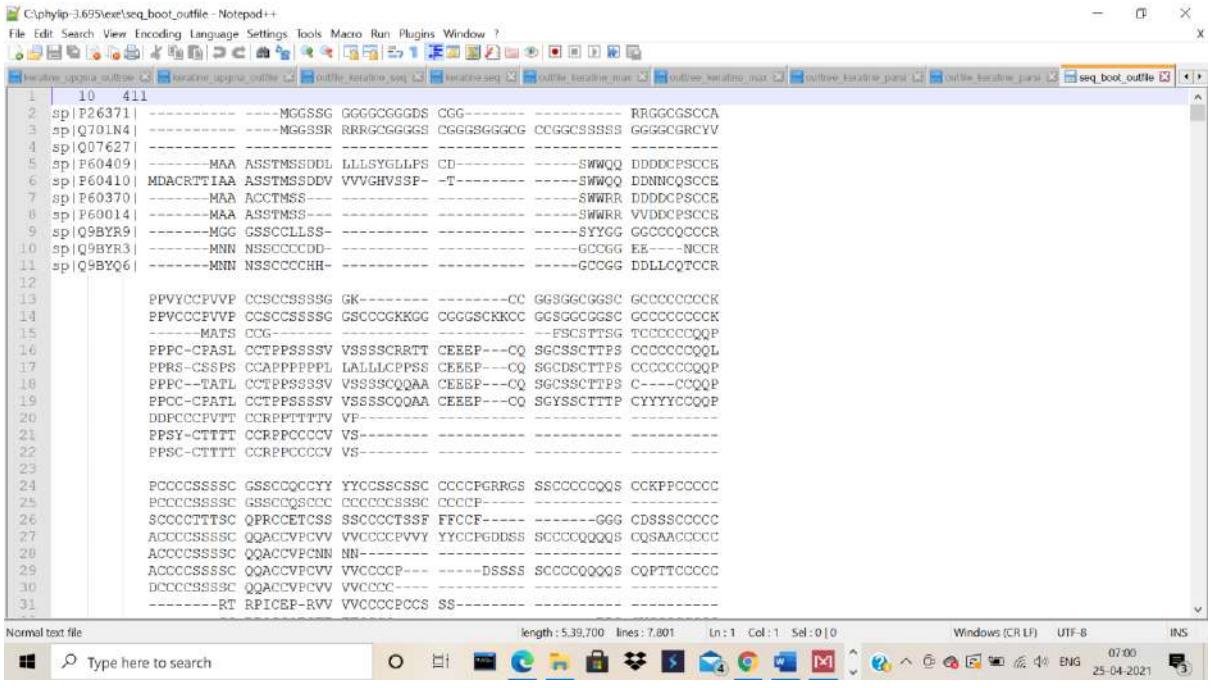


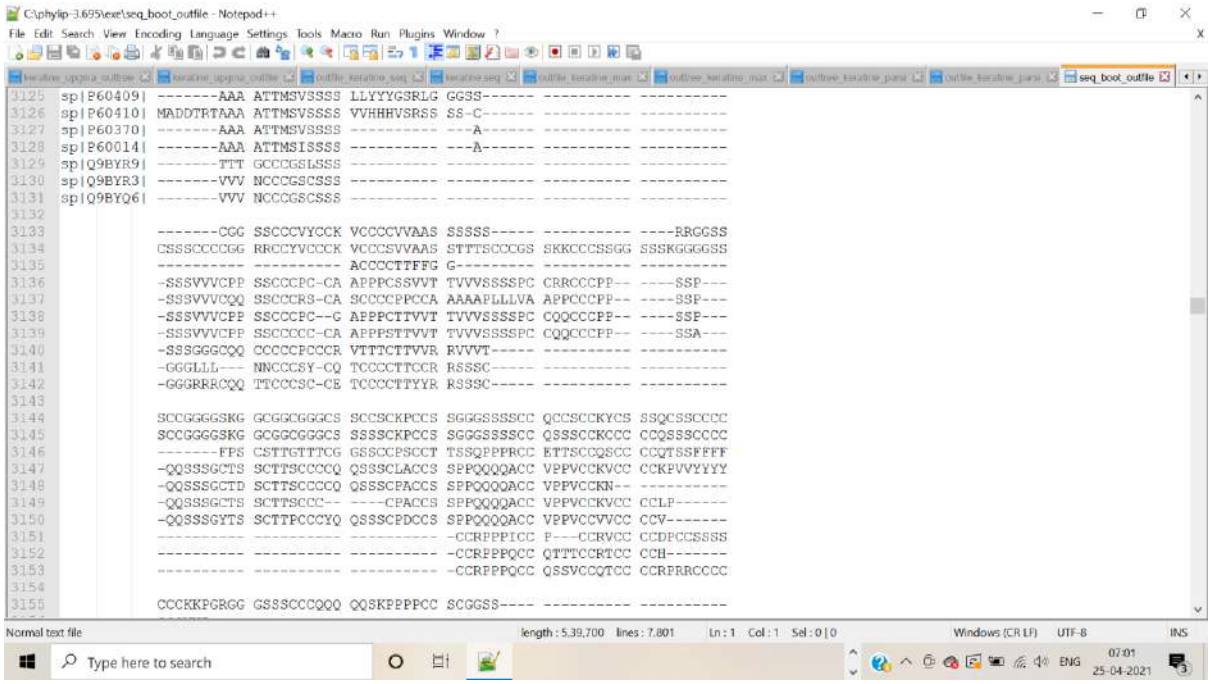
Step 2: Enter the saved file name = keratine.seq

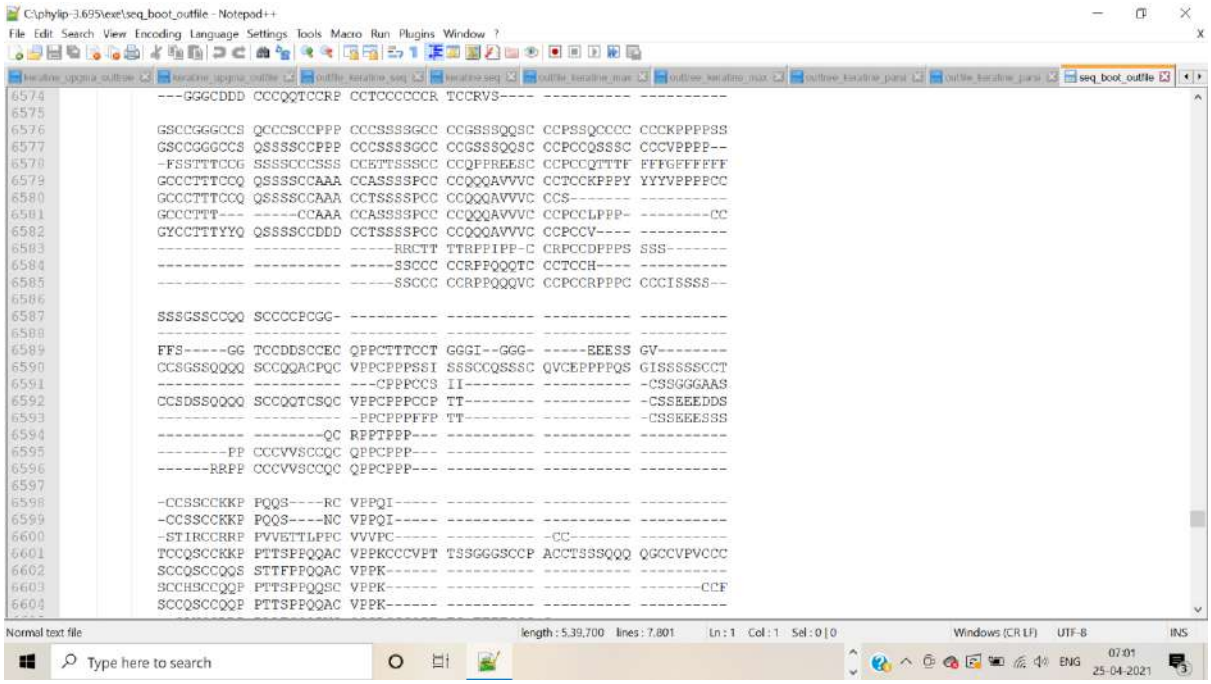




Step 3: Once opened the file in notpade ++ you will get 100 different alignment sets for the sequences.



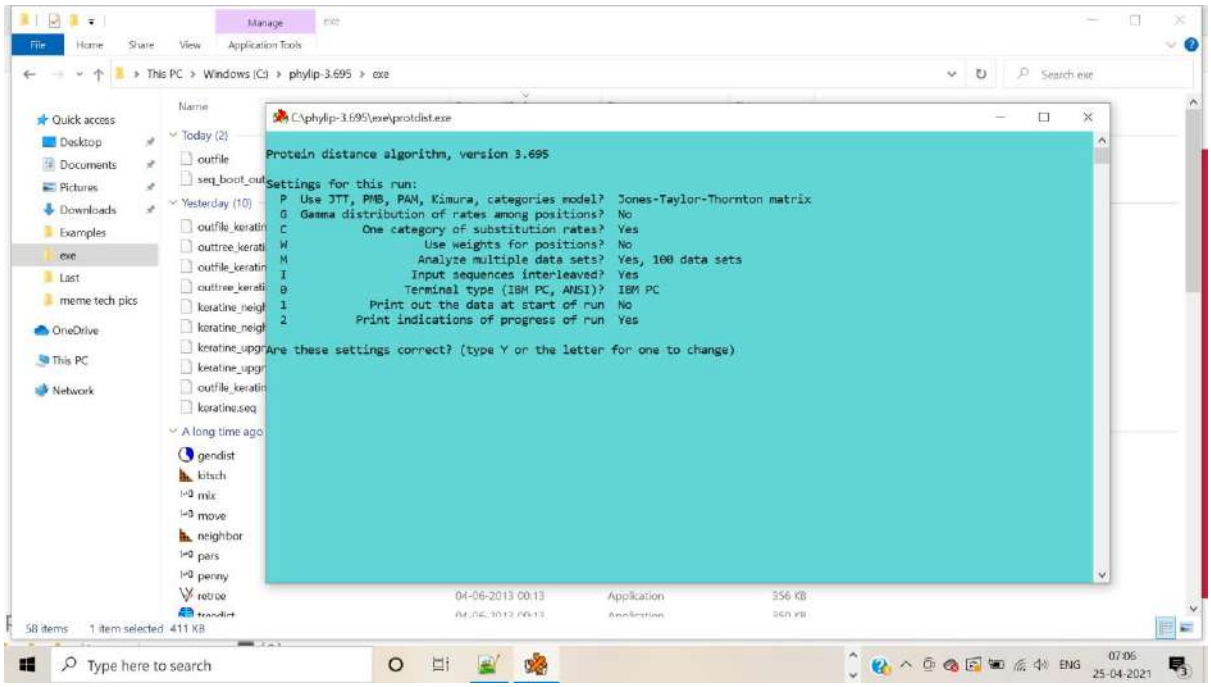

C:\phylip-3.695\exe\seq_boot_outfile - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
seq_boot_outfile
sp|P60409| -------AAA ATTMSVSSSS LLYYYGSRLG GGSS------ ---------- ----------
sp|P60410| MADDTRTAAA ATTMSVSSSS VVHHHVSRSS SS-C------ ---------- ----------
sp|P60370| -------AAA ATTMSVSSSS ---------- ---A------ ---------- ----------
sp|P60014| -------AAA ATTMSISSSS ---------- ---A------ ---------- ----------
sp|Q9BYR9| -------TTT GCCCGSLSSS ---------- ---------- ---------- ----------
sp|Q9BYR3| -------VVV NCCCGSCSSS ---------- ---------- ---------- ----------
sp|Q9BYQ6| -------VVV NCCCGSCSSS ---------- ---------- ---------- ----------
-------CGG SSCCCVYCCK VCCCCVVAAS SSSSS----- ---------- ----RRGGSS
CSSSCCCCGG RRCCYVCCCK VCCCSVVAAS STTTSCCCGS SKKCCCSSGG SSSKGGGGSS
---------- ---------- ACCCCTTFFG G--------- ---------- ----------
-SSSVVVCPP SSCCCPC-CA APPPCSSVVT TVVVSSSSPC CRRCCCPP-- ----SSP---
-SSSVVVCQQ SSCCCRS-CA SCCCCPPCCA AAAAPLLLVA APPCCCPP-- ----SSP---
-SSSVVVCPP SSCCCPC--G APPPCTTVVT TVVVSSSSPC CQQCCCPP-- ----SSP---
-SSSVVVCPP SSCCCCC-CA APPPSTTVVT TVVVSSSSPC CQQCCCPP-- -----SSA---
-SSSGGGCQQ CCCCCPCCCR VTTTCTTVVR RVVVT----- ---------- ----------
-GGGLLL--- NNCCCSY-CQ TCCCCTTCCR RSSSC----- ---------- ----------
-GGGRRRCQQ TTCCCSC-CE TCCCCTTYYR RSSSC----- ---------- ----------
SCCGGGGSKG GCGGCGGGCS SCCSCKPCCS SGGGSSSSCC QCCSCCKYCS SSQCSSCCCC
SCCGGGGSKG GCGGCGGGCS SSSSCKPCCS SGGGSSSSCC QSSSCCKCCC CCQSSSCCCC
-------FPS CSTTGTTTCG GSSCCPSCCT TSSQPPPRCC ETTSCCQSCC CCQTSSFFFF
-QQSSSGCTS SCTTSCCCCQ QSSSCLACCS SPPQQQQACC VPPVCCKVCC CCKPVVYYYY
-QQSSSGCTD SCTTSCCCCQ QSSSCPACCS SPPQQQQACC VPPVCCKN-- ----------
-QQSSSGCTS SCTTSCCC-- ----CPACCS SPPQQQQACC VPPVCCKVCC CCLP------
-QQSSSGYTS SCTTPCCCYQ QSSSCPDCCS SPPQQQQACC VPPVCCVVCC CCV-------
---------- ---------- ---------- -CCRPPPICC P---CCRVCC CCDPCCSSSS
---------- ---------- ---------- -CCRPPPQCC QTTTCCRTCC CCH-------
---------- ---------- ---------- -CCRPPPQCC QSSVCCQTCC CCRPRRCCCC
CCCKKPGRGG GSSSCCCQQQ QQSKPPPPCC SCGGSS---- ---------- ----------
Normal text file
length : 5,39,700 lines : 7,801
Ln : 1 Col : 1 Sel : 0 | 0
Windows (CR LF)
UTF-8
INS
Type here to search
ENG
07:01
25-04-2021


C:\phylip-3.695\exe\seq_boot_outfile - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
seq_boot_outfile
---GGGCDDD CCCQQTCCRP CCTCCCCCCR TCCRVS---- ---------- ----------
GSCCGGGCCS QCCCSCCPPP CCCSSSSGCC CCGSSSQQSC CCPSSQCCCC CCCKPPPPSS
GSCCGGGCCS QSSSSCCPPP CCCSSSSGCC CCGSSSQQSC CCPCCQSSSC CCCVPPPP--
-FSSTTTCCG SSSSCCCSSS CCETTSSSCC CCQPPREESC CCPCCQTTTF FFFGFFFFFF
GCCCTTTCCQ QSSSSCCAAA CCASSSSPCC CCQQQAVVVC CCTCCKPPPY YYYVPPPCC
GCCCTTTCCQ QSSSSCCAAA CCTSSSSPCC CCQQQAVVVC CCS------- ----------
GCCCTTT--- -----CCAAA CCASSSSPCC CCQQQAVVVC CCPCCLPPP- --------CC
GYCCTTTYYQ QSSSSCCDDD CCTSSSSPCC CCQQQAVVVC CCPCCV---- ----------
---------- ---------- -----RRCTT TTRPPIPP-C CRPCCDPPPS SSS-------
---------- ---------- -----SSCCC CCRPPQQQTC CCTCCH---- ----------
---------- ---------- -----SSCCC CCRPPQQQVC CCPCCRPPPC CCCISSSS--
SSSGSSCCQQ SCCCCPCGG- ---------- ---------- ---------- ----------
FFS-----GG TCCDDSCCEC QPPCTTTCCT GGGI--GGG- -----EEESS GV--------
CCSGSSQQQQ SCCQQACPQC VPPCPPPSSI SSSCCQSSSC QVCEPPPPQS GISSSSSCCT
---------- ---------- ---CPPPCCS II-------- ---------- -CSSGGGAAS
CCSDSSQQQQ SCCQQTCSQC VPPCPPPCCP TT-------- ---------- -CSSEEEDDS
---------- ---------- -PPCPPPFFP TT-------- ---------- -CSSEEESSS
---------- --------QC RPPTPPP--- ---------- ---------- ----------
--------PP CCCVVSCCQC QPPCPPP--- ---------- ---------- ----------
------RRPP CCCVVSCCQC QPPCPPP--- ---------- ---------- ----------
-CCSSCCKKP PQQS----RC VPPQI----- ---------- ---------- ----------
-CCSSCCKKP PQQS----NC VPPQI----- ---------- ---------- ----------
-STIRCCRRP PVVETTLPPC VVVPC----- ---------- -CC------- ----------
TCCQSCCKKP PTTSPPQQAC VPPKCCCVPT TSSGGGSCCP ACCTSSSQQQ QGCCVPVCCC
SCCQSCCQQS STTFPPQQAC VPPK------ ---------- ---------- ----------
SCCHSCCQQP PTTSPPQQSC VPPK------ ---------- ---------- -------CCF
SCCQSCCQQP PTTSPPQQAC VPPK------ ---------- ---------- ----------
Normal text file
length : 5,39,700 lines : 7,801
Ln : 1 Col : 1 Sel : 0 | 0
Windows (CR LF)
UTF-8
INS
Type here to search
ENG
07:01
25-04-2021

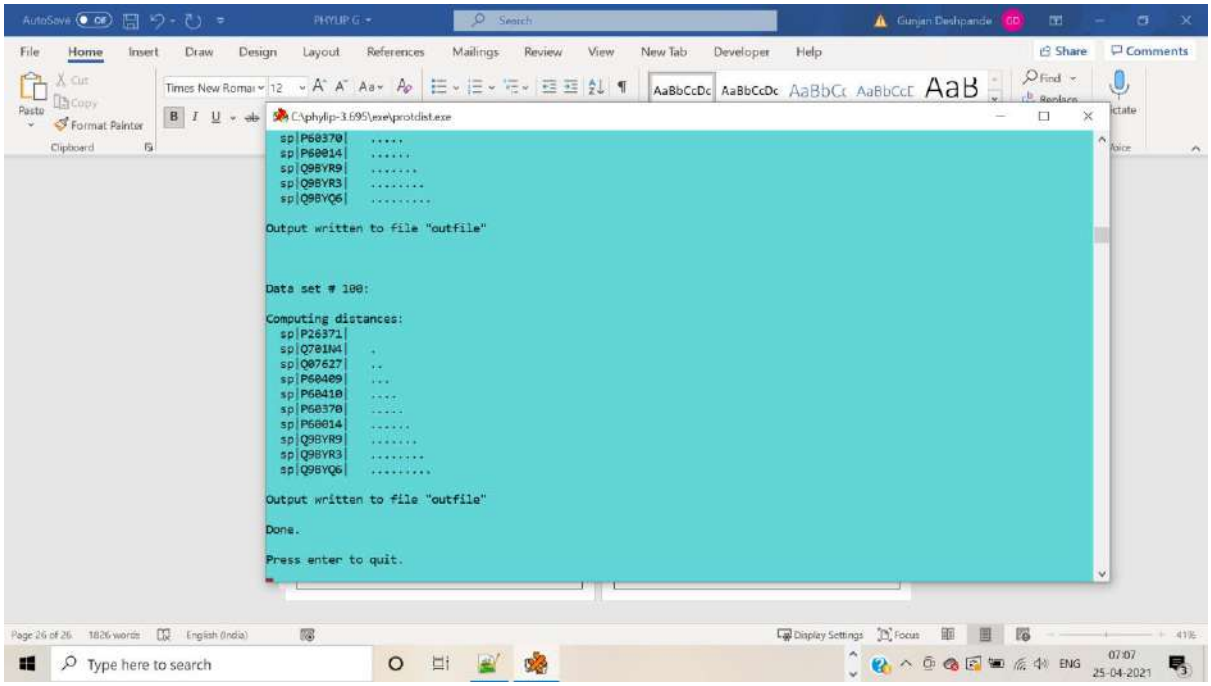
- ✓ **Performing Bootstrapping by distance method :**

Creating distance for 100 datasets:

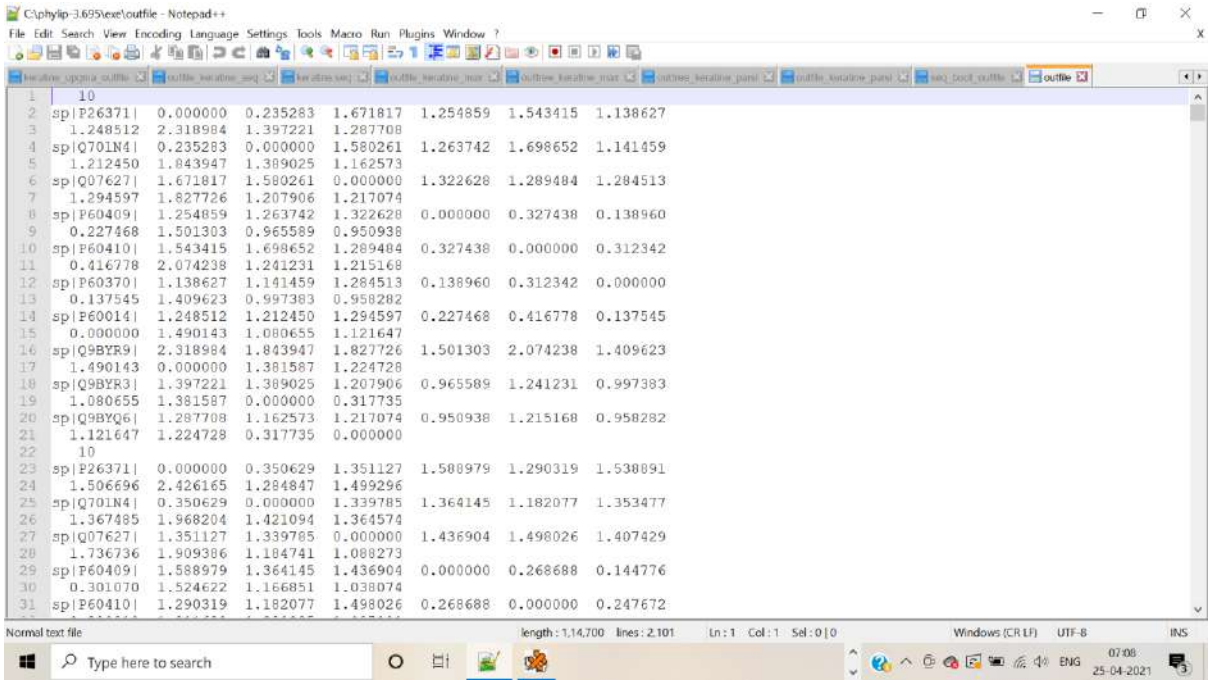
Step 1: Select the prodist file and Modify/Change the parameters.



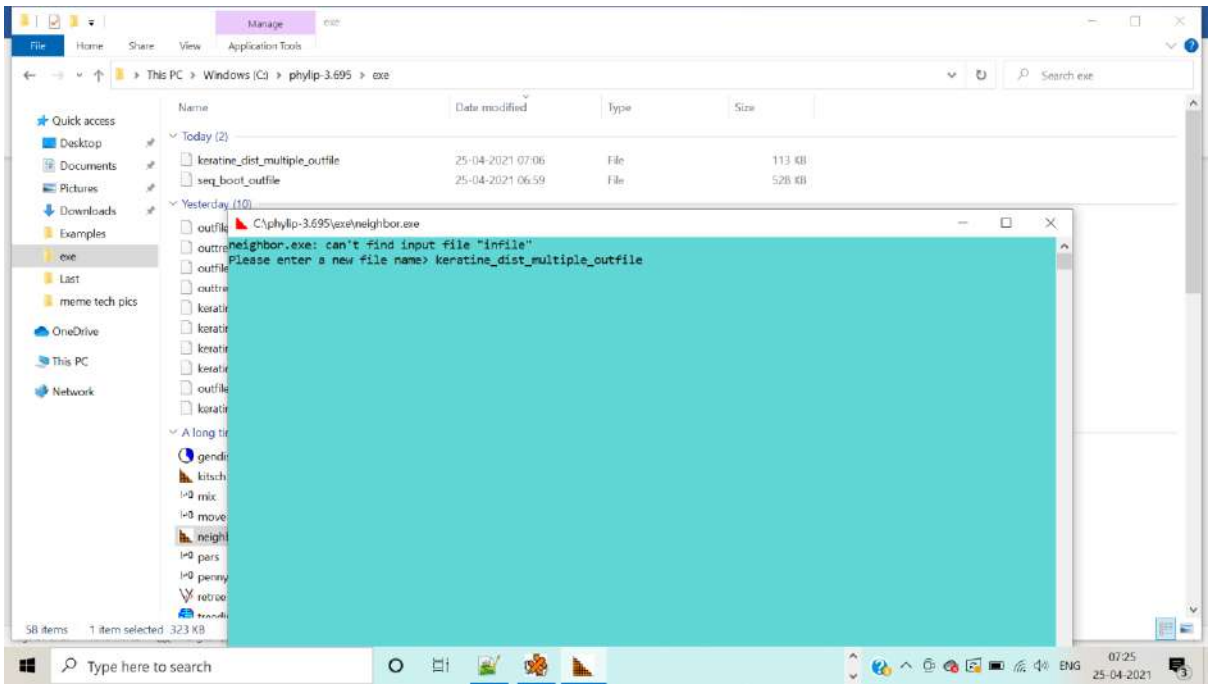
Step 2: Observe distance for 100 datasets



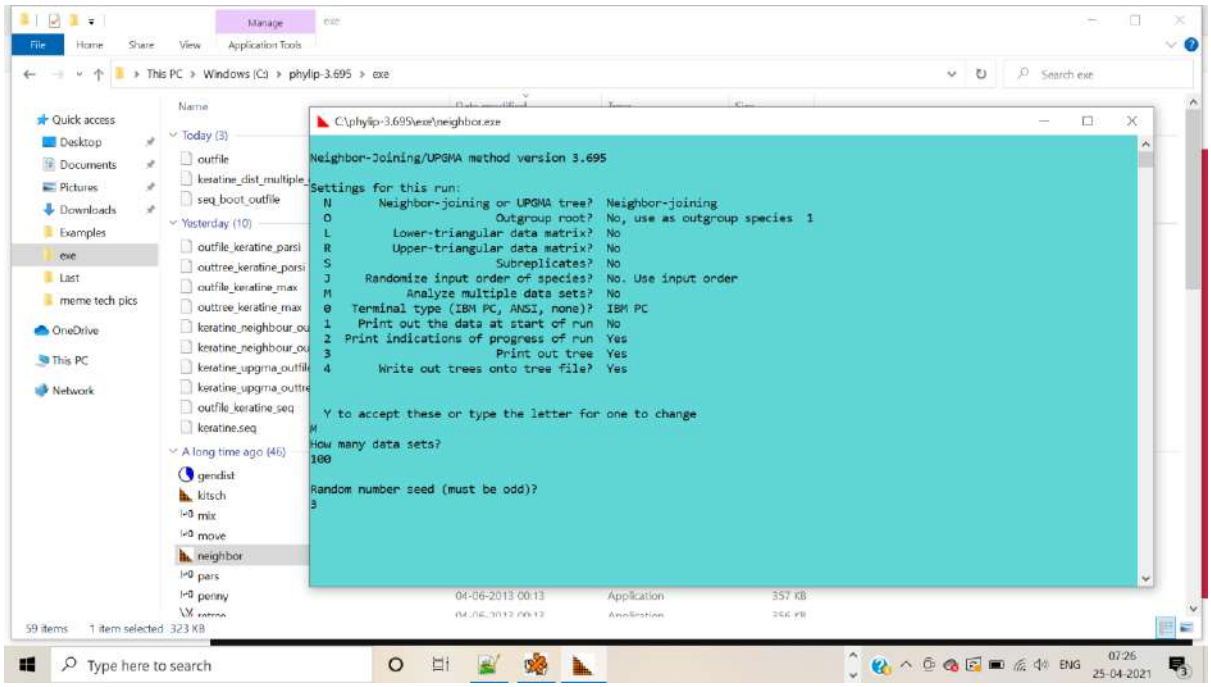
Step 3: Obtaining 100 distance matrices

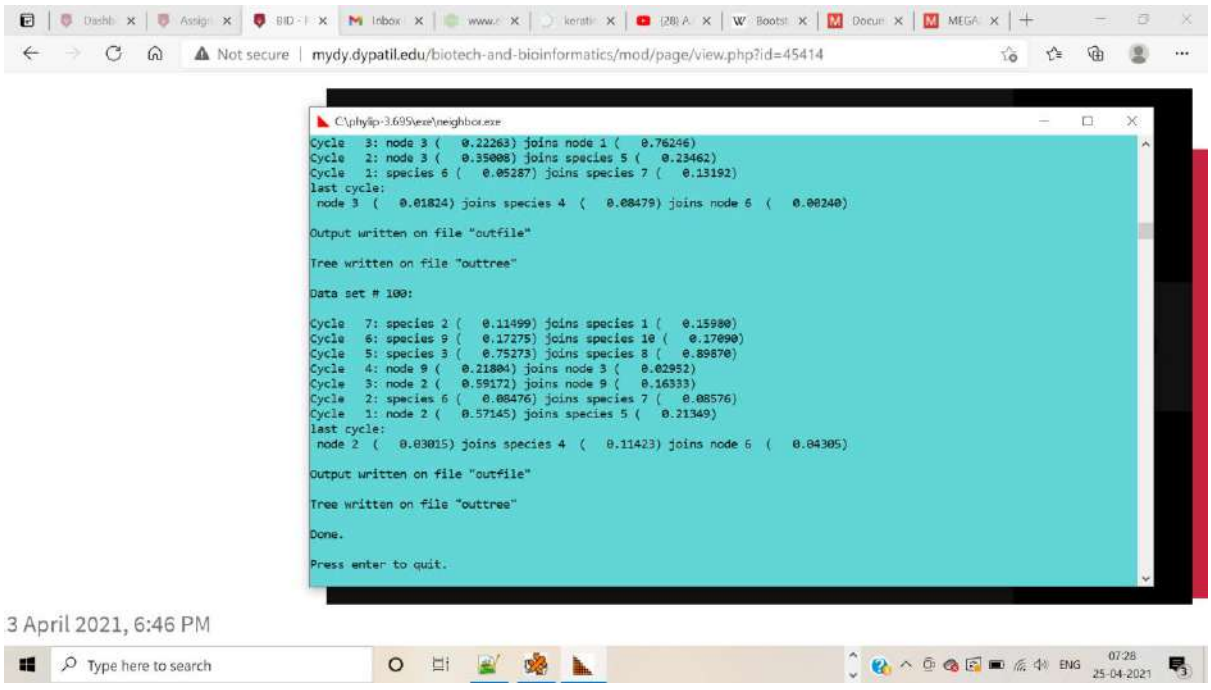


Step 4 : Obtaining Phylogenetic tree by Neighbour method.

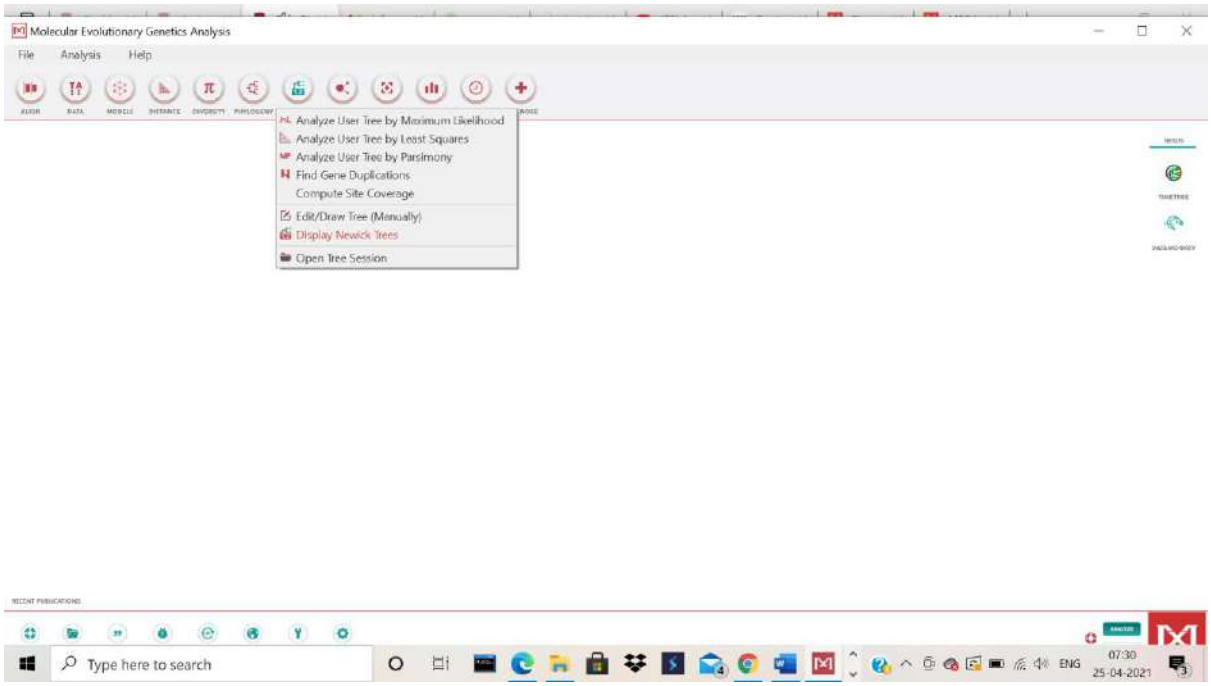


Setting the parameters:

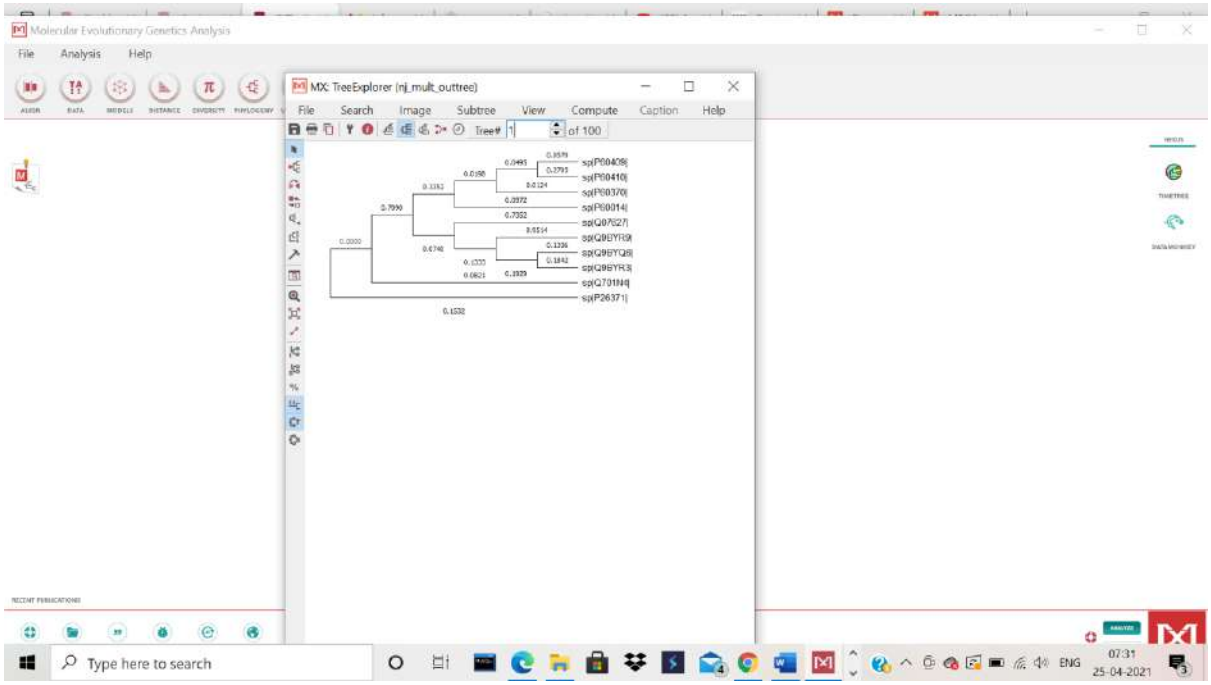




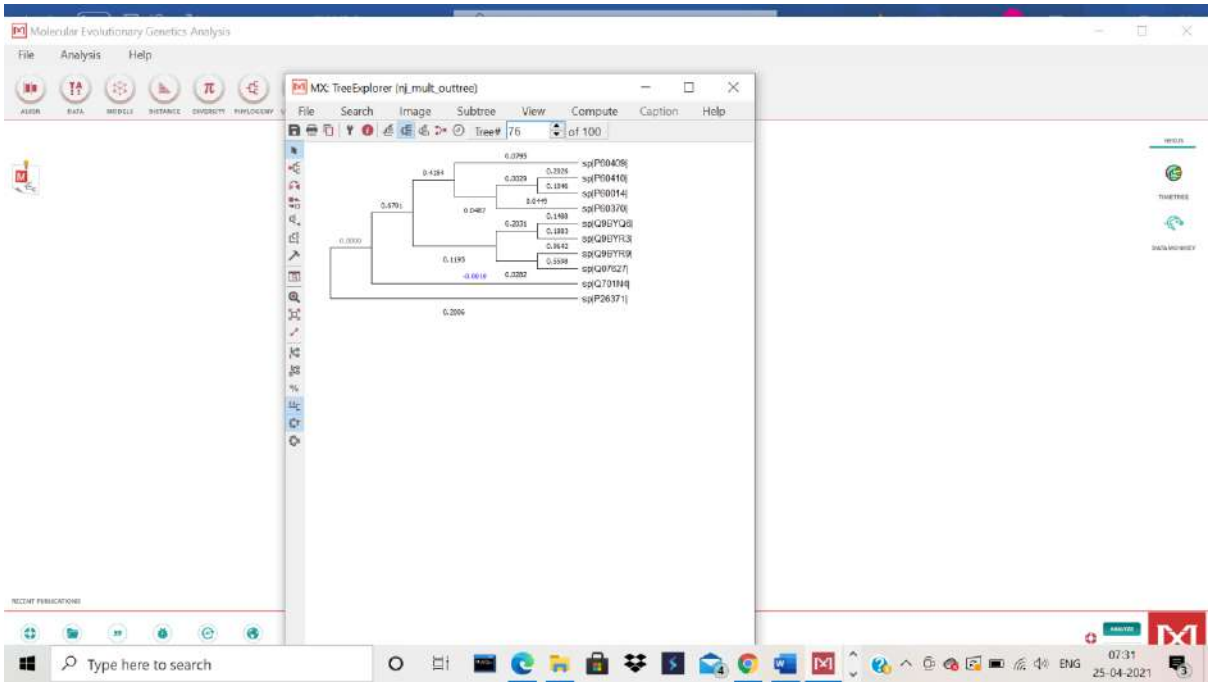
Step 5: To compare all 100 trees go to MEGA:



Step 6: Go to user tree and click on display newrick tree and see all the 1 of 100 trees.



76 of 100

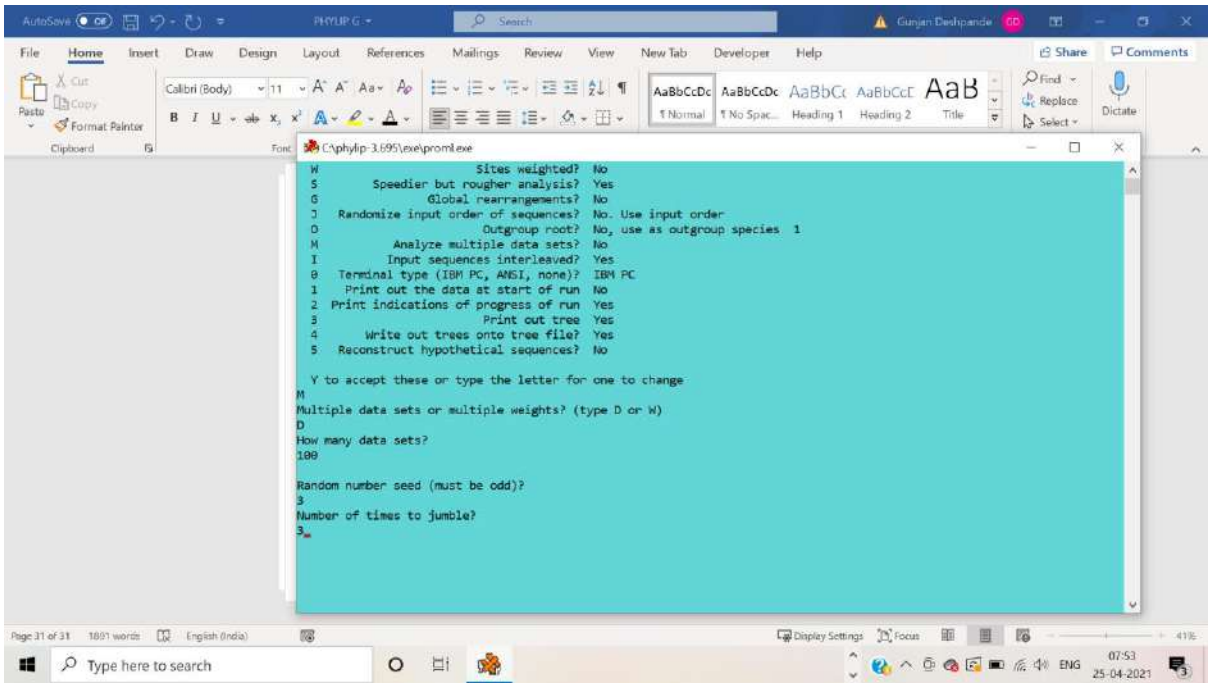

Molecular Evolutionary Genetics Analysis
File
Analysis
Help
MX: TreeExplorer (nj_mult_outtree)
File
Search
Image
Subtree
View
Compute
Caption
Help
Tree# 76
of 100
Type here to search
ENG
07:31
25-04-2021

\

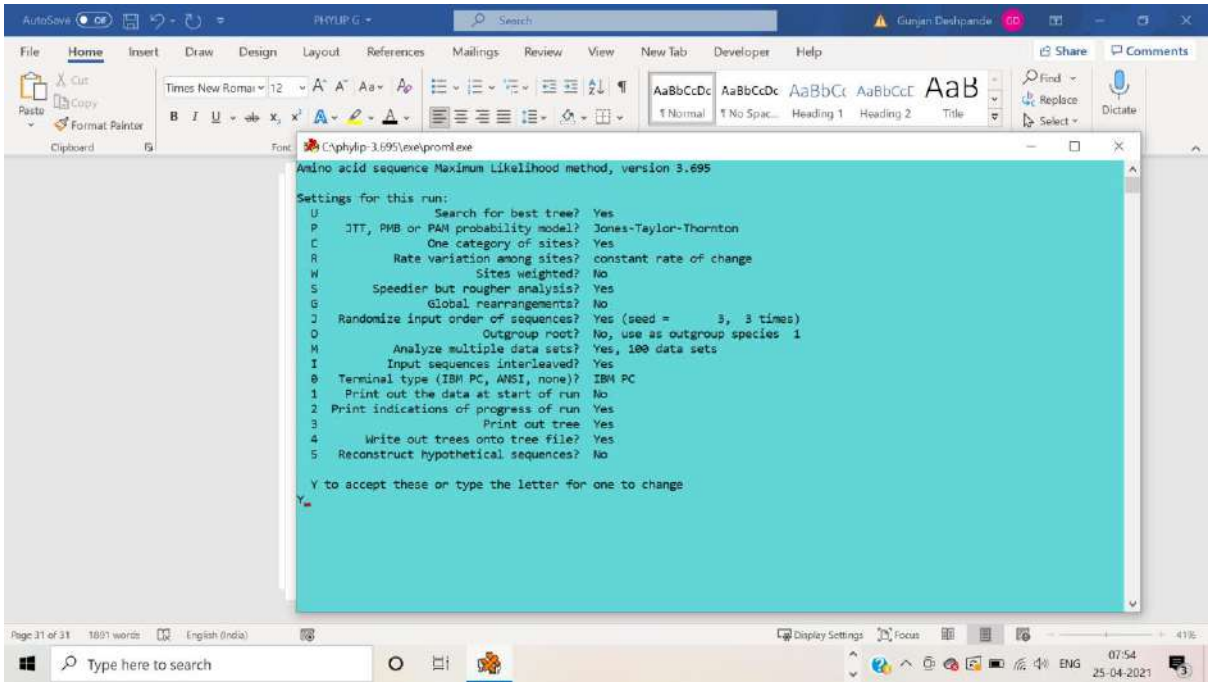
- ✓ **Performing Bootstrapping for Character based methods:-**

Maximum Likelihood: Where proml is the program used for proteins in ML.

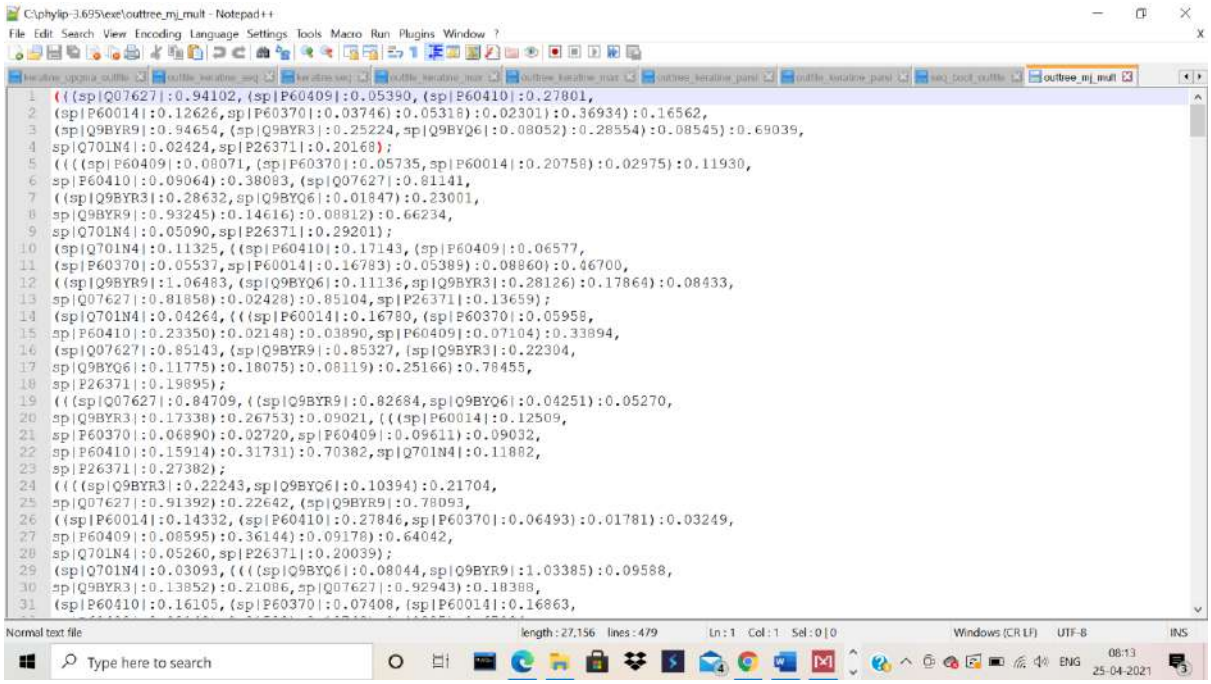
Step 1: Select parameters:



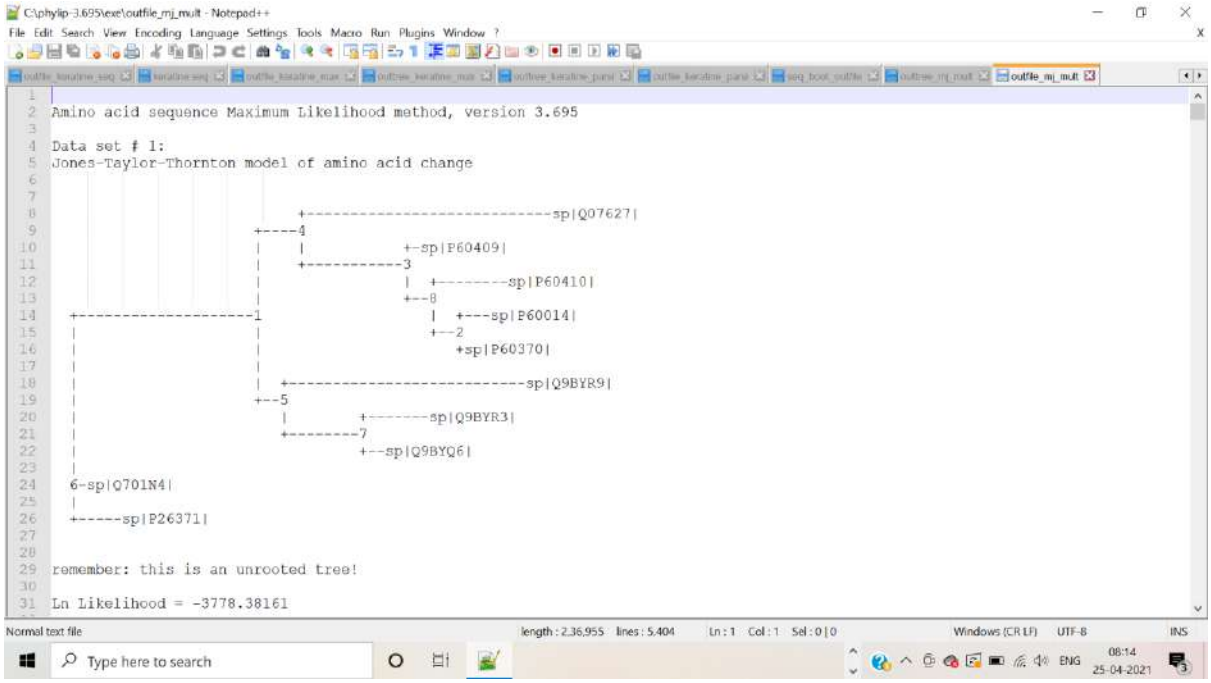
Step 2: Accept all the respective modified parameters by typing Y.

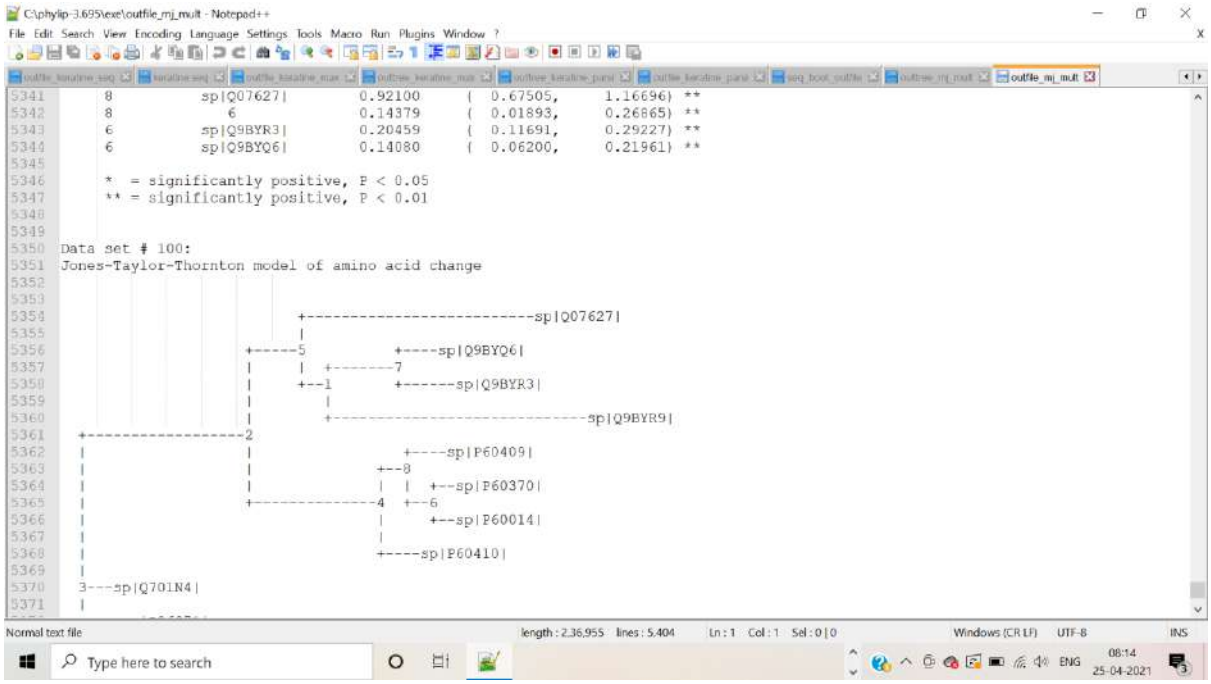


Step 3: Out tree obtained.

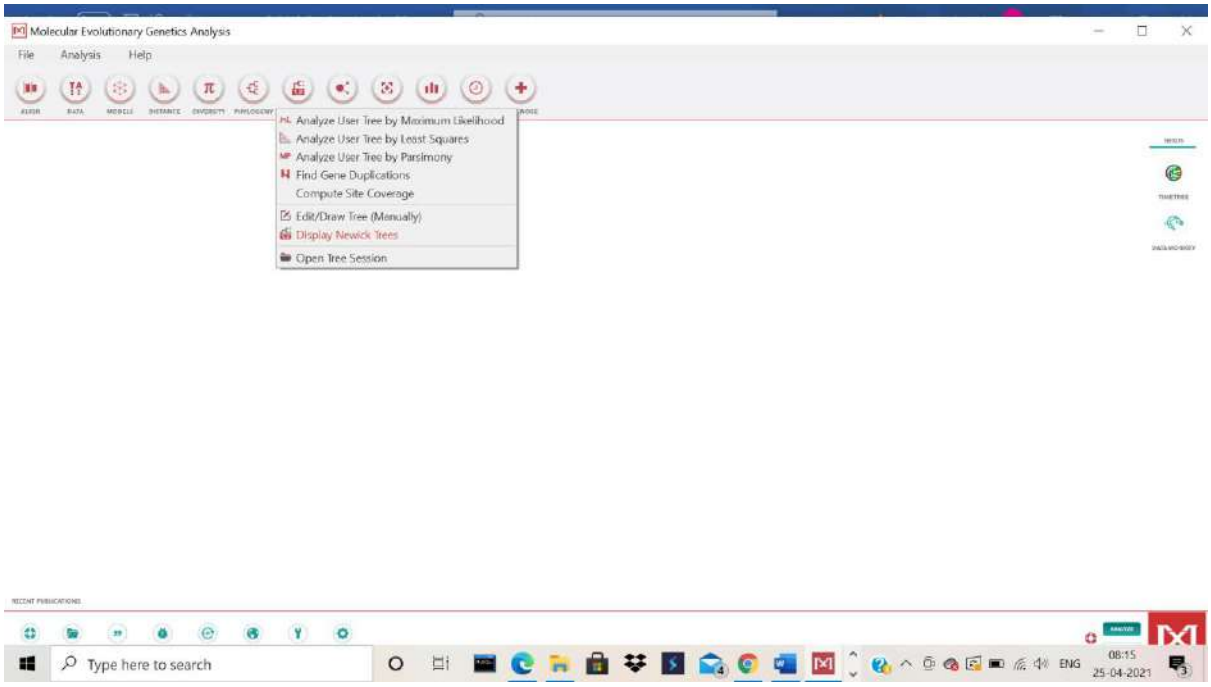


Step 4: Outfile for all 100 results.

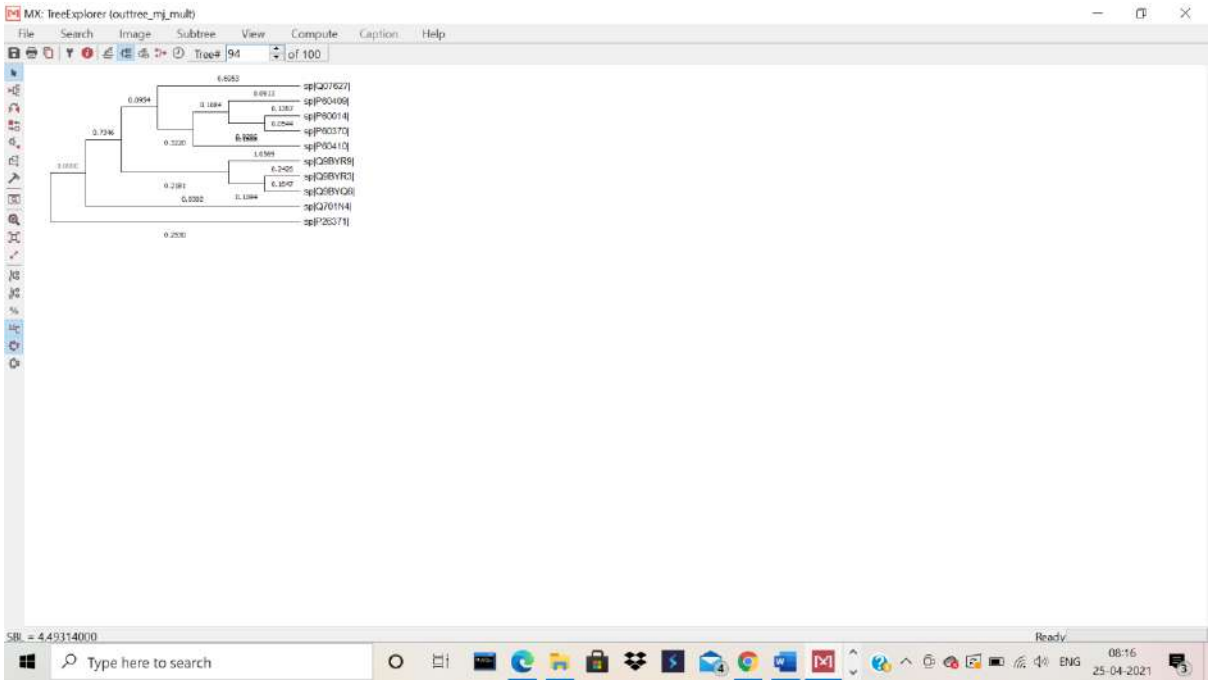




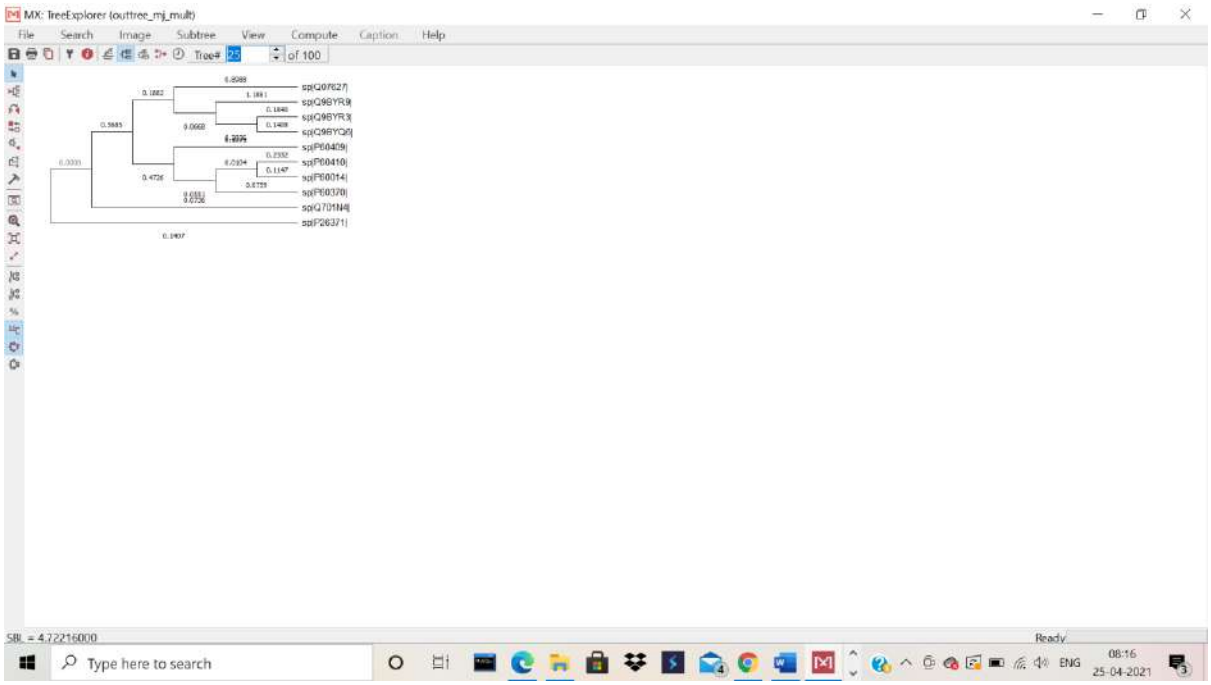
Step 5: Go to MEGA and select user tree and click on display netwrick tree.



94 of 100

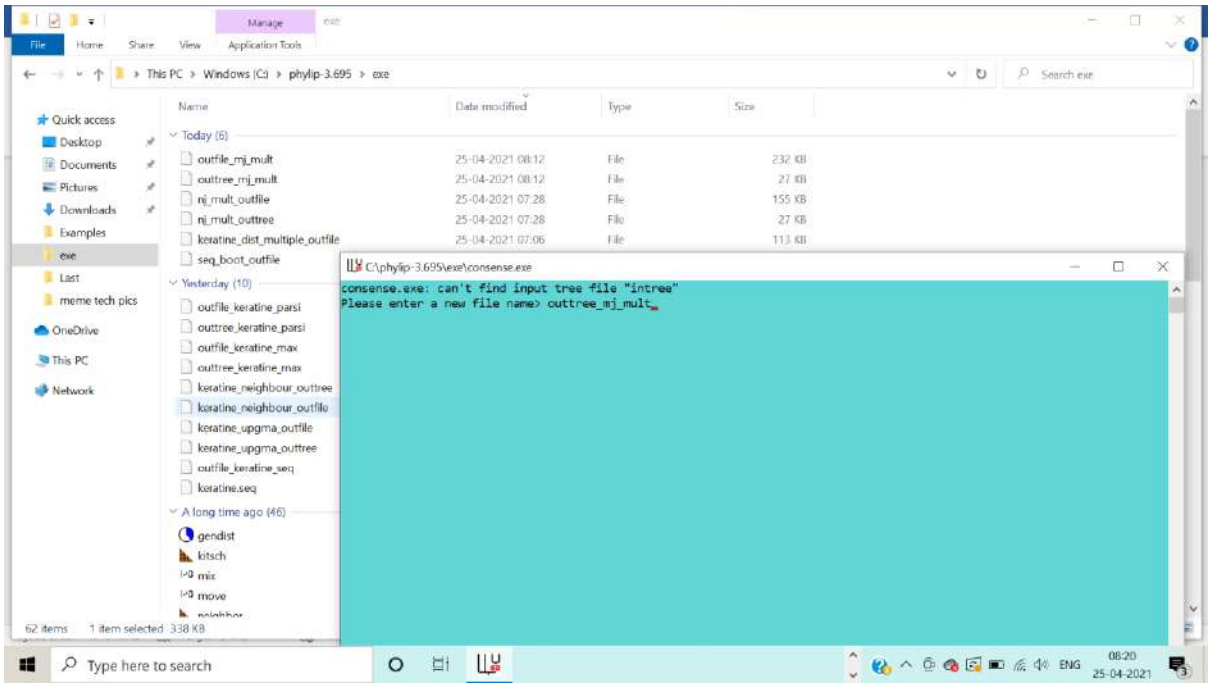


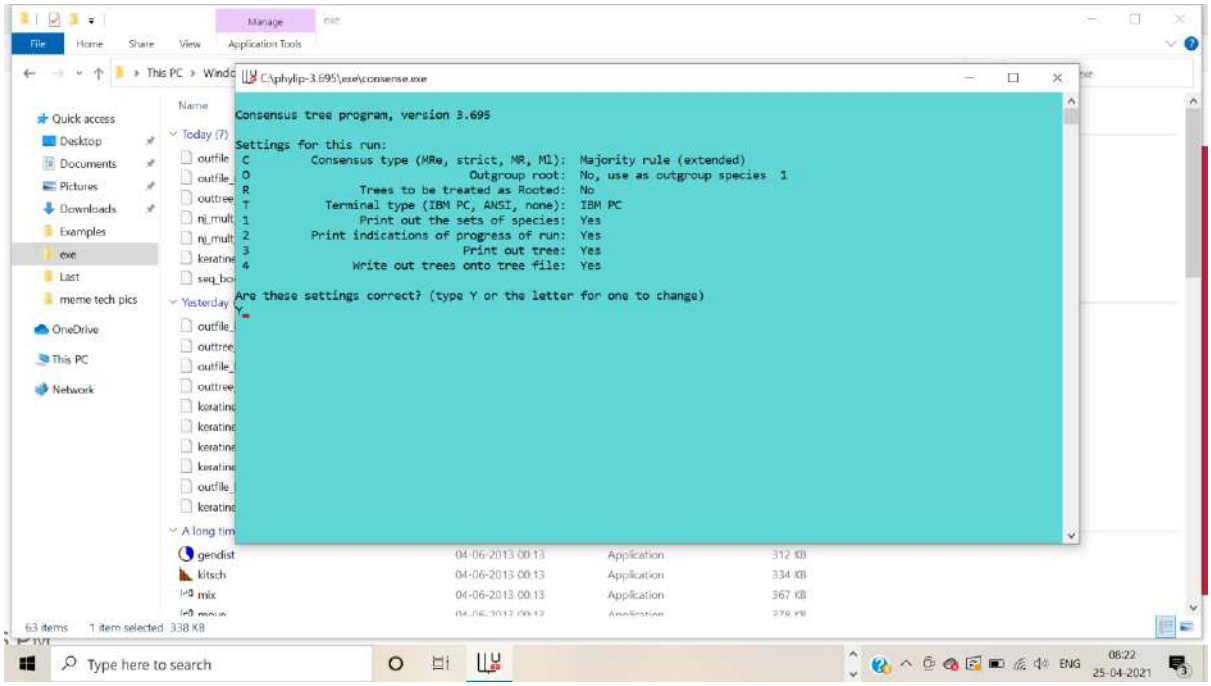
25 of 100



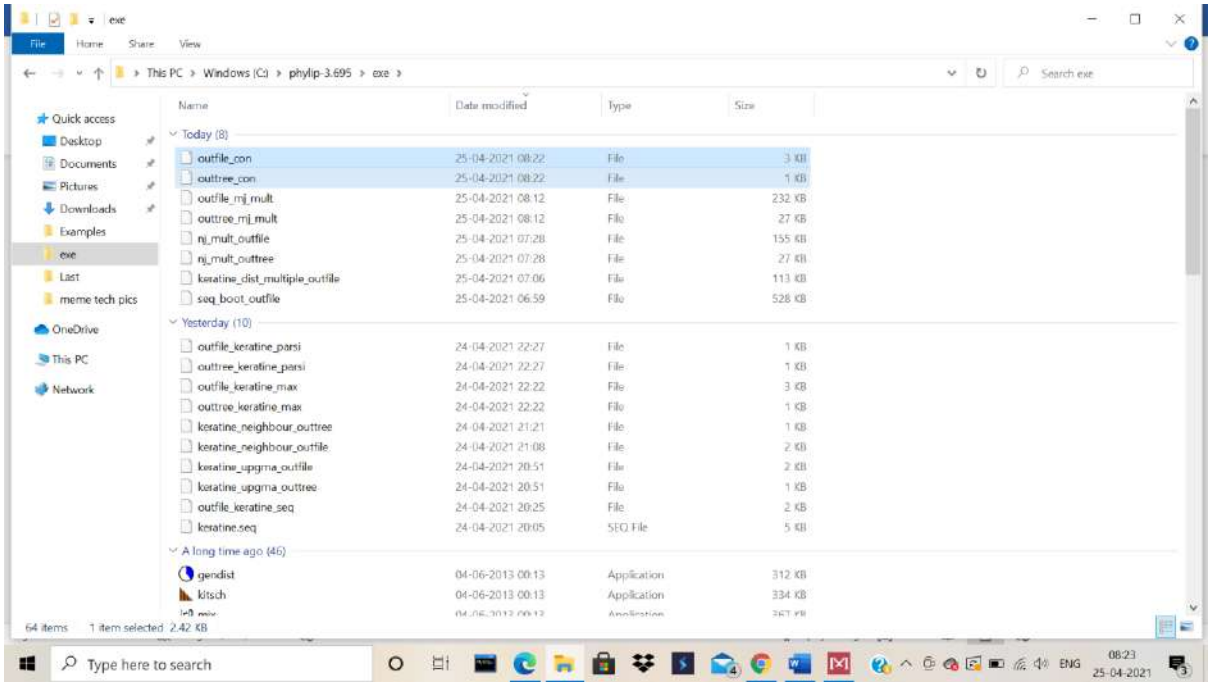
- ✓ **Performing consensus tree of NJ and ML:**

Step 1: Select a file which has atleast 100 trees.

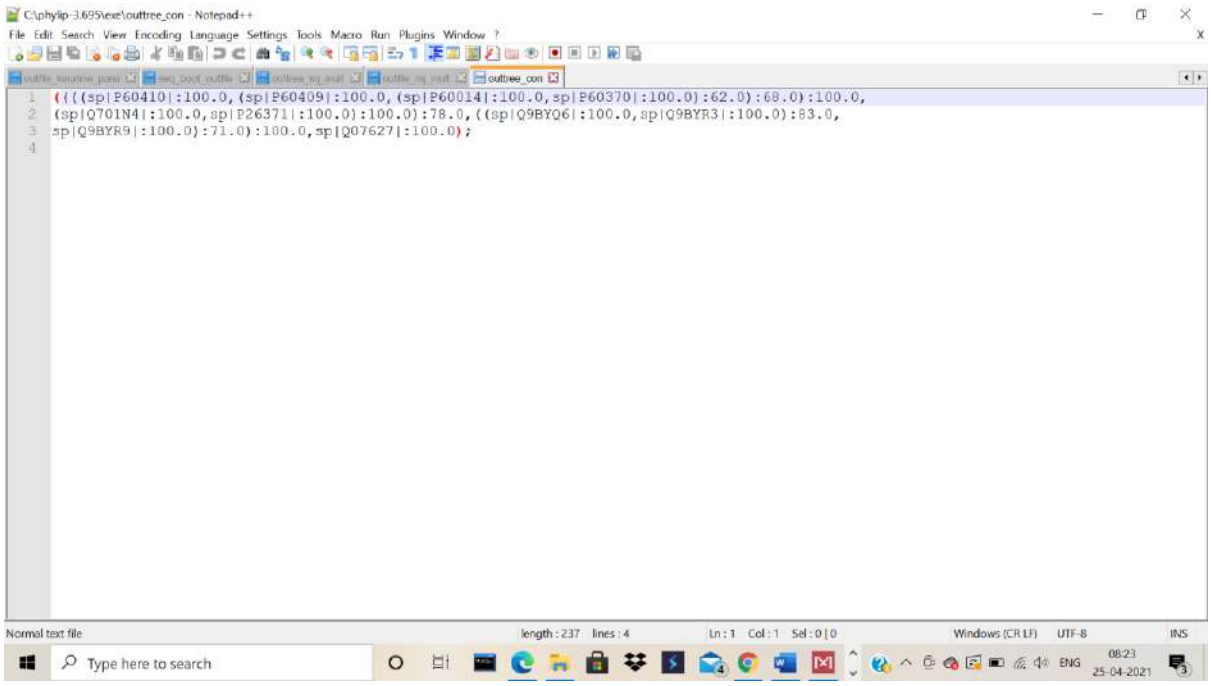




Select 2: Outfile and outtree appears as soon as you press Y and agree their terms.

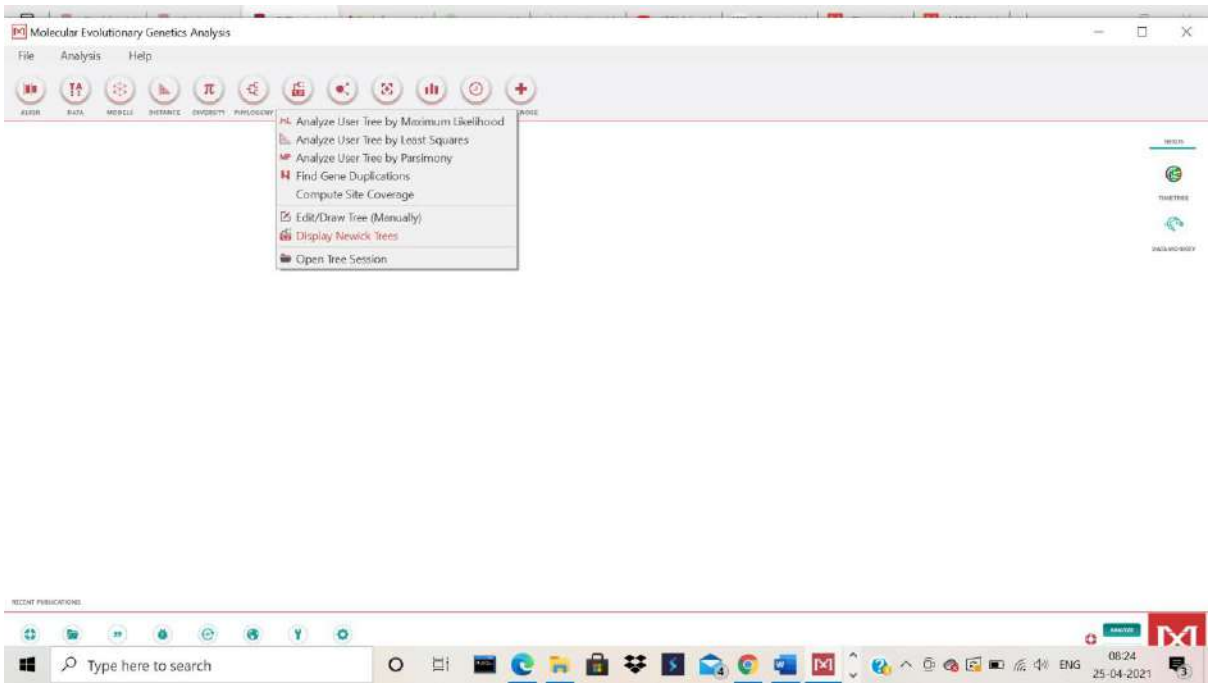


Step 3: This is the outtree for all the sequence evaluated.



This outtree is conscenses of those 100 trees.

Step 4: Open MEGA go to user tree and click on display newrick tree.



Step 5: We successfully got **consensus tree.**