HOMEWORK-2

NAME: Lipi Vikram Thakker
BIOL 5385: Molecular Evolution.
Collaboration with: Ketki Joshi & Freya Hammer

Assignment Report: Ancestral Reconstruction of Sequences

This assignment is a brief task to determine the ancestral relationship between the given sequences as a form of MSA. Here, in this assignment, two MSA file are provided under the name Prion family and Ebola family from Pfam. The tasks performed in this assignment are as follows:

- Installation of FastML on mac.
- Running the FastML for given MSA text file.
- Collect the FastML results.
- Analyze the results in MATLAB.

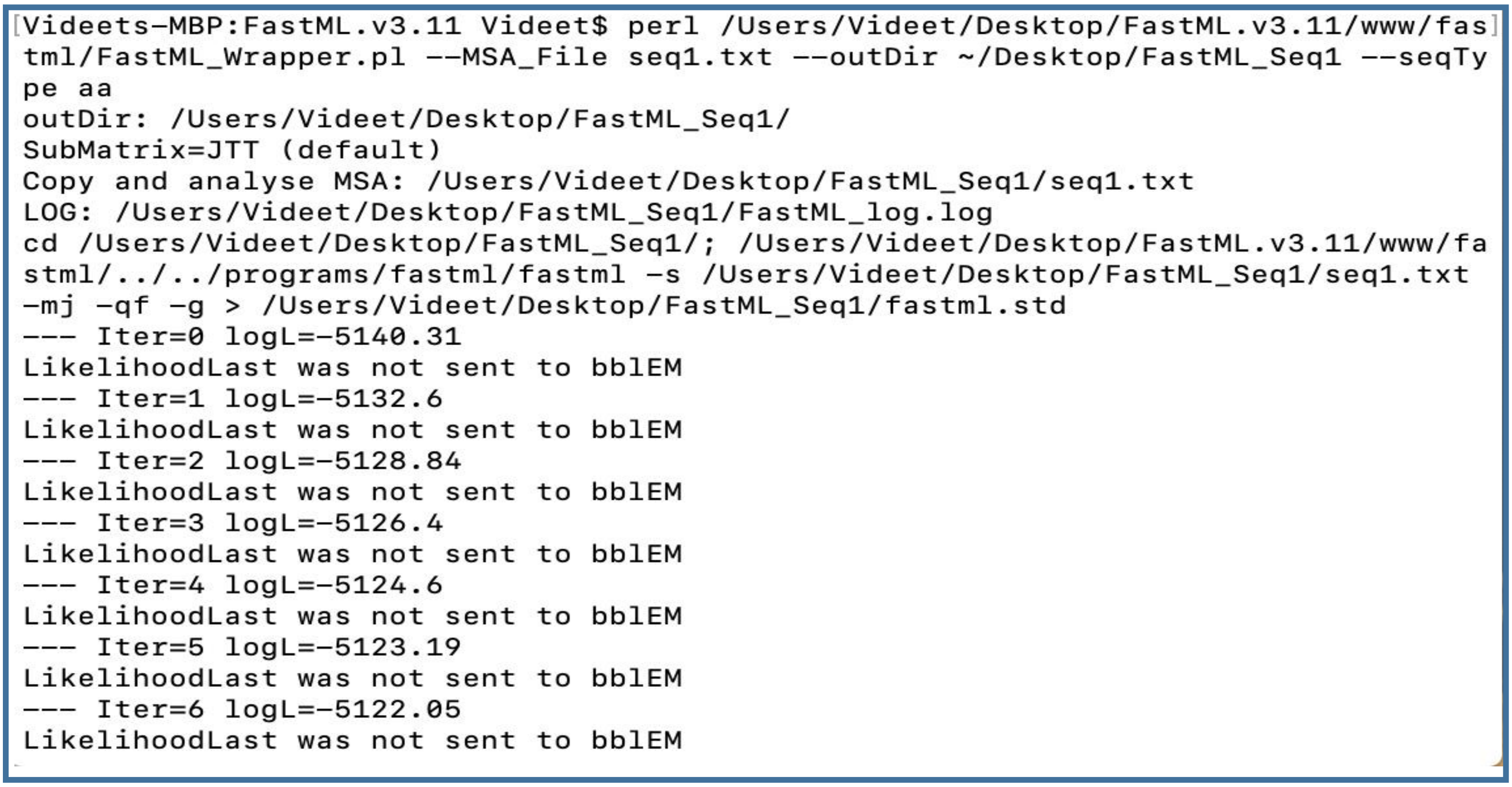
Step 1: Install FastML on Mac.

- Installation of FastML on mac was easier, as all the dependencies are found in the system beforehand.
    - Download the source code from the site.
    - Unzip the downloaded folder using the command line function tar -xzf FastML.v3.11.tgz. This will creat a directory named FastML.v3.11.
    - Enter the newly formed directory FastML.v3.11 using cd command
    - Compile using make command. This step is an installation of FastML dependencies
    - Check for perl in the system using "perl -v" command. In Mac, perl is already present.

All set with the FastML software installation. Moving to step 2.

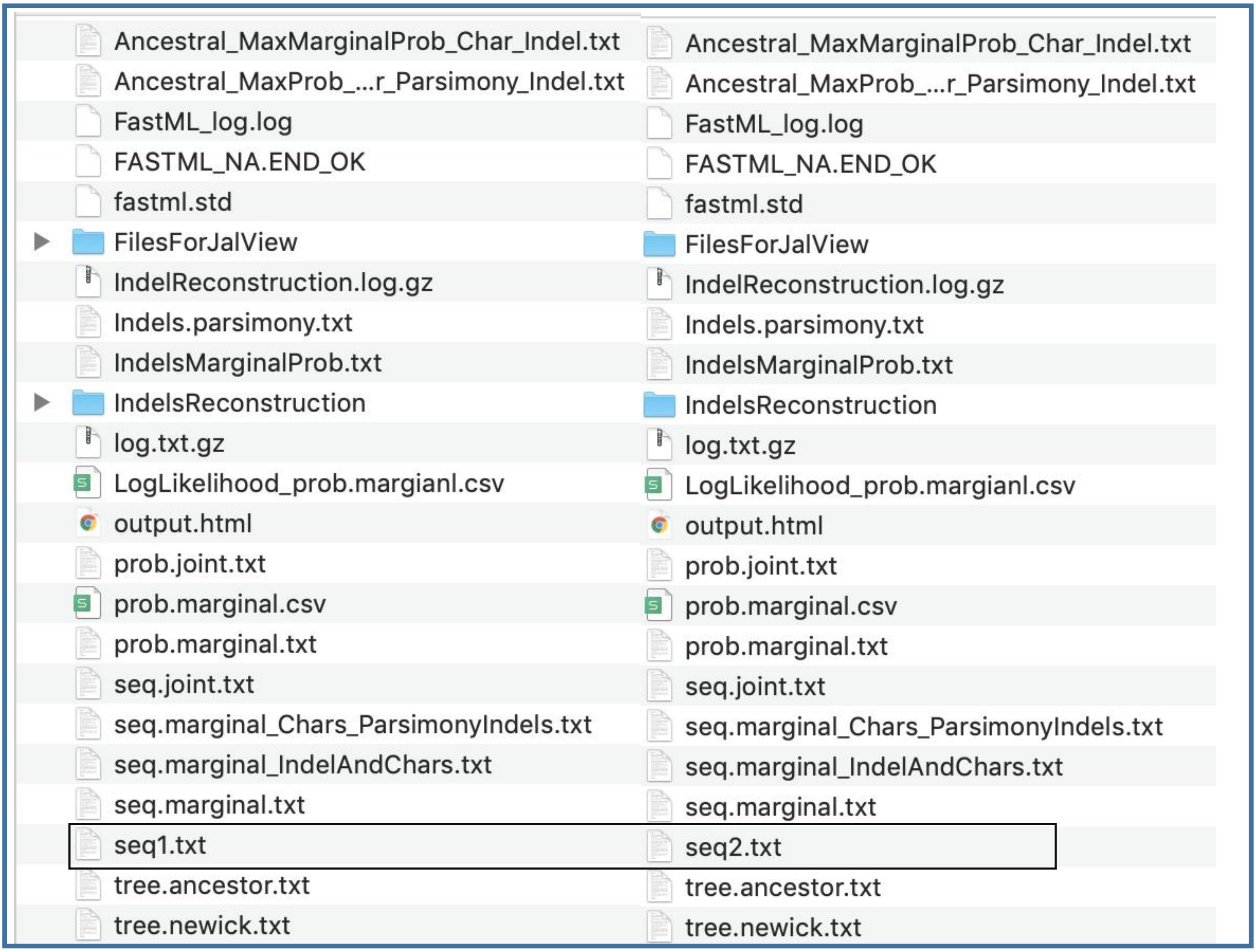
Step 2: Running FastML with the given sequences files.
Using the below command, gives the results.

- FastML.v3.11/www/fastml/FastML_Wrapper.pl --MSA_File MSA_File --seqType [AA|NUC|CODON] --outDir OUTDIR.
    - The key components in this command is,
        - path directory in the FastML.v3.11 folder.
        - The MSA file of the sequences.
        - Sequence type- amino-acid, codon, nucleotide.
        - The location of the output folder/ directory.
    - The command differs on each system, as the path of to directories is different.



```
Videets-MBP:FastML.v3.11 Videet$ perl /Users/Videet/Desktop/FastML.v3.11/www/fas
tml/FastML_Wrapper.pl --MSA_File seq1.txt --outDir ~/Desktop/FastML_Seq1 --seqTy
pe aa
outDir: /Users/Videet/Desktop/FastML_Seq1/
SubMatrix=JTT (default)
Copy and analyse MSA: /Users/Videet/Desktop/FastML_Seq1/seq1.txt
LOG: /Users/Videet/Desktop/FastML_Seq1/FastML_log.log
cd /Users/Videet/Desktop/FastML_Seq1/; /Users/Videet/Desktop/FastML.v3.11/www/fa
stml/../../programs/fastml/fastml -s /Users/Videet/Desktop/FastML_Seq1/seq1.txt
-mj -qf -g > /Users/Videet/Desktop/FastML_Seq1/fastml.std
--- Iter=0 logL=-5140.31
LikelihoodLast was not sent to bblEM
--- Iter=1 logL=-5132.6
LikelihoodLast was not sent to bblEM
--- Iter=2 logL=-5128.84
LikelihoodLast was not sent to bblEM
--- Iter=3 logL=-5126.4
LikelihoodLast was not sent to bblEM
--- Iter=4 logL=-5124.6
LikelihoodLast was not sent to bblEM
--- Iter=5 logL=-5123.19
LikelihoodLast was not sent to bblEM
--- Iter=6 logL=-5122.05
LikelihoodLast was not sent to bblEM
```

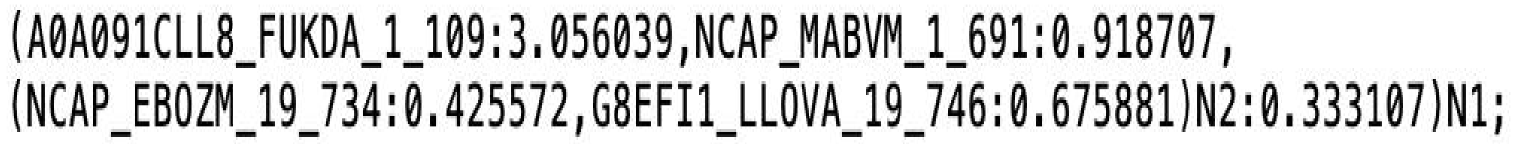
Step 3: Collect the FastML results.
The above step will create result folder named FastML_Seq1 and FastML_Seq2 on the desktop, consisting of the results for both the given MSA files.



Ancestral_MaxMarginalProb_Char_Indel.txt	Ancestral_MaxMarginalProb_Char_Indel.txt
Ancestral_MaxProb_...r_Parsimony_Indel.txt	Ancestral_MaxProb_...r_Parsimony_Indel.txt
FastML_log.log	FastML_log.log
FASTML_NA.END_OK	FASTML_NA.END_OK
fastml.std	fastml.std
▶ FilesForJalView	FilesForJalView
IndelReconstruction.log.gz	IndelReconstruction.log.gz
Indels.parsimony.txt	Indels.parsimony.txt
IndelsMarginalProb.txt	IndelsMarginalProb.txt
▶ IndelsReconstruction	IndelsReconstruction
log.txt.gz	log.txt.gz
LogLikelihood_prob.margianl.csv	LogLikelihood_prob.margianl.csv
output.html	output.html
prob.joint.txt	prob.joint.txt
prob.marginal.csv	prob.marginal.csv
prob.marginal.txt	prob.marginal.txt
seq.joint.txt	seq.joint.txt
seq.marginal_Chars_ParsimonyIndels.txt	seq.marginal_Chars_ParsimonyIndels.txt
seq.marginal_IndelAndChars.txt	seq.marginal_IndelAndChars.txt
seq.marginal.txt	seq.marginal.txt
seq1.txt	seq2.txt
tree.ancestor.txt	tree.ancestor.txt
tree.newick.txt	tree.newick.txt

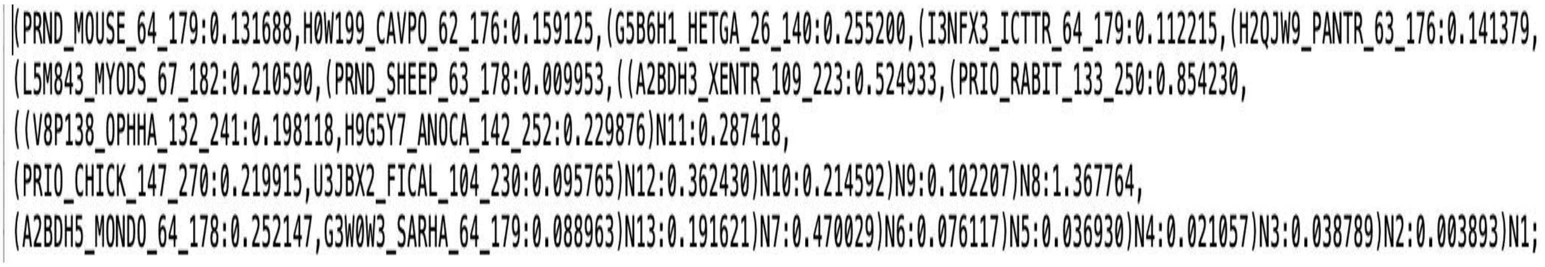
The tree results:
NEWICK TREE

For Ebola family



```
(A0A091CLL8_FUKDA_1_109:3.056039,NCAP_MABVM_1_691:0.918707,
(NCAP_EBOZM_19_734:0.425572,G8EFI1_LLOVA_19_746:0.675881)N2:0.333107)N1;
```

For Ebola family

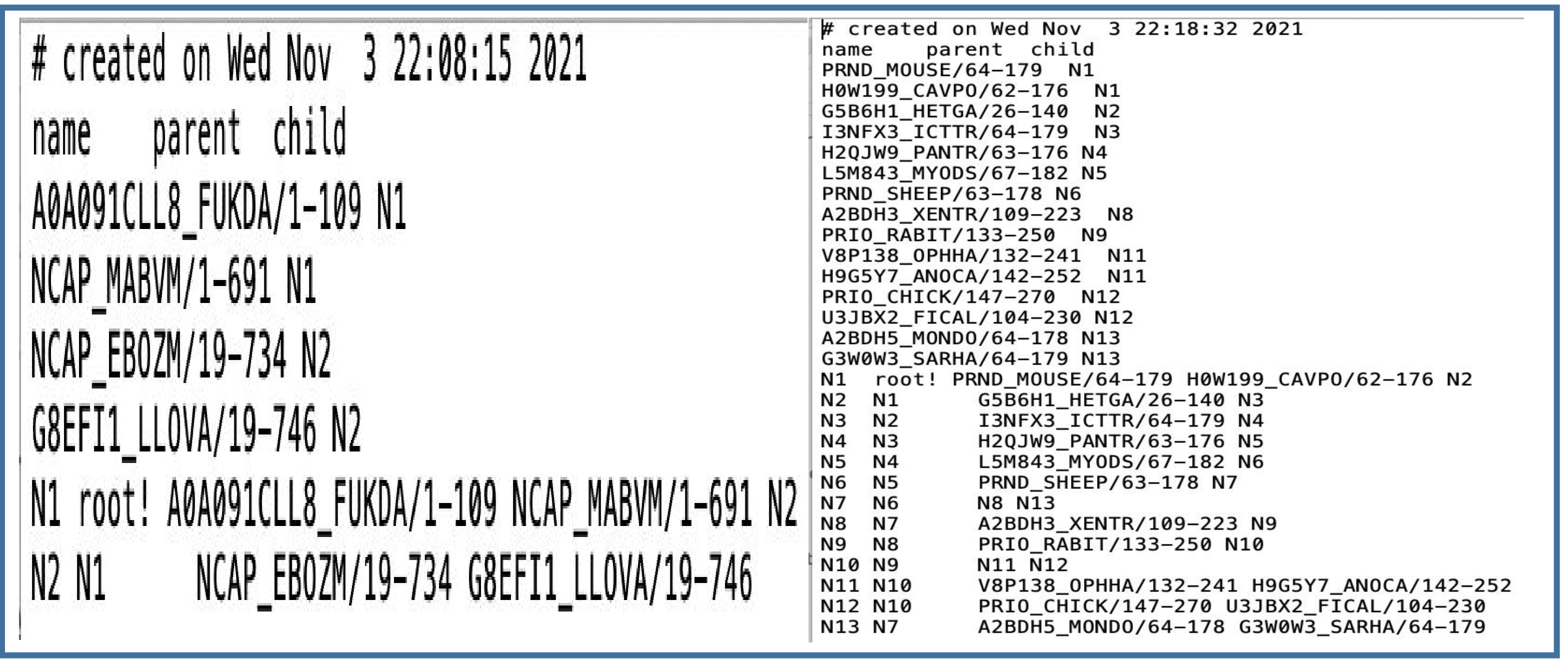


```
(PRND_MOUSE_64_179:0.131688,H0W199_CAVPO_62_176:0.159125,(G5B6H1_HETGA_26_140:0.255200,(I3NFX3_ICTTR_64_179:0.112215,(H2QJW9_PANTR_63_176:0.141379,
(L5M843_MYODS_67_182:0.210590,(PRND_SHEEP_63_178:0.009953,((A2BDH3_XENTR_109_223:0.524933,(PRIO_RABIT_133_250:0.854230,
((V8P138_OPHHA_132_241:0.198118,H9G5Y7_ANOCA_142_252:0.229876)N11:0.287418,
(PRIO_CHICK_147_270:0.219915,U3JBX2_FICAL_104_230:0.095765)N12:0.362430)N10:0.214592)N9:0.102207)N8:1.367764,
(A2BDH5_MONDO_64_178:0.252147,G3W0W3_SARHA_64_179:0.088963)N13:0.191621)N7:0.470029)N6:0.076117)N5:0.036930)N4:0.021057)N3:0.038789)N2:0.003893)N1;
```

Tree for ancestors:

For Ebola family

For Prion family



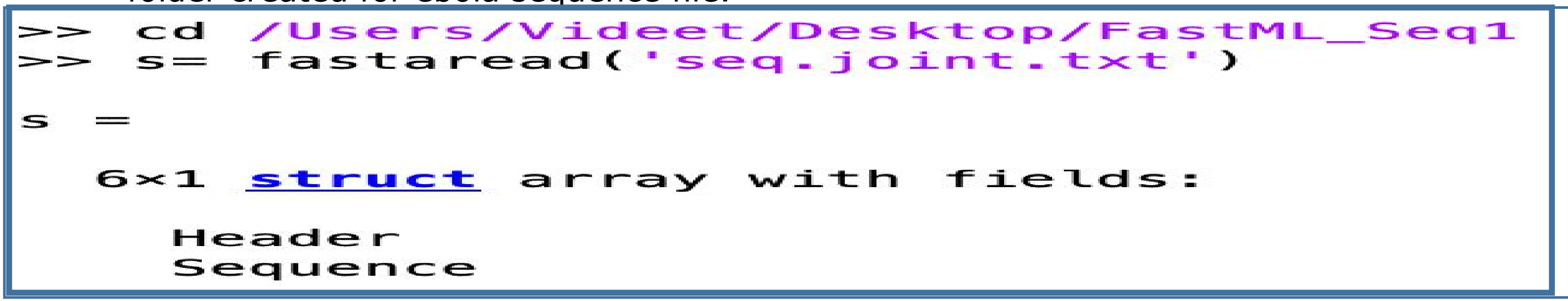
```
# created on Wed Nov  3 22:08:15 2021
name    parent  child
A0A091CLL8_FUKDA/1-109 N1
NCAP_MABVM/1-691 N1
NCAP_EBOZM/19-734 N2
G8EFI1_LLOVA/19-746 N2
N1 root! A0A091CLL8_FUKDA/1-109 NCAP_MABVM/1-691 N2
N2 N1      NCAP_EBOZM/19-734 G8EFI1_LLOVA/19-746
```

```
# created on Wed Nov  3 22:18:32 2021
name    parent  child
PRND_MOUSE/64-179  N1
H0W199_CAVPO/62-176  N1
G5B6H1_HETGA/26-140  N2
I3NFX3_ICTTR/64-179  N3
H2QJW9_PANTR/63-176 N4
L5M843_MYODS/67-182 N5
PRND_SHEEP/63-178 N6
A2BDH3_XENTR/109-223  N8
PRIO_RABIT/133-250  N9
V8P138_OPHHA/132-241  N11
H9G5Y7_ANOCA/142-252  N11
PRIO_CHICK/147-270  N12
U3JBX2_FICAL/104-230 N12
A2BDH5_MONDO/64-178 N13
G3W0W3_SARHA/64-179 N13
N1  root! PRND_MOUSE/64-179 H0W199_CAVPO/62-176 N2
N2  N1      G5B6H1_HETGA/26-140 N3
N3  N2      I3NFX3_ICTTR/64-179 N4
N4  N3      H2QJW9_PANTR/63-176 N5
N5  N4      L5M843_MYODS/67-182 N6
N6  N5      PRND_SHEEP/63-178 N7
N7  N6      N8 N13
N8  N7      A2BDH3_XENTR/109-223 N9
N9  N8      PRIO_RABIT/133-250 N10
N10 N9      N11 N12
N11 N10     V8P138_OPHHA/132-241 H9G5Y7_ANOCA/142-252
N12 N10     PRIO_CHICK/147-270 U3JBX2_FICAL/104-230
N13 N7      A2BDH5_MONDO/64-178 G3W0W3_SARHA/64-179
```

Step 4: Analyze the results in MATLAB

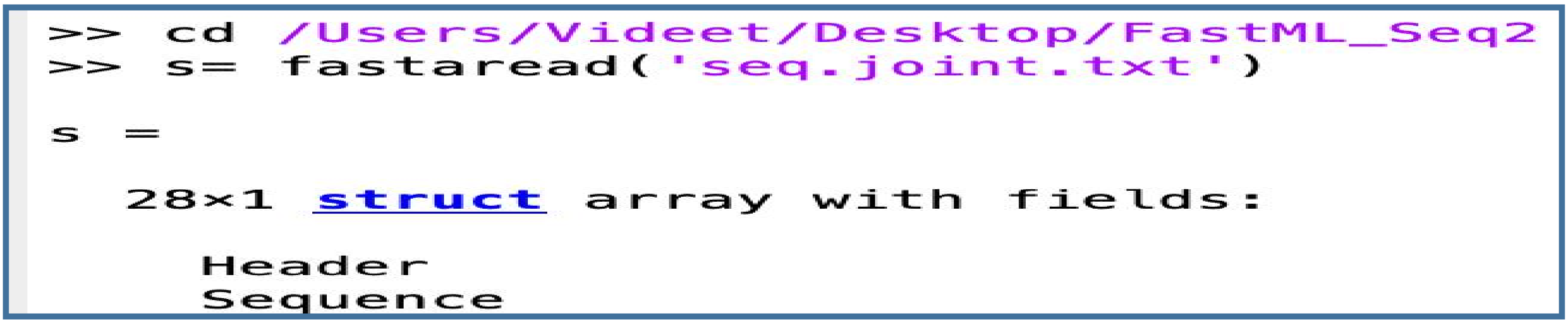
The analysis of the results is done using MATLAB

- Enter the folder created for each MSA file, namely FastML_Seq1 and FastML_Seq2.
- Using the MATLAB command line, entering each folder to retrieve trees and sequence analysis.
- Sequence analysis of the seq.joint.txt.
  - Ebola sequence file: Using the fastaread command on the seq.joint.txt, found in the FastML folder created for ebola sequence file.



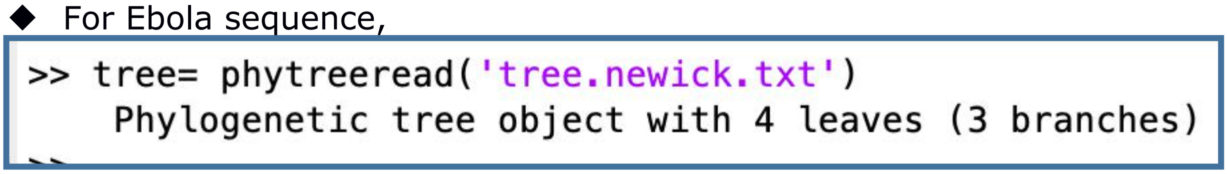
```
>> cd /Users/Videet/Desktop/FastML_Seq1
>> s= fastaread('seq.joint.txt')

s =

  6×1 struct array with fields:

    Header
    Sequence
```

  - Prion sequence file: Using the fastaread command on the seq.joint.txt, found in the FastML folder created for Prion sequence file.



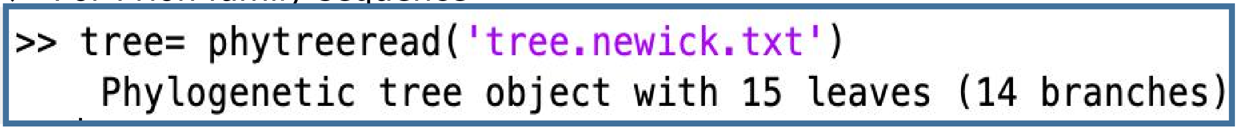
```
>> cd /Users/Videet/Desktop/FastML_Seq2
>> s= fastaread('seq.joint.txt')

s =

  28×1 struct array with fields:

    Header
    Sequence
```

- Sequence analysis of the seq.joint.txt.
  - Using the command seqalignviewer.
  - The result of this step is attached as other file with the homework submission.
- Tree structure analysis in MATLAB:
  - Using the command phytreeread, to read the tree structure.
    - For Ebola sequence,



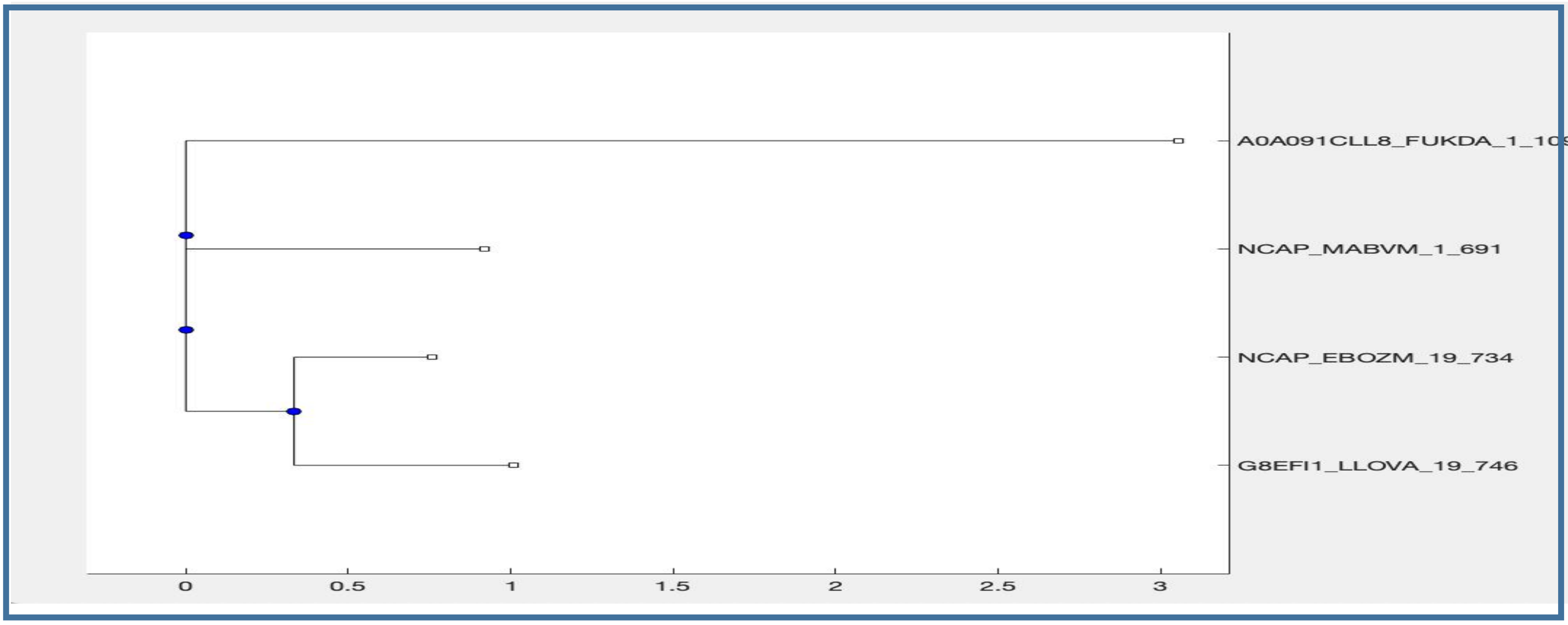
```
>> tree= phytreeread('tree.newick.txt')
    Phylogenetic tree object with 4 leaves (3 branches)
```

    - For Prion family sequence

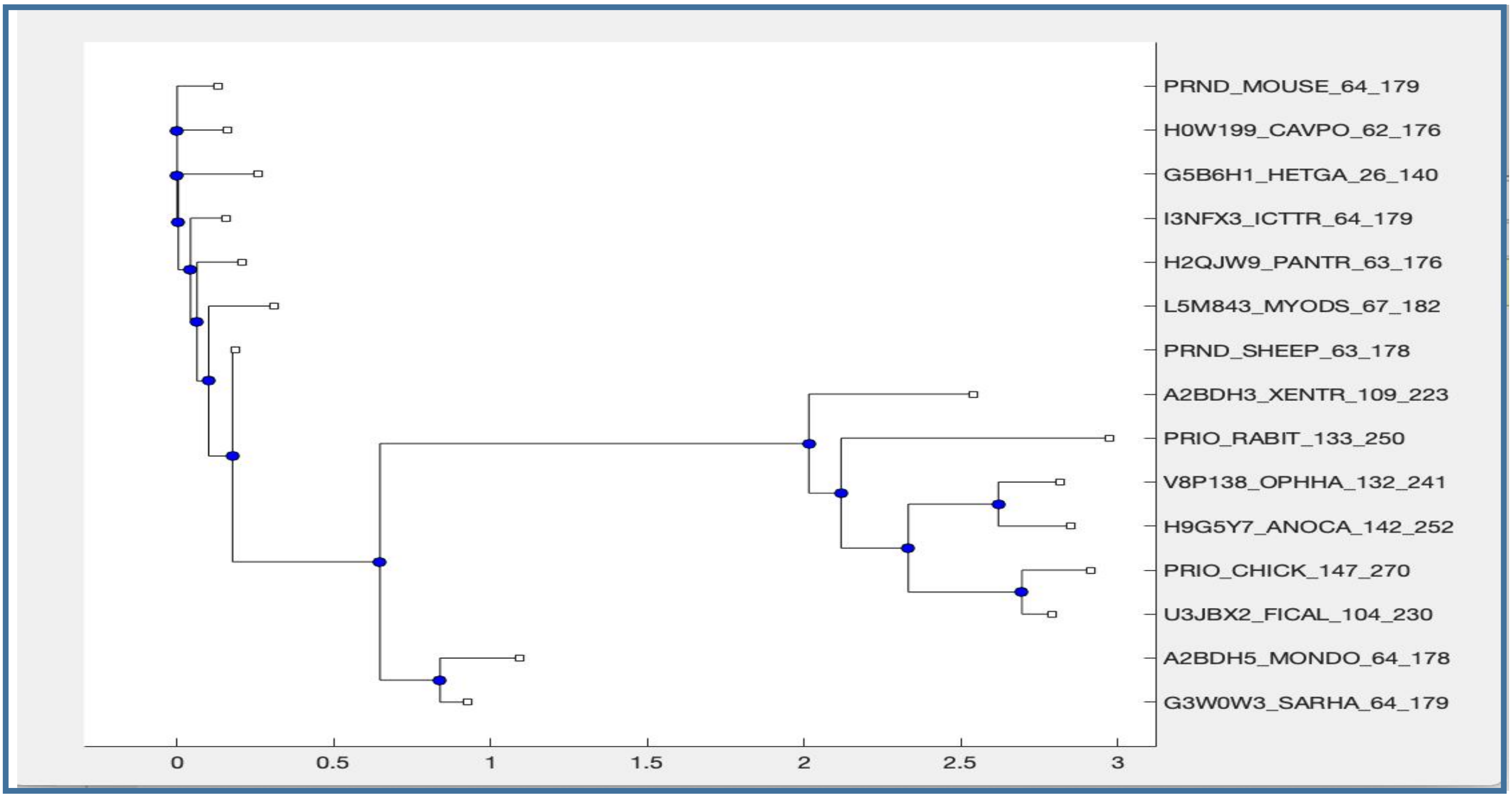


```
>> tree= phytreeread('tree.newick.txt')
    Phylogenetic tree object with 15 leaves (14 branches)
```

- Lastly, phytreeviewer, to display the tree.
  - For Ebola sequence, the tree is in the below image
  - From the below image, it is clear that N1 and N2 are the HTU's and N1 is the root(ancestor), the species are the otu.
  - The most related sequence is ncap_ebozm_19_734 and G8EFI1_LLOVA_19_746 and these two are quite distinct to a0a091cll8_fukda_1_109.



- For Prion family sequence, , the tree is in the below image.
- From the below image, it is clear that N1 to N13 are the HTU's and N1 is the root(ancestor), the species are the otu.
- The tree is quite vast, so some of the most related sequence is PRIO_CHICK_147_270 and U3JBX2_FICAL_104_230 and these two are quite distinct to PRND_MOUSE_64_179.


PRND_MOUSE_64_179
H0W199_CAVPO_62_176
G5B6H1_HETGA_26_140
I3NFX3_ICTTR_64_179
H2QJW9_PANTR_63_176
L5M843_MYODS_67_182
PRND_SHEEP_63_178
A2BDH3_XENTR_109_223
PRIO_RABIT_133_250
V8P138_OPHHA_132_241
H9G5Y7_ANOCA_142_252
PRIO_CHICK_147_270
U3JBX2_FICAL_104_230
A2BDH5_MONDO_64_178
G3W0W3_SARHA_64_179
0
0.5
1
1.5
2
2.5
3