BT3040 – BIOINFORMATICS – Assignment 9

Submitted by Sahana (BE17B038)

Question 1

Algorithm –

(i) Multiple sequence alignment using MAFFT for the given sequences in Set 1 and 2 separately.
   a. MSA was done in Automode and not G-iNSI.

(ii) Reformat and download the data (MSA) in Phylip format

(iii) Bootstrapping is done using **Seqboot** program.

(iv) Maximum likelihood method is done using **proml** program

(v) Data for Consensus tree is got using **Consense** program

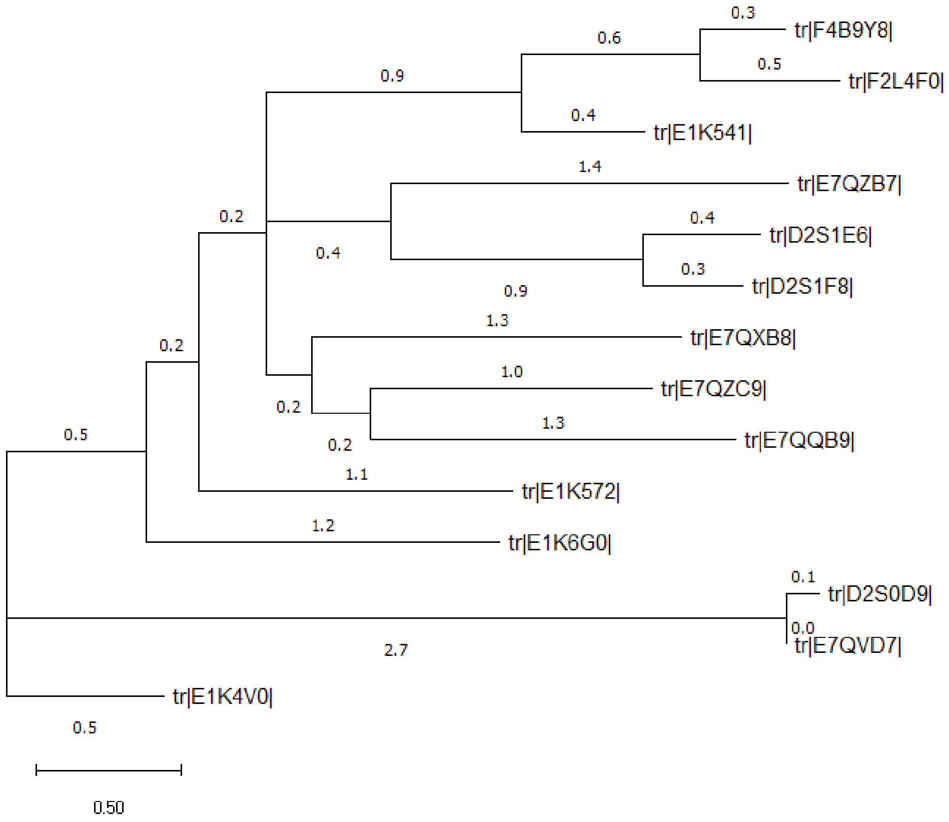
(vi) Use Treeview/MegaX to view the tree. Save images of the obtained phylogenic tree.

(vii) NJ and UPGMA methods are obtained using **protdist** and **neighbor** programs

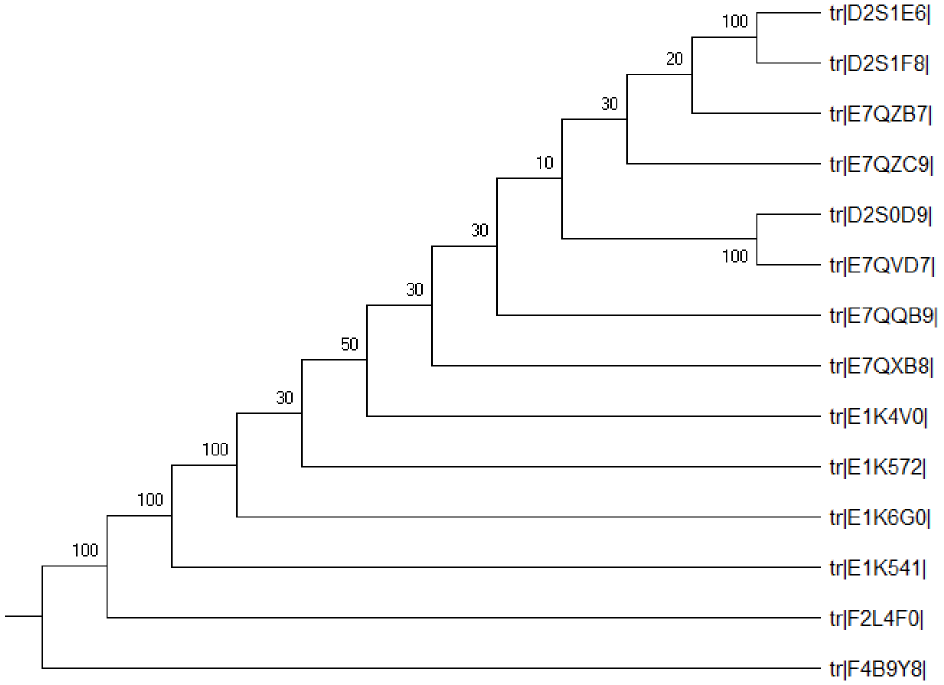
The results are shown as follows. Only the phylogenetic trees are attached.

Set 1 – tim.dat

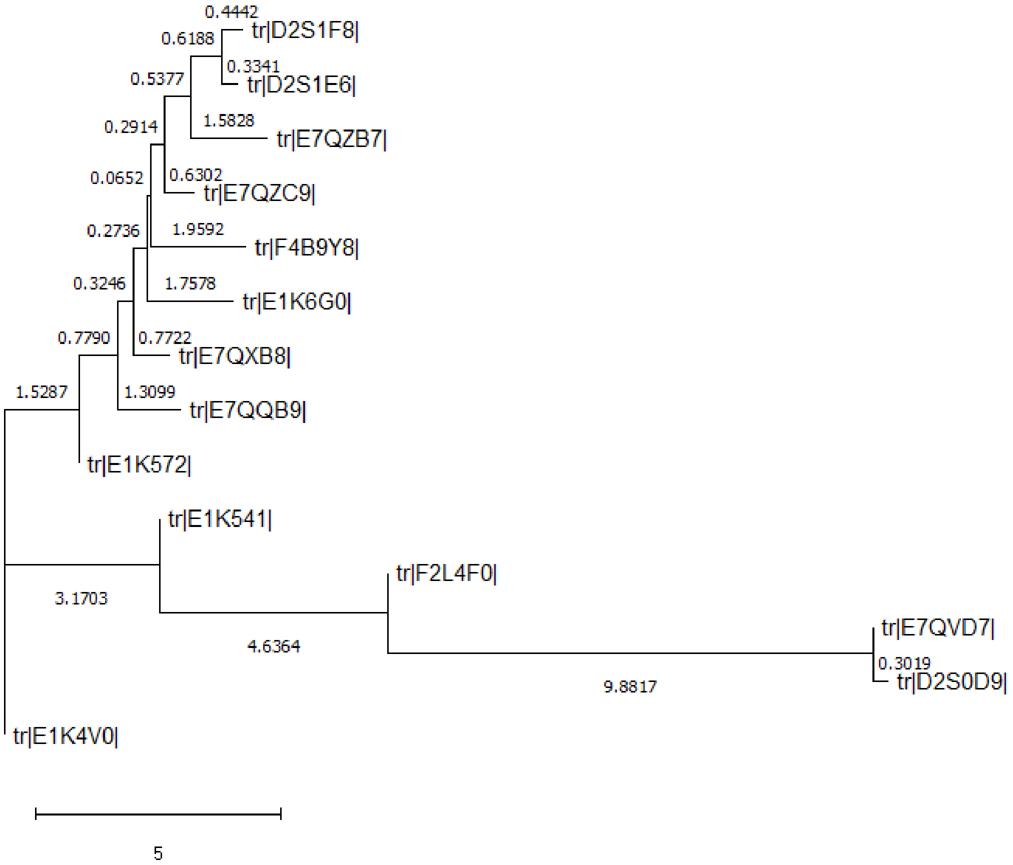
A. Output Ph.tree based on MSA and bootstrapping



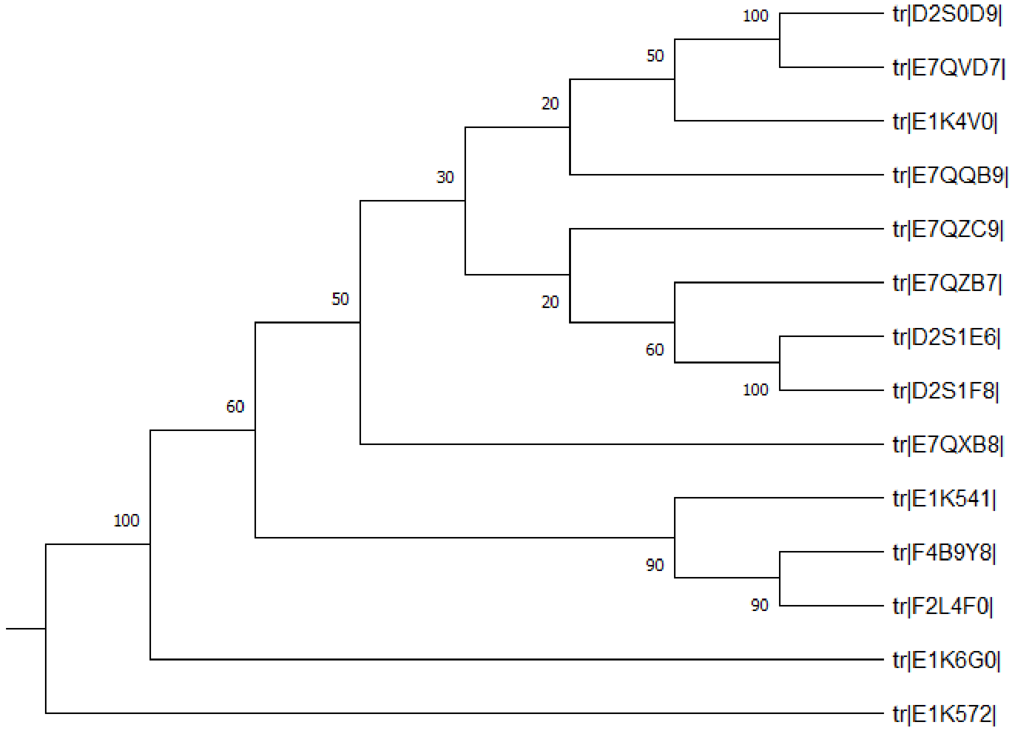
B. Consensus tree based on Maximum Likelihood



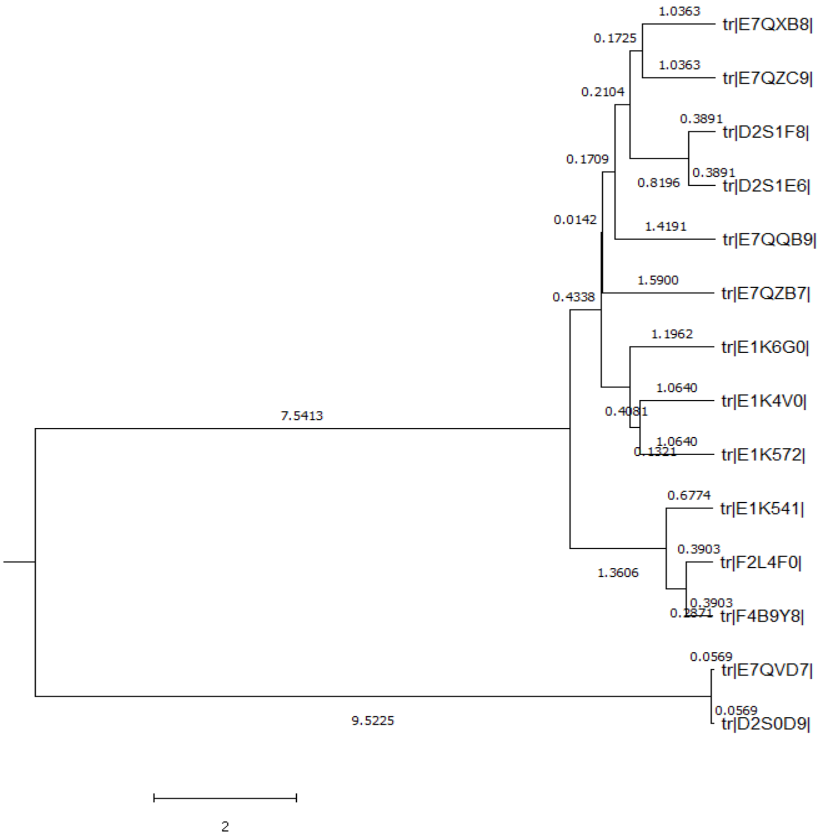
C. Output Ph.tree based on Neighbor joining



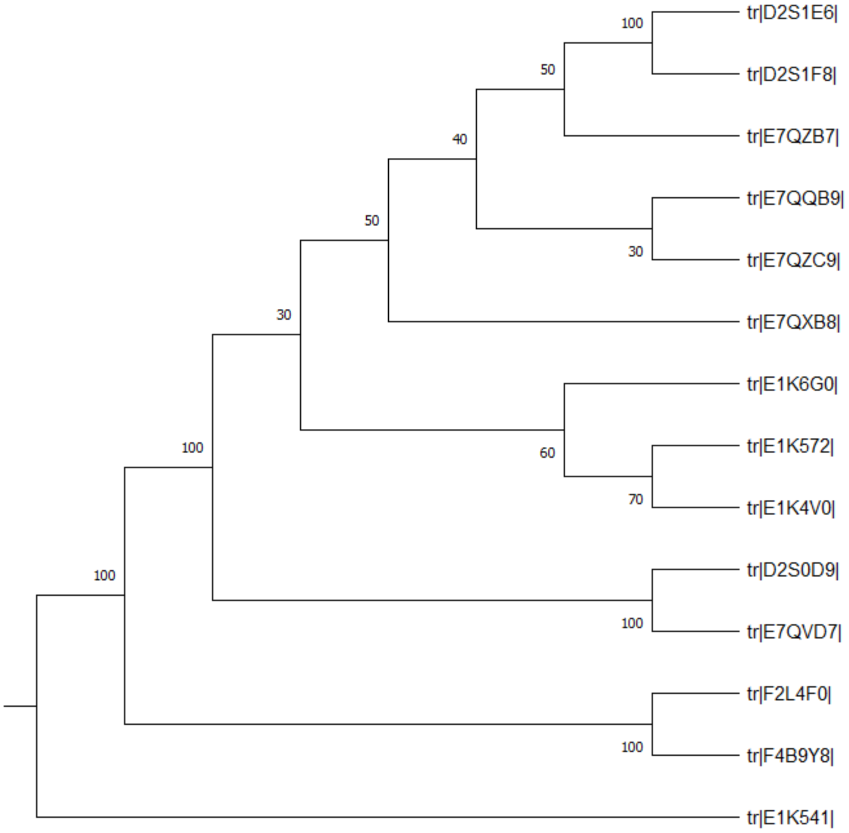
D. Consensus tree based on Neighbor joining



E. Output Ph.tree based on Neighbor joining

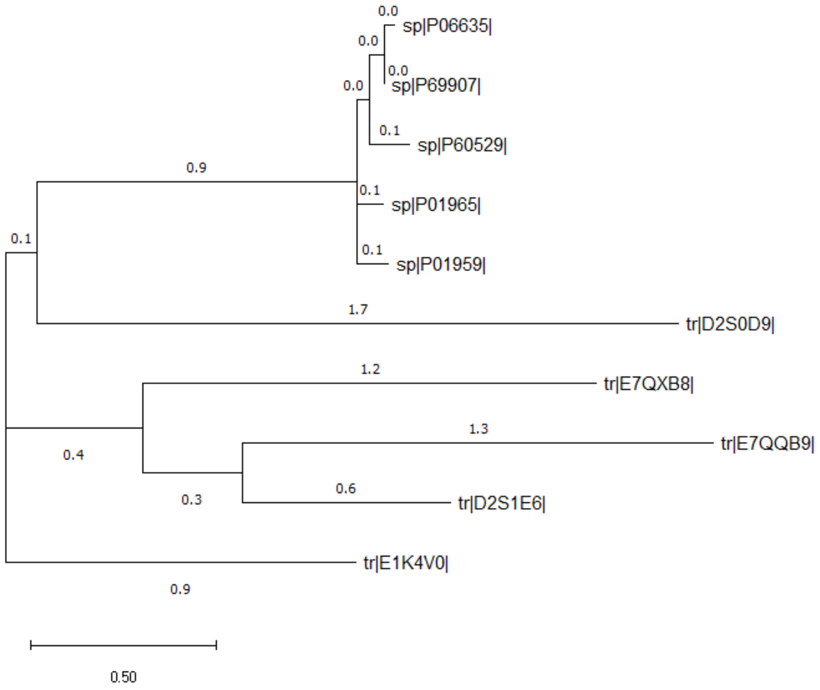


F. Consensus tree based on Neighbor joining

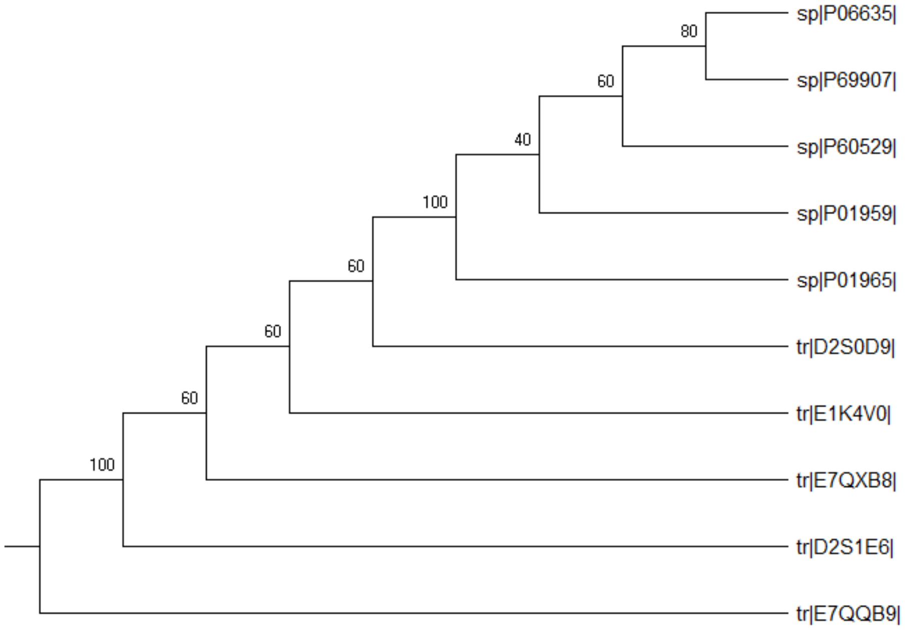


Set 2 – tim-hemo.dat

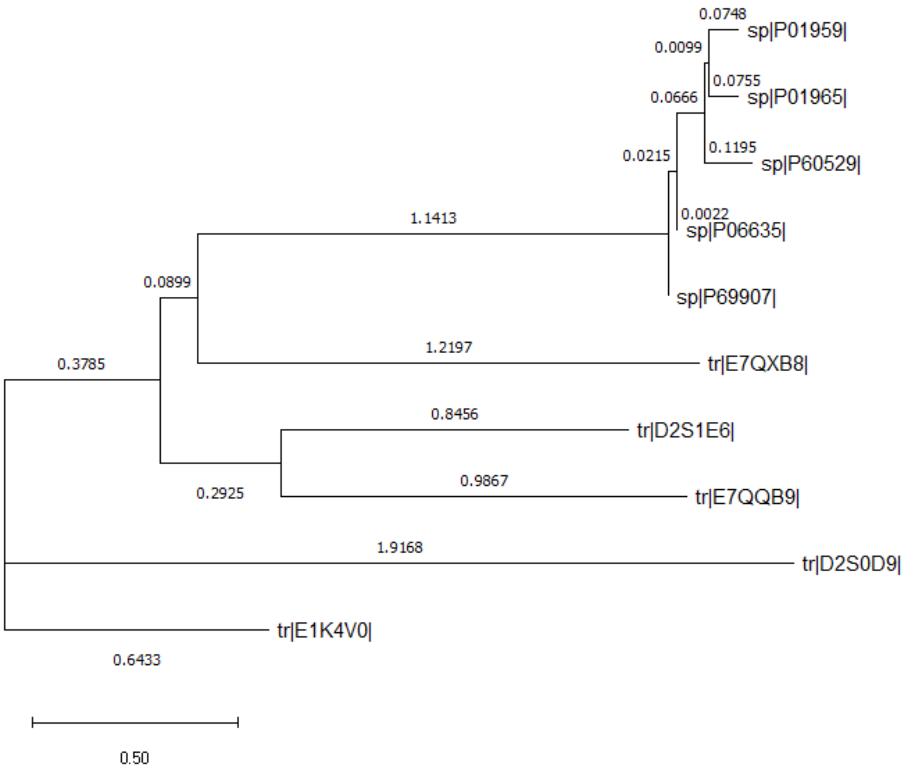
A. Output Ph.tree based on MSA and bootstrapping



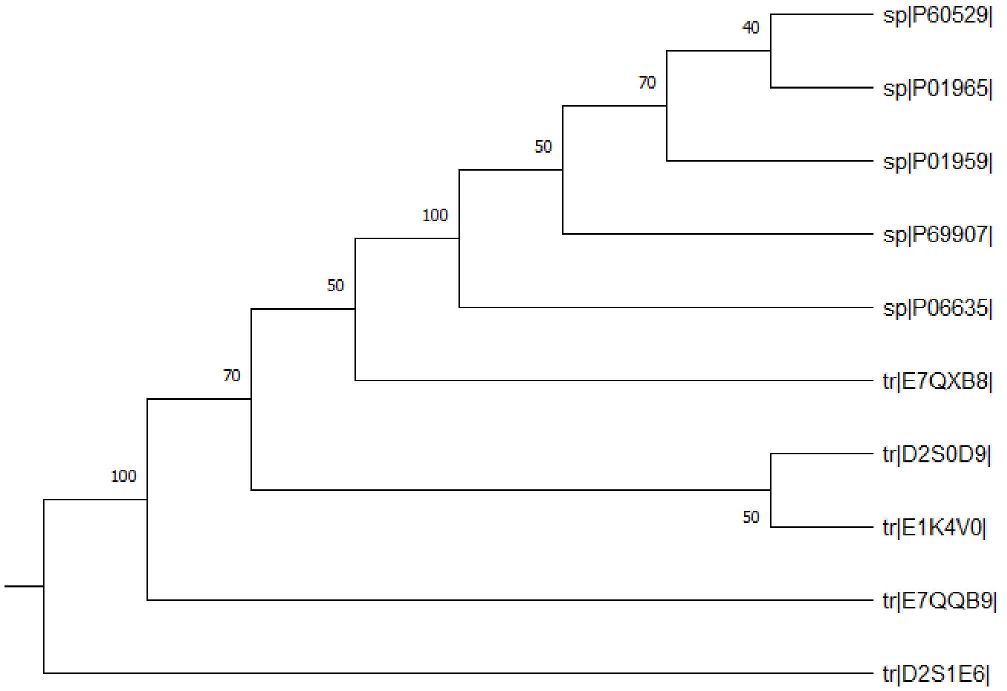
B. Consensus tree based on Maximum Likelihood



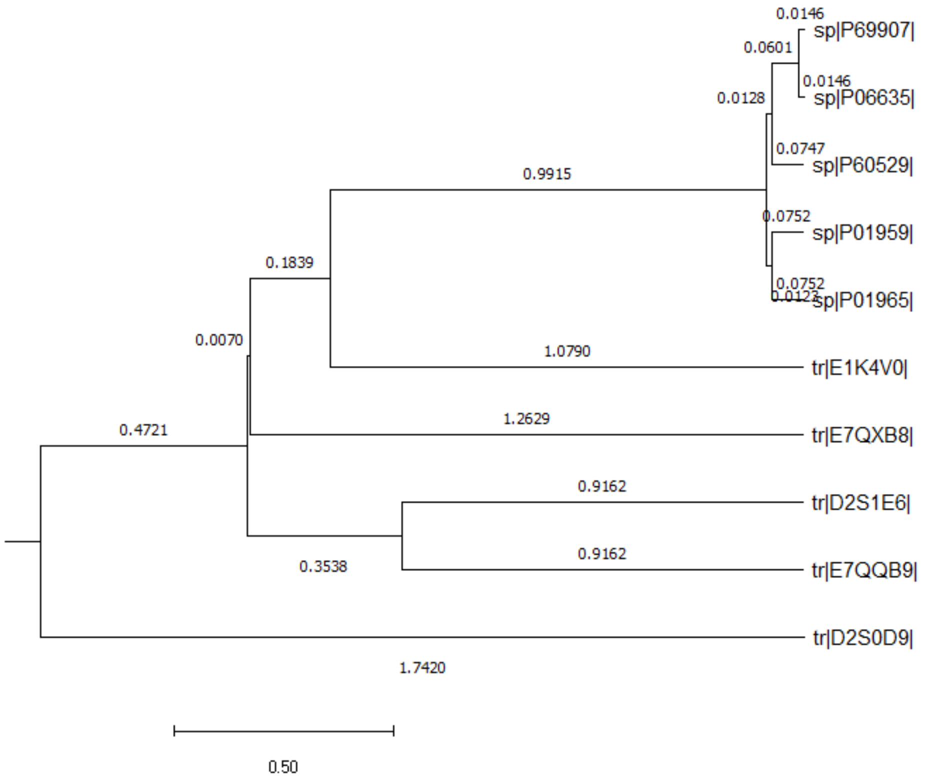
C. Output Ph.tree based on Neighbor joining



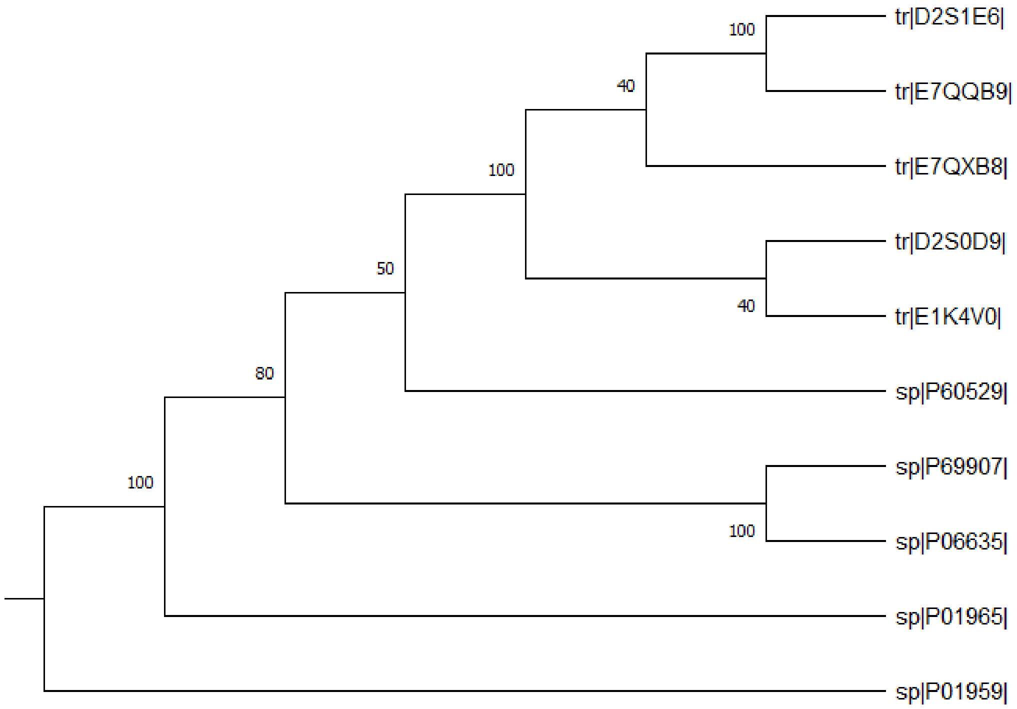
D. Consensus tree based on Neighbor joining



E. Output Ph.tree based on UPGMA



F. Consensus tree based on UPGMA



Question 2

Sequences -

```
MVLSPADKTNVKAAWGKVGAHAGEYGA
MKRLPADPPCVKTTWGKVKAKAGDYGA
MALSAADKTNVKATSSKVGGHAGEYGA
MVLSAADKTNVKAAWSKAGGNAGEWWA
MVLSAADKTNVKAAWSKVLANAGEFGA
ALLPIRTTYHKKNNVCASGHIPEEKDL
DEASSLKGHHIKASSKLEADALLIPLS
```

Algorithm –

- First construct an alignment matrix based on the positional occurrence of AAs in the given sequences.
- Each element in the weight matrix is computed as

Weight_matrix(i,j) = ln[(Nij + p)/(p)*(N+1)]

Nij = value at position (i,j) in the alignment matrix
p = probability of an AA in that position = 1/20
N = total number of sequences

Code –

```
import math as m
import pandas
import numpy as np

def PSSM(seq):
    N = len(seq)
    l = len(seq[0])
    align_m = [[0 for i in range(l)] for i in range(20)]
    weight_m = [[0.000 for i in range(l)] for i in range(20)]
    AA_all
=['A','C','D','E','F','G','H','I','K','L','M','N','P','Q','R','S','T','V','W','Y'
]
    position = [str(i+1) for i in range(27)]

    p = 1/20

    for i in range(N):
        for j in range(l):
            aa = seq[i][j]
            ind = AA_all.index(aa)
            align_m[ind][j]+=1
    data1 = np.array(align_m)

    print('The alignment matrix for the given sequences where rows represent AA
and columns are positional occurances - ')
    print(pandas.DataFrame(data1, AA_all, position))

    for i in range(20):
        for j in range(l):
            c = align_m[i][j] + p
            d = (p)*(N + 1)
            e = c/d
            weight_m[i][j] = float('%.3f'%(m.log(e)))
    data2 = np.array(weight_m)

    print('The weight matrix for the given sequences where rows represent AA and
columns are positional occurances - ')
    print(pandas.DataFrame(data2, AA_all, position))

s = ['MVLSPADKTNVKAAWGKVGAHAGEYGA', 'MKRLPADPPCVKTTWGKVKAKAGDYGA',
'MALSAADKTNVKATSSKVGGHAGEYGA', 'MVLSAADKTNVKAAWSKAGGNAGEWWA',
'MVLSAADKTNVKAAWSKVLANAGEFGA', 'ALLPIRTTYHKKNNVCASGHIPEEKDL',
'DEASSLKGHHIKASSKLEADALLIPLS']

PSSM(s)
```

Output –

Position	1	2	3	4	5	6	7	8	9	10
A	0.965	0.965	0.965	-2.079	2.031	2.536	-2.079	-2.079	-2.079	-2.079
C	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	0.965
D	0.965	-2.079	-2.079	-2.079	-2.079	-2.079	2.536	-2.079	-2.079	-2.079
E	-2.079	0.965	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079
F	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079
G	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	0.965	-2.079	-2.079
H	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	0.965	1.634
I	-2.079	-2.079	-2.079	-2.079	0.965	-2.079	-2.079	-2.079	-2.079	-2.079
K	-2.079	0.965	-2.079	-2.079	-2.079	-2.079	0.965	2.315	-2.079	-2.079
L	-2.079	0.965	2.536	0.965	-2.079	0.965	-2.079	-2.079	-2.079	-2.079
M	2.536	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079
N	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	2.315

P	-2.079	-2.079	-2.079	0.965	1.634	-2.079	-2.079	0.965	0.965	-2.079
Q	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079
R	-2.079	-2.079	0.965	-2.079	-2.079	0.965	-2.079	-2.079	-2.079	-2.079
S	-2.079	-2.079	-2.079	2.536	0.965	-2.079	-2.079	-2.079	-2.079	-2.079
T	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	0.965	0.965	2.315	-2.079
V	-2.079	2.031	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079
W	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079
Y	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	0.965	-2.079

Position	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
A	-2.079	-2.079	2.536	2.031	-2.079	-2.079	0.965	0.965	0.965	2.031
C	-2.079	-2.079	-2.079	-2.079	-2.079	0.965	-2.079	-2.079	-2.079	-2.079
D	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	0.965
E	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	0.965	-2.079	-2.079
F	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079
G	-2.079	-2.079	-2.079	-2.079	-2.079	1.634	-2.079	-2.079	2.315	1.634
H	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	0.965
I	0.965	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079
K	0.965	2.869	-2.079	-2.079	-2.079	0.965	2.536	-2.079	0.965	-2.079
L	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	0.965	-2.079	0.965	-2.079
M	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079
N	-2.079	-2.079	0.965	0.965	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079
P	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079
Q	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079
R	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079
S	-2.079	-2.079	-2.079	0.965	1.634	2.031	-2.079	0.965	-2.079	-2.079
T	-2.079	-2.079	0.965	1.634	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079
V	2.536	-2.079	-2.079	-2.079	0.965	-2.079	-2.079	2.315	-2.079	-2.079
W	-2.079	-2.079	-2.079	-2.079	2.315	-2.079	-2.079	-2.079	-2.079	-2.079
Y	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079

Position	**21**	**22**	**23**	**24**	**25**	**26**	**27**
A	0.965	2.536	-2.079	-2.079	-2.079	-2.079	2.536
C	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079
D	-2.079	-2.079	-2.079	0.965	-2.079	0.965	-2.079
E	-2.079	-2.079	0.965	2.536	-2.079	-2.079	-2.079
F	-2.079	-2.079	-2.079	-2.079	0.965	-2.079	-2.079
G	-2.079	-2.079	2.536	-2.079	-2.079	2.315	-2.079
H	1.634	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079
I	0.965	-2.079	-2.079	0.965	-2.079	-2.079	-2.079
K	0.965	-2.079	-2.079	-2.079	0.965	-2.079	-2.079
L	-2.079	0.965	0.965	-2.079	-2.079	0.965	0.965
M	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079
N	1.634	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079

P	-2.079	0.965	-2.079	-2.079	0.965	-2.079	-2.079
Q	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079
R	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079
S	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	0.965
T	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079
V	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079	-2.079
W	-2.079	-2.079	-2.079	-2.079	0.965	0.965	-2.079
Y	-2.079	-2.079	-2.079	-2.079	2.031	-2.079	-2.079