BT 3040: BIOINFORMATICS

Assignment 10

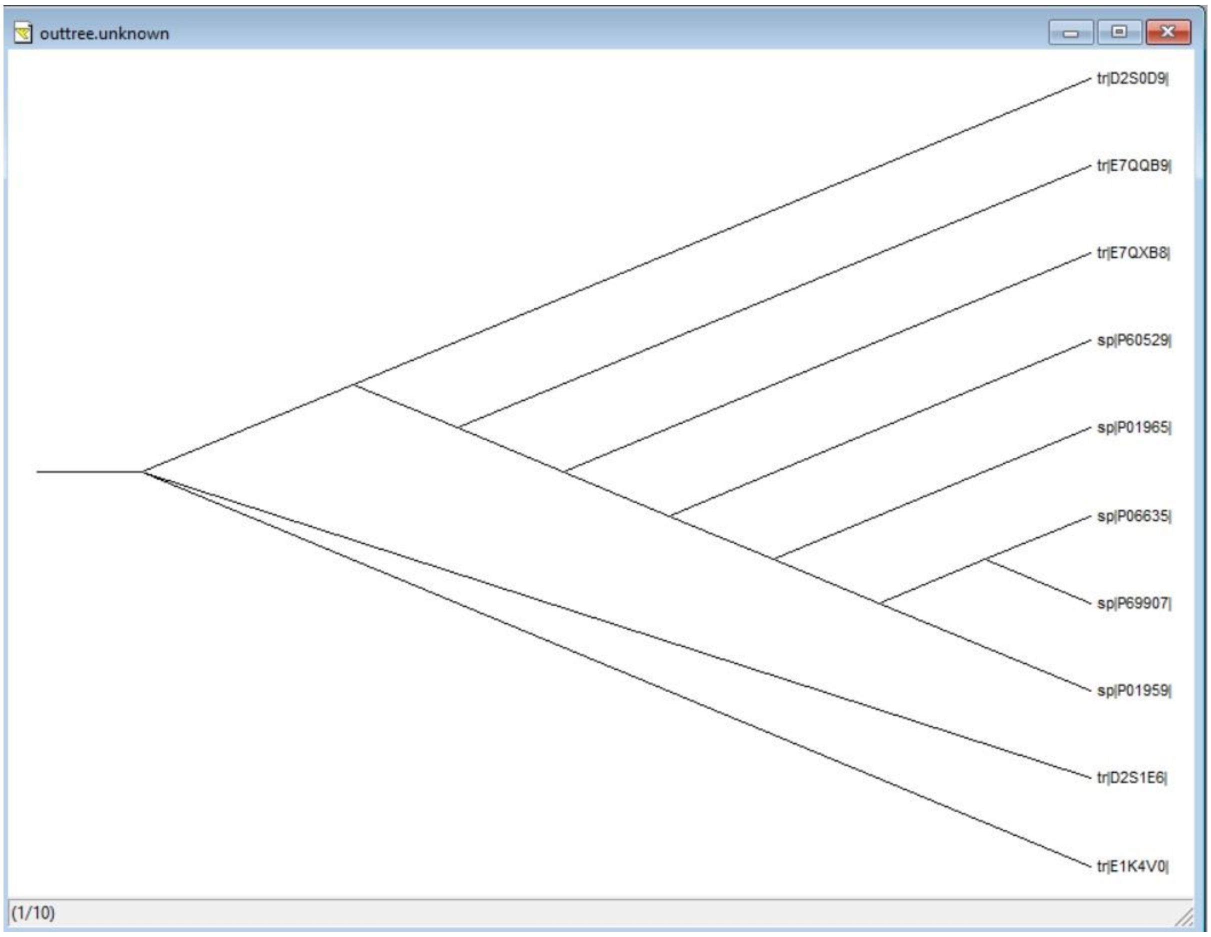
Atharva Mandar Phatak | BE21B009
Department of Biotechnology

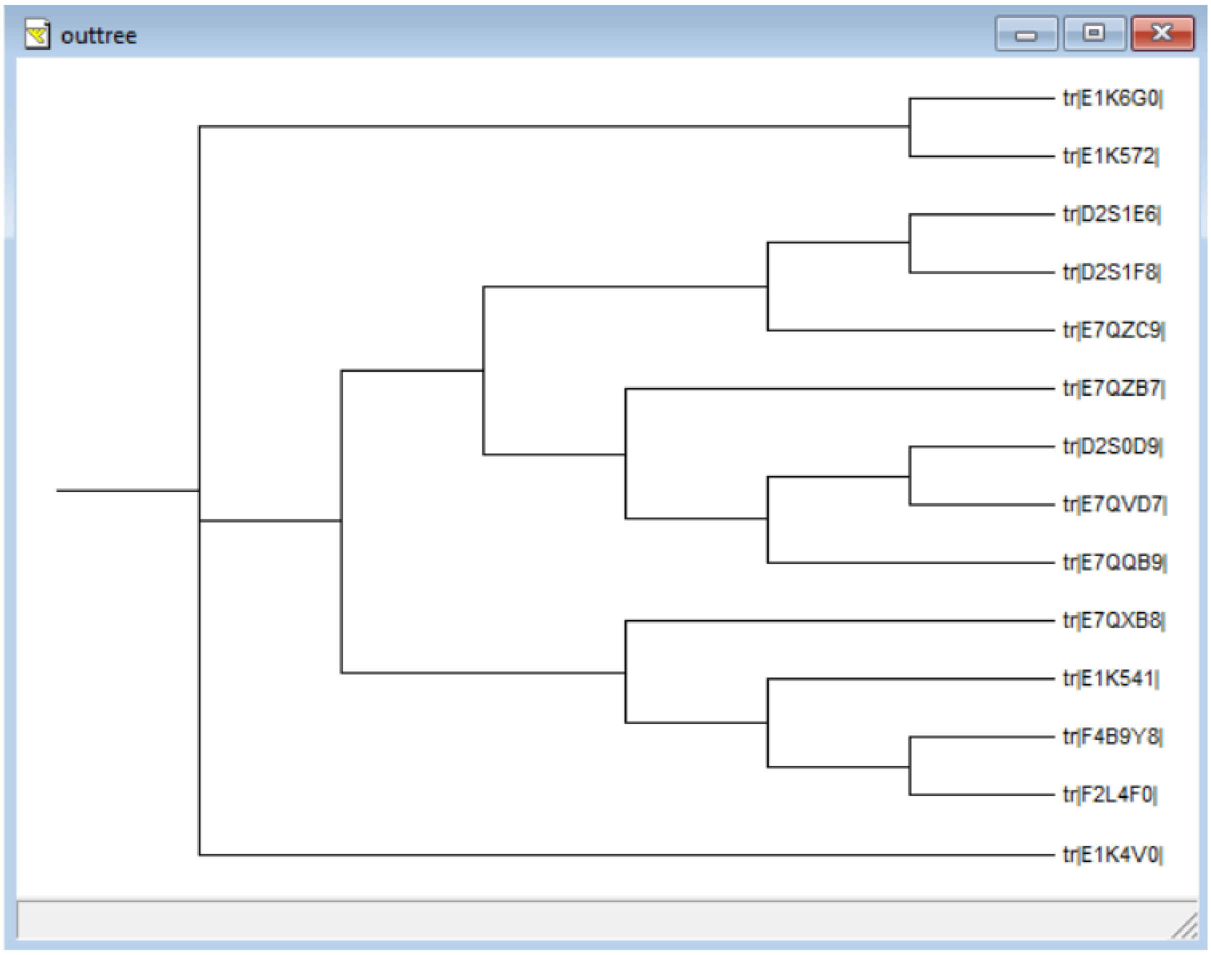
Indian Institute of Technology
Madras

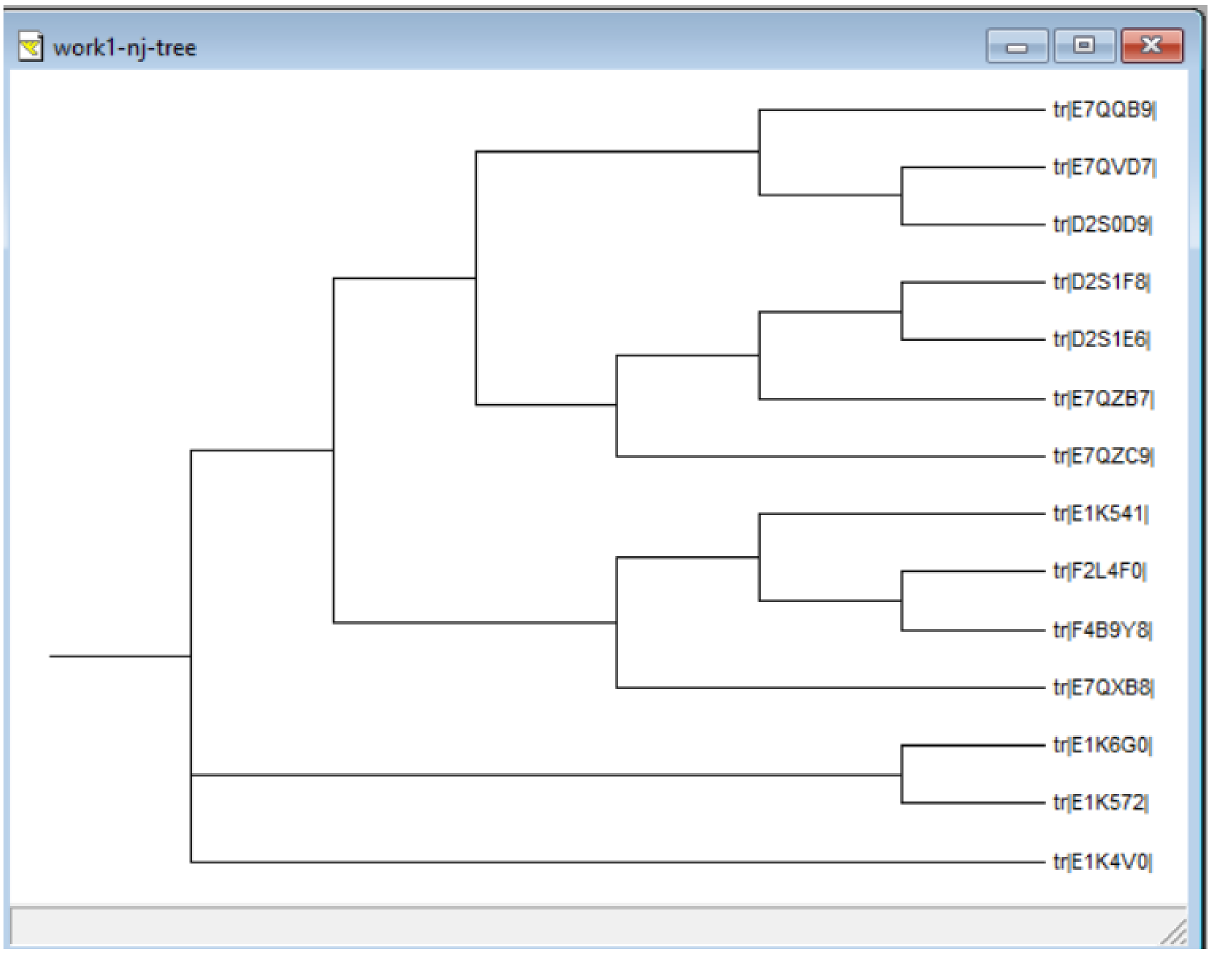
Q1) Obtain the consensus phylogenetic tree for the following two sets of sequences:

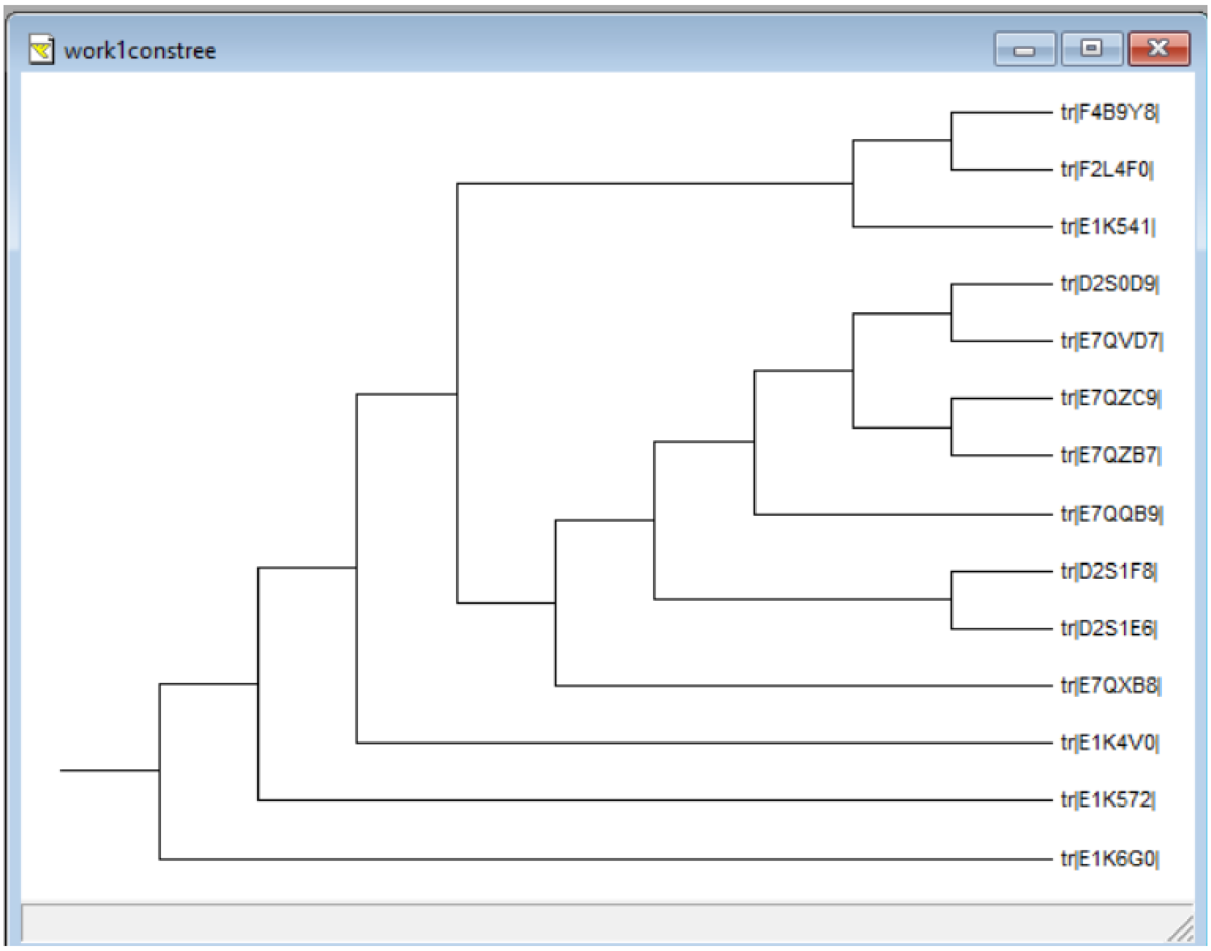
Set 1: tim.dat

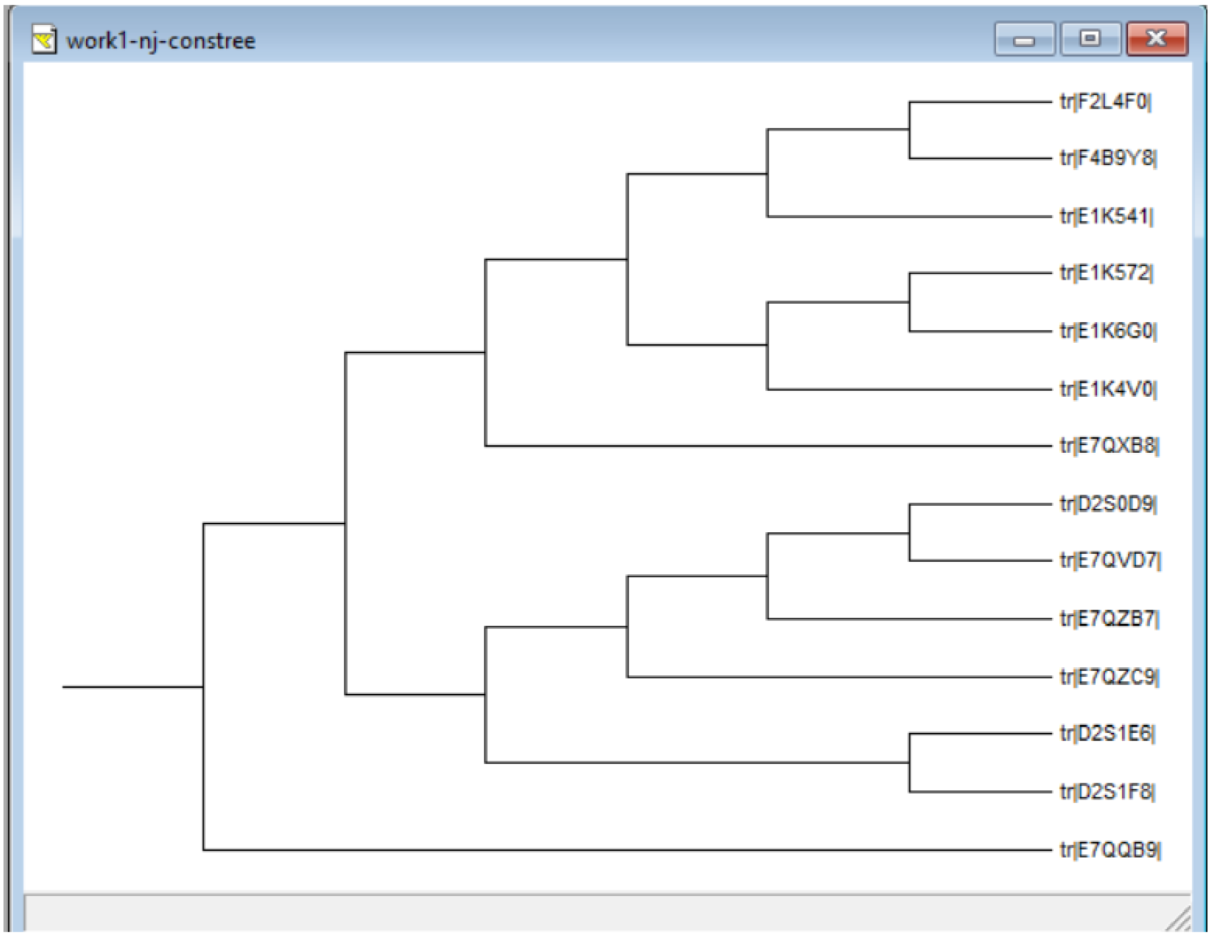
All the files in 'tim_dat' folder




outtree
tr|E1K6G0|
tr|E1K572|
tr|D2S1E6|
tr|D2S1F8|
tr|E7QZC9|
tr|E7QZB7|
tr|D2S0D9|
tr|E7QVD7|
tr|E7QQB9|
tr|E7QXB8|
tr|E1K541|
tr|F4B9Y8|
tr|F2L4F0|
tr|E1K4V0|

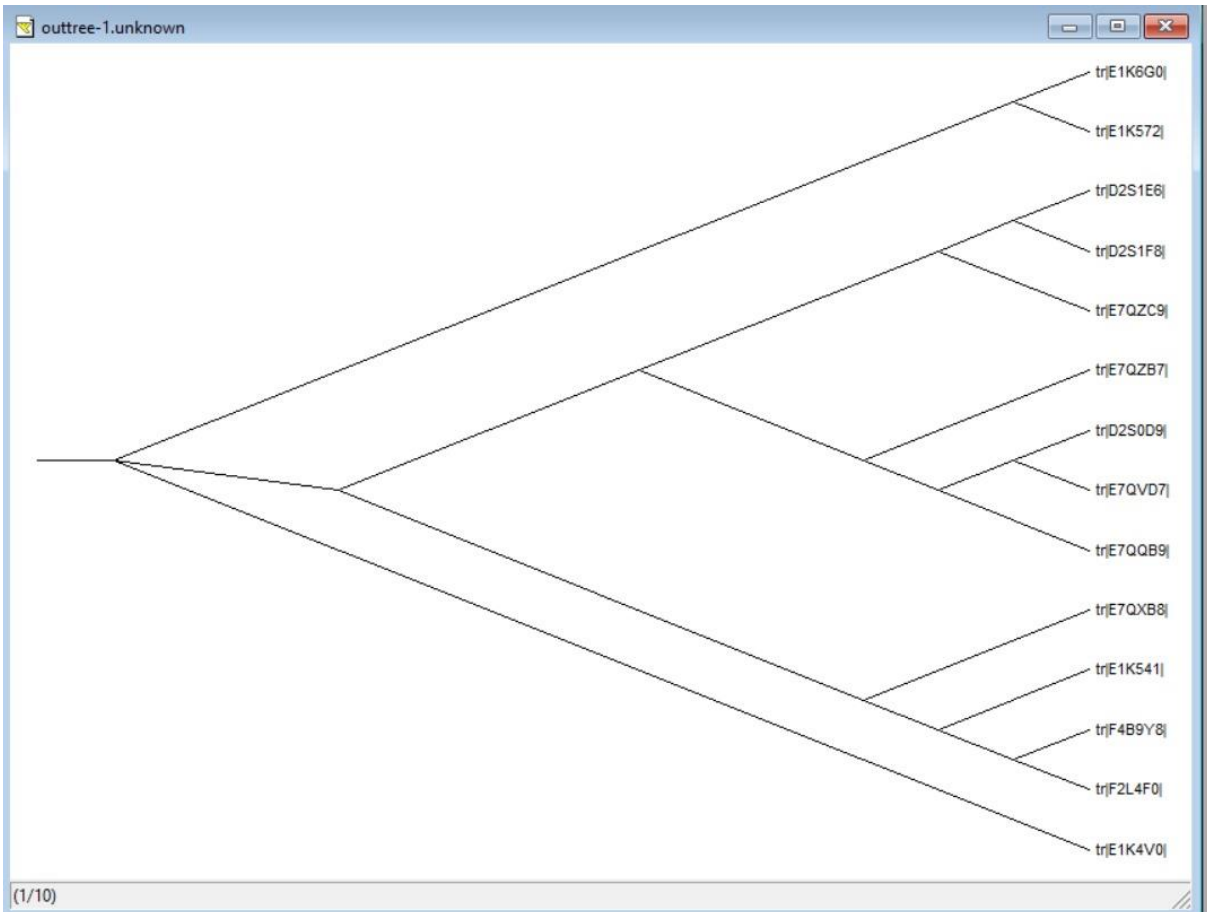

work1-nj-tree
tr|E7QQB9|
tr|E7QVD7|
tr|D2S0D9|
tr|D2S1F8|
tr|D2S1E6|
tr|E7QZB7|
tr|E7QZC9|
tr|E1K541|
tr|F2L4F0|
tr|F4B9Y8|
tr|E7QXB8|
tr|E1K6G0|
tr|E1K572|
tr|E1K4V0|

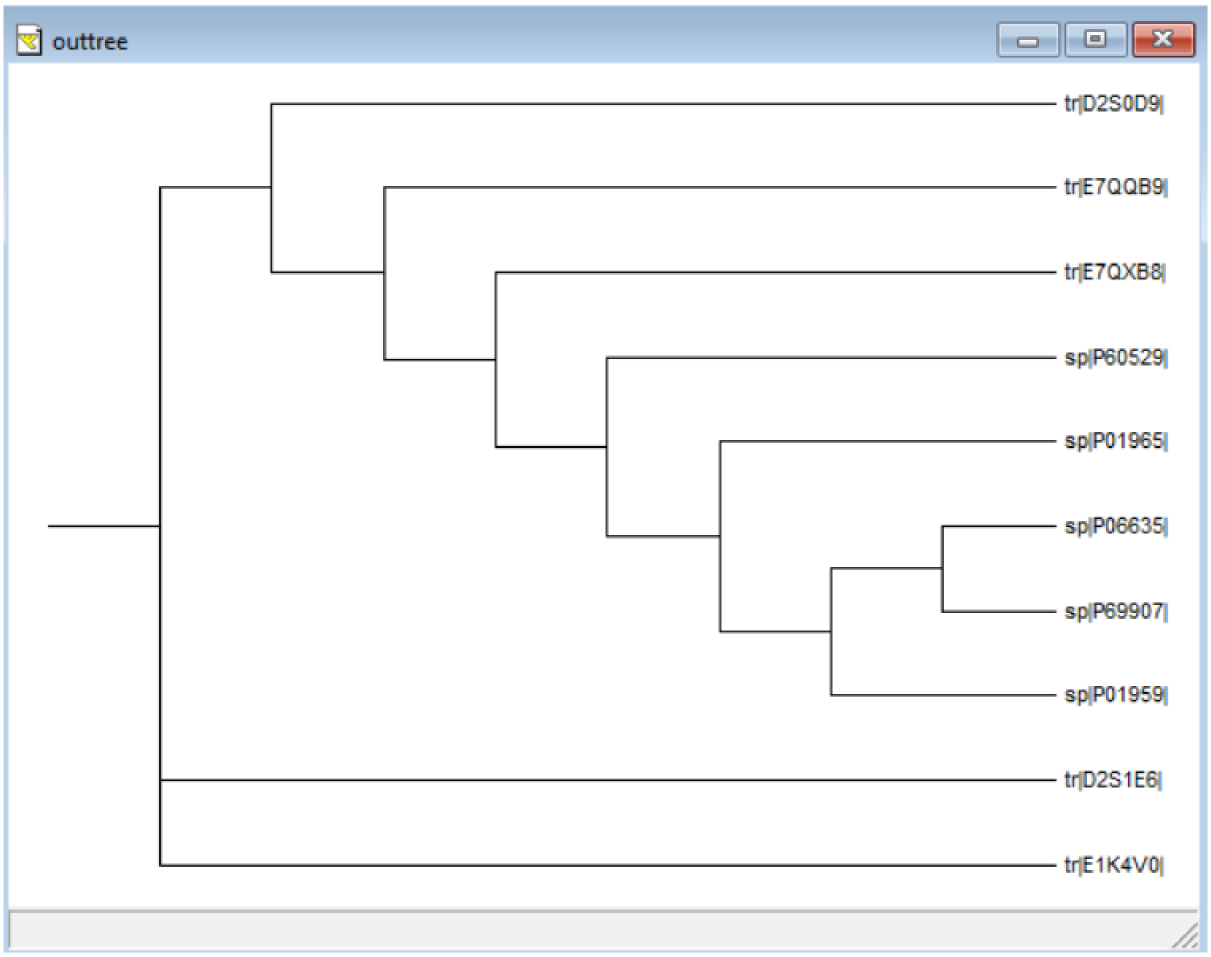

work1constree
tr|F4B9Y8|
tr|F2L4F0|
tr|E1K541|
tr|D2S0D9|
tr|E7QVD7|
tr|E7QZC9|
tr|E7QZB7|
tr|E7QQB9|
tr|D2S1F8|
tr|D2S1E6|
tr|E7QXB8|
tr|E1K4V0|
tr|E1K572|
tr|E1K6G0|

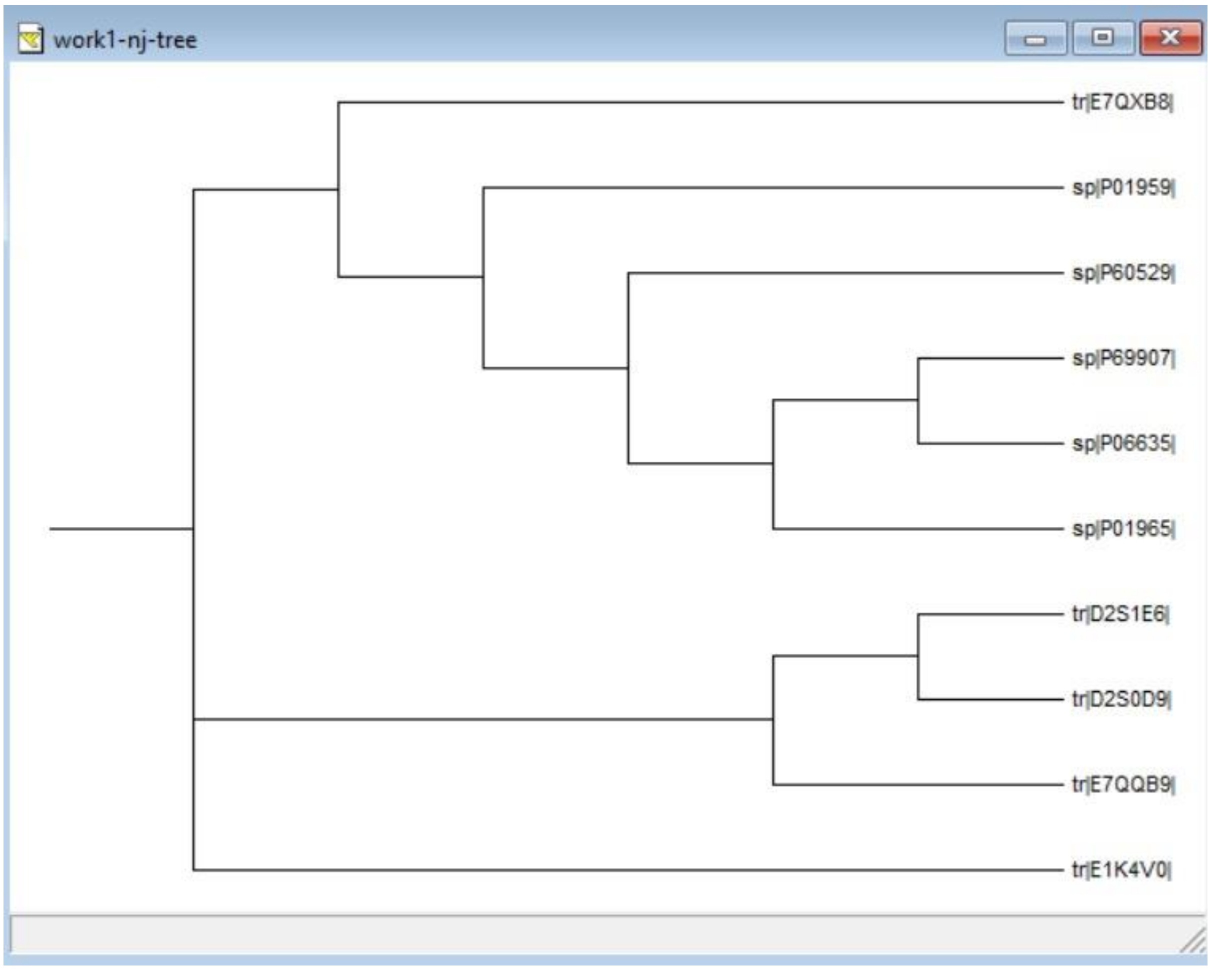

work1-nj-constree
tr|F2L4F0|
tr|F4B9Y8|
tr|E1K541|
tr|E1K572|
tr|E1K6G0|
tr|E1K4V0|
tr|E7QXB8|
tr|D2S0D9|
tr|E7QVD7|
tr|E7QZB7|
tr|E7QZC9|
tr|D2S1E6|
tr|D2S1F8|
tr|E7QQB9|

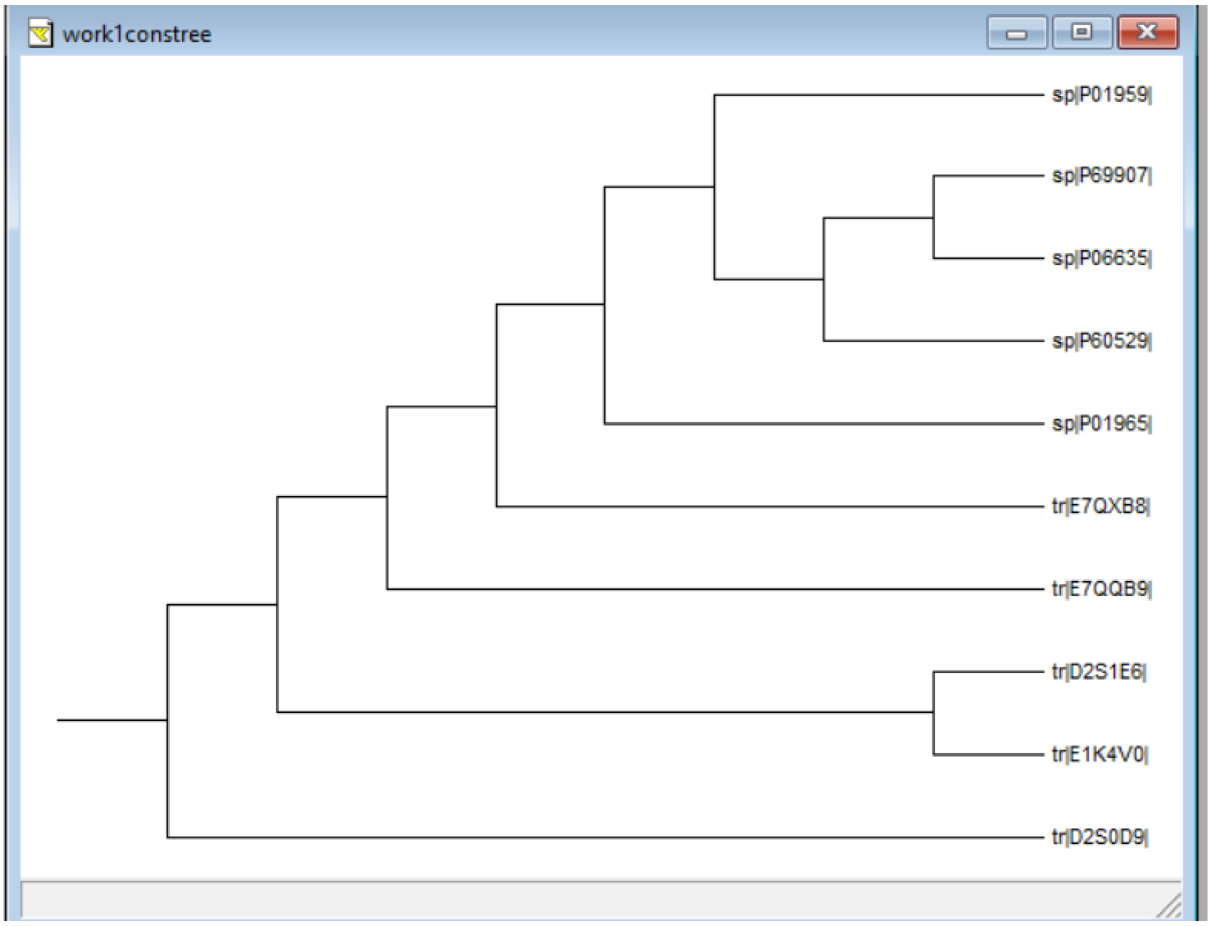
Set 2: tim-hemo.dat

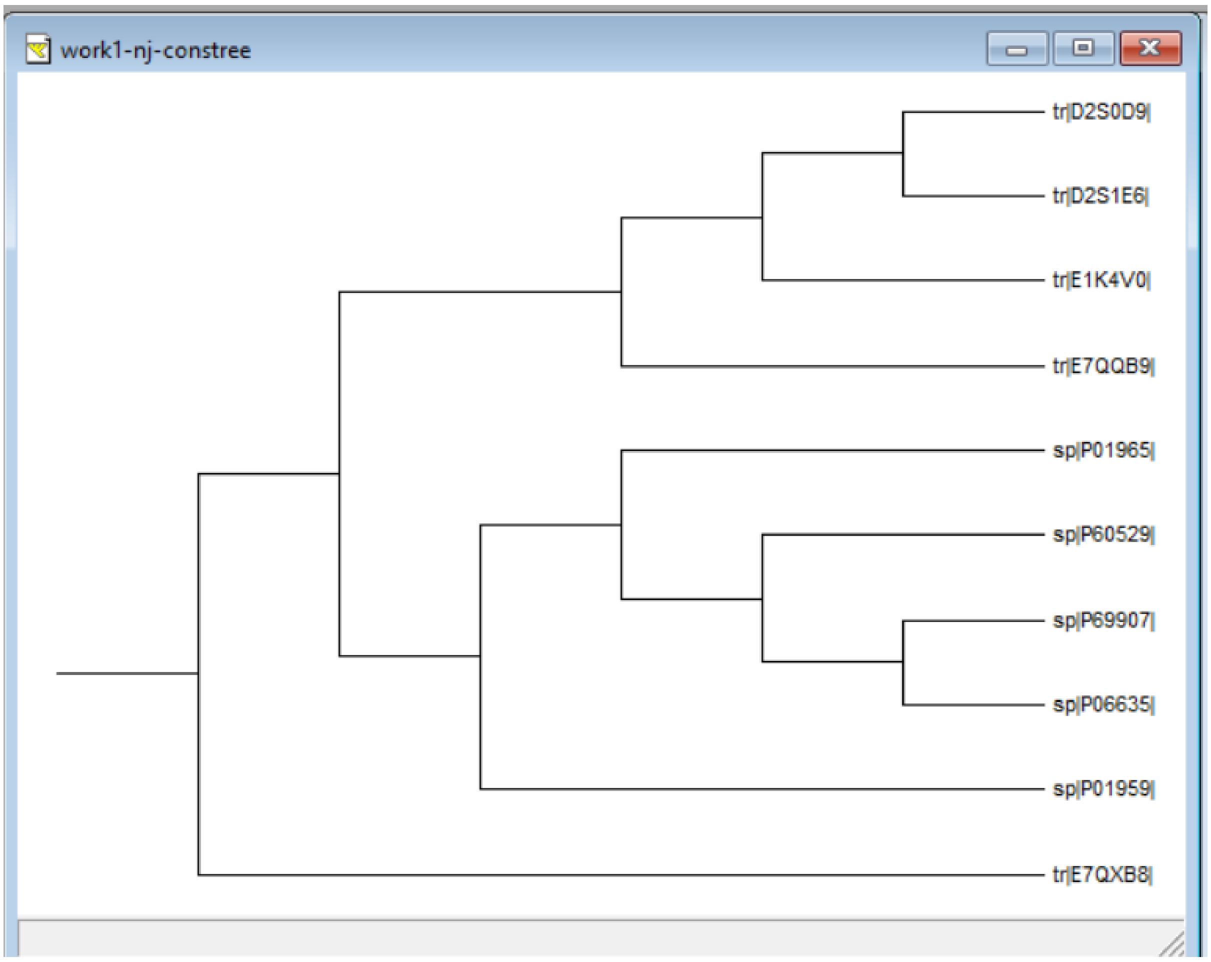
All the files in 'tim_dat_hemo' folder




outtree
tr|D2S0D9|
tr|E7QQB9|
tr|E7QXB8|
sp|P60529|
sp|P01965|
sp|P06635|
sp|P69907|
sp|P01959|
tr|D2S1E6|
tr|E1K4V0|


work1-nj-tree
tr|E7QXB8|
sp|P01959|
sp|P60529|
sp|P69907|
sp|P06635|
sp|P01965|
tr|D2S1E6|
tr|D2S0D9|
tr|E7QQB9|
tr|E1K4V0|


work1constree
sp|P01959|
sp|P69907|
sp|P06635|
sp|P60529|
sp|P01965|
tr|E7QXB8|
tr|E7QQB9|
tr|D2S1E6|
tr|E1K4V0|
tr|D2S0D9|


work1-nj-constree
tr|D2S0D9|
tr|D2S1E6|
tr|E1K4V0|
tr|E7QQB9|
sp|P01965|
sp|P60529|
sp|P69907|
sp|P06635|
sp|P01959|
tr|E7QXB8|

Q2) Obtain the weight matrix for the following sequences

MVLSPADKTNVKGKVGAHAGEYGAAAW
MKRLPADPPCVKGKVKAKAGDYGATTW
MALSAADKTNVKSKVGGHAGEYGAATS
MVLSAADKTNVKSKAGGNAGEWWAAAW
MVLSAADKTNVKSKVLANAGEFGAAAW
ALLPIRTTYHKKCASGHIPEEKDLNNV
DEASSLKGHHIKKLEADALLIPLSASS

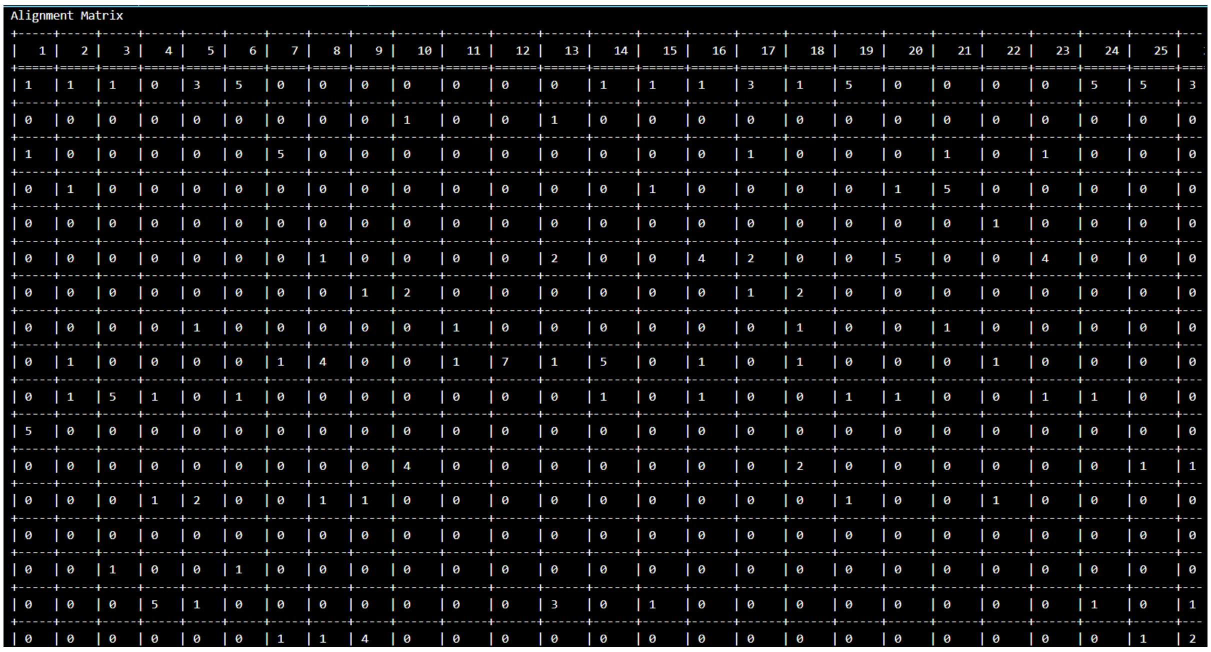
Code:

```
#BT3040| Assignemnt 10 | Q2 | Atharva Mandar Phatak | BE21B009 |

import math

def sequence_to_matrix(input_sequence, sequence_length):
    amino_acids = ['A', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'K', 'L', 'M', 'N',
'P', 'Q', 'R', 'S', 'T', 'V', 'W', 'Y']
    matrix = []

    for i in range(0, len(input_sequence), sequence_length):
        subsequence = input_sequence[i:i+sequence_length]
        row = [0] * len(subsequence)
        for j, aa in enumerate(subsequence):
            row[j] = aa
        matrix.append(row)

    return matrix

input_sequence =
'MVLSPADKTNVKGKVGAHAGEYGAAAWMKRLPADPPCVKGKVKAKAGDYGATTWMALSAADKTNVKSKVGGHAGEYG
AATSMVLSAADKTNVKSKAGGNAGEWWAAAWMVLSAADKTNVKSKVLANAGEFGAAAWALLPIRTTYHKKCASGHIPE
EKDLNNVDEASSLKGHHIKKLEADALLIPLSASS'
sequence_length = 27
amino_acid_matrix = sequence_to_matrix(input_sequence, sequence_length)


def count_amino_acids(matrix):
    amino_acids = ['A', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'K', 'L', 'M', 'N',
'P', 'Q', 'R', 'S', 'T', 'V', 'W', 'Y']
    amino_acid_counts = [[0] * len(matrix[0]) for _ in
range(len(amino_acids))]
```

```python
    for row in matrix:
        for i, amino_acid in enumerate(row):
            amino_acid_counts[amino_acids.index(amino_acid)][i] += 1

    return amino_acid_counts

def calculate_weight_matrix(amino_acid_counts):
    num_rows = len(amino_acid_counts)
    num_cols = len(amino_acid_counts[0])

    weight_matrix = [[0] * num_cols for _ in range(num_rows)]

    for i in range(num_rows):
        for j in range(num_cols):
            n = amino_acid_counts[i][j]
            weight_matrix[i][j] = round(math.log((n + 1/20) / (8/20)),2)

    return weight_matrix


# Count amino acids
amino_acid_counts = count_amino_acids(amino_acid_matrix)

# Calculate weight matrix
weight_matrix = calculate_weight_matrix(amino_acid_counts)


from tabulate import  tabulate
print("Alignment Matrix")
print(tabulate(amino_acid_counts, headers=["{}".format(i+1) for i in 
range(len(amino_acid_matrix[0]))], tablefmt="grid",  numalign="centre"))
print("")
print("")
print("Weight Matrix")
print(tabulate(weight_matrix, headers=["{}".format(i+1) for i in 
range(len(weight_matrix[0]))], tablefmt="grid",  numalign="centre"))
```

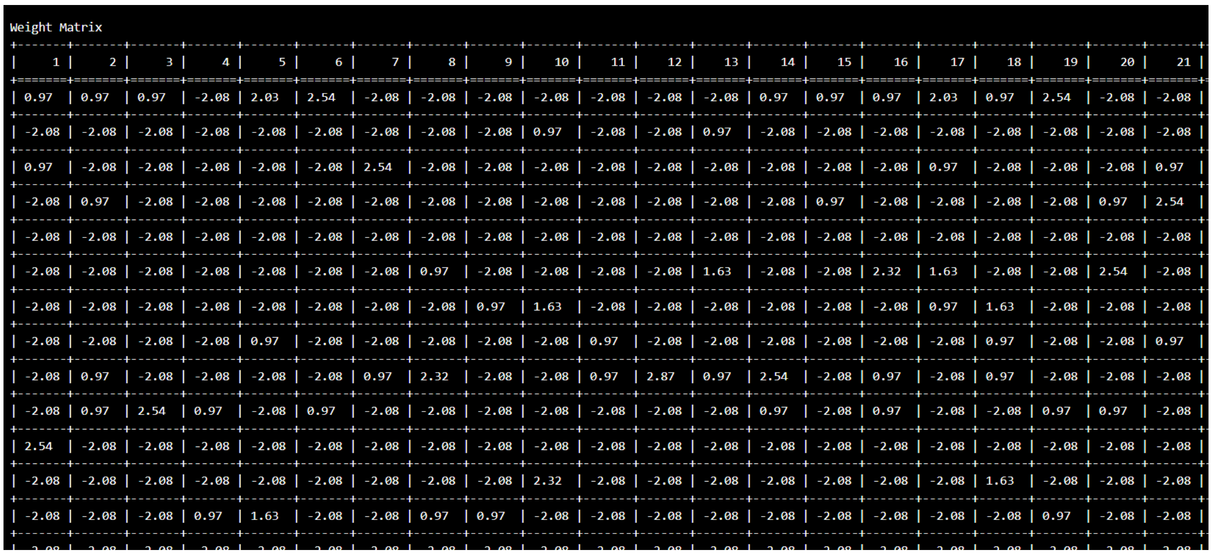
Output:

a) Alignment Matrix



```
Alignment Matrix
+-----+-----+-----+-----+-----+-----+-----+-----+-----+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+---
|   1 |   2 |   3 |   4 |   5 |   6 |   7 |   8 |   9 |   10 |   11 |   12 |   13 |   14 |   15 |   16 |   17 |   18 |   19 |   20 |   21 |   22 |   23 |   24 |   25 |
+=====+=====+=====+=====+=====+=====+=====+=====+=====+======+======+======+======+======+======+======+======+======+======+======+======+======+======+======+======+===
| 1   | 1   | 1   | 0   | 3   | 5   | 0   | 0   | 0   | 0    | 0    | 0    | 0    | 1    | 1    | 1    | 3    | 1    | 5    | 0    | 0    | 0    | 0    | 5    | 5    | 3
+-----+-----+-----+-----+-----+-----+-----+-----+-----+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+---
| 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 1    | 0    | 0    | 1    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0
+-----+-----+-----+-----+-----+-----+-----+-----+-----+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+---
| 1   | 0   | 0   | 0   | 0   | 0   | 5   | 0   | 0   | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 1    | 0    | 0    | 0    | 1    | 0    | 1    | 0    | 0    | 0
+-----+-----+-----+-----+-----+-----+-----+-----+-----+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+---
| 0   | 1   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0    | 0    | 0    | 0    | 0    | 1    | 0    | 0    | 0    | 0    | 1    | 5    | 0    | 0    | 0    | 0    | 0
+-----+-----+-----+-----+-----+-----+-----+-----+-----+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+---
| 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 1    | 0    | 0    | 0    | 0
+-----+-----+-----+-----+-----+-----+-----+-----+-----+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+---
| 0   | 0   | 0   | 0   | 0   | 0   | 0   | 1   | 0   | 0    | 0    | 0    | 2    | 0    | 0    | 4    | 2    | 0    | 0    | 5    | 0    | 0    | 4    | 0    | 0    | 0
+-----+-----+-----+-----+-----+-----+-----+-----+-----+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+---
| 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 1   | 2    | 0    | 0    | 0    | 0    | 0    | 0    | 1    | 2    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0
+-----+-----+-----+-----+-----+-----+-----+-----+-----+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+---
| 0   | 0   | 0   | 0   | 1   | 0   | 0   | 0   | 0   | 0    | 1    | 0    | 0    | 0    | 0    | 0    | 0    | 1    | 0    | 0    | 1    | 0    | 0    | 0    | 0    | 0
+-----+-----+-----+-----+-----+-----+-----+-----+-----+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+---
| 0   | 1   | 0   | 0   | 0   | 0   | 1   | 4   | 0   | 0    | 1    | 7    | 1    | 5    | 0    | 1    | 0    | 1    | 0    | 0    | 0    | 1    | 0    | 0    | 0    | 0
+-----+-----+-----+-----+-----+-----+-----+-----+-----+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+---
| 0   | 1   | 5   | 1   | 0   | 1   | 0   | 0   | 0   | 0    | 0    | 0    | 0    | 1    | 0    | 1    | 0    | 0    | 1    | 1    | 0    | 0    | 1    | 1    | 0    | 0
+-----+-----+-----+-----+-----+-----+-----+-----+-----+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+---
| 5   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0
+-----+-----+-----+-----+-----+-----+-----+-----+-----+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+---
| 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 4    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 2    | 0    | 0    | 0    | 0    | 0    | 0    | 1    | 1
+-----+-----+-----+-----+-----+-----+-----+-----+-----+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+---
| 0   | 0   | 0   | 1   | 2   | 0   | 0   | 1   | 1   | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 1    | 0    | 0    | 1    | 0    | 0    | 0    | 0
+-----+-----+-----+-----+-----+-----+-----+-----+-----+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+---
| 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0
+-----+-----+-----+-----+-----+-----+-----+-----+-----+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+---
| 0   | 0   | 1   | 0   | 0   | 1   | 0   | 0   | 0   | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0
+-----+-----+-----+-----+-----+-----+-----+-----+-----+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+---
| 0   | 0   | 0   | 5   | 1   | 0   | 0   | 0   | 0   | 0    | 0    | 0    | 3    | 0    | 1    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 1    | 0    | 1
+-----+-----+-----+-----+-----+-----+-----+-----+-----+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+------+---
| 0   | 0   | 0   | 0   | 0   | 0   | 1   | 1   | 4   | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 1    | 2
```

b) Weight Matrix



```
Weight Matrix
+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+
|     1 |     2 |     3 |     4 |     5 |     6 |     7 |     8 |     9 |    10 |    11 |    12 |    13 |    14 |    15 |    16 |    17 |    18 |    19 |    20 |    21 |
+=======+=======+=======+=======+=======+=======+=======+=======+=======+=======+=======+=======+=======+=======+=======+=======+=======+=======+=======+=======+=======+
| 0.97  | 0.97  | 0.97  | -2.08 | 2.03  | 2.54  | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | 0.97  | 0.97  | 0.97  | 2.03  | 0.97  | 2.54  | -2.08 | -2.08 |
+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+
| -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | 0.97  | -2.08 | -2.08 | 0.97  | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 |
+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+
| 0.97  | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | 2.54  | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | 0.97  | -2.08 | -2.08 | -2.08 | 0.97  |
+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+
| -2.08 | 0.97  | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | 0.97  | -2.08 | -2.08 | -2.08 | -2.08 | 0.97  | 2.54  |
+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+
| -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 |
+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+
| -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | 0.97  | -2.08 | -2.08 | -2.08 | -2.08 | 1.63  | -2.08 | -2.08 | 2.32  | 1.63  | -2.08 | -2.08 | 2.54  | -2.08 |
+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+
| -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | 0.97  | 1.63  | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | 0.97  | 1.63  | -2.08 | -2.08 | -2.08 |
+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+
| -2.08 | -2.08 | -2.08 | -2.08 | 0.97  | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | 0.97  | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | 0.97  | -2.08 | -2.08 | 0.97  |
+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+
| -2.08 | 0.97  | -2.08 | -2.08 | -2.08 | -2.08 | 0.97  | 2.32  | -2.08 | -2.08 | 0.97  | 2.87  | 0.97  | 2.54  | -2.08 | 0.97  | -2.08 | 0.97  | -2.08 | -2.08 | -2.08 |
+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+
| -2.08 | 0.97  | 2.54  | 0.97  | -2.08 | 0.97  | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | 0.97  | -2.08 | 0.97  | -2.08 | -2.08 | 0.97  | 0.97  | -2.08 |
+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+
| 2.54  | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 |
+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+
| -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | 2.32  | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | 1.63  | -2.08 | -2.08 | -2.08 |
+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+
| -2.08 | -2.08 | -2.08 | 0.97  | 1.63  | -2.08 | -2.08 | 0.97  | 0.97  | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | -2.08 | 0.97  | -2.08 | -2.08 |
+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+
```

Complete output in ‘Q2_Output.txt’