BT 3040: BIOINFORMATICS

Assignment 11

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:

```python
#Q1|Assignment 11 | BT3040 | Atharva Mandar Phatak | BE21B009
import numpy as np
import pandas as pd
import pprint as pprint

amino_acids = [
    "A",  # Alanine
    "R",  # Arginine
    "N",  # Asparagine
    "D",  # Aspartic Acid
    "C",  # Cysteine
    "Q",  # Glutamine
    "E",  # Glutamic Acid
    "G",  # Glycine
    "H",  # Histidine
    "I",  # Isoleucine
    "L",  # Leucine
    "K",  # Lysine
    "M",  # Methionine
    "F",  # Phenylalanine
    "P",  # Proline
    "S",  # Serine
    "T",  # Threonine
    "W",  # Tryptophan
    "Y",  # Tyrosine
    "V"   # Valine
]

prop_dict = {
    "A": 0,  # Alanine
    "R": 0,  # Arginine
    "N": 0,  # Asparagine
    "D": 0,  # Aspartic Acid
    "C": 0,  # Cysteine
    "Q": 0,  # Glutamine
    "E": 0,  # Glutamic Acid
    "G": 0,  # Glycine
    "H": 0,  # Histidine
    "I": 0,  # Isoleucine
    "L": 0,  # Leucine
    "K": 0,  # Lysine
    "M": 0,  # Methionine
    "F": 0,  # Phenylalanine
```

```python
    "P": 0,  # Proline
    "S": 0,  # Serine
    "T": 0,  # Threonine
    "W": 0,  # Tryptophan
    "Y": 0,  # Tyrosine
    "V": 0   # Valine
}


def calculate_propensity(residues, structure):
    # Create a DataFrame to store residue counts
    df = pd.DataFrame({'Residue': list(residues), 'Structure': list(structure)})

#N
    N = len(df['Residue'])




    alpha_conformation_df = df[df['Structure'] == 'H']

#N_alpha
    n_alpha= len(alpha_conformation_df['Residue'])


#n_alpha_i
    n_alpha_i =alpha_conformation_df['Residue'].value_counts()

#N_alpha_i
    N_alpha_i =df['Residue'].value_counts()

    for i in amino_acids:
        if(i in n_alpha_i):
            percent_residue=n_alpha_i[i]/N_alpha_i[i]
            percent_all = n_alpha/N

            prop_of_i = percent_residue/percent_all

            prop_dict[i]=prop_of_i
        else: continue




propensities = calculate_propensity(seq1_prim, seq1_sec)

sorted_prop_dict = {k: v for k, v in sorted(prop_dict.items())}

pprint.pprint(sorted_prop_dict)
```

Output :

```
{'A': 1.5510204081632653,
 'C': 1.0204081632653061,
 'D': 0,
 'E': 0.40816326530612246,
 'F': 1.0204081632653061,
 'G': 1.0612244897959184,
 'H': 0.8746355685131195,
 'I': 1.2004801920768309,
 'K': 0,
 'L': 1.2244897959183674,
 'M': 1.530612244897959,
 'N': 1.3605442176870748,
 'P': 0.22675736961451246,
 'Q': 0,
 'R': 0.6802721088435374,
 'S': 1.3605442176870748,
 'T': 0.7653061224489796,
 'V': 0.5830903790087464,
 'W': 1.1131725417439702,
 'Y': 0.5830903790087464}
```

Q2) Find the propensity of alpha helices manually for the sequence in question 1

Total number of AA in the given sequence (nH): **200**

Total number of AA which have helical confirmation in the given sequence (N): **98**

Denominator value of Propensity formula = 98/200 = **0.49**

Residue	n_alpha_1	N_alpha_i	Percent_Residue	Percent_All_Residue	Ratio (Propensity)
A	19.00	25.00	0.76	0.49	1.55
C	2.00	4.00	0.50	0.49	1.02
D	0.00	0.00	0.00	0.49	0.00
E	1.00	5.00	0.20	0.49	0.41
F	10.00	20.00	0.50	0.49	1.02
G	13.00	25.00	0.52	0.49	1.06
H	3.00	7.00	0.43	0.49	0.87
I	10.00	17.00	0.59	0.49	1.20
K	0.00	0.00	0.00	0.49	0.00
L	12.00	20.00	0.60	0.49	1.22
M	3.00	4.00	0.75	0.49	1.53
N	2.00	3.00	0.67	0.49	1.36
P	1.00	9.00	0.11	0.49	0.23
Q	0.00	0.00	0.00	0.49	0.00
R	3.00	9.00	0.33	0.49	0.68
S	6.00	9.00	0.67	0.49	1.36
T	3.00	8.00	0.38	0.49	0.77
V	2.00	7.00	0.29	0.49	0.58
W	6.00	11.00	0.55	0.49	1.11
Y	2.00	7.00	0.29	0.49	0.58

Q3) Using the rules for helices and strands, identify the helical and strand segments in the following sequence

KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQI NSRWWCNDGRTPGSRNLCNIPCSALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCK GTDVQAWIRGCRL

```
CHOFAS predicts protein secondary structure
version 2.0u66 September 1998
Please cite:
Chou and Fasman (1974) Biochem., 13:222-245


Chou-Fasman plot of @, 129 aa;
seq1


        .         .         .         .         .         .         .
      KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINS
helix    <------->              <------->             <>
sheet              EEEEEEE       EEEEEEE       EEEEEEEE
turns T                         T   T   T   T T          T


        .         .         .         .         .         .         .
      RWWCNDGRTPGSRNLCNIPCSALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDV
helix                   <-------------->   <---->       <>
sheet          EEEE                                     EEE
turns     TT   T T            T        T    T T          T


        QAWIRGCRL
helix
sheet EEEE
turns

Residue totals: H: 47   E: 33   T: 16
      percent: H: 36.4 E: 25.6 T: 12.4
```

←--→ : Represents helix

EEE.. : Represents Sheet

TTT.. : Represents Turn

Helix:

RCELAAAMKR

WVCAAKFESN

IL

LSSDITASVNCAKKIVS

MNAWVA

DV

Sheet

YSLGNWV

FNTQATN

TDYGILQI

LCNI

TDV

Turns

K

S

T

R

G

T

S

DG

G

R

S

K

D

N

N

Q4) Verify one of the helical and strand segments, manually

→Q4)

Alpha Helix

KVFGRC $= a + \delta\delta + o\delta +$

$= 0.5 + 1 + 1 + (-1) + 0 + 0$

$= 1.5 \ (<4)$

VFGRCE $= 1 + 1 + (-1) + 0 + 0 + 1$

$= 2 \ (<4)$

R·

FGRCEL $= 1 + (-1) + 0 + 0 + 1 + 1$

$= 2$

GRCELA $= (-1) + 0 + 0 + 1 + 1 + 1$

$= 2$

RCGLAA $= 0 + 0 + 1 + 1 + 1 + 1$

$= 44$

Extending Further

ELAA → $1.53 + 1.34 + 1.45 + 1.45 = 5.77 \ (>4)$
LAAA → $1.34 + 1.45 + 1.45 + 1.45 = 5.69 \ (>4)$
AAAM → $1.45 + 1.45 + 1.45 + 1.20 = 5.55 \ (>4)$
AAMK → $1.45 + 1.45 + 1.20 + 1.07 = 5.17 \ (>4)$
AMKR → $1.45 + 1.20 + 1.07 + 0.77 = 4.51 \ (>4)$
MKRH → $1.20 + 1.07 + 0.79 + 1.24 = 4.3 \ (>4)$
KRHG → $1.07 + 0.79 + 1.24 + 0.53 = \boxed{3.63 \ (<4)}$

∴ The potential α-helix segment can be

RCELAAAMKRHG

for β-strand

$$KVFGR = -1 + 1 + 1 + 0 + 0 = 1 \, (<3)$$

$$VFGRC = 1 + 1 + 0 + 0 + 1 = 3 \, (=3)$$

condition
satisfied

Extending Segment

$$GRC = 0.81 + 0.9 + 1.3 = 3.01 \, (>3)$$

$$RCE = 0.9 + 1.3 + 0.26 = 2.46 \, (<3)$$

∴ β-strand can be $\boxed{\text{"VFGRC"}}$