BT3040 – BIOINFORMATICS – Assignment 8

Submitted by Sahana (BE17B038)

Question 1

Algorithm –

- Count absolute difference between compositions of a pair of sequences for Hamming distance
- Sum the square of difference between compositions of a pair of sequence and then take square root of it for Euclidian distance

Code –

```
def distance(a,b,c):
    cA = composition(a)
    cB = composition(b)
    cC = composition(c)

    H_AB = 0
    H_BC = 0
    H_AC = 0
    E_AB = 0
    E_BC = 0
    E_AC = 0
    for i in range(20):
        H_AB+=abs(cA[i]-cB[i])
        H_BC+=abs(cB[i]-cC[i])
        H_AC+=abs(cA[i]-cC[i])
        E_AB+=(cA[i]-cB[i])**2
        E_BC+=(cC[i]-cB[i])**2
        E_AC+=(cA[i]-cC[i])**2
    E_AB=E_AB**0.5
    E_BC=E_BC**0.5
    E_AC=E_AC**0.5
    Hmin = min(H_AB,H_BC,H_AC)
    Emin = min(E_AB,E_BC,E_AC)
#     print(H_AB,H_BC,H_AC)
#     print(E_AB,E_BC,E_AC)
    print('Sequences that are close to each other based on Hamming distance
= %f' %Hmin)
    print('Sequences that are close to each other based on Eucledian
distance = %f' %Emin)

s1 =
'MALLPAAPGAPARATPTRWPVGCFNRPWTKWSYDEALEFIKAAGYAWTGLLTASKPSLHHATATPEYLAALKQK
SRHAA'
s2 =
'AMENLNMDLLYMAAAVMMGLAAIGAAIGIGILGGKFLEAFARQPDLIPLLRTQFFIVMGLVDAIPMIAVGLGLY
VFAVA'
s3 = 'AADVSAAVGATGQSGMTYRLGLSWDWDKSWWQTHATGIRLTGYWDAGYTYWEGGDEGAGKHSLSF'

distance(s1,s2,s3)
```

Output –

Pair	Hamming distance	Euclidian distance
Seq 1 and Seq 2	86.0759493670886	21.924693766694645
Seq 2 and Seq 3	91.99610516066208	24.67582831988431
Seq 1 and Seq 3	67.03018500486853	21.26072178817923
Result	Seq 1 and Seq 3	Seq 1 and Seq 3

Question 2

Algorithm –

- Get the manually curated sequences from UniProt (690 sequences)
- At the webserver of CD-HIT, set % identity as 0.4/0.4/0.75/0.9
- Analyse the results.

Result -

Percentage identity	Total number of clusters (representative sequences)	Cluster with largest number of sequences	Number of sequences in the aforementioned cluster
40%	235	0 (i.e., 1st)	69
50%	296	0 (i.e., 1st)	66
75%	420	0 (i.e., 1st)	66
90%	499	0 (i.e., 1st)	47

Question 3 and Question 4 – No results since PISCES server isn't functioning.

Question 5

The total number of beta barrel membrane proteins is 51,656. Of which, 690 proteins have ben reviewed (SwissProt) and the remaining are unreviewed (TrEMBL). (The manually annotated sequences have been used for the above question.) 50% similarity identity from Uniref for the above sequences yields 357 sequences. However, Uniref does not differentiate between manually annotated and unreviewed sequences.

Methods	Total number of clusters (representative sequences)	Cluster with largest number of sequences	Number of sequences in the aforementioned cluster
Uniref similarity cutoff = 50% (out of 51,656 sequences)	357	UniRef50_O03042	17,287
CD-HIT (out of 690 sequences)	296	0 (i.e., 1st)	66