Practical 9

## Date: 27-03-2022 Name: Aswin. V

## Day: Sunday Roll Number: BS19B003

**Q1 –**

def Q1(A, B):  
 AA\_A = {'A': 0, 'C': 0, 'D': 0, 'E': 0, 'F': 0, 'G': 0, 'H': 0, 'I': 0, 'K': 0, 'L': 0,  
 'M': 0, 'N': 0, 'P': 0, 'Q': 0, 'R': 0, 'S': 0, 'T': 0, 'V': 0, 'W': 0, 'Y': 0}  
 AA\_B = {'A': 0, 'C': 0, 'D': 0, 'E': 0, 'F': 0, 'G': 0, 'H': 0, 'I': 0, 'K': 0, 'L': 0,  
 'M': 0, 'N': 0, 'P': 0, 'Q': 0, 'R': 0, 'S': 0, 'T': 0, 'V': 0, 'W': 0, 'Y': 0}  
  
 H\_AB, E\_AB = 0, 0  
 d1 = ['A', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'K', 'L',  
 'M', 'N', 'P', 'Q', 'R', 'S', 'T', 'V', 'W', 'Y']  
  
 for i in A:  
 AA\_A[i] += 100 / len(A)  
  
 for i1 in B:  
 AA\_B[i1] += 100 / len(B)  
  
 for i2 in range(20):  
 d = AA\_A[d1[i2]] - AA\_B[d1[i2]]  
 H\_AB += abs(d)  
 E\_AB += d \*\* 2  
 E\_AB = E\_AB \*\* 0.5  
  
 return H\_AB, E\_AB  
  
if \_\_name\_\_ == '\_\_main\_\_':  
 A = 'AMENLNMDLLYMAAAVMMGLAAIGAAIGIGILGGKFLEGAARQPDLIPLLRTQFFIVMGLVDAI' \  
 'PMIAVGLGLYVMFAVA'  
 B = 'AADVSAAVGATGQSGMTYRLGLSWDWDKSWWQTSTGRLTGYWDAGYTYWEGGDEGAGKHSLSFAP' \  
 'VFVYEFAGDSIKPFIEAGIGVAAFSGTRVGDQNLGSSLNFEDRIGAGLKFANGQSVGVRAIHYSNAGLKQPN' \  
 'DGIESYSLFYKIPI'  
 C = 'MALLPAAPGAPARATPTRWPVGCFNRPWTKWSYDEALDGIKAAGYAWTGLLTASKPSLHHATATPEY' \  
 'LAALKQKSRHAA'  
  
 H\_12, E\_12 = Q1(A, B)  
 H\_23, E\_23 = Q1(B, C)  
 H\_31, E\_31 = Q1(C, A)  
  
 st = '{:<20} {:<20} {:<20}'  
 print(st.format('Pair', 'Hamming Distance', 'Euclidian Distance'))  
 print(st.format('Seq 1 and Seq 2', H\_12, E\_12))  
 print(st.format('Seq 2 and Seq 3', H\_23, E\_23))  
 print(st.format('Seq 3 and Seq 1', H\_31, E\_31))  
 print(st.format('Result', min(H\_12, H\_23, H\_31), min(E\_12, E\_23, E\_31)))

Output:

Pair Hamming Distance Euclidian Distance

Seq 1 and Seq 2 66.5728476821192 20.1062168421535

Seq 2 and Seq 3 72.6632576075111 20.112952107271113

Seq 3 and Seq 1 84.33544303797467 22.086816691389572

Result 66.5728476821192 20.1062168421535

**Q2 –**

Algorithm:

* Get the UniProt sequences that have been manually curated. (703 sequences)
* In the parameters in CD-HIT web server change the % identity to 0.4, 0.5, 0.75, 0.9.

Results:

|  |  |  |  |
| --- | --- | --- | --- |
| % Identity | Total number of clusters | Cluster with the largest number of sequences | Number of sequences in the cluster |
| 40 | 245 | 1st | 69 |
| 50 | 304 | 1st | 66 |
| 75 | 430 | 1st | 66 |
| 90 | 509 | 1st | 46 |

**Q3 –**

PISCES is not responding.

**Q4 –**

Comparison between 40% and 50% cut-offs in CD-HIT:

Among the above obtained clusters;

* 25 were related to Homo sapiens
* 15 were related to E.Coli
* 14 were related to K2 strain of E.coli
* 2 were related to coronavirus.

**Q5 –**



There are 63,789 beta barrel membrane proteins in all. 703 proteins have been evaluated and can be found in SwissProt, while the rest have not been reviewed and can be found in TrEMBL. For the above sequences, Uniref produces 357 sequences with a 50 percent similarity identity. Uniref, on the other hand, does not distinguish between manually annotated and unreviewed sequences.

|  |  |  |  |
| --- | --- | --- | --- |
| Method | Total number of clusters | Clusters with the largest number of sequences | Number of sequences in the cluster mentioned |
| Uniref similarity cut-off 50% | 365 | <https://www.uniprot.org/uniref/UniRef50_O03042> | 15,299 |
| CD-HIT  (703 sequences) | 304 | 1st | 66 |