Practical 9

Name: Pavani Kudari
Roll no: BE19B023

1.

```
def Q1(Seq1, Seq2):
    AA_Seq1 = {'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_Seq2 = {'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 Seq1:
        AA_Seq1[i] += 100 / len(Seq1)

    for j in Seq2:
        AA_Seq2[j] += 100 / len(Seq2)

    for k in range(20):
        d = AA_Seq1[d1[k]] - AA_Seq2[d1[k]]
        H_AB += abs(d)
        E_AB += d ** 2
    E_AB = E_AB ** 0.5

    return H_AB, E_AB

if __name__ == '__main__':
    Seq1 =
'AMENLNMDLLYMAAAVMMGLAAIGAAIGIGILGGKFLEGAARQPDLIPLLRTQFFIVMGLVDAI' \
        'PMIAVGLGLYVMFAVA'
    Seq2 =
'AADVSAAVGATGQSGMTYRLGLSWDWDKSWWQTSTGRLTGYWDAGYTYWEGGDEGAGKH
SLSFAP' \

'VFVYEFAGDSIKPFIEAGIGVAAFSGTRVGDQNLGSSLNFEDRIGAGLKFANGQSVGVRAIHYS
NAGLKQPN' \
        'DGIESYSLFYKIPI'
    Seq3 =
'MALLPAAPGAPARATPTRWPVGCFNRPWTKWSYDEALDGIKAAGYAWTGLLTASKPSLHHATA
TPEY' \
        'LAALKQKSRHAA'
```

```
H_12, E_12 = Q1(Seq1, Seq2)
H_23, E_23 = Q1(Seq2, Seq3)
H_31, E_31 = Q1(Seq3, Seq1)

st = '{:<25} {:<25} {:<25}'
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))
```

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
```

Seq 1 and Seq 2 are close to each other using Hamming and Euclidean distance methods.

In [ ]:

2.

Algorithm:

- Download the manually curated UniProt sequences. (703 variations)
- Change the percent identity to 0.4, 0.5, 0.75, and 0.9 in the CD-HIT web server parameters.

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

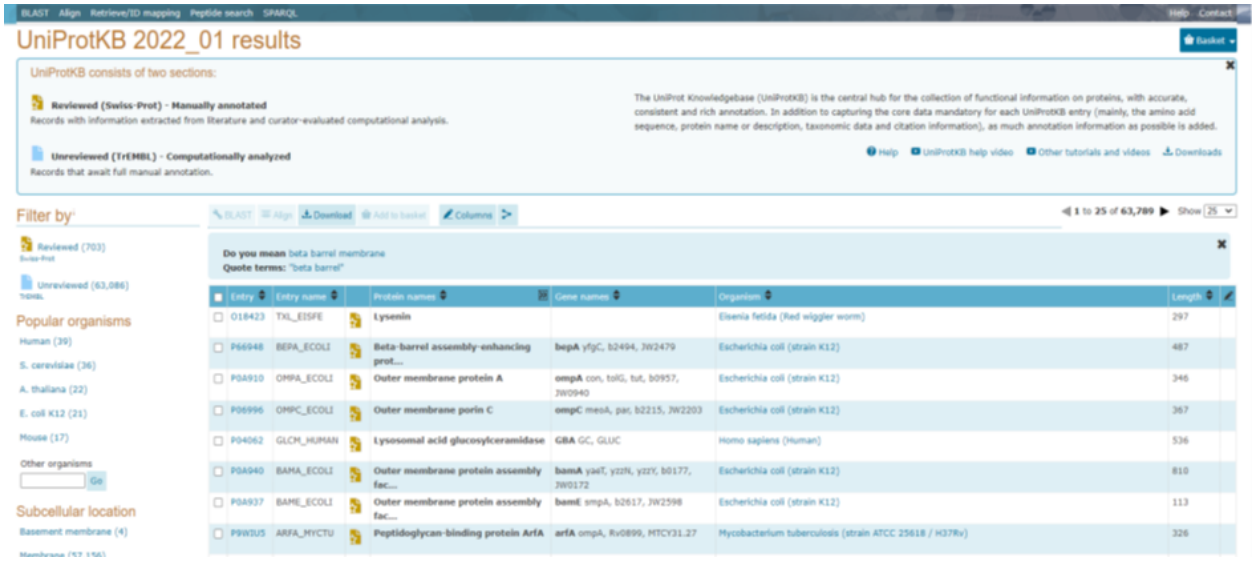
90	509	1st	46

3. PISCES is not responding.

4. Compare the results obtained with the cut-offs 40% and 50%.

- 25 are related to Homo sapiens
- 14 are related to K2 strain of E.coil
- 15 are related to E.coli
- 2 are related to coronavirus

5.



There are a total of 63,789 beta barrel membrane proteins. 703 proteins have been reviewed and are available in SwissProt, whereas the remaining proteins have not been reviewed and are available in TrEMBL. Uniref generates 357 sequences with a 50 percent similarity identity for the above sequences. Uniref, on the other hand, does not distinguish between sequences that have been manually annotated and those that have not been evaluated.

Method: Uniref similarity cut-off 50%

Total no of clusters: 365

Clusters with the largest number of sequence: https://www.uniprot.org/uniref/UniRef50_O03042

Number of sequences in the cluster mentioned: 15,299

For Method: CD- HIT

Total number of clusters: 304

Number of sequences in cluster mentioned: 66