

## BT3040 – Bioinformatics

## Practical 9

1

The Python code to find the pair of sequences which are close to each other using Hamming and Euclidean distance methods is given below:

```
import pandas as pd

residues = ["G", "A", "V", "L", "I", "P", "F", "Y", "W", "S", "T", "C", "M", "N", "Q",
"D", "E", "K", "R", "H"]

def composition(sequence):
    composition = {residue: 100 * sequence.count(residue) / len(sequence) for residue
in residues}

    return composition

def hamming(sequence1, sequence2):
    comp1 = composition(sequence1)
    comp2 = composition(sequence2)

    ham_dist = 0

    for residue in residues:
        ham_dist += abs(comp1[residue] - comp2[residue])

    return ham_dist

def euclidean(sequence1, sequence2):
    comp1 = composition(sequence1)
    comp2 = composition(sequence2)

    euc_dist = 0

    for residue in residues:
        euc_dist += (comp1[residue] - comp2[residue]) ** 2

    euc_dist = euc_dist ** 0.5

    return euc_dist

seq1 =
"AMENLNMDLLYMAAAVMMGLAAIGAAGIGILGGKFLEGAARQPDLIPLLRQTQFFIVMGLVDAIPMIAVGLGLYVMFAVA"
seq2 =
"AADVSAAVGATGQSGMTYRLGLSWDWDKSWWQTSTGRLTGYWDAGYTYWEGGDEGAGKHSLSFAPVFVYEFAGDSIKPFIEAGIGV
AAFSGTRVGDQNLGSSLNFEDRIGAGLKFANGQSVGVRAIHYSNAGLKQPNDGIESYSLFYKIPI"
seq3 =
"MALLPAAPGAPARATPTRWPVGCFNRPWTKWSYDEALDGIKAAGYAWTGLLTASKPSLHHATATPEYLAALKQKSRHAA"

distances = pd.DataFrame({"Hamming Distance": [hamming(seq1, seq2), hamming(seq1,
seq3), hamming(seq2, seq3)],
                        "Euclidean Distance": [euclidean(seq1, seq2), euclidean(seq1,
seq3), euclidean(seq2, seq3)]})
distances.index = ["1 and 2", "1 and 3", "2 and 3"]

print(distances)
```

```
print("")

print(f"Based on Hamming distance, sequences {distances[distances.columns[0]].idxmin()}
are the closest.")
print(f"Based on Euclidean distance, sequences
{distances[distances.columns[1]].idxmin()} are the closest.")
```

The output of this is shown below:

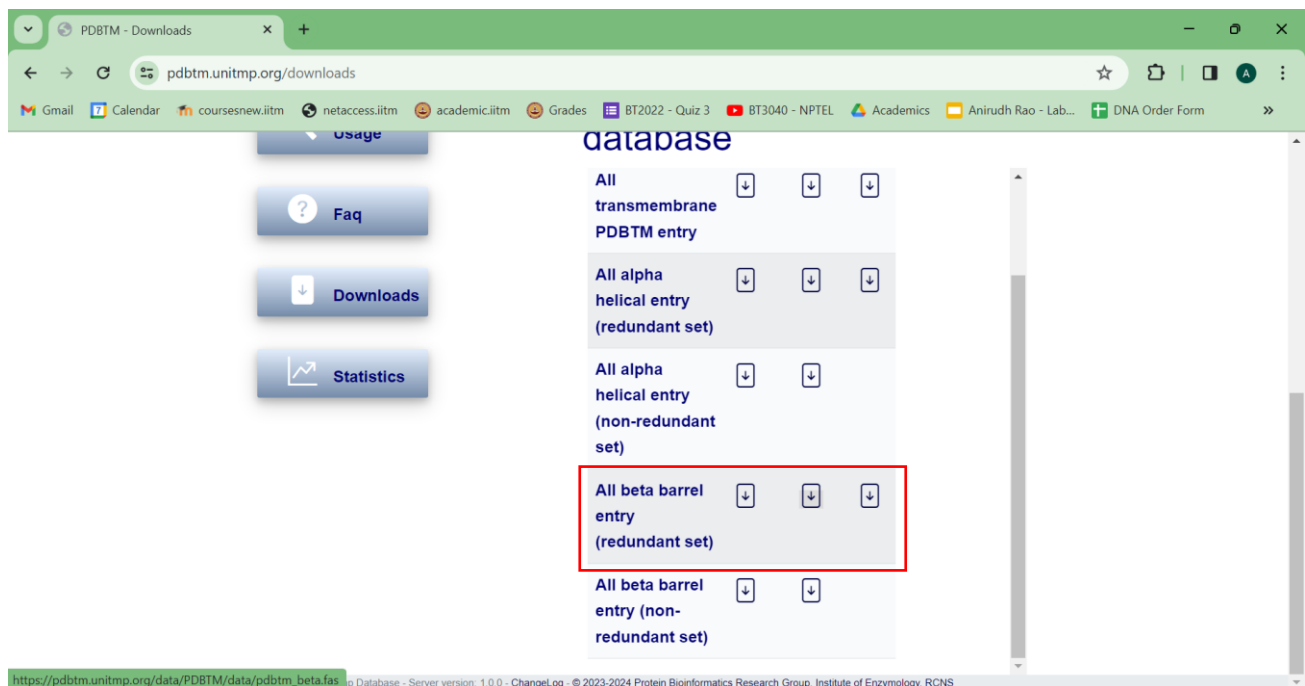
```
Run - Anirudh-Python
workboard x
C:\Users\anira\PycharmProjects\Anirudh-Python\venv\Scripts\python.
Hamming Distance Euclidean Distance
1 and 2 66.572848 20.106217
1 and 3 84.335443 22.086817
2 and 3 72.663258 20.112952

Based on Hamming distance, sequences 1 and 2 are the closest.
Based on Euclidean distance, sequences 1 and 2 are the closest.

Process finished with exit code 0
```

## 2

First, a FASTA file (beta.fasta) containing all beta barrel membrane proteins was downloaded from the PDBTM database. This contains 2878 redundant sequences.



Next, CD-HIT was installed in WSL by running the command:

```
sudo apt-get install cd-hit
```

CD-HIT was then used to obtain the non-redundant sequences with sequence identities of less than 40%, 50%, 75% and 90%. This was done by running the commands:

```
cd-hit -I beta.fasta -o beta_40.txt -c 0.4 -n 2
cd-hit -I beta.fasta -o beta_50.txt -c 0.5 -n 2
cd-hit -I beta.fasta -o beta_75.txt -c 0.75 -n 2
cd-hit -I beta.fasta -o beta_90.txt -c 0.9 -n 2
```

```
anirao@CRISPR: ~
Command: cd-hit -i beta.fasta -o beta_90.txt -c 0.9 -n 2

Started: Tue Apr 9 09:39:00 2024
=====
                        Output
=====
Your word length is 2, using 5 may be faster!
total seq: 2878
longest and shortest : 2124 and 11
Total letters: 983136
Sequences have been sorted

Approximated minimal memory consumption:
Sequence      : 1M
Buffer        : 1 X 10M = 10M
Table         : 1 X 0M = 0M
Miscellaneous  : 0M
Total         : 12M

Table limit with the given memory limit:
Max number of representatives: 1531928
Max number of word counting entries: 98450154

comparing sequences from      0 to      2878
..      2878 finished      370 clusters

Approximated maximum memory consumption: 12M
writing new database
writing clustering information
program completed !

Total CPU time 0.60
```

The results from CD-HIT are:

Identity	Number of Clusters
40%	240
50%	265
75%	330
90%	370

The PISCES webserver was used to get the non-redundant sequences of beta barrel membrane proteins with sequence identities of less than 20%, 30%, 40% and 50%. Default parameters were used.

**PISCES: A Protein Sequence Culling Server**

**Step 1: Input PDB list**

Paste or type in your list of PDB chains in the following textbox [Help?](#)

```
4a8d_A,4a8d_B,4a8d_C,4a8d_D,4a8d_E,4a8d_F,4a8d_G,4a8d_H,4a8d_I,4a8d_J,4a8d_K,4a8d_L,4a8d_M,3b07_A,3b07_C,3b07_E,3b07_G,3b07_B,3b07_D,3b07_F,3b07_H,8b4i_A,8b4i_B,8b4i_C,8b4i_D,8b4i_E,8b4i_F,8b4i_G,8b4i
```

**Step 2: Choose your desired thresholds**

Maximum pairwise percent sequence identity:	<input type="text" value="20"/>
Minimum resolution (X-ray and EM):	<input type="text" value="0.0"/>
Maximum resolution (X-ray and EM):	<input type="text" value="2.0"/>
Maximum R-value (X-ray only):	<input type="text" value="0.25"/>
Minimum chain length:	<input type="text" value="40"/>
Maximum chain length:	<input type="text" value="10000"/>

The results were obtained via email.

results from PISCES cullpdb\_pc20.0\_res0.0-2.0\_len40-10000\_R0.25\_Xray\_d2024\_04\_08\_chains31

roland.dunbrack@fcc.edu  
to me

Mon, Apr 8, 1:28 PM (1 day ago)

Your thresholds for culling the PDB

- Resolution : 0.0 - 2.0
- R-factor : 0.25
- Sequence length : 40 - 10000
- Sequence percentage identity: <= 20.0
- X-ray entries: Included
- EM entries: Excluded
- NMR entries: Excluded
- Chains with chain breaks: Included
- Chains with disorder: Included

Number of chains: 31

The results from PISCES are:

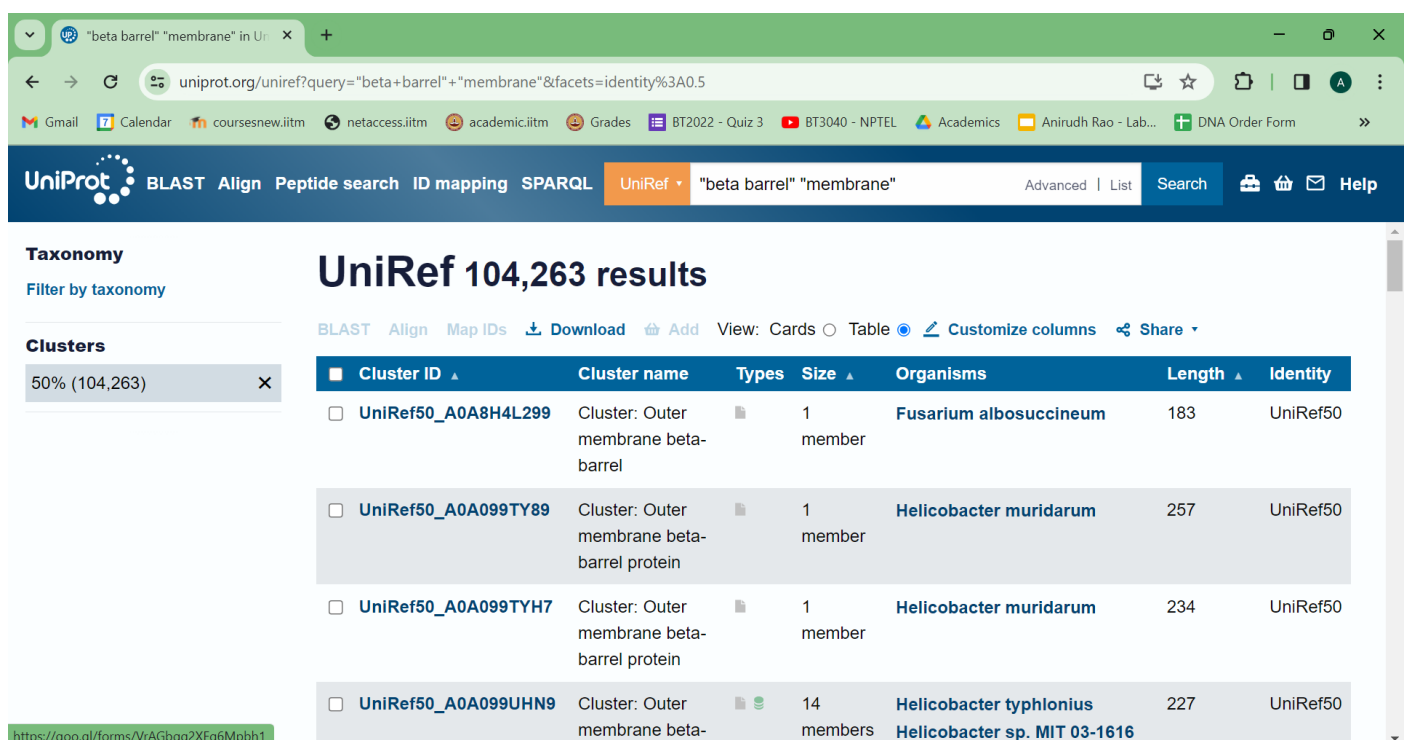
Identity	Number of Sequences
20%	31
30%	39
40%	42
50%	46

4

- In CD-HIT, the number of clusters increases as the cutoff is changed from 40% to 50%.
- In PISCES, the number of sequences increases as the cutoff is changed from 40% to 50%.
- In CD-HIT, the size of the largest clusters with identity 40% is 322 and with identity 50% is 318.
- In PISCES, the new sequences in the sequences with 50% identity are 2Y2X\_A, 1QJ8\_A, 4FRX\_A, and 5FP1\_A, which are not present in the sequences with 40% identity.

5

UniRef was used to obtain the beta barrel membrane proteins with 50% identity cutoff.



The screenshot shows the UniProt website interface. The search query is "beta barrel" "membrane". The results are displayed as a table with 7 columns: Cluster ID, Cluster name, Types, Size, Organisms, Length, and Identity. The table shows 4 clusters, all of which are "Cluster: Outer membrane beta-barrel protein". The organisms listed are *Fusarium albosuccineum*, *Helicobacter muridarum*, and *Helicobacter typhlonius*. The lengths range from 183 to 257, and the identities are all UniRef50.

Cluster ID	Cluster name	Types	Size	Organisms	Length	Identity
UniRef50_A0A8H4L299	Cluster: Outer membrane beta-barrel		1 member	<i>Fusarium albosuccineum</i>	183	UniRef50
UniRef50_A0A099TY89	Cluster: Outer membrane beta-barrel protein		1 member	<i>Helicobacter muridarum</i>	257	UniRef50
UniRef50_A0A099TYH7	Cluster: Outer membrane beta-barrel protein		1 member	<i>Helicobacter muridarum</i>	234	UniRef50
UniRef50_A0A099UHN9	Cluster: Outer membrane beta-		14 members	<i>Helicobacter typhlonius</i> <i>Helicobacter sp. MIT 03-1616</i>	227	UniRef50

The results were downloaded in .xlsx format.

Cluster ID	Cluster Name	Types	Size	Organism	Length	Identity
UniRef50_Cluster: O UniProtKB	7	Acinetoba	228	0.5		
UniRef50_Cluster: O UniProtKB	5	Candidatu	270	0.5		
UniRef50_Cluster: O UniProtKB	3	Sphingobi	433	0.5		
UniRef50_Cluster: O UniProtKB	19	Comamon	118	0.5		
UniRef50_Cluster: A UniProtKB	108	Erwinia m	1290	0.5		
UniRef50_Cluster: O UniProtKB	5	Bacteroid	535	0.5		
UniRef50_Cluster: O UniProtKB	2	Deinococc	164	0.5		
UniRef50_Cluster: O UniProtKB	3	Hylemone	218	0.5		
UniRef50_Cluster: O UniProtKB	6	Hylemone	247	0.5		
UniRef50_Cluster: O UniProtKB	1	Hylemone	218	0.5		
UniRef50_Cluster: O UniProtKB	1	Hylemone	204	0.5		
UniRef50_Cluster: O UniProtKB	1	Hylemone	207	0.5		
UniRef50_Cluster: O UniProtKB	4	Hylemone	194	0.5		
UniRef50_Cluster: O UniProtKB	263	Limimari	215	0.5		
UniRef50_Cluster: O UniProtKB	5	Limimari	218	0.5		
UniRef50_Cluster: O UniProtKB	2	Rubellim	203	0.5		
UniRef50_Cluster: O UniProtKB	2	Rubellim	238	0.5		
UniRef50_Cluster: O UniProtKB	2	Chondron	268	0.5		
UniRef50_Cluster: O UniProtKB	1	Chondron	270	0.5		
UniRef50_Cluster: O UniProtKB	2	Chondron	270	0.5		

- The number of clusters was obtained as 104263, containing 506879 sequences.
- The largest cluster (Ail/Lom family outer membrane beta-barrel protein) has a size of 8551.