Anirudh Rao
BE21B004

BT3040 – Bioinformatics

Practical 9

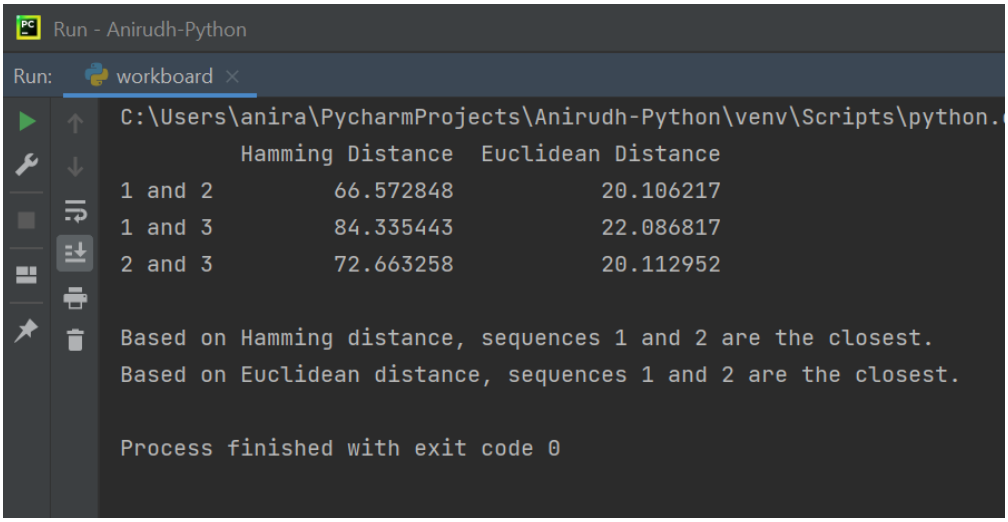
1

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 =
"AMENLNMDLLYMAAAVMMGLAAIGAAIGIGILGGKFLEGAARQPDLIPLLRTQFFIVMGLVDAIPMIAVGLGLYVMFAVA"
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)
```

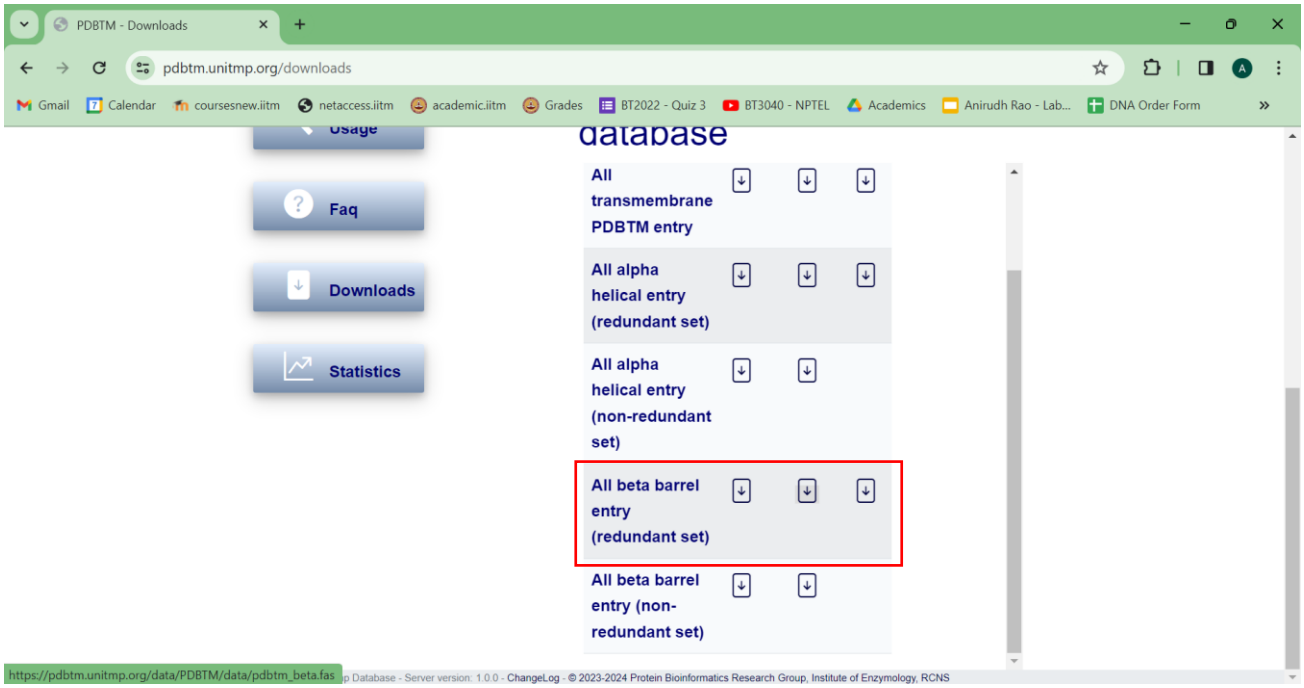
```
print("")

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:



2

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

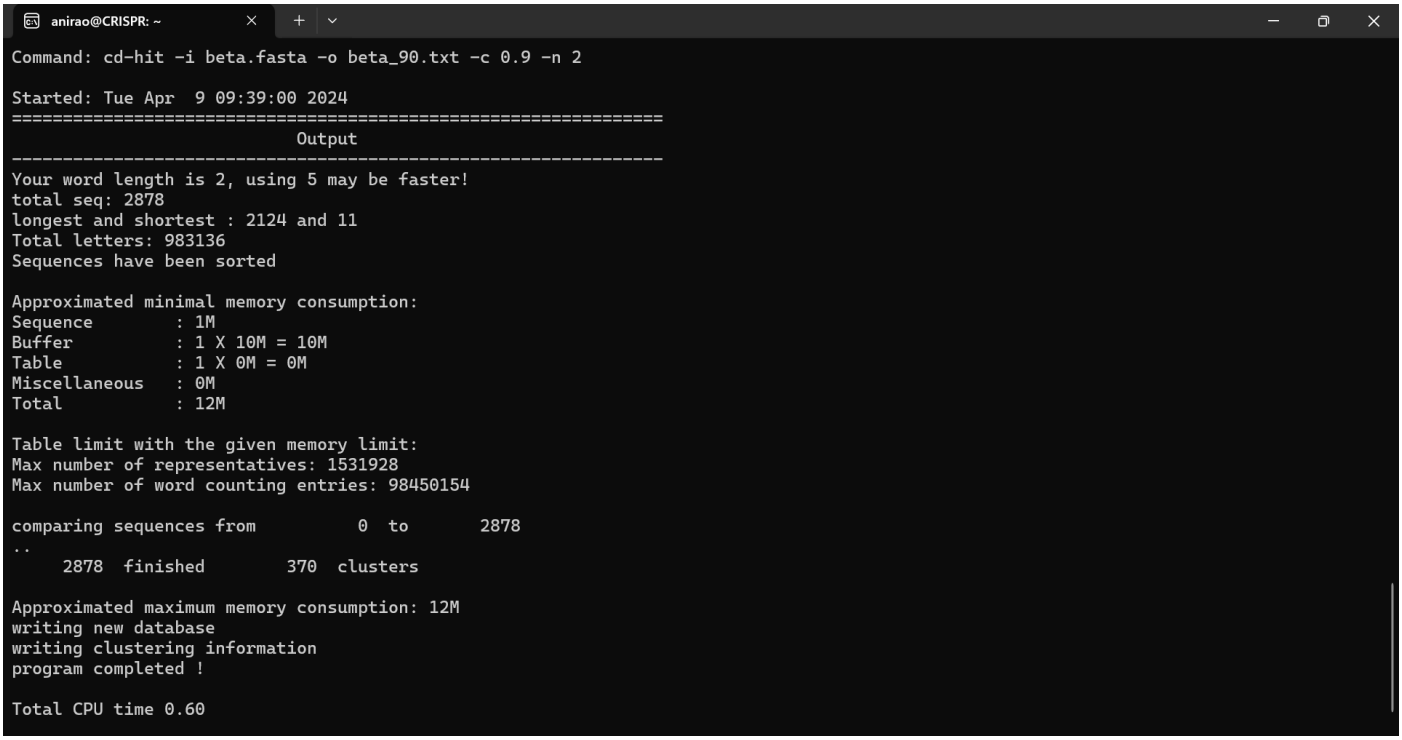


Next, CD-HIT was installed in WSL by running the command:

```
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
```

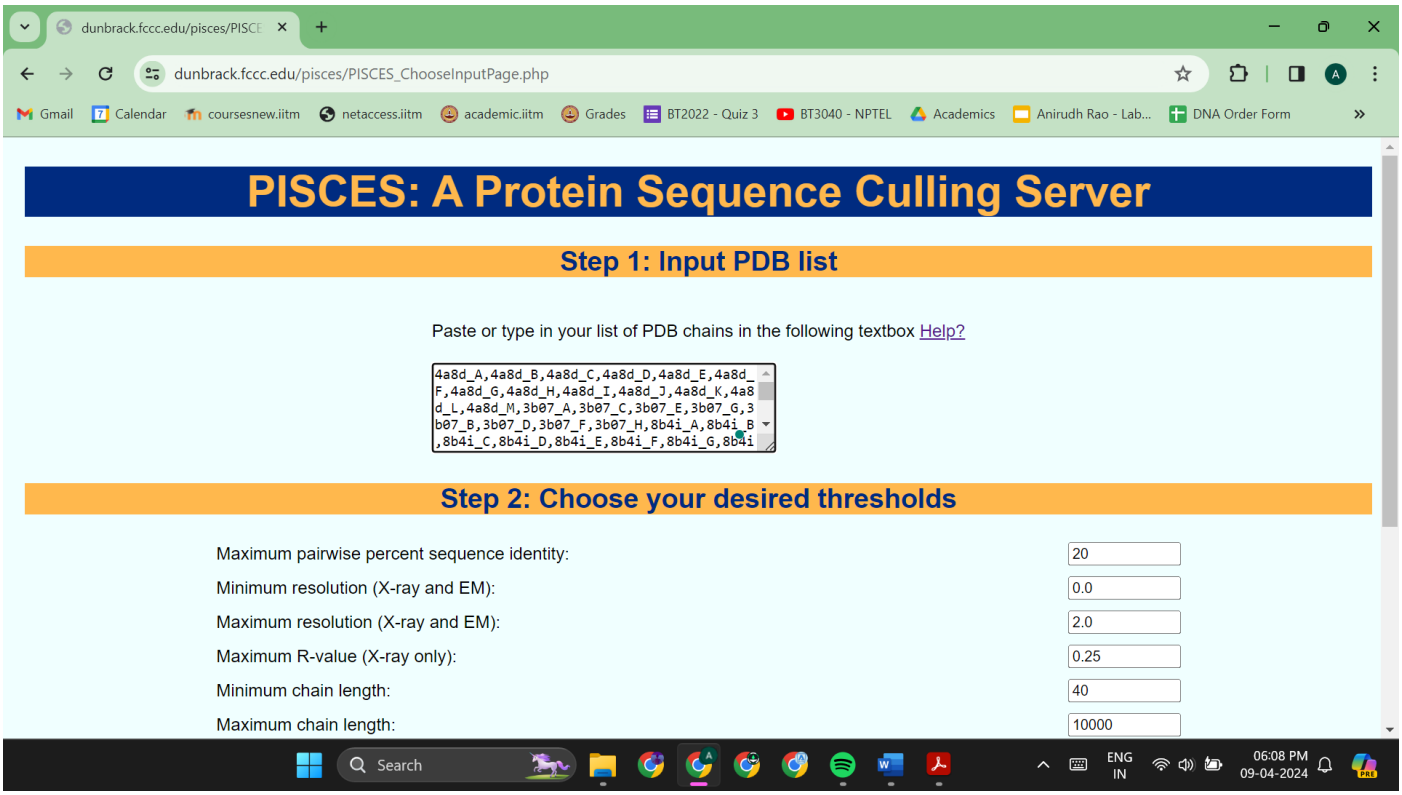


```
anirao@CRISPR: ~

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
```

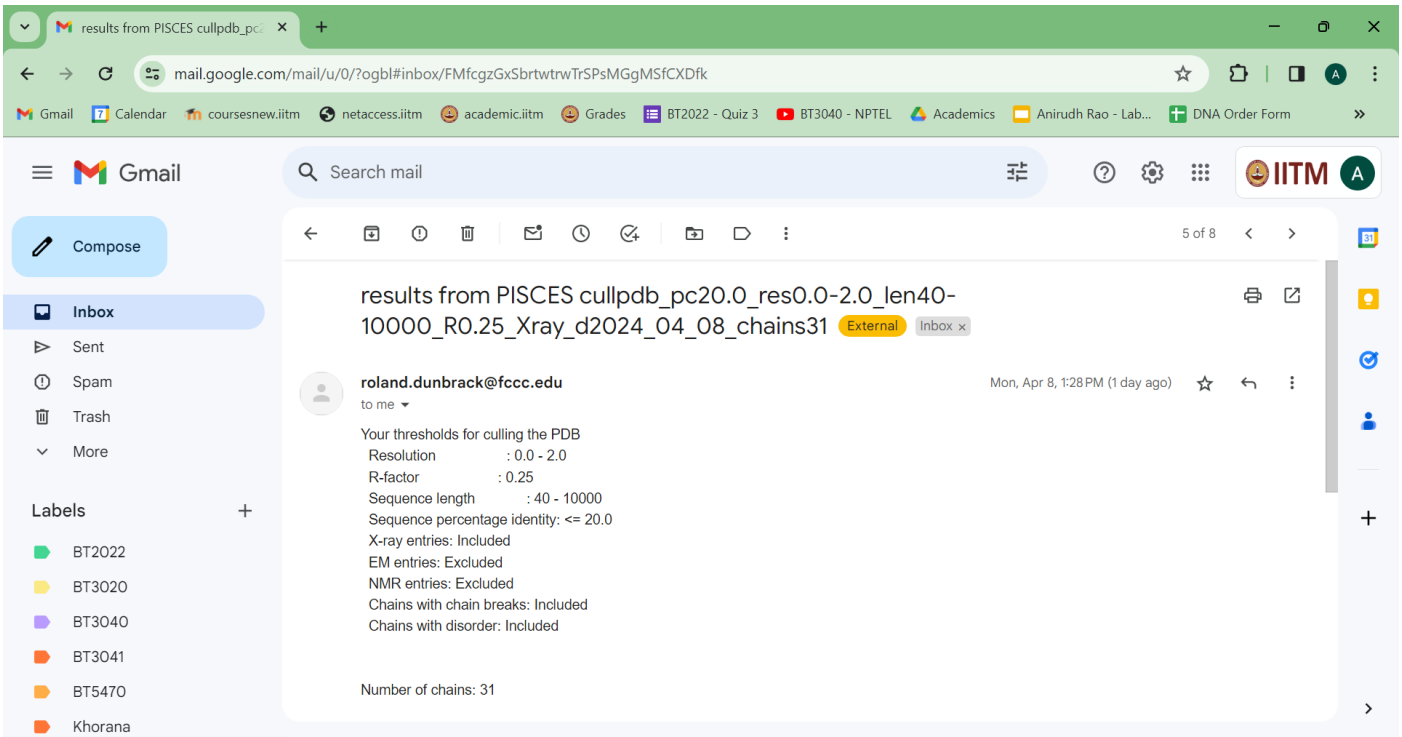
The results from CD-HIT are:

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.



The results were obtained via email.



The results from PISCES are:

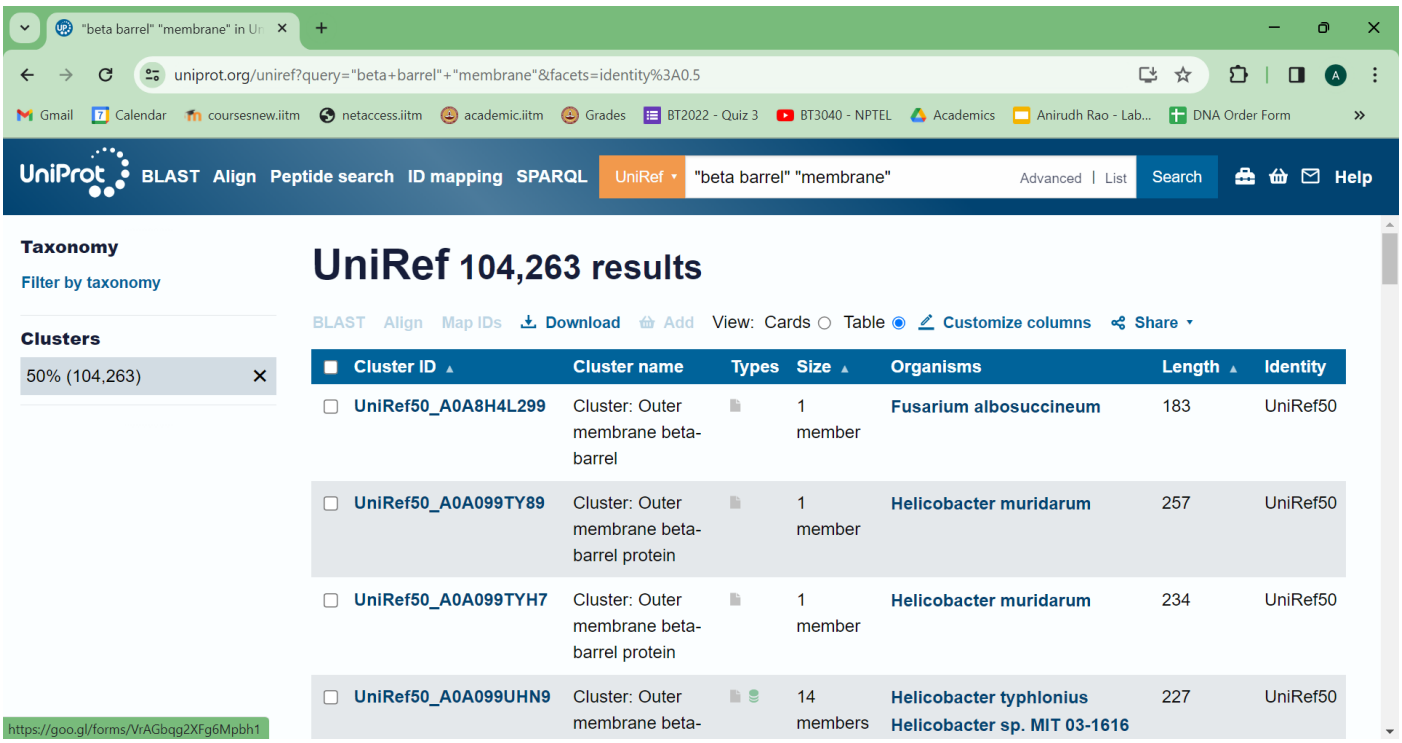
Identity	Number of Sequences
20%	31
30%	39
40%	42
50%	46

4

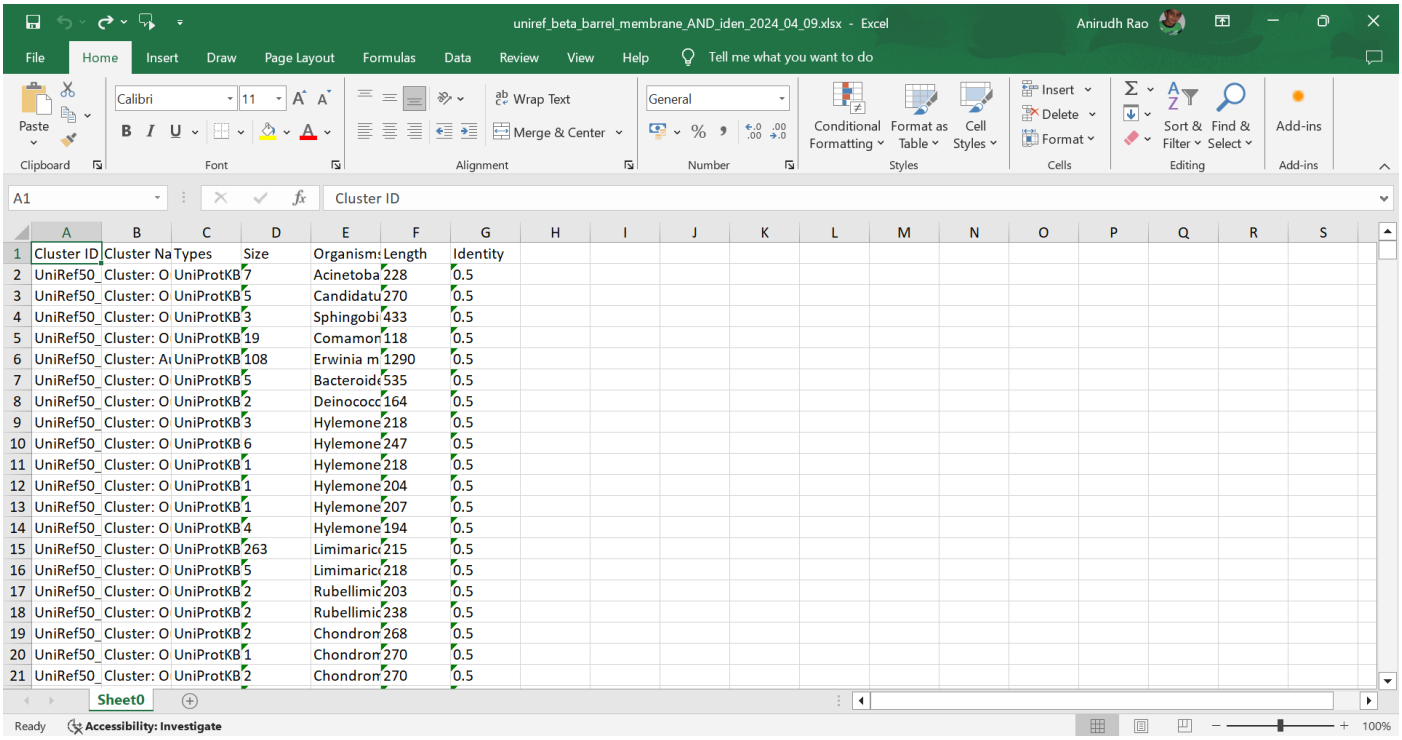
- In CD-HIT, the number of clusters increases as the cutoff is changed from 40% to 50%.
- In PISCES, the number of sequences increases as the cutoff is changed from 40% to 50%.
- In CD-HIT, the size of the largest clusters with identity 40% is 322 and with identity 50% is 318.
- In PISCES, the new sequences 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 results were downloaded in .xlsx format.



Cluster ID	Cluster Na	Types	Size	Organisms	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	Limimaric	215	0.5
UniRef50_	Cluster: O	UniProtKB	5	Limimaric	218	0.5
UniRef50_	Cluster: O	UniProtKB	2	Rubellimic	203	0.5
UniRef50_	Cluster: O	UniProtKB	2	Rubellimic	238	0.5
UniRef50_	Cluster: O	UniProtKB	2	Chondrom	268	0.5
UniRef50_	Cluster: O	UniProtKB	1	Chondrom	270	0.5
UniRef50_	Cluster: O	UniProtKB	2	Chondrom	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.