Anirudh Rao
BE21B004

BT4110 – Computational Biology Lab

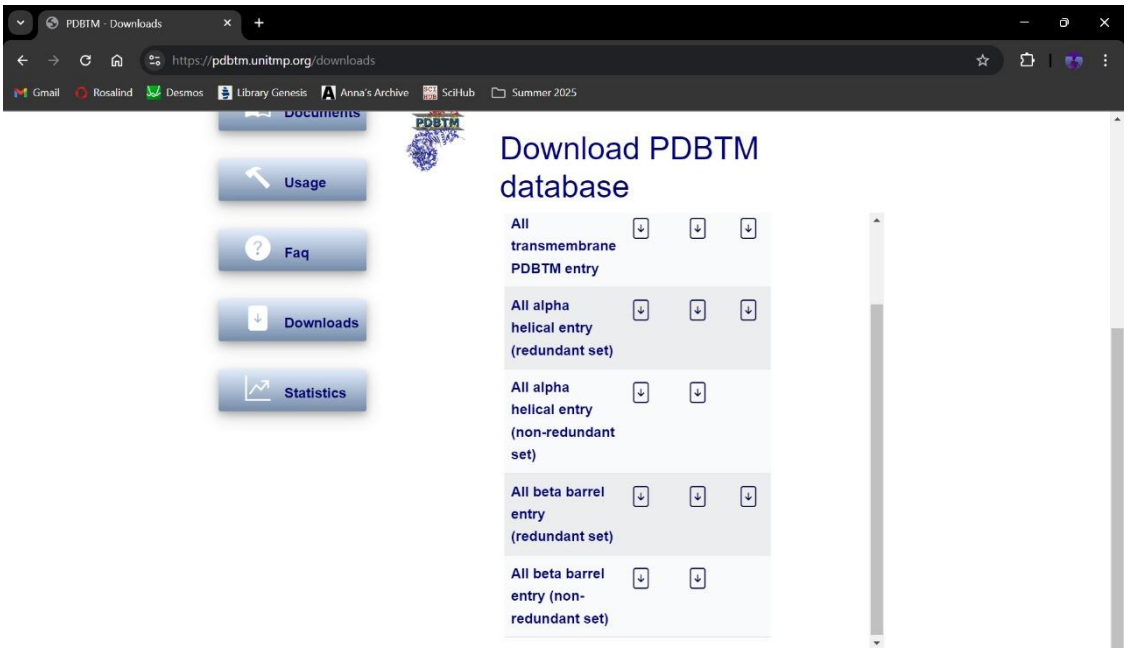
Practical 1

1 to 4

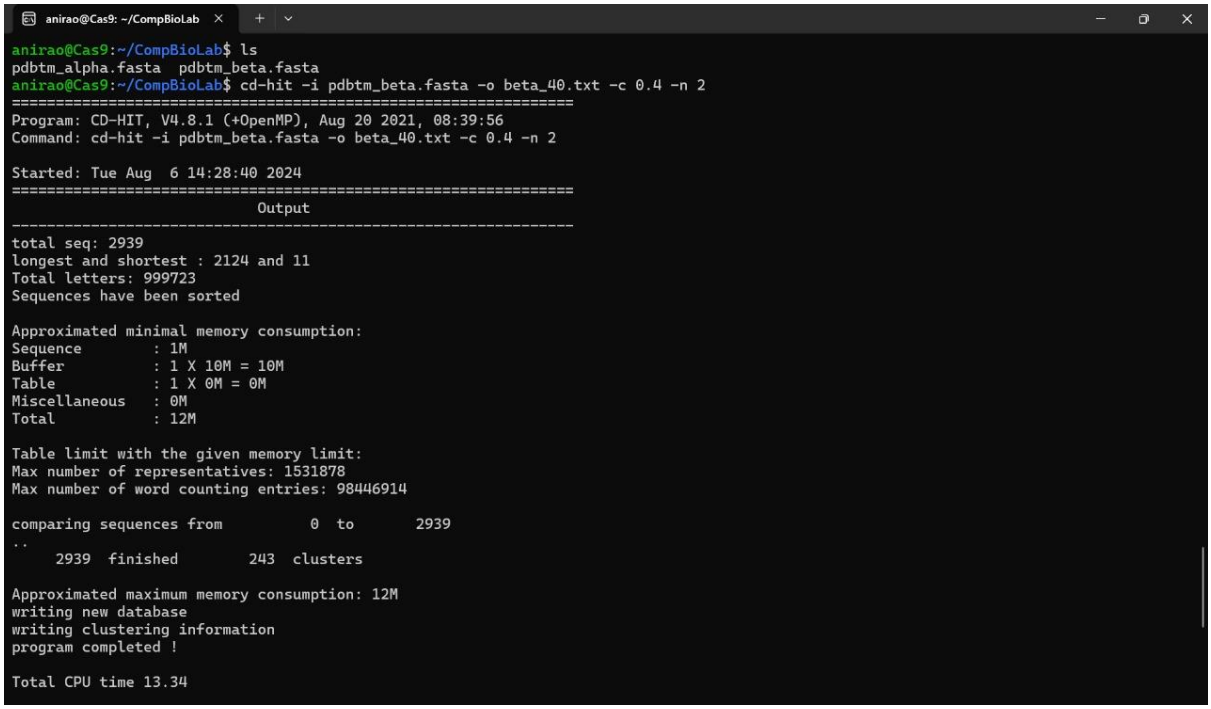
The required sequences were downloaded from PDBTM and were converted into a non-redundant set of sequences using CD-HIT by running the commands

```
cd-hit -i pdbtm_alpha.fasta -o alpha_40.txt -c 0.4 -n 2
```

```
cd-hit -i pdbtm_beta.fasta -o beta_40.txt -c 0.4 -n 2
```




```
anirao@Cas9: ~/CompBioLab

Command: cd-hit -i pdbtm_alpha.fasta -o alpha_40.txt -c 0.4 -n
         2

Started: Tue Aug  6 14:29:06 2024
============================================================
                           Output
------------------------------------------------------------
total seq: 63910
longest and shortest : 3453 and 11
Total letters: 17486153
Sequences have been sorted

Approximated minimal memory consumption:
Sequence        : 25M
Buffer          : 1 X 11M = 11M
Table           : 1 X 1M = 1M
Miscellaneous   : 0M
Total           : 39M

Table limit with the given memory limit:
Max number of representatives: 1502441
Max number of word counting entries: 95114510

comparing sequences from          0  to      63910
.....     10000  finished       499  clusters
.....     20000  finished       895  clusters
.....     30000  finished      1309  clusters
.....     40000  finished      1687  clusters
.....     50000  finished      2057  clusters
.....     60000  finished      2340  clusters
...
    63910  finished      2353  clusters

Approximated maximum memory consumption: 42M
writing new database
writing clustering information
program completed !
```



5

A Python code was used to obtain the overall amino acid composition in TMH and TMB.

Amino acid	TMH (alpha)	TMB (beta)
A	0.084248416	0.07577245
C	0.014871707	0.006502623
D	0.04041765	0.064329523
E	0.048175789	0.048041296
F	0.056482718	0.042362057
G	0.069296461	0.089411069
H	0.019150721	0.015348724
I	0.068706252	0.044526079
K	0.045452545	0.050321436
L	0.118444282	0.081325015
M	0.026793665	0.016763256
N	0.036992886	0.061321004
P	0.042514445	0.038065681
Q	0.033742855	0.04335434
R	0.046282203	0.049666952
S	0.06226314	0.07594135
T	0.054149064	0.066641332
V	0.077431246	0.06391783
W	0.018340478	0.017217173
Y	0.036243477	0.04917081

6

A Python script was used to compute the Fisher discriminant ratio for each of the 20 amino acids based on the composition of the two groups.

Amino acid	FDR
D	0.631626467
N	0.561164032
L	0.47762507
I	0.435399278
M	0.259043878
S	0.208341644
G	0.182661386
Q	0.17027019
T	0.15931243
Y	0.157397886
F	0.155993115
C	0.15088709
V	0.116224848
H	0.04141087
P	0.028192376
A	0.013990474
R	0.007067958
W	0.003626732
E	0.000135943
K	0.000105448

Asp (D) was identified as the amino acid with most importance for discrimination.

7 to 9

A Python script was used to perform the discrimination and report the performance. The final results were obtained as:

TP: 1884

TN: 203

FP: 40

FN: 469

Sensitivity: 0.800679983000425

Specificity: 0.8353909465020576

Accuracy: 0.8039291217257319

The same process was repeated using a 50% train-test split. The split was done randomly after setting a random seed to ensure replicability. The results for this were obtained on the test set as:

TP: 1133

TN: 125

FP: 27

FN: 305

Sensitivity: 0.7878998609179416

Specificity: 0.8223684210526315

Accuracy: 0.7911949685534592

Sensitivity, specificity, and accuracy have all reduced after the train-test split.