BT3040 : Bioinformatics - Practical 8

ADARSH D

BS18B011

1. Refer "**p8_q1.py**" for code.

	Hamming Distance	Euclidean Distance
Sequences 1 & 2	**0.665728476821192**	**0.20106216842153501**
Sequences 2 & 3	**0.726632576075111**	**0.20112952107271115**
Sequences 3 & 1	**0.8433544303797469**	**0.2208681669138957**

It is known that Sequences 1 & 2 are the most close to one another with minimum Hamming and Euclidean distances.

2. Redundant protein sequences obtained from PDB was fed to CD-HIT.

% Identity	No. of non-redundant Clusters
40	134
50	152
75	177
90	189

4. With 40% identity there were 134 clusters, whereas, in case of 50% identity there were 152 clusters. The number of clusters were more for higher cutoffs.

5. When "beta barrel membrane protein" sequences were extracted and fed to CD-HIT, 507 clusters were generated at 50% identity cutoff, whereas, Uniprot with 50% identity cutoff generated 3792 clusters:

https://www.uniprot.org/uniref/?query=uniprot:(beta+barrel+membrane+proteins)+identity:0.5

Uniprot generated more clusters than that of CD-HIT.