Practicals – 9

-BS19B032

-R. Vasantha Kumar

1) I wrote a code to calculate the Hamming and Euclidean distances between the given three sequences. I attached the code with submission.

The results are:

```
Hamming distance between sequences 1 and 2: 0.665728476821192
Hamming distance between sequences 1 and 3: 0.8433544303797469
Hamming distance between sequences 2 and 3: 0.726632576075111

Euclidean distance between sequences 1 and 2: 0.20106216842153501
Euclidean distance between sequences 1 and 3: 0.2208681669138957
Euclidean distance between sequences 2 and 3: 0.20112952107271115
```

From the results, it is clear that sequences 1 and 2 are close to each other, as they have less Hamming and Euclidean distances.

2) I found the non-redundant sequences of beta barrel membrane proteins using CD-HIT. For the beta barrel membrane sequences, I got the sequences from Uniprot, with SWISS-Prot. There was a total of 703 sequences. I have attached the sequences as text files with submission.

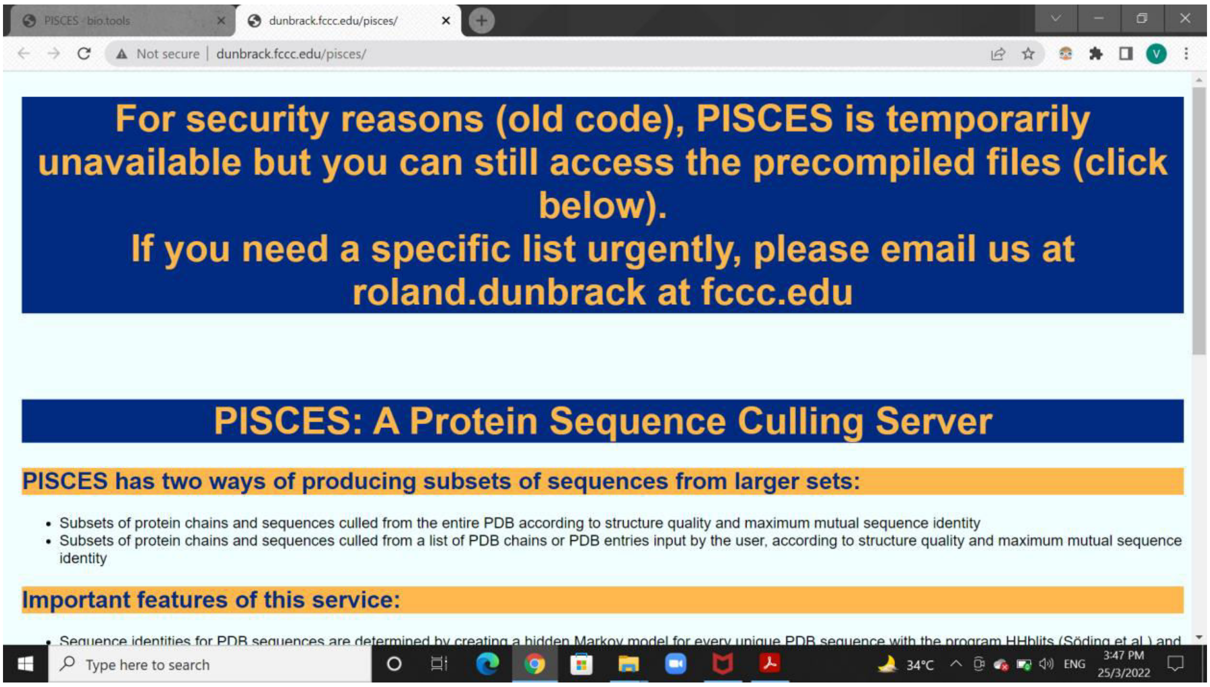
90% - file name – 90%_cd-hit

75% - file name – 75%_cd-hit

50% - file name – 50%_cd-hit

40% - file name – 40%_cd-hit

3) As instructed, since the PISCES server is down, this question cannot be done.


PISCES · bio.tools
dunbrack.fccc.edu/pisces/
Not secure | dunbrack.fccc.edu/pisces/
Type here to search
34°C
ENG
3:47 PM
25/3/2022

For security reasons (old code), PISCES is temporarily unavailable but you can still access the precompiled files (click below).
If you need a specific list urgently, please email us at roland.dunbrack at fccc.edu

PISCES: A Protein Sequence Culling Server

PISCES has two ways of producing subsets of sequences from larger sets:

- Subsets of protein chains and sequences culled from the entire PDB according to structure quality and maximum mutual sequence identity
- Subsets of protein chains and sequences culled from a list of PDB chains or PDB entries input by the user, according to structure quality and maximum mutual sequence identity

Important features of this service:

- Sequence identities for PDB sequences are determined by creating a hidden Markov model for every unique PDB sequence with the program HHblits (Söding et al.) and

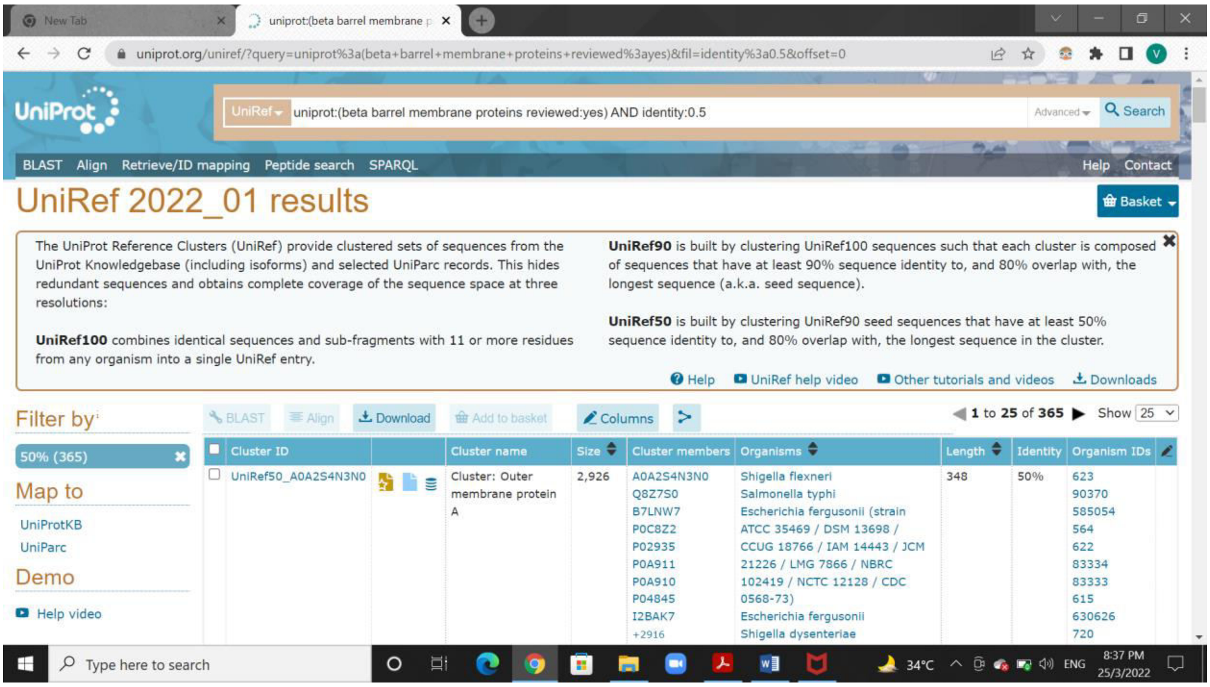
4) As expected, one has to get less number of non-reductant sequence with lower cut off, as threshold decreases, it would be easy to become a reductant sequence.

So, when I found the non-reductant sequences for the given cut off, I got a total of 304 non-reductant sequences for cut off of 50%.

Then, for the cut off of 40%, I got a total of 245 non-reductant sequences.

So, as the cut off decreases, number of non-reductant sequences also decreases.

5) I extracted the data with the cut-off of 50% from Uniprot. I got a total of 365 sequences passing the threshold. But the initial number of sequences was 703. Therefore, total number of non-reductant sequences obtained is 703-365 = 338,i.e.., 338 non-reductant sequences.



uniprot:(beta barrel membrane proteins reviewed:yes) AND identity:0.5

UniRef 2022_01 results

The UniProt Reference Clusters (UniRef) provide clustered sets of sequences from the UniProt Knowledgebase (including isoforms) and selected UniParc records. This hides redundant sequences and obtains complete coverage of the sequence space at three resolutions:

UniRef100 combines identical sequences and sub-fragments with 11 or more residues from any organism into a single UniRef entry.

UniRef90 is built by clustering UniRef100 sequences such that each cluster is composed of sequences that have at least 90% sequence identity to, and 80% overlap with, the longest sequence (a.k.a. seed sequence).

UniRef50 is built by clustering UniRef90 seed sequences that have at least 50% sequence identity to, and 80% overlap with, the longest sequence in the cluster.

1 to **25** of **365**

□	Cluster ID		Cluster name	Size	Cluster members	Organisms	Length	Identity	Organism IDs
□	UniRef50_A0A2S4N3N0		Cluster: Outer membrane protein A	2,926	A0A2S4N3N0 Q8Z7S0 B7LNW7 P0C8Z2 P02935 P0A911 P0A910 P04845 I2BAK7 +2916	Shigella flexneri Salmonella typhi Escherichia fergusonii (strain ATCC 35469 / DSM 13698 / CCUG 18766 / IAM 14443 / JCM 21226 / LMG 7866 / NBRC 102419 / NCTC 12128 / CDC 0568-73) Escherichia fergusonii Shigella dysenteriae	348	50%	623 90370 585054 564 622 83334 83333 615 630626 720

When this is compared with the results obtained via CD-HIT, I got a total a 304 non-reductant sequences. So, there is a slight difference. So, in Uniprot, 34 extra sequences passes the threshold. Therefore, it is safe to say that CD-HIT, algorithm is more efficient.