***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.

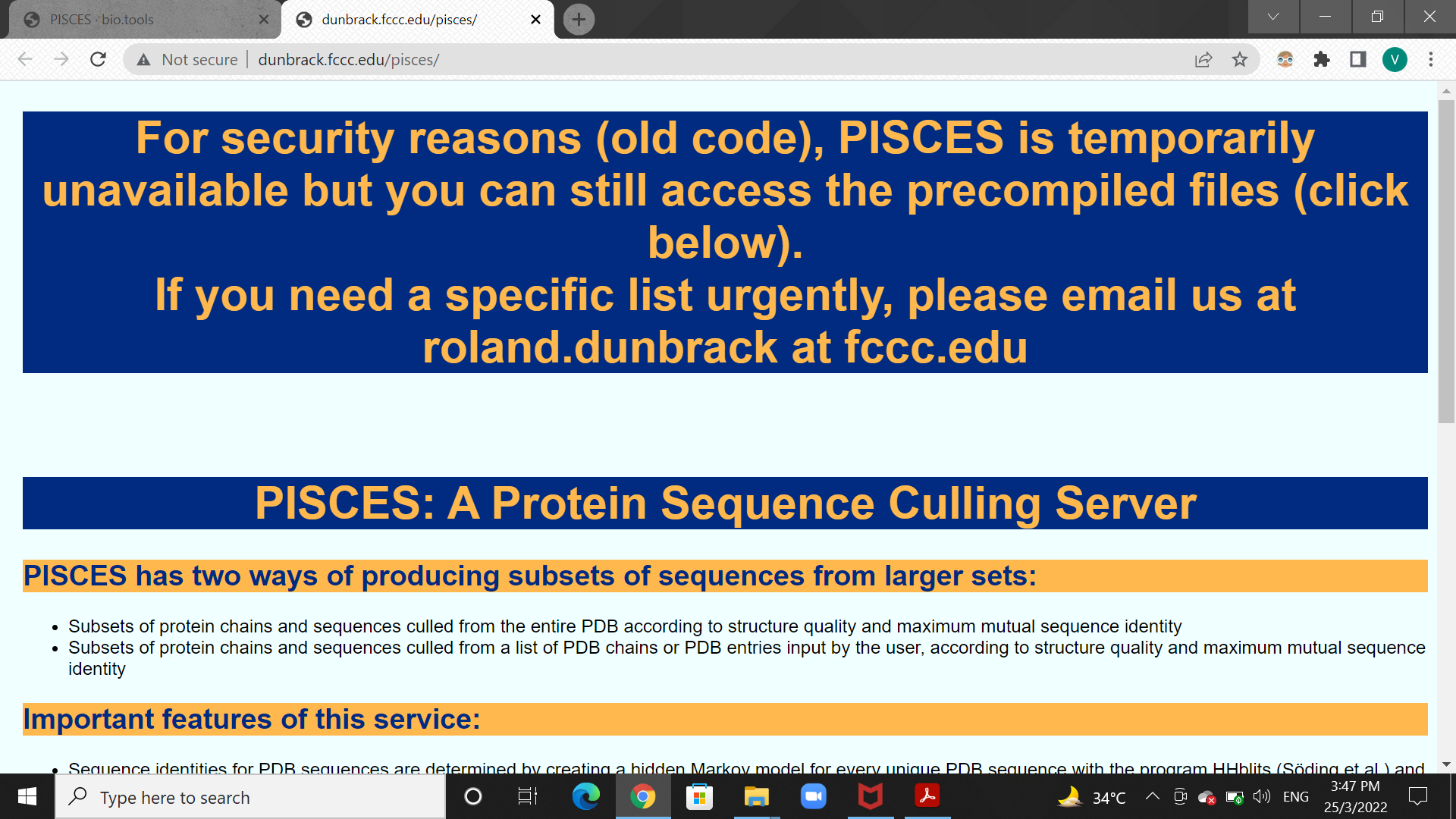
90% - file name – 90%\_cd-hit

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.



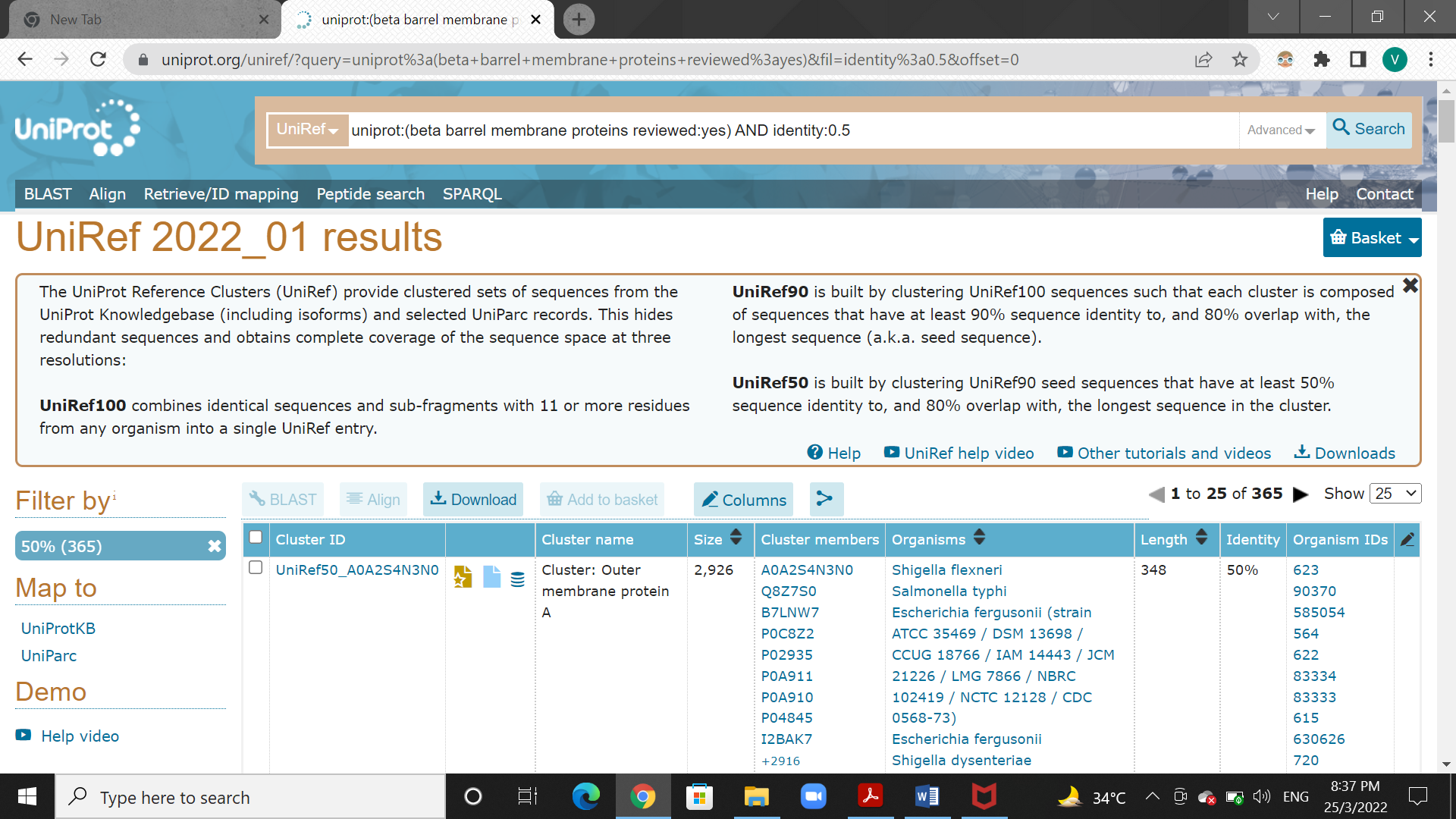
4) As expected, one has to get less number of non-reductant sequence with lower cut off, as threshold decreases, it would be easy to become a reductant sequence.

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.



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.