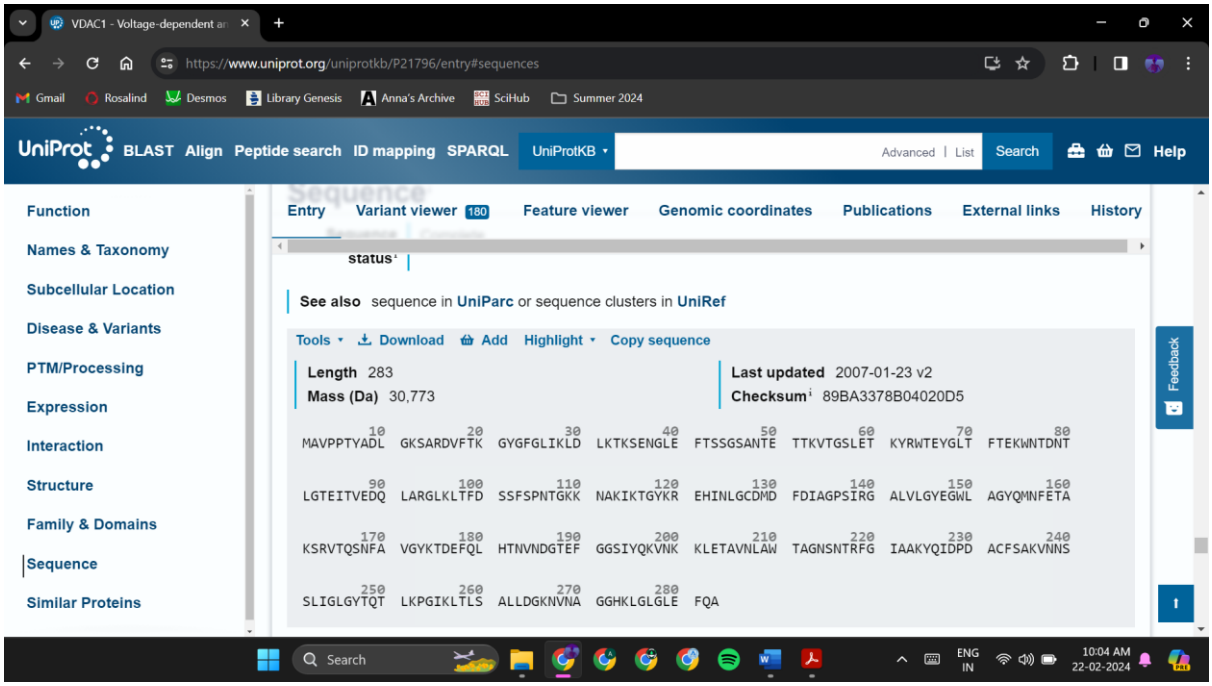
Anirudh Rao
BE21B004

BT3040 – Bioinformatics

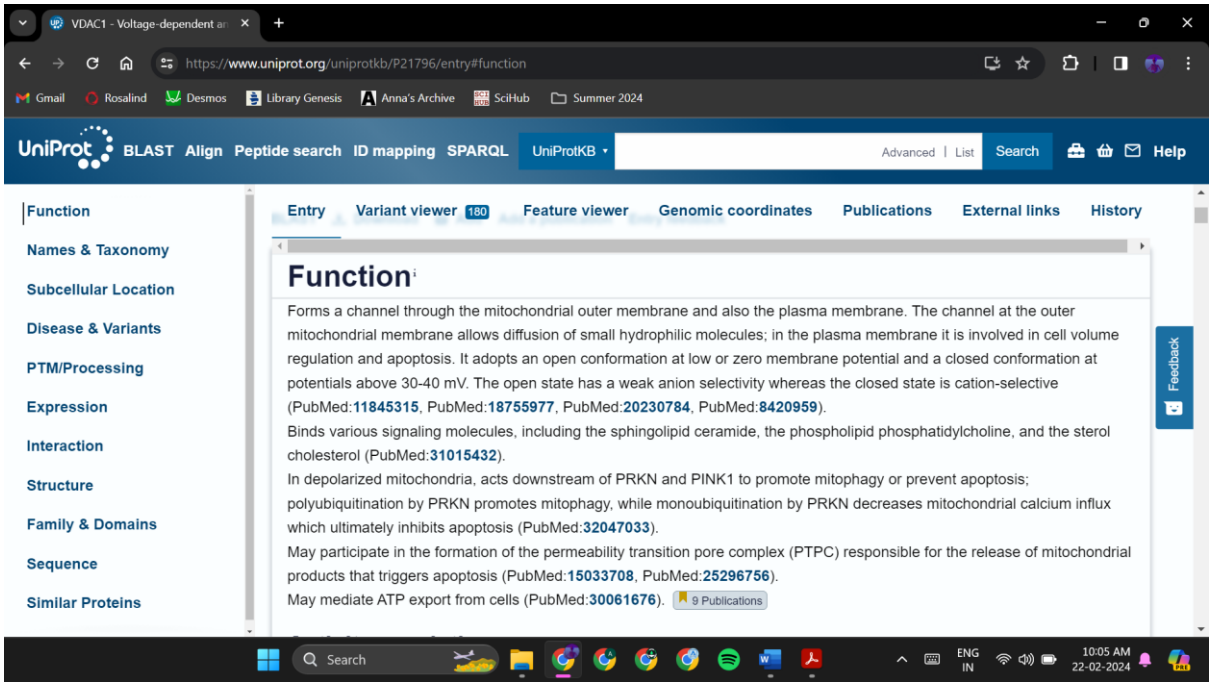
Practical 3

1

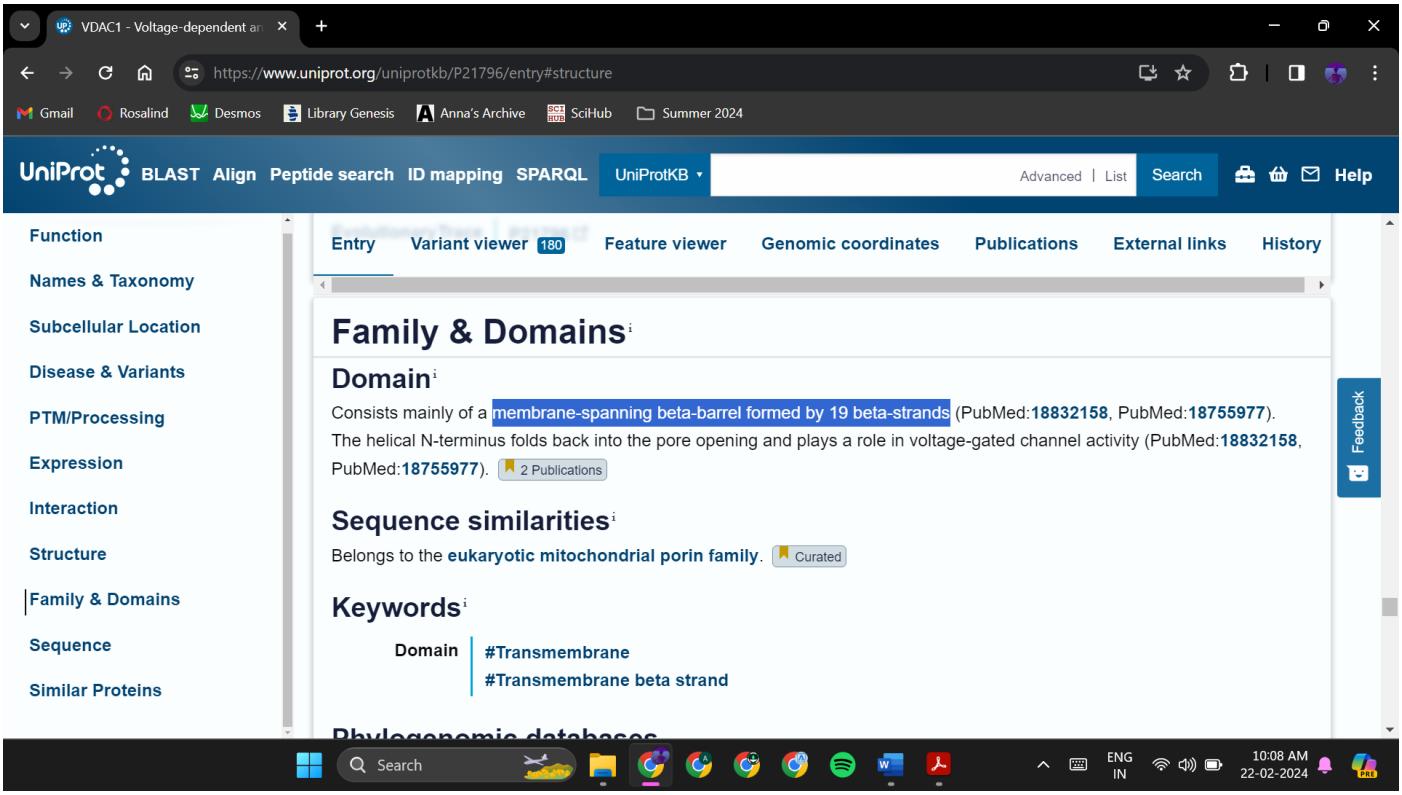
The sequence of the protein is shown below:



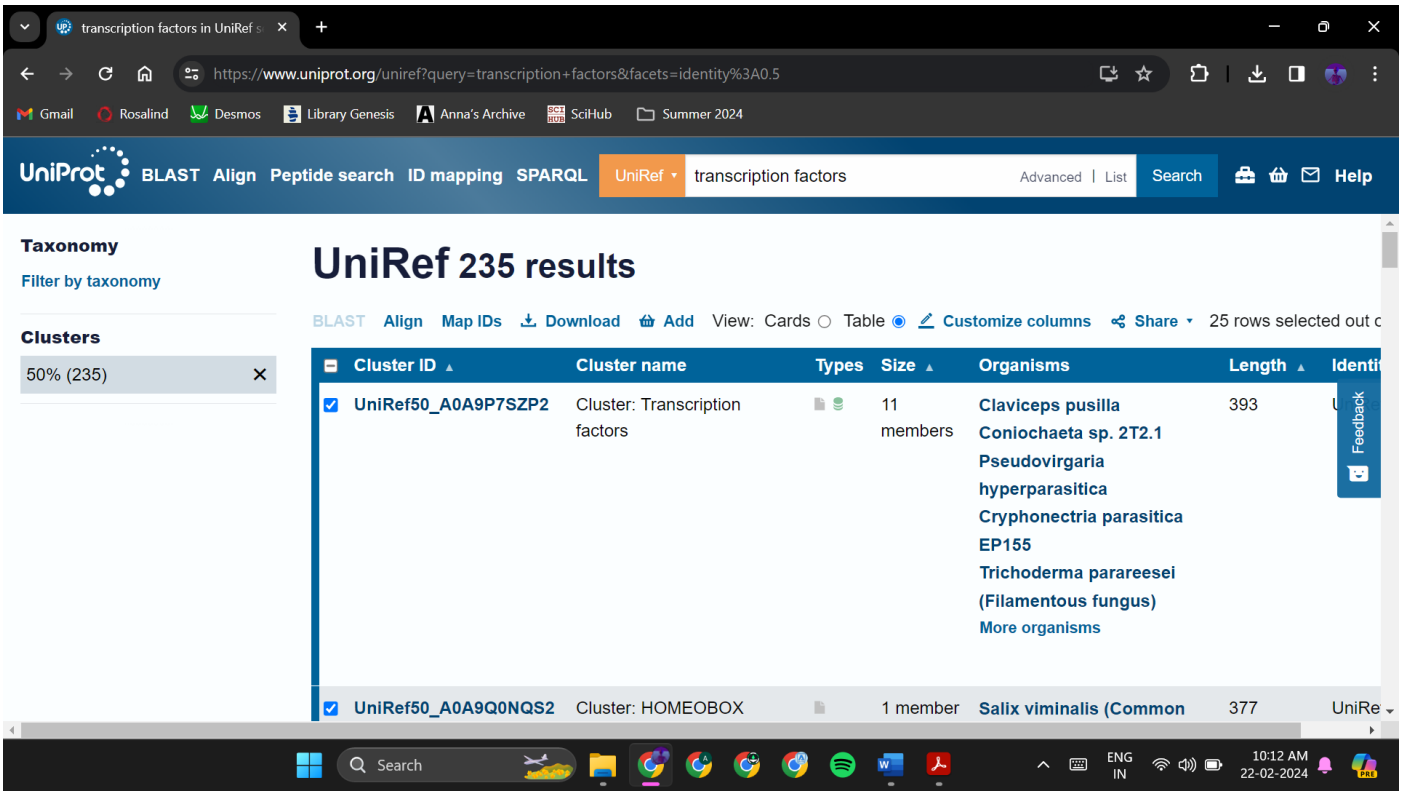
The protein forms a channel through the mitochondrial outer membrane and also the plasma membrane. The channel at the outer mitochondrial membrane allows diffusion of small hydrophilic molecules. The function of the protein is shown below:



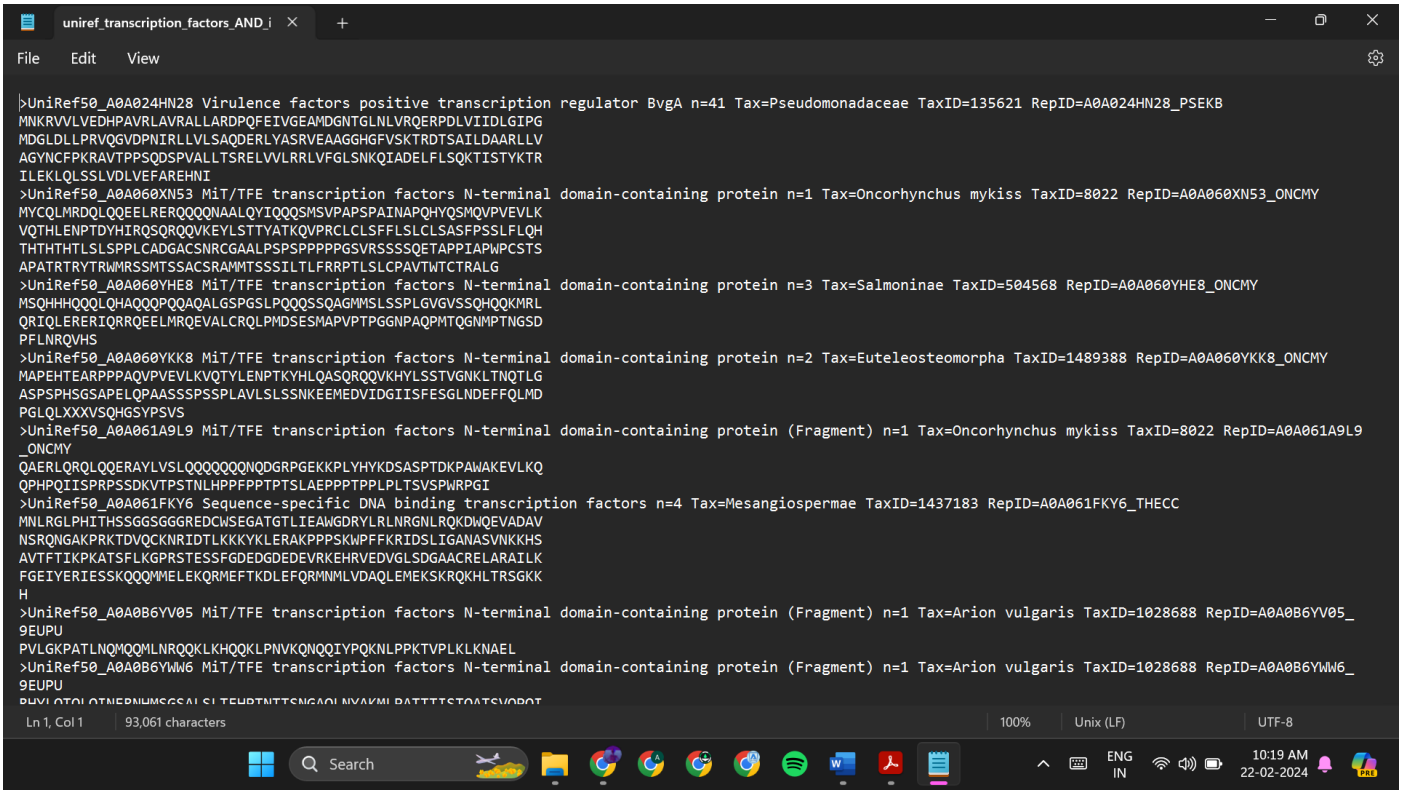
The protein has 19 transmembrane segments.



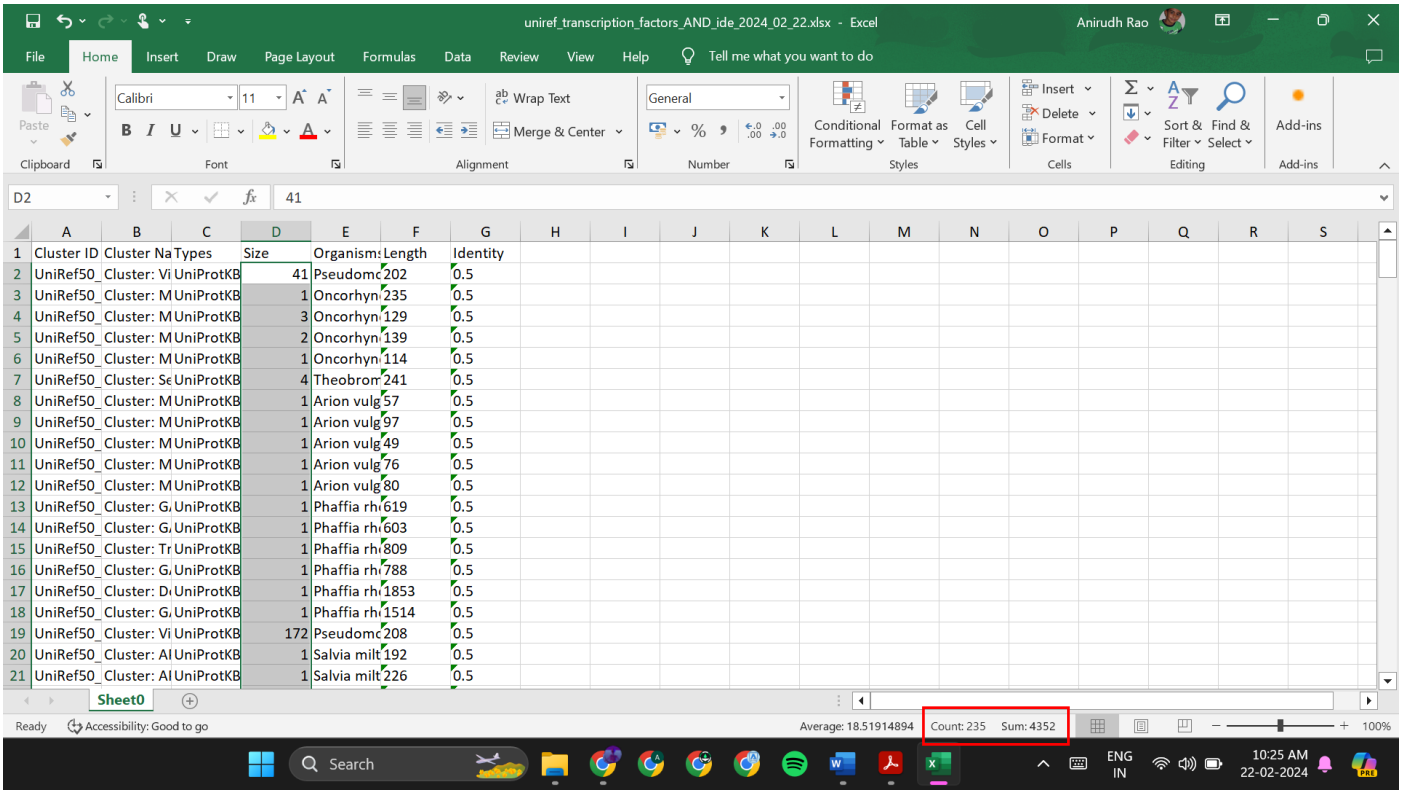
2



These sequences were downloaded in FASTA format.

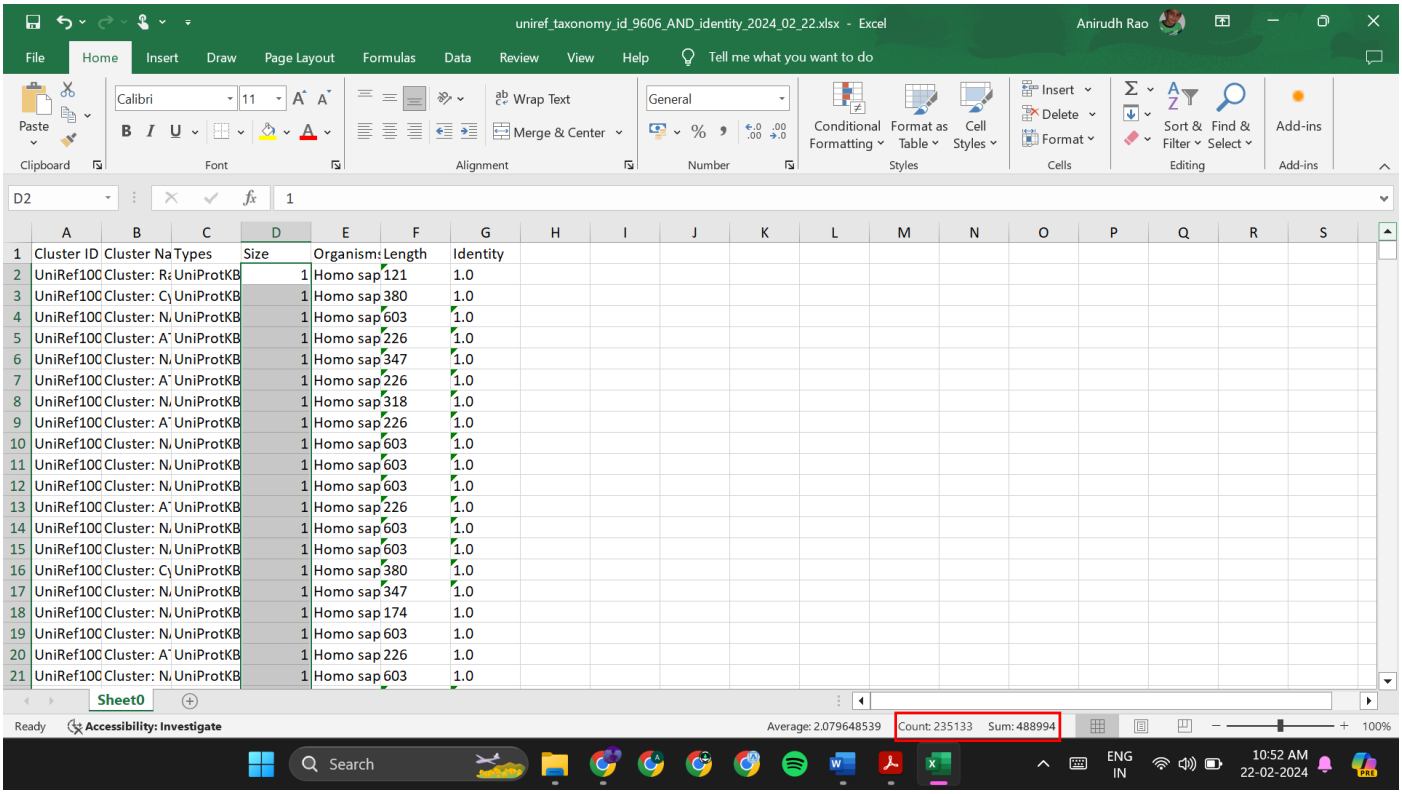


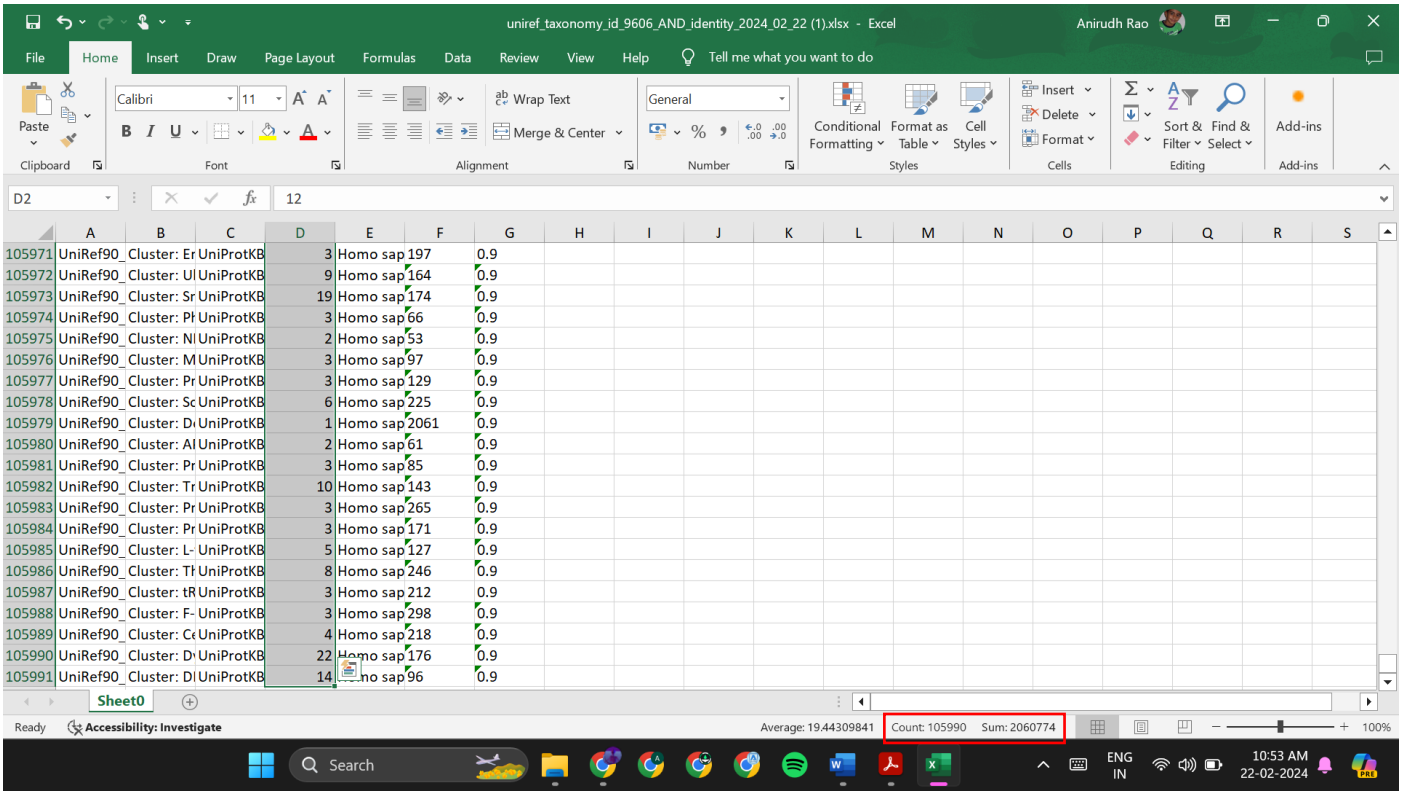
There are 235 clusters and 4352 sequences of transcription factors with 50% sequence identity. This was obtained by downloading the data in .xlsx format.

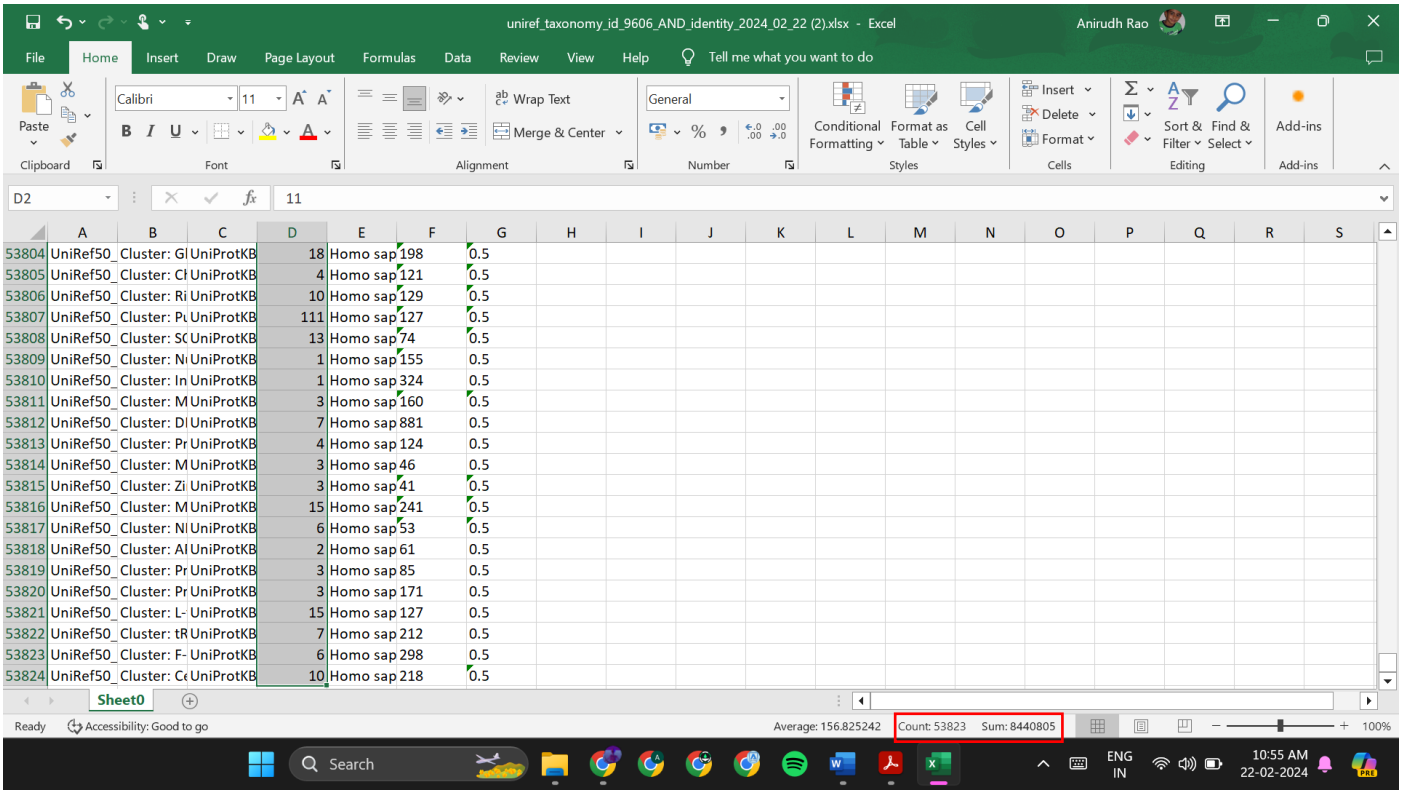


Among all the protein sequences from *Homo sapiens*, there are

- 235,513 clusters with 100% identity with a total of 488,994 sequences.
- 105,990 clusters with 90% identity with a total of 2,060,774 sequences.
- 53,823 clusters with 50% identity with a total of 8,440,805 sequences.

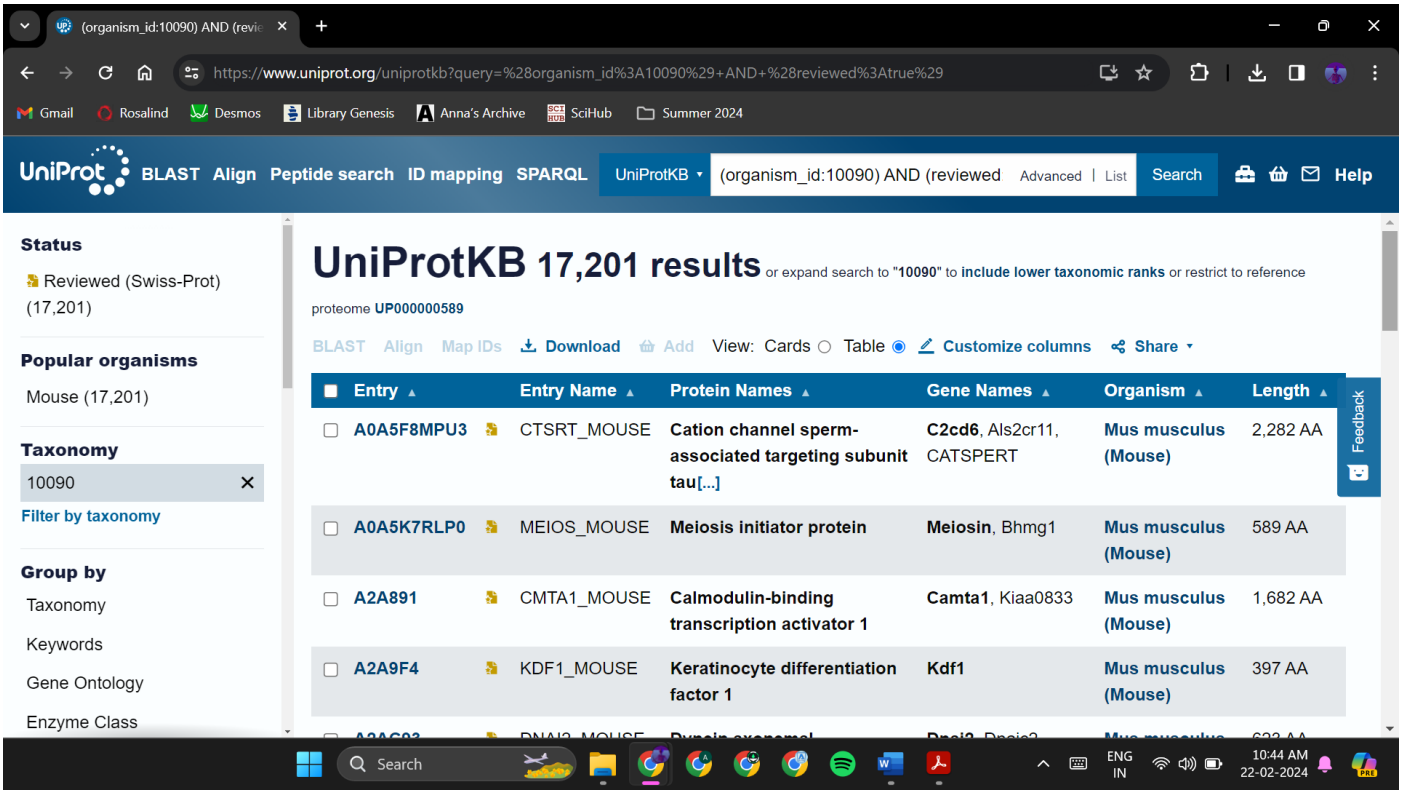




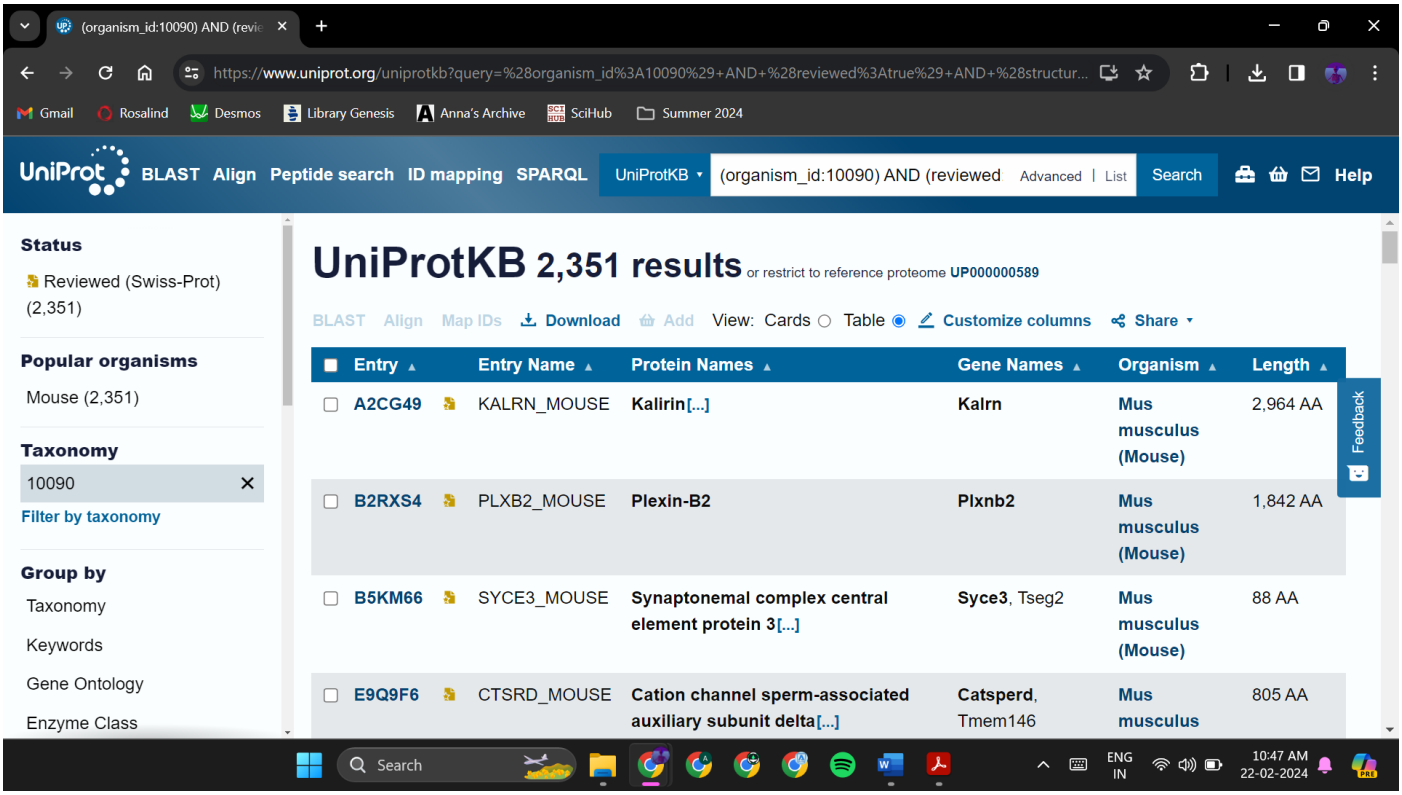


4

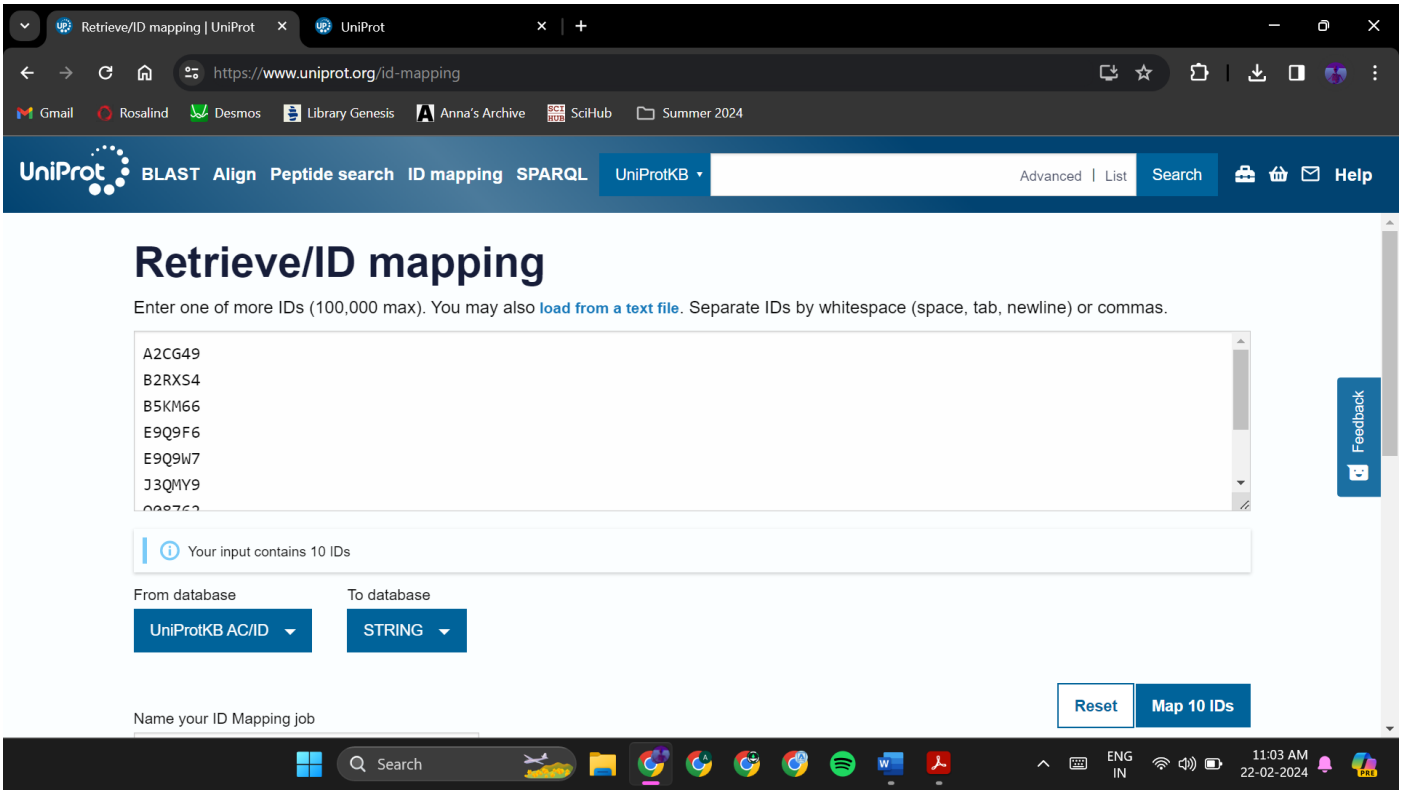
There are 17,201 *Mus musculus* protein sequences that have been manually annotated / reviewed.



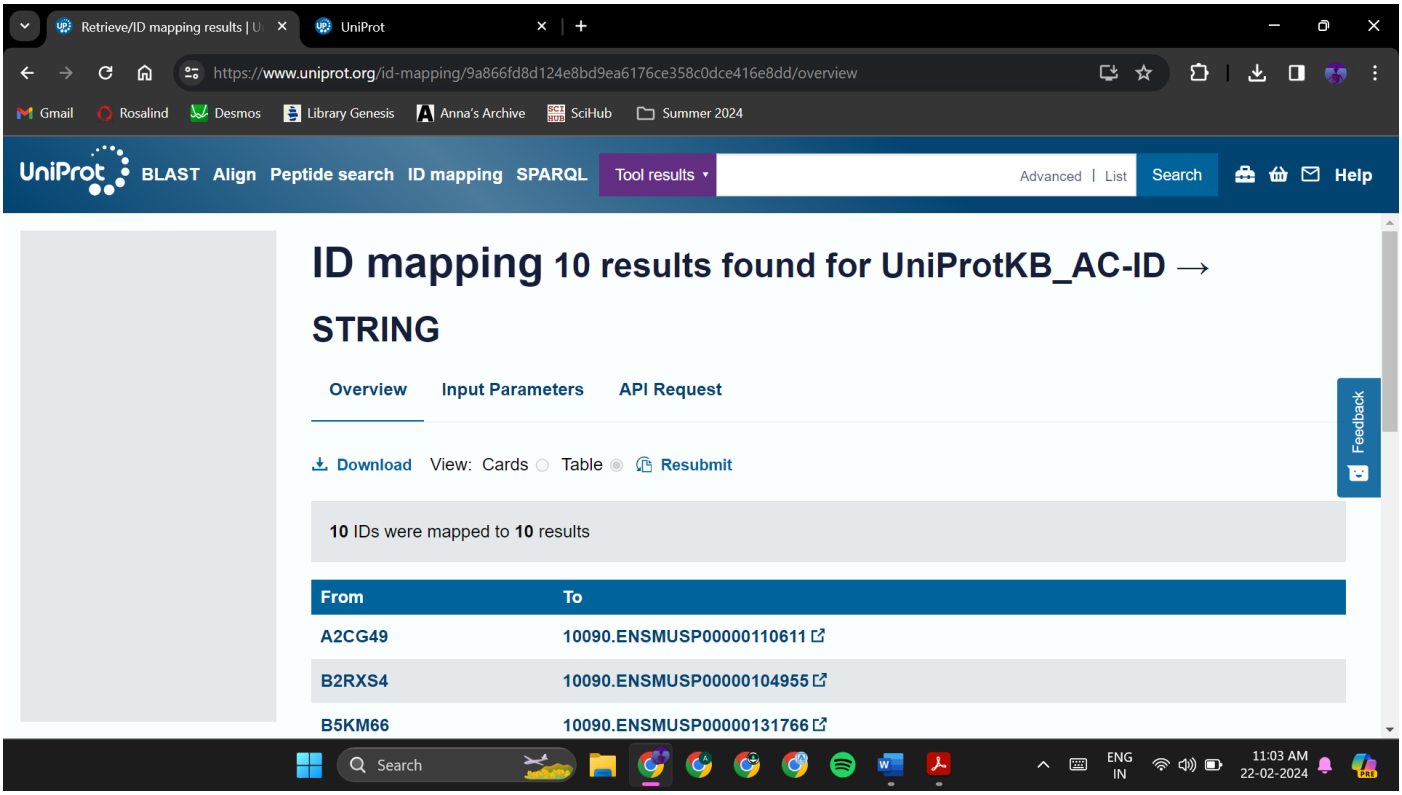
Of these manually annotated sequences, 2,351 are associated with a 3D structure in PDB.



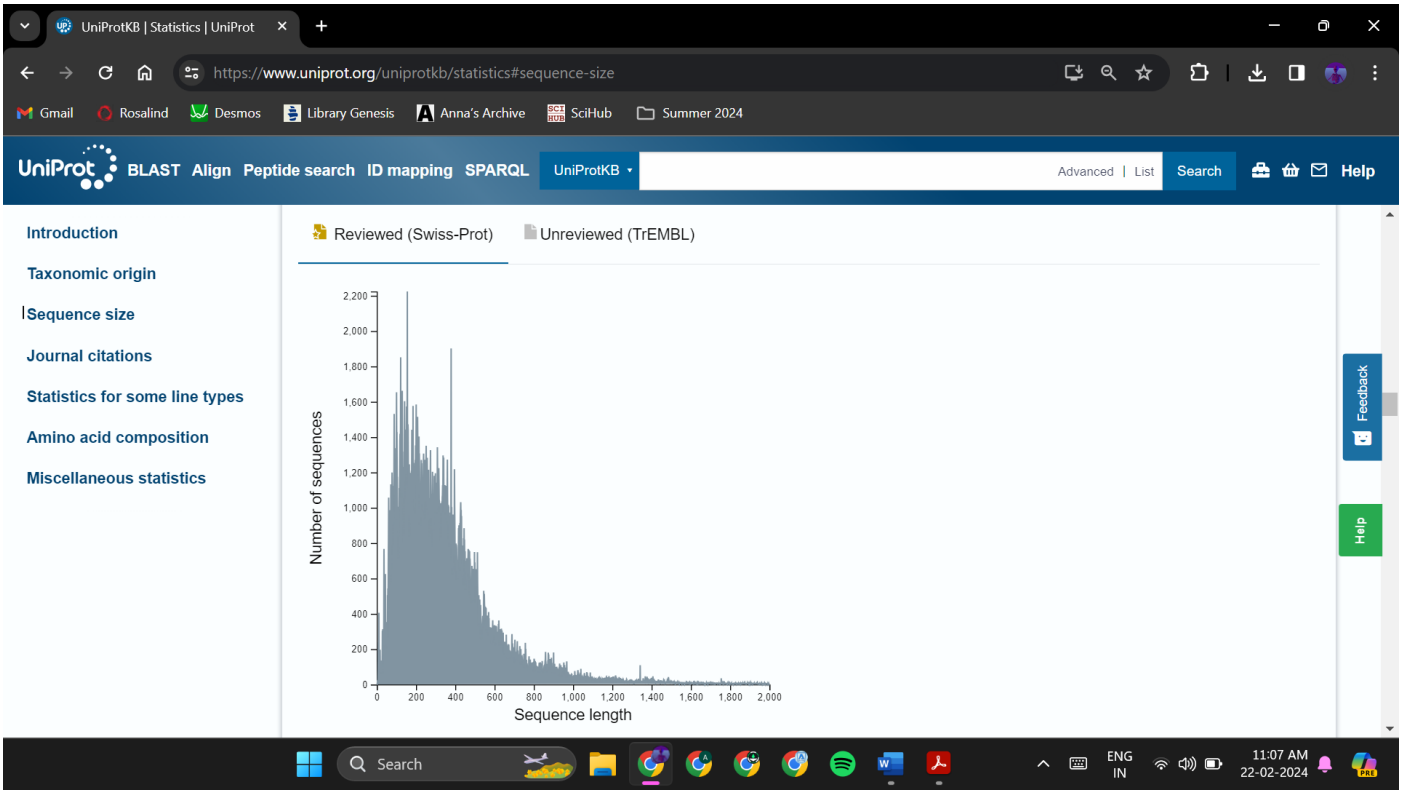
The first 10 entries obtained above were mapped to the STRING database.



10 STRING IDs were mapped.



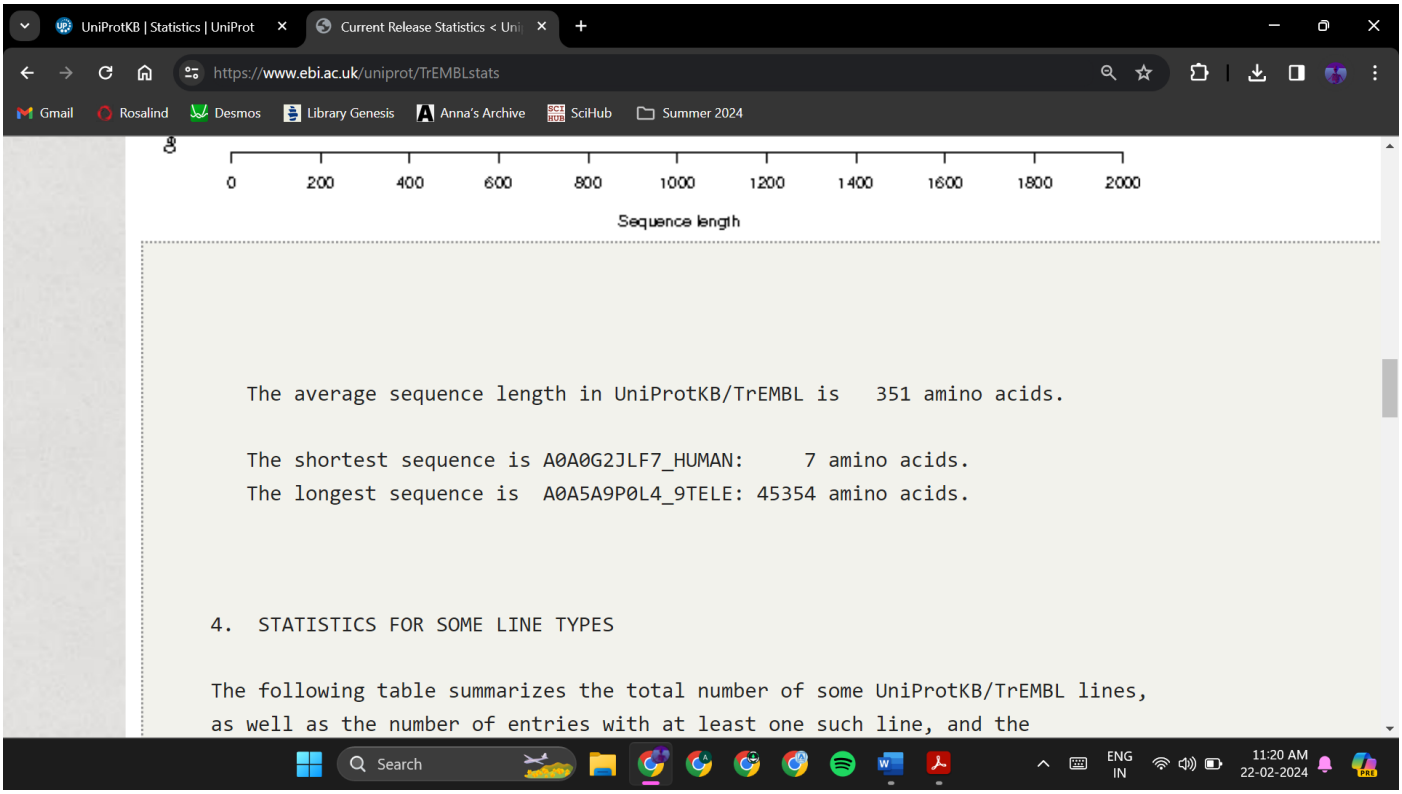
From the UniProt statistics, we can see that most sequences are 200 amino acids long. Longer sequences are less frequently occurring in UniProt.



The average sequence length in UniProtKB/TrEMBL is 351 amino acids.

The shortest sequence is A0A0G2JLF7_HUMAN: 7 amino acids.

The longest sequence is A0A5A9P0L4_9TELE: 45354 amino acids.



The amino acid composition of the UniProt database is shown below:

Amino acid	Count	Percent	Entries with amino acid	Average count per reviewed entry
Leu	19,932,861	9.65%	567,649	34.92
Ala	17,051,081	8.26%	566,853	29.87
Gly	14,609,708	7.07%	567,747	25.59
Val	14,163,491	6.86%	567,085	24.81
Glu	13,880,281	6.72%	562,180	24.32
Ser	13,743,324	6.65%	567,138	24.08
Ile	12,207,933	5.91%	565,003	21.39
Lys	11,983,288	5.80%	563,811	20.99
Arg	11,420,988	5.53%	564,631	20.01
Asp	11,282,060	5.46%	561,315	19.76
Thr	11,076,320	5.36%	565,017	19.40
Pro	9,799,572	4.74%	561,695	17.17
Asn	8,391,346	4.06%	560,439	14.70

Amino acid	Count	Percent	Entries with amino acid	Average count per reviewed entry
Gln	8,121,397	3.93%	557,812	14.23
Phe	7,989,951	3.87%	559,855	14.00
Tyr	6,036,657	2.92%	550,456	10.58
Met	4,983,743	2.41%	564,767	8.73
His	4,706,093	2.28%	539,008	8.24
Cys	2,864,637	1.39%	473,267	5.02
Trp	2,279,505	1.10%	454,465	3.99