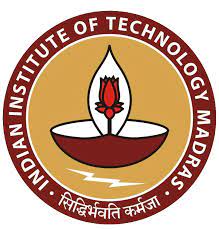
..

**BT 3040: Bioinformatics**

**Assignment 3**



Indian Institute of Technology Madras

Atharva Mandar Phatak | BE21B009

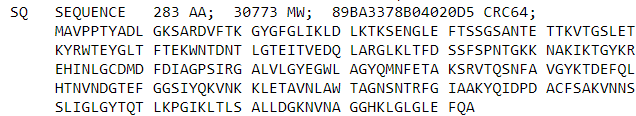
Department of Biotechnology

**Q1)** **Find the amino acid sequence of human mitochondrial β barrel membrane**

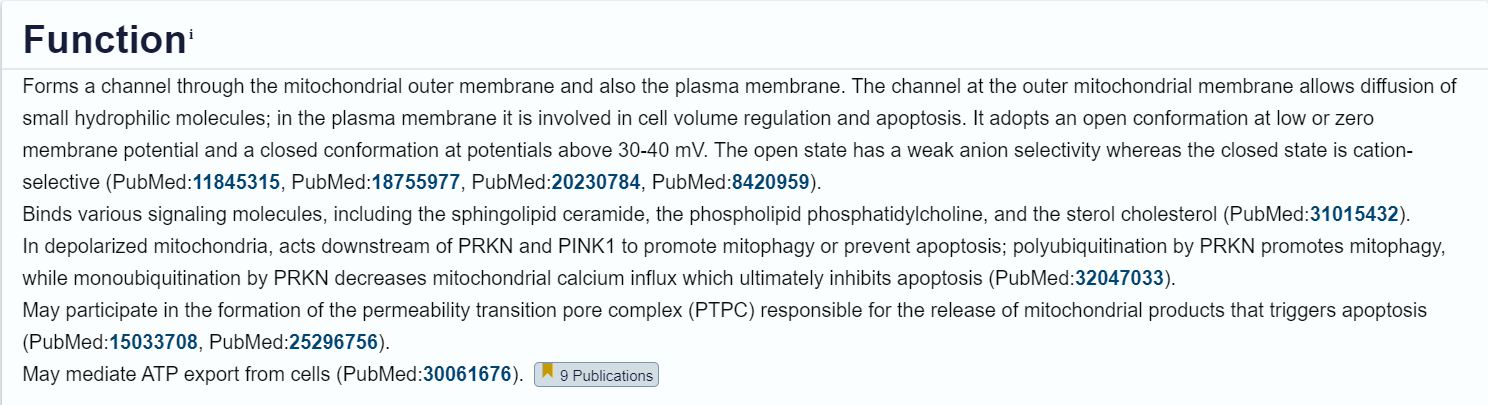
**protein VDAC1 and its function? How many transmembrane segments are**

**present in the protein?**

### A1) Sequence of human mitochondrial beta barallel membrane protein VDAC



### A2) Function:



### B) Number of transmembrane sequence present in the protein = **19**

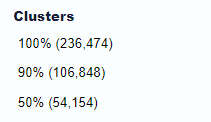


## **Q2) Obtain the sequences of “transcription factors” with 50% sequence identity in FASTA format. List the count of sequences and count of clusters.** [**http://www.uniprot.org/uniprot/**](http://www.uniprot.org/uniprot/)

### 

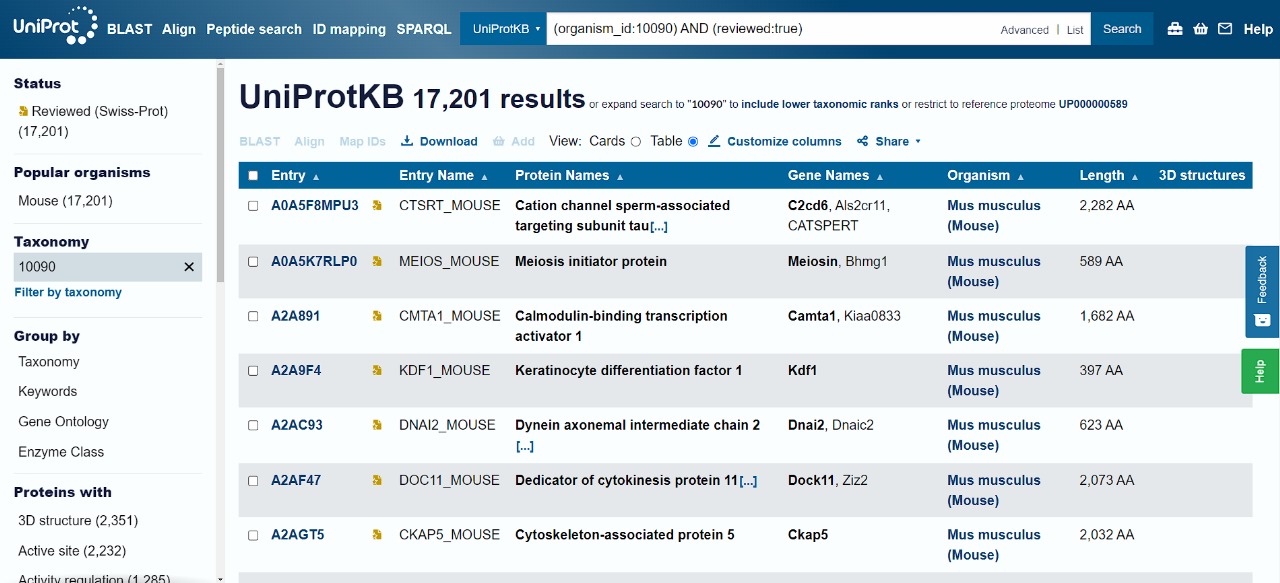
The sequences’ FASTA file is in the folder named ‘Q2\_ transcription\_factors’

## **Q3)** **How many protein sequences from Homo sapiens are obtained at identity cutoff of 100%, 90% and 50% sequence identity?**

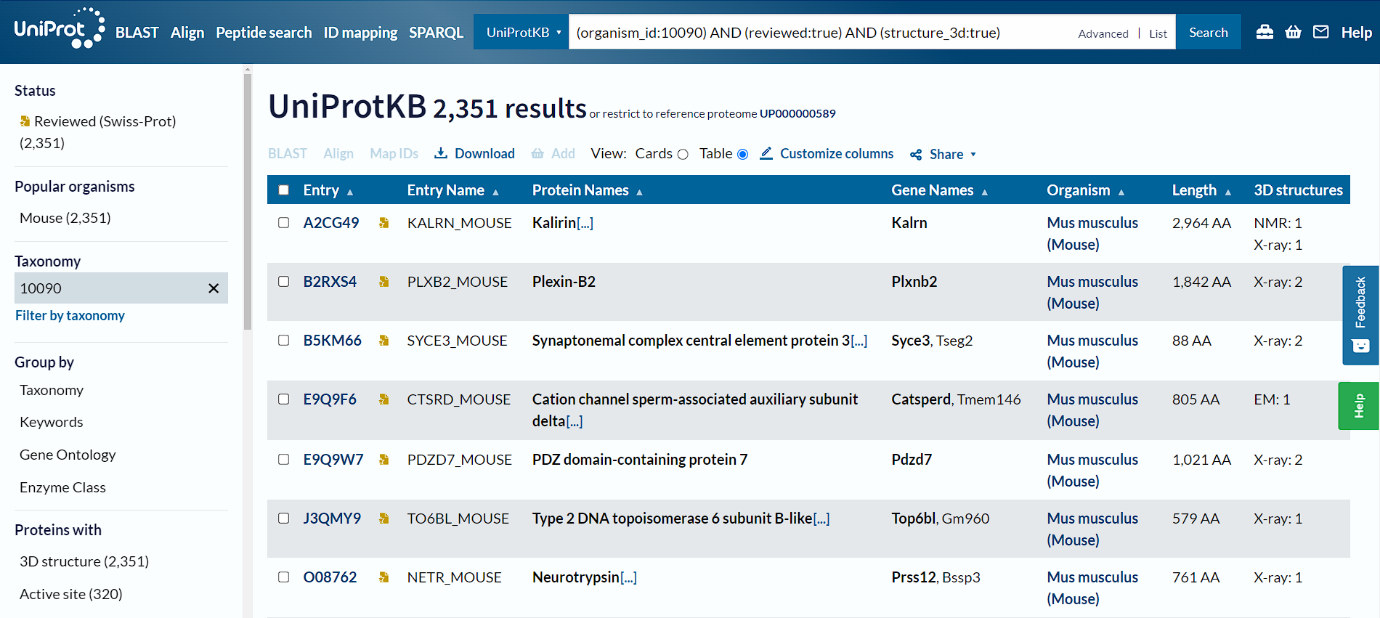


**Q4) In UniProt, how many mouse (Mus musculus) protein sequences are manually annotated? And how many of these manually annotated protein sequences are associated with PDB (3D structures)?**

1. Manually annotated sequences are **17807** of (Mus musculus) protein, under the Swiss-Prot Database.

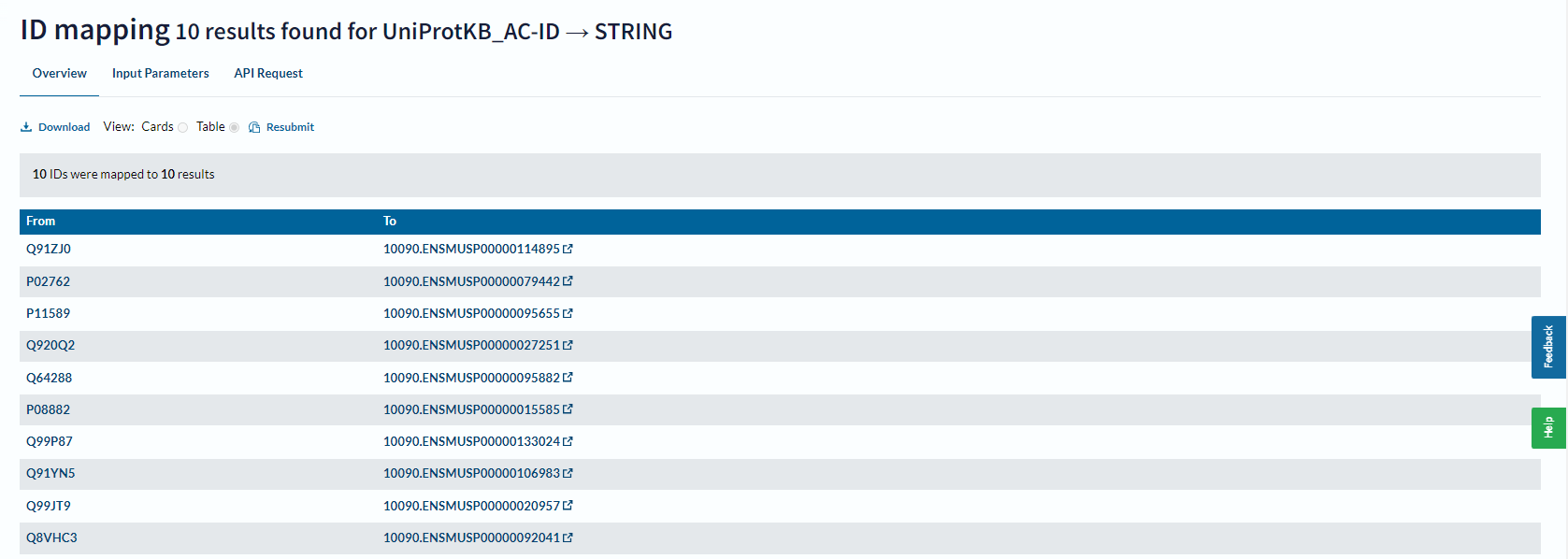


1. There are **2351** manually annotated protein sequences are associated with PDB (3D structures)



## **Q5) Map first 10 UniProt IDs of above manually curated mouse protein sequences with 3D structures to STRING database. How many STRING IDs are**

## **mapped?**



10 STRING IDs are mapped.

## **Q6) Using UniProt Statistics data, answer the following**

## **a) What do you infer from the distribution of sequence length in UniProt?**

## **b) The shortest and longest sequence in UniProtKB**

## **c) Amino acid composition in percent for the complete database**

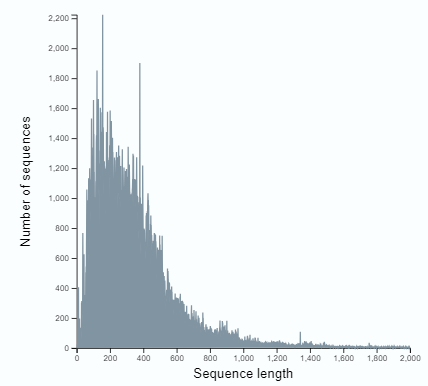
The following link has UniProt Statistics as of Feb 2024: <https://www.uniprot.org/uniprotkb/statistics#statistics-for-some-line-type>

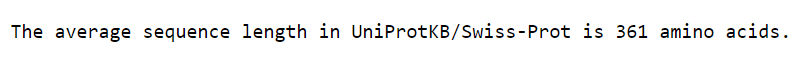
UniProtKB/TrEMBL : <https://www.ebi.ac.uk/uniprot/TrEMBLstats>

UniProtKB/Swiss-Prot: <https://web.expasy.org/docs/relnotes/relstat.html>

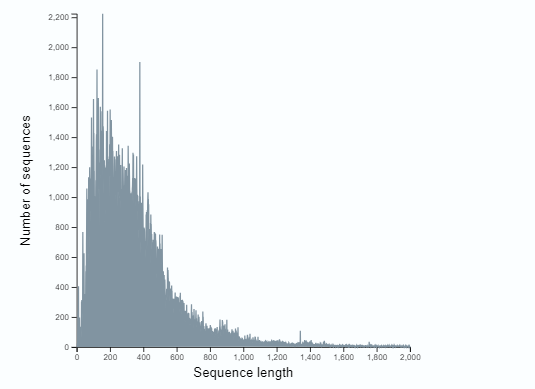
a) The sequence distribution chart plots the Number of sequences vs Sequence Length. We see a peak in the graph at range ~351 (TrEMBL) and ~361 (Swiss-Prot). The distribution is right-skewed, indicating more amino acids lie in the range of 0-600. We can also infer that it indicates the number of amino acids in the canonical sequence displayed by default in the entry's Sequence section.

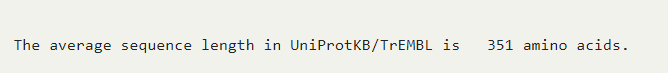
a1) Reviewed (Swiss-Prot)



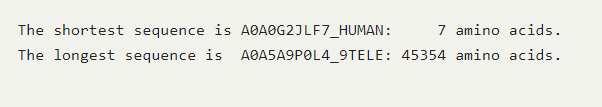


a2) Unreviewed (TrEMBL)

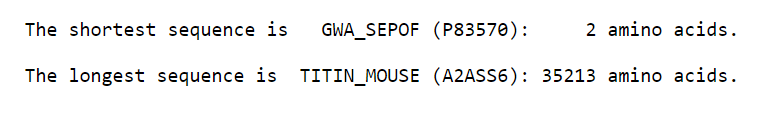




b)

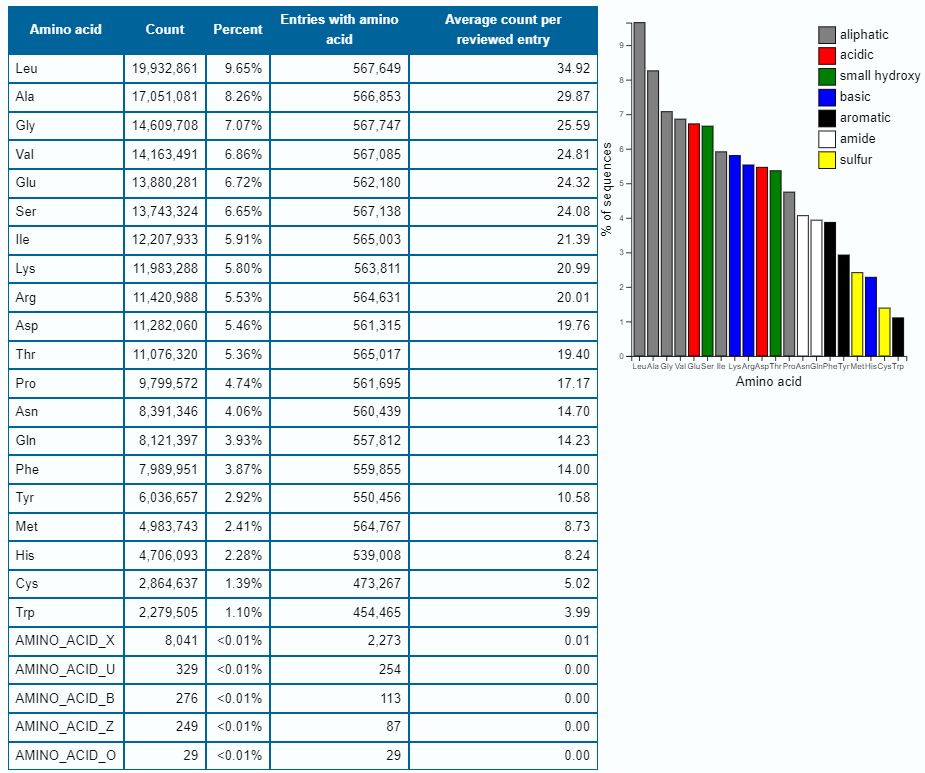
b1) The shortest and longest sequence (in TrEMBL) are as follows:

b2) The shortest and longest sequence (in SwissProt) are as follows:



c) The amino acid composition is as follows:

c1) Reviewed (Swiss-Prot)



c2) Unreviewed (TrEMBL)

