Done by:B Jagadishwaran
BE18B001
Bioinformatics
Practical 3
3 Mar 2021

1. 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?

MAVPPTYADLGKSARDVFTKGYGFGLIKLDLKTKSENGLEFTSSGSANTETTKV
TGSLETKYRWTEYGLTFTEKWNTDNTLGTEITVEDQLARGLKLTFDSSFSPNTG
KKNAKIKTGYKREHINLGCDMDFDIAGPSIRGALVLGYEGWLAGYQMNFETAKS
RVTQSNFAVGYKTDEFQLHTNVNDGTEFGGSIYQKVNKKLETAVNLAWTAGNS
NTRFGIAAKYQIDPDACFSAKVNNSSLIGLGYTQTLKPGIKLTLSALLDGKNVNA
GGHKLGLGLEFQA.

This is the amino acid sequence of β barrel membrane protein VDAC1 and its functions include

- Forms a channel through the mitochondrial outer membrane and also the plasma membrane. This channel let outer mitochondrial membrane for diffusion of small hydrophilic molecules.
- cell volume regulation and apoptosis are taken care by this protein.
- Binds various signaling molecules, including the sphingolipid ceramide, the phospholipid phosphatidylcholine, and the sterol cholesterol
- May participate in the formation of the permeability transition pore complex (PTPC) responsible for the release of mitochondrial products that triggers apoptosis

There are **19** transmembrane segments present in the protein

2. Obtain the sequences of "transcription factors" with less than 50% sequence identity in FASTA format. List the count of sequences and count of clusters.

https://www.uniprot.org/uniref/?query=uniprot:(transcription+factors)+identity:0.5

I have visited the above link and found there are 30,397 50% identity sequences

count of sequences -218606
count of clusters -30397

3. How many clusters of protein sequences from Homo sapiens are obtained at identity cutoff of 100%, 90% and 50% sequence identity?

I have visited the link
--https://www.uniprot.org/uniref/?query=homo+sapiens&sort=score

And I got answer:
Using the search ---uniprot:(organism:homo sapiens) AND identity:0.5/0.9

100% : 143,929
90% : 86,911
50% : 66,881

4. 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)? Hint: Use "Advanced" search

I visited the website
https://www.uniprot.org/uniprot/?query=*&fil=organism%3A%22Mus+musculus+%28Mouse%29+%5B10090%5D%22+AND+reviewed%3Ayes

And found there are--**17,063**

And using the search key-

database:(type:pdb) AND reviewed:yes AND organism:"Mus musculus (Mouse) [10090]"

I found there are **1980** associated with 3D structure

5. Map UniProt IDs of above manually curated mouse protein sequences with 3D structures to STRING database. How many STRING IDs are mapped? Hint: Use "Retrieve/ID mapping"

I have downloaded the fasta file of previous 3D PDB and uploaded in the link **Retrieve/ID mapping** and converted From UniprotKB AC/I toSTRING AND got the result

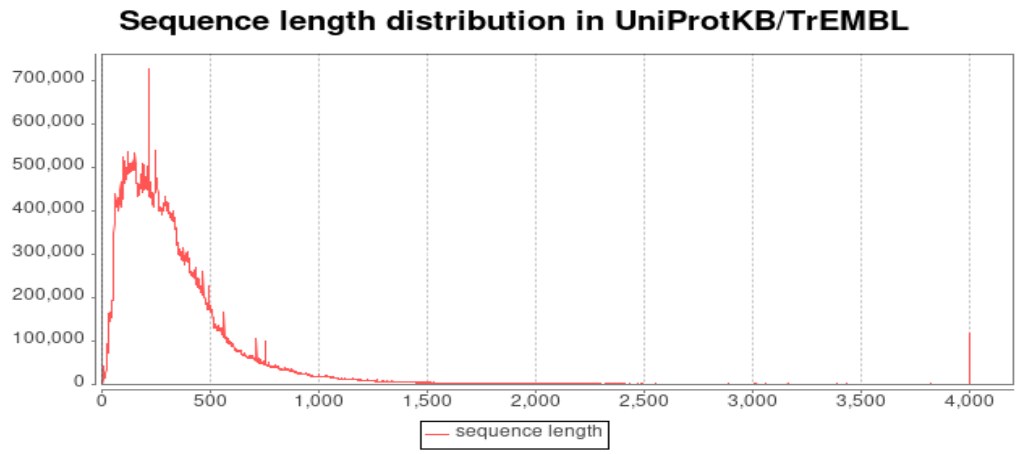
3,530 out of 26,413 identifiers from UniProtKB AC/ID were successfully mapped to 1,765 STRING IDs.

Link used:

https://www.uniprot.org/mapping/M20210309216DA2B77BFBD2E6699CA9B6D1C41EB218BD618

6. 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

a)
Sequence length distribution tells .How protein number varies with sequence length in simple words tells how many number of protein in certain range of sequence length.



We can see maximum number of proteins in the range of 250 sequence length

b)
https://www.uniprot.org/statistics/Swiss-Prot- short sequence
https://www.uniprot.org/statistics/TrEMBL -longest sequence
The shortest sequence is **P83570 at 2 AA** while the longest sequence is **A0A5A9P0L4 at 45,354 AA**

c)

https://www.uniprot.org/statistics/TrEMBL -I used this link to get composition