BT3 040 Practical-III Solutions

Name: Kartik, Roll No: BS19B017

Question 1:

=>> Amino Acid Sequence:

>sp|P21796|VDAC1_HUMAN Voltage-dependent anion-selective channel protein 1 OS=Homo sapiens OX=9606 GN=VDAC1 PE=1 SV=2

MAVPPTYADLGKSARDVFTKGYGFGLIKLDLKTKSENGLEFTSSGSANTETTKVTGSLET
KYRWTEYGLTFTEKWNTDNTLGTEITVEDQLARGLKLTFDSSFSPNTGKKNAKIKTGYKR
EHINLGCDMDFDIAGPSIRGALVLGYEGWLAGYQMNFETAKSRVTQSNFAVGYKTDEFQL
HTNVNDGTEFGGSIYQKVNKKLETAVNLAWTAGNSNTRFGIAAKYQIDPDACFSAKVNNS
SLIGLGYTQTLKPGIKLTLSALLDGKNVNAGGHKLGLGLEFQA.

Procedure: Visit Uniprot website and search for “human mitochondrial β barrel membrane protein VDAC” in search box. Then open the search result: https://www.uniprot.org/uniprot/P21796.

Function: It forms a channel through outer membrane of mitochondria along with the plasma membrane whereby the outer mitochondrial membrane channel permits the diffusion of small hydrophilic molecules; in the plasma membrane which is involved in the apoptosis, cell and volume regulation. It accepts an open conformation at very low or zero membrane potential or adopts an closed conformation when potential is above 30-40mV. The closed has a cation selectivity whereas the open state is weak anion-selective. It may also participate in the production of permeability transition pore complex (PTPC) and it could be the triggering agent for release of mitochondrial products that initiates apoptosis.

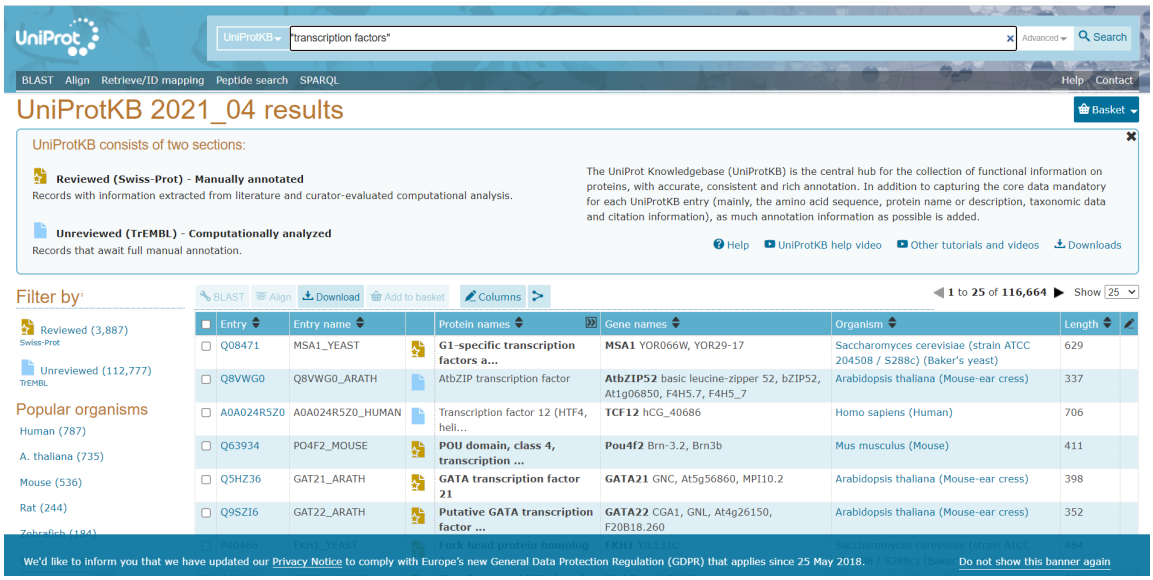
Number of Transmembrane segments present in this protein: 19

Question 2:

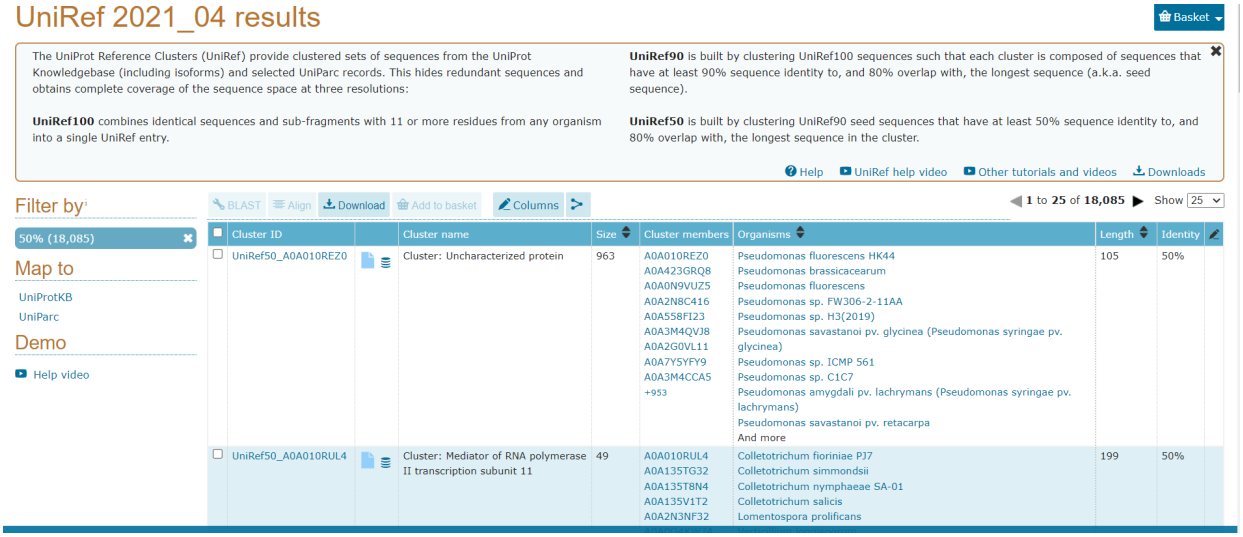
=>> Process:

1. Search for “transcription factors” in UniProt.
https://www.uniprot.org/uniprot/?query=%22transcription+factors%22&sort=score

2. Select cluster identities for specific values.



Total number of sequences: 116,664



50% Identity Clusters: 18,085 clusters

Question 3:

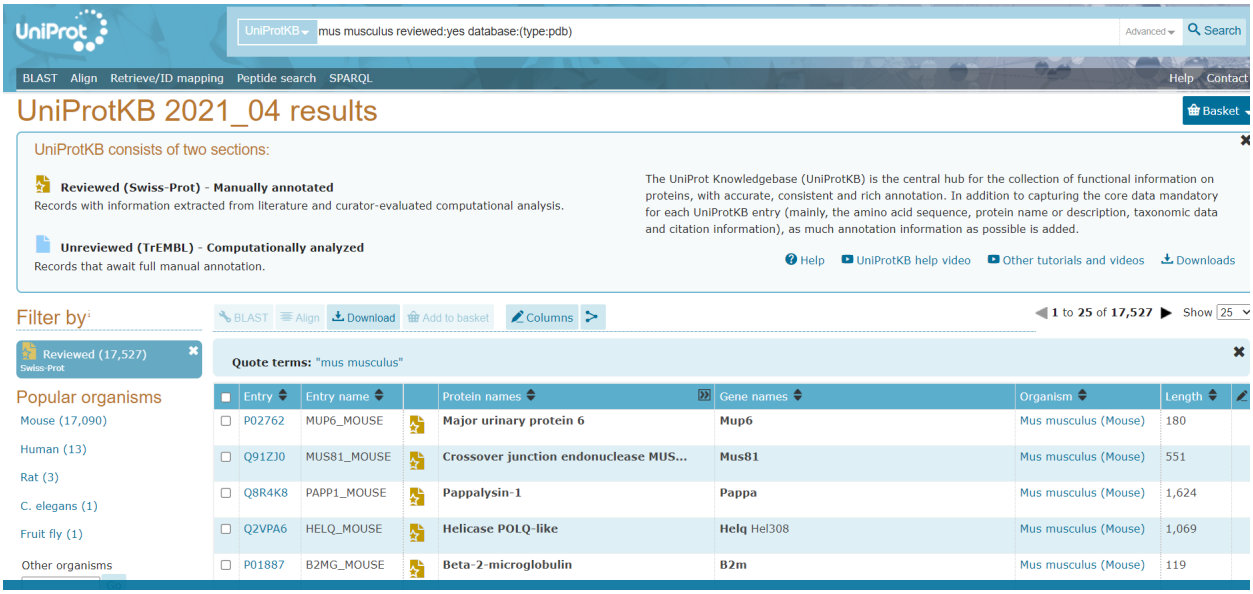
=>> There are a total of 1,926,090 sequences of “homo sapiens” in UniProt.

Sequence Identity	Number of clusters
50%	162,680
90%	450,164
100%	1,358,983

Question 4:

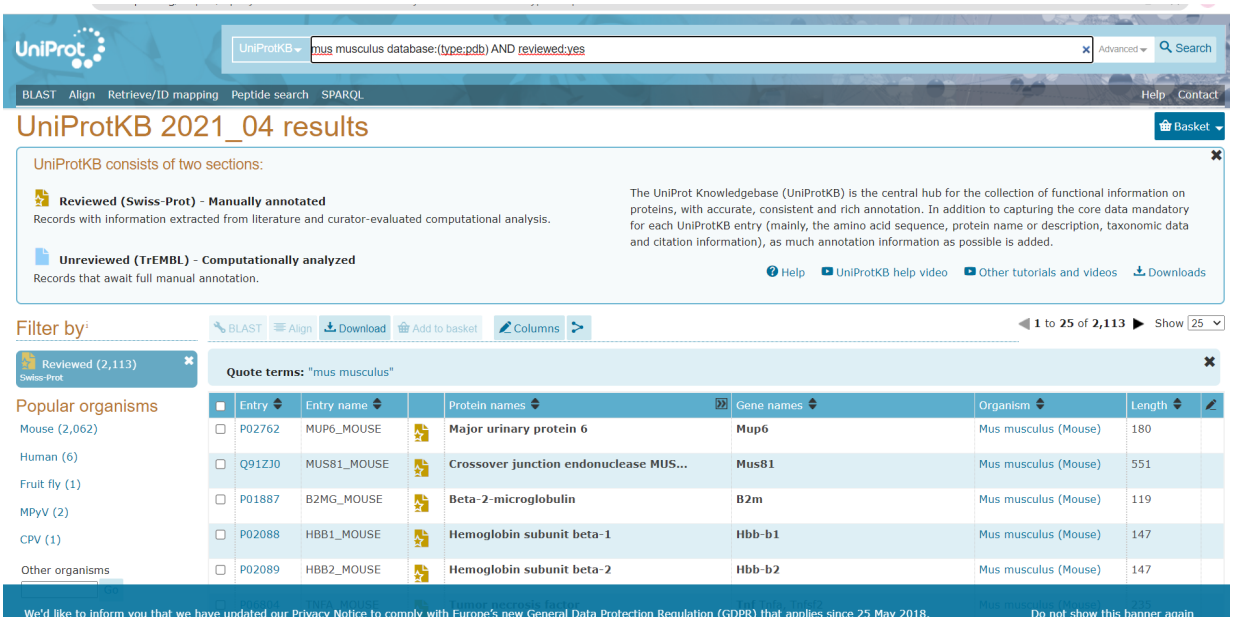
=>> Search query = reviewed:yes AND organism:"Mus musculus (Mouse) [10090]"

17,527 sequences are manually annotated for “Mus Musculus” from search query as shown below.



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

There are **2,113** sequences from the search results shown in image below, which also have 3D structure in PDB



Question 5:

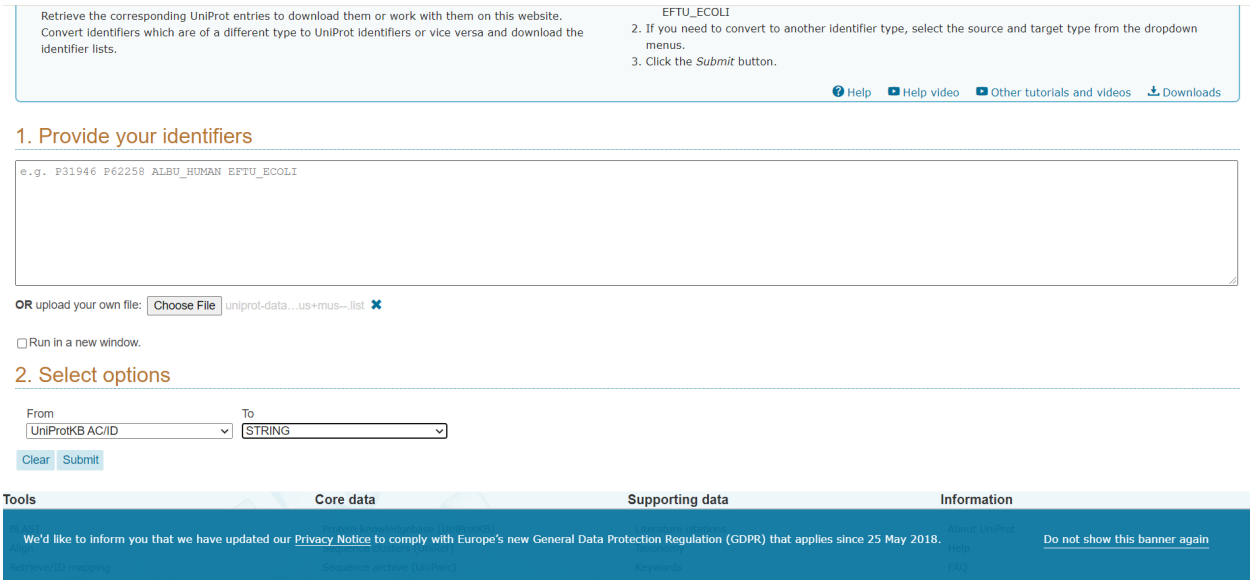
=>> Procedure:

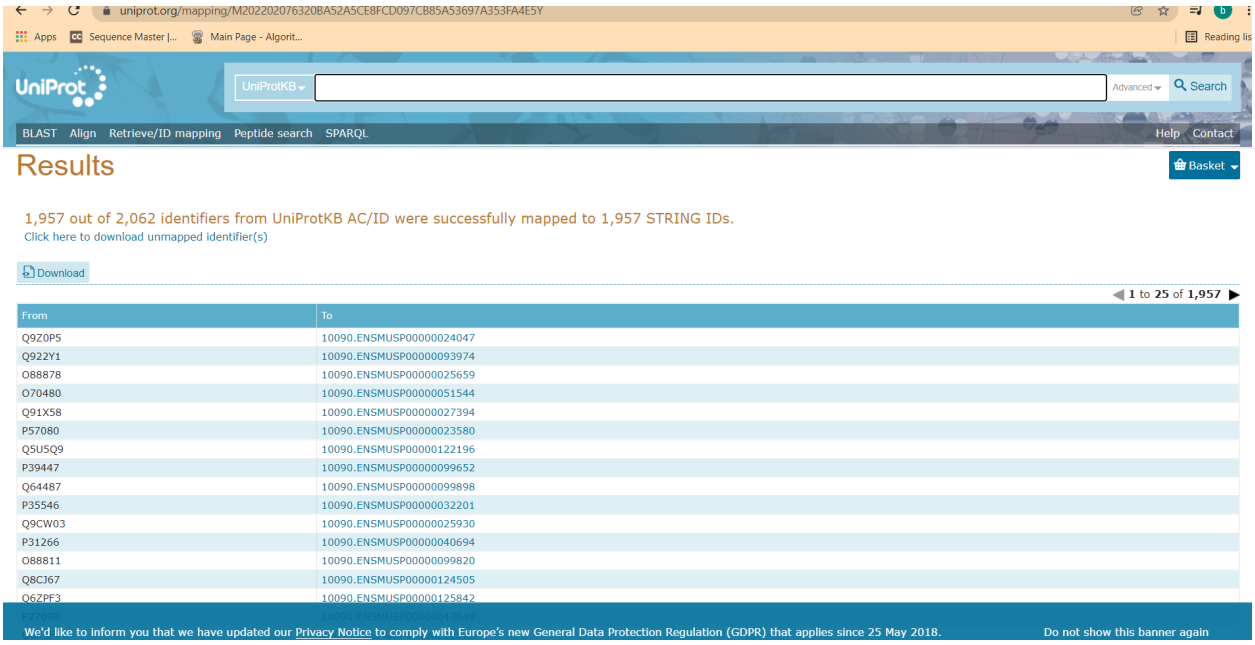
1). Search in UniProt with search query: database:(type:pdb) AND reviewed: yes AND organism:"Mus musculus (Mouse) [10090]"

2. Select only the entry column and download the identifiers as a list as a file.

3. In Retrieve/ID Mapping, paste these identifiers from the downloaded file or direct paste.

4.Under select options: From : **UniProtKB** To : **STRING** and then submit.

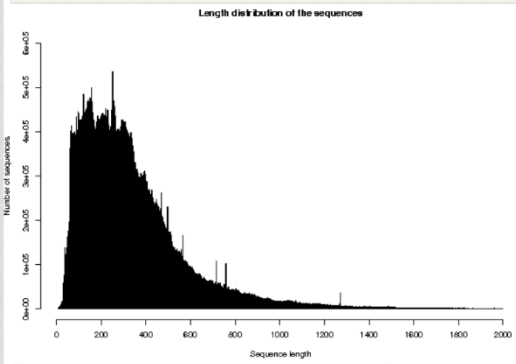




1,957 out of 2,062 identifiers from UniProtKB AC/ID were successfully mapped to **1,957 STRING IDs** as shown in above image.

Question 6: a) =>> Visit (https://www.uniprot.org/statistics/Swiss-Prot) and https://www.ebi.ac.uk/uniprot/TrEMBLstats for statistics on sequence lengths.

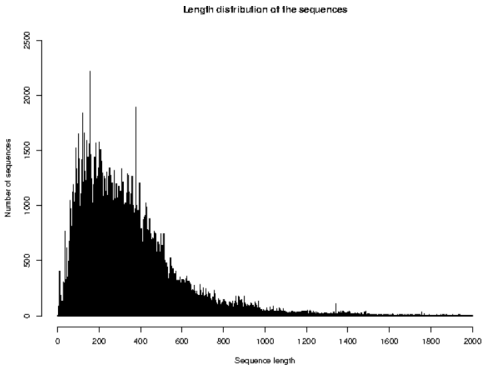
Sequence Length Distribution in UniProtKB/TrEMBL:




```
The average sequence length in UniProtKB/TrEMBL is   348 amino acids.

The shortest sequence is A0A0G2JLF7_HUMAN:     7 amino acids.
The longest sequence is  A0A5A9P0L4_9TELE: 45354 amino acids.
```

Sequence Length Distribution in UniProtKB/Swiss-Prot:




```
The average sequence length in UniProtKB/Swiss-Prot is 360 amino acids.

The shortest sequence is   GWA_SEPOF (P83570):     2 amino acids.
The longest sequence is  TITIN_MOUSE (A2ASS6): 35213 amino acids.
```

Inference: Most of the sequences contain 0 – 800 Amino Acids (AAs). There is very less probability of existence of a sequence with more than 1400 AAs and this probability decreases with increase in sequence length.

b)

As per given in TrEMBL:

The shortest sequence in UniProtKB = **7 AAs**, with sequence ID = **A0A1Y7VI41**

The longest sequence in UniProtKB = **45,354 AAs** with sequence ID = **A0A5A9P0L4**

As per given in Swiss_Prot:

The shortest sequence in UniProtKB = **2 AAs**, which sequence ID = **P0DPR3**

The longest sequence in UniProtKB = **35,213 AAs** with sequence ID = **A2ASS6**

c)

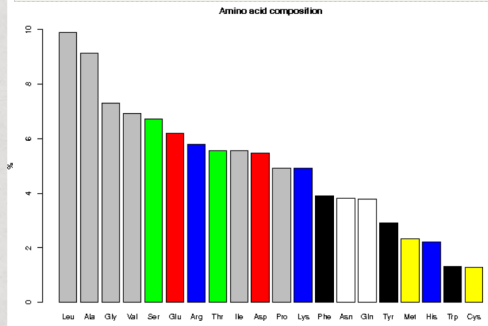
As per given in TrEMBL:

```
5.  AMINO ACID COMPOSITION

   5.1  Composition in percent for the complete database

   Ala (A) 9.13   Gln (Q) 3.77   Leu (L) 9.88   Ser (S) 6.71
   Arg (R) 5.80   Glu (E) 6.18   Lys (K) 4.91   Thr (T) 5.57
   Asn (N) 3.80   Gly (G) 7.30   Met (M) 2.34   Trp (W) 1.30
   Asp (D) 5.47   His (H) 2.20   Phe (F) 3.90   Tyr (Y) 2.90
   Cys (C) 1.27   Ile (I) 5.56   Pro (P) 4.92   Val (V) 6.93

   Asx (B) 0      Glx (Z) 0      Xaa (X) 0.08
```



As per given in Swiss_Prot:

```
6.1  Composition in percent for the complete database

Ala (A) 8.25   Gln (Q) 3.93   Leu (L) 9.65   Ser (S) 6.64
Arg (R) 5.53   Glu (E) 6.72   Lys (K) 5.80   Thr (T) 5.35
Asn (N) 4.06   Gly (G) 7.07   Met (M) 2.41   Trp (W) 1.10
Asp (D) 5.46   His (H) 2.27   Phe (F) 3.86   Tyr (Y) 2.92
Cys (C) 1.38   Ile (I) 5.91   Pro (P) 4.74   Val (V) 6.86

Asx (B) 0.000  Glx (Z) 0.000  Xaa (X) 0.00
```