

BT3 040 Practical-III Solutions

Name: Kartik, Roll No: BS19B017

Question 1:

=>> Amino Acid Sequence:

```
<  →  ↻  uniprot.org/uniprot/P21796.fasta  ⌵  ☆  👤  
>sp|P21796|VDAC1_HUMAN Voltage-dependent anion-selective channel protein 1 OS=Homo sapiens OX=9606 GN=VDAC1 PE=1 SV=2  
MAVPPTYADLGKSARDVFTKGYGFLIKDLKTKSENGLEFTSSGSANTETTKVTGSLET  
KYRWTEYGLTFTEKWNNTDNLGTEITVEDQLARGLKLTDFDSSFSPNTGKKNAIKITGYKR  
EHINLGCDMDFDIAGPSIRGALVLGYEGWLAGYQMNFETAKSRVTQSNFAVGKYKDEFQL  
HTNVNDGTEFGGSIYQKVNKKLETAVNLAWTAGNSNTRFGIAAKYQIDPDACFSKVNNS  
SLIGLGYTQTLKPGIKLTLSALLDGKNVNAGGHKLGLGLEFQA
```

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

MAVPPTYADLGKSARDVFTKGYGFLIKDLKTKSENGLEFTSSGSANTETTKVTGSLET
KYRWTEYGLTFTEKWNNTDNLGTEITVEDQLARGLKLTDFDSSFSPNTGKKNAIKITGYKR
EHINLGCDMDFDIAGPSIRGALVLGYEGWLAGYQMNFETAKSRVTQSNFAVGKYKDEFQL
HTNVNDGTEFGGSIYQKVNKKLETAVNLAWTAGNSNTRFGIAAKYQIDPDACFSKVNNS
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+facto%22&sort=score>

2. Select cluster identities for specific values.

UniProtKB 2021_04 results

UniProtKB consists of two sections:

- Reviewed (Swiss-Prot) - Manually annotated**
Records with information extracted from literature and curator-evaluated computational analysis.
- Unreviewed (TrEMBL) - Computationally analyzed**
Records that await full manual annotation.

The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. In addition to capturing the core data mandatory for each UniProtKB entry (mainly, the amino acid sequence, protein name or description, taxonomic data and citation information), as much annotation information as possible is added.

Filter by:

- Reviewed (3,887)
- Unreviewed (112,777)

Popular organisms

- Human (787)
- A. thaliana (735)
- Mouse (536)
- Rat (244)
- Zebrafish (184)

Entry	Entry name	Protein names	Gene names	Organism	Length
<input type="checkbox"/> Q08471	MSA1_YEAST	G1-specific transcription factors a...	MSA1 YOR066W, YOR29-17	Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)	629
<input type="checkbox"/> Q8VWG0	Q8VWG0_ARATH	AtbZIP transcription factor	AtbZIP52 basic leucine-zipper 52, bZIP52, At1g06850, F4H5.7, F4H5_7	Arabidopsis thaliana (Mouse-ear cress)	337
<input type="checkbox"/> A0A024R5Z0	A0A024R5Z0_HUMAN	Transcription factor 12 (HTF4, heli...	TCF12 hCG_40686	Homo sapiens (Human)	706
<input type="checkbox"/> Q63934	PO4F2_MOUSE	POU domain, class 4, transcription ...	Pou4f2 Brn-3.2, Brn3b	Mus musculus (Mouse)	411
<input type="checkbox"/> Q5HZ36	GAT21_ARATH	GATA transcription factor 21	GATA21 GNC, At5g56860, MP110.2	Arabidopsis thaliana (Mouse-ear cress)	398
<input type="checkbox"/> Q9SZ16	GAT22_ARATH	Putative GATA transcription factor ...	GATA22 CGA1, GNL, At4g26150, F20B18.260	Arabidopsis thaliana (Mouse-ear cress)	352

We'd like to inform you that we have updated our [Privacy Notice](#) to comply with Europe's new General Data Protection Regulation (GDPR) that applies since 25 May 2018. [Do not show this banner again](#)

Total number of sequences: 116,664

UniRef 2021_04 results

The UniProt Reference Clusters (UniRef) provide clustered sets of sequences from the UniProt Knowledgebase (including isoforms) and selected UniParc records. This hides redundant sequences and obtains complete coverage of the sequence space at three resolutions:

- UniRef100** combines identical sequences and sub-fragments with 11 or more residues from any organism into a single UniRef entry.
- UniRef90** is built by clustering UniRef100 sequences such that each cluster is composed of sequences that have at least 90% sequence identity to, and 80% overlap with, the longest sequence (a.k.a. seed sequence).
- UniRef50** is built by clustering UniRef90 seed sequences that have at least 50% sequence identity to, and 80% overlap with, the longest sequence in the cluster.

Filter by:

- 50% (18,085)

Map to

- UniProtKB
- UniParc

Demo

- Help video

Cluster ID	Cluster name	Size	Cluster members	Organisms	Length	Identity
<input type="checkbox"/> UniRef50_A0A010REZ0	Cluster: Uncharacterized protein	963	A0A010REZ0 A0A423GRQ8 A0A0N9VUZ5 A0A2N8C416 A0A558FI23 A0A3M4QVJ8 A0A2GOVL11 A0A7Y5YFY9 A0A3M4CCA5 +953	Pseudomonas fluorescens HK44 Pseudomonas brassicacearum Pseudomonas fluorescens Pseudomonas sp. FW306-2-11AA Pseudomonas sp. H3(2019) Pseudomonas savastanoi pv. glycinea (Pseudomonas syringae pv. glycinea) Pseudomonas sp. ICMP 561 Pseudomonas sp. C1C7 Pseudomonas amygdali pv. lachrymans (Pseudomonas syringae pv. lachrymans) Pseudomonas savastanoi pv. retacarpa And more	105	50%
<input type="checkbox"/> UniRef50_A0A010RUL4	Cluster: Mediator of RNA polymerase II transcription subunit 11	49	A0A010RUL4 A0A135TG32 A0A135T8N4 A0A135V1T2 A0A2N3NF32	Colletotrichum fioriniae P17 Colletotrichum simmondsii Colletotrichum nymphaeae SA-01 Colletotrichum salicis Lomentospora prolificans	199	50%

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.

UniProtKB 2021_04 results

UniProtKB consists of two sections:

- Reviewed (Swiss-Prot) - Manually annotated**
Records with information extracted from literature and curator-evaluated computational analysis.
- Unreviewed (TrEMBL) - Computationally analyzed**
Records that await full manual annotation.

The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. In addition to capturing the core data mandatory for each UniProtKB entry (mainly, the amino acid sequence, protein name or description, taxonomic data and citation information), as much annotation information as possible is added.

Filter by: **Reviewed (17,527)** Swiss-Prot

Popular organisms: Mouse (17,090), Human (13), Rat (3), C. elegans (1), Fruit fly (1), Other organisms

Quote terms: "mus musculus"

Entry	Entry name	Protein names	Gene names	Organism	Length
<input type="checkbox"/> P02762	MUP6_MOUSE	Major urinary protein 6	Mup6	Mus musculus (Mouse)	180
<input type="checkbox"/> Q91ZJ0	MUS81_MOUSE	Crossover junction endonuclease MUS...	Mus81	Mus musculus (Mouse)	551
<input type="checkbox"/> Q8R4K8	PAPP1_MOUSE	Pappalysin-1	Pappa	Mus musculus (Mouse)	1,624
<input type="checkbox"/> Q2VPA6	HELQ_MOUSE	Helicase POLQ-like	Helq Hel308	Mus musculus (Mouse)	1,069
<input type="checkbox"/> P01887	B2MG_MOUSE	Beta-2-microglobulin	B2m	Mus musculus (Mouse)	119

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

UniProtKB 2021_04 results

UniProtKB consists of two sections:

- Reviewed (Swiss-Prot) - Manually annotated**
Records with information extracted from literature and curator-evaluated computational analysis.
- Unreviewed (TrEMBL) - Computationally analyzed**
Records that await full manual annotation.

The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. In addition to capturing the core data mandatory for each UniProtKB entry (mainly, the amino acid sequence, protein name or description, taxonomic data and citation information), as much annotation information as possible is added.

Filter by:

Quote terms: "mus musculus"

Entry	Entry name	Protein names	Gene names	Organism	Length
P02762	MUP6_MOUSE	Major urinary protein 6	Mup6	Mus musculus (Mouse)	180
Q91ZJ0	MUS81_MOUSE	Crossover junction endonuclease MUS...	Mus81	Mus musculus (Mouse)	551
P01887	B2MG_MOUSE	Beta-2-microglobulin	B2m	Mus musculus (Mouse)	119
P02088	HBB1_MOUSE	Hemoglobin subunit beta-1	Hbb-b1	Mus musculus (Mouse)	147
P02089	HBB2_MOUSE	Hemoglobin subunit beta-2	Hbb-b2	Mus musculus (Mouse)	147

We'd like to inform you that we have updated our Privacy Notice to comply with Europe's new General Data Protection Regulation (GDPR) that applies since 25 May 2018. Do not show this banner again.

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.

Retrieve the corresponding UniProt entries to download them or work with them on this website. Convert identifiers which are of a different type to UniProt identifiers or vice versa and download the identifier lists.

EFTU_ECOLI

2. If you need to convert to another identifier type, select the source and target type from the dropdown menus.
3. Click the *Submit* button.

[Help](#) [Help video](#) [Other tutorials and videos](#) [Downloads](#)

1. Provide your identifiers

e.g. P31946 P62258 ALBU_HUMAN EFTU_ECOLI

OR upload your own file: uniprot-data...us+mus--list ✕

☐ Run in a new window.

2. Select options

From To

Tools	Core data	Supporting data	Information
We'd like to inform you that we have updated our Privacy Notice to comply with Europe's new General Data Protection Regulation (GDPR) that applies since 25 May 2018. Do not show this banner again			
Sequence ID mapping	Sequence annotation (UniProt)	Proteomics	UniProt help

← → ↻ uniprot.org/mapping/M2022020763208A52A5CE8FCD097CB85A53697A353FA4E5Y

Apps [Sequence Master J...](#) [Main Page - Algorit...](#) [Reading lis](#)

UniProt

[BLAST](#) [Align](#) [Retrieve/ID mapping](#) [Peptide search](#) [SPARQL](#) [Help](#) [Contact](#)

Results

[Basket](#)

1,957 out of 2,062 identifiers from UniProtKB AC/ID were successfully mapped to 1,957 STRING IDs.
[Click here to download unmapped identifier\(s\)](#)

◀ 1 to 25 of 1,957 ▶

From	To
Q9Z0P5	10090.ENSMUSP00000024047
Q9Z2Y1	10090.ENSMUSP00000093974
O88878	10090.ENSMUSP00000025659
O70480	10090.ENSMUSP00000051544
Q91XS8	10090.ENSMUSP00000027394
P57080	10090.ENSMUSP00000023580
Q5USQ9	10090.ENSMUSP000000122196
P39447	10090.ENSMUSP00000099652
Q64487	10090.ENSMUSP00000099898
P35546	10090.ENSMUSP00000032201
Q9CW03	10090.ENSMUSP00000025930
P31266	10090.ENSMUSP00000040694
O88811	10090.ENSMUSP00000099820
Q8CJ67	10090.ENSMUSP00000124505
Q6ZPF3	10090.ENSMUSP00000125842

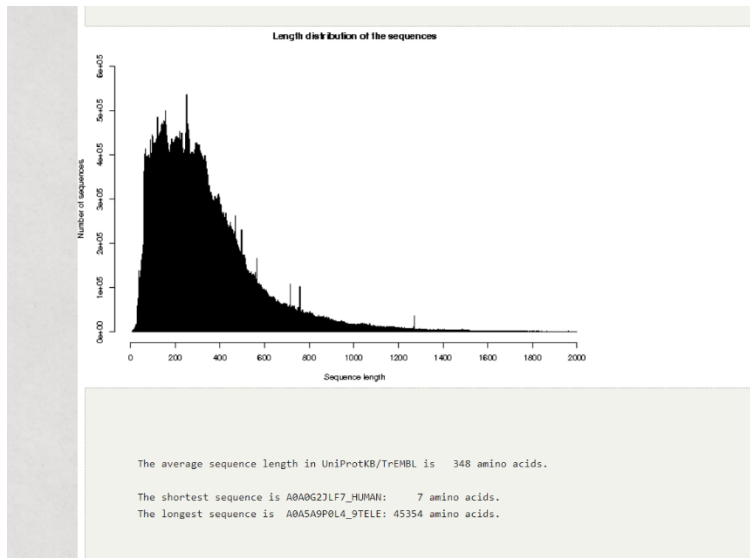
We'd like to inform you that we have updated our Privacy Notice to comply with Europe's new General Data Protection Regulation (GDPR) that applies since 25 May 2018.

[Do not show this banner again](#)

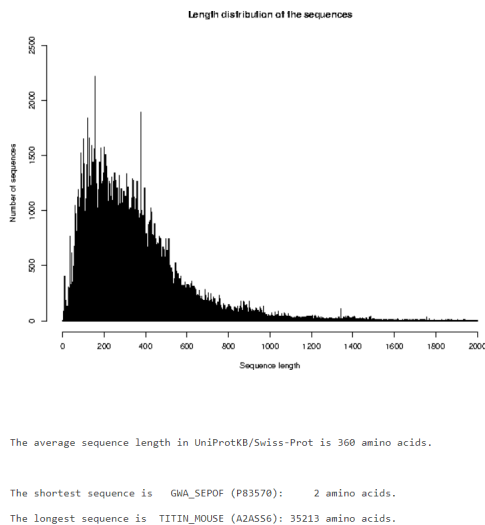
1,957 out of 2,062 identifiers from UniProtKB AC/ID were successfully mapped to **1,957 STRING IDs** as shown in above image.

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:



Sequence Length Distribution in UniProtKB/Swiss-Prot:



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**

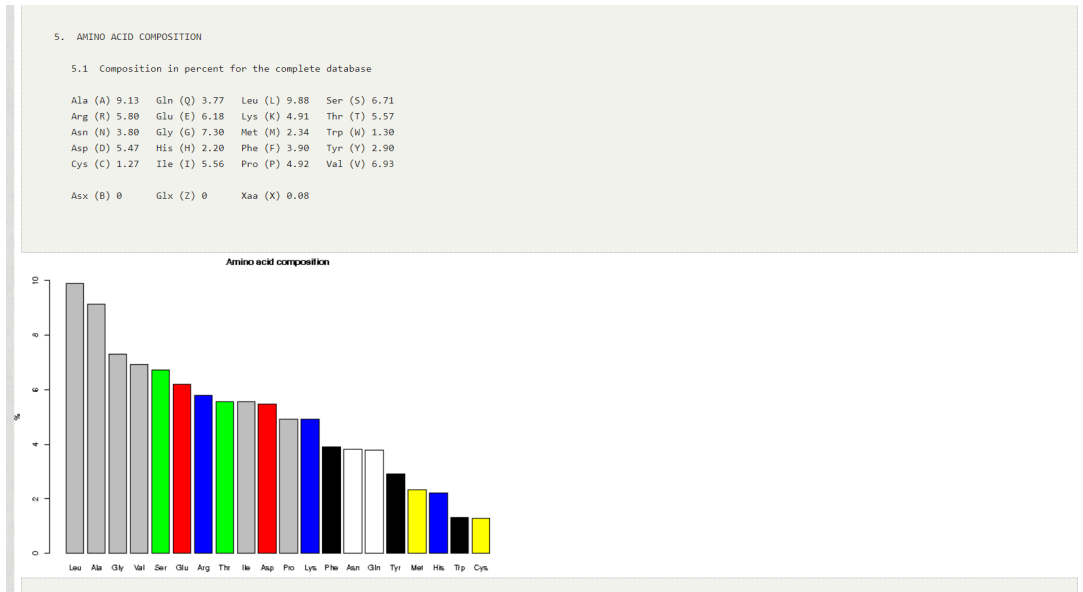
As per given in Swiss Prot:

The shortest sequence in UniProtKB = **2 AAs**, which sequence ID = **PODPR3**

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

c)

As per given in TrEMBL:



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

