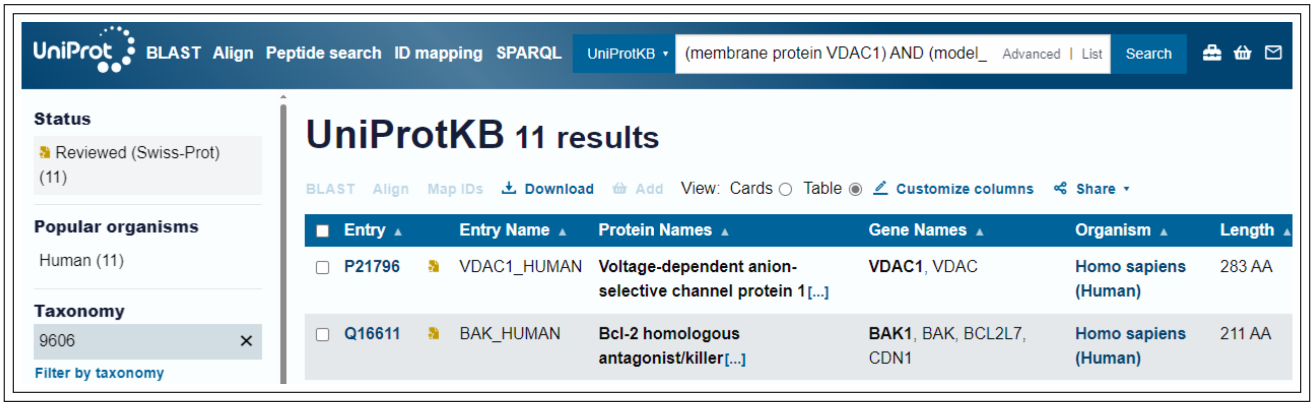
Practical 3

Question 1. Find the amino acid sequence of **human mitochondrial beta-barrel membrane protein VDAC1** and its function. How many **transmembrane segments** are present in the protein?

Solution. Using an advanced query search on the **UniProtKB**, we can find the desired amino acid sequence.



(membrane protein VDAC1) AND (model_

Status

Reviewed (Swiss-Prot) (11)

Popular organisms

Human (11)

Taxonomy

9606 ✕

UniProtKB 11 results

View: Cards ○ Table ●

□	Entry	Entry Name	Protein Names	Gene Names	Organism	Length
□	P21796	VDAC1_HUMAN	Voltage-dependent anion-selective channel protein 1[...]	VDAC1, VDAC	Homo sapiens (Human)	283 AA
□	Q16611	BAK_HUMAN	Bcl-2 homologous antagonist/killer[...]	BAK1, BAK, BCL2L7, CDN1	Homo sapiens (Human)	211 AA

Figure 1. Advanced search on UniProtKB

Human mitochondrial beta barrel membrane protein VDAC1

```
# The advanced query used was:

(membrane protein VDAC1) AND (model_organism:9606) AND (proteins_with:7)
    AND (keyword:KW-0496)

# The amino acid sequence

MAVPPTYADLGKSARDVFTKGYGFGLIKLDLKTKSENGLEFTSSGSANTETTKVTGSLETKYRWTEYGLTFT
EKWNTDNTLGTEITVEDQLARGLKLTFDSSFSPNTGKKNAKIKTGYKREHINLGCDMDFDIAGPSIRGALVL
GYEGWLAGYQMNFETAKSRVTQSNFAVGYKTDEFQLHTNVNDGTEFGGSIYQKVNKKLETAVNLAWTAGNSN
TRFGIAAKYQIDPDACFSAKVNNSSLIGLGYTQTLKPGIKLTLSALLDGKNVNAGGHKLGLGLEFQA
```

The functions of the human mitochondrial beta-barrel membrane protein VDAC1 are:

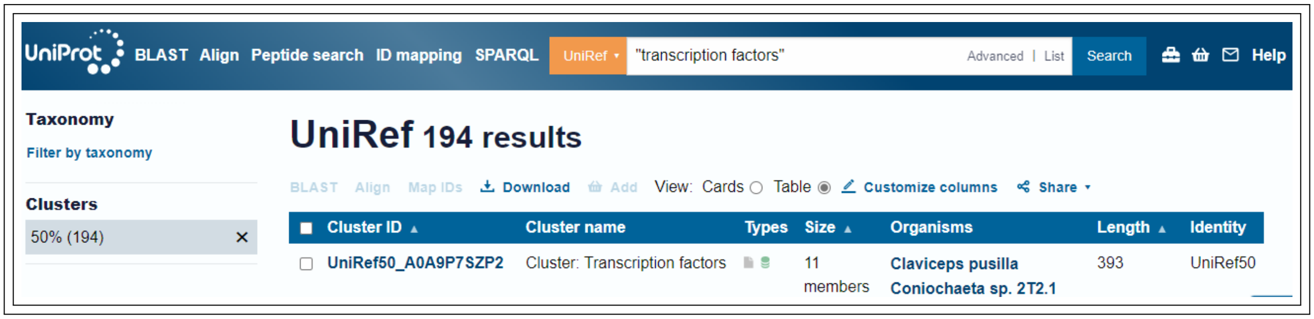
- Forms a channel through the mitochondrial outer membrane and also the plasma membrane. The channel at the outer mitochondrial membrane allows diffusion of small hydrophilic molecules; in the plasma membrane, it is involved in cell volume regulation and apoptosis. It adopts an open conformation at low or zero membrane potential and a closed conformation at potentials above 30–40 mV. The open state has weak anion selectivity, whereas the closed state is cation-selective.
- Binds various signaling molecules, including the sphingolipid ceramide, the phospholipid phosphatidylcholine, and the sterol cholesterol.
- In depolarized mitochondria, it acts downstream of PRKN and PINK1 to promote mitophagy or prevent apoptosis; polyubiquitination by PRKN promotes mitophagy, while monoubiquitination by PRKN decreases mitochondrial calcium influx, which ultimately inhibits apoptosis.

- May participate in the formation of the permeability transition pore complex (PTPC), which is responsible for the release of mitochondrial products that trigger the process of apoptosis.
- May mediate ATP export from cells.

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

Question 2. Obtain the sequences of "transcription factors" with 50% sequence identity in FASTA format. List the count of sequences and the count of clusters.
http://www.uniprot.org/uniprot/

Solution. I performed an advanced search on the **UniRef** platform and chose the cluster with 50% sequence identity.



"transcription factors"

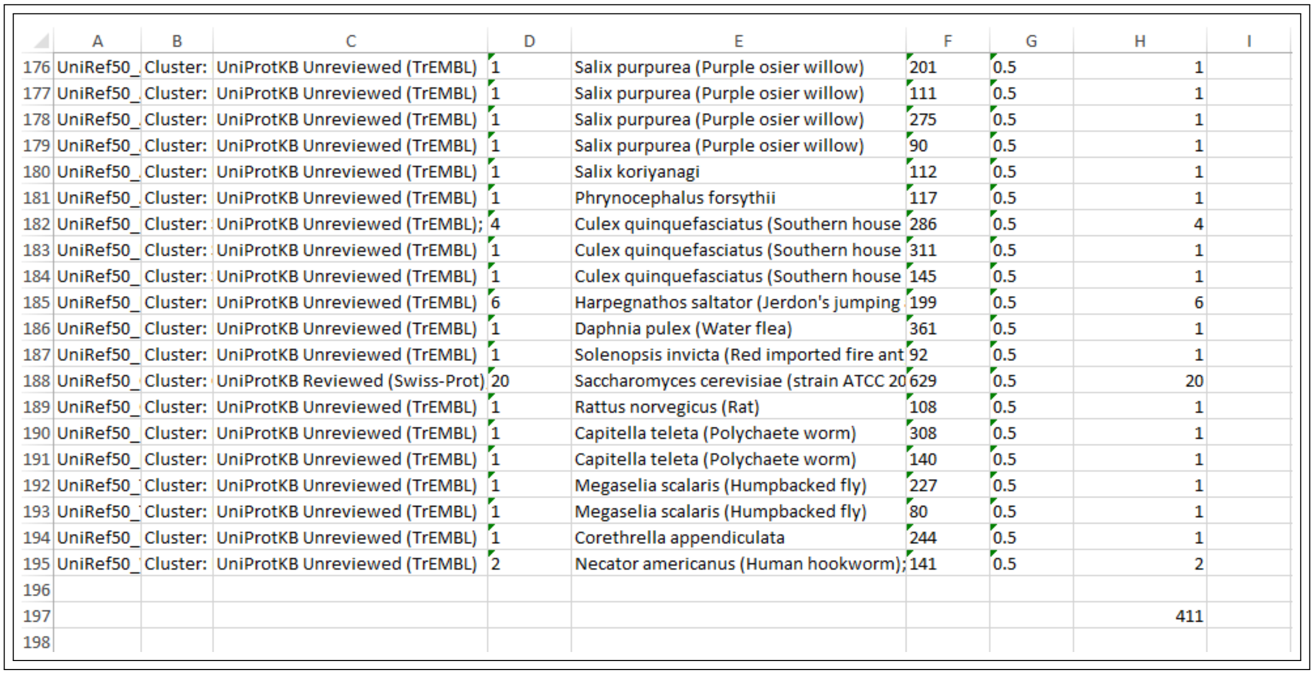
50% (194) ×

UniRef 194 results

☐	Cluster ID ▲	Cluster name	Types	Size ▲	Organisms	Length ▲	Identity
☐	UniRef50_A0A9P7SZP2	Cluster: Transcription factors		11 members	Claviceps pusilla Coniochaeta sp. 2T2.1	393	UniRef50

FIGURE 2. Advanced search on UniRef with clusters

The count of clusters obtained is **194**. Now, this database with 194 clusters is downloaded in Excel format. The database has a feature called cluster size. This denotes the number of sequences present in each cluster. So, the downloaded database has a column called cluster size. The sum of all entries in this column will give us a count of the number of sequences. Seen below is a screenshot of the Excel file with the summation operation. The count of the number of sequences is **411**.

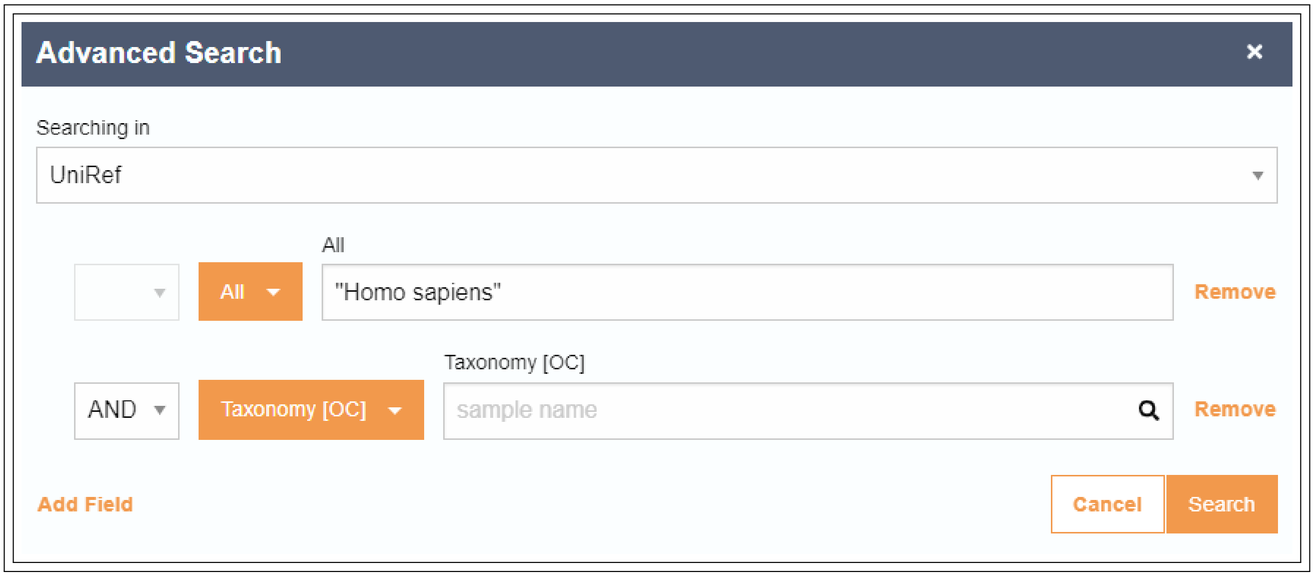


	A	B	C	D	E	F	G	H	I
176	UniRef50_	Cluster:	UniProtKB Unreviewed (TrEMBL)	1	Salix purpurea (Purple osier willow)	201	0.5	1	
177	UniRef50_	Cluster:	UniProtKB Unreviewed (TrEMBL)	1	Salix purpurea (Purple osier willow)	111	0.5	1	
178	UniRef50_	Cluster:	UniProtKB Unreviewed (TrEMBL)	1	Salix purpurea (Purple osier willow)	275	0.5	1	
179	UniRef50_	Cluster:	UniProtKB Unreviewed (TrEMBL)	1	Salix purpurea (Purple osier willow)	90	0.5	1	
180	UniRef50_	Cluster:	UniProtKB Unreviewed (TrEMBL)	1	Salix koriyanagi	112	0.5	1	
181	UniRef50_	Cluster:	UniProtKB Unreviewed (TrEMBL)	1	Phrynocephalus forsythii	117	0.5	1	
182	UniRef50_	Cluster:	UniProtKB Unreviewed (TrEMBL);	4	Culex quinquefasciatus (Southern house	286	0.5	4	
183	UniRef50_	Cluster:	UniProtKB Unreviewed (TrEMBL)	1	Culex quinquefasciatus (Southern house	311	0.5	1	
184	UniRef50_	Cluster:	UniProtKB Unreviewed (TrEMBL)	1	Culex quinquefasciatus (Southern house	145	0.5	1	
185	UniRef50_	Cluster:	UniProtKB Unreviewed (TrEMBL)	6	Harpegnathos saltator (Jerdon's jumping	199	0.5	6	
186	UniRef50_	Cluster:	UniProtKB Unreviewed (TrEMBL)	1	Daphnia pulex (Water flea)	361	0.5	1	
187	UniRef50_	Cluster:	UniProtKB Unreviewed (TrEMBL)	1	Solenopsis invicta (Red imported fire ant	92	0.5	1	
188	UniRef50_	Cluster:	UniProtKB Reviewed (Swiss-Prot)	20	Saccharomyces cerevisiae (strain ATCC 20	629	0.5	20	
189	UniRef50_	Cluster:	UniProtKB Unreviewed (TrEMBL)	1	Rattus norvegicus (Rat)	108	0.5	1	
190	UniRef50_	Cluster:	UniProtKB Unreviewed (TrEMBL)	1	Capitella teleta (Polychaete worm)	308	0.5	1	
191	UniRef50_	Cluster:	UniProtKB Unreviewed (TrEMBL)	1	Capitella teleta (Polychaete worm)	140	0.5	1	
192	UniRef50_	Cluster:	UniProtKB Unreviewed (TrEMBL)	1	Megaselia scalaris (Humpbacked fly)	227	0.5	1	
193	UniRef50_	Cluster:	UniProtKB Unreviewed (TrEMBL)	1	Megaselia scalaris (Humpbacked fly)	80	0.5	1	
194	UniRef50_	Cluster:	UniProtKB Unreviewed (TrEMBL)	1	Corethrella appendiculata	244	0.5	1	
195	UniRef50_	Cluster:	UniProtKB Unreviewed (TrEMBL)	2	Necator americanus (Human hookworm);	141	0.5	2	
196									
197								411	
198									

FIGURE 3. Counting the number of sequences

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

Solution. I performed an advanced search on the **UniRef** platform, where I filtered by taxonomy. In the taxonomy field, I entered **Homo sapiens**.



Advanced Search

Searching in: UniRef

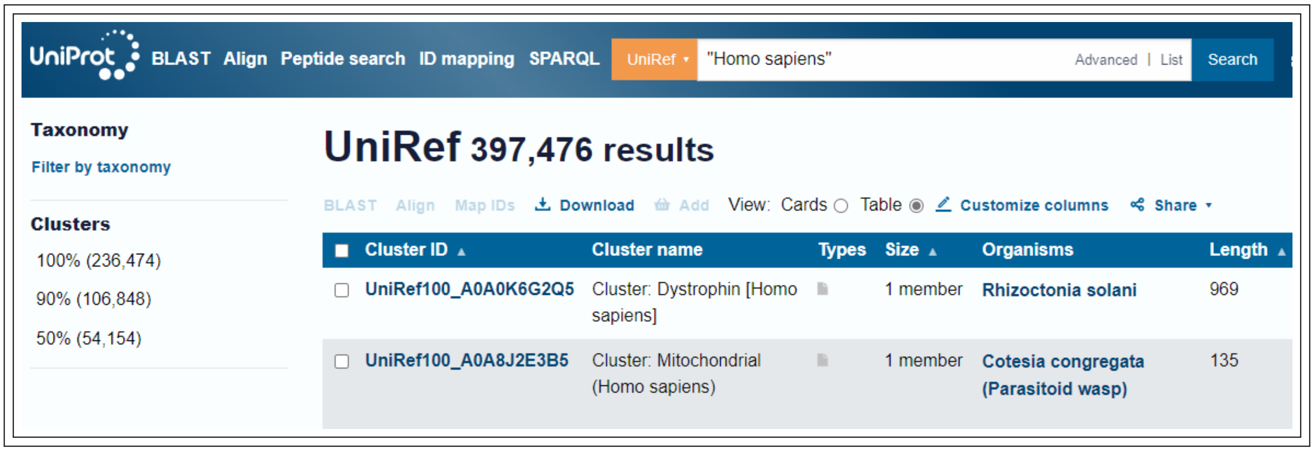
All: "Homo sapiens"

AND Taxonomy [OC]: sample name

FIGURE 4. Filter search by taxonomy

On filtering by taxonomy, I obtained the number of protein sequences at identity cutoffs of 100%, 90% and 50% sequence identity.

100%	90%	50%
236,474	106,848	54,154



Taxonomy

Filter by taxonomy

Clusters

100% (236,474)

90% (106,848)

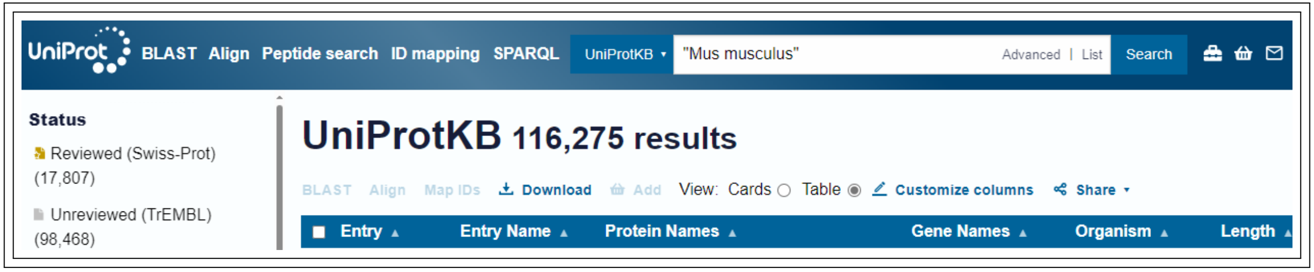
50% (54,154)

UniRef 397,476 results

Cluster ID	Cluster name	Types	Size	Organisms	Length
UniRef100_A0A0K6G2Q5	Cluster: Dystrophin [Homo sapiens]		1 member	Rhizoctonia solani	969
UniRef100_A0A8J2E3B5	Cluster: Mitochondrial (Homo sapiens)		1 member	Cotesia congregata (Parasitoid wasp)	135

FIGURE 5. Filter search by taxonomy

Question 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


UniProt
BLAST Align Peptide search ID mapping SPARQL
UniProtKB ▾
Advanced | List
Search
BLAST Align Map IDs Download Add
Customize columns
Share ▾

"Mus musculus"

UniProtKB 116,275 results

Status

Reviewed (Swiss-Prot) (17,807)

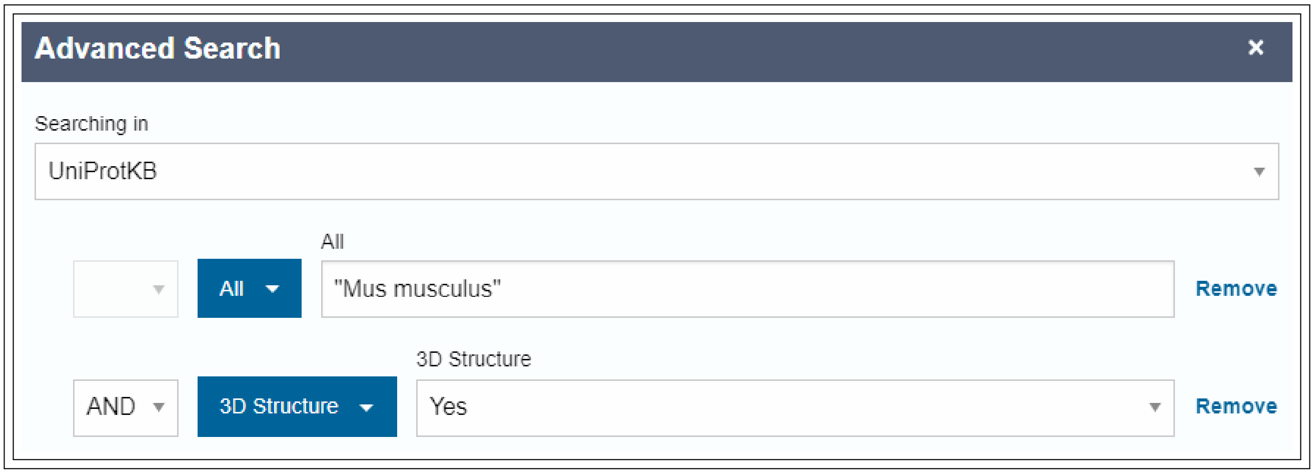
Unreviewed (TrEMBL) (98,468)

View: Cards ○ Table ●

■	Entry ▲	Entry Name ▲	Protein Names ▲	Gene Names ▲	Organism ▲	Length ▲

FIGURE 6. UniProt search for Mus musculus

Solution. Firstly, I obtained the number of manually annotated protein sequences in mouse, by searching for **Mus musculus** in the UniProtKB. This gave me the result that there are **116,275** manually annotated protein sequences of *Mus musculus* in UniProt.
In order to obtain the number of protein sequences associated with 3D structures, I used an advanced search. Along with the search of *Mus musculus*, I added another field called **3D structure** and set it to **Yes**. This gave me the number of manually annotated protein sequences that are associated with 3D structures. This number is **2,610**.

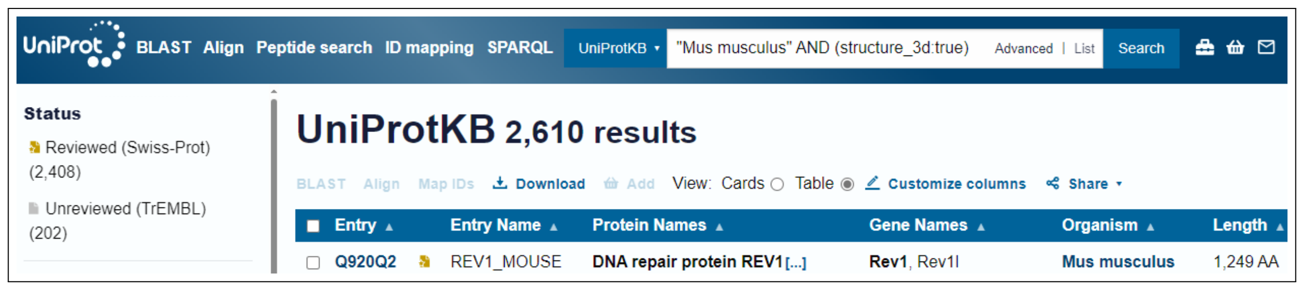


Searching in: UniProtKB

All: "Mus musculus"

AND 3D Structure: Yes

FIGURE 7. UniProt advanced search for Mus musculus 3D structures


UniProt
BLAST Align Peptide search ID mapping SPARQL
UniProtKB ▾
Advanced | List
Search
BLAST Align Map IDs Download Add
Customize columns
Share ▾

"Mus musculus" AND (structure_3d:true)

UniProtKB 2,610 results

Status

Reviewed (Swiss-Prot) (2,408)

Unreviewed (TrEMBL) (202)

View: Cards ○ Table ●

■	Entry ▲		Entry Name ▲	Protein Names ▲	Gene Names ▲	Organism ▲	Length ▲
□	Q920Q2		REV1_MOUSE	DNA repair protein REV1[...]	Rev1, Rev1l	Mus musculus	1,249 AA

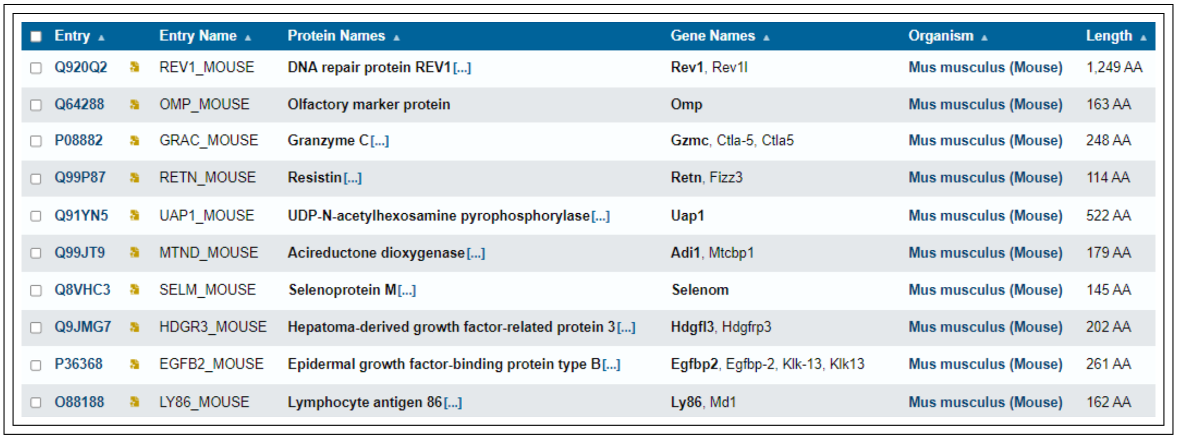
FIGURE 8. UniProt search for Mus musculus 3D structures

Mus Musculus

The number of manually annotated protein sequences: 116,275
The number of manually annotated protein sequences associated with 3D structures: 2,610

Question 5. Map first 10 UniProt IDs of above manually curated mouse protein sequences with 3D structures to the STRING database. How many STRING IDs are mapped?
Hint: Use "Retrieve/ID mapping" and upload Uniprot IDs as text file

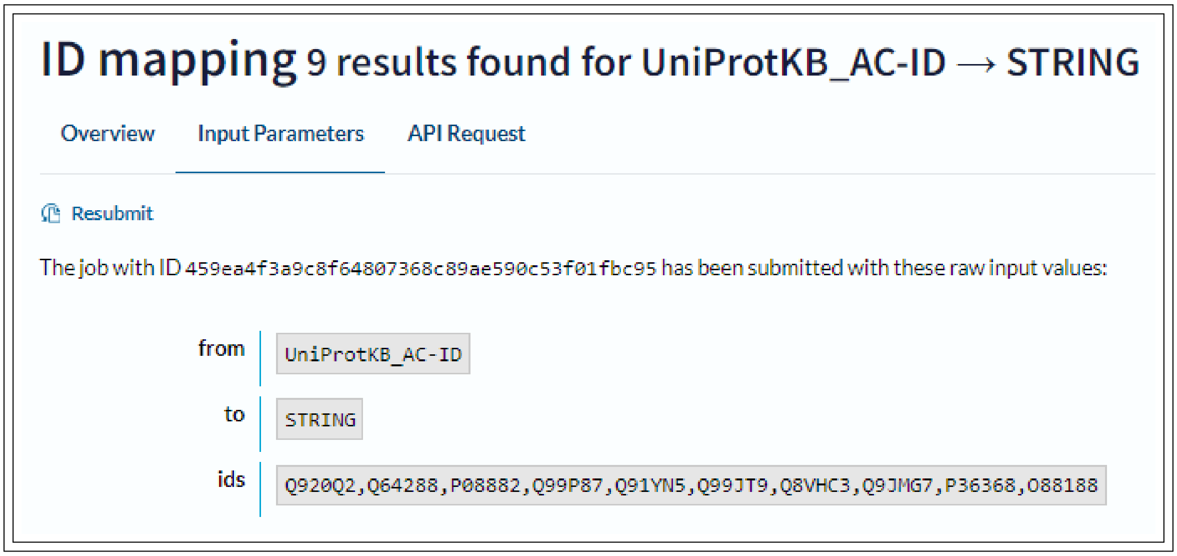
Solution. From the previous result, we have the top 10 UniProt IDs of manually curated mouse protein sequences with 3D structures.



■ Entry ▲		Entry Name ▲	Protein Names ▲	Gene Names ▲	Organism ▲	Length ▲
□ Q920Q2		REV1_MOUSE	DNA repair protein REV1 [...]	**Rev1**, Rev1l	Mus musculus (Mouse)	1,249 AA
□ Q64288		OMP_MOUSE	Olfactory marker protein	**Omp**	Mus musculus (Mouse)	163 AA
□ P08882		GRAC_MOUSE	Granzyme C [...]	**Gzmc**, Ctla-5, Ctla5	Mus musculus (Mouse)	248 AA
□ Q99P87		RETN_MOUSE	Resistin [...]	**Retn**, Fizz3	Mus musculus (Mouse)	114 AA
□ Q91YN5		UAP1_MOUSE	UDP-N-acetylhexosamine pyrophosphorylase [...]	**Uap1**	Mus musculus (Mouse)	522 AA
□ Q99JT9		MTND_MOUSE	Acireductone dioxygenase [...]	**Adi1**, Mtcbp1	Mus musculus (Mouse)	179 AA
□ Q8VHC3		SELM_MOUSE	Selenoprotein M [...]	**Selenom**	Mus musculus (Mouse)	145 AA
□ Q9JMG7		HDGR3_MOUSE	Hepatoma-derived growth factor-related protein 3 [...]	**Hdgfl3**, Hdgfrp3	Mus musculus (Mouse)	202 AA
□ P36368		EGFB2_MOUSE	Epidermal growth factor-binding protein type B [...]	**Egfbp2**, Egfbp-2, Klk-13, Klk13	Mus musculus (Mouse)	261 AA
□ O88188		LY86_MOUSE	Lymphocyte antigen 86 [...]	**Ly86**, Md1	Mus musculus (Mouse)	162 AA

FIGURE 9. Top 10 UniProt IDs satisfying criteria

I went to the **Retrieve/ID mapping** section of the UniProt. Here, I entered the top 10 UniProt IDs from above in the form of a text file separated by commas.



ID mapping 9 results found for UniProtKB_AC-ID → STRING

The job with ID 459ea4f3a9c8f64807368c89ae590c53f01fbc95 has been submitted with these raw input values:

from `UniProtKB_AC-ID`

to `STRING`

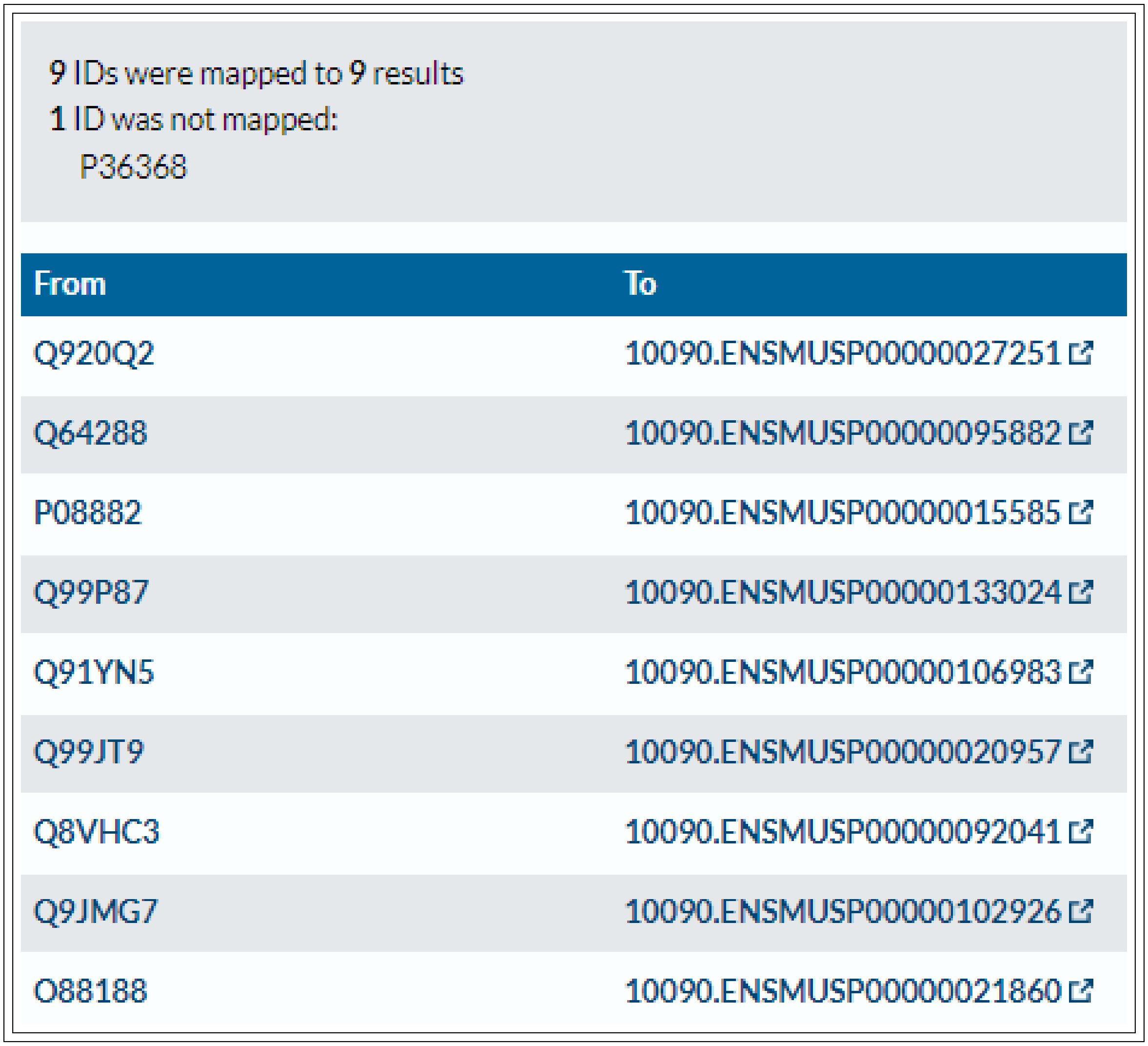
ids `Q920Q2,Q64288,P08882,Q99P87,Q91YN5,Q99JT9,Q8VHC3,Q9JMG7,P36368,O88188`

FIGURE 10. Input to the ID Mapping

I set the **From database** to **UniProtKB AC/ID**. I also set the **To database** to **STRING**. Then I ran the command **Map IDs**. Out of the 10 IDs, 9 IDs were mapped to 9 results, and 1 ID was not mapped.

UniProtKB to STRING database ID Mapping

```
Q920Q2                              10090.ENSMUSP00000027251
Q64288                              10090.ENSMUSP00000095882
P08882                              10090.ENSMUSP00000015585
Q99P87                              10090.ENSMUSP00000133024
Q91YN5                              10090.ENSMUSP00000106983
Q99JT9                              10090.ENSMUSP00000020957
Q8VHC3                              10090.ENSMUSP00000092041
Q9JMG7                              10090.ENSMUSP00000102926
O88188                              10090.ENSMUSP00000021860
```



9 IDs were mapped to 9 results
1 ID was not mapped:
P36368

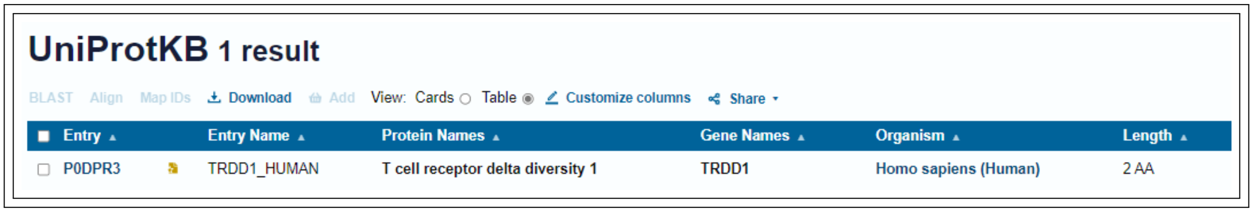
From	To
Q920Q2	10090.ENSMUSP00000027251
Q64288	10090.ENSMUSP00000095882
P08882	10090.ENSMUSP00000015585
Q99P87	10090.ENSMUSP00000133024
Q91YN5	10090.ENSMUSP00000106983
Q99JT9	10090.ENSMUSP00000020957
Q8VHC3	10090.ENSMUSP00000092041
Q9JMG7	10090.ENSMUSP00000102926
O88188	10090.ENSMUSP00000021860

FIGURE 11. ID Mapping Output

Question 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

Solution. Following are the solutions using the UniProt Statistics data

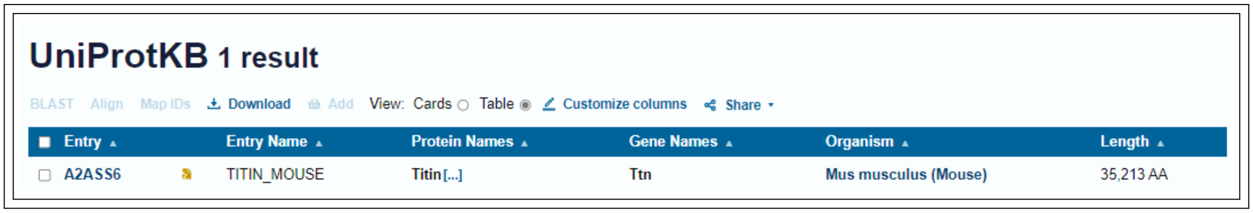
- (a) This indicates the number of **amino acids** in the **canonical sequence** displayed by default in the entry's Sequence section.
- (b) Shortest sequence in UniProtKB is the **P0DPR3**, **T cell receptor delta diversity 1**, belonging to *Homo sapiens*. It is **2 amino acid** long.
  Longest sequence in UniProtKB is the **A2ASS6**, **Titin, 2.7.11.1, Connectin**. It belongs to *Mus musculus*. It is **35,213 amino acid** long.



UniProtKB 1 result

Entry	Entry Name	Protein Names	Gene Names	Organism	Length
☐ P0DPR3	TRDD1_HUMAN	T cell receptor delta diversity 1	TRDD1	Homo sapiens (Human)	2 AA

FIGURE 12. Shortest sequence in UniProtKB

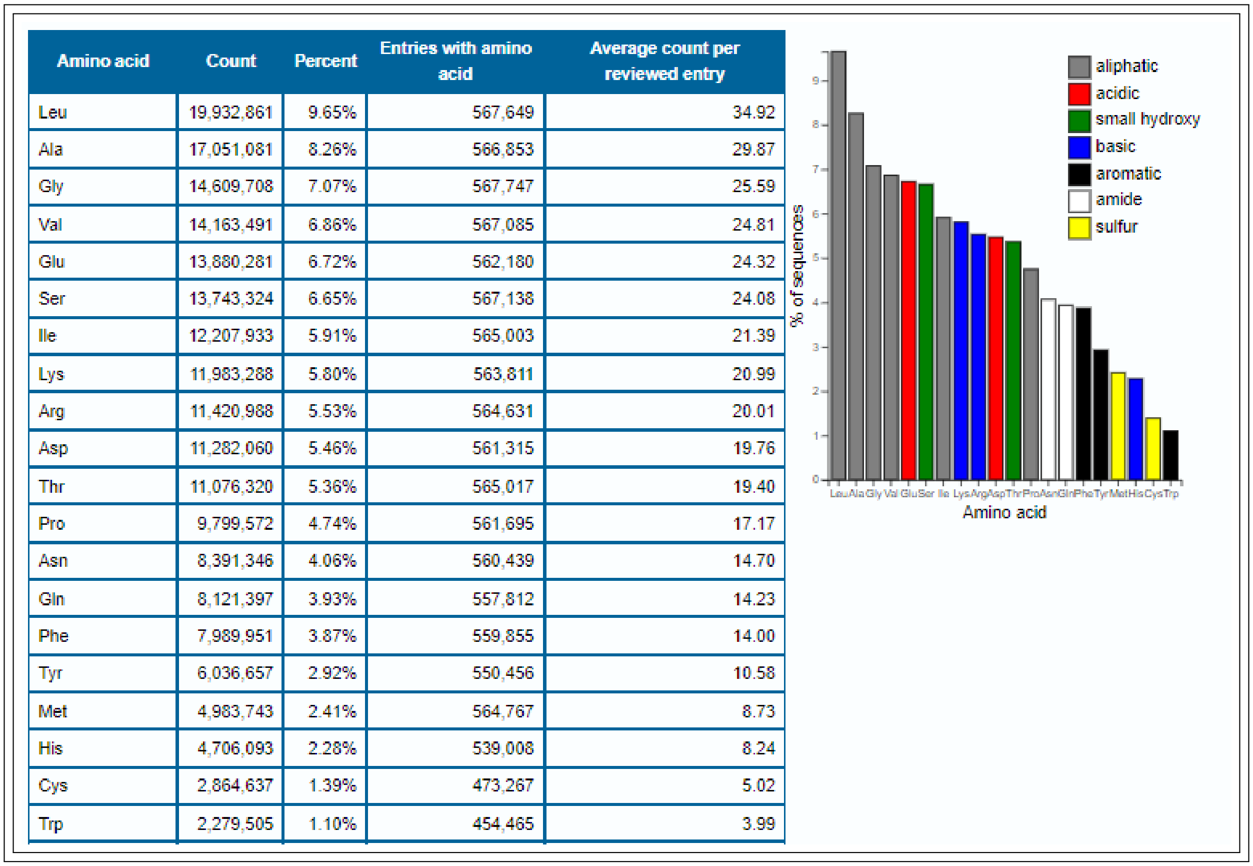


UniProtKB 1 result

■ Entry ▲		Entry Name ▲	Protein Names ▲	Gene Names ▲	Organism ▲	Length ▲
□ A2ASS6		TITIN_MOUSE	Titin [...]	Ttn	Mus musculus (Mouse)	35,213 AA

Figure 13. Longest sequence in UniProtKB

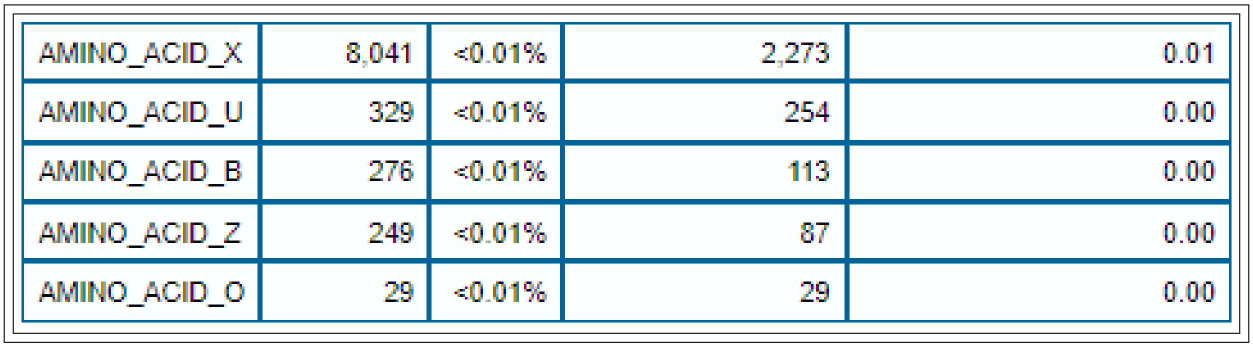
- (c) The table below shows the percentage composition of every amino acid.



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
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

Figure 14. Amino Acid Composition

- (c) Along with the above 20 amino acids, there are also a few other non-trivial amino acids as shown in table below.



AMINO_ACID_X	8,041	<0.01%	2,273	0.01
AMINO_ACID_U	329	<0.01%	254	0.00
AMINO_ACID_B	276	<0.01%	113	0.00
AMINO_ACID_Z	249	<0.01%	87	0.00
AMINO_ACID_O	29	<0.01%	29	0.00

Figure 15. Non-Trivial Amino Acid Composition