

BT3040 – Bioinformatics

Practical 3

1

The sequence of the protein is shown below:

The screenshot displays the UniProt entry for VDAC1 (P21796). The 'Sequence' tab is selected, showing the protein's amino acid sequence. The sequence is as follows:

```
MAVPPTYADL GKSARDVFTK GYGFGLIKLD LTKSENGLE FTSSGSANTE TTKVTGSLET KYRWTEYGLT FTEKWNTDNT
LGTEITVEDQ LARGKLKTFD SSFSPNTGKK NAKIKTGKVR EHINLGCDND FDIAGPSIRG ALVLGYEGWL AGYQMINFETA
KSRVTQSNFA VGYKTDEFQL HTNVNDGTEF GGSYQKVNK KLETAVNLAW TAGNSNTRFG IAAKYQIDPD ACFSKVNNS
SLIGLGYTQT LKPGIKLTLS ALLDGKNVNA GGHLGLGLE FQA
```

Additional information shown includes: Length 283, Mass (Da) 30,773, Last updated 2007-01-23 v2, and Checksum 89BA3378B04020D5.

The protein 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. The function of the protein is shown below:

The screenshot displays the UniProt entry for VDAC1 (P21796) with the 'Function' tab selected. The function is described as follows:

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 a weak anion selectivity whereas the closed state is cation-selective (PubMed:11845315, PubMed:18755977, PubMed:20230784, PubMed:8420959).

Binds various signaling molecules, including the sphingolipid ceramide, the phospholipid phosphatidylcholine, and the sterol cholesterol (PubMed:31015432).

In depolarized mitochondria, 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 (PubMed:32047033).

May participate in the formation of the permeability transition pore complex (PTPC) responsible for the release of mitochondrial products that triggers apoptosis (PubMed:15033708, PubMed:25296756).

May mediate ATP export from cells (PubMed:30061676).

The protein has 19 transmembrane segments.

The screenshot shows the UniProt entry for P21796 (VDAC1). The 'Family & Domains' section is highlighted, indicating that the protein consists mainly of a membrane-spanning beta-barrel formed by 19 beta-strands. It also mentions that the helical N-terminus folds back into the pore opening and plays a role in voltage-gated channel activity. The 'Keywords' section lists '#Transmembrane' and '#Transmembrane beta strand'. The 'Sequence similarities' section states that it belongs to the eukaryotic mitochondrial porin family.

2

The screenshot shows the UniRef 235 results page for a query of 'transcription factors'. The results are displayed in a table with columns for Cluster ID, Cluster name, Types, Size, Organisms, Length, and Identity. The first cluster, UniRef50_A0A9P7SZP2, is selected and shows 11 members, including Claviceps pusilla, Coniochaeta sp. 2T2.1, Pseudovirgaria hyperparasitica, Cryphonectria parasitica EP155, and Trichoderma parareesei (Filamentous fungus). The second cluster, UniRef50_A0A9Q0NQS2, is also selected and shows 1 member, Salix viminalis (Common).

Cluster ID	Cluster name	Types	Size	Organisms	Length	Identity
UniRef50_A0A9P7SZP2	Cluster: Transcription factors		11 members	Claviceps pusilla Coniochaeta sp. 2T2.1 Pseudovirgaria hyperparasitica Cryphonectria parasitica EP155 Trichoderma parareesei (Filamentous fungus) More organisms	393	
UniRef50_A0A9Q0NQS2	Cluster: HOMEBOX		1 member	Salix viminalis (Common)	377	

These sequences were downloaded in FASTA format.

Among all the protein sequences from *Homo sapiens*, there are

- 235,513 clusters with 100% identity with a total of 488,994 sequences.
- 105,990 clusters with 90% identity with a total of 2,060,774 sequences.
- 53,823 clusters with 50% identity with a total of 8,440,805 sequences.

uniref_taxonomy_id_9606_AND_identity_2024_02_22.xlsx - Excel

File Home Insert Draw Page Layout Formulas Data Review View Help Tell me what you want to do

Clipboard Font Alignment Number Styles Cells Editing Add-ins

D2

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Cluster ID	Cluster Na	Types	Size	Organism:	Length	Identity												
2	UniRef100 Cluster: R	UniProtKB		1	Homo sap	121	1.0												
3	UniRef100 Cluster: C	UniProtKB		1	Homo sap	380	1.0												
4	UniRef100 Cluster: N	UniProtKB		1	Homo sap	603	1.0												
5	UniRef100 Cluster: A	UniProtKB		1	Homo sap	226	1.0												
6	UniRef100 Cluster: N	UniProtKB		1	Homo sap	347	1.0												
7	UniRef100 Cluster: A	UniProtKB		1	Homo sap	226	1.0												
8	UniRef100 Cluster: N	UniProtKB		1	Homo sap	318	1.0												
9	UniRef100 Cluster: A	UniProtKB		1	Homo sap	226	1.0												
10	UniRef100 Cluster: N	UniProtKB		1	Homo sap	603	1.0												
11	UniRef100 Cluster: N	UniProtKB		1	Homo sap	603	1.0												
12	UniRef100 Cluster: N	UniProtKB		1	Homo sap	603	1.0												
13	UniRef100 Cluster: A	UniProtKB		1	Homo sap	226	1.0												
14	UniRef100 Cluster: N	UniProtKB		1	Homo sap	603	1.0												
15	UniRef100 Cluster: N	UniProtKB		1	Homo sap	603	1.0												
16	UniRef100 Cluster: C	UniProtKB		1	Homo sap	380	1.0												
17	UniRef100 Cluster: N	UniProtKB		1	Homo sap	347	1.0												
18	UniRef100 Cluster: N	UniProtKB		1	Homo sap	174	1.0												
19	UniRef100 Cluster: N	UniProtKB		1	Homo sap	603	1.0												
20	UniRef100 Cluster: A	UniProtKB		1	Homo sap	226	1.0												
21	UniRef100 Cluster: N	UniProtKB		1	Homo sap	603	1.0												

Sheet0

Ready Accessibility: Investigate Average: 2.079648539 Count: 235133 Sum: 488994

Search

uniref_taxonomy_id_9606_AND_identity_2024_02_22 (1).xlsx - Excel

File Home Insert Draw Page Layout Formulas Data Review View Help Tell me what you want to do

Clipboard Font Alignment Number Styles Cells Editing Add-ins

D2

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
105971	UniRef90 Cluster: Er	UniProtKB		3	Homo sap	197	0.9												
105972	UniRef90 Cluster: Ul	UniProtKB		9	Homo sap	164	0.9												
105973	UniRef90 Cluster: Sr	UniProtKB		19	Homo sap	174	0.9												
105974	UniRef90 Cluster: Pl	UniProtKB		3	Homo sap	66	0.9												
105975	UniRef90 Cluster: Ni	UniProtKB		2	Homo sap	53	0.9												
105976	UniRef90 Cluster: M	UniProtKB		3	Homo sap	97	0.9												
105977	UniRef90 Cluster: Pr	UniProtKB		3	Homo sap	129	0.9												
105978	UniRef90 Cluster: Sc	UniProtKB		6	Homo sap	225	0.9												
105979	UniRef90 Cluster: Di	UniProtKB		1	Homo sap	2061	0.9												
105980	UniRef90 Cluster: Al	UniProtKB		2	Homo sap	61	0.9												
105981	UniRef90 Cluster: Pr	UniProtKB		3	Homo sap	85	0.9												
105982	UniRef90 Cluster: Tr	UniProtKB		10	Homo sap	143	0.9												
105983	UniRef90 Cluster: Pr	UniProtKB		3	Homo sap	265	0.9												
105984	UniRef90 Cluster: Pr	UniProtKB		3	Homo sap	171	0.9												
105985	UniRef90 Cluster: L	UniProtKB		5	Homo sap	127	0.9												
105986	UniRef90 Cluster: Tl	UniProtKB		8	Homo sap	246	0.9												
105987	UniRef90 Cluster: tr	UniProtKB		3	Homo sap	212	0.9												
105988	UniRef90 Cluster: F	UniProtKB		3	Homo sap	298	0.9												
105989	UniRef90 Cluster: C	UniProtKB		4	Homo sap	218	0.9												
105990	UniRef90 Cluster: D	UniProtKB		22	Homo sap	176	0.9												
105991	UniRef90 Cluster: DI	UniProtKB		14	Homo sap	96	0.9												

Sheet0

Ready Accessibility: Investigate Average: 19.44309841 Count: 105990 Sum: 2060774

Search

uniref_taxonomy_id_9606_AND_identity_2024_02_22 (2).xlsx - Excel

File Home Insert Draw Page Layout Formulas Data Review View Help Tell me what you want to do

Clipboard Font Alignment Number Styles Cells Editing Add-ins

Calibri 11 A⁺ B I U Wrap Text General Conditional Formatting Format as Table Cell Styles Insert Delete Format Sort & Filter Find & Select Add-ins

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
53804	UniRef50_Cluster: GI	UniProtKB	18	Homo sap	198	0.5													
53805	UniRef50_Cluster: CI	UniProtKB	4	Homo sap	121	0.5													
53806	UniRef50_Cluster: RI	UniProtKB	10	Homo sap	129	0.5													
53807	UniRef50_Cluster: PI	UniProtKB	111	Homo sap	127	0.5													
53808	UniRef50_Cluster: SI	UniProtKB	13	Homo sap	74	0.5													
53809	UniRef50_Cluster: NI	UniProtKB	1	Homo sap	155	0.5													
53810	UniRef50_Cluster: LI	UniProtKB	1	Homo sap	324	0.5													
53811	UniRef50_Cluster: MI	UniProtKB	3	Homo sap	160	0.5													
53812	UniRef50_Cluster: DI	UniProtKB	7	Homo sap	881	0.5													
53813	UniRef50_Cluster: Pr	UniProtKB	4	Homo sap	124	0.5													
53814	UniRef50_Cluster: M	UniProtKB	3	Homo sap	46	0.5													
53815	UniRef50_Cluster: Zi	UniProtKB	3	Homo sap	41	0.5													
53816	UniRef50_Cluster: M	UniProtKB	15	Homo sap	241	0.5													
53817	UniRef50_Cluster: NI	UniProtKB	6	Homo sap	53	0.5													
53818	UniRef50_Cluster: AI	UniProtKB	2	Homo sap	61	0.5													
53819	UniRef50_Cluster: Pr	UniProtKB	3	Homo sap	85	0.5													
53820	UniRef50_Cluster: Pr	UniProtKB	3	Homo sap	171	0.5													
53821	UniRef50_Cluster: L-	UniProtKB	15	Homo sap	127	0.5													
53822	UniRef50_Cluster: tr	UniProtKB	7	Homo sap	212	0.5													
53823	UniRef50_Cluster: F-	UniProtKB	6	Homo sap	298	0.5													
53824	UniRef50_Cluster: Cc	UniProtKB	10	Homo sap	218	0.5													

Sheet0

Ready Accessibility: Good to go Average: 156.825242 Count: 53823 Sum: 8440805

10:55 AM 22-02-2024

4

There are 17,201 *Mus musculus* protein sequences that have been manually annotated / reviewed.

(organism_id:10090) AND (reviewed) x

https://www.uniprot.org/uniprotkb?query=%28organism_id%3A10090%29+AND+%28reviewed%3Atrue%29

Gmail Rosalind Desmos Library Genesis Anna's Archive SciHub Summer 2024

UniProtKB BLAST Align Peptide search ID mapping SPARQL (organism_id:10090) AND (reviewed) Advanced List Search Help

Status
Reviewed (Swiss-Prot)
(17,201)

Popular organisms
Mouse (17,201)

Taxonomy
10090 x
Filter by taxonomy

Group by
Taxonomy
Keywords
Gene Ontology
Enzyme Class

UniProtKB 17,201 results

proteome UP000000589

BLAST Align Map IDs Download Add View: Cards Table Customize columns Share

Entry	Entry Name	Protein Names	Gene Names	Organism	Length
<input type="checkbox"/> A0A5F8MPU3	CTSRT_MOUSE	Cation channel sperm-associated targeting subunit tau[...]	C2cd6, Als2cr11, CATSPERT	Mus musculus (Mouse)	2,282 AA
<input type="checkbox"/> A0A5K7RLP0	MEIOS_MOUSE	Meiosis Initiator protein	Meiosin, Bhmg1	Mus musculus (Mouse)	589 AA
<input type="checkbox"/> A2A891	CMTA1_MOUSE	Calmodulin-binding transcription activator 1	Camta1, Kiaa0833	Mus musculus (Mouse)	1,682 AA
<input type="checkbox"/> A2A9F4	KDF1_MOUSE	Keratinocyte differentiation factor 1	Kdf1	Mus musculus (Mouse)	397 AA

10:44 AM 22-02-2024

Of these manually annotated sequences, 2,351 are associated with a 3D structure in PDB.

The screenshot shows the UniProtKB search results page. The search query is "(organism_id:10090) AND (reviewed)". The results are displayed in a table with columns: Entry, Entry Name, Protein Names, Gene Names, Organism, and Length. The first four results are shown:

Entry	Entry Name	Protein Names	Gene Names	Organism	Length
A2CG49	KALRN_MOUSE	Kalirin[...]	Kalrn	Mus musculus (Mouse)	2,964 AA
B2RXS4	PLXB2_MOUSE	Plexin-B2	Plxnb2	Mus musculus (Mouse)	1,842 AA
B5KM66	SYCE3_MOUSE	Synaptonemal complex central element protein 3[...]	Syce3, Tseg2	Mus musculus (Mouse)	88 AA
E9Q9F6	CTSRD_MOUSE	Cation channel sperm-associated auxiliary subunit delta[...]	Catsperd, Tmem146	Mus musculus	805 AA

5

The first 10 entries obtained above were mapped to the STRING database.

The screenshot shows the UniProt Retrieve/ID mapping page. The page title is "Retrieve/ID mapping". Below the title, there is a text input field containing the following IDs: A2CG49, B2RXS4, B5KM66, E9Q9F6, E9Q9W7, J3QMY9, and 008763. Below the input field, there is a message: "Your input contains 10 IDs". Below the message, there are two dropdown menus: "From database" (set to "UniProtKB AC/ID") and "To database" (set to "STRING"). At the bottom right, there are two buttons: "Reset" and "Map 10 IDs".

10 STRING IDs were mapped.

Retrieve/ID mapping results | UniProt

https://www.uniprot.org/id-mapping/9a866fd8d124e8bd9ea6176ce358c0dce416e8dd/overview

UniProt BLAST Align Peptide search ID mapping SPARQL Tool results Advanced | List Search

ID mapping 10 results found for UniProtKB_AC-ID → STRING

Overview Input Parameters API Request

Download View: Cards Table Resubmit

10 IDs were mapped to 10 results

From	To
A2CG49	10090.ENSMUSP00000110611
B2RXS4	10090.ENSMUSP00000104955
B5KM66	10090.ENSMUSP00000131766

6

From the UniProt statistics, we can see that most sequences are 200 amino acids long. Longer sequences are less frequently occurring in UniProt.

UniProtKB | Statistics | UniProt

https://www.uniprot.org/uniprotkb/statistics#sequence-size

UniProt BLAST Align Peptide search ID mapping SPARQL UniProtKB Advanced | List Search

Introduction Taxonomic origin Sequence size Journal citations Statistics for some line types Amino acid composition Miscellaneous statistics

Reviewed (Swiss-Prot) Unreviewed (TrEMBL)

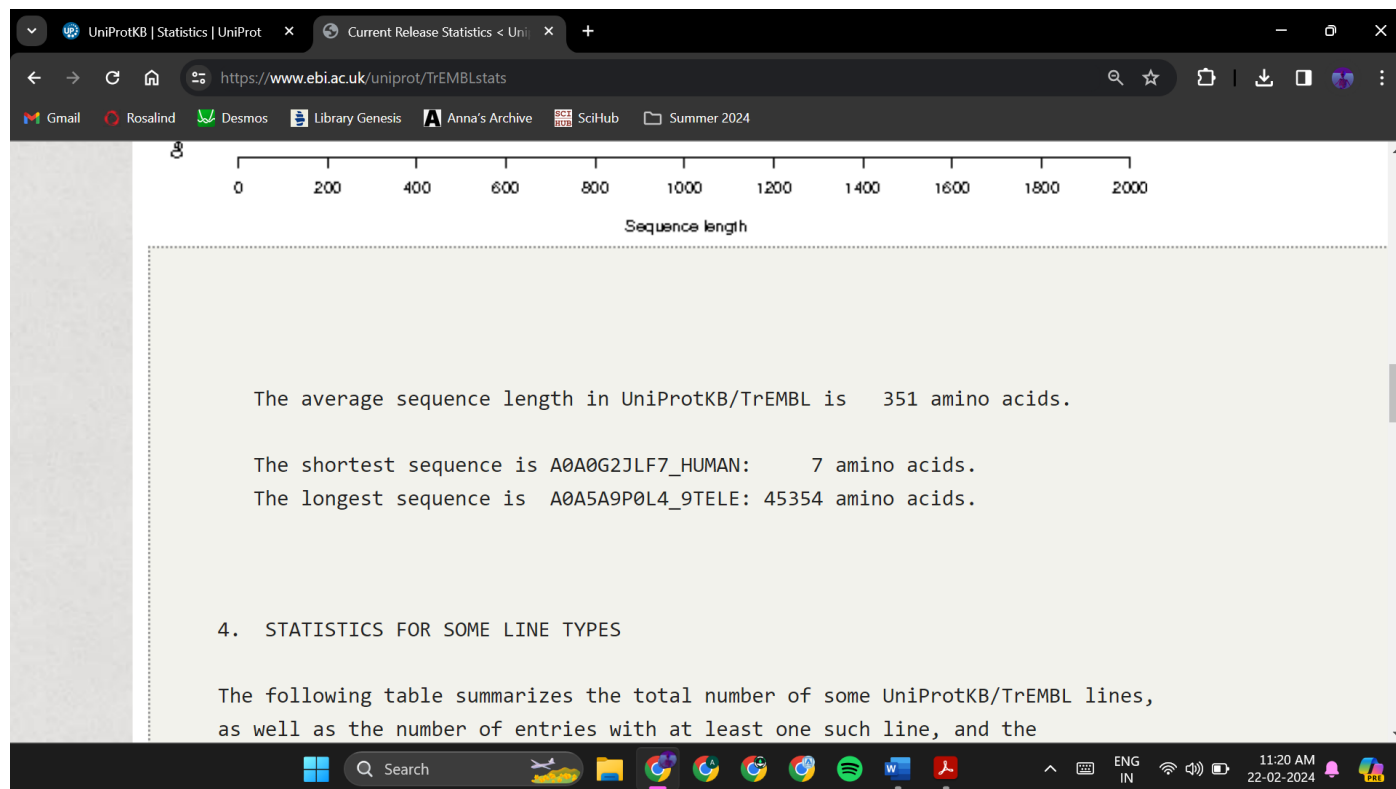
Number of sequences

Sequence length

The average sequence length in UniProtKB/TrEMBL is 351 amino acids.

The shortest sequence is A0A0G2JLF7_HUMAN: 7 amino acids.

The longest sequence is A0A5A9P0L4_9TELE: 45354 amino acids.



The amino acid composition of the UniProt database is shown below:

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

Amino acid	Count	Percent	Entries with amino acid	Average count per reviewed entry
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

UniProtKB | Statistics | UniProt

[https://www.uniprot.org/uniprotkb/statistics#amino-acid-composition](#)

[Gmail](#)
[Rosalind](#)
[Desmos](#)
[Library Genesis](#)
[Anna's Archive](#)
[SciHub](#)
[Summer 2024](#)

UniProt
BLAST
Align
Peptide search
ID mapping
SPARQL
UniProtKB

Advanced
List
Search

Feedback
Help

Introduction
Taxonomic origin
Sequence size
Journal citations
Statistics for some line types
Amino acid composition
Miscellaneous statistics

Amino acid composition

Reviewed (Swiss-Prot)

Unreviewed (TrEMBL)

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

% of sequences

aliphatic

acidic

small hydroxy

basic

aromatic

amide

sulfur

Search

ENG

IN

11:13 AM 22-02-2024