# BT 3040: BIOINFORMATICS

**Assignment 3**

Atharva Mandar Phatak | BE21B009
Department of Biotechnology

Indian Institute of Technology
Madras

**Q1) 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?**

A1) Sequence of human mitochondrial beta barallel membrane protein VDAC

```
SQ   SEQUENCE   283 AA;  30773 MW;  89BA3378B04020D5 CRC64;
     MAVPPTYADL GKSARDVFTK GYGFGLIKLD LKTKSENGLE FTSSGSANTE TTKVTGSLET
     KYRWTEYGLT FTEKWNTDNT LGTEITVEDQ LARGLKLTFD SSFSPNTGKK NAKIKTGYKR
     EHINLGCDMD FDIAGPSIRG ALVLGYEGWL AGYQMNFETA KSRVTQSNFA VGYKTDEFQL
     HTNVNDGTEF GGSIYQKVNK KLETAVNLAW TAGNSNTRFG IAAKYQIDPD ACFSAKVNNS
     SLIGLGYTQT LKPGIKLTLS ALLDGKNVNA GGHKLGLGLE FQA
```

A2) Function:

## Function

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**). 9 Publications

B) Number of transmembrane sequence present in the protein = **19**

| | | | | |
|---|---|---|---|---|
| Transmembrane | 26-35 | Beta stranded | 2 Publications | BLAST Add |
| Transmembrane | 39-47 | Beta stranded | 2 Publications | BLAST Add |
| Transmembrane | 54-64 | Beta stranded | 2 Publications | BLAST Add |
| Transmembrane | 69-76 | Beta stranded | 2 Publications | BLAST Add |
| Transmembrane | 80-89 | Beta stranded | 2 Publications | BLAST Add |
| Transmembrane | 95-104 | Beta stranded | 2 Publications | BLAST Add |
| Transmembrane | 111-120 | Beta stranded | 2 Publications | BLAST Add |
| Transmembrane | 123-130 | Beta stranded | 2 Publications | BLAST Add |
| Transmembrane | 137-145 | Beta stranded | 2 Publications | BLAST Add |
| Transmembrane | 150-158 | Beta stranded | 2 Publications | BLAST Add |
| Transmembrane | 163-175 | Beta stranded | 2 Publications | BLAST Add |
| Transmembrane | 178-185 | Beta stranded | 2 Publications | BLAST Add |
| Transmembrane | 189-198 | Beta stranded | 2 Publications | BLAST Add |
| Transmembrane | 202-211 | Beta stranded | 2 Publications | BLAST Add |
| Transmembrane | 218-227 | Beta stranded | 2 Publications | BLAST Add |
| Transmembrane | 231-238 | Beta stranded | 2 Publications | BLAST Add |
| Transmembrane | 242-251 | Beta stranded | 2 Publications | BLAST Add |
| Transmembrane | 254-263 | Beta stranded | 2 Publications | BLAST Add |
| Transmembrane | 273-282 | Beta stranded | 2 Publications | BLAST Add |

**Q2) Obtain the sequences of "transcription factors" with 50% sequence identity in FASTA format. List the count of sequences and count of clusters.**
**http://www.uniprot.org/uniprot/**

| Cluster ID | Cluster name | Types | Size | Organisms | Length | Identity |
|---|---|---|---|---|---|---|
| | | | | More organisms | | |
| UniRef50_A0A0M4FLP9 | Cluster: NAC transcription factors 38 | | 3 members | Manihot esculenta (Cassava)<br>Corchorus capsularis (Jute)<br>Hevea brasiliensis | 200 | UniRef50 |
| UniRef50_A0A0M4FLS3 | Cluster: NAC transcription factors 88 | | 1 member | Manihot esculenta (Cassava) | 264 | UniRef50 |
| UniRef50_A0A9Q0QM67 | Cluster: HOMEOBOX PROTEIN TRANSCRIPTION FACTORS | | 1 member | Salix koriyanagi | 351 | UniRef50 |
| UniRef50_A0A9Q0W5E4 | Cluster: HOMEOBOX PROTEIN TRANSCRIPTION FACTORS | | 1 member | Salix purpurea (Purple osier willow) | 90 | UniRef50 |
| UniRef50_A0A0M4FET5 | Cluster: NAC transcription factors 45 | | 1 member | Manihot esculenta (Cassava) | 82 | UniRef50 |
| UniRef50_A0A0M4FKY5 | Cluster: NAC transcription factors 39 | | 1 member | Manihot esculenta (Cassava) | 129 | UniRef50 |
| UniRef50_A0A0M5JF80 | Cluster: NAC transcription factors 42 | | 1 member | Manihot esculenta (Cassava) | 124 | UniRef50 |
| UniRef50_A0A2P2JCL1 | Cluster: NAC transcription factors 13 | | 1 member | Rhizophora mucronata (Asiatic mangrove) | 289 | UniRef50 |
| UniRef50_A0A9Q0P6B9 | Cluster: HOMEOBOX PROTEIN TRANSCRIPTION FACTORS | | 1 member | Salix koriyanagi | 384 | UniRef50 |
| UniRef50_A0A9Q0VVN3 | Cluster: HOMEOBOX PROTEIN TRANSCRIPTION FACTORS | | 1 member | Salix purpurea (Purple osier willow) | 275 | UniRef50 |
| UniRef50_A0A0G2SJ85 | Cluster: AP3/EREBP Transcription Factors (Fragment) | | 1 member | Salvia miltiorrhiza (Chinese sage) | 192 | UniRef50 |
| UniRef50_A0A5B6ZXL9 | Cluster: Putative NAC transcription factors 48 | | 1 member | Davidia involucrata (Dove tree) | 187 | UniRef50 |
| UniRef50_A0A0G2SJB9 | Cluster: AP11/EREBP Transcription Factors (Fragment) | | 1 member | Salvia miltiorrhiza (Chinese sage) | 226 | UniRef50 |
| UniRef50_A0A0M3R857 | Cluster: NAC transcription factors 43 (Fragment) | | 4 members | Manihot esculenta (Cassava) | 367 | UniRef50 |

The sequences' FASTA file is in the folder named 'Q2_ transcription_factors'

**Q3) How many protein sequences from Homo sapiens are obtained at identity cutoff of 100%, 90% and 50% sequence identity?**

**Clusters**

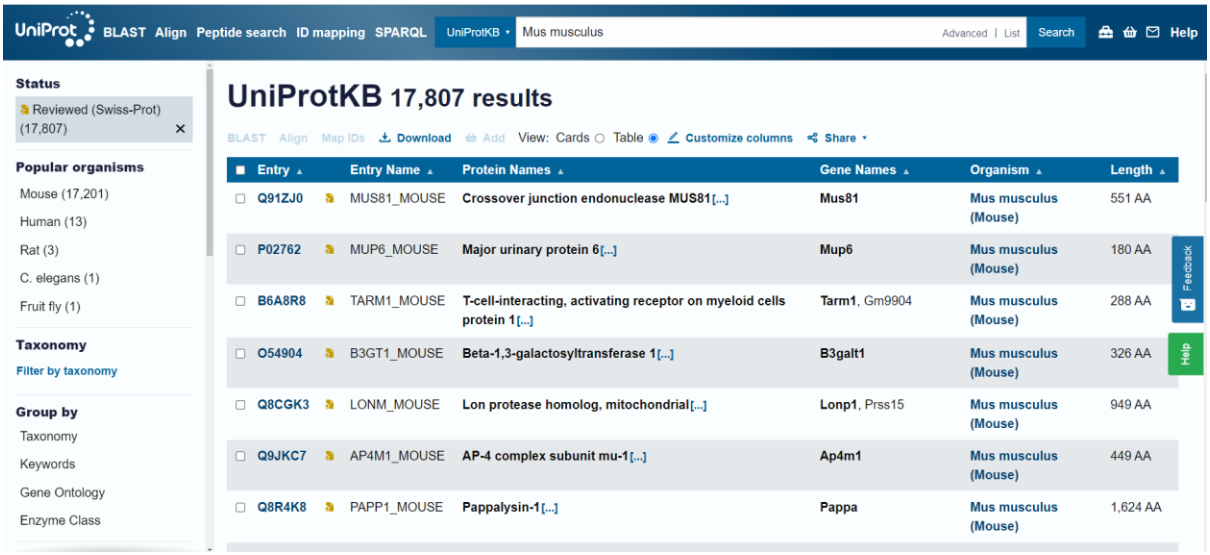
100% (236,474)

90% (106,848)

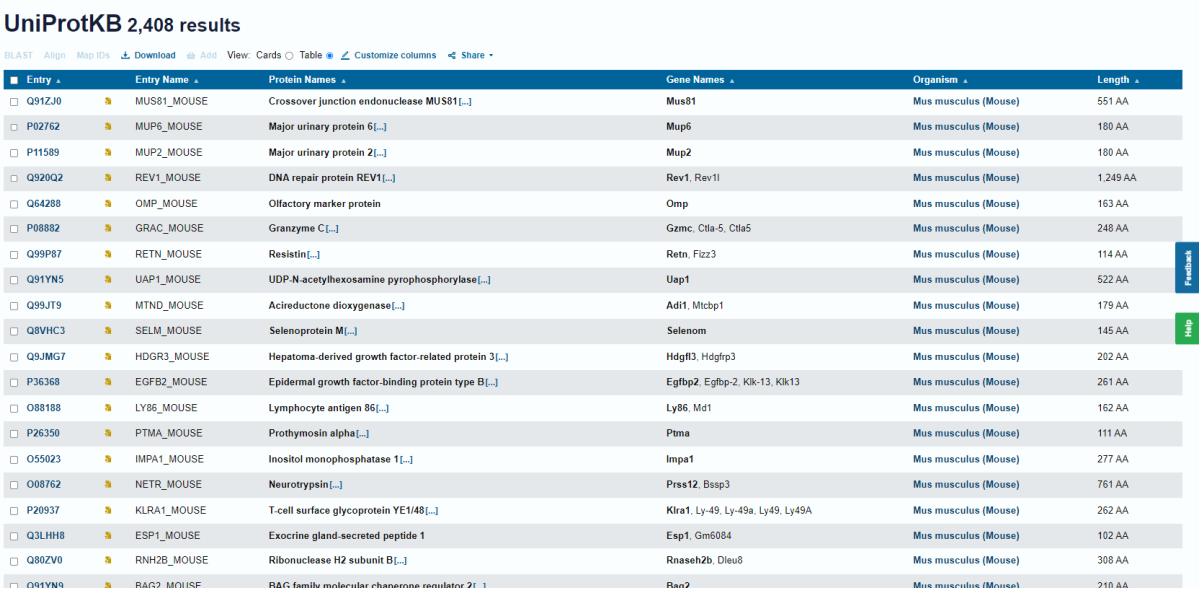
50% (54,154)

**Q4) 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)?**

a) Manually annotated sequences are **17807** of (Mus musculus) protein, under the Swiss-Prot Database.




b) There are **2408** manually annotated protein sequences are associated with PDB (3D structures)



**Q5) Map first 10 UniProt IDs of above manually curated mouse protein sequences with 3D structures to STRING database. How many STRING IDs are mapped?**

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 |
|---|---|
| Q91ZJ0 | 10090.ENSMUSP00000114895 |
| P02762 | 10090.ENSMUSP00000079442 |
| P11589 | 10090.ENSMUSP00000095655 |
| Q920Q2 | 10090.ENSMUSP00000027251 |
| Q64288 | 10090.ENSMUSP00000095882 |
| P08882 | 10090.ENSMUSP00000015585 |
| Q99P87 | 10090.ENSMUSP00000133024 |
| Q91YN5 | 10090.ENSMUSP00000106983 |
| Q99JT9 | 10090.ENSMUSP00000020957 |
| Q8VHC3 | 10090.ENSMUSP00000092041 |

Feedback

Help

10 STRING IDs are mapped.

**Q6) 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**

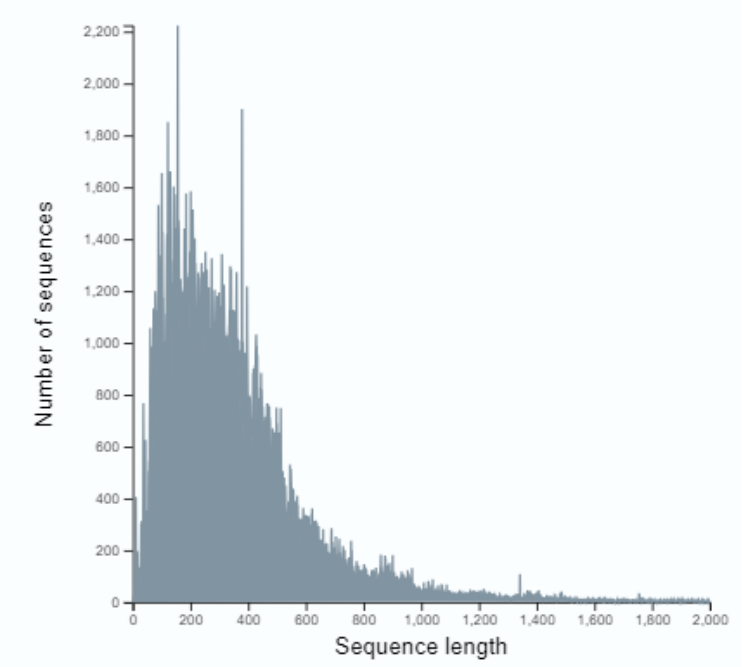
The following link has UniProt Statistics as of Feb 2024:
https://www.uniprot.org/uniprotkb/statistics#statistics-for-some-line-type

UniProtKB/TrEMBL : https://www.ebi.ac.uk/uniprot/TrEMBLstats

UniProtKB/Swiss-Prot: https://web.expasy.org/docs/relnotes/relstat.html

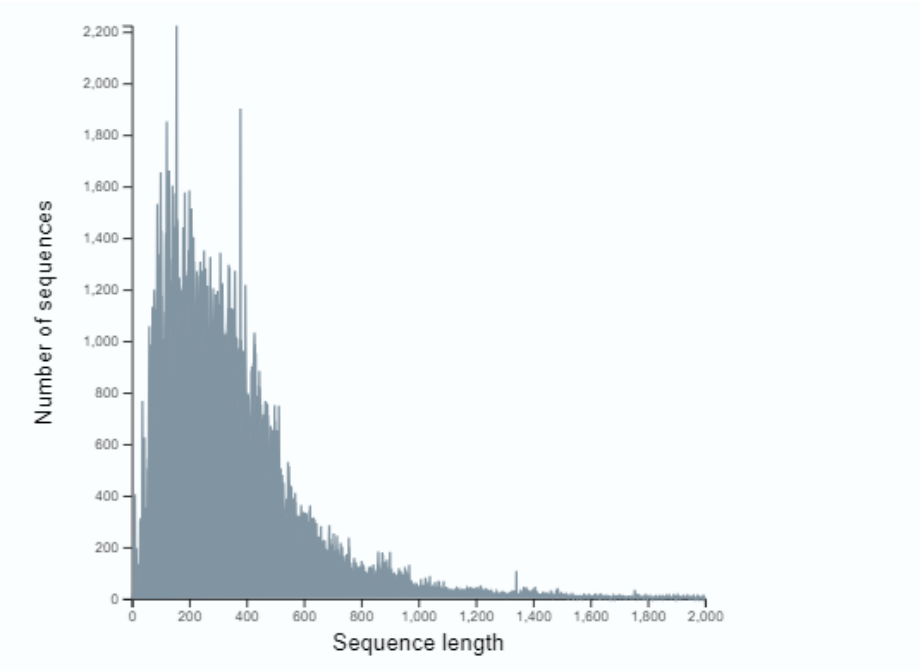
a) The sequence distribution chart plots the Number of sequences vs Sequence Length. We see a peak in the graph at range ~351 (TrEMBL) and ~361 (Swiss-Prot). The distribution is right-skewed, indicating more amino acids lie in the range of 0-600. We can also infer that it indicates the number of amino acids in the canonical sequence displayed by default in the entry's Sequence section.

a1) Reviewed (Swiss-Prot)




The average sequence length in UniProtKB/Swiss-Prot is 361 amino acids.

a2) Unreviewed (TrEMBL)



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

b)

b1) The shortest and longest sequence (in TrEMBL) are as follows:

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

b2) The shortest and longest sequence (in SwissProt) are as follows:

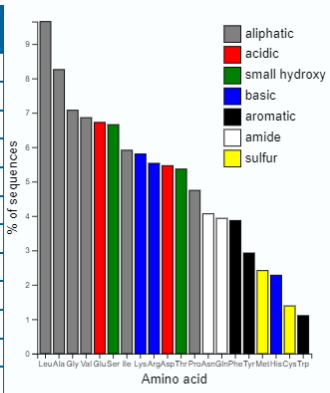
```
The shortest sequence is   GWA_SEPOF (P83570):     2 amino acids.

The longest sequence is  TITIN_MOUSE (A2ASS6): 35213 amino acids.
```

c) The amino acid composition is as follows:

c1) Reviewed (Swiss-Prot)

| 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 |
| 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 |



c2) Unreviewed (TrEMBL)

| Amino acid | Count | Percent | Entries with amino acid | Average count per unreviewed entry |
|---|---|---|---|---|
| Leu | 8,646,856,405 | 9.84% | 249,303,314 | 34.62 |
| Ala | 7,915,577,778 | 9.01% | 248,721,380 | 31.69 |
| Gly | 6,378,270,739 | 7.26% | 248,495,438 | 25.54 |
| Val | 6,027,240,933 | 6.86% | 248,735,049 | 24.13 |
| Ser | 6,012,325,446 | 6.85% | 248,832,843 | 24.07 |
| Glu | 5,489,092,503 | 6.25% | 247,424,325 | 21.98 |
| Arg | 5,135,951,097 | 5.85% | 247,778,023 | 20.56 |
| Thr | 4,881,727,427 | 5.56% | 248,187,256 | 19.55 |
| Ile | 4,854,708,881 | 5.53% | 247,736,715 | 19.44 |
| Asp | 4,809,963,037 | 5.48% | 247,082,650 | 19.26 |
| Pro | 4,399,981,178 | 5.01% | 246,327,937 | 17.62 |
| Lys | 4,342,592,769 | 4.94% | 243,345,982 | 17.39 |
| Phe | 3,413,266,470 | 3.89% | 245,203,498 | 13.67 |
| Gln | 3,345,659,092 | 3.81% | 244,843,857 | 13.40 |
| Asn | 3,339,513,148 | 3.80% | 243,271,710 | 13.37 |
| Tyr | 2,529,285,941 | 2.88% | 240,531,652 | 10.13 |
| Met | 2,047,944,763 | 2.33% | 247,749,460 | 8.20 |
| His | 1,956,609,024 | 2.23% | 235,782,076 | 7.83 |
| Trp | 1,143,655,671 | 1.30% | 215,592,125 | 4.58 |
| Cys | 1,139,776,125 | 1.30% | 207,129,597 | 4.56 |
| AMINO_ACID_X | 22,129,989 | 0.03% | 2,929,800 | 0.09 |
| AMINO_ACID_U | 20,769 | <0.01% | 19,804 | 0.00 |
| AMINO_ACID_B | 19,992 | <0.01% | 17,869 | 0.00 |
| AMINO_ACID_Z | 7,598 | <0.01% | 7,192 | 0.00 |
| AMINO_ACID_O | 311 | <0.01% | 305 | 0.00 |