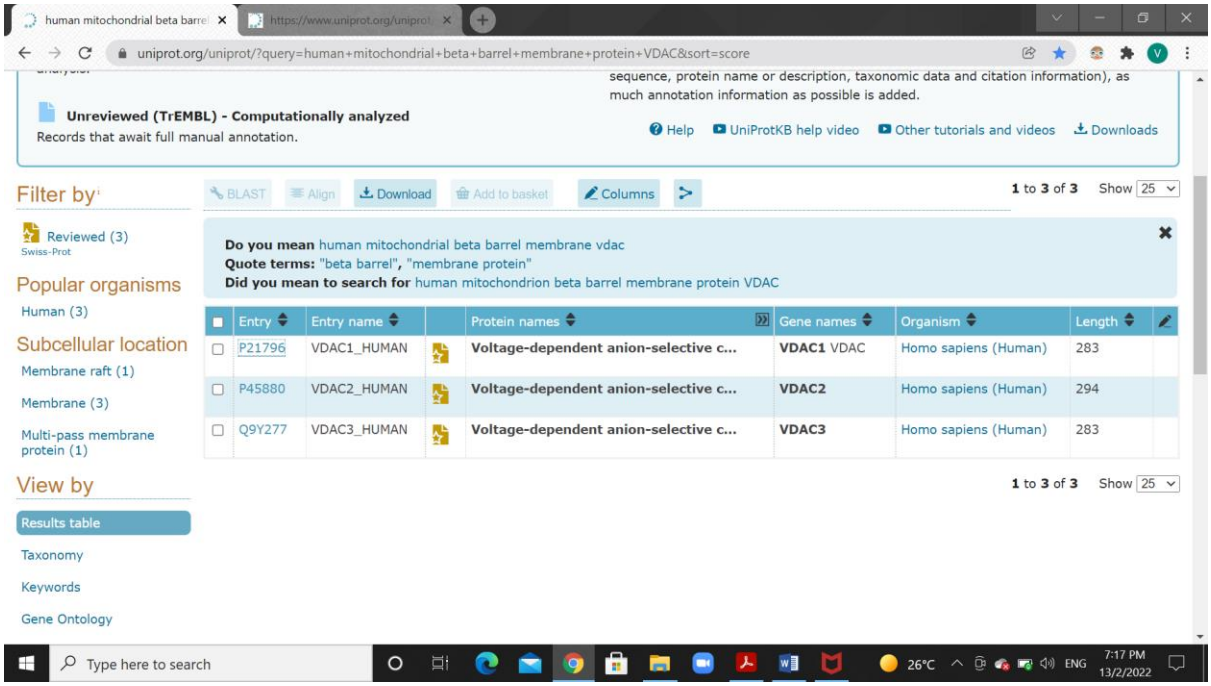
Practicals 3

-BS19B032

-R. Vasantha Kumar

1) There are three human mitochondrial β barrel membrane protein VDAC: VDAC 1, VDAC 2, VDAC 3.



VDAC 1:

MAVPPTYADLGKSARDVFTKGYGFGLIKLDLKTKSENGLEFTSSGSANTETTKVTGSLET
KYRWTEYGLTFTEKWNTDNTLGTEITVEDQLARGLKLTFDSSFSPNTGKKNAKIKTGYKR
EHINLGCDMDFDIAGPSIRGALVLGYEGWLAGYQMNFETAKSRVTQSNFAVGYKTDEFQL
HTNVNDGTEFGGSIYQKVNKKLETAVNLAWTAGNSNTRFGIAAKYQIDPDACFSAKVNNS
SLIGLGYTQTLKPGIKLTLSALLDGKNVNAGGHKLGLGLEFQA

Its function is to form channels in mitochondrial outer membrane and plasma membrane for diffusion of hydrophilic molecules.

VDAC 2:

MATHGQTCARPMCIPPSYADLGKAARDIFNKGFGFGLVKLDVKTKSCSGVEFSTSGSSNT
DTGKVTGTLETKYKWCEYGLTFTEKWNTDNTLGTEIAIEDQICQGLKLTFDTTFSPNTGK
KSGKIKSSYKRECINLGCDVDFDFAGPAIHGSAVFGYEGWLAGYQMTFDSAKSKLTRNNF
AVGYRTGDFQLHTNVNDGTEFGGSIYQKVCEDLDTSVNLAWTSGTNCTRFGIAAKYQLDP
TASISAKVNNSSLIGVGYTQTLRPGVKLTLSALVDGKSINAGGHKVGLALELEA

Its function is similar to VDAC 1 to form channels in plasma membrane.

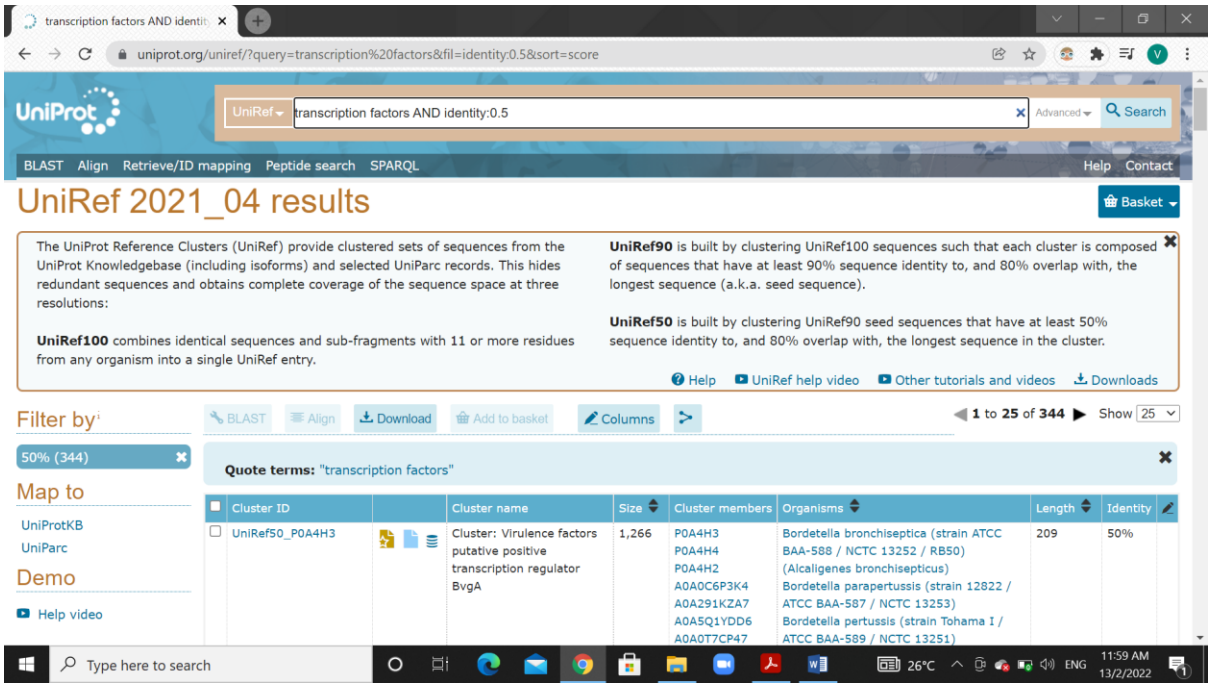
VDAC 3:

MCNTPTYCDLGKAAKDVFNKGYGFGMVKIDLKTKSCSGVEFSTSGHAYTDTGKASGNLET
KYKVCNYGLTFTQKWNTDNTLGTEISWENKLAEGLKLTLDTIFVPNTGKKSGKLKASYKR
DCFSVGSNVDIDFSGPTIYGWAVLAFEGWLAGYQMSFDTAKSKLSQNNFALGYKAADFQL
HTHVNDGTEFGGSIYQKVNEKIETSINLAWTAGSNNTRFGIAAKYMLDCRTSLSAKVNNA
SLIGLGYTQTLRPGVKLTLSALIDGKNFSAGGHKVGLGFELEA

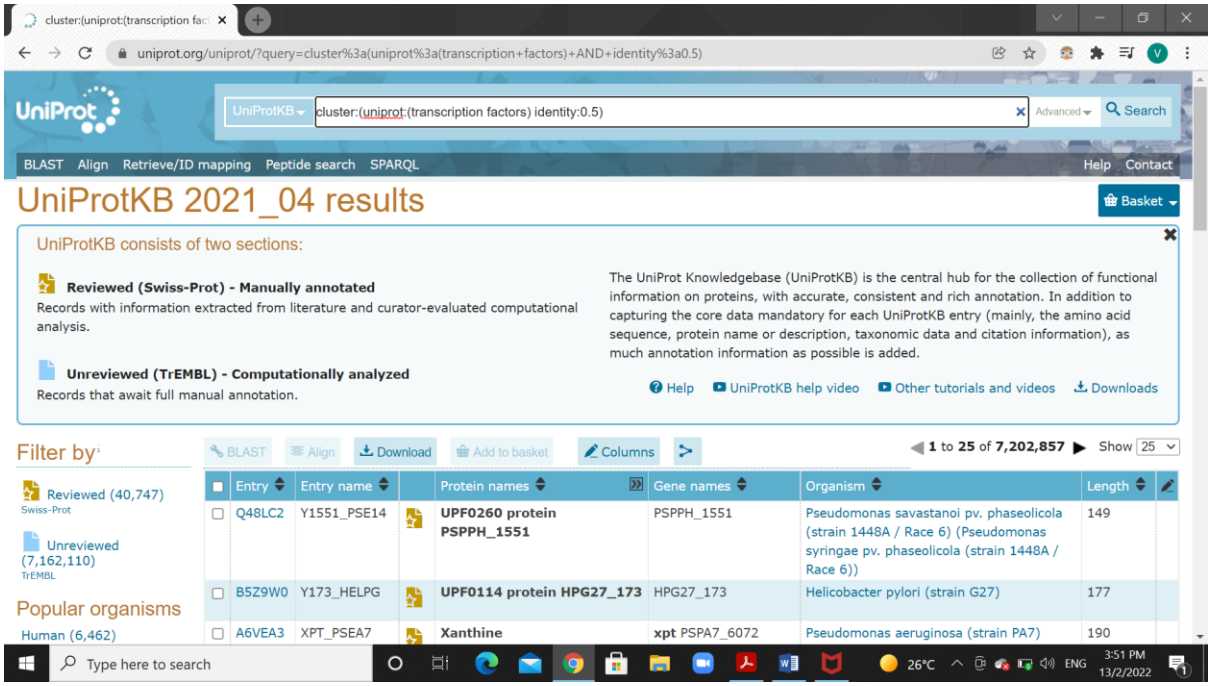
Function similar to VDAC 1 and VDAC 2.

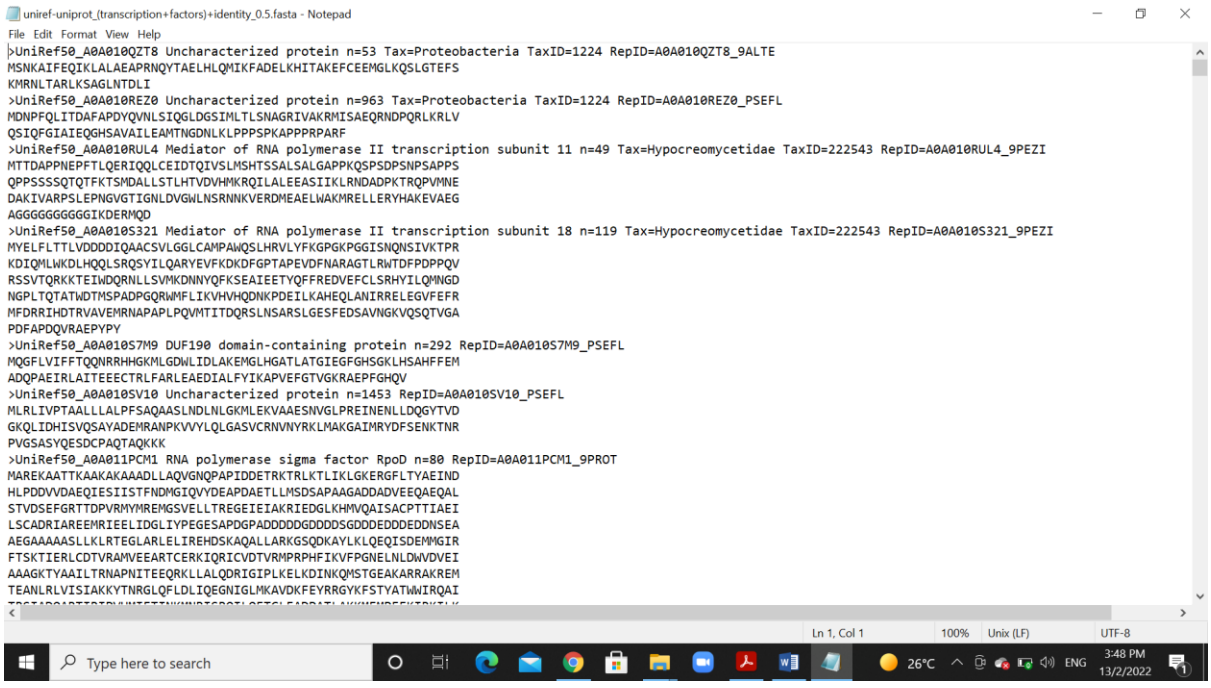
Only one transmembrane segments.

2) The number of clusters for transcription factors with 50% sequence identity is 344.



The number of sequences for transcription factors with 50% sequence identity is 7,202,857.



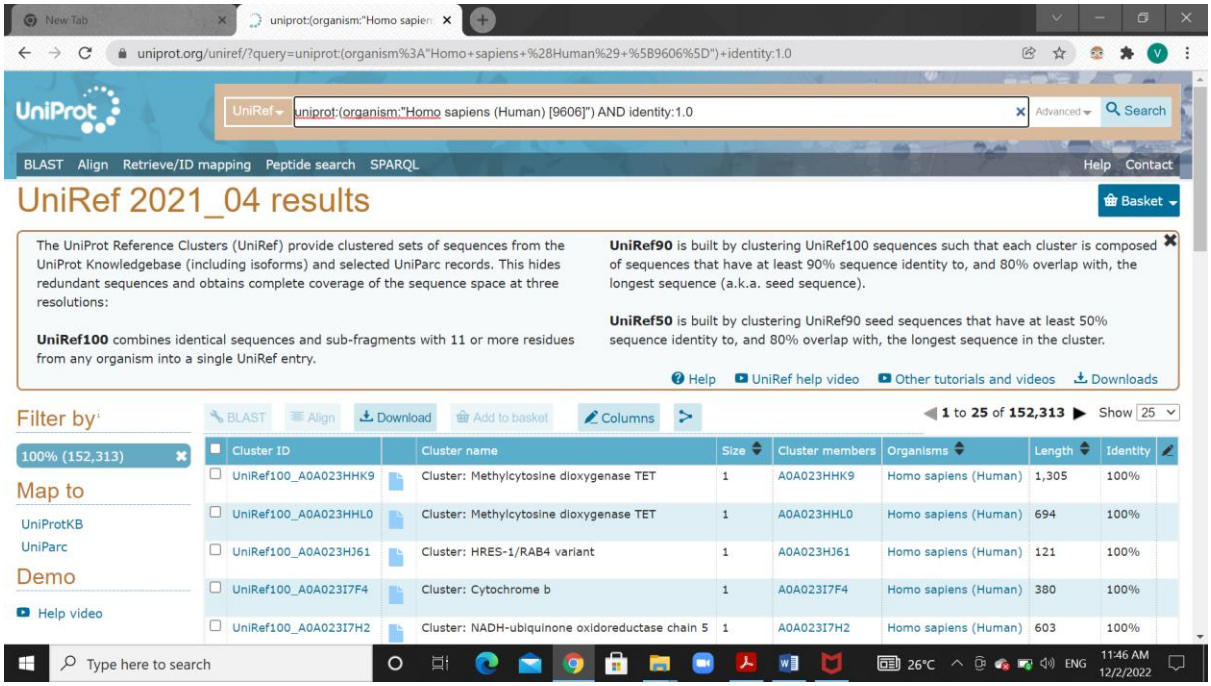


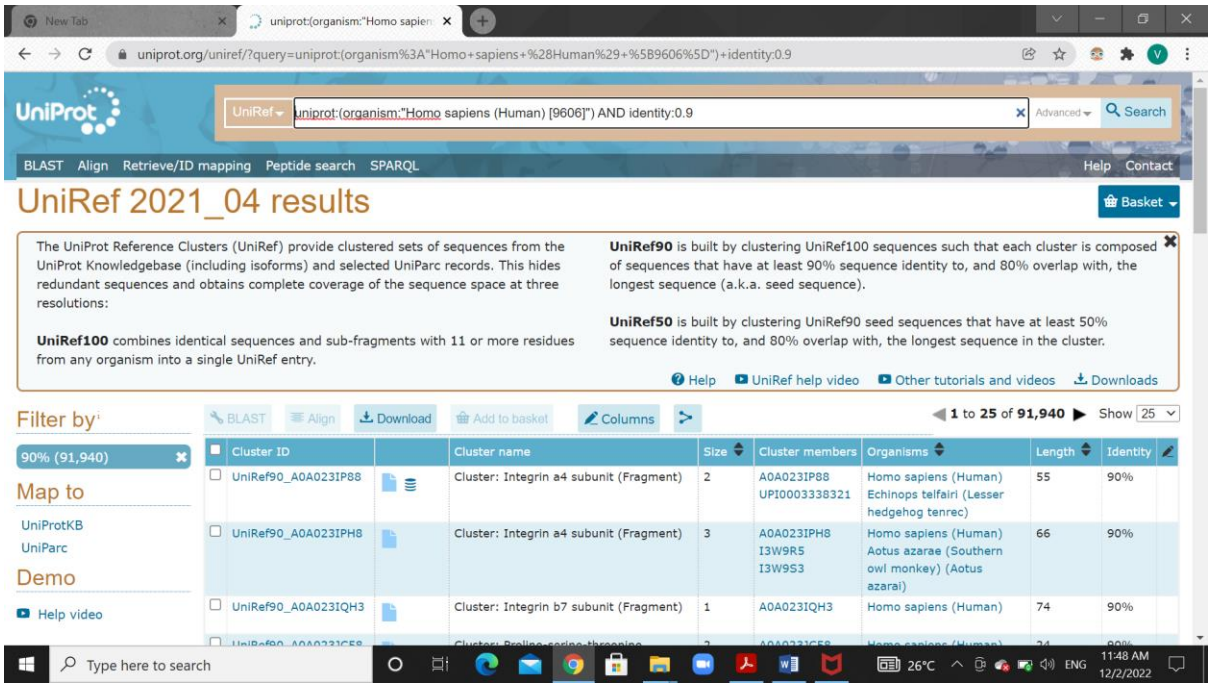
3) The number of protein sequences from Homo sapiens that are obtained at identity cutoff of 100%, 90% and 50% sequence identity are:

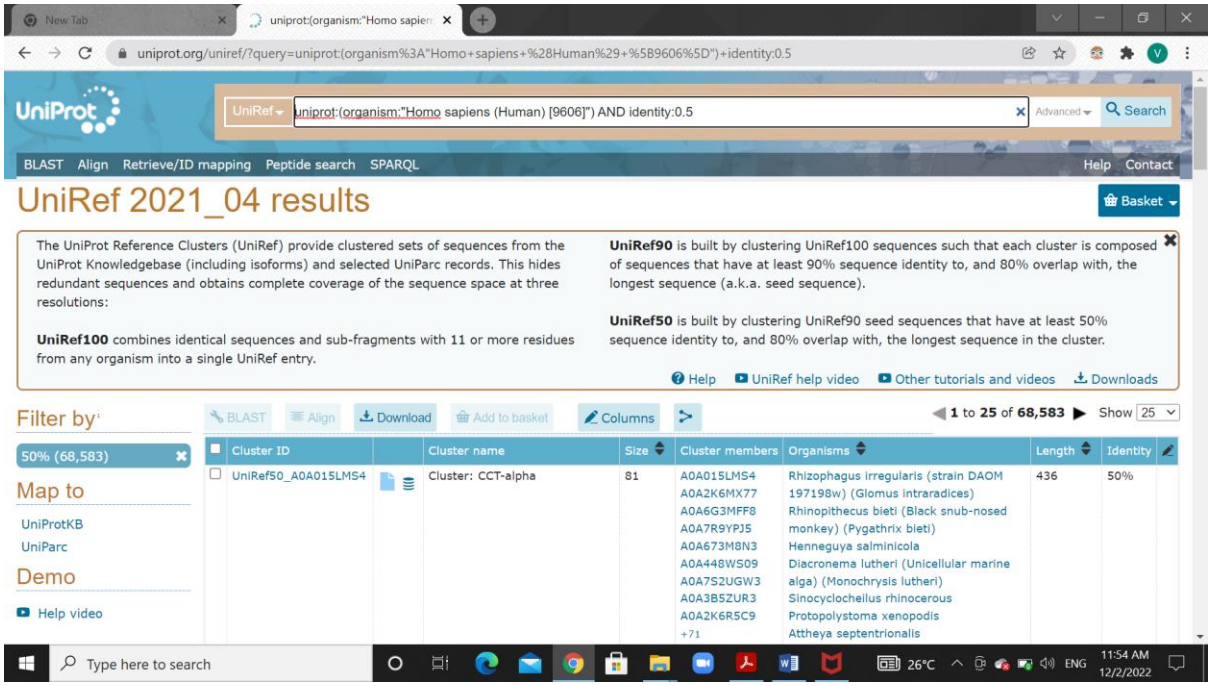
100%: 152,313

90%: 91,940

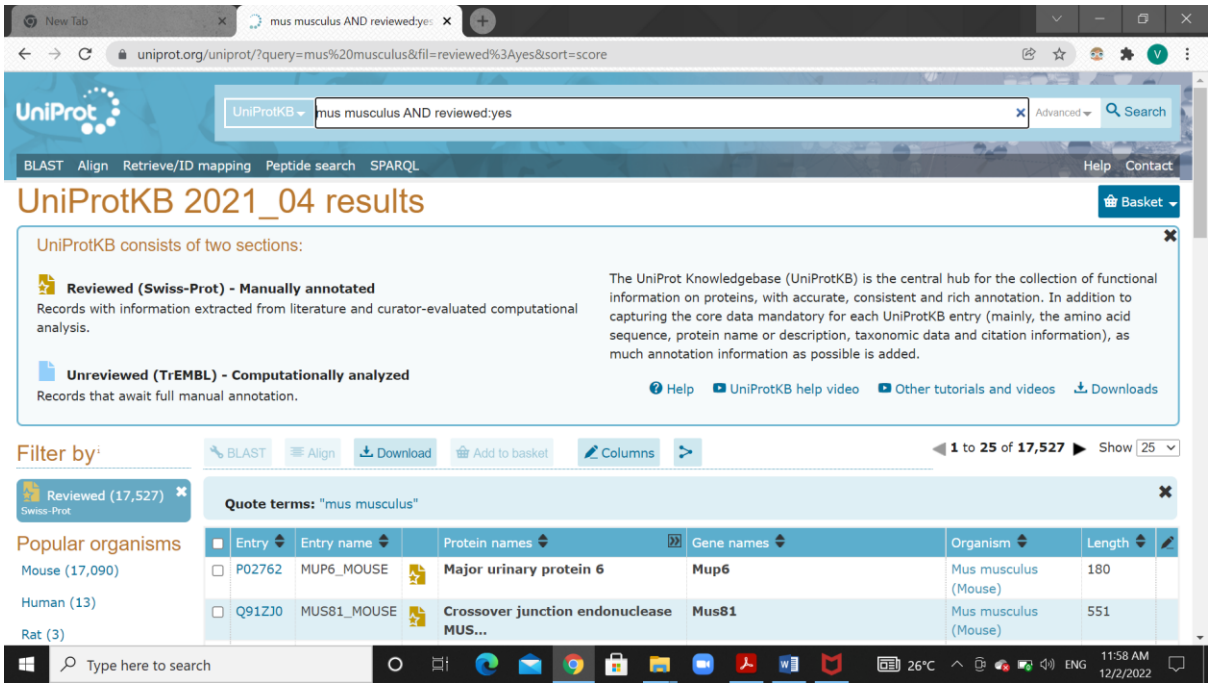
50%: 68,583


New Tab
uniprot:(organism:"Homo sapien
uniprot.org/uniref/?query=uniprot:(organism%3A"Homo+sapiens+%28Human%29+%5B9606%5D")+identity:1.0
UniProt
UniRef
uniprot:(organism:"Homo sapiens (Human) [9606]") AND identity:1.0
Advanced
Search
BLAST Align Retrieve/ID mapping Peptide search SPARQL
Help Contact
UniRef 2021_04 results
Basket
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.
Help
UniRef help video
Other tutorials and videos
Downloads
Filter by
100% (152,313)
Map to
UniProtKB
UniParc
Demo
Help video
BLAST
Align
Download
Add to basket
Columns
1 to 25 of 152,313
Show 25
Cluster ID
Cluster name
Size
Cluster members
Organisms
Length
Identity
UniRef100_A0A023HHK9
Cluster: Methylcytosine dioxygenase TET
1
A0A023HHK9
Homo sapiens (Human)
1,305
100%
UniRef100_A0A023HHL0
Cluster: Methylcytosine dioxygenase TET
1
A0A023HHL0
Homo sapiens (Human)
694
100%
UniRef100_A0A023HJ61
Cluster: HRES-1/RAB4 variant
1
A0A023HJ61
Homo sapiens (Human)
121
100%
UniRef100_A0A023I7F4
Cluster: Cytochrome b
1
A0A023I7F4
Homo sapiens (Human)
380
100%
UniRef100_A0A023I7H2
Cluster: NADH-ubiquinone oxidoreductase chain 5
1
A0A023I7H2
Homo sapiens (Human)
603
100%
Type here to search
26°C
ENG
11:46 AM
12/2/2022

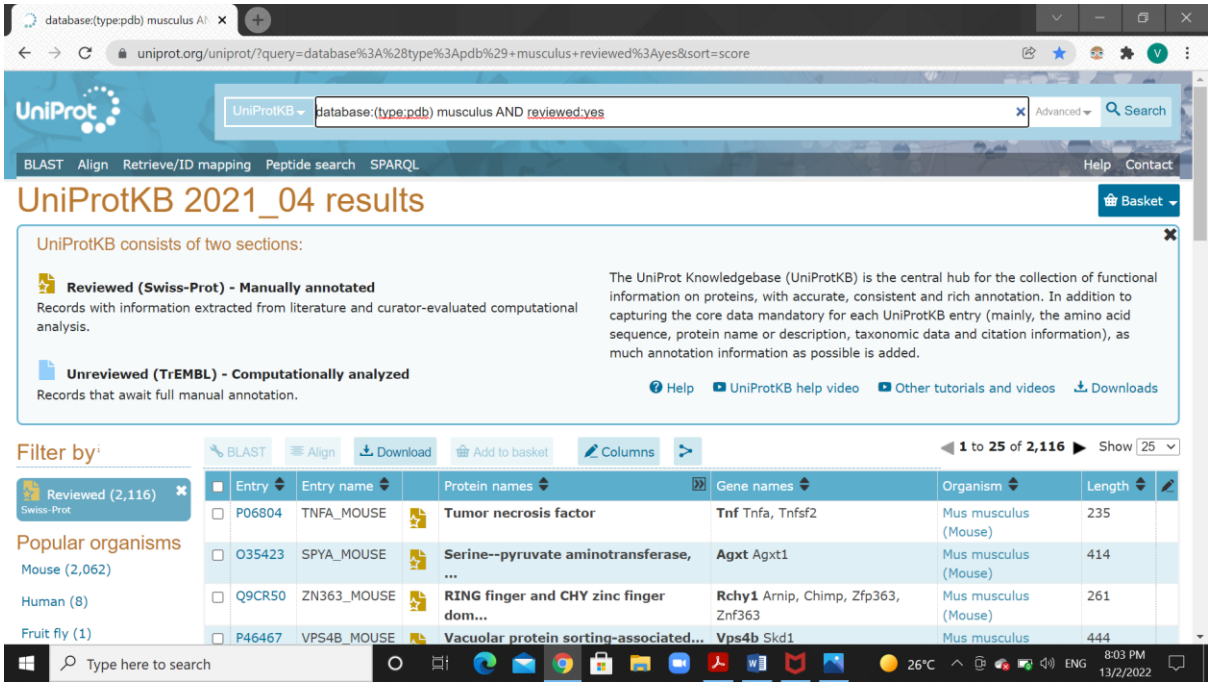

New Tab
uniprot:(organism:"Homo sapien
uniprot.org/uniref/?query=uniprot:(organism%3A"Homo+sapiens+%28Human%29+%5B9606%5D")+identity:0.9
UniProt
UniRef
uniprot:(organism:"Homo sapiens (Human) [9606]") AND identity:0.9
Advanced
Search
BLAST Align Retrieve/ID mapping Peptide search SPARQL
Help Contact
UniRef 2021_04 results
Basket
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.
Help
UniRef help video
Other tutorials and videos
Downloads
Filter by
90% (91,940)
Map to
UniProtKB
UniParc
Demo
Help video
BLAST
Align
Download
Add to basket
Columns
1 to 25 of 91,940
Show 25
Cluster ID
Cluster name
Size
Cluster members
Organisms
Length
Identity
UniRef90_A0A023IP88
Cluster: Integrin a4 subunit (Fragment)
2
A0A023IP88 UPI0003338321
Homo sapiens (Human) Echinops telfairi (Lesser hedgehog tenrec)
55
90%
UniRef90_A0A023IPH8
Cluster: Integrin a4 subunit (Fragment)
3
A0A023IPH8 I3W9R5 I3W9S3
Homo sapiens (Human) Aotus azarae (Southern owl monkey) (Aotus azarai)
66
90%
UniRef90_A0A023IQH3
Cluster: Integrin b7 subunit (Fragment)
1
A0A023IQH3
Homo sapiens (Human)
74
90%
Type here to search
26°C
ENG
11:48 AM
12/2/2022



4) In UniProt, the number of mouse (Mus musculus) protein sequences that are manually annotated is 17,527.

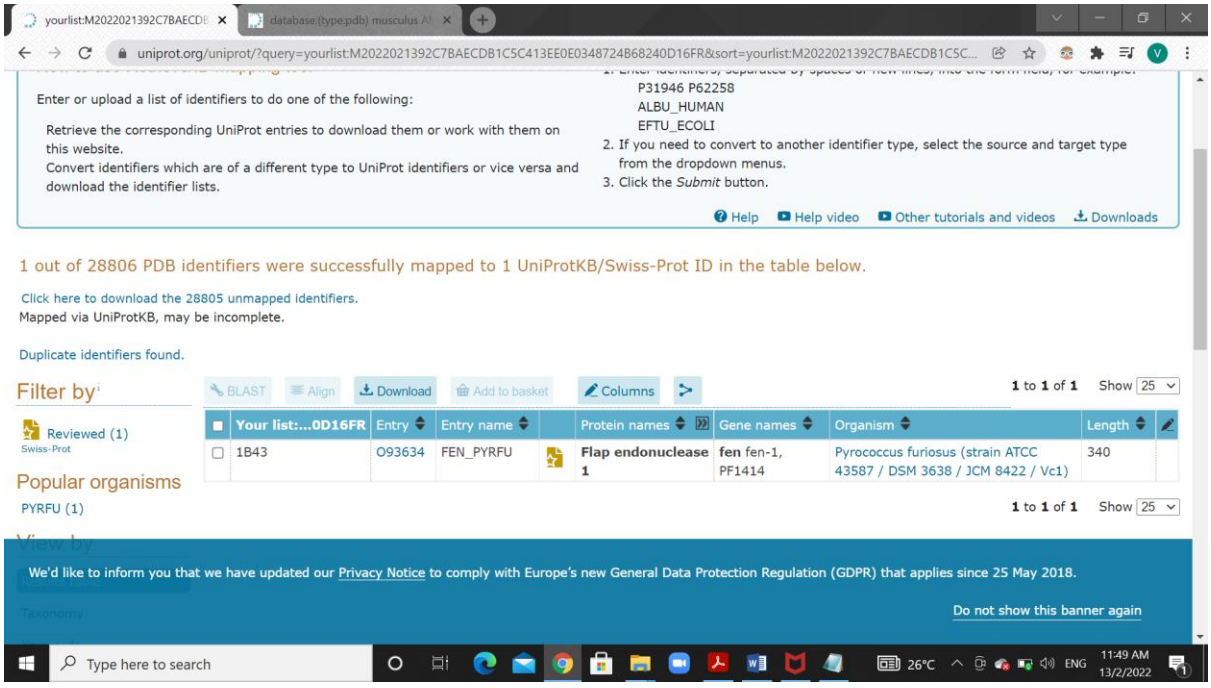


The number of these manually annotated protein sequences that are associated with PDB (3D structures) is 2,116.

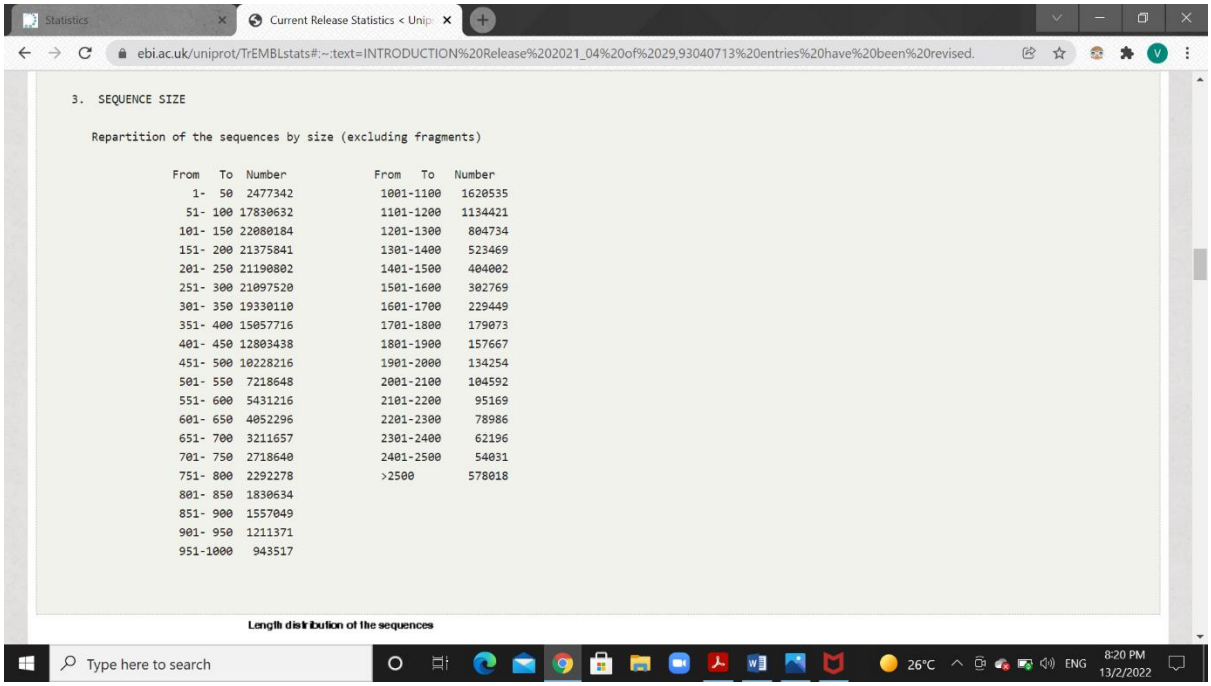


5) Using Retrieve/ID mapping, I mapped the UniProt IDs of the manually curated mouse protein sequences with 3D structures to STRING database.

Only one of them successfully mapped.



6) a) When we look at the distribution of sequence length in UniProt, we could see that most of the sequences are in range 101-150. The average sequence length of the sequences is 348 amino acids.

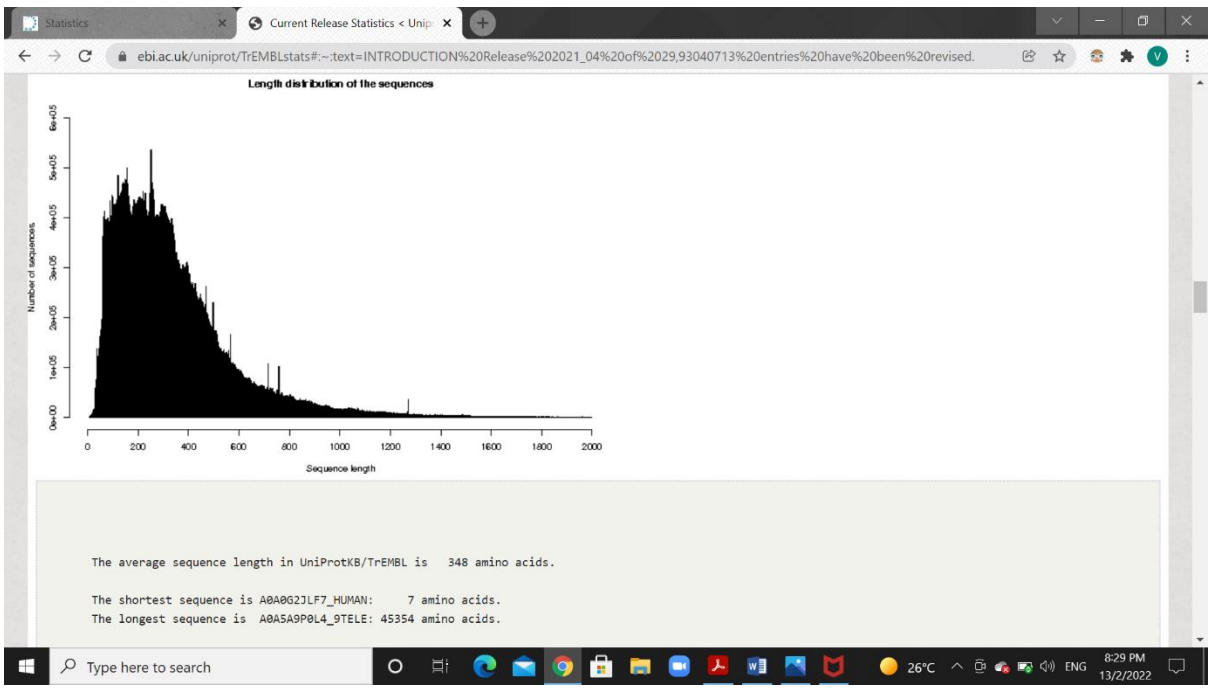


3. SEQUENCE SIZE

Repartition of the sequences by size (excluding fragments)

From	To	Number	From	To	Number
1-	50	2477342	1001-1100		1620535
51-	100	17830632	1101-1200		1134421
101-	150	22080184	1201-1300		804734
151-	200	21375841	1301-1400		523469
201-	250	21190802	1401-1500		404002
251-	300	21097520	1501-1600		302769
301-	350	19330110	1601-1700		229449
351-	400	15057716	1701-1800		179073
401-	450	12803438	1801-1900		157667
451-	500	10228216	1901-2000		134254
501-	550	7218648	2001-2100		104592
551-	600	5431216	2101-2200		95169
601-	650	4052296	2201-2300		78986
651-	700	3211657	2301-2400		62196
701-	750	2718640	2401-2500		54031
751-	800	2292278	>2500		578018
801-	850	1830634			
851-	900	1557049			
901-	950	1211371			
951-	1000	943517			

Length distribution of the sequences



Length distribution of the sequences

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.

b) The shortest sequence is A0A0G2JLF7_HUMAN of length 7 amino acids.

The longest sequence is A0A5A9P0L4_9TELE of length 45354 amino acids.

c) The amino acid composition in percent for the complete database is:

Ala (A) - 9.13 Gln (Q) - 3.77 Leu (L) - 9.88 Ser (S) - 6.71

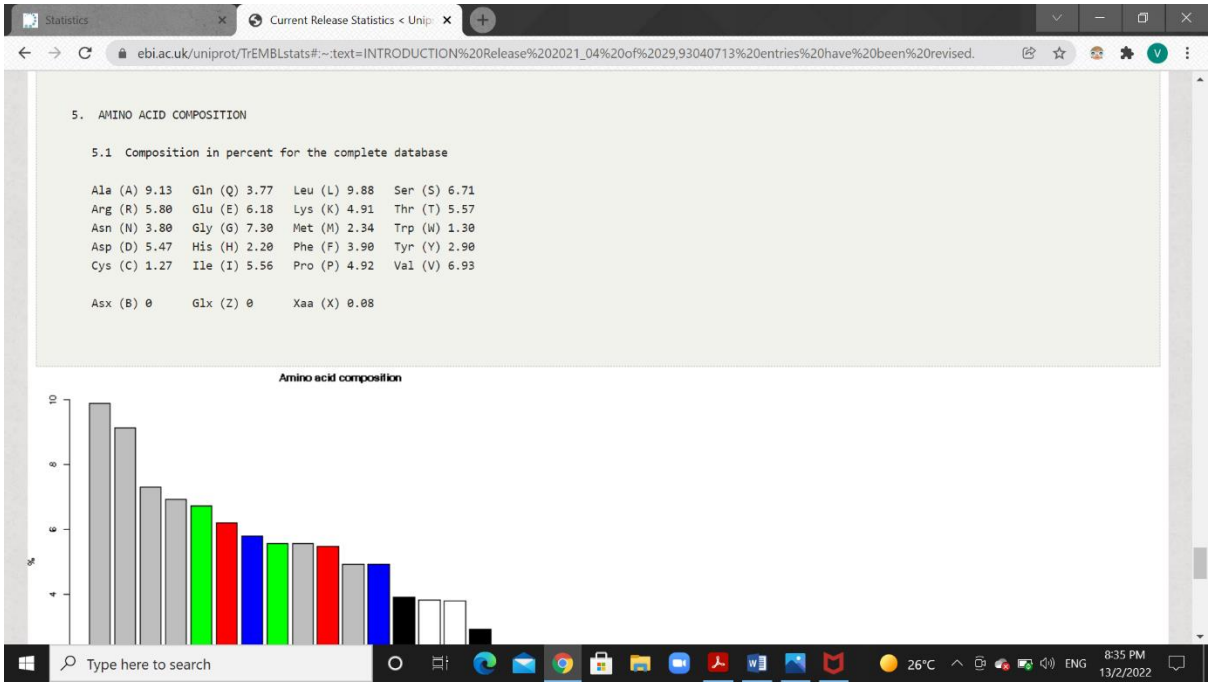
Arg (R) - 5.80 Glu (E) - 6.18 Lys (K) - 4.91 Thr (T) - 5.57

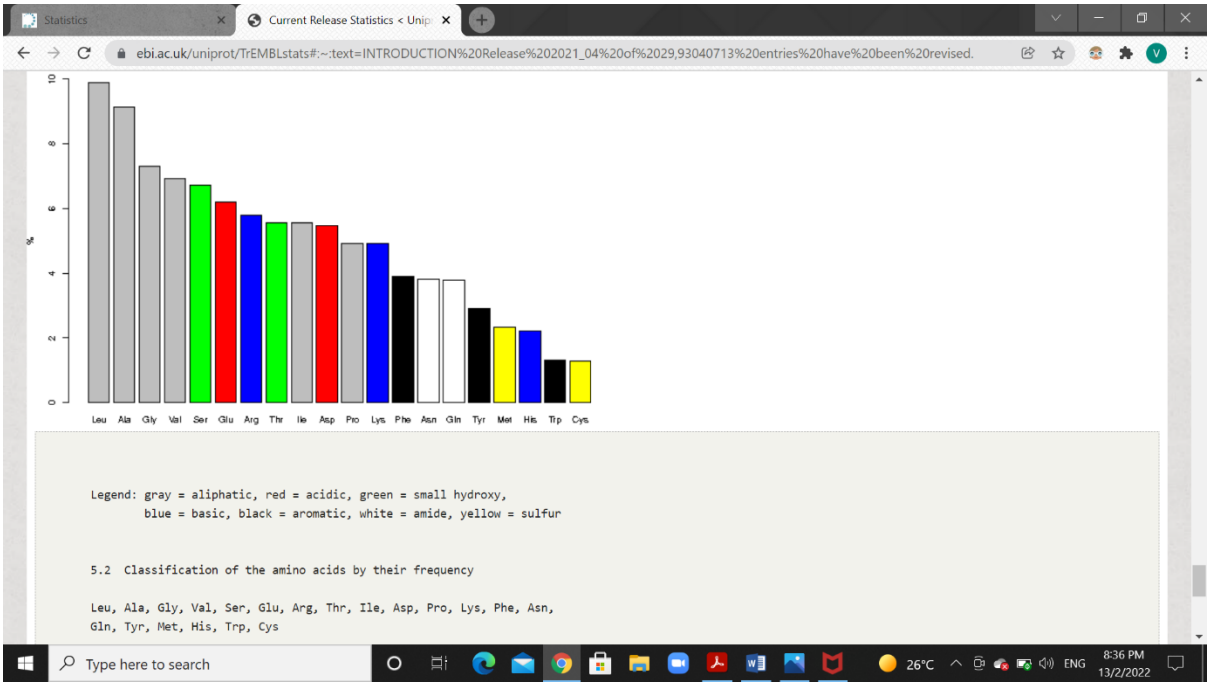
Asn (N) - 3.80 Gly (G) - 7.30 Met (M) - 2.34 Trp (W) - 1.30

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





Legend: gray = aliphatic, red = acidic, green = small hydroxy,
blue = basic, black = aromatic, white = amide, yellow = sulfur

5.2 Classification of the amino acids by their frequency

Leu, Ala, Gly, Val, Ser, Glu, Arg, Thr, Ile, Asp, Pro, Lys, Phe, Asn,
Gln, Tyr, Met, His, Trp, Cys