

Practicals 3

-BS19B032

-R. Vasantha Kumar

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

human mitochondrial beta barrel x https://www.uniprot.org/uniprot/ x

uniprot.org/uniprot/?query=human+mitochondrial+beta+barrel+membrane+protein+VDAC&sort=score

sequence, protein name or description, taxonomic data and citation information), as much annotation information as possible is added.

Unreviewed (TrEMBL) - Computationally analyzed
Records that await full manual annotation.

Filter by: Reviewed (3) Swiss-Prot

Popular organisms: Human (3)

Subcellular location: Membrane raft (1), Membrane (3), Multi-pass membrane protein (1)

View by: Results table, Taxonomy, Keywords, Gene Ontology

Do you mean human mitochondrial beta barrel membrane vdc
Quote terms: "beta barrel", "membrane protein"
Did you mean to search for human mitochondrion beta barrel membrane protein VDAC

Entry	Entry name	Protein names	Gene names	Organism	Length
P21796	VDAC1_HUMAN	Voltage-dependent anion-selective c...	VDAC1 VDAC	Homo sapiens (Human)	283
P45880	VDAC2_HUMAN	Voltage-dependent anion-selective c...	VDAC2	Homo sapiens (Human)	294
Q9Y277	VDAC3_HUMAN	Voltage-dependent anion-selective c...	VDAC3	Homo sapiens (Human)	283

1 to 3 of 3 Show 25

VDAC 1:

MAVPPTYADLGKSARDVFTKGYGFGLIKLDLKTSENGLEFTSSGSANTETTKVTGSLET
KYRWTEYGLTFTEKWNTDNTLGTEITVEDQLARGLKLTDFDSSFSPTGKKNKIKTGYKR
EHINLGCMDMFDIAGPSIRGALVLGYEGWLAGYQMNFFETAKSRVTQSNFAVGKYKTDEFQL
HTNVNDGTEFGGSIYQKVNKKLETAVNLAWTAGNSNTRFGIAAKYQIDPDACFSKVNNS
SLIGLGYTQTLKPGIKLTLTSLALLDGKNVNAGGHKLGLEFQA

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

VDAC 2:

MATHGQTCARPMCIIPSYADLGKAARDIFNKGFGFGLVKLDVKTSCSGVEFSTSGSSNT
DTGKVTGTLETKYKWCEYGLTFTEKWNTDNTLGTEIAIEDQICQGLKLTDFDTTFSPNTGK
KSGKIKSSYKRECINLGCDDVDFDFAGPAIHGSAVFGYEGWLAGYQMTFDSKSKLTRNNE
AVGYRTGDFQLHTNVNDGTEFGGSIYQKVEDLDTSVNLAWTSGTNCTRFGIAAKYQLDP
TASISAKVNNSSLIGVGYTQTLRPGVKLTLTSLALVDGKSINAGGHKVGLALELEA

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

VDAC 3:

MCNTPTYCDLGKAAKDVFNKGYGFGMVKIDLTKSCSGVEFSTSGHAYTDTGKASGNLET
 KYKVCNYGLTFTQKWNNTDNTLGTEISWENKLAEGGLKLTLDITFVFNTPGKKSGLKASYKR
 DCFVSGSNVDIDFSGPTIYGWAVLAFEGWLAGYQMSFDTAKSKLSQNNFALGYKAAADFQL
 HTHVNDGTEFGGSIYQKVNEKIETSINLAWTAGSNNTFRGIAAKYMLDCRTSLSAKVNNA
 SLIGLGYTQTLRPGVKLTLSALIDGKNFSAGGHKVGGLGFELEA

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.*

transcription factors AND identity: x

uniprot.org/uniref/?query=transcription%20factors&fil=identity:0.5&sort=score

UniProt

transcription factors AND identity:0.5

Advanced Search

BLAST Align Retrieve/ID mapping Peptide search SPARQL Help Contact

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.

Help UniRef help video Other tutorials and videos Downloads

Filter by: 50% (344)

Map to: UniProtKB UniParc Demo Help video

Quote terms: "transcription factors"

Cluster ID	Cluster name	Size	Cluster members	Organisms	Length	Identity
UniRef50_P0A4H3	Cluster: Virulence factors putative positive transcription regulator BvgA	1,266	P0A4H3 P0A4H4 P0A4H2 A0A0C6P3K4 A0A291KZA7 A0A5Q1YDD6 A0A077CP47	Bordetella bronchiseptica (strain ATCC BAA-588 / NCTC 13252 / RB50) (Alcaligenes bronchisepticus) Bordetella parapertussis (strain 12822 / ATCC BAA-587 / NCTC 13253) Bordetella pertussis (strain Tohama I / ATCC BAA-589 / NCTC 13251)	209	50%

1 to 25 of 344 Show 25

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

cluster:(uniprot:(transcription factors)+identity:0.5)

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 (40,747) ☐ Unreviewed (7,162,110) ☐ Popular organisms

Human (6,462)

Entry	Entry name	Protein names	Gene names	Organism	Length
<input type="checkbox"/> Q48LC2	Y1551_PSE14	UPF0260 protein PSPPH_1551	PSPPH_1551	Pseudomonas savastanoi pv. phaseolicola (strain 1448A / Race 6) (Pseudomonas syringae pv. phaseolicola (strain 1448A / Race 6))	149
<input type="checkbox"/> B5Z9W0	Y173_HELPG	UPF0114 protein HPG27_173	HPG27_173	Helicobacter pylori (strain G27)	177
<input type="checkbox"/> A6VEA3	XPT_PSE7	Xanthine	xpt PSA7_6072	Pseudomonas aeruginosa (strain PA7)	190

1 to 25 of 7,202,857 Show 25

uniref-uniprot_(transcription+factors)+identity_0.5.fasta - Notepad

```

>UniRef50_A0A010QZT8 Uncharacterized protein n=53 Tax=Proteobacteria TaxID=1224 RepID=A0A010QZT8_9ALTE
MSNKATFEQIKLALAEAPNQYTAELHLQMIKFADELKHITAKEFCEEMGLKQSLGTEFS
KMRNLTKRLSAGLNTDLI
>UniRef50_A0A010REZ0 Uncharacterized protein n=963 Tax=Proteobacteria TaxID=1224 RepID=A0A010REZ0_PSEFL
MDNPFQLITDAFADPYQVNLISQIGLDGSLMILSNAGRIYAKRMISAEQRNDPQRLKRLV
QSIQGIATIEQHSVAIVALEMTNGDNKLPPSPKAPPPRPARF
>UniRef50_A0A010RUL4 Mediator of RNA polymerase II transcription subunit 11 n=49 Tax=Hypocreomycetidae TaxID=222543 RepID=A0A010RUL4_9PEZI
MTTDAPNPEFTLQERIQQLCEIDTQIVSLMSHTSSALGAPPKQSPSPSNPSAPPS
QPSSSSQTFKTSMDALLSTLHTVDVHMKQILALEASIIKLRNDADPKTRQPMVNE
DAKIVARPSLEPNVGITIGNLVGWLNSRNKVERDMEAEALWAKMRELLERYHAKEVAEG
AGGGGGGGGGGKIDERMQD
>UniRef50_A0A010S321 Mediator of RNA polymerase II transcription subunit 18 n=119 Tax=Hypocreomycetidae TaxID=222543 RepID=A0A010S321_9PEZI
MYELFTTLVDDDDIQAACSVLGLCAMPANQSLHRLVYFKGPGKPGISNQNSIVKTPR
KDIQMLWKLHQQLSRQSYILQARYEVFKDKDFGPTAPEVDNARAGTLRWDFDPPQV
RSSVTRQKKTETDQRLNLSVMKNNYQFSEATEETQYQFREDVEFCLSRHYILQWMD
NGPLTQTATMDTHSPADPGQRWMLIKVHVQDNKPOEILKAHEQLANIRELEGVFEFR
MFDRIHDTAVEMRNAPAPLPQVMTITDQRLNSARSLGESFEDSAVNGKVQSQTVGA
PDFADQVRAEPPY
>UniRef50_A0A010S7M9 DUF190 domain-containing protein n=292 RepID=A0A010S7M9_PSEFL
MQGFLVIFTQNRHGHGKMLGDWLDLAKEMGLHATLATGIEFGHSGKLSAHFFEM
ADQPAEIRLAIETEECTLFARLEADIALFYIKAPVEFTVGKRAEPFGHQV
>UniRef50_A0A010SV10 Uncharacterized protein n=1453 RepID=A0A010SV10_PSEFL
MLRLIVPTAALLLALPFSQAASLNDNLGKMLEKVAESNVGLPREINENLDQGYTVD
GKQLIDHISVQSAYADEMRANPKVVYLQLGASVCRNVYRKLMAKGAIMRYDFSENKTR
PVGSAQSQSDCPAQTAKKK
>UniRef50_A0A011PCM1 RNA polymerase sigma factor RpoD n=80 RepID=A0A011PCM1_9PROT
MAREKAATTAKAKAAADLLAQVGNQAPIDDETRKTRKTLIKLGERGFLTYAEIND
HLPDDVVDAEQIESIISTFNDMGIVQYDEAPDAETLHMSDAPAGADDADVEEQAEAL
STVDSEFGRTTDPVRMYRMHSGVELLTREGEIETAKRIEDGLKHMVQAISACPTTIAEI
LSCADRIAREEMRIEELIDGLIYPEGESAPDGPADDDDDGDDDDGDDDDDDDDNEA
AEGAAAAASLLKRLTEGLARLEIREHDSKAQALLARKGSQKAYLKLQEQISDEMNGIR
FTSKTIERLCDTVRAMVEEARTCEKRIQICVDVTRMPRPHIKVFPNGNELNDWVDVEI
AAAGKYAAITLRNAPNITEQRKLLALQDRIQIPKELKDINKQMSGEAKARRAKREM
TEANLRLVSIKAKYTNRLQFLDLIQENIGLMKAVDKFEYRRGYKFSTYATWIRQAI
  
```

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

uniportal(organism:"Homo sapiens")

uniportal.org/uniportal/?query=uniportal(organism%3A"Homo sapiens (Human) [9606]") AND identity:1.0

UniRef

uniportal(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

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

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_A0A02317F4	Cluster: Cytochrome b	1	A0A02317F4	Homo sapiens (Human)	380	100%
UniRef100_A0A02317H2	Cluster: NADH-ubiquinone oxidoreductase chain 5	1	A0A02317H2	Homo sapiens (Human)	603	100%

1 to 25 of 152,313 Show 25

uniportal(organism:"Homo sapiens")

uniportal.org/uniportal/?query=uniportal(organism%3A"Homo sapiens (Human) [9606]") AND identity:0.9

UniRef

uniportal(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

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

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 azarae)	66	90%
UniRef90_A0A023IQH3	Cluster: Integrin b7 subunit (Fragment)	1	A0A023IQH3	Homo sapiens (Human)	74	90%

1 to 25 of 91,940 Show 25

UniProt

UniRef

UniRef50_A0A015LMS4

Cluster: CCT-alpha

Size: 81

Cluster members: A0A015LMS4, A0A2K6MX77, A0A6G3MFF8, A0A7R9YPJ5, A0A673M8N3, A0A448WS09, A0A752UGW3, A0A3B5ZUR3, A0A2K6R5C9, +71

Organisms: Rhizophagus irregularis (strain DAOM 197198w) (Glomus intraradices), Rhinopithecus bieti (Black snub-nosed monkey) (Pygathrix bieti), Henneguya salminicola, Diacronema lutheri (Unicellular marine alga) (Monochrysis lutheri), Sinocyclocheilus rhinoceros, Protopolystoma xenopodis, Attheya septentrionalis

Length: 436

Identity: 50%

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

UniProt

UniProtKB

mus musculus AND reviewed:yes

UniProtKB 2021_04 results

Reviewed (Swiss-Prot) - Manually annotated

Unreviewed (TrEMBL) - Computationally analyzed

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
Q91Z30	MUS81_MOUSE	Crossover junction endonuclease MUS...	Mus81	Mus musculus (Mouse)	551

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

database:(type:pdb) musculus AND reviewed:yes

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 (2,116)

Popular organisms: Mouse (2,062), Human (8), Fruit fly (1)

Entry	Entry name	Protein names	Gene names	Organism	Length
P06804	TNFA_MOUSE	Tumor necrosis factor	Tnf Tnfa, Tnfsf2	Mus musculus (Mouse)	235
O35423	SPYA_MOUSE	Serine--pyruvate aminotransferase, ...	Agxt Agxt1	Mus musculus (Mouse)	414
Q9CR50	ZN363_MOUSE	RING finger and CHY zinc finger dom...	Rchy1 Arrip, Chimp, Zfp363, Znf363	Mus musculus (Mouse)	261
P46467	VPS4B_MOUSE	Vacuolar protein sorting-associated...	Vps4b Skd1	Mus musculus	444

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.

yourlist:M2022021392C7BAECD81C5C413EE0E0348724B68240D16FR&sort=yourlist:M2022021392C7BAECD81C5C413EE0E0348724B68240D16FR

Enter or upload a list of identifiers to do one of the following:

- 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.

1 out of 28806 PDB identifiers were successfully mapped to 1 UniProtKB/Swiss-Prot ID in the table below.

Click here to download the 28805 unmapped identifiers. Mapped via UniProtKB, may be incomplete.

Duplicate entries found.

Filter by: Reviewed (1)

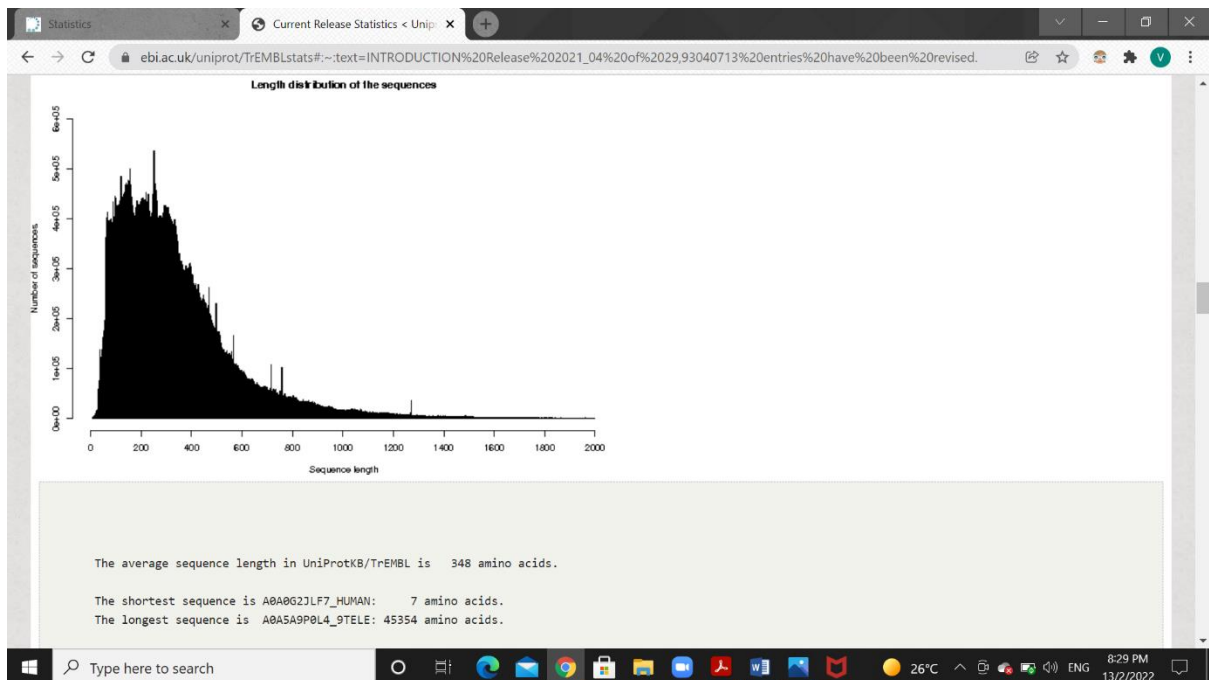
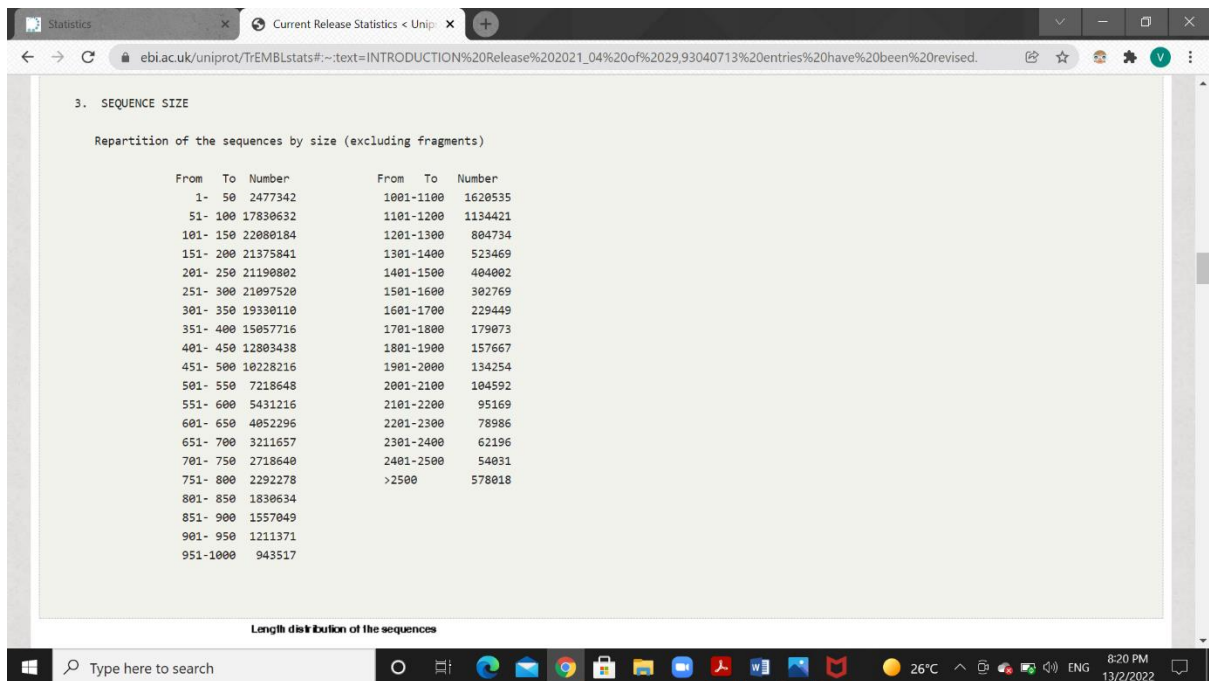
Popular organisms: PYRFU (1)

Your list...	Entry	Entry name	Protein names	Gene names	Organism	Length
1B43	O93634	FEN_PYRFU	Flap endonuclease 1	fen fen-1, PF1414	Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1)	340

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

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.



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

