

Question 1

In [16]:

```
import string
import pandas as pd
import numpy as np

aas = set(string.ascii_uppercase) - set("BJOZXU")
a = "AMENLNMDLLYMAAAVMMGLAAIGAIGILGGKFLEGAARQPDLIPLLRTOFFIVMGLVDAIPMIAVG LGLYVMFAVA".replace(" ", '')
b = "AADVSAAVGATGQSGMTYRLGLSWDWDKSWWQTSTGRLTGYWDAGYTYWEGGDEGAGKHSLSFAP VFVYEFAGDSIKPFIEAGIGVAAFSGTRVGDQNLGSSSLNFEDR"
c = "MALLPAAPGAPARATPTRWPVGCNRPWTKWSYDEALDGIKAAGYAWTGLLTASKPSLHHATATPEY LAALKQKSRHAA".replace(" ", '')

def comp(a):
    return {x : a.count(x)/len(a) for x in set(aas)}

def hamm(a,b):
    d1 = comp(a)
    d2 = comp(b)
    return sum([abs(d1[x]-d2[x]) for x in aas])

def euclidean(a,b):
    d1 = comp(a)
    d2 = comp(b)
    return sum([(d1[x] - d2[x])**2 for x in aas])**0.5


eu = [euclidean(x,y) for x,y in [(a,b),(b,c),(c,a)]]
ha = [hamm(x,y) for x,y in [(a,b),(b,c),(c,a)]]
df = pd.DataFrame([eu,ha], index = ['Euclidean', 'Hamming'], columns = ['1,2','2,3','1,3'])
df
```

Out[16]:

	1,2	2,3	1,3
Euclidean	0.201062	0.201130	0.220868
Hamming	0.665728	0.726633	0.843354

Question 2

703 Reviewed sequences from uniprot were taken




Advanced Search


[BLAST](#)
[Align](#)
[Retrieve/ID mapping](#)
[Peptide search](#)
[SPARQL](#)
Help
Contact

UniProtKB 2022_01 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.

Help
UniProtKB help video
Other tutorials and videos
Downloads

Filter by

Reviewed (703) Swiss-Prot

Popular organisms




Human (36)
S. cerevisiae (36)
E. coli K12 (21)
A. thaliana (19)

1 to 25 of 703
Show 25

BLAST
Align
Download
Add to basket
Columns

Do you mean beta barrel membrane reviewed:yes

Quote terms: "beta barrel"

Entry	Entry name	Protein names	Gene names	Organism	Length
<input type="checkbox"/> O18423	TXL_EISFE	 Lysenin		Eisenia fetida (Red wiggler worm)	297
<input type="checkbox"/> P66948	BEPA_ECOLI	 Beta-barrel assembly-enhancing prot...	bepA yfG, b2494, JW2479	Escherichia coli (strain K12)	487
<input type="checkbox"/> P0A910	OMPA_ECOLI	 Outer membrane protein A	ompA con. tolG, tut. b0957.	Escherichia coli (strain K12)	346

In [21]:

```
pd.DataFrame(np.transpose([[245,304,430,504,639]]),index = [0.4,0.5,0.75,0.9,1], columns = ["Number of Clusters"])
```

Out[21]:

	Number of Clusters
0.40	245
0.50	304
0.75	430
0.90	504
1.00	639

Header of each file in CD-HIT Cluster is attached

Raw output	>Cluster 0	0	810aa, >sp P0A940 BAMA_ECOLI...	at 78.77%
Download all files		1	803aa, >sp B5F8T8 BAMA_SALA4...	at 79.83%
Browse clusters by size		2	810aa, >sp A9MPI4 BAMA_SALAR...	at 79.01%
Browse clusters by length		3	810aa, >sp B520F6 BAMA_ECO5E...	at 78.77%
Distribution of clusters		4	810aa, >sp B7MP37 BAMA_ECO81...	at 78.64%
		5	797aa, >sp P46024 BAMA1_HAEIF...	at 45.17%
		6	810aa, >sp B7MBF8 BAMA_ECO45...	at 78.77%
		7	810aa, >sp P0A942 BAMA_ECO57...	at 78.77%
		8	809aa, >sp B5Y1J4 BAMA_KLEP3...	at 78.86%
		9	809aa, >sp A6T4X9 BAMA_KLEP7...	at 79.23%
		10	814aa, >sp Q6D8D5 BAMA_PECAS...	*
		11	805aa, >sp Q57T31 BAMA_SALCH...	at 78.88%
		12	810aa, >sp B7LGN8 BAMA_ECO55...	at 78.77%
		13	810aa, >sp B7M1Y0 BAMA_ECO8A...	at 78.77%
		14	809aa, >sp C6DAJ1 BAMA_PECPP...	at 96.42%
		15	793aa, >sp O32629 BAMA2_HAEIF...	at 41.99%
		16	797aa, >sp Q7N8N9 BAMA_PHOLL...	at 76.16%
		17	797aa, >sp Q9S341 BAMA_PHOLU...	at 75.91%
		18	810aa, >sp A7ZHR7 BAMA_ECO24...	at 78.77%
		19	810aa, >sp C4ZRR9 BAMA_ECOBW...	at 78.77%
		20	810aa, >sp B1XD46 BAMA_ECODH...	at 78.77%
		21	810aa, >sp B1LGX9 BAMA_ECOSM...	at 78.77%
		22	810aa, >sp Q1RG12 BAMA_ECOUT...	at 78.77%
		23	802aa, >sp B7LW74 BAMA_ESCF3...	at 79.55%
		24	805aa, >sp C0Q6K0 BAMA_SALPC...	at 79.01%
		25	803aa, >sp Q8Z9A3 BAMA_SALTI...	at 79.70%
		26	810aa, >sp Q32JT2 BAMA_SHIDS...	at 78.77%
		27	810aa, >sp Q0T832 BAMA_SHIF8...	at 78.77%
		28	810aa, >sp P0A943 BAMA_SHIFL...	at 78.77%
		29	810aa, >sp Q3Z5I1 BAMA_SHISS...	at 78.77%
		30	808aa, >sp Q2NRL5 BAMA_SODGM...	at 81.56%
		31	795aa, >sp A7FFH7 BAMA_YERP3...	at 81.76%
		32	795aa, >sp Q1CAM6 BAMA_YERPA...	at 81.76%
		33	795aa, >sp B2JZ26 BAMA_YERPB...	at 81.76%
		34	795aa, >sp Q8ZH58 BAMA_YERPE...	at 81.76%
		35	795aa, >sp Q667J7 BAMA_YERPS...	at 81.76%
		36	810aa, >sp B7UIM2 BAMA_ECO27...	at 78.77%
		37	795aa, >sp C5B7R5 BAMA_EDWI9...	at 78.49%
		38	805aa, >sp A4W6S2 BAMA_ENT38...	at 80.12%
		39	803aa, >sp B4SV06 BAMA_SALNS...	at 79.38%
		40	803aa, >sp Q5PD65 BAMA_SALPA...	at 79.70%

Raw output	>Cluster 0	0	810aa, >sp P0A940 BAMA_ECOLI...	at 78.77%
Download all files		1	803aa, >sp B5F8T8 BAMA_SALA4...	at 79.83%
Browse clusters by size		2	810aa, >sp A9MPI4 BAMA_SALAR...	at 79.01%
Browse clusters by length		3	810aa, >sp B520F6 BAMA_ECO5E...	at 78.77%
Distribution of clusters		4	810aa, >sp B7MP37 BAMA_ECO81...	at 78.64%
		5	810aa, >sp B7MBF8 BAMA_ECO45...	at 78.77%
		6	810aa, >sp P0A942 BAMA_ECO57...	at 78.77%
		7	809aa, >sp B5Y1J4 BAMA_KLEP3...	at 78.86%
		8	809aa, >sp A6T4X9 BAMA_KLEP7...	at 79.23%
		9	814aa, >sp Q6D8D5 BAMA_PECAS...	*
		10	805aa, >sp Q57T31 BAMA_SALCH...	at 78.88%
		11	810aa, >sp B7LGN8 BAMA_ECO55...	at 78.77%
		12	810aa, >sp B7M1Y0 BAMA_ECO8A...	at 78.77%
		13	809aa, >sp C6DAJ1 BAMA_PECPP...	at 96.42%
		14	797aa, >sp Q7N8N9 BAMA_PHOLL...	at 76.16%
		15	797aa, >sp Q9S341 BAMA_PHOLU...	at 75.91%
		16	810aa, >sp A7ZHR7 BAMA_ECO24...	at 78.77%
		17	810aa, >sp C4ZRR9 BAMA_ECOBW...	at 78.77%
		18	810aa, >sp B1XD46 BAMA_ECODH...	at 78.77%
		19	810aa, >sp B1LGX9 BAMA_ECOSM...	at 78.77%
		20	810aa, >sp Q1RG12 BAMA_ECOUT...	at 78.77%
		21	802aa, >sp B7LW74 BAMA_ESCF3...	at 79.55%
		22	805aa, >sp C0Q6K0 BAMA_SALPC...	at 79.01%
		23	803aa, >sp Q8Z9A3 BAMA_SALTI...	at 79.70%
		24	810aa, >sp Q32JT2 BAMA_SHIDS...	at 78.77%
		25	810aa, >sp Q0T832 BAMA_SHIF8...	at 78.77%
		26	810aa, >sp P0A943 BAMA_SHIFL...	at 78.77%
		27	810aa, >sp Q3Z5I1 BAMA_SHISS...	at 78.77%
		28	808aa, >sp Q2NRL5 BAMA_SODGM...	at 81.56%
		29	795aa, >sp A7FFH7 BAMA_YERP3...	at 81.76%
		30	795aa, >sp Q1CAM6 BAMA_YERPA...	at 81.76%
		31	795aa, >sp B2JZ26 BAMA_YERPB...	at 81.76%
		32	795aa, >sp Q8ZH58 BAMA_YERPE...	at 81.76%
		33	795aa, >sp Q667J7 BAMA_YERPS...	at 81.76%
		34	810aa, >sp B7UIM2 BAMA_ECO27...	at 78.77%
		35	795aa, >sp C5B7R5 BAMA_EDWI9...	at 78.49%
		36	805aa, >sp A4W6S2 BAMA_ENT38...	at 80.12%
		37	810aa, >sp B4SV06 BAMA_SALNS...	at 79.38%
		38	803aa, >sp Q5PD65 BAMA_SALPA...	at 79.70%
		39	803aa, >sp B4SV06 BAMA_SALNS...	at 79.38%

Raw output	>Cluster 0	0	810aa, >sp P0A940 BAMA_ECOLI...	at 78.77%
Download all files		1	803aa, >sp B5F8T8 BAMA_SALA4...	at 79.83%
Browse clusters by size		2	810aa, >sp A9MPI4 BAMA_SALAR...	at 79.01%
Browse clusters by length		3	810aa, >sp B520F6 BAMA_ECO5E...	at 78.77%
Distribution of clusters		4	810aa, >sp B7MP37 BAMA_ECO81...	at 78.64%
		5	810aa, >sp B7MBF8 BAMA_ECO45...	at 78.77%
		6	810aa, >sp P0A942 BAMA_ECO57...	at 78.77%
		7	809aa, >sp B5Y1J4 BAMA_KLEP3...	at 78.86%
		8	809aa, >sp A6T4X9 BAMA_KLEP7...	at 79.23%
		9	814aa, >sp Q6D8D5 BAMA_PECAS...	*
		10	805aa, >sp Q57T31 BAMA_SALCH...	at 78.88%
		11	810aa, >sp B7LGN8 BAMA_ECO55...	at 78.77%
		12	810aa, >sp B7M1Y0 BAMA_ECO8A...	at 78.77%
		13	809aa, >sp C6DAJ1 BAMA_PECPP...	at 96.42%
		14	797aa, >sp Q7N8N9 BAMA_PHOLL...	at 76.16%
		15	797aa, >sp Q9S341 BAMA_PHOLU...	at 75.91%
		16	810aa, >sp A7ZHR7 BAMA_ECO24...	at 78.77%
		17	810aa, >sp C4ZRR9 BAMA_ECOBW...	at 78.77%
		18	810aa, >sp B1XD46 BAMA_ECODH...	at 78.77%
		19	810aa, >sp B1LGX9 BAMA_ECOSM...	at 78.77%
		20	810aa, >sp Q1RG12 BAMA_ECOUT...	at 78.77%
		21	802aa, >sp B7LW74 BAMA_ESCF3...	at 79.55%
		22	805aa, >sp C0Q6K0 BAMA_SALPC...	at 79.01%
		23	803aa, >sp Q8Z9A3 BAMA_SALTI...	at 79.70%
		24	810aa, >sp Q32JT2 BAMA_SHIDS...	at 78.77%
		25	810aa, >sp Q0T832 BAMA_SHIF8...	at 78.77%
		26	810aa, >sp P0A943 BAMA_SHIFL...	at 78.77%
		27	810aa, >sp Q3Z5I1 BAMA_SHISS...	at 78.77%
		28	808aa, >sp Q2NRL5 BAMA_SODGM...	at 81.56%
		29	795aa, >sp A7FFH7 BAMA_YERP3...	at 81.76%
		30	795aa, >sp Q1CAM6 BAMA_YERPA...	at 81.76%
		31	795aa, >sp B2JZ26 BAMA_YERPB...	at 81.76%
		32	795aa, >sp Q8ZH58 BAMA_YERPE...	at 81.76%
		33	795aa, >sp Q667J7 BAMA_YERPS...	at 81.76%
		34	810aa, >sp B7UIM2 BAMA_ECO27...	at 78.77%
		35	795aa, >sp C5B7R5 BAMA_EDWI9...	at 78.49%
		36	805aa, >sp A4W6S2 BAMA_ENT38...	at 80.12%
		37	810aa, >sp B4SV06 BAMA_SALNS...	at 79.38%
		38	803aa, >sp Q5PD65 BAMA_SALPA...	at 79.70%
		39	803aa, >sp B4SV06 BAMA_SALNS...	at 79.38%

Raw output	>Cluster 0	0	810aa, >sp P0A940 BAMA_ECOLI...	*
Download all files		1	803aa, >sp B5F8T8 BAMA_SALA4...	at 93.52%
Browse clusters by size		2	810aa, >sp A9MPI4 BAMA_SALAR...	at 95.93%
Browse clusters by length		3	810aa, >sp B520F6 BAMA_ECO5E...	at 100.00%
Distribution of clusters		4	810aa, >sp B7MP37 BAMA_ECO81...	at 99.88%
		5	810aa, >sp B7MBF8 BAMA_ECO45...	at 100.00%
		6	810aa, >sp P0A942 BAMA_ECO57...	at 100.00%
		7	809aa, >sp B5Y1J4 BAMA_KLEP3...	at 91.47%
		8	809aa, >sp A6T4X9 BAMA_KLEP7...	at 91.59%
		9	805aa, >sp Q57T31 BAMA_SALCH...	at 93.54%
		10	810aa, >sp B7LGN8 BAMA_ECO55...	at 100.00%
		11	810aa, >sp B7M1Y0 BAMA_ECO8A...	at 100.00%
		12	810aa, >sp A7ZHR7 BAMA_ECO24...	at 100.00%
		13	810aa, >sp C4ZRR9 BAMA_ECOBW...	at 100.00%
		14	810aa, >sp B1XD46 BAMA_ECODH...	at 100.00%
		15	810aa, >sp B1LGX9 BAMA_ECOSM...	at 98.89%
		16	810aa, >sp Q1RG12 BAMA_ECOUT...	at 100.00%
		17	802aa, >sp B7LW74 BAMA_ESCF3...	at 93.77%
		18	805aa, >sp C0Q6K0 BAMA_SALPC...	at 93.66%
		19	803aa, >sp Q8Z9A3 BAMA_SALTI...	at 93.90%
		20	810aa, >sp Q32JT2 BAMA_SHIDS...	at 99.75%
		21	810aa, >sp Q0T832 BAMA_SHIF8...	at 100.00%
		22	810aa, >sp P0A943 BAMA_SHIFL...	at 100.00%
		23	810aa, >sp Q3Z5I1 BAMA_SHISS...	at 100.00%
		24	810aa, >sp B7UIM2 BAMA_ECO27...	at 100.00%
		25	810aa, >sp B4SV06 BAMA_SALNS...	at 95.31%
		26	803aa, >sp Q5PD65 BAMA_SALPA...	at 93.77%
		27	810aa, >sp B2U320 BAMA_SHIB3...	at 100.00%
		28	810aa, >sp Q3Z5W3 BAMA_SHIBS...	at 100.00%
		29	810aa, >sp A7ZWC3 BAMA_ECOHS...	at 100.00%
		30	810aa, >sp A1A7M1 BAMA_ECOK1...	at 99.88%
		31	810aa, >sp Q0TLF6 BAMA_ECOL5...	at 100.00%
		32	810aa, >sp P0A941 BAMA_ECOL6...	at 100.00%
		33	810aa, >sp B1IQG4 BAMA_ECOLC...	at 100.00%
		34	810aa, >sp B7N844 BAMA_ECOLU...	at 100.00%
		35	810aa, >sp B6HZF1 BAMA_ECOSE...	at 100.00%
		36	804aa, >sp B5FUJ24 BAMA_SALDC...	at 93.78%
		37	804aa, >sp B5R3J0 BAMA_SALEP...	at 93.78%
		38	810aa, >sp B5RHG2 BAMA_SALG2...	at 95.31%
		39	803aa, >sp B4SV06 BAMA_SALNS...	at 95.31%

Question 3 and 4

Since Pisces server is down, these questions could not be done. Instead the comaparision is done between uniport and CD-HIT at cutoffs in question 5

Question 5

```
In [25]: pd.DataFrame(np.transpose([[245,304,430,504,639],[ '-',365,'-',542,640]]),index = [0.4,0.5,0.75,0.9,1], columns = [
```

Out[25]:

	Number of Clusters in CD-HIT	Number of Clusters in Uniprot
0.40	245	-
0.50	304	365
0.75	430	-
0.90	504	542
1.00	639	640