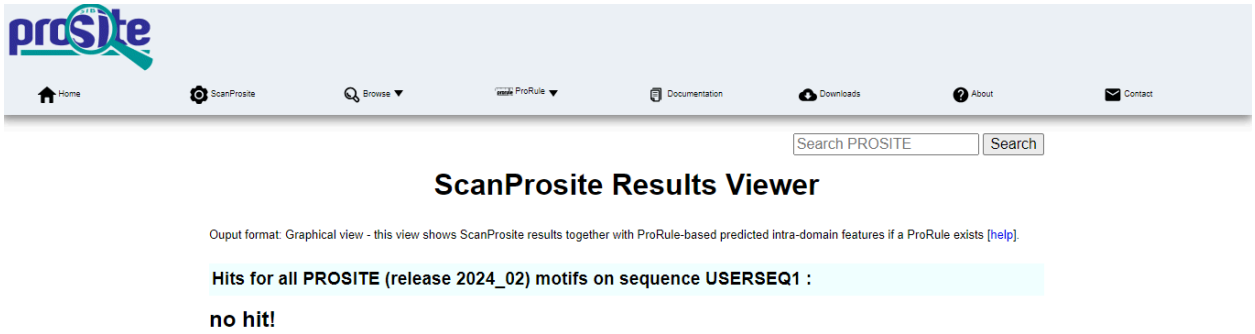
BSE322 Practical Assignment 2

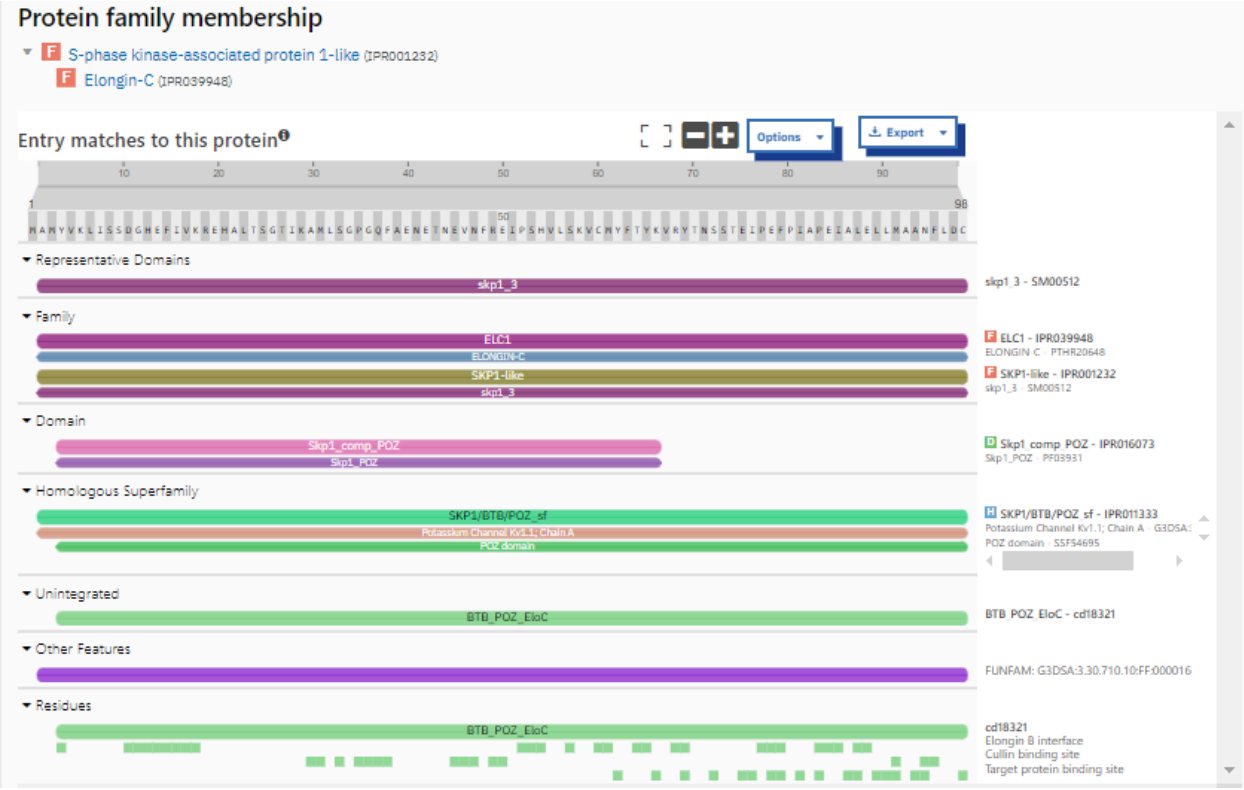
Name: Pradeep Kumar
Roll Number: 220777

1. Using your protein sequence as the query, explore two protein family/domain databases listed below. Summarize what you get in each case, and compare the results from the two databases.
(a) PROSITE: http://prosite.expasy.org/
(b) InterPro: http://www.ebi.ac.uk/interpro/
Interpret the result: what does the presence or absence of matches in these databases represent?
Answer 1.
(a)PROSITE
PROSITE serves as a valuable tool for researchers investigating proteins, offering an extensive repository of information regarding protein families, functional domains, and active sites. Through its compilation of characteristic amino acid patterns, it aids scholars in comprehending protein functionality and classification.



Upon querying my sequence on the PROSITE database, the outcome indicated no hit, as given in the provided image. This implies that the PROSITE protein database may have limitations, leading to the absence of a corresponding protein match for my sequence.
Source:
https://prosite.expasy.org/cgi-bin/prosite/scanprosite/ScanView.cgi?scanfile=701289525067.scan.gz
(b) InterPro: InterPro serves as a universal platform for protein investigations, providing a standardized terminology for protein domains and functional regions. This enables researchers to gather data from multiple resources, facilitating the comparison of protein attributes across various sources. Essentially, InterPro plays a crucial role in

bioinformatics research by facilitating seamless comparisons of protein properties across different species.



List of Hits on InterPro:-

ACCESSION	NAME	SOURCE DATABASE	MATCHES
IPR011333	SKP1/BTB/POZ domain superfamily	InterPro	50
IPR039948	Elongin-C	InterPro	50
IPR001232	S-phase kinase-associated protein 1-like	InterPro	50
IPR016073	SKP1 component, POZ domain	InterPro	50

Interpretation of the result:
Sphingomyelin synthase, also referred to as phosphatidylcholine:ceramide cholinephosphotransferase, is an enzyme with bidirectional lipid cholinephosphotransferase activity. It can convert phosphatidylcholine (PC) and ceramide into sphingomyelin (SM) and diacylglycerol (DAG), and vice versa. This category also encompasses proteins related to sphingomyelin synthase, such as SAMD8 in humans, SMSr in Drosophila melanogaster, and Protein PHLOEM UNLOADING MODULATOR in Arabidopsis.

This domain is present in sphingomyelin synthase, also known as phosphatidylcholine:ceramide cholinephosphotransferase, and other associated proteins. Sphingomyelin synthase functions bidirectionally as a lipid cholinephosphotransferase, facilitating the conversion of phosphatidylcholine (PC) and ceramide into sphingomyelin (SM) and diacylglycerol (DAG), and vice versa. It shares similarity with the C-terminal region of phosphatidic acid phosphatase type 2 (PAP2).

Source:
https://www.ebi.ac.uk/interpro/result/InterProScan/iprscan5-R20240401-111735-0744-38460661-p1m/

2. Use your protein sequence as query in a BLAST search against the RefSeq database. All other parameters can be taken as default parameters. Find two sets of 5 homologs from different organisms using the following criteria:

Set #1: Get the hits which have E-value between 10-10 and 0.01 and select 4 homologs from these hits. Add your sequence in this set (so the set has 5 sequences in total). List the sequence, % similarity and E-value for each homolog.

Set #2: Get the hits which have E-value more than 0.01 and select 4 homologs from these hits. Add your sequence in this set. Show the sequence, % similarity and E-value for each homolog.

If you don't find the required hits with the above criteria, you can relax the criteria or vary parameters (such as BLOSUM/PAM matrices, or the organisms in which search is being conducted) to vary the number of hits. If you still don't get desired e-values, use hits with other e-values as close as possible to the desired e-values; just explain what you are doing in your report.

Answer:
SET#1
Given Sequence:

>8IJ1_C Chain C, Elongin-C [Homo sapiens]
MAMYVKLISSDGHEFIVKREHALTSGTIKAMLSGPGQFAENETNEVNFREIPSHVLSKV
CMYFTYKVRYTNSSTEIPEFPIAPEIALELLMAANFLDC

Homolog 1: Opisthorchis viverrini

hypothetical protein T265_04891 [Opisthorchis viverrini]

Sequence ID: XP_009168015.1 Length: 357 Number of Matches: 2

Range 1: 79 to 124 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
62.4 bits(150)	6e-09	Composition-based stats.	29/46(63%)	34/46(73%)	0/46(0%)

```
Query  2    AMYVKLISSDGHEFIVKREHALTSGTIKAMLSGPGQFAENETNEVN  47
            AMYVKL+SSD HEF V+RE+AL SGTIKAMLSGP     +    +N
Sbjct  79   AMYVKLVSSDDHEFYVRREYALISGTIKAMLSGPASARPSSVTTLN  124
```

E-value: 6e-09
Percentage Identity: 63.04%

Sequence:
>XP_009168015.1 hypothetical protein T265_04891 [Opisthorchis viverrini]
MTEVQAQPSSASEQKYGGAEGADAMYVKLVSSDDHEFYVRREYALISGTIKAMLSGPEVQAQPSSASEQK
YGGAEGADAMYVKLVSSDDHEFYVRREYALISGTIKAMLSGPASARPSSVTTLNTSKSVFKPRKPVYLAT
FNVRTLKQAGQQVAFARTLDSLCIDVCCLSETRTQYASVVIKQTAPSLSYRFRLRTSGDAKAAVAGYAGI
PVDSRLCAGRLATLVRESRGSEVHRTLFIVSAYAPTACSSESGRDSFYDALDALQQQAKSSDIVVVAGDM
NAQVKVNVRADQGAWWIRKAQEMEDAKNTGDVRKLFHLIRSTDPRKPLVSEIIRDQNGSLKCSKAERLAC
WAQYFEQ

Homolog 2: Penicillium samsonianum

SKP1 component POZ [Penicillium samsonianum]

Sequence ID: XP_057129221.1 Length: 101 Number of Matches: 1

Range 1: 9 to 100 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
63.5 bits(153)	1e-10	Compositional matrix adjust.	34/94(36%)	52/94(55%)	2/94(2%)

```
Query  4   YVKLISSDGHEFIVKREHALTSGTIKAMLSGPGQFAENETNEVNFREIPSHVLSKVCMYF  63
           +V ++SSDG EFI+ R  A  S T +  LS    F E  + E    +    ++ K+C Y
Sbjct  9   FVTIVSSDGFEFIIPRSAAYVSETFRVALSS-TNFPEGVSGEYVLGDYSGVIVEKICEYL  67

Query  64  TYKVRYTNSSTEIPEFPIAPEIALELLMAANFLD  97
            Y  ++ +    +P+  I PE+ LELLMAA+FL+
Sbjct  68  CYNEKHKDQ-VNVPDMDIPPELCLELLMAADFLN  100
```

E-Value: 1e-10
Percent Similarity: 36%
Sequence:
>XP_057129221.1 SKP1 component POZ [Penicillium samsonianum]
MAPSTDSEFVTIVSSDGFEFIIPRSAAYVSETFRVALSSTNFPEGVSGEYVLGDYSGVIV
EKICEYLCYN
EKHKDQVNVPDMDIPPELCLELLMAADFLNT

Homolog 3: Kazachstania africana CBS 2517

hypothetical protein KAFR_0H01430 [Kazachstania africana CBS 2517]

Sequence ID: XP_003958688.1 Length: 101 Number of Matches: 1

Range 1: 8 to 101 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
62.0 bits(149)	5e-10	Compositional matrix adjust.	34/97(35%)	55/97(56%)	6/97(6%)

```
Query  5   VKLISSDGHEFIVKREHALTSGTIKAMLSGPGQFAENETNEVNFREIPSHVLSKVCMYFT  64
           V L++SDG E  +  E AL S T+K ML GP +   ++    NF     HV+ K   Y
Sbjct  8   VNLVASDGSEHTISIEAALLSPTLKTMLEGPFKKDGSKIELTNFE---PHVVQKAAEYLQ  64

Query  65  YKVRYTN---SSTEIPEFPIAPEIALELLMAANFLDC  98
           +K++Y +      ++PEF +  E++LELL+ A++L+
Sbjct  65  HKLKYQDVDVKKEDVPEFVVPTEMSLELLLIADYLNI  101
```

E-Value: 5e-10
Percent Similarity:35%
Sequence:
>XP_003958688.1 hypothetical protein KAFR_0H01430 [Kazachstania africana CBS 2517]
MSDTDGLVNLVASDGSEHTISIEAALLSPTLKTMLEGPFKKDGSKIELTNFEPHVVQKAAEYLQHKLKYQ
DVDVKKEDVPEFVVPTEMSLELLLIADYLNI

Homolog 4: Suillus paluster

BTB/POZ protein, partial [Suillus paluster]

Sequence ID: XP_041172880.1 **Length:** 101 **Number of Matches:** 1

Range 1: 2 to 94 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
62.8 bits(151)	2e-10	Compositional matrix adjust.	38/94(40%)	54/94(57%)	7/94(7%)

```
Query  4   YVKLISSDGHEFIVKREHALTSGTIKAMLSGPGQFAENETNEVNFREIPSHVLSKVCMYF  63
           +VK+ SSDG+ F+VKR  A+TSGT+K MLS    F E   N     E  + V+ KVC Y
Sbjct  2   WVKITSSDGYSFLVKRSVAVTSGTLKNMLSEDSSFKEAIANTCPISE-RAAVVEKVCEYM  60

Query  64  TYKVRY----TNSSTEIPEFP--IAPEIALELLM  91
           +++  Y    +    ++ EF   I PE+ALEL +
Sbjct  61  SFRAYYEGPGSKEGVDVNEFTERIPPEVALELCV  94
```

E-Value: 2e-10
Percent Similarity: 40%
Sequence:

```
>XP_041172880.1 BTB/POZ protein, partial [Suillus paluster]
DWVKITSSDGYSFLVKRSVAVTSGTLKNMLSEDSSFKEAIANTCPISERAAVVEKVCEYMSFRAYYEGPG
SKEGVDVNEFTERIPPEVALELCVTLLSVPL
```

SET#2

Given Sequence:
MAMYVKLISSDGHEFIVKREHALTSGTIKAMLSGPGQFAENETNEVNFREIPSHVLSKV
CMYFTYKVRYTNSSTEIPEFPIAPEIALELLMAANFLDC

Homolog 1: Kazachstania africana CBS 2517

uncharacterized protein C9374_006263 [Naegleria lovaniensis]

Sequence ID: XP_044546954.1 **Length:** 129 **Number of Matches:** 1

Range 1: 31 to 129 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
43.5 bits(101)	0.011	Compositional matrix adjust.	27/102(26%)	50/102(49%)	10/102(9%)

```
Query  4    YVKLISSDGHEFIVKREHALTSGTIKAMLSGPGQFAENETNEVNFREIPSHVLSKVCMYF  63
            Y+ LIS++  EFI+ ++ A  S  +  +++         +N +   +I + VL  +C +
Sbjct  31   YITLISAEKFEFILSKKAAAQSKYLHQLITDD---VFQTSNRITLHDISTDVLELLCQFL  87

Query  64   TYKVRYTNSSTEI-------PEFPIAPEIALELLMAANFLDC  98
              K    N  +         P+ P   +I +ELL+A+N+LDC
Sbjct  88   VDKSIKGNFMSTFNPLQDLDPQNPDHRQIVIELLLASNYLDC  129
```

E-Value: 0.011
Percent Similarity: 26%
Sequence:

>XP_003958688.1 hypothetical protein KAFR_0H01430 [Kazachstania africana CBS 2517]
MSDTDGLVNLVASDGSEHTISIEAALLSPTLKTMLEGPFKKDGSKIELTNFEPHVVQKAAEYLQHKLKYQ DVDVKKEDVPEFVVPTEMSLELLLIADYLNI

Homolog 2: Puccinia graminis f. sp. tritici CRL 75-36-700-3

E3 ubiquitin ligase complex SCF subunit sconC [Puccinia graminis f. sp. tritici CRL 75-36-700-3]

Sequence ID: XP_003335788.2 **Length:** 158 **Number of Matches:** 1

Range 1: 2 to 106 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
43.1 bits(100)	0.025	Compositional matrix adjust.	30/109(28%)	50/109(45%)	20/109(18%)

```
Query  5    VKLISSDGHEFIVKREHALTSGTIKAMLSGPGQFAENETNEVNFREIPSHVLSKVCMYFT  64
            V +++SDG EFIV++E A  S  IK M+   G+      N +    + + VL KV  +
Sbjct  2    VLMVTSDGEEFIVEKEVATRSALIKNMIEDLGE----SDNPIPLPNVSASVLKKVLEWCE  57

Query  65   YKVRYTNSSTEIPE----------------FPIAPEIALELLMAANFLD  97
            +  +    S E P+                 +  E+  E+++AAN+LD
Sbjct  58   HHKKDPEPSAEDPDDARKRATEISDWDTKFINVDQEMLFEIILAANYLD  106
```

E-Value: 0.025
Percent Similarity: 28%
Sequence:
>XP_003335788.2 E3 ubiquitin ligase complex SCF subunit sconC [Puccinia graminis f. sp. tritici CRL 75-36-700-3]
MVLMVTSDGEEFIVEKEVATRSALIKNMIEDLGESDNPIPLPNVSASVLKKVLEWCEHHKKDPEPSAEDP DDARKRATEISDWDTKFINVDQEMLFEIILAANYLDIKPLLDVGCKSVANMIKGKQPEEIRKLFNIANDF TPEEEAQIKKENEWAEDR

Homolog 3:

uncharacterized protein TraAM80_06691 [Trypanosoma rangeli]

Sequence ID: XP_029236641.1 **Length:** 172 **Number of Matches:** 1

Range 1: 48 to 172 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
43.1 bits(100)	0.027	Compositional matrix adjust.	39/127(31%)	53/127(41%)	34/127(26%)

```
Query  4    YVKLISSDGHEFIVKREHALTSGTIKAML-------------------SGPGQFAENETN  44
            YV ++S DG EFI+    A  S  I ++L                   S PG  A N  N
Sbjct  48   YVCMLSGDGMEFIIPEAAARQSKMISSLLDAIYSLPNRGGFGESLQKKSTPGVLAVNNVN  107

Query  45   E---VNFREIPSHVLSKVCMYFTYKVRYTNSSTEIPEF-------PIAPE---IALELLM  91
                +    + S  L  VC Y   +     +STE  EF       P++ E   I  ELL+
Sbjct  108  MMPMIPLEPLSSRTLELVCRYLLQRSTGDPNSTE--EFSLLGELDPMSDEDQDIVSELLL  165

Query  92   AANFLDC  98
            AA+F+DC
Sbjct  166  AADFIDC  172
```

E-Value: 0.027

Percent Similarity: 31%
Sequence:
>XP_029236641.1 uncharacterized protein TraAM80_06691 [Trypanosoma rangeli]
MAEERGVDIPVNSEANLQTDTDMCAARLGGQPSTEVREPAWPVEPLPYVCMLSGDGMEFIIPE
AAARQSKMISSLLDAIYSLPNRGGFGESLQKKSTPGVLAVNNVNMMPMIPLEPLSSRTLELVCR
YLLQRSTGDPNSTEEFSLLGELDPMSDEDQDIVSELLLAADFIDC

Homolog 4: Tetrapisispora phaffii CBS 4417

hypothetical protein TPHA_0C01720 [Tetrapisispora phaffii CBS 4417]

Sequence ID: XP_003684762.1 **Length:** 97 **Number of Matches:** 1

Range 1: 4 to 97 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
42.7 bits(99)	0.013	Compositional matrix adjust.	35/97(36%)	56/97(57%)	6/97(6%)

```
Query  5   VKLISSDGHEFIVKREHALTSGTIKAMLSGPGQFAENETNEVNFREIPSHVLSKVCMYFT  64
           + L+S D  EF V +E  + S T+KAM+  P  F EN + ++      S VL+ +  Y
Sbjct  4   ITLVSKDNVEFEVPKEVIIISQTLKAMVDSP--FIEN-SGKITLTNFDSPVLAVIVDYLN  60

Query  65  YKVRYTN---SSTEIPEFPIAPEIALELLMAANFLDC  98
           Y  +Y +   +  +IPEF I  E++LELL+AA++L+
Sbjct  61  YNFKYKDEDPTKVDIPEFEIPTELSLELLLAADYLNI  97
```

E-Value: 0.013
Percent Similarity: 36%
Sequence:
>XP_003684762.1 hypothetical protein TPHA_0C01720 [Tetrapisispora phaffii CBS 4417]
MDIITLVSKDNVEFEVPKEVIIISQTLKAMVDSPFIENSGKITLTNFDSPVLAVIVDYLNYNFKYKDE
DPTKVDIPEFEIPTELSLELLLAADYLNI

Note: I have increased the Max target sequences by 5000 and the expected threshold to 10 to get the result.

3. Use the Clustal Omega Server (default options) available from EBI to do multiple sequence alignment of Set #1 along with your query.

https://www.ebi.ac.uk/jdispatcher/msa/clustalo
How many positions are absolutely conserved? Save the alignment as it will be needed later.

Answer:

```
CLUSTAL O(1.2.4) multiple sequence alignment


XP_041172880.1
------------------------------------------------------------   0
8IJ1_C
------------------------------------------------------------   0
```

```
XP_009168015.1
MTEVQAQPSSASEQKYGGAEGADAMYVKLVSSDDHEFYVRREYALISGTIKAMLSGPEVQ   60
XP_057129221.1
------------------------------------------------------------   0
XP_003958688.1
------------------------------------------------------------   0
XP_041172880.1
-------------------DWVKITSSDGYSFLVKRSVAVTSGTLKNMLSEDSSFKEAIA   41
8IJ1_C
-----------------MAMYVKLISSDGHEFIVKREHALTSGTIKAMLSGPGQFAENET   43
XP_009168015.1
AQPSSASEQKYGGAEGADAMYVKLVSSDDHEFYVRREYALISGTIKAMLSGPASARPSSV   120
XP_057129221.1
------------MAPSTDSEFVTIVSSDGFEFIIPRSAAYVSETFRVALSSTN-FPEGVS   47
XP_003958688.1
--------------MSDTDGLVNLVASDGSEHTISIEAALLSPTLKTMLEGPF-KKDGS-   44
                     *.: :**. .. :  . *  * *::  *.
XP_041172880.1
NTCPISERA-AVVEKVCEYMSFRAYYEGPGSKEGVDVNEFTERIPPEVALELCVTLLSVP   100
8IJ1_C
NEVNFREIPSHVLSKVCMYFTYKVRYTNSSTE----IPEF--PIAPEIALELLMAANFL-   96
XP_009168015.1
TTLNTSKSVFKP-RKPVYLATFNVRTLKQAGQ----QVAF----------ARTLDSLCID   165
XP_057129221.1
GEYVLGDYSGVIVEKICEYL----CYNEKH-KDQVNVPDM--DIPPELCLELLMAADFLN   100
XP_003958688.1
-KIELTNFEPHVVQKAAEYLQHKLKYQDVD-VKKEDVPEF--VVPTEMSLELLLIADYLN   100
      .       *                        :             :    :
XP_041172880.1
L-----------------------------------------------------------   101
8IJ1_C
-DC---------------------------------------------------------   98
XP_009168015.1
VCCLSETRTQYASVVIKQTAPSLSYRFRLRTSGDAKAAVAGYAGIPVDSRLCAGRLATLV   225
XP_057129221.1
T-----------------------------------------------------------   101
XP_003958688.1
I-----------------------------------------------------------   101
```

```
XP_041172880.1
------------------------------------------------------------   101
8IJ1_C
------------------------------------------------------------   98
XP_009168015.1
RESRGSEVHRTLFIVSAYAPTACSSESGRDSFYDALDALQQQAKSSDIVVVAGDMNAQVK   285
XP_057129221.1
------------------------------------------------------------   101
XP_003958688.1
------------------------------------------------------------   101
XP_041172880.1
------------------------------------------------------------   101
8IJ1_C
------------------------------------------------------------   98
XP_009168015.1
VNVRADQGAWWIRKAQEMEDAKNTGDVRKLFHLIRSTDPRKPLVSEIIRDQNGSLKCSKA   345
XP_057129221.1
------------------------------------------------------------   101
XP_003958688.1
------------------------------------------------------------   101
XP_041172880.1      ------------  101
8IJ1_C              ------------  98
XP_009168015.1      ERLACWAQYFEQ  357
XP_057129221.1      ------------  101
XP_003958688.1      ------------  101
```

There are a total of 8 positions that are absolutely conserved.
Source:
https://www.ebi.ac.uk/jdispatcher/msa/clustalo/summary?jobId=clustalo-I20240401-155820-0951-43818169-p1m

4. Repeat the above analysis for Set #2, show the results, and explain any differences in the number of absolutely conserved positions seen for Set #1 and Set #2. Save the alignment.

Answer:

```
CLUSTAL O(1.2.4) multiple sequence alignment
```

```
XP_029236641.1
MAEERGVDIPVNSEANLQTDTDMCAARLGGQPSTEVREPAWPVEPLPYVCMLSGDGMEFI   60
XP_003335788.2
-----------------------------------------------MVLMVTSDGEEFI   13
8IJ1_C
---------------------------------------------MAMYVKLISSDGHEFI   16
XP_003958688.1
----------------------------------------MSDTDGLVNLVASDGSEHT   19
XP_003684762.1
----------------------------------------------MDIITLVSKDNVEFE   15
                                                : ::: *. *.
XP_029236641.1
IPEAAARQSKMISSLLDAIYSLPNRGGFGESLQKKSTPGVLAVNNVNMMPMIPLEPLSSR   120
XP_003335788.2
VEKEVATRSALIKNMIEDLGESD--------------------------NPIPLPNVSAS   47
8IJ1_C
VKREHALTSGTIKAMLSGPGQFA-------------------E---NETNEVNFREIPSH   54
XP_003958688.1
ISIEAALLSPTLKTMLEGPFKKD-------------------G------SKIELTNFEPH   54
XP_003684762.1
VPKEVIIISQTLKAMVDSPFIEN-------------------S------GKITLTNFDSP   50
:       *  :. ::.                                  : :  .

XP_029236641.1
TLELVCRYLLQRSTGDPNST---E-----EFSLLGELDPMSDEDQDIVSELLLAADFIDC   172
XP_003335788.2
VLKKVLEWCEHHKKDPEPSAEDPDDARKRATEISDWDTKFINVDQEMLFEIILAANYLDI   107
8IJ1_C
VLSKVCMYFTYKVRYTNS----------------STEIPEFPIAPEIALELLMAANFLDC   98
XP_003958688.1
VVQKAAEYLQHKLKYQDVDV-------------KKEDVPEFVVPTEMSLELLLIADYLNI   101
XP_003684762.1
VLAVIVDYLNYNFKYKDEDP-------------TKVDIPEFEIPTELSLELLLAADYLNI   97
.:     :   .                                  ::  *::: *::::

XP_029236641.1
---------------------------------------------------- 172
XP_003335788.2
KPLLDVGCKSVANMIKGKQPEEIRKLFNIANDFTPEEEAQIKKENEWAEDR 158
8IJ1_C
---------------------------------------------------- 98
XP_003958688.1
---------------------------------------------------- 101
XP_003684762.1
---------------------------------------------------- 97
```

There are a total of 5 positions that are absolutely conserved.

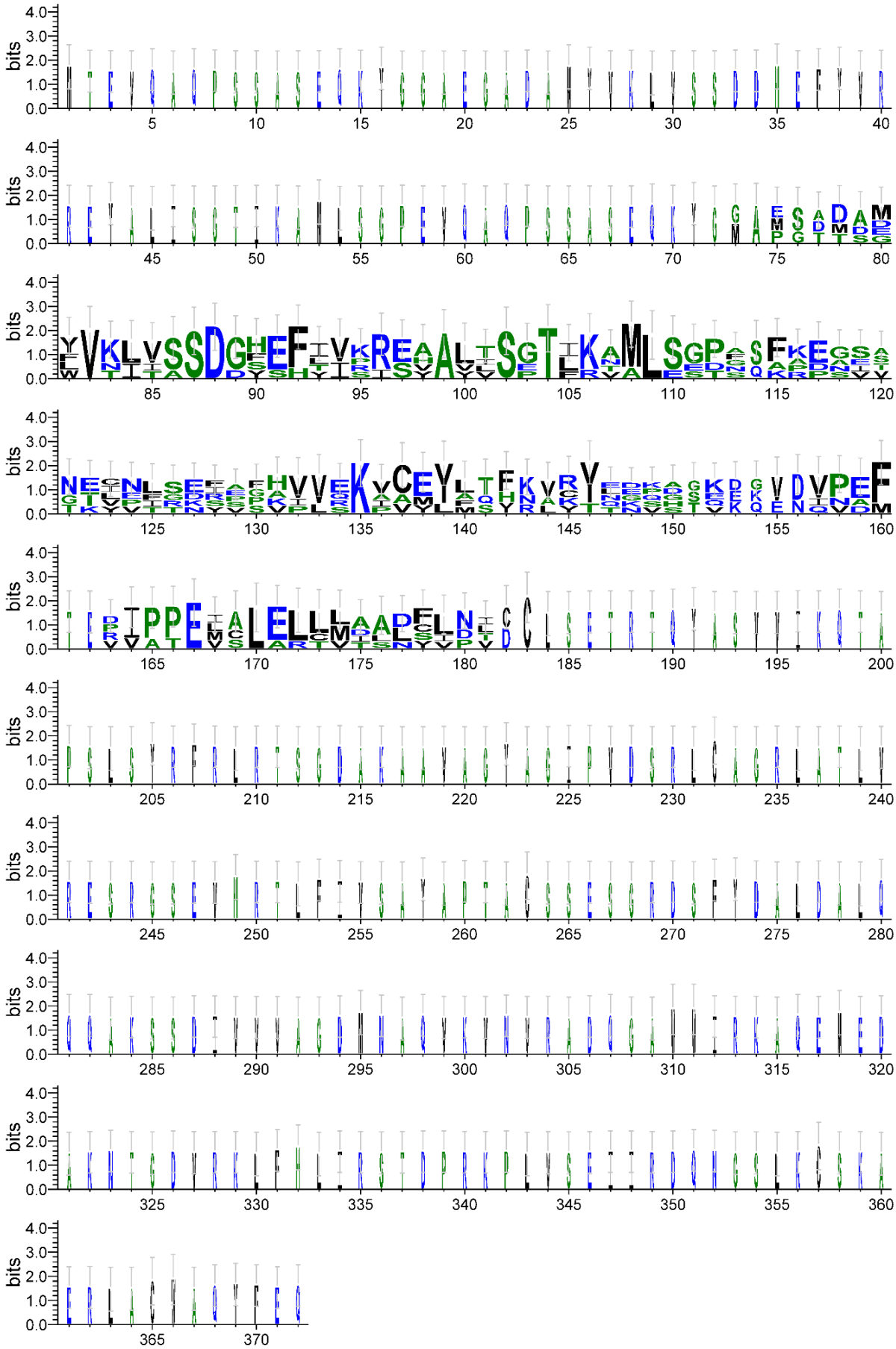
The presence of positions that are absolutely conserved in a multiple sequence alignment indicates their critical role in maintaining protein function or structure. The higher number of absolutely conserved positions in Set #1 compared to Set #2 implies that the sequences in Set #1 are more functionally or structurally significant. This could be attributed to the closer relationship between the sequences in Set #1, as closely related sequences are more likely to share common functional or structural features, resulting in more conserved positions. This is supported by all sequences in Set #1 having BLAST E values <0.01, indicating high similarity. Conversely, sequences in Set #2 may be more divergent and less related, resulting in fewer absolutely conserved positions. The higher BLAST E values (>0.01) for Set #2 suggest lower similarity among its sequences compared to those in Set #1.

Source:
https://www.ebi.ac.uk/jdispatcher/msa/clustalo/summary?jobId=clustalo-I20240401-161635-0329-78815151-p1m

5. Compute and show the sequence logos for the each of the above two multiple alignments, using the WebLogo3 program. http://weblogo.threeplusone.com/create.cgi. Adjust the display parameters to obtain a clear and usable image. Comment on the difference between the sequence logos obtained for Set #1 and Set #2. If the webLogo link http://weblogo.threeplusone.com/create.cgi is not working, you can use the older version of the server which is available at http://weblogo.berkeley.edu/logo.cgi (in this tool, make sure to uncheck the option of "Small sample correction").

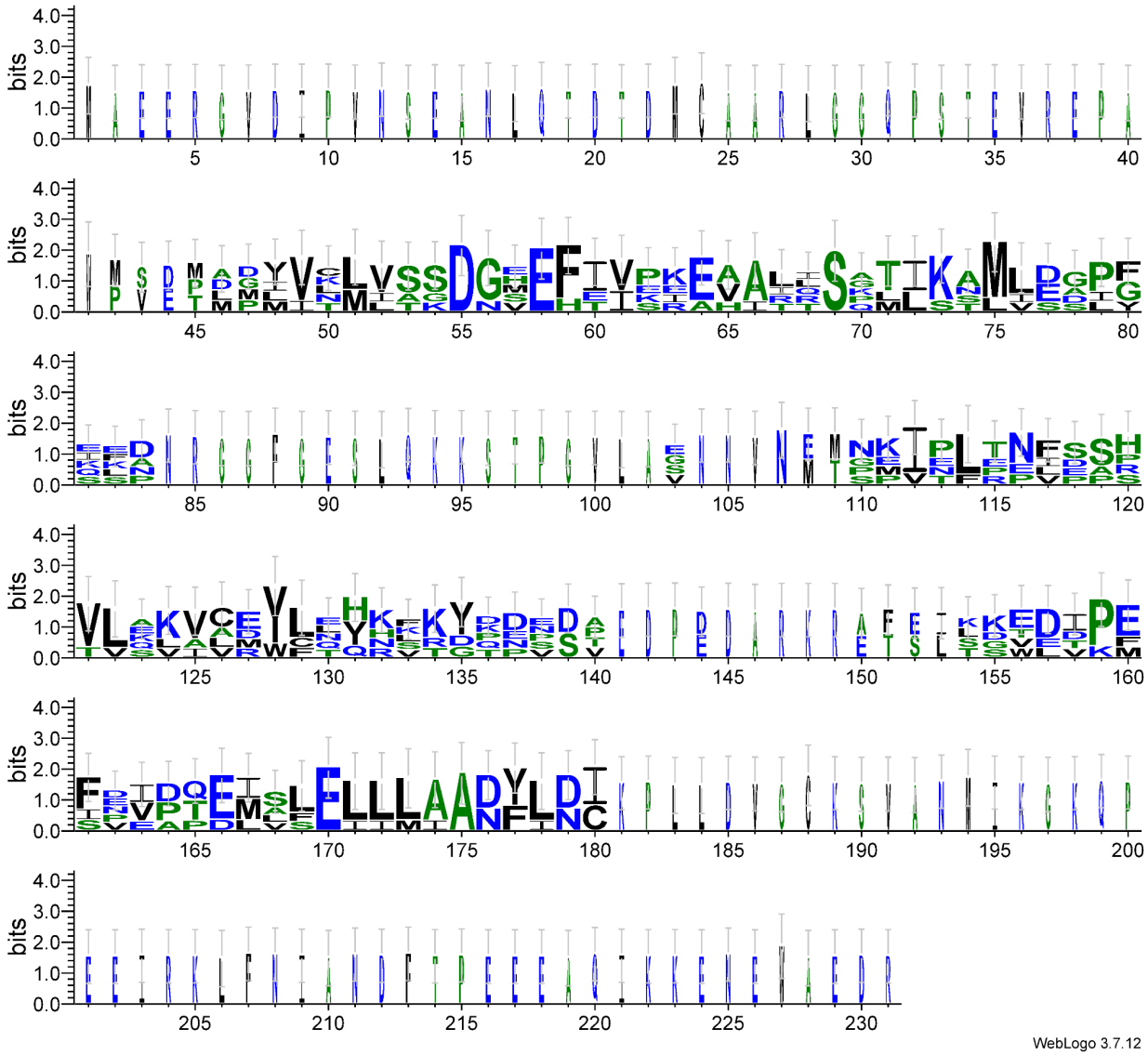
Answer:

MSA Logo for Set #1



WebLogo 3.7.12

MSA Logo for Set #2:



Looking at the two Multiple Sequence Alignments (MSAs), it's evident that Set 1 has slightly more similar amino acids compared to Set 2. This difference is because Set 2 has undergone more mutations, indicating that the homologs in Set 2 are further apart compared to those in Set 1.

6. Using the above sequence logos, can you discover any conserved motif in your protein? How does it relate to the information you found in Question 1?

Answer:

By the analysing set1, we see bigger letter at place (82 to 170) which implies that that a trend of higher conservation towards this position, indicating potential motifs located in this position.
By the analysing set 2, we see bigger letter at place (55 to 175) which implies that that a trend of higher conservation towards this position, indicating potential motifs located in this position.

This observation does not align completely with the findings from Question 1, however some initial portion are only aligiend. where the some of domains were detected closer to the sequence's starting point.

7. Use the MEME program (http://meme-suite.org/tools/meme) to find motifs in Set #1. Show the 3 best motifs identified. How do these motifs relate to the sequence logo for Set #1 identified in the Question 5.

Answer:

DISCOVERED MOTIFS

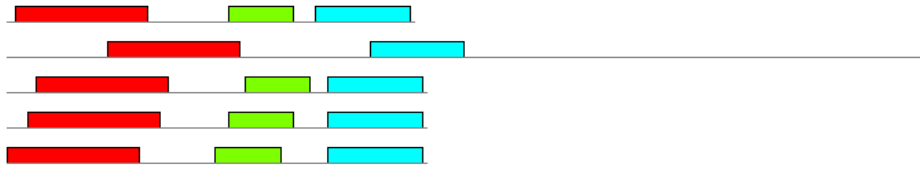
	Logo	E-value	Sites	Width	More	Submit/Download
1.		8.2e-023	5	32		
2.		1.1e-003	5	23		
3.		2.0e-003	4	16		

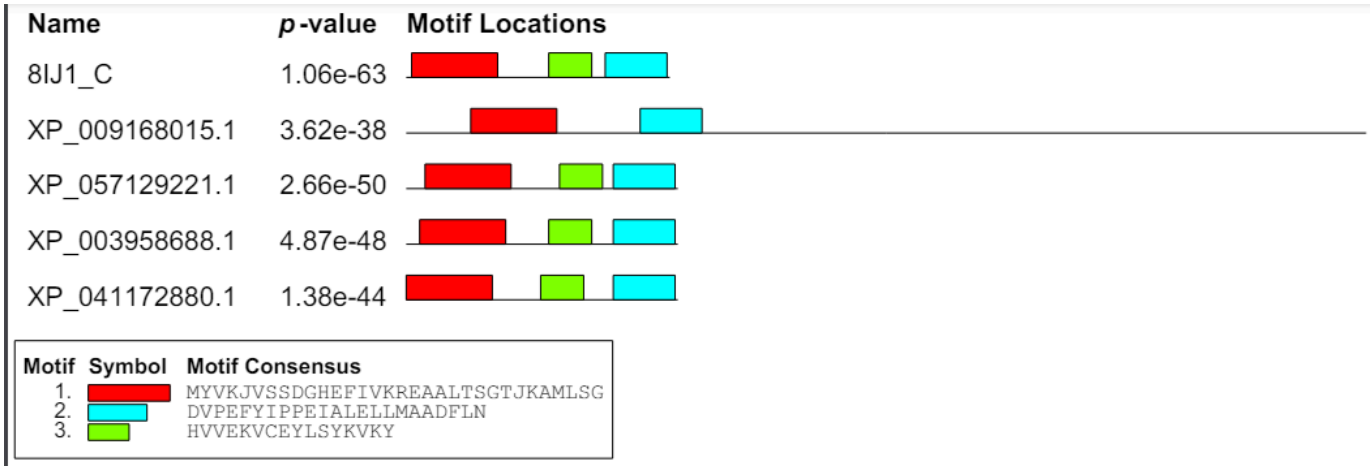
Stopped because requested number of motifs (3) found.

MOTIF LOCATIONS

Only Motif Sites Motif Sites+Scanned Sites All Sequences Download PDF Download SVG

	Name	p-value	Motif Locations
1.	8IJ1_C	1.06e-63	
2.	XP_009168015.1	3.62e-38	
3.	XP_057129221.1	2.66e-50	
4.	XP_003958688.1	4.87e-48	
5.	XP_041172880.1	1.38e-44	



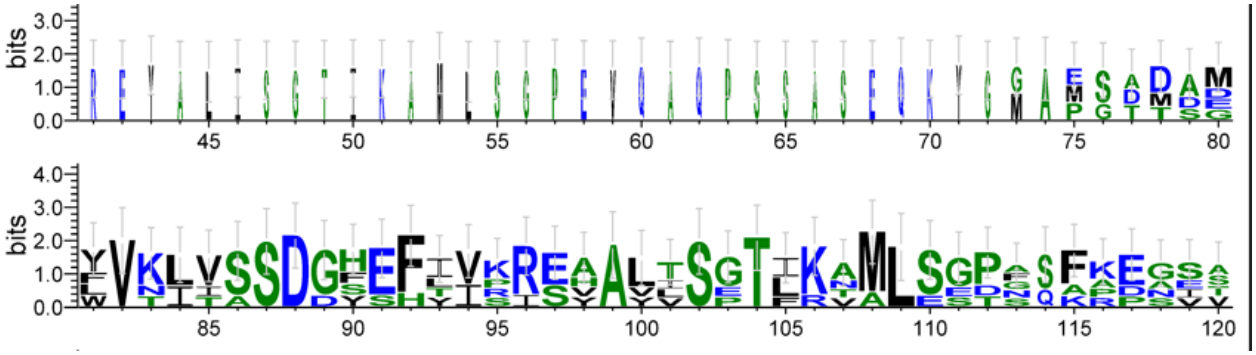


Motif 1: MYVKJVSSDGHEFIVKREAALTSGTJKAMLSG

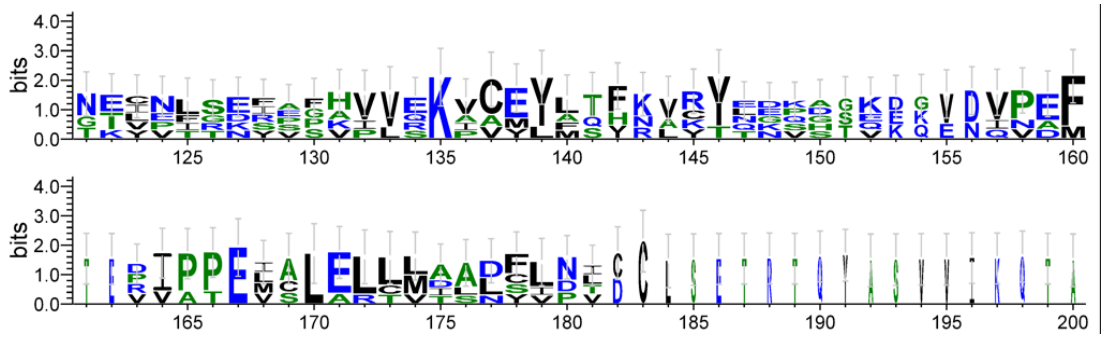
Motif 2: DVPEFYIPPEIALELLMAADFLN

Motif 3: HVVEKVCEYLSYKVKY

The **motif 1** can be seen in the set #1 at (80 to 111), here only with a exception of 2 amino acid in sequence in motif it is both J but in MSA it is L and I else other are same in both.



The **motif 2** can be seen similar sequence in MSA Set#1 at place (156 180), having some no match at place 161, 162, 163.



The **motif 3** also can be seen similar sequence in the set #1of MSA at place (131 to 146) having some differene at position in MSA at 141, 142, 145.

Source:
https://meme-suite.org/meme//opal-jobs/appMEME_5.5.51711988933759-1677616242/meme.html

8. Can you find out whether a structure has been determined for your protein? If so, what is the PDB ID? Which experimental method is used to solve this structure? If it is X-ray, what is the resolution of the protein?
Answer:
Yes, a match for my protein is found in the PDB database, indicating that the protein's structure has been identified.

Chain C, Elongin-C [Homo sapiens]

Sequence ID: 8IJ1_C **Length:** 98 **Number of Matches:** 1

See 4 more title(s) See all Identical Proteins(IPG)

Range 1: 1 to 98 GenPept Graphics ▼ Next Match ▲ Previous Match

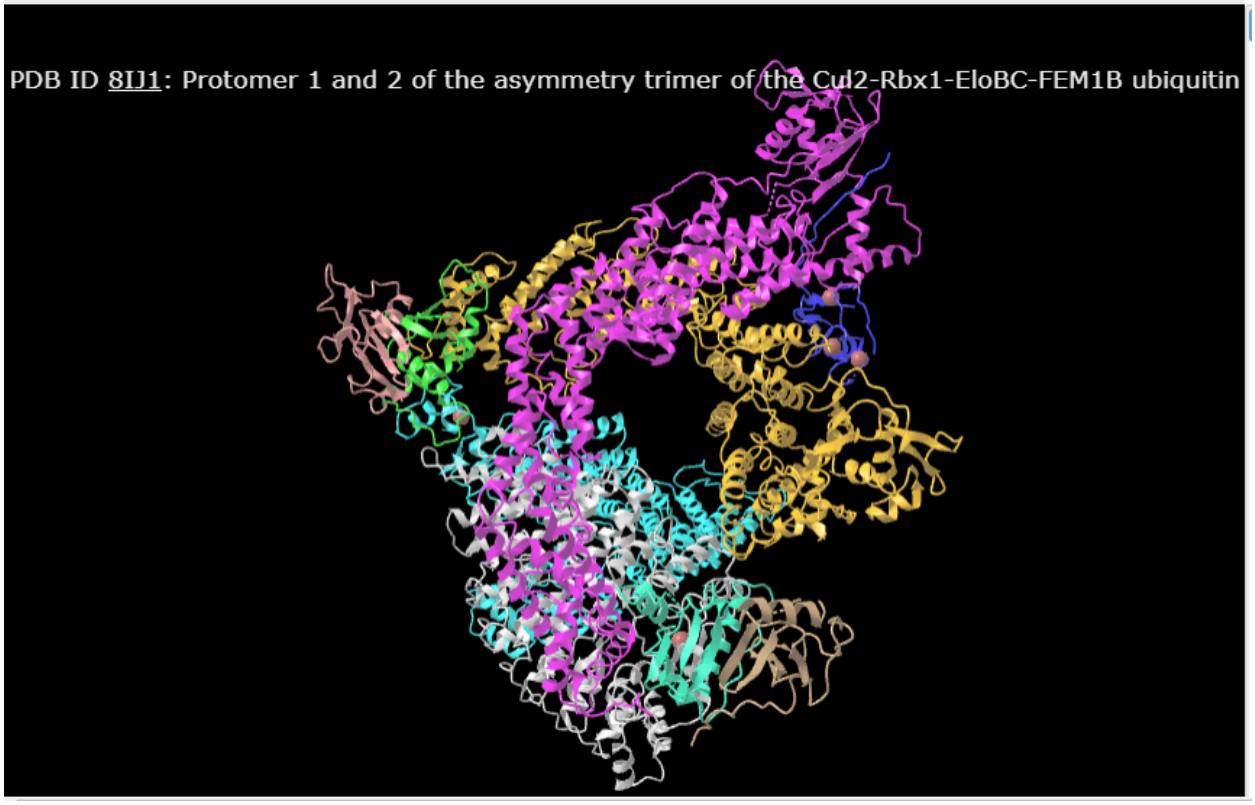
Score	Expect	Method	Identities	Positives	Gaps
205 bits(521)	4e-70	Compositional matrix adjust.	98/98(100%)	98/98(100%)	0/98(0%)

```
Query  1   MAMYVKLISSDGHEFIVKREHALTSGTIKAMLSGPGQFAENETNEVNFREIPSHVLSKVC  60
           MAMYVKLISSDGHEFIVKREHALTSGTIKAMLSGPGQFAENETNEVNFREIPSHVLSKVC
Sbjct  1   MAMYVKLISSDGHEFIVKREHALTSGTIKAMLSGPGQFAENETNEVNFREIPSHVLSKVC  60

Query  61  MYFTYKVRYTNSSTEIPEFPIAPEIALELLMAANFLDC  98
           MYFTYKVRYTNSSTEIPEFPIAPEIALELLMAANFLDC
Sbjct  61  MYFTYKVRYTNSSTEIPEFPIAPEIALELLMAANFLDC  98
```

PDB ID: 81J1

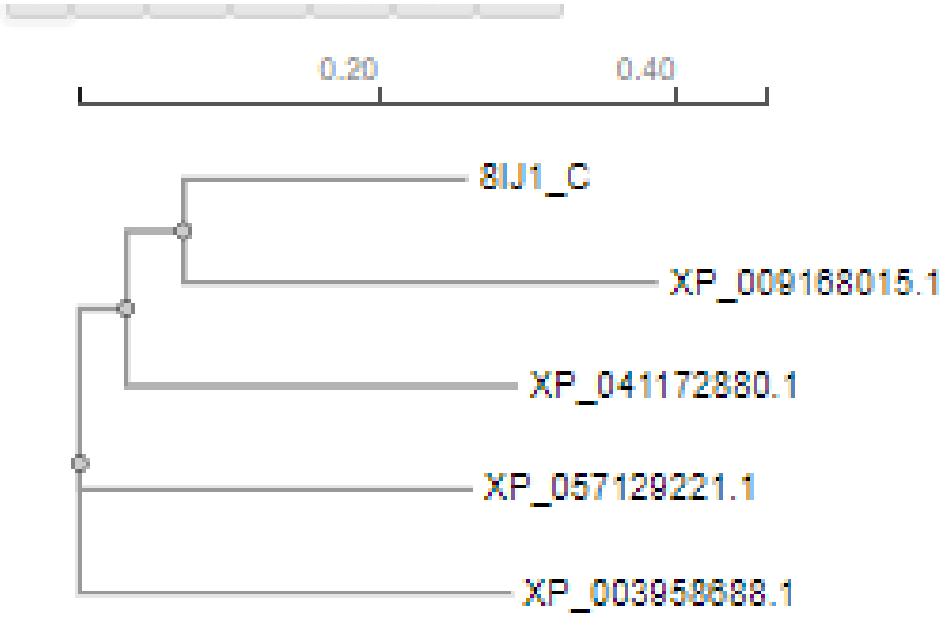
Method: electron microscopy
Structure:



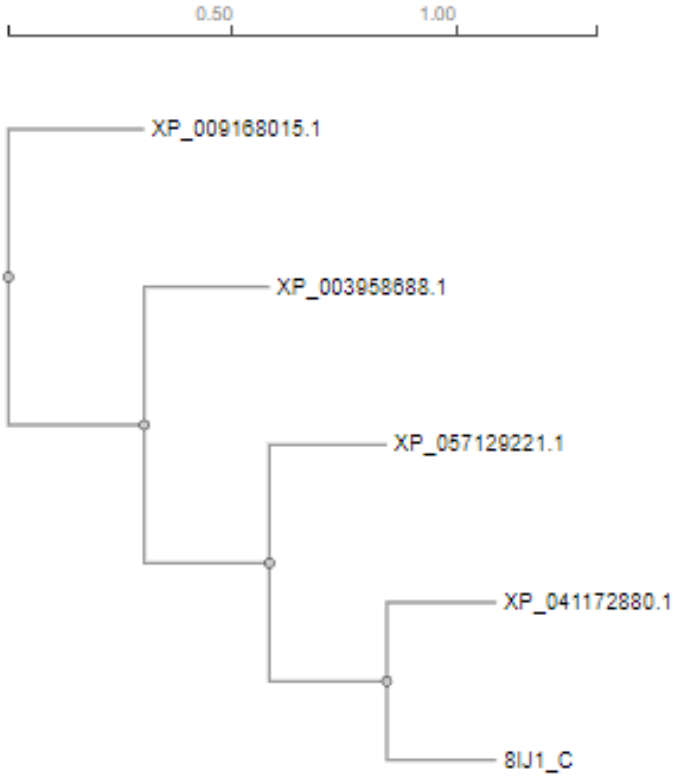
Source:
https://www.ncbi.nlm.nih.gov/Structure/mmdb/mmdbsrv.cgi?Dopt=s&uid=243259
https://blast.ncbi.nlm.nih.gov/Blast.cgi#alnHdr_8IJ1_C

9. Create a phylogenetic tree for the multiple sequence alignment of Set #1 using the EBI tool (https://www.ebi.ac.uk/jdispatcher/phylogeny/simple_phylogeny). Compare the results from UPGMA and Neighbor-joining method.

Answer:
Neighbor-joining method on Set #1:



UPGMA on Set #1:

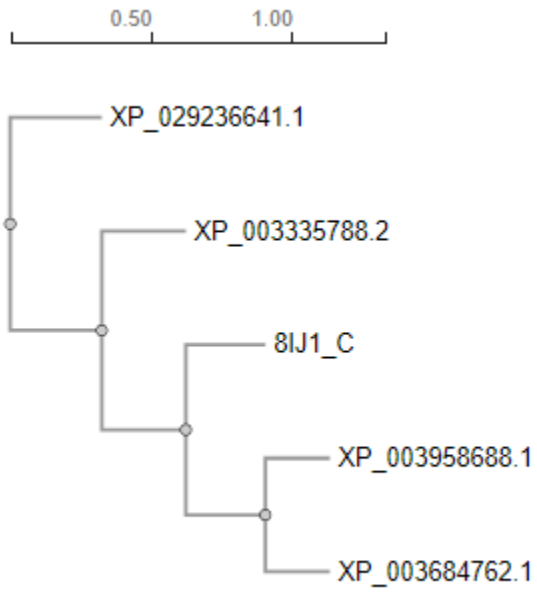


Result obtained from Neighbour joining method describes that my protein 81J1_C is more related to XP_009168015.1 and have same ancestor. Whereas, result obtained from UPGMA

method define that my protein is more related to XP_041172880.1 and descendant of all 3 other homolog.
Source:
https://www.ebi.ac.uk/jdispatcher/phylogeny/simple_phylogeny/summary?jobId=simple_phylogeny-I20240403-075341-0751-58162204-p1m

10. Create a phylogenetic tree using the multiple sequence alignment of Set #2 with UPGMA method. Compare the resulting tree with that obtained with UPGMA in Q9.

Answer:
UPGMA SET #2:-



Result obtained from UPGMA set#1 say that my protein 8IJ1_C is more related with only one homolog XP_009168015 and descendant of 3 other homolog. but result obtained from UPGMA set#2 say that my protein is more closely related with parent of two homolog and it is ancestor of two homolog and descendant of other two homolog.

Source:
https://www.ebi.ac.uk/jdispatcher/phylogeny/simple_phylogeny/summary?jobId=simple_phylogeny-I20240403-075725-0823-62176631-p1m