Hugh (Trevor) Redford
A17067426
htredford@ucsd.edu

Find a Gene Project

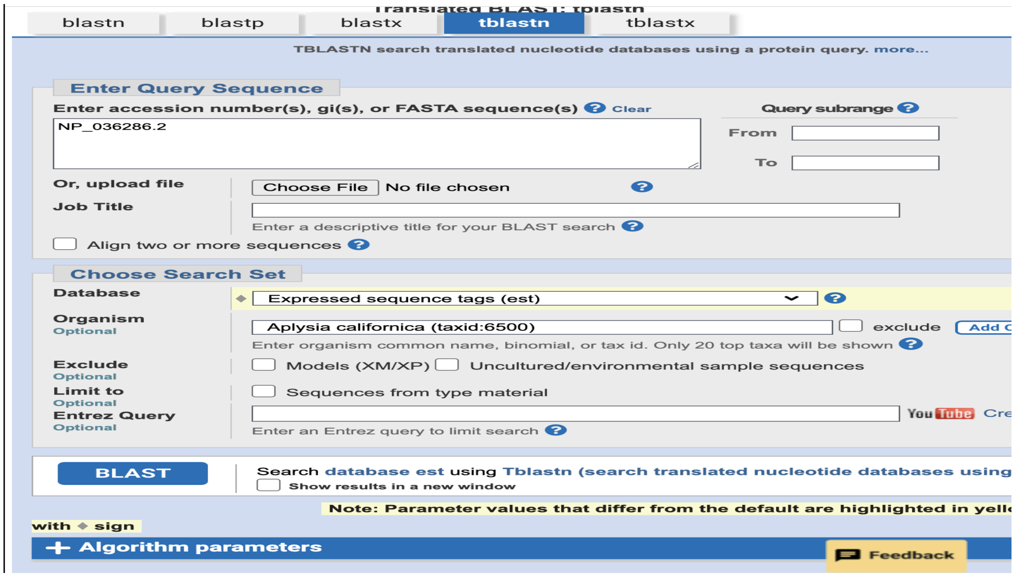
Questions:

[Q1] Tell me the name of a protein you are interested in. Include the species and the accession number. This can be a human protein or a protein from any other species as long as its function is known. If you do not have a favorite protein, select human RBP4 or KIF11. Do not use beta globin as this is in the worked example report that I provide you with online.

Name: Argonaute-2 (AGO2)
Accession Number: NP_036286.2
Species: Homo sapiens (Human)
Function: Argonaute-2 is a key protein in the RNA-induced silencing complex that facilitates gene silencing by guiding microRNAs (miRNAs) and small interfering RNAs (siRNAs) to degrade target messenger RNAs (mRNAs) or inhibit their translation.

[Q2] Perform a BLAST search against a DNA database, such as a database consisting of genomic DNA or ESTs. The BLAST server can be at NCBI or elsewhere. Include details of the BLAST method used, database searched and any limits applied (e.g. Organism).

Method: TBLASTN (2.7.1)
Database: Expressed Sequence Tags (est)
Organism: Aplysia californica (Sea Slug) - TaxID:6500



Translated BLAST: tblastn

blastn | blastp | blastx | tblastn | tblastx

TBLASTN search translated nucleotide databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) ? Clear

NP_036286.2

Query subrange ?

From

To

Or, upload file: Choose File No file chosen ?

Job Title:

Enter a descriptive title for your BLAST search ?

□ Align two or more sequences ?

Choose Search Set

Database: ◆ Expressed sequence tags (est) ?

Organism (Optional): Aplysia californica (taxid:6500) □ exclude Add o

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown ?

Exclude (Optional): □ Models (XM/XP) □ Uncultured/environmental sample sequences

Limit to (Optional): □ Sequences from type material

Entrez Query (Optional): YouTube Cre

Enter an Entrez query to limit search ?

BLAST

Search database est using Tblastn (search translated nucleotide databases using

□ Show results in a new window

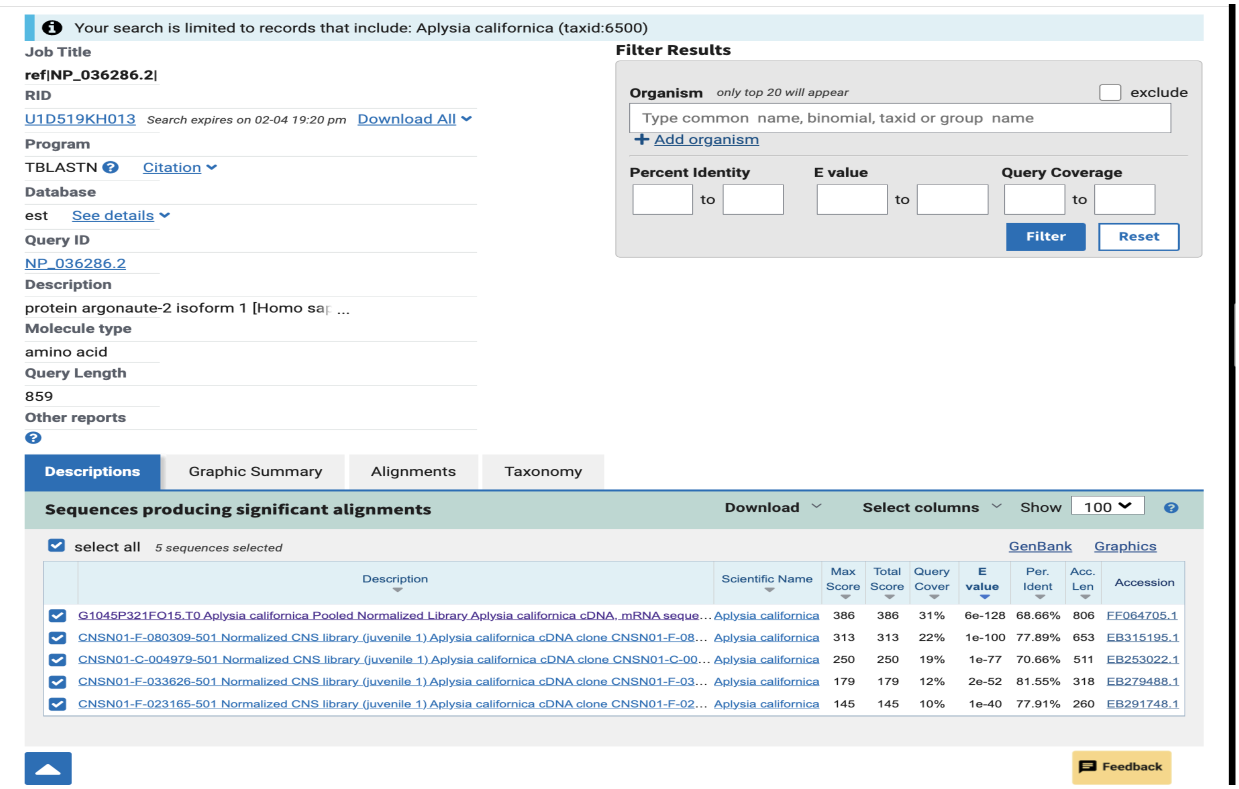
Note: Parameter values that differ from the default are highlighted in yell

with ◆ sign

+ Algorithm parameters

Feedback

Chosen Match: Accession FF064705.1. This gene is a 806 base pair clone of G1045P321FO15.T0 from Aplysia californica represents a putative homolog of Argonaute-2 (AGO2).



Your search is limited to records that include: Aplysia californica (taxid:6500)

Job Title
ref|NP_036286.2|

RID
U1D519KH013 *Search expires on 02-04 19:20 pm* Download All

Program
TBLASTN Citation

Database
est See details

Query ID
NP_036286.2

Description
protein argonaute-2 isoform 1 [Homo sa ...

Molecule type
amino acid

Query Length
859

Other reports

Filter Results

Organism *only top 20 will appear* □ exclude
Type common name, binomial, taxid or group name
+ Add organism

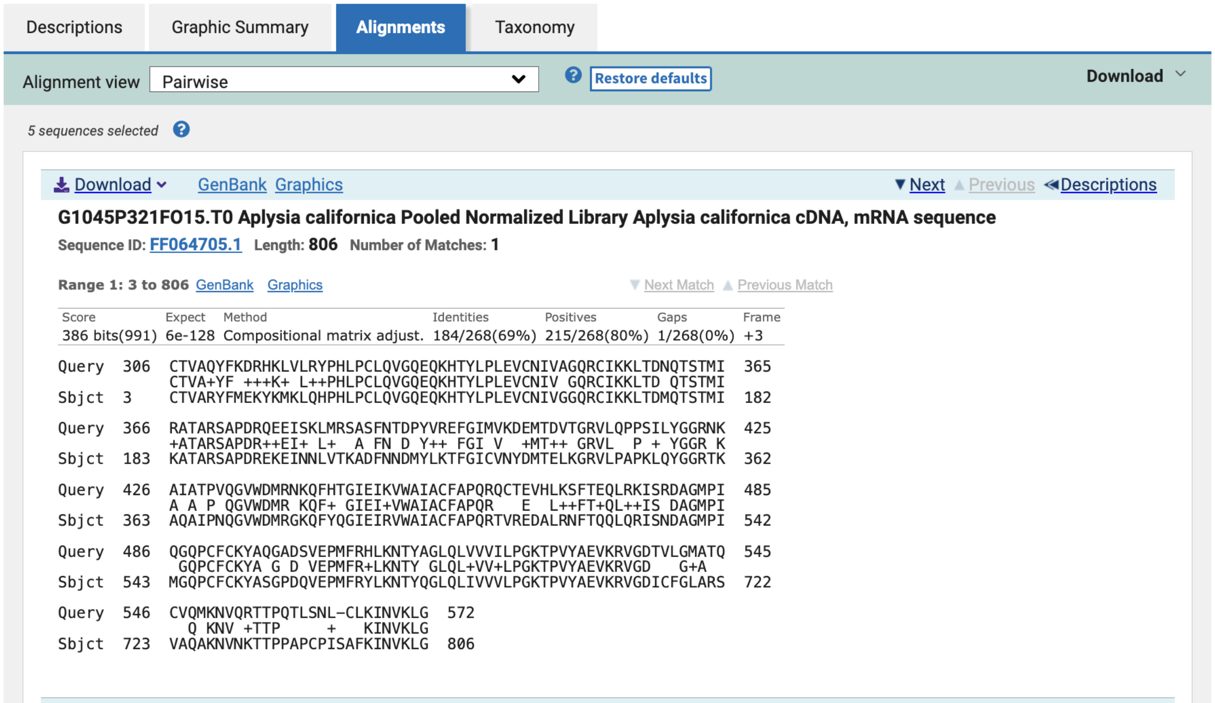
Percent Identity ___ to ___ E value ___ to ___ Query Coverage ___ to ___

Filter Reset

Sequences producing significant alignments

☑ select all *5 sequences selected* GenBank Graphics

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	G1045P321FO15.T0 Aplysia californica Pooled Normalized Library Aplysia californica cDNA, mRNA seque...	Aplysia californica	386	386	31%	6e-128	68.66%	806	FF064705.1
☑	CNSN01-F-080309-501 Normalized CNS library (juvenile 1) Aplysia californica cDNA clone CNSN01-F-08...	Aplysia californica	313	313	22%	1e-100	77.89%	653	EB315195.1
☑	CNSN01-C-004979-501 Normalized CNS library (juvenile 1) Aplysia californica cDNA clone CNSN01-C-00...	Aplysia californica	250	250	19%	1e-77	70.66%	511	EB253022.1
☑	CNSN01-F-033626-501 Normalized CNS library (juvenile 1) Aplysia californica cDNA clone CNSN01-F-03...	Aplysia californica	179	179	12%	2e-52	81.55%	318	EB279488.1
☑	CNSN01-F-023165-501 Normalized CNS library (juvenile 1) Aplysia californica cDNA clone CNSN01-F-02...	Aplysia californica	145	145	10%	1e-40	77.91%	260	EB291748.1



Alignment view Pairwise

5 sequences selected

G1045P321FO15.T0 Aplysia californica Pooled Normalized Library Aplysia californica cDNA, mRNA sequence

Sequence ID: FF064705.1 Length: **806** Number of Matches: **1**

Range 1: 3 to 806 GenBank Graphics

```
Score          Expect  Method                       Identities    Positives     Gaps       Frame
386 bits(991)  6e-128  Compositional matrix adjust. 184/268(69%)  215/268(80%)  1/268(0%)  +3

Query  306  CTVAQYFKDRHKLVLRYPHLPCLQVGQEQKHTYLPLEVCNIVAGQRCIKKLTDNQTSTMI  365
            CTVA+YF +++K+ L++PHLPCLQVGQEQKHTYLPLEVCNIV GQRCIKKLTD QTSTMI
Sbjct  3    CTVARYFMEKYKMKLQHPHLPCLQVGQEQKHTYLPLEVCNIVGGQRCIKKLTDMQTSTMI  182

Query  366  RATARSAPDRQEEISKLMRSASFNTDPYVREFGIMVKDEMTDVTGRVLQPPSILYGGRNK  425
            +ATARSAPDR++EI+ L+  A FN D Y++ FGI V  +MT++ GRVL  P + YGGR K
Sbjct  183  KATARSAPDREKEINNLVTKADFNNDMYLKTFGICVNYDMTELKGRVLPAPKLQYGGRTK  362

Query  426  AIATPVQGVWDMRNKQFHTGIEIKVWAIACFAPQRQCTEVHLKSFTEQLRKISRDAGMPI  485
            A A P QGVWDMR KQF+ GIEI+VWAIACFAPQR   E  L++FT+QL++IS DAGMPI
Sbjct  363  AQAIPNQGVWDMRGKQFYQGIEIRVWAIACFAPQRTVREDALRNFTQQLQRISNDAGMPI  542

Query  486  QGQPCFCKYAQGADSVEPMFRHLKNTYAGLQLVVVILPGKTPVYAEVKRVGDTVLGMATQ  545
             GQPCFCKYA G D VEPMFR+LKNTY GLQL+VV+LPGKTPVYAEVKRVGD   G+A
Sbjct  543  MGQPCFCKYASGPDQVEPMFRYLKNTYQGLQLIVVVLPGKTPVYAEVKRVGDICFGLARS  722

Query  546  CVQMKNVQRTTPQTLSNL-CLKINVKLG  572
              Q KNV +TTP     +   KINVKLG
Sbjct  723  VAQAKNVNKTTPPAPCPISAFKINVKLG  806
```

Alignment Details:

Score	Expect	Method	Identities	Positives	Gaps	Frame
386 bits(991)	6e-128	Compositional matrix adjust.	184/268(69%)	215/268(80%)	1/268(0%)	+3

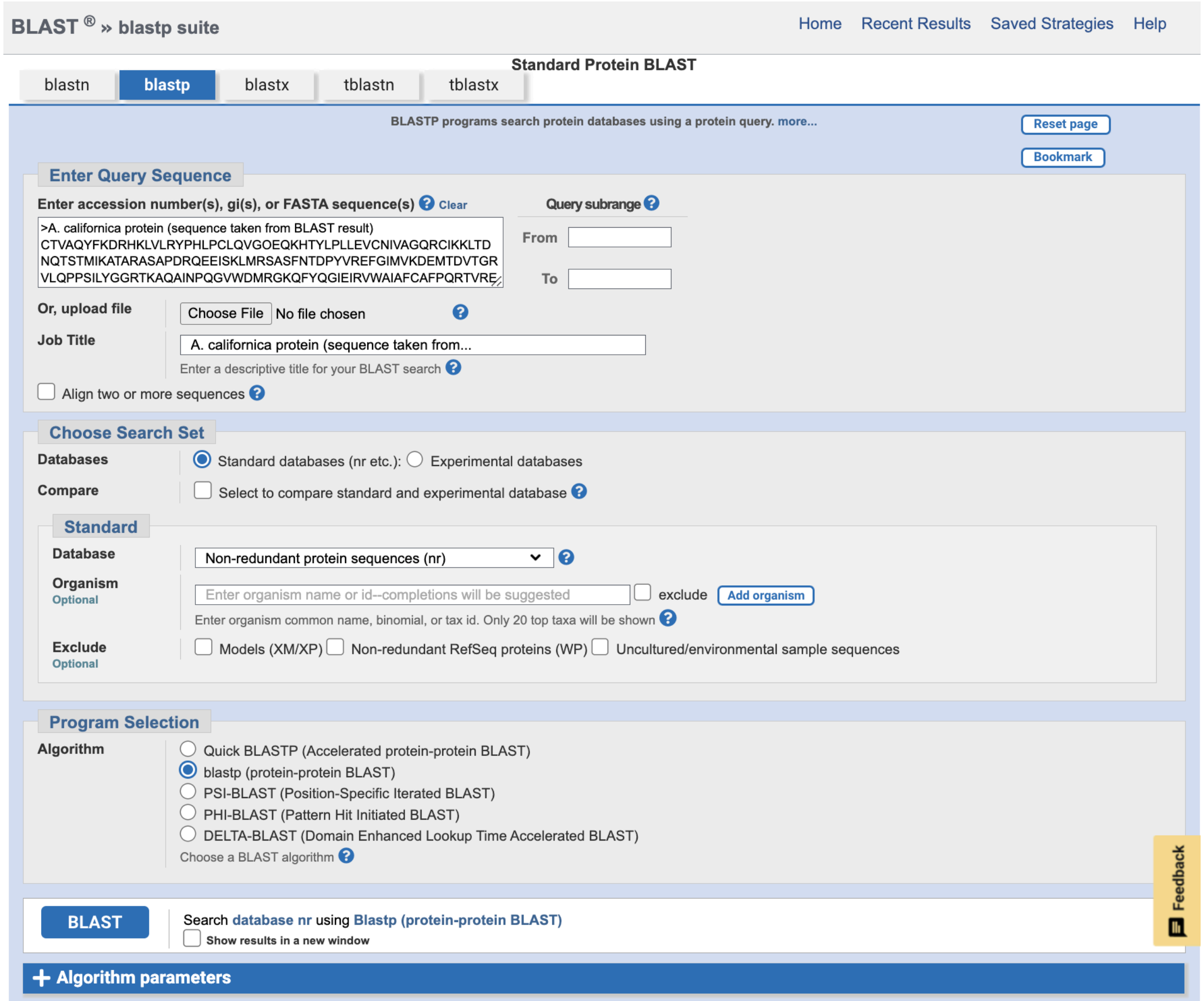
```
Query  306
CTVAQYFKDRHKLVLRYPHLPCLQVGQEQKHTYLPLEVCNIVAGQRCIKKLTDNQTSTMI  365
            CTVA+YF +++K+ L++PHLPCLQVGQEQKHTYLPLEVCNIV GQRCIKKLTD
QTSTMI
Sbjct  3
CTVARYFMEKYKMKLQHPHLPCLQVGQEQKHTYLPLEVCNIVGGQRCIKKLTDMQTSTMI  182

Query  366
RATARSAPDRQEEISKLMRSASFNTDPYVREFGIMVKDEMTDVTGRVLQPPSILYGGRNK  425
            +ATARSAPDR++EI+ L+  A FN D Y++ FGI V  +MT++ GRVL  P +
YGGR K
Sbjct  183
KATARSAPDREKEINNLVTKADFNNDMYLKTFGICVNYDMTELKGRVLPAPKLQYGGRTK  362

Query  426
AIATPVQGVWDMRNKQFHTGIEIKVWAIACFAPQRQCTEVHLKSFTEQLRKISRDAGMPI  485
            A A P QGVWDMR KQF+ GIEI+VWAIACFAPQR   E  L++FT+QL++IS
DAGMPI
Sbjct  363
AQAIPNQGVWDMRGKQFYQGIEIRVWAIACFAPQRTVREDALRNFTQQLQRISNDAGMPI  542

Query  486
QGQPCFCKYAQGADSVEPMFRHLKNTYAGLQLVVVILPGKTPVYAEVKRVGDTVLGMATQ  545
             GQPCFCKYA G D VEPMFR+LKNTY GLQL+VV+LPGKTPVYAEVKRVGD
G+A
Sbjct  543
MGQPCFCKYASGPDQVEPMFRYLKNTYQGLQLIVVVLPGKTPVYAEVKRVGDICFGLARS  722

Query  546  CVQMKNVQRTTPQTLSNL-CLKINVKLG  572
              Q KNV +TTP     +   KINVKLG

Sbjct  723  VAQAKNVNKTTPPAPCPISAFKINVKLG  806
```

[Q3] Gather information about this “novel” protein. At a minimum, show me the protein sequence of the “novel” protein as displayed in your BLAST results from [Q2] as FASTA format (you can copy and paste the aligned sequence subject lines from your BLAST result page if necessary) or translate your novel DNA sequence using a tool called EMBOSS Transeq at the EBI. Don’t forget to translate all six reading frames; the ORF (open reading frame) is likely to be the longest sequence without a stop codon. It may not start with a methionine if you don’t have the complete coding region. Make sure the sequence you provide includes a header/subject line and is in traditional FASTA format.

```
>A. californica protein (sequence taken from BLAST result)
CTVAQYFKDRHKLVLRYPHLPCLQVGOEQKHTYLPLLEVCNIVAGQRCIKKLTDNQTST
MIKATARASAPDRQEEISKLMRSASFNTDPYVREFGIMVKDEMTDVTGRVLQPPSILYGG
RTKAQAINPQGVWDMRGKQFYQGIEIRVWAIAFCAFPQRTVREDALRNFTQQLQRISND
AGMPIMGQPCFCKYASGPDQVEPFMFRKLKNTYQGLQILVVVLPGKTYVAEVKRVGDIC
FGLARS VQAQAKNVNKTTTPPAPCPISAFKINVKLG
```

Name: Aplysia Argonaute-like protein
Species: Alpysia californica
Taxonomy: Eukaryota; Metazoa; Mollusca; Gastropoda; Heterobranchia; Aplysiida; Aplysiidae; Aplysia.

[Q4] Prove that this gene, and its corresponding protein, are novel. For the purposes of this project, “novel” is defined as follows. Take the protein sequence (your answer to [Q3]), and use it as a query in a blastp search of the nr database at NCBI.



Standard Protein BLAST

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

```
>A. californica protein (sequence taken from BLAST result)
CTVAQYFKDRHKLVLRYPHLPCLQVGOEQKHTYLPLLEVCNIVAGQRCIKKLTD
NQTSTMIKATARASAPDRQEEISKLMRSASFNTDPYVREFGIMVKDEMTDVTGR
VLQPPSILYGGRTKAQAINPQGVWDMRGKQFYQGIEIRVWAIAFCAFPQRTVRE
```

Query subrange

From

To

Or, upload file: Choose File No file chosen

Job Title: A. californica protein (sequence taken from...

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Databases: ● Standard databases (nr etc.): ○ Experimental databases

Compare: □ Select to compare standard and experimental database

Standard

Database: Non-redundant protein sequences (nr)

Organism Optional: Enter organism name or id--completions will be suggested □ exclude Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude Optional: □ Models (XM/XP) □ Non-redundant RefSeq proteins (WP) □ Uncultured/environmental sample sequences

Program Selection

Algorithm:

○ Quick BLASTP (Accelerated protein-protein BLAST)
● blastp (protein-protein BLAST)
○ PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

BLAST Search database nr using Blastp (protein-protein BLAST)

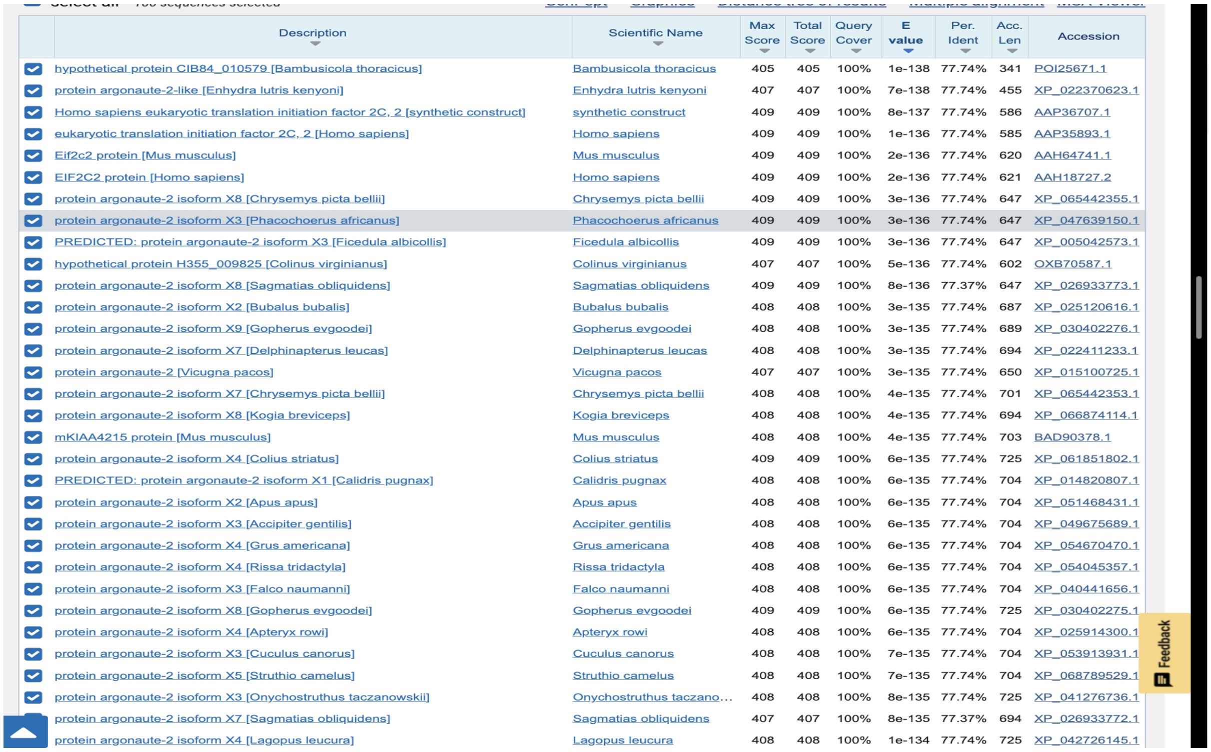
□ Show results in a new window

+ Algorithm parameters

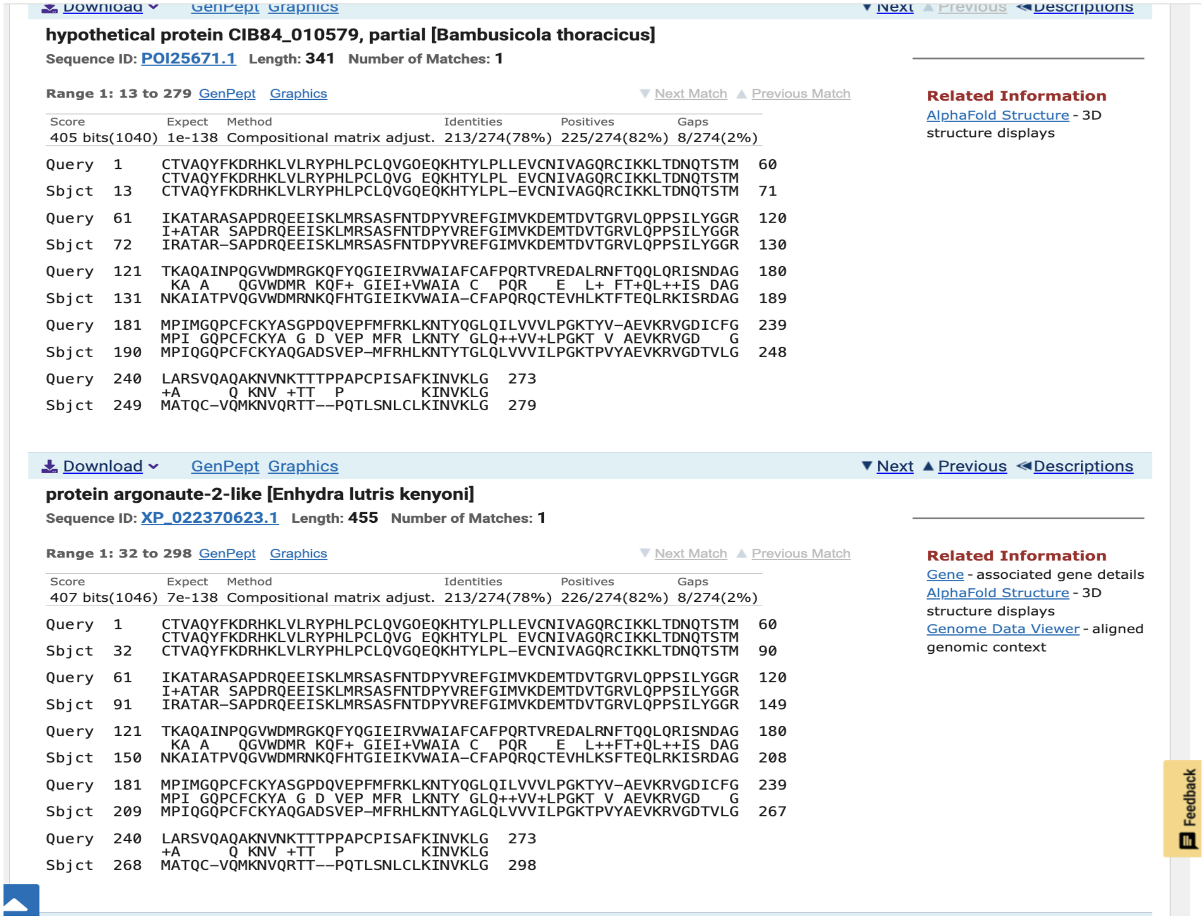
The top matches are Hypothetical protein (Bambusicola thoracicus) – 77.74% identity
Argonaute-2-like (Enhydra lutris kenyoni) – 77.74% identity
EIF2C2 (Homo sapiens, Mus musculus, multiple species) – 77.74% identity
Multiple Argonaute homologs in vertebrates (Chrysemys picta bellii, Phacochoerus africanus, Gophers, Kogia, Apus, etc.) – 77.74% identity.
Query Coverage: 100% for top matches
E-value: Best hits in the range of 6e-138 to 6e-135 (highly significant)

Since no 100% identical protein exists, the Aplysia californica protein is likely novel in this species. However, it shares homology with Argonaute-2 proteins in mammals, birds, reptiles, and other vertebrates, confirming it belongs to the Argonaute protein family.

See below for more details:



	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	hypothetical protein CIB84_010579 [Bambusicola thoracicus]	Bambusicola thoracicus	405	405	100%	1e-138	77.74%	341	POI25671.1
☑	protein argonaute-2-like [Enhydra lutris kenyoni]	Enhydra lutris kenyoni	407	407	100%	7e-138	77.74%	455	XP_022370623.1
☑	Homo sapiens eukaryotic translation initiation factor 2C, 2 [synthetic construct]	synthetic construct	409	409	100%	8e-137	77.74%	586	AAP36707.1
☑	eukaryotic translation initiation factor 2C, 2 [Homo sapiens]	Homo sapiens	409	409	100%	1e-136	77.74%	585	AAP35893.1
☑	Eif2c2 protein [Mus musculus]	Mus musculus	409	409	100%	2e-136	77.74%	620	AAH64741.1
☑	EIF2C2 protein [Homo sapiens]	Homo sapiens	409	409	100%	2e-136	77.74%	621	AAH18727.2
☑	protein argonaute-2 isoform X8 [Chrysemys picta bellii]	Chrysemys picta bellii	409	409	100%	3e-136	77.74%	647	XP_065442355.1
☑	protein argonaute-2 isoform X3 [Phacochoerus africanus]	Phacochoerus africanus	409	409	100%	3e-136	77.74%	647	XP_047639150.1
☑	PREDICTED: protein argonaute-2 isoform X3 [Ficedula albicollis]	Ficedula albicollis	409	409	100%	3e-136	77.74%	647	XP_005042573.1
☑	hypothetical protein H355_009825 [Colinus virginianus]	Colinus virginianus	407	407	100%	5e-136	77.74%	602	OXB70587.1
☑	protein argonaute-2 isoform X8 [Sagmatias obliquidens]	Sagmatias obliquidens	409	409	100%	8e-136	77.37%	647	XP_026933773.1
☑	protein argonaute-2 isoform X2 [Bubalus bubalis]	Bubalus bubalis	408	408	100%	3e-135	77.74%	687	XP_025120616.1
☑	protein argonaute-2 isoform X9 [Gopherus evgoodei]	Gopherus evgoodei	408	408	100%	3e-135	77.74%	689	XP_030402276.1
☑	protein argonaute-2 isoform X7 [Delphinapterus leucas]	Delphinapterus leucas	408	408	100%	3e-135	77.74%	694	XP_022411233.1
☑	protein argonaute-2 [Vicugna pacos]	Vicugna pacos	407	407	100%	3e-135	77.74%	650	XP_015100725.1
☑	protein argonaute-2 isoform X7 [Chrysemys picta bellii]	Chrysemys picta bellii	408	408	100%	4e-135	77.74%	701	XP_065442353.1
☑	protein argonaute-2 isoform X8 [Kogia breviceps]	Kogia breviceps	408	408	100%	4e-135	77.74%	694	XP_066874114.1
☑	mKIAA4215 protein [Mus musculus]	Mus musculus	408	408	100%	4e-135	77.74%	703	BAD90378.1
☑	protein argonaute-2 isoform X4 [Colius striatus]	Colius striatus	409	409	100%	6e-135	77.74%	725	XP_061851802.1
☑	PREDICTED: protein argonaute-2 isoform X1 [Calidris pugnax]	Calidris pugnax	408	408	100%	6e-135	77.74%	704	XP_014820807.1
☑	protein argonaute-2 isoform X2 [Apus apus]	Apus apus	408	408	100%	6e-135	77.74%	704	XP_051468431.1
☑	protein argonaute-2 isoform X3 [Accipiter gentilis]	Accipiter gentilis	408	408	100%	6e-135	77.74%	704	XP_049675689.1
☑	protein argonaute-2 isoform X4 [Grus americana]	Grus americana	408	408	100%	6e-135	77.74%	704	XP_054670470.1
☑	protein argonaute-2 isoform X4 [Rissa tridactyla]	Rissa tridactyla	408	408	100%	6e-135	77.74%	704	XP_054045357.1
☑	protein argonaute-2 isoform X3 [Falco naumanni]	Falco naumanni	408	408	100%	6e-135	77.74%	704	XP_040441656.1
☑	protein argonaute-2 isoform X8 [Gopherus evgoodei]	Gopherus evgoodei	409	409	100%	6e-135	77.74%	725	XP_030402275.1
☑	protein argonaute-2 isoform X4 [Apteryx rowi]	Apteryx rowi	408	408	100%	6e-135	77.74%	704	XP_025914300.1
☑	protein argonaute-2 isoform X3 [Cuculus canorus]	Cuculus canorus	408	408	100%	7e-135	77.74%	704	XP_053913931.1
☑	protein argonaute-2 isoform X5 [Struthio camelus]	Struthio camelus	408	408	100%	7e-135	77.74%	704	XP_068789529.1
☑	protein argonaute-2 isoform X3 [Onychostruthus taczanowskii]	Onychostruthus taczano...	408	408	100%	8e-135	77.74%	725	XP_041276736.1
☑	protein argonaute-2 isoform X7 [Sagmatias obliquidens]	Sagmatias obliquidens	407	407	100%	8e-135	77.37%	694	XP_026933772.1
	protein argonaute-2 isoform X4 [Lagopus leucura]	Lagopus leucura	408	408	100%	1e-134	77.74%	725	XP_042726145.1



Download GenPept Graphics ▼ Next ▲ Previous ◀Descriptions

hypothetical protein CIB84_010579, partial [Bambusicola thoracicus]

Sequence ID: POI25671.1 Length: 341 Number of Matches: 1

Range 1: 13 to 279 GenPept Graphics ▼ Next Match ▲ Previous Match

```
Score            Expect  Method                          Identities     Positives      Gaps
405 bits(1040)  1e-138  Compositional matrix adjust.  213/274(78%)  225/274(82%)  8/274(2%)

Query  1    CTVAQYFKDRHKLVLRYPHLPCLQVGOEQKHTYLPLLEVCNIVAGQRCIKKLTDNQTSTM  60
            CTVAQYFKDRHKLVLRYPHLPCLQVG EQKHTYLPL EVCNIVAGQRCIKKLTDNQTSTM
Sbjct  13   CTVAQYFKDRHKLVLRYPHLPCLQVGQEQKHTYLPL-EVCNIVAGQRCIKKLTDNQTSTM  71

Query  61   IKATARASAPDRQEEISKLMRSASFNTDPYVREFGIMVKDEMTDVTGRVLQPPSILYGGR  120
            I+ATAR SAPDRQEEISKLMRSASFNTDPYVREFGIMVKDEMTDVTGRVLQPPSILYGGR
Sbjct  72   IRATAR-SAPDRQEEISKLMRSASFNTDPYVREFGIMVKDEMTDVTGRVLQPPSILYGGR  130

Query  121  TKAQAINPQGVWDMRGKQFYQGIEIRVWAIAFCAFPQRTVREDALRNFTQQLQRISNDAG  180
             KA A   QGVWDMR KQF+ GIEI+VWAIA C  PQR   E  L+ FT+QL++IS DAG
Sbjct  131  NKAIATPVQGVWDMRNKQFHTGIEIKVWAIA-CFAPQRQCTEVHLKTFTEQLRKISRDAG  189

Query  181  MPIMGQPCFCKYASGPDQVEPFMFRKLKNTYQGLQILVVVLPGKTYV-AEVKRVGDICFG  239
            MPI GQPCFCKYA G D VEP MFR LKNTY GLQ++VV+LPGKT V AEVKRVGD   G
Sbjct  190  MPIQGQPCFCKYAQGADSVEP-MFRHLKNTYTGLQLVVVILPGKTPVYAEVKRVGDTVLG  248

Query  240  LARSVQAQAKNVNKTTTPPAPCPISAFKINVKLG  273
            +A     Q KNV +TT  P        KINVKLG
Sbjct  249  MATQC-VQMKNVQRTT--PQTLSNLCLKINVKLG  279
```

Related Information

AlphaFold Structure - 3D structure displays

Download GenPept Graphics ▼ Next ▲ Previous ◀Descriptions

protein argonaute-2-like [Enhydra lutris kenyoni]

Sequence ID: XP_022370623.1 Length: 455 Number of Matches: 1

Range 1: 32 to 298 GenPept Graphics ▼ Next Match ▲ Previous Match

```
Score            Expect  Method                          Identities     Positives      Gaps
407 bits(1046)  7e-138  Compositional matrix adjust.  213/274(78%)  226/274(82%)  8/274(2%)

Query  1    CTVAQYFKDRHKLVLRYPHLPCLQVGOEQKHTYLPLLEVCNIVAGQRCIKKLTDNQTSTM  60
            CTVAQYFKDRHKLVLRYPHLPCLQVG EQKHTYLPL EVCNIVAGQRCIKKLTDNQTSTM
Sbjct  32   CTVAQYFKDRHKLVLRYPHLPCLQVGQEQKHTYLPL-EVCNIVAGQRCIKKLTDNQTSTM  90

Query  61   IKATARASAPDRQEEISKLMRSASFNTDPYVREFGIMVKDEMTDVTGRVLQPPSILYGGR  120
            I+ATAR SAPDRQEEISKLMRSASFNTDPYVREFGIMVKDEMTDVTGRVLQPPSILYGGR
Sbjct  91   IRATAR-SAPDRQEEISKLMRSASFNTDPYVREFGIMVKDEMTDVTGRVLQPPSILYGGR  149

Query  121  TKAQAINPQGVWDMRGKQFYQGIEIRVWAIAFCAFPQRTVREDALRNFTQQLQRISNDAG  180
             KA A   QGVWDMR KQF+ GIEI+VWAIA C  PQR   E  L++FT+QL++IS DAG
Sbjct  150  NKAIATPVQGVWDMRNKQFHTGIEIKVWAIA-CFAPQRQCTEVHLKSFTEQLRKISRDAG  208

Query  181  MPIMGQPCFCKYASGPDQVEPFMFRKLKNTYQGLQILVVVLPGKTYV-AEVKRVGDICFG  239
            MPI GQPCFCKYA G D VEP MFR LKNTY GLQ++VV+LPGKT V AEVKRVGD   G
Sbjct  209  MPIQGQPCFCKYAQGADSVEP-MFRHLKNTYAGLQLVVVILPGKTPVYAEVKRVGDTVLG  267

Query  240  LARSVQAQAKNVNKTTTPPAPCPISAFKINVKLG  273
            +A     Q KNV +TT  P        KINVKLG
Sbjct  268  MATQC-VQMKNVQRTT--PQTLSNLCLKINVKLG  298
```

Related Information

Gene - associated gene details

AlphaFold Structure - 3D structure displays

Genome Data Viewer - aligned genomic context

[Q5] Generate a multiple sequence alignment with your novel protein, your original query protein, and a group of other members of this family from different species. A typical number of proteins to use in a multiple sequence alignment for this assignment purpose is a minimum of 5 and a maximum of 20 - although the exact number is up to you. Include the multiple sequence alignment in your report. Use Courier font with a size appropriate to fit page width. Side-note: Indicate your sequence in the alignment by choosing an appropriate name for each sequence in the input unaligned sequence file (i.e. edit the sequence file so that the species, or short common, names (rather than accession numbers) display in the output alignment and in the subsequent answers below). The goal in this step is to create an interesting an alignment for building a phylogenetic tree that illustrates species divergence

Relabeled sequences for alignment:

```
>Human_AGO2 gi|4F3T_A|ref|Chain_A| Protein argonaute-2 [Homo sapiens]
MYSGAGPALAPPAPPPPIQGYAFKPPPRPDFGTSGRTIKLQANFFEMDIPKIDIYHYELDIKPEKCPRRVNREIVEHMV
QHFKTQIFGDRKPVFDGRKNLYTAMPLPIGRDKVELEVTLPGEGKDRIFKVSIKWVSCVSLQALHDALSGRLPSVPFET
IQALDVVMRHLPSMRYTPVGRSFFTASEGCSNPLGGGREVWFGFHQSVRPSLWKMMLNIDVSATAFYKAQPVIEFVCEV
LDFKSIEEQQKPLTDSQRVKFTKEIKGLKVEITHCGQMKRKYRVCNVTRRPASHQTFPLQQESGQTVECTVAQYFKDRH
KLVLRYPHLPCLQVGQEQKHTYLPLEVCNIVAGQRCIKKLTDNQTSTMIRATARSAPDRQEEISKLMRSASFNTDPYVR
EFGIMVKDEMTDVTGRVLQPPSILYGGRNKAIATPVQGVWDMRNKQFHTGIEIKVWAIACFAPQRQCTEVHLKSFTEQL
RKISRDAGMPIQGQPCFCKYAQGADSVEPMFRHLKNTYAGLQLVVVILPGKTPVYAEVKRVGDTVLGMATQCVQMKNVQ
RTTPQTLSNLCLKINVKLGGVNNILLPQGRPPVFQQPVIFLGADVTHPPAGDGKKPSIAAVVGSMDAHPNRYCATVRVQ
QHRQEIIQDLAAMVRELLIQFYKSTRFKPTRIIFYRDGVSEGQFQQVLHHELLAIREACIKLEKDYQPGITFIVVQKRH
HTRLFCTDKNERVGKSGNIPAGTTVDTKITHPTEFDFYLCSHAGIQGTSRPSHYHVLWDDNRFSSDELQILTYQLCHTY
VRCTRSVSIPAPAYYAHLVAFRARYHLVDKEHDSAEGSHTSGQSNGRDHQALAKAVQVHQDTLRTMYA

>A. californica protein (sequence taken from BLAST result)
CTVAQYFKDRHKLVLRYPHLPCLQVGOEQKHTYLPLLEVCNIVAGQRCIKKLTDNQTST
MIKATARASAPDRQEEISKLMRSASFNTDPYVREFGIMVKDEMTDVTGRVLQPPSILYGG
RTKAQAINPQGVWDMRGKQFYQGIEIRVWAIAFCAFPQRTVREDALRNFTQQLQRISND
AGMPIMGQPCFCKYASGPDQVEPFMFRKLKNTYQGLQILVVVLPGKTYVAEVKRVGDIC
FGLARSVQAQAKNVNKTTTPPAPCPISAFKINVKLG


>Dolphin_AGO2 gi|XP_059851221.1|ref|Protein argonaute-2 isoform X3 [Delphinus
delphis]
MYSGAGPEKENRSVRGKHGNQRVLAPPPPPPPPVQGYAFKPPPRPDFGTSGRTIKLQANFFEMDIPKIDIYHYELDIKP
EKCPRRVNREIVEHMVQHFKTQIFGDRKPVFDGRKNLYTAMPLPIGRDKVELEVTLPGEGKDRIFKVSIKWVSCVSLQA
LHDALSGRLPSVPFETIQALDVVMRHLPSMRYTPVGRSFFTASEGCSNPLGGGREVWFGFHQSVRPSLWKMMLNIDVSA
TAFYKAQPVIEFVCEVLDFKSIEEQQKPLTDSQRVKFTKEIKGLKVEITHCGQMKRKYRVCNVTRRPASHQTFPLQQES
GQTVECTVAQYFKDRHKLVLRYPHLPCLQVGQEQKHTYLPLEVCNIVAGQRCIKKLTDNQTSTMIRATARSAPDRQEEI
SKLMRSASFNTDPYVREFGIMVRDEMTDVTGRVLQPPSILYGGRNKAIATPVQGVWDMRNKQFHTGIEIKVWAIACFAP
QRQCTEVHLKSFTEQLRKISRDAGMPIQGQPCFCKYAQGADSVEPMFRHLKNTYAGLQLVVVILPGKTPVYAEVKRVGD
TVLGMATQCVQMKNVQRTTPQTLSNLCLKINVKLGGVNNILLPQGRPPVFQQPVIFLGADVTHPPAGDGKKPSIAAVVG
SMDAHPNRYCATVRVQQHRQEIIQDLAAMVRELLIQFYKSTRFKPTRIIFYRDGVSEGQFQQVLHHELLAIREACIKLE
```

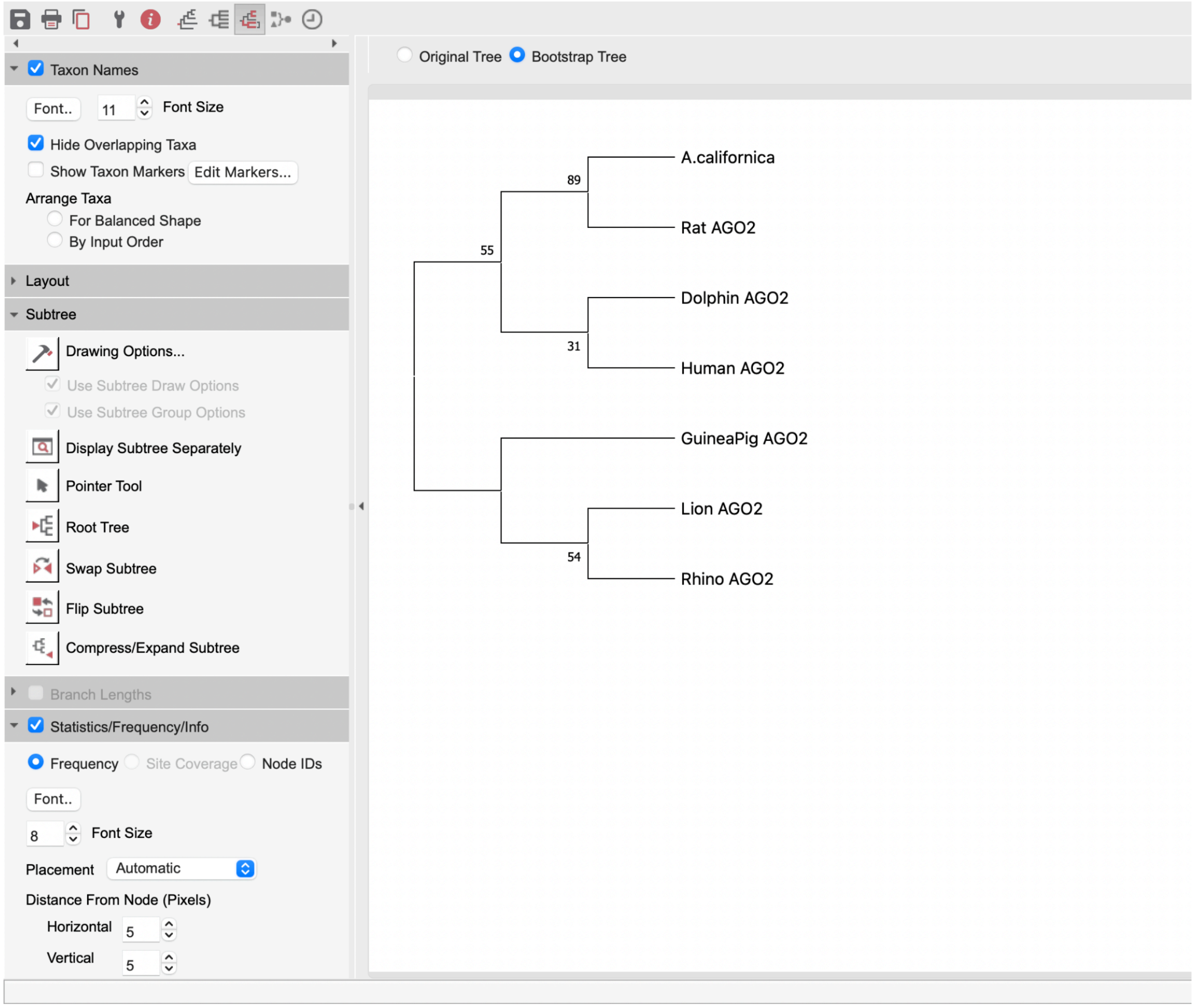
```
KDYQPGITFIVVQKRHHTRLFCTDKNERVGKSGNIPAGTTVDTKITHPTEFDFYLCSHAGIQGTSRPSHYHVLWDDNRF
SSDELQILTYQLCHTYVRCTRSVSIPAPAYYAHLVAFRARYHLVDKEHDSAEGSHTSGQSNGRDHQALAKAVQVHQTLR
TMYFA

>Lion_AGO2 gi|XP_042780216.1|ref|Protein argonaute-2 isoform X1 [Panthera leo]
PVLAPPAPPPPPIQGYAFKPPPRPDFGTSGRTIKLQANFFEMDIPKIDIYHYELDIKPEKCPRRVNEIVEHMVQHFKTQ
IFGDRKPVFDGRKNLYTAMPLPIGRDKVELEVTLPGEGKDRIFKVSIKWVSCVSLQALHDALSGRLPSVPFETIQALDV
VMRHLPSMRYTPVGRSFFTASEGCSNPLGGGREVWFGFHQSVRPSLWKMMLNIDVSATAFYKAQPVIEFVCEVLDFKSI
EEQQKPLTDSQRVKFTKEIKGLKVEITHCGQMKRKYRVCNVTRRPASHQTFPLQQESGQTVECTVAQYFKDRHKLVLRY
PHLPCLQVGQEQKHTYLPLEVCNIVAGQRCIKKLTDNQTSTMIRATARSAPDRQEEISKLMRSASFNTDPYVREFGIMV
KDEMTDVTGRVLQPPSILYGGRNKAIATPVQGVWDMRNKQFHTGIEIKVWAIACFAPQRQCTEVHLKSFTEQLRKISRD
AGMPIQGQPCFCKYAQGADSVEPMFRHLKNTYAGLQLVVVILPGKTPVYAEVKRVGDTVLGMATQCVQMKNVQRTTPQT
LSNLCLKINVKLGGVNNILLPQGRPPVFQQPVIFLGADVTHPPAGDGKKPSIAAVVGSMDAHPNRYCATVRVQQHRQEI
IQDLATMVRELLIQFYKSTRFKPTRIIFYRDGVSEGQFQQVLHHELLAIREACIKLEKDYQPGITFIVVQKRHHTRLFC
TDKNERVGKSGNIPAGTTVDTKITHPTEFDFYLCSHAGIQGTSRPSHYHVLWDDNRFSSDELQILTYQLCHTYVRCTRS
VSIPAPAYYAHLVAFRARYHLVDKEHDSAEGSHTSGQSNGRDHQALAKAVQVHQDTLRTMYFA

>Rat_AGO2 gi|AAF12800.1|ref|GERp95 [Rattus norvegicus]
MYSGAGPVLASPAPTTSPIPGYAFKPPPRPDFGTTGRTIKLQANFFEMDIPKIDIYHYELDIKPEKCPRRVNREIVEHM
VQHFKTQIFGDRKPVFDGRKNLYTAMPLPIGRDKVELEVTLPGEGKDRIFKVSIKWVSCVSLQALHDALSGRLPSVPFE
TIQALDVVMRHLPSMRYTPVGRSFFTASEGCSNPLGGGREVWFGFHQSVRPSLWKMMLNIDVSATAFYKAQPVIEFVCE
VLDFKSIEEQQKPLTDSQRVKFTKEIKGLKVEITHCGQMKRKYRVCNVTRRPASHQTFPLQQESGQTVECTVAQYFKDR
HKLVLRYPHLPCLQVGQEQKHTYLPLEVCNIVAGQRCIKKLTDNQTSTMIRATARSAPDRQEEISKLMRSASFNTDPYV
REFGIMVKDEMTDVTGRVLQPPSILYGGRNKAIATPVQGVWDMRNKQFHTGIEIKVWAIACFAPQRQCTEVHLKSFTEQ
LRKISRDAGMPIQGQPCFCKYAQGADSVEPMFRHLKNTYAGLQLVVVILPGKTPVYAEVKRVGDTVLGMATQCVQMKNV
QRTTPQTLSNLCLKINVKLGGVNNILLPQGRPPVFQQPVIFLGADVTHPPAGDGKKPSIAAVVGSMDAHPNRYCATVRV
QQHRQEIIQDLAAMVRELLIQFYKSTRFKPTRIIFYRDGVSEGQFQQVLHHELLAIREACIKLEKEYQPGITFIVVQKR
HHTRLFCTDKNERVGKSGNIPAGTTVDTKITHPTEFDFYLCSHAGIQGTSRPSHYHVLWDDNRFSSDELQILTYQLCHT
YVRCTRSVSIPAPAYYAHLVAFRARYHLVDKEHDSAEGSHTSGQSNGRDHQALAKAVQVHQDTLRTMYFA

>GuineaPig_AGO2 gi|XP_063087086.1|ref|Protein argonaute-2 isoform X1 [Cavia
porcellus]
PMLAPPAPPPPPIQGYAFKPPPRPDFGTSGRTIKLQANFFEMDIPKIDIYHYELDIKPEKCPRRVNREIVEHMVQHFKT
QIFGDRKPVFDGRKNLYTAMPLPIGRDKVELEVTLPGEGKDRIFKVSIKWVSCVSLQALHDALSGRLPSVPFETIQALD
VVMRHLPSMRYTPVGRSFFTASEGCSNPLGGGREVWFGFHQSVRPSLWKMMLNIDVSATAFYKAQPVIEFVCEVLDFKS
IEEQQKPLTDSQRVKFTKEIKGLKVEITHCGQMKRKYRVCNVTRRPASHQTFPLQQESGQTVECTVAQYKDRHKLVLRY
PHLPCLQVGQEQKHTYLPLEVCNIVAGQRCIKKLTDNQTSTMIRATARSAPDRQEEISKLMRSASFNTDPYVREFGIMV
KDEMTDVTGRVLQPPSILYGGRNKAIATPVQGVWDMRNKQFHTGIEIKVWAIACFAPQRQCTEVHLKSFTEQLRKISRD
AGMPIQGQPCFCKYAQGADSVEPMFRHLKNTYAGLQLVVVILPGKTPVYAEVKRVGDTVLGMATQCVQMKNVQRTTPQT
LSNLCLKINVKLGGVNNILLPQGRPPVFQQPVIFLGADVTHPPAGDGKKPSIAAVVGSMDAHPNRYCATVRVQQHRQEI
IQDLAAMVRELLIQFYKSTRFKPTRIIFYRDGVSEGQFQQVLHHELLAIREACIKLEKDYQPGITFIVVQKRHHTRLFC
TDKNERVGKSGNIPAGTTVDTKITHPTEFDFYLCSHAGIQGTSRPSHYHVLWDDNRFSSDELQILTYQLCHTYVRCTRS
VSIPAPAYYAHLVAFRARYHLVDKEHDSAEGSHTSGQSNGRDHQALAKAVQVHQDTLRTMYFA

>Rhino_AGO2 gi|XP_004431098.1|ref|PREDICTED: protein argonaute-2 [Ceratotherium
simum simum]
MFSLLLAVLAPPAPPPPPIQGYAFKPPPRPDFGTSGRTIKLQANFFEMDIPKIDIYHYELDIKPEKCPRRVNREIVEHM
VQHFKTQIFGDRKPVFDGRKNLYTAMPLPIGRDKVELEVTLPGEGKDRIFKVSIKWVSCVSLQALHDALSGRLPSVPFE
TIQALDVVMRHLPSMRYTPVGRSFFTASEGCSNPLGGGREVWFGFHQSVRPSLWKMMLNIDVSATAFYKAQPVIEFVCE
```

```
VLDFKSIEEQQKPLTDSQRVKFTKEIKGLKVEITHCGQMKRKYRVCNVTRRPASHQTFPLQQESGQTVECTVAQYFKDR
HKLVLRYPHLPCLQVGQEQKHTYLPLEVCNIVAGQRCIKKLTDNQTSTMIRATARSAPDRQEEISKLMRSASFNTDPYV
REFGIMVKDEMTDVTGRVLQPPSILYGGRNKAIATPVQGVWDMRNKQFHTGIEIKVWAIACFAPQRQCTEVHLKSFTEQ
LRKISRDAGMPIQGQPCFCKYAQGADSVEPMFRHLKNTYAGLQLVVVILPGKTPVYAEVKRVGDTVLGMATQCVQMKNV
QRTTPQTLSNLCLKINVKLGGVNNILLPQGRPPVFQQPVIFLGADVTHPPAGDGKKPSIAAVVGSMDAHPNRYCATVRV
QQHRQEIIQDLAAMVRELLIQFYKSTRFKPTRIIFYRDGVSEGQFQQVLHHELLAIREACIKLEKDYQPGITFIVVQKR
HHTRLFCTDKNERVGKSGNIPAGTTVDTKITHPTEFDFYLCSHAGIQGTSRPSHYHVLWDDNRFSSDELQILTYQLCHT
YVRCTRSVSIPAPAYYAHLVAFRARYHLVDKEHDSAEGSHTSGQSNGRDHQALAKAVQVHQDTLRTMYFA
```

Alignment:

```
CLUSTAL multiple sequence alignment by MUSCLE (3.8)

Human_AGO2          MYSGAGPALAPPAPPPPIQGYAFKPPPRPDFGTSGRTIKLQANFFEMDIPKIDIYHYELDIKPE
Dolphin_AGO2        MYSGAGPEKENRSVRGKHGNQRVLAPPPPPPPPVQGYAFKPPPRPDFGTSGRTIKLQANFFEMD
Lion_AGO2           PVLAPPAPPPPPIQGYAFKPPPRPDFGTSGRTIKLQANFFEMDIPKIDIYHYELDIKPEKCPRR
Rat_AGO2            MYSGAGPVLASPAPTTSPIPGYAFKPPPRPDFGTTGRTIKLQANFFEMDIPKIDIYHYELDIKP
GuineaPig_AGO2      PMLAPPAPPPPPIQGYAFKPPPRPDFGTSGRTIKLQANFFEMDIPKIDIYHYELDIKPEKCPRR
Rhino_AGO2          MFSLLLAVLAPPAPPPPPIQGYAFKPPPRPDFGTSGRTIKLQANFFEMDIPKIDIYHYELDIKP
A.californica       CTVAQYFKDRHKLVLRYPHLPCLQVGOEQKHTYLPLLEVCNIVAGQRCIKKLTDNQTSTMIRAT
                    *  * *   *  *     *  *    ** ** **     ** **     ** ** ** **  **


Human_AGO2          VAKHGKVKLGAHFSDGLAHLNDLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
Dolphin_AGO2        VAKHGTVVMGGLDRAIQNMDDIKNAYRELSVMHSEKLHVDDPNFRLLSEHTLCMAAKFG
Lion_AGO2           AKHGTVVMGGLDRAIQNMDDIKNAYRQLSVMHSEKLHVDDPNFRLLAEHTLCMAAKFG
Rat_AGO2            VAKHGTVVMGGLERAIKMNNDVKNTYAALSVMHSEKLHVDDPNFRLLADCTIVCAA MKFG
GuineaPig_AGO2      VAKHGKTVMHGLDRAVQNLD DIKNTYTALSVMHSEKLHVDDPNFRLLADCTIVCAAKLG
Rhino_AGO2          VAKHGKTVMHGLDRAVQNLD DIKNTYTALSVMHSEKLHVDDPNFRLLADCTIVCAAKLG
A.californica       CTVAQYFKDRHKLVLRYPHLPCLQVGOEQKHTYLPLLEVCNIVAGQRCIKKLTDNQTSTMIRAT
                    *  ** *: .:. . : ::*::*:.  *  ** :*:.******.:** **.: .   *::*


Human_AGO2          KE-FTPPVQAAYQKVVAGVANALAHKYH
Dolphin_AGO2        PSVFTPEVHETWQKFLNVVVAA LKGQYH
Lion_AGO2           PTEFTADVQEAWQKFLMATVSALGRQYH
Rat_AGO2            PTEFTADVQEAWQKFLMATVSALGRQYH
GuineaPig_AGO2      QAGFNADVQEAWQKFLAVVVSA LCRQYH
Rhino_AGO2          PAVFSADTQEA FQKFLAVVVSA LGRQYH
A.californica       CTVAQYFKDRHKLVLRYPHLPCLQVGOEQKHTYLPLLEVCNIVAGQRCIKKLTDNQTSTMIRAT
                    *..    : ::**:.   * . ** .:**
```

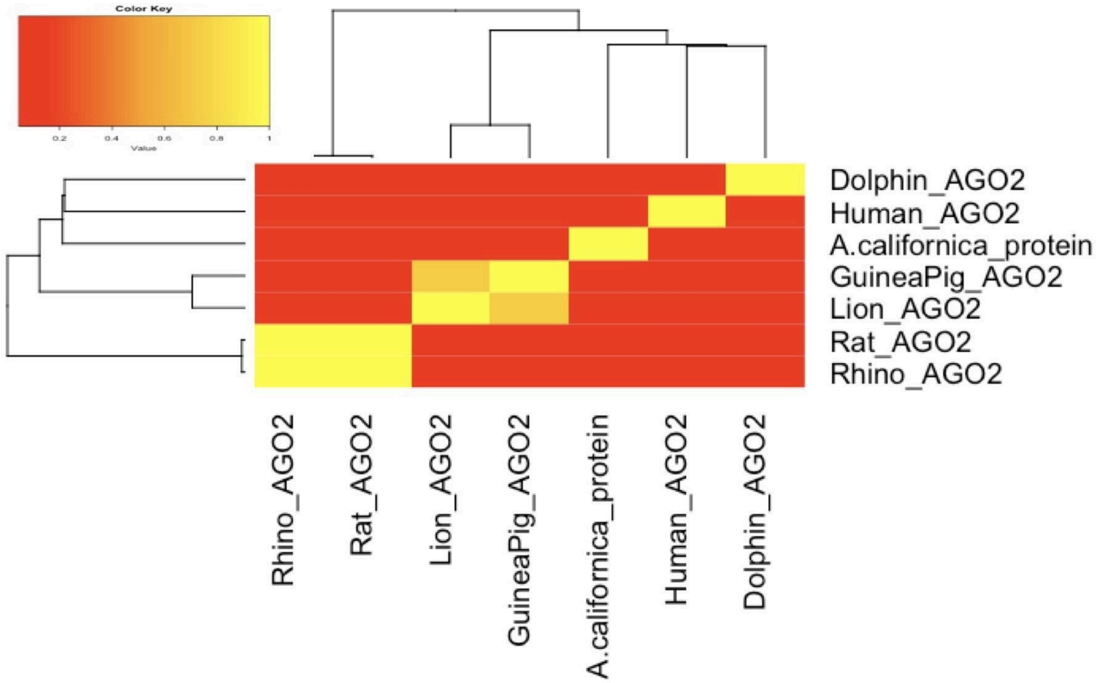
[Q6] Create a phylogenetic tree, using either a parsimony or distance-based approach. Bootstrapping and tree rooting are optional. Use “simple phylogeny” online from the EBI or any respected phylogeny program (such as MEGA, PAUP, or Phylip). Paste an image of your Cladogram or tree output in your report.



○ Original Tree ● Bootstrap Tree

☑ Taxon Names
11 Font Size
☑ Hide Overlapping Taxa
□ Show Taxon Markers
Arrange Taxa
○ For Balanced Shape
○ By Input Order
☑ Use Subtree Draw Options
☑ Use Subtree Group Options
□ Branch Lengths
☑ Statistics/Frequency/Info
● Frequency ○ Site Coverage ○ Node IDs
8 Font Size
Placement Automatic
Distance From Node (Pixels)
Horizontal 5
Vertical 5

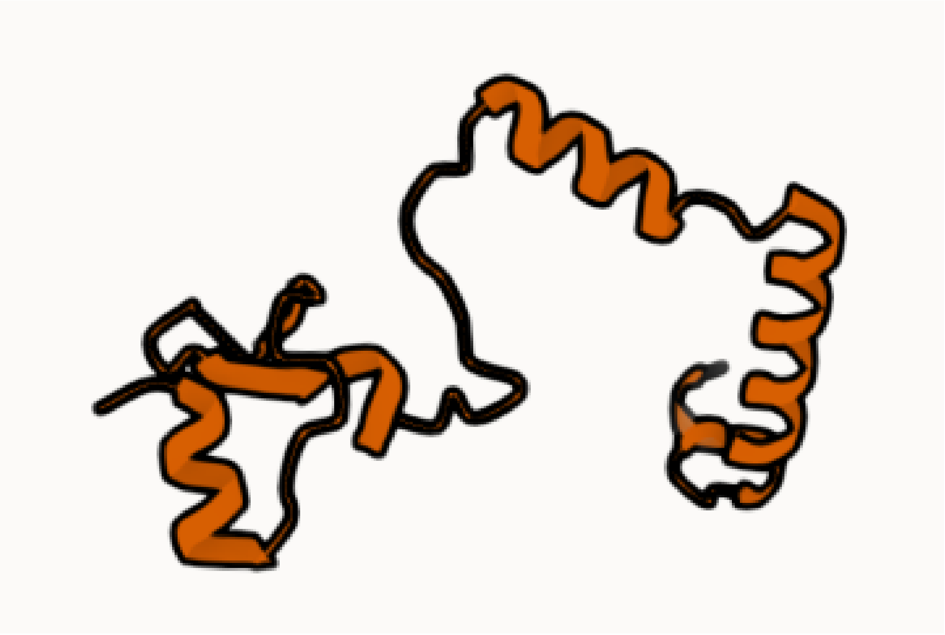
[Q7] Generate a sequence identity based heatmap of your aligned sequences using R. If necessary convert your sequence alignment to the ubiquitous FASTA format (Seaview can read in clustal format and “Save as” FASTA format for example). Read this FASTA format alignment into R with the help of functions in the Bio3D package. Calculate a sequence identity matrix (again using a function within the Bio3D package). Then generate a heatmap plot and add to your report. Do make sure your labels are visible and not cut at the figure margins.



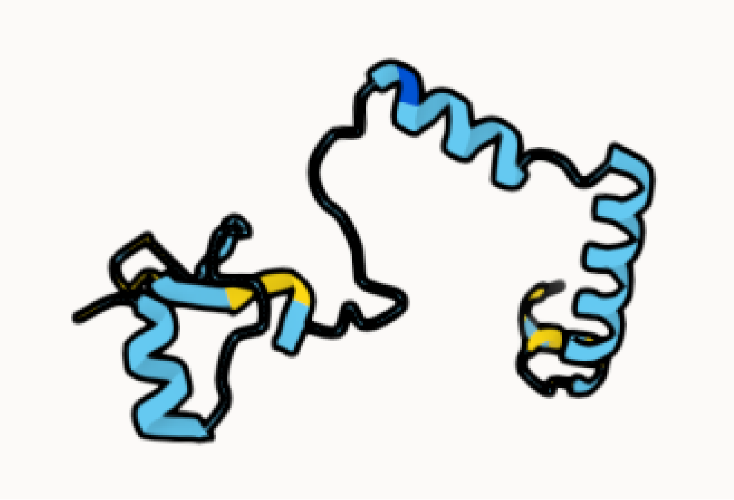
[Q8] Using R/Bio3D (or an online blast server if you prefer), search the main protein structure database for the most similar atomic resolution structures to your aligned sequences. List the top 3 unique hits (i.e. not hits representing different chains from the same structure) along with their Evalue and sequence identity to your query. Please also add annotation details of these structures. For example include the annotation terms PDB identifier (structureId), Method used to solve the structure (experimentalTechnique), resolution (resolution), and source organism (source).

ID	Technique	Resolution	Source	Evalue	Identity
4F3T_A	X-Ray Diffraction	2.250	Homosapiens	0	99.07
4OLA_A	X-Ray Diffraction	2.300	Homosapiens	0	98.95
5T7B_A	X-Ray Diffraction	2.529	Homosapiens	0	98.84

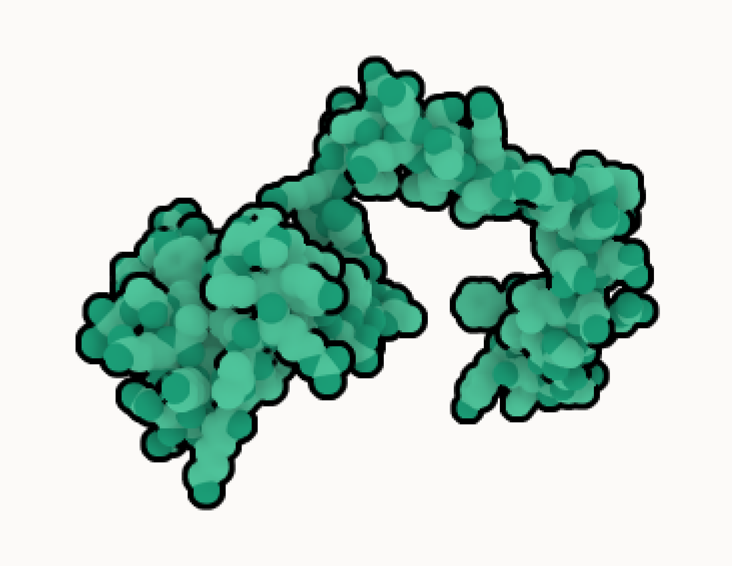
[Q9] Using AlphaFold notebook generate a structural model using the default parameters for your novel protein sequence. Note that this can take some time depending upon your sequence length. If your model is taking many hours to generate or your input sequence yields a “too many amino acids” (i.e. length) error you can focus on a single domain from your sequence - identify region by searching for PFAM domain matches. Once complete save the resulting PDB format file for your records. Finally, generate a molecular figure of your generated PDB structure using the Mol* viewer online (or VMD/PyMol/Chimera if you prefer). To complete your analysis you can optionally highlight conserved residues that are likely to be functional as spacefill and the protein as cartoon colored by local alpha fold pLDDT quality score. This score is contained in the B-factor column of your PDB downloaded file. Please use a white or transparent background for your figure (i.e. not the default black in PyMol/VMD/Chimera etc.).



A. californica protein structure



The A. californica protein structure is colored by pLDDT score, indicating AlphaFold's confidence. Blue regions (pLDDT > 70) are highly reliable, while yellow and black suggest lower confidence, indicating flexibility or disorder.



The conserved residues are shown in spacefill representation, highlighting their potential functional and structural importance.

[Q10] Perform a “Target” search of ChEMBEL (https://www.ebi.ac.uk/chembl/) with your novel sequence. Are there any Target Associated Assays and ligand efficiency data reported that may be useful starting points for exploring potential inhibition of your novel protein? If there are no assays listed here simply list “non available as of [date]”.

ChEMBL details 1 Binding Assay (CHEMBL5137313); No ligand efficiency data.

https://www.ebi.ac.uk/chembl/explore/assay/CHEMBL5137313

Target: CHEMBL612545 (Unchecked)
Target type: UNCHECKED

ChEMBL details 1 Binding Assay (CHEMBL2183115); No ligand efficiency data.

https://www.ebi.ac.uk/chembl/explore/assay/CHEMBL2183115

Target: CHEMBL1944497 (Soluble acetylcholine receptor)
Target type: SINGLE PROTEIN