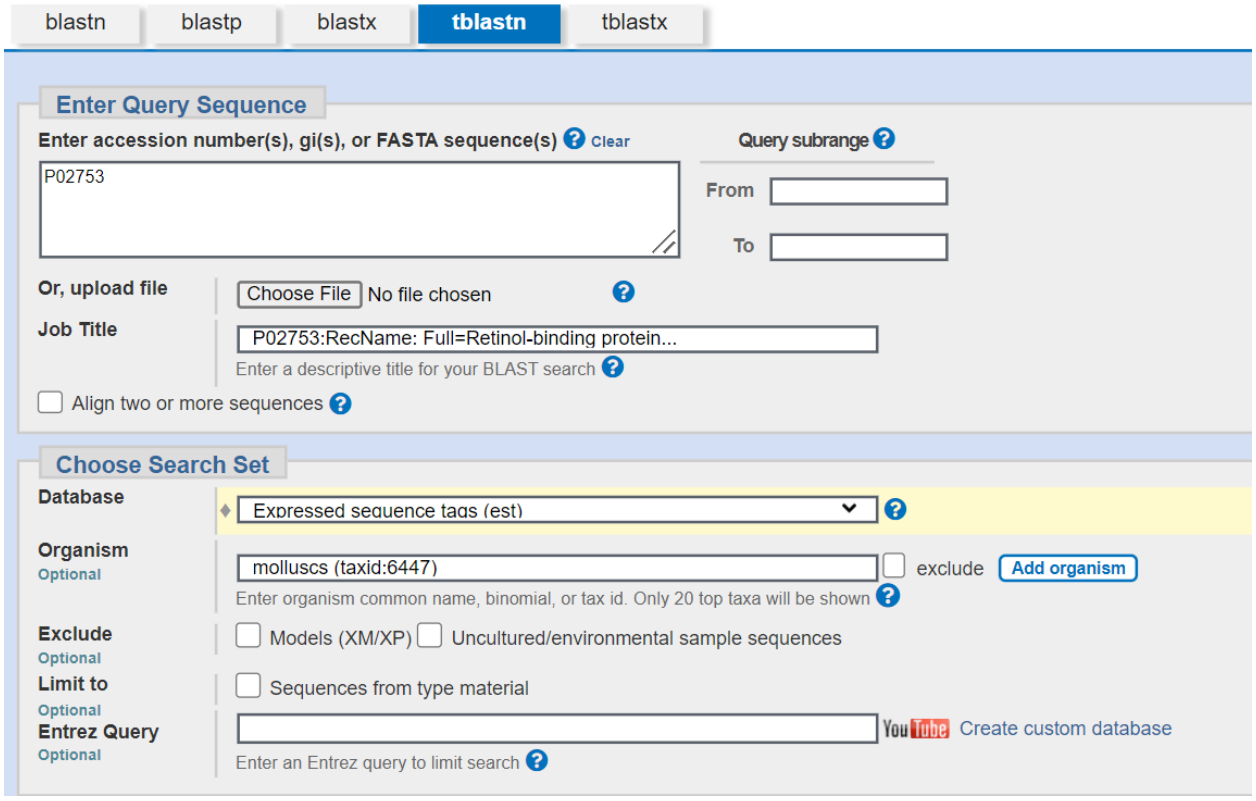
FIND A GENE - MEG ROBINSON

[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 it's 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: Retinol-binding protein 4
Accession: P02753
Species: Homo sapiens (Human)

[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). Also include the output of that BLAST search in your document. On the BLAST results, clearly indicate a match that represents a protein sequence, encoded from some DNA sequence, that is homologous to your query protein. I need to be able to inspect the pairwise alignment you have selected, including the E value and score. It should be labeled a "genomic clone" or "mRNA sequence", etc. - but include no functional annotation.

Method: TBLASTN (2.7.1) search against zebra fish ESTs
Database: Expressed Sequence Tags (est)
Organism : molluscs (taxid:6447))

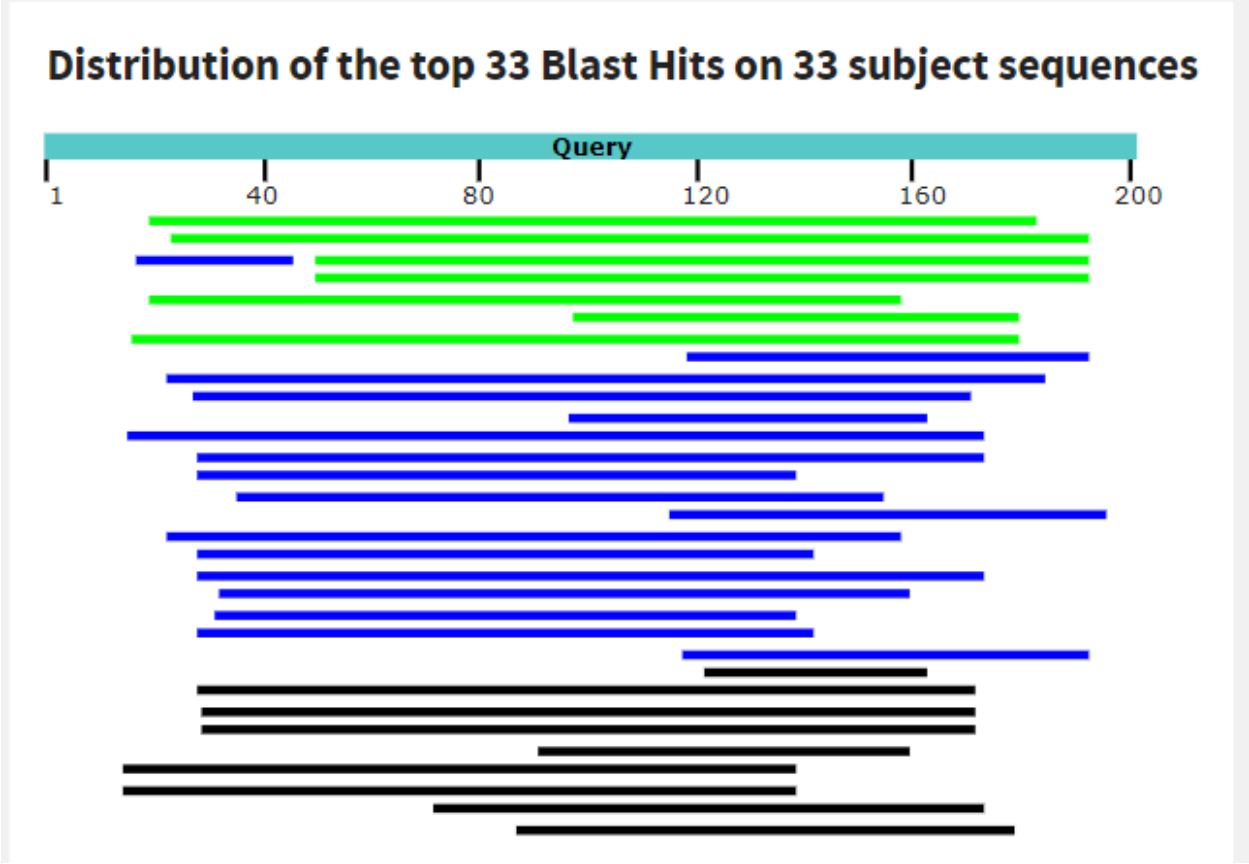


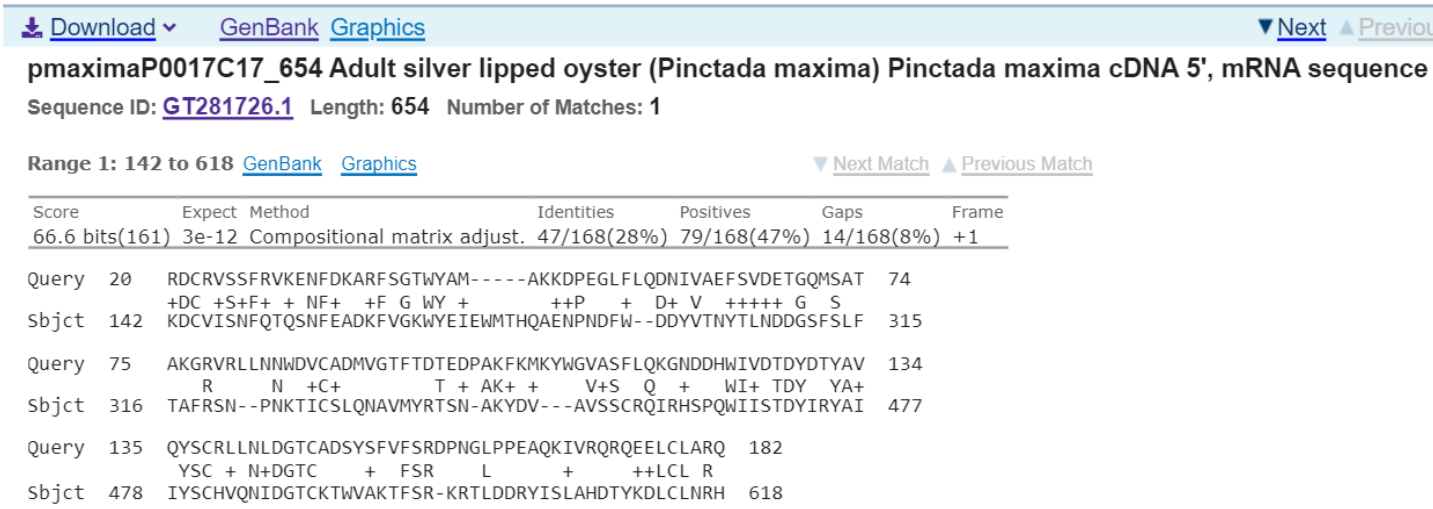
Sequences producing significant alignments Download Select columns Show 100

select all *33 sequences selected* GenBank Graphics

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
pmaximaP0017C17_654 Adult silver lipped oyster (Pinctada maxima) Pinctada maxima cDNA 5', mRNA sequence	Pinctada maxima	66.6	66.6	81%	3e-12	27.98%	654	GT281726.1
FQ660773 Crassostrea gigas library (Genoscope - CEA) Crassostrea gigas cDNA clone WY0AAA57YO16FM1, ...	Crassostrea gigas	63.2	63.2	84%	1e-10	30.43%	818	FQ660773.1
CU987947 Pascal Favrel cDNA library all stages (wy0aaa) Crassostrea gigas cDNA clone wy0aaa16c22, mRNA ...	Crassostrea gigas	59.3	59.3	70%	1e-09	30.20%	636	CU987947.1
CU996704 Pascal Favrel cDNA library all stages (wy0aaa) Crassostrea gigas cDNA clone wy0aaa30b08, mRNA ...	Crassostrea gigas	60.1	60.1	70%	1e-09	30.67%	1039	CU996704.1
EST1436 cDNA library in Pinctada martensi Pinctada martensii cDNA clone CHIP_05_stomach_08_93.SP6_E12...	Pinctada imbricata	58.2	58.2	68%	9e-09	28.37%	1550	FG592628.1

Chosen match: Accession GT281726.1 Adult silver lipped oyster (Pinctada maxima) Pinctada maxima cDNA 5', mRNA sequence. See below for alignment details.





[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.

Translated Sequence:

>GT281726.1 | pmaximaP0017C17_654 Adult silver lipped oyster (Pinctada maxima) Pinctada maxima cDNA 5', mRNA sequence

KDCVISNFQTQSNFEADKFVGKWYEIEWMTHQAENPNDFW--DDYVTNYTLNDDGSFSLF

TAFRSN--PNKTICSLQNAVMYRTSN-AKYDV---AVSSCRQIRHSPQWIISTDYIRYAI

IYSCHVQNIDGTCKTWVAKTFSR-KRTLDDRYISLAHDTYKDLCLNRH

>Human RBP4 | NP_001310447.1 | retinol-binding protein 4 isoform b [Homo sapiens]

RDCRVSSFRVKENFDKARFSGTWYAM-----AKKDPEGLFLQDNIVAEFSVDETGQMSAT

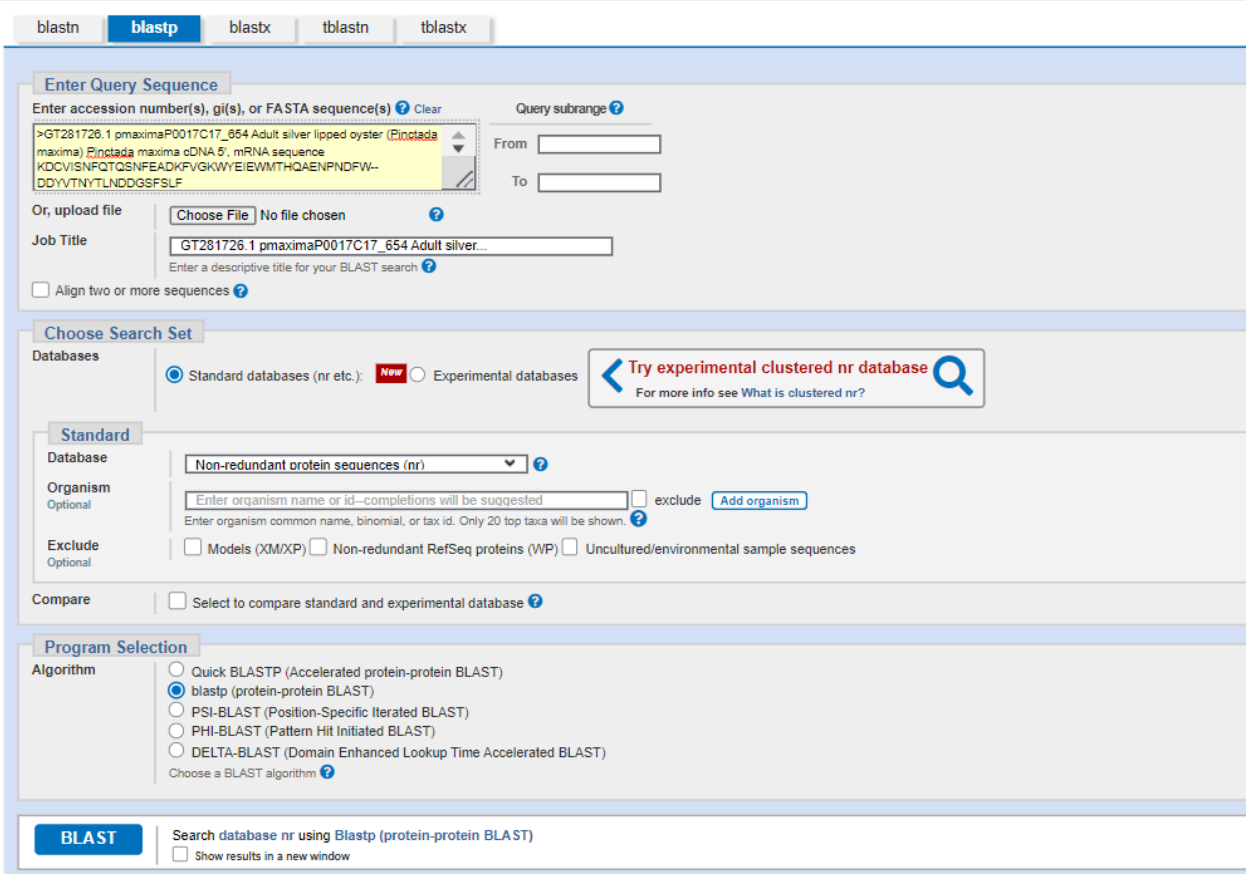
AKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAV

QYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQ

SOURCE Pinctada maxima (silver-lipped pearl oyster)
Eukaryota; Metazoa; Spiralia; Lophotrochozoa; Mollusca; Bivalvia; Autobranchia; Pteriomorphia; Pterioida; Pterioidea; Pteriidae; Pinctada.

[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. • If there is a match with 100% amino acid identity to a protein in the database, from the same species, then your protein is NOT novel (even if the match is to a protein with a name such as “unknown”). Someone has already found and annotated this sequence, and assigned it an accession number. • If the top match reported has less than 100% identity, then it is likely that your protein is novel, and you have succeeded. • If there is a match with 100% identity, but to a different species than the one you started with, then you have likely succeeded in finding a novel gene. • If there are no database matches to the original query from [Q1], this indicates that you have partially succeeded: yes, you may have found a new gene, but no, it is not actually homologous to the original query. You should probably start over.

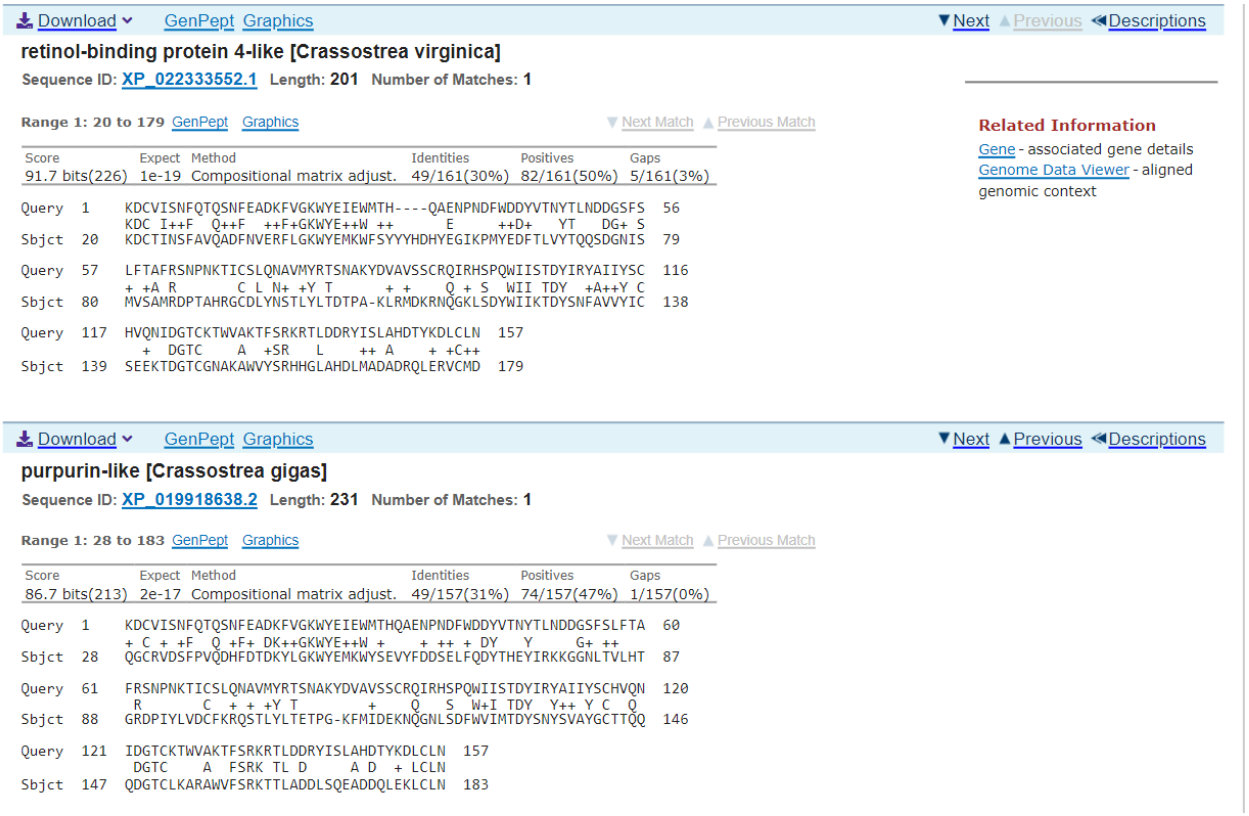
A BLASTP search against NR database (see setup in first screen-shot below) yielded a top hit result is to a protein from Crassostrea virginica (eastern oyster) but with not 100% identity (30.43%). See additional screen shots below for top hits and selected alignment details:



Sequences producing significant alignments Download Select columns Show 100

select all *100 sequences selected* GenPept Graphics Distance tree of results Multiple alignment MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
retinol-binding protein 4-like [Crassostrea virginica]	Crassostrea virg…	91.7	91.7	98%	1e-19	30.43%	201	XP_022333552.1
purpurin-like [Crassostrea gigas]	Crassostrea gigas	86.7	86.7	98%	2e-17	31.21%	231	XP_019918638.2
apolipoprotein D [Lingula anatina]	Lingula anatina	86.3	86.3	97%	3e-17	28.85%	219	XP_013390868.1
uncharacterized protein LOC124124228 [Haliotis rufescens]	Haliotis rufescens	87.8	151	99%	7e-17	31.01%	412	XP_046343315.1
purpurin-like isoform X1 [Crassostrea gigas]	Crassostrea gigas	84.3	84.3	96%	1e-16	27.92%	202	XP_034326755.1
retinol-binding protein 4-A-like [Crassostrea gigas]	Crassostrea gigas	83.2	83.2	96%	3e-16	30.52%	202	XP_034326754.1



Download GenPept Graphics ▼Next ▲Previous ◀◀Descriptions

retinol-binding protein 4-like [Crassostrea virginica]

Sequence ID: XP_022333552.1 Length: 201 Number of Matches: 1

Range 1: 20 to 179 GenPept Graphics ▼Next Match ▲Previous Match

Score	Expect	Method	Identities	Positives	Gaps
91.7 bits(226)	1e-19	Compositional matrix adjust.	49/161(30%)	82/161(50%)	5/161(3%)

```
Query  1    KDCVISNFQTQSNFEADKFVGKWYEIEWMTH----QAENPNDFWDDYVTNYTLNDDGSFS  56
            KDC I++F  Q++F  ++F+GKWYE++W ++      E     ++D+   YT   DG+ S
Sbjct  20   KDCTINSFAVQADFNVERFLGKWYEMKWFSYYYHDHYEGIKPMYEDFTLVYTQQSDGNIS  79

Query  57   LFTAFRSNPNKTICSLQNAVMYRTSNAKYDVAVSSCRQIRHSPQWIISTDYIRYAIIYSC  116
            + +A R        C L N+ +Y T      + +     Q + S  WII TDY  +A++Y C
Sbjct  80   MVSAMRDPTAHRGCDLYNSTLYLTDTPA-KLRMDKRNQGKLSDYWIIKTDYSNFAVVYIC  138

Query  117  HVQNIDGTCKTWVAKTFSRKRTLDDRYISLAHDTYKDLCLN  157
             +  DGTC    A  +SR   L    ++ A    + +C++
Sbjct  139  SEEKTDGTCGNAKAWVYSRHHGLAHDLMADADRQLERVCMD  179
```


Download GenPept Graphics ▼Next ▲Previous ◀◀Descriptions

purpurin-like [Crassostrea gigas]

Sequence ID: XP_019918638.2 Length: 231 Number of Matches: 1

Range 1: 28 to 183 GenPept Graphics ▼Next Match ▲Previous Match

Score	Expect	Method	Identities	Positives	Gaps
86.7 bits(213)	2e-17	Compositional matrix adjust.	49/157(31%)	74/157(47%)	1/157(0%)

```
Query  1    KDCVISNFQTQSNFEADKFVGKWYEIEWMTHQAENPNDFWDDYVTNYTLNDDGSFSLFTA  60
            + C + +F  Q +F+ DK++GKWYE++W +    + ++ + DY   Y     G+ ++
Sbjct  28   QGCRVDSFPVQDHFDTDKYLGKWYEMKWYSEVYFDDSELFQDYTHEYIRKKGGNLTVLHT  87

Query  61   FRSNPNKTICSLQNAVMYRTSNAKYDVAVSSCRQIRHSPQWIISTDYIRYAIIYSCHVQN  120
             R       C  + + +Y T          +    Q   S  W+I TDY  Y++ Y C  Q
Sbjct  88   GRDPIYLVDCFKRQSTLYLTETPG-KFMIDEKNQGNLSDFWVIMTDYSNYSVAYGCTTQQ  146

Query  121  IDGTCKTWVAKTFSRKRTLDDRYISLAHDTYKDLCLN  157
              DGTC    A  FSRK TL D     A D  + LCLN
Sbjct  147  QDGTCLKARAWVFSRKTTLADDLSQEADDQLEKLCLN  183
```

[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 alignment for building a phylogenetic tree that illustrates species divergence.

Re-labeled sequences for alignment:

```
>Human_original_seq | NP_001310447.1 | retinol-binding protein 4
isoform b [Homo sapiens]
RDCRVSSFRVKENFDKARFSGTWYAM-----AKKDPEGLFLQDNIVAEFSVDETGQMSAT
AKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAV
QYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQ

> Silver_lipped_oyster_novel_protein | taken from BLAST | Pinctada
maxima cDNA 5', mRNA sequence | GT281726.1
KDCVISNFQTQSNFEADKFVGKWYEIEWMTHQAENPNDFW--DDYVTNYTLNDDGSFSLF
TAFRSN--PNKTICSLQNAVMYRTSN-AKYDV---AVSSCRQIRHSPQWIISTDYIRYAI
IYSCHVQNIDGTCKTWVAKTFSR-KRTLDDRYISLAHDTYKDLCLNRH

> Eastern_Oyster |  Crassostrea virginica
KDCVISNFQTQSNFEADKFVGKWYEIEWMTH----QAENPNDFWDDYVTNYTLNDDGSFS
LFTAFRSNPNKTICSLQNAVMYRTSNAKYDVAVSSCRQIRHSPQWIISTDYIRYAIIYSC
HVQNIDGTCKTWVAKTFSRKRTLDDRYISLAHDTYKDLCLN

> Scallop | XP_021358553.1 | retinol-binding protein 4-like
[Mizuhopecten yessoensis]
RDCRLSSFQLQPDFNLAQFEGYWYSLTINRYWLAVPRWFPVRQSNVQVNYRLLSDGSLEV
KTGGEFMFPFCDYIAGKGYIPDKTQPQKLEVQFDTLLTRTSRKNPYWVVSTDYEGFAVIY
SCWKEREDGTC

> Snail | XP_041352779.1 | retinol-binding protein 4-like
[Gigantopelta aegis]
KDCNVHNITVQPDFDLQKYAGTWYEMKWLATVYIPPNQLYQDYRHIYTYAGRGENVTVDI
AGRDPANINKCFYNTARLMETDDDAK---MVFATVKERYS-YWVVKTDYTSYAVVYGCYS
VTSDGACNGTRSWIWSRTKTLSQDKMAIAETVISETLCVNQ

> Mandarin_fish | XP_044034678.1 |retinol-binding protein 4
[Siniperca chuatsi]
QDCQVANIQVMQNFDKTRYAGTWYVIA-----KKDPEGLFLLDNIMAQFTVADDGKMTAT
AKGRVIILNNWEMCADMLATFEETSDPAKFRMKYWGVASYLQTGNDDHWVIDTDYDNYAI
HYSCRLQDSDGTCLDSYSMIFSRHLDGLRPEDQRTV---HQKKMDLCL

> Painted_turtle | XP_005301331.1 | retinol-binding protein 4
[Chrysemys picta bellii]
RDCRVSNFRVQENFDKARYTGTWYAIA-----KKDPEGLFLQDNVVAQFTIDENGQMSAT
AKGRVRLFNNWDVCADMIGSFTDTEDPAKFKMKYWGVASFLQKGNDDHWVVDTDYDTYAL
HYSCRQLNDDGTCADSYSFVFSRDPKGLSPEVQRIIRQRQVDLCLDR
```

Alignment:

Obtained using MUSCLE (version 3.8) at EBI:

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


Snail                                KDCNVHNITVQPDFDLQKYAGTWYEMKWLATVYIPPNQLY-----QDYRHIYTYAGRGE-
Mandarin_fish                        QDCQVANIQVMQNFDKTRYAGTWYVIA-----KKDPEGLF---LLDNIMAQFTVADDGKM
Human_original_seq                   RDCRVSSFRVKENFDKARFSGTWYAMA-----KKDPEGLF---LQDNIVAEFSVDETGQM
Painted_turtle                       RDCRVSNFRVQENFDKARYTGTWYAIA-----KKDPEGLF---LQDNVVAQFTIDENGQM
Silver_lipped_oyster_novel_prote     KDCVISNFQTQSNFEADKFVGKWYEIEWMTHQAENPNDFW-----DDYVTNYTLNDDGSF
Eastern_Oyster                       KDCVISNFQTQSNFEADKFVGKWYEIEWMTHQAENPNDFW-----DDYVTNYTLNDDGSF
Scallop                              RDCRLSSFQLQPDFNLAQFEGYWYSLTINRYWLAVPR--WFPVRQSNVQVNYRLLSDGSL
                                     .** : .:    :*:  .: * ** :         *   :     .:    :     *.

Snail                                NVTVDIAGRDPANINKCFYNTARLMETDDDAKMVFA-------TVKERYSYWVVKTDYTS
Mandarin_fish                        TATAKGRVIILNNWEMCADMLATFEETSDPAKFRMKYWGVASYLQTGNDDHWVIDTDYDN
Human_original_seq                   SATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDT
Painted_turtle                       SATAKGRVRLFNNWDVCADMIGSFTDTEDPAKFKMKYWGVASFLQKGNDDHWVVDTDYDT
Silver_lipped_oyster_novel_prote     SLFTAFRSNPNKTICSLQNAVMY---RTSNAKYDVA---VSSCRQIRHSPQWIISTDYIR
Eastern_Oyster                       SLFTAFRSNPNKTICSLQNAVMY---RTSNAKYDVA---VSSCRQIRHSPQWIISTDYIR
Scallop                              EVKTG--GEFMFPFCDYIAGKGYIPDKTQPQKLEVQFDTLLT-RTSRKNPYWVVSTDYEG
                                        .                         .  *  .            .   *::.***

Snail                                YAVVYGCYSVTSDGACNGTRSWIWSRTKTLSQDKMAIAETVISETLCVNQ-
Mandarin_fish                        YAIHYSCRLQDSDGTCLDSYSMIFSRHLDGLRPEDQRTVHQKKMDLCL---
Human_original_seq                   YAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQ
Painted_turtle                       YALHYSCRQLNDDGTCADSYSFVFSRDPKGLSPEVQRIIRQRQVDLCLDR-
Silver_lipped_oyster_novel_prote     YAIIYSCHVQNIDGTCKTWVAKTFSRKRT-LDDRYISLAHDTYKDLCLNRH
Eastern_Oyster                       YAIIYSCHVQNIDGTCKTWVAKTFSRKRT-LDDRYISLAHDTYKDLCLN--
Scallop                              FAVIYSCWKEREDGTC-----------------------------------
                                     :*: *.*     **:*
```

[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.

Q7-Q10

Meg Robinson

3/11/2022

Question 7

[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.

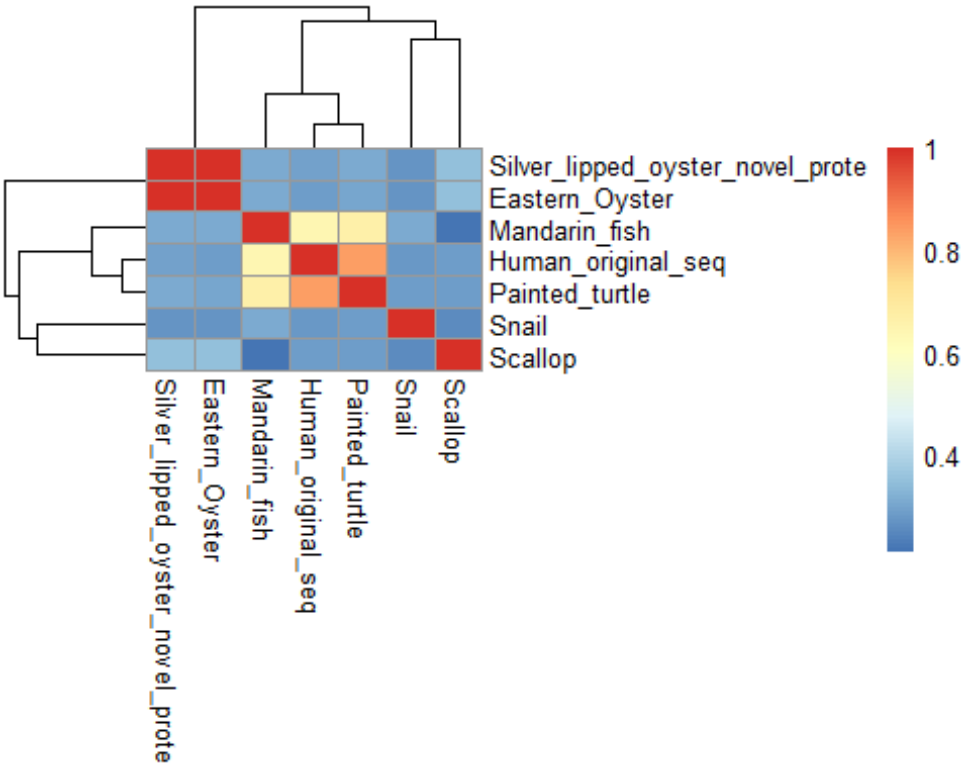
```
library(bio3d)

## Warning: package 'bio3d' was built under R version 4.1.2

library(pheatmap)

## Warning: package 'pheatmap' was built under R version 4.1.2

aln <- read.fasta("cluster.fasta")
iden <- seqidentity(aln)
pheatmap(iden, margins = c(12,12))
```



Question 8

[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).

Use the sequence with the highest identity

```
rowmean <- rowMeans(iden)
rowmean
```

```
##                            Snail                      Mandarin_fish
##                        0.3811429                          0.4931429
##                Human_original_seq                     Painted_turtle
##                        0.5180000                          0.5265714
## Silver_lipped_oyster_novel_prote                     Eastern_Oyster
##                        0.5040000                          0.5032857
##                          Scallop
##                        0.3904286
```

```
highest.name <- names(which.max(rowmean))
highest.val <- max(rowMeans(iden))*100
```

```
print(paste("The species with the highest average row identity is",
highest.name))
```

```
## [1] "The species with the highest average row identity is Painted_turtle"
```

```
print(paste("The average row identity of this species is",
round(highest.val,2), "%"))
```

```
## [1] "The average row identity of this species is 52.66 %"
```

Now we will continue with the FASTA of the turtle

```
turtle <- read.fasta("turtle.fasta")
turtle.blast <- blast.pdb(turtle)
```

```
##  Searching ... please wait (updates every 5 seconds) RID = 2RPPBTPA016
##  .........
##  Reporting 111 hits
```

```
# top 3 hits ID
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
df.blast <- turtle.blast$hit.tbl
df.blast.id <- head(df.blast$subjectids,3)

df.eval.iden <- select(head(df.blast,3), c("evalue","identity"))
rownames(df.eval.iden) <-
as.vector(select(head(df.blast,3),"subjectids"))[1:3,]

df.eval.iden
```

```
##           evalue identity
## 1IIU_A 1.52e-113   91.304
## 4O9S_A 9.07e-106   83.951
## 3FMZ_A 9.23e-106   83.951
```

```
df.pdb <- lapply(df.blast.id,pdb.annotate) %>% bind_rows()
df.pdb.select <- select(df.pdb,
c("structureId","experimentalTechnique","resolution","source"))
df.pdb.select
```

```
##        structureId experimentalTechnique resolution        source
## 1IIU_A        1IIU                X-ray        2.5 Gallus gallus
## 4O9S_A        4O9S                X-ray        2.3  Homo sapiens
## 3FMZ_A        3FMZ                X-ray        2.9  Homo sapiens
```

Now we can combine our two dataframes

```
df.final <- cbind(df.eval.iden, df.pdb.select)
colnames(df.final) <- c("Evalue", "Identity", "ID", "Technique",
"Resolution", "Source")
df.final
```

```
##           Evalue Identity   ID Technique Resolution        Source
## 1IIU_A 1.52e-113   91.304 1IIU     X-ray        2.5 Gallus gallus
## 4O9S_A 9.07e-106   83.951 4O9S     X-ray        2.3  Homo sapiens
## 3FMZ_A 9.23e-106   83.951 3FMZ     X-ray        2.9  Homo sapiens
```

Question 9

> [Q9] Generate a molecular figure of one of your identified PDB structures using VMD. You can optionally highlight conserved residues that are likely to be functional. Please use a white or transparent background for your figure (i.e. not the default black). Based on sequence similarity. How likely is this structure to be similar to your "novel" protein?

Based on the sequence similarity, this is very likely to be similar in structure to my novel protein Pinctada maxima due to the high sequence similarity (>91%).

(I used 1IIU for Q9)

Question 10

[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?

CHEMBL details 3 Binding Assays (CHEMBL3707846, CHEMBL3707845,CHEMBL3705836 – all the same) and 1 Inhibition Asssay (CHEMBL3887722); No ligand efficiency data.

https://www.ebi.ac.uk/chembl/g/#search_results/assays/query=%3E%20Silver_lipped_oyster_novel_protein%20%7C%20taken%20from%20BLAST%20%7C%20Pinctada%20maxima%20cDNA%205%26%23x27%3B%2C%20mRNA%20sequence%20%7C%20GT281726.1%20KDCVISNFQTQSNFEADKFVGKWYEIEWMTHQAENPNDFW--DDYVTNYTLNDDGSFSLF%20TAFRSN--PNKTICSLQNAVMYRTSN-AKYDV---AVSSCRQIRHSPQWIISTDYIRYAI%20IYSCHVQNIDGTCKTWVAKTFSR-KRTLDDRYISLAHDTYKDLCLNRH

In the binding assay they tested the affinity of various compounds at the NE, DA and 5HT transporters of HEK293E cell lines. There are no references provided or useful information.