BIMM-143: INTRODUCTION TO BIOINFORMATICS

The find-a-gene project assignment
http://thegrantlab.org/bimm143
Dr. Barry Grant

Version: 2025-04-07 (13:23:59 PDT on Mon, Apr 07)

Overview:

The find-a-gene project is a required assignment for BIMM-143. You should prepare a written report in **PDF** format that has responses to each question labeled **[Q1] - [Q10]** below. You may wish to consult the scoring rubric at the end of this document and the example report provided online (note that the example report is from a *previous quarter* and the questions may differ).

The objective with this assignment is for you to demonstrate your grasp of database searching, sequence analysis, structure analysis and the R environment that we have covered in class.

Due Date:

Your responses to questions Q1-Q4 are due at 12pm on the **Monday of Week 5** (see the Assignments and Grading section of our website for details). Note that these first set of answers can be obtained very quickly (at best within 15 or 20 minutes), so if you don't succeed at first, just keep trying.

The complete assignment, including responses to all questions, is due at 12pm on the **Monday of Week 10**.

Submission instructions:

Your report formatted as a **PDF document** should be uploaded to ***GradeScope***. Please make sure to include your UCSD email and PID number on the first page.

Be sure to include your UCSD email and PID number on the first page of your report.

Submit your preliminary report with answers to Q1-Q4 as soon as you can so we can determine if you have found a novel gene. Submit this preliminary report as one document with screen shots of the results inserted appropriately.

See the demonstration report linked to on the course website for an example of format. Note again that example questions may differ. I will indicate on GradeScope my decision (1pt indicating all is good, 0pts revisions required). You should proceed with subsequent questions only after we are sure you have found a novel gene (and thus be successful in the later stages of the project).

For the final report add your results for Q5-Q10 to the preliminary report and submit the final

document containing your results for all questions.
Please do not send only Q5-Q10 answers as the final report.
Questions:

[Q1] Tell me the name of a protein you are interested in. Include the species, accession number and known function. 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

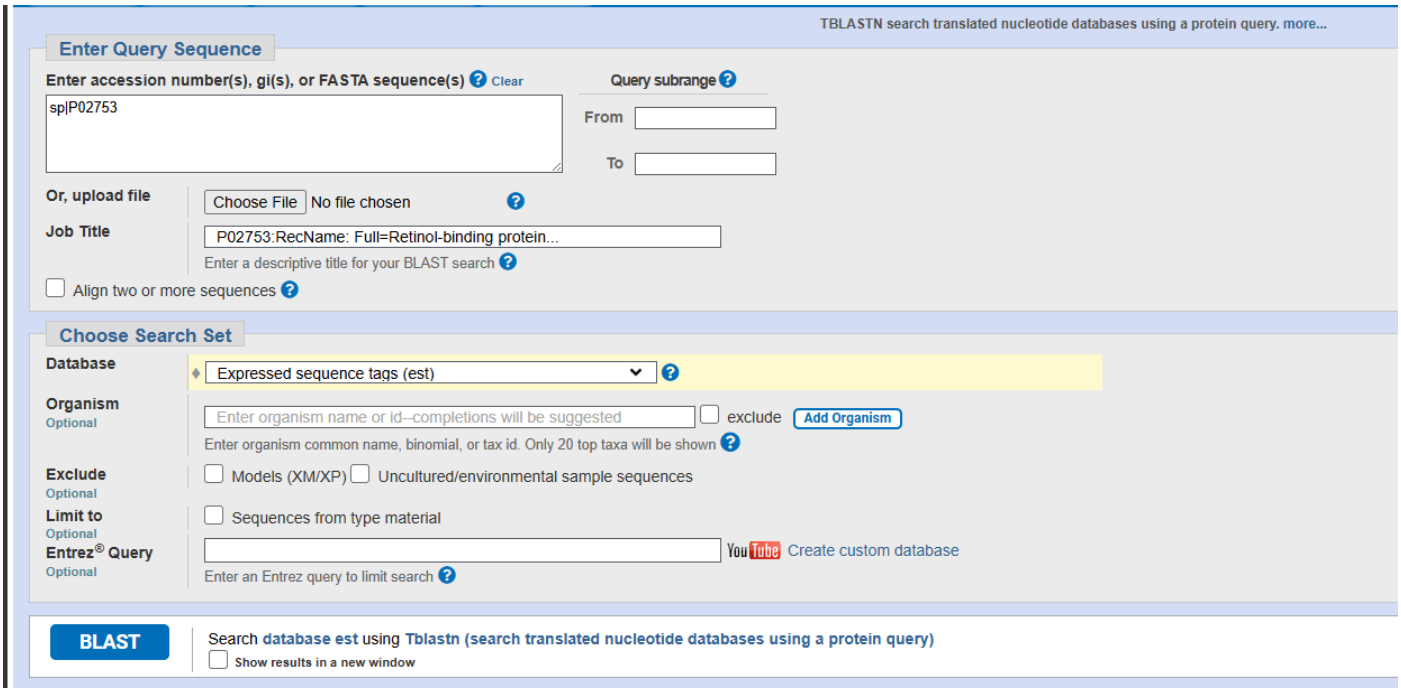
Species: Homo Sapiens

[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

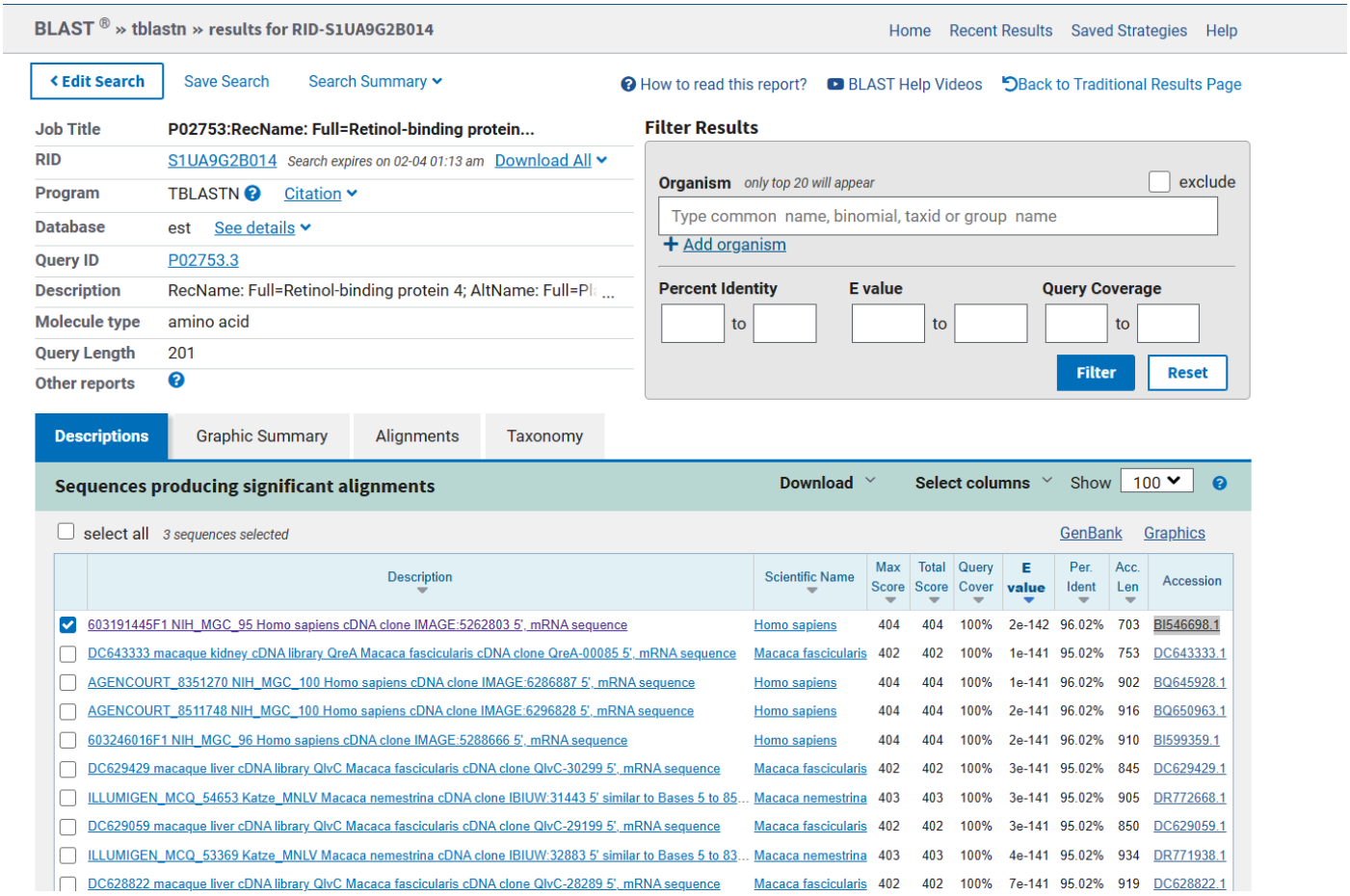
Database: Expressed Sequence Tags (est)

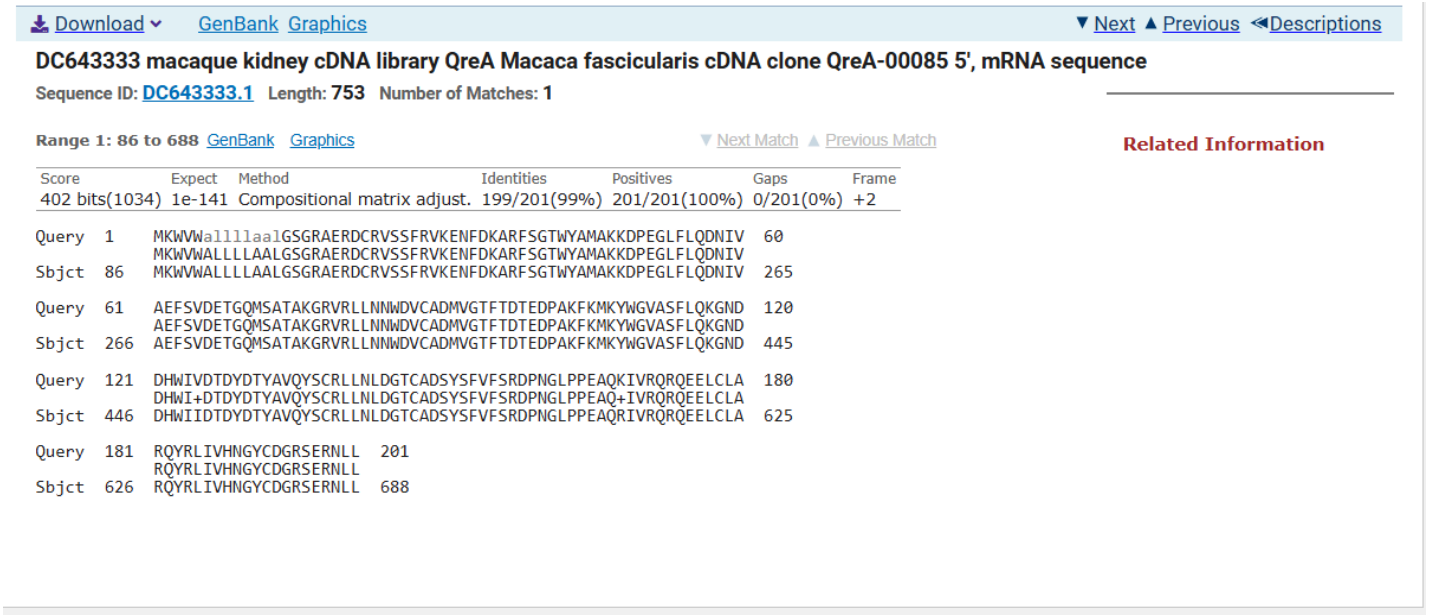
Also include the output of that BLAST search in your document. If appropriate, change the font to `Courier size 10` so that the results are displayed neatly. You can also screen capture a BLAST output (e.g. alt print screen on a PC or on a MAC press ⌘-shift-4. The pointer becomes a bulls eye. Select the area you wish to capture and release. The image is saved as a file called `Screen Shot [].png` in your Desktop directory). It is **not** necessary to print out all of the blast results if there are many pages.



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.

Chosen match: DC643333.1, 753 base pair clone from Macaca fascicularis





In general, [Q2] is the most difficult for students because it requires you to have a “feel”

for how to interpret BLAST results. You need to distinguish between a perfect match to your query (i.e. a sequence that is not “novel”), a near match (something that might be “novel”, depending on the results of [Q4]), and a non-homologous result.

If you are having trouble finding a novel gene try restricting your search to an organism that is poorly annotated.

[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. Here, tell me the name of the novel protein, and the species from which it derives. It is very unlikely (but still definitely possible) that you will find a novel gene from an organism such as *S. cerevisiae*, human or mouse, because those genomes have already been thoroughly annotated. It is more likely that you will discover a new gene in a genome that is currently being sequenced, such as bacteria or plants or protozoa.

Chosen Sequence:

>M. Fascicularis protein (sequence taken from BLAST result)
MKWVWallllaalGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV
MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQ
DNIVMKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGL
FLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWG
VASFLQKGNDAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKM
KYWGVASFLQKGNDAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPA
KFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDP
NGLPPEAQKIVRQRQEELCLADHWI+DTDYDTYAVQYSCRLLNLDGTCADSYSFVFSR
DPNGLPPEAQ+IVRQRQEELCLADHWIIDTDYDTYAVQYSCRLLNLDGTCADSYSFVFS
RDPNGLPPEAQRIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL
RQYRLIVHNGYCDGRSERNLLRQYRLIVHNGYCDGRSERNLL

Name:

Macaca kidney protein

Species

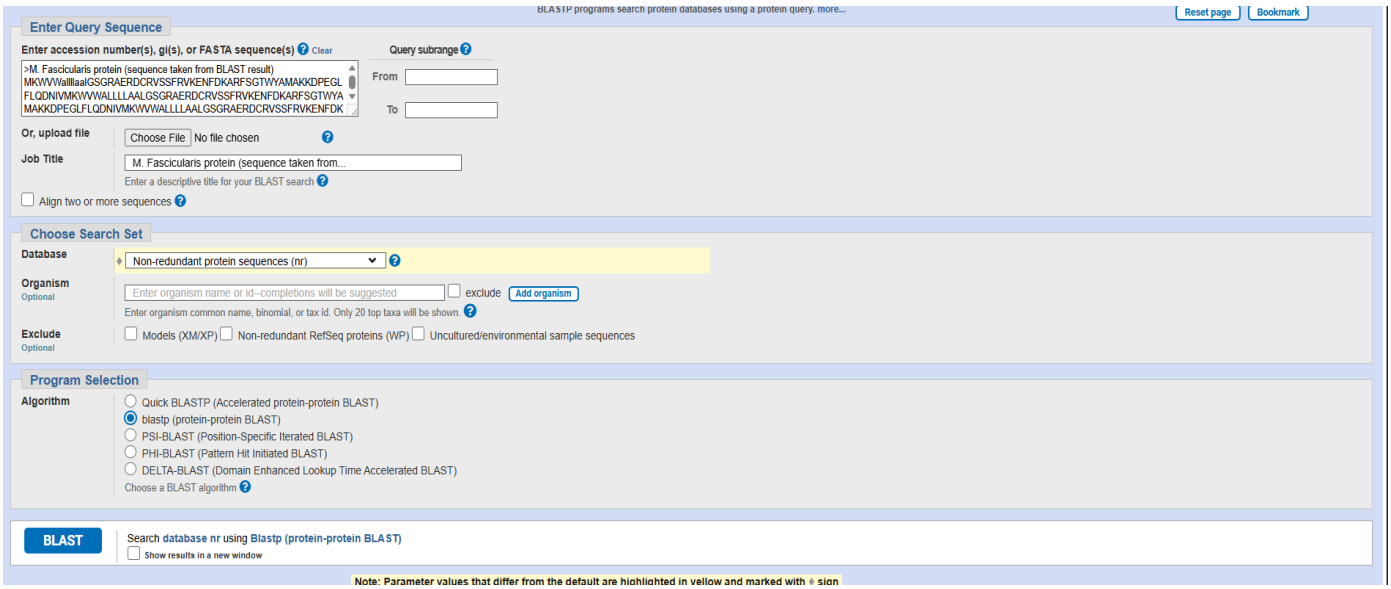
:Macaca fascicularis

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Cercopithecidae; Cercopithecinae; Macaca.

[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 yielded a top hit result to a protein from sciurus carolinensis (Eastern Grey Squirrel).



Download GenPept Graphics Sort by: E value Next Previous Descriptions

Retinol-binding protein 4 [Sciurus carolinensis]

Sequence ID: MBZ3872464.1 Length: 262 Number of Matches: 5

Related Information

Range 1: 1 to 250 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
335 bits(859)	2e-107	Compositional matrix adjust.	184/299(62%)	210/299(70%)	49/299(16%)

```
Query  121  MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV  180
            M+WVWAL+LLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYA+AKKDPEGLFLQDNIV
Sbjct  1    MEWVWALVLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAIAKKDPEGLFLQDNIV  60

Query  181  AEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGND  240
            AEFSVDE G MSATAKGRVRLL+NW+VCADMVGTFTDTEDPAKFKMKYWGVASFLQ+GND
Sbjct  61   AEFSVDEYGHMSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGND  120

Query  241  AEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGND  300
              + +D T   +   +   RLLN    CAD                  SF+
Sbjct  121  DHWIID-TDYDTFALQYSCRLLNLDGTCADSY-------------------SFV-----  154

Query  301  AEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGND  360
              F+ D  G +S   +  VR     ++C D              +   W     ++  +D
Sbjct  155  --FARDPNG-LSPETRKLVRQRQE-ELCLD--------------RQYRW-----IEHNDD  191

Query  361  DHWIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCL  419
            DHWI+DTDYDT+A+QYSCRLLNLDGTCADSYSFVF+RDPNGL PE +K+VRQRQEELCL
Sbjct  192  DHWIIDTDYDTFALQYSCRLLNLDGTCADSYSFVFARDPNGLSPETRKLVRQRQEELCL  250
```

Range 2: 34 to 250 GenPept Graphics Next Match Previous Match First Match

Score	Expect	Method	Identities	Positives	Gaps
328 bits(841)	1e-104	Compositional matrix adjust.	163/217(75%)	181/217(83%)	19/217(8%)

```
Query  280  DPAKFKMKYWGVAS------FLQKGNDAEFSVDETGQMSATAKGRVRLLNNWDVCADMVG  333
            D A+F   ++ +A        FLQ    AEFSVDE G MSATAKGRVRLL+NW+VCADMVG
Sbjct  34   DKARFSGTWYAIAKKDPEGLFLQDNIVAEFSVDEYGHMSATAKGRVRLLSNWEVCADMVG  93

Query  334  TFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRLLNLDGTCADSYSF  393
            TFTDTEDPAKFKMKYWGVASFLQ+GNDDHWI+DTDYDT+A+QYSCRLLNLDGTCADSYSF
Sbjct  94   TFTDTEDPAKFKMKYWGVASFLQRGNDDHWIIDTDYDTFALQYSCRLLNLDGTCADSYSF  153

Query  394  VFSRDPNGLPPEAQKIVRQRQEELCL-----------ADHW-IDTDYDTYAVQYSCRLLN  441
            VF+RDPNGL PE +K+VRQRQEELCL           DHW IDTDYDT+A+QYSCRLLN
Sbjct  154  VFARDPNGLSPETRKLVRQRQEELCLDRQYRWIEHNDDDHWIIDTDYDTFALQYSCRLLN  213

Query  442  LDGTCADSYSFVFSRDPNGLPPEA-QIVRQRQEELCL  477
            LDGTCADSYSFVF+RDPNGL PE  ++VRQRQEELCL
Sbjct  214  LDGTCADSYSFVFARDPNGLSPETRKLVRQRQEELCL  250
```

Range 3: 34 to 261 GenPept Graphics Next Match Previous Match First Match

Score	Expect	Method	Identities	Positives	Gaps
225 bits(574)	5e-65	Compositional matrix adjust.	130/236(55%)	155/236(65%)	35/236(14%)

Range 3: 34 to 261 GenPept Graphics ▼ Next Match ▲ Previous Match First Match

Score	Expect	Method	Identities	Positives	Gaps
225 bits(574)	5e-65	Compositional matrix adjust.	130/236(55%)	155/236(65%)	35/236(14%)

```
Query  340  DPAKFKMKYWGVAS------FLQKGNDDHWIVDTDYDTYA-----VQYSCRLLNLDGTCA  388
            D A+F   ++ +A       FLQ    D+ + +   D Y       +   RLL+    CA
Sbjct  34   DKARFSGTWYAIAKKDPEGLFLQ----DNIVAEFSVDEYGHMSATAKGRVRLLSNWEVCA  89

Query  389  DSY-SFVFSRDPNGLPPEAQKIVR--QRQEELCLADHWI-DTDYDTYAVQYSCRLLNLDG  444
            D   +F  + DP     +   +    QR  +    DHWI DTDYDT+A+QYSCRLLNLDG
Sbjct  90   DMVGTFTDTEDPAKFKMKYWGVASFLQRGND----DHWIIDTDYDTFALQYSCRLLNLDG  145

Query  445  TCADSYSFVFSRDPNGLPPEA-QIVRQRQEELCLA-----------DHWIIDTDYDTYAV  492
            TCADSYSFVF+RDPNGL PE  ++VRQRQEELCL            DHWIIDTDYDT+A+
Sbjct  146  TCADSYSFVFARDPNGLSPETRKLVRQRQEELCLDRQYRWIEHNDDDHWIIDTDYDTFAL  205

Query  493  QYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQRIVRQRQEELCLARQYRLIVHNG  548
            QYSCRLLNLDGTCADSYSFVF+RDPNGL PE +++VRQRQEELCL RQYR I HNG
Sbjct  206  QYSCRLLNLDGTCADSYSFVFARDPNGLSPETRKLVRQRQEELCLDRQYRWIEHNG  261
```

Range 4: 1 to 159 GenPept Graphics ▼ Next Match ▲ Previous Match First Match

Score	Expect	Method	Identities	Positives	Gaps
122 bits(306)	8e-27	Compositional matrix adjust.	81/167(49%)	100/167(59%)	14/167(8%)

```
Query  61   MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV  120
            M+WVWAL+LLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYA+AKKDPEGLFLQDNIV
Sbjct  1    MEWVWALVLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAIAKKDPEGLFLQDNIV  60

Query  121  MKWV---WALLLLAALGSGRAERDCRVSSFRV---KENFDKARFSGTWYAMAKKDPEGLF  174
             ++    +  +   A G  R   +  V +  V    +  D A+F   ++ +A       F
Sbjct  61   AEFSVDEYGHMSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVAS------F  114

Query  175  LQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDP  221
            LQ      + +D T   +   +   RLLN    CAD   +F    DP
Sbjct  115  LQRGNDDHWIID-TDYDTFALQYSCRLLNLDGTCADSY-SFVFARDP  159
```

Range 5: 1 to 63 GenPept Graphics ▼ Next Match ▲ Previous Match First Match

Score	Expect	Method	Identities	Positives	Gaps
121 bits(304)	1e-26	Compositional matrix adjust.	57/63(90%)	62/63(98%)	0/63(0%)

```
Query  1   MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV  60
           M+WVWAL+LLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYA+AKKDPEGLFLQDNIV
Sbjct  1   MEWVWALVLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAIAKKDPEGLFLQDNIV  60

Query  61  MKW  63
            ++
Sbjct  61  AEF  63
```

[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

Sequences Before Alignment

>MBZ3872464.1 Retinol-binding protein 4 [Sciurus carolinensis] (Novel)

MEWVWALVLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAIAKKDPEGLFLQDNIVAEFSVDEYGH

MSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDHWIIDTDYDTFALQYSCRL

LNLDGTCADSYSFVFARDPNGLSPETRKLVRQRQEELCLDRQYRWIEHNDDDHWIIDTDYDTFALQYSCR

LLNLDGTCADSYSFVFARDPNGLSPETRKLVRQRQEELCLDRQYRWIEHNGE

>M. Fascicularis protein (Initial)

MKWVWallllaalGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVMKWVWALLLLAALGSGRAERDCRVSSF
RVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVMKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPE
GLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDAEFSVDETGQMSATA
KGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDT
EDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLADH
WI+DTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQ+IVRQRQEELCLADHWIIDTDYDTYAVQYSCRLLNLDGTCAD
SYSFVFSRDPNGLPPEAQRIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL
RQYRLIVHNGYCDGRSERNLLRQYRLIVHNGYCDGRSERNLL

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

MNYSKIPAQVDLRRQTERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMS

ATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRLLN

LDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL

>XP_062070582.1 retinol-binding protein 4 [Lepus europaeus]

MEWVWALVLLAALGSGRAERDCRVSSFRVKENFDKARFAGTWYAMAKKDPEGLFLQDNIVAEFSVDENGH

MSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDHWIIDTDYDTYAVQYSCRL

LNFDGTCADSYSFVFSRDPHGLPPDVQKLVRQRQEELCLSRQYRLIVHNGYCDDKSVRNLL

>XP_008268128.1 retinol-binding protein 4 isoform X1 [Oryctolagus cuniculus]

MEWVWALVLLAALGSGRAERDCRVSSFRVKENFDKARFAGTWYAMAKKDPEGLFLQDNIVAEFSVDENGH

MSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDHWIIDTDYDTFAVQYSCRL

LNFDGTCADSYSFVFSRDPHGLPPDVQKLVRQRQEELCLSRQYRLIVHNGYCDDKSVRNLL

>XP_037012605.2 retinol-binding protein 4 [Artibeus jamaicensis]

MEWVWALVLLAALGSARAERDCRVSSFRVKENFDKARFSGIWYAVAKKDPEGLFLQDNIIAEFSVDENGQ

MSATAKGRVTLLNNWEVCADMLGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIIDTDYDTYAVQYSCRL

LNLDGTCADSYSFVFARDPHGLPPEVQRIVRQRQEELCLARQYRLIVHNGYCDGKSEANLL

>XP_032987299.1 retinol-binding protein 4 [Rhinolophus ferrumequinum]

MEWVWALVLLAVLGSARAERNCRVSSFRVKENFDKARFAGTWYAMAKKDPEGLFLQDNIIAEFSVDESGQ

MSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIIDTDYDTYAVQYSCRL

LNLDGTCADSYSFVFARNPHGLPPEVQRIVRRRQEELCLARQYRLIVHNGYCDRNSERNLL

Sequences Relabeled and Aligned

CLUSTAL multiple sequence alignment by MUSCLE (3.8)

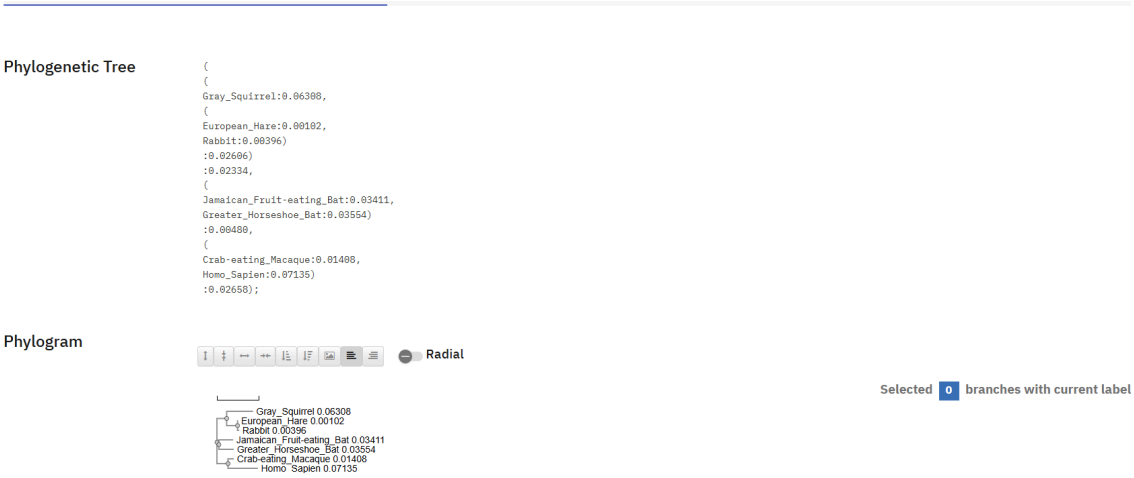
```
Gray_Squirrel               MEWVWALVLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAIAKKDPEGLFLQDNIV
European_Hare               MEWVWALVLLAALGSGRAERDCRVSSFRVKENFDKARFAGTWYAMAKKDPEGLFLQDNIV
Rabbit                      MEWVWALVLLAALGSGRAERDCRVSSFRVKENFDKARFAGTWYAMAKKDPEGLFLQDNIV
Jamaican_Fruit-eating_Bat   MEWVWALVLLAALGSARAERDCRVSSFRVKENFDKARFSGIWYAVAKKDPEGLFLQDNII
Greater_Horseshoe_Bat       MEWVWALVLLAVLGSARAERNCRVSSFRVKENFDKARFAGTWYAMAKKDPEGLFLQDNII
Crab-eating_Macaque         MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV
Homo_Sapien                 MNYSKIPAQVDL--RRQTERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV
                            *:: .   :     ..**.******************:* ***.**************:

Gray_Squirrel               AEFSVDEYGHMSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGND
European_Hare               AEFSVDENGHMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGND
Rabbit                      AEFSVDENGHMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGND
Jamaican_Fruit-eating_Bat   AEFSVDENGQMSATAKGRVTLLNNWEVCADMLGTFTDTEDPAKFKMKYWGVASFLQKGND
Greater_Horseshoe_Bat       AEFSVDESGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGND
Crab-eating_Macaque         AEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGND
Homo_Sapien                 AEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGND
                            ******* *.********* **.**.*****.**********************
Gray_Squirrel               DHWIIDTDYDTFALQYSCRLLNLDGTCADSYSFVFARDPNGLSPETRKLVRQRQEELCLD
European_Hare               DHWIIDTDYDTYAVQYSCRLLNFDGTCADSYSFVFSRDPHGLPPDVQKLVRQRQEELCLS
Rabbit                      DHWIIDTDYDTFAVQYSCRLLNFDGTCADSYSFVFSRDPHGLPPDVQKLVRQRQEELCLS
Jamaican_Fruit-eating_Bat   DHWIIDTDYDTYAVQYSCRLLNLDGTCADSYSFVFARDPHGLPPEVQRIVRQRQEELCLA
Greater_Horseshoe_Bat       DHWIIDTDYDTYAVQYSCRLLNLDGTCADSYSFVFARNPHGLPPEVQRIVRRRQEELCLA
Crab-eating_Macaque         DHWIIDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQRIVRQRQEELCLA
Homo_Sapien                 DHWIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLA
                            ****.*******.*.********.************.*.*.*.** *. .:.** .*******

Gray_Squirrel               RQYRWIEHNGE-------------------------------------------------
European_Hare               RQYRLIVHNGYCDDKSVRNLL---------------------------------------
Rabbit                      RQYRLIVHNGYCDDKSVRNLL---------------------------------------
Jamaican_Fruit-eating_Bat   RQYRLIVHNGYCDGKSEANLL-----------------------------------------
Greater_Horseshoe_Bat       RQYRLIVHNGYCDRNSERNLL---------------------------------------
Crab-eating_Macaque         RQYRLIVHNGYCDGRSERNLLRQYRLIVHNGYCDGRSERNLLRQYRLIVHNGYCDGRSER
Homo_Sapien                 RQYRLIVHNGYCDGRSERNLL---------------------------------------
                            **** * ***
```

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.

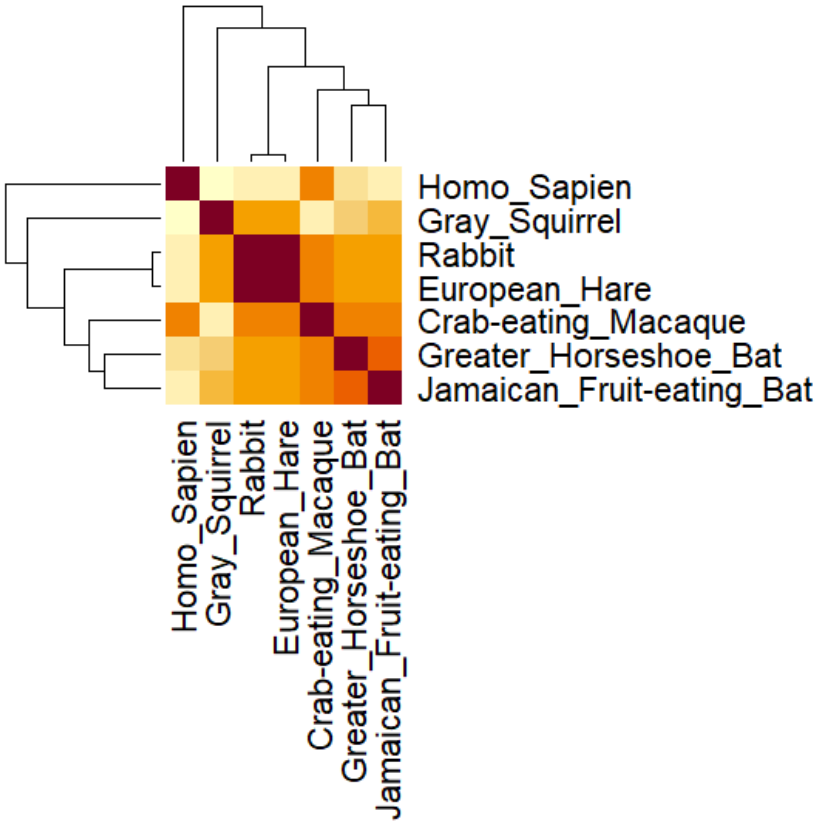
[Q6] Create a phylogenetic tree, using either a parsimony or distance-based approach. Bootstrapping and tree rooting are optional. Use “simple phylogeny” online from the EBI or any respected phylogeny program (such as MEGA, PAUP, or Phylip). Paste an image of your Cladogram or tree output in your report.

Phylogenetic Tree created using “simple phylogeny” online from EBI



[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).

PDB_ID	E.value	Identity	Method	Resolution	Source
1AQB	8.39e-22	93.333	X-ray	1.65	Sus scrofa domesticus
1BRP	8.36e-22	93.333	X-ray	2.50	Homo sapiens

1ERB	1.94e-71	76.712	X-ray	1.90	Bos taurus

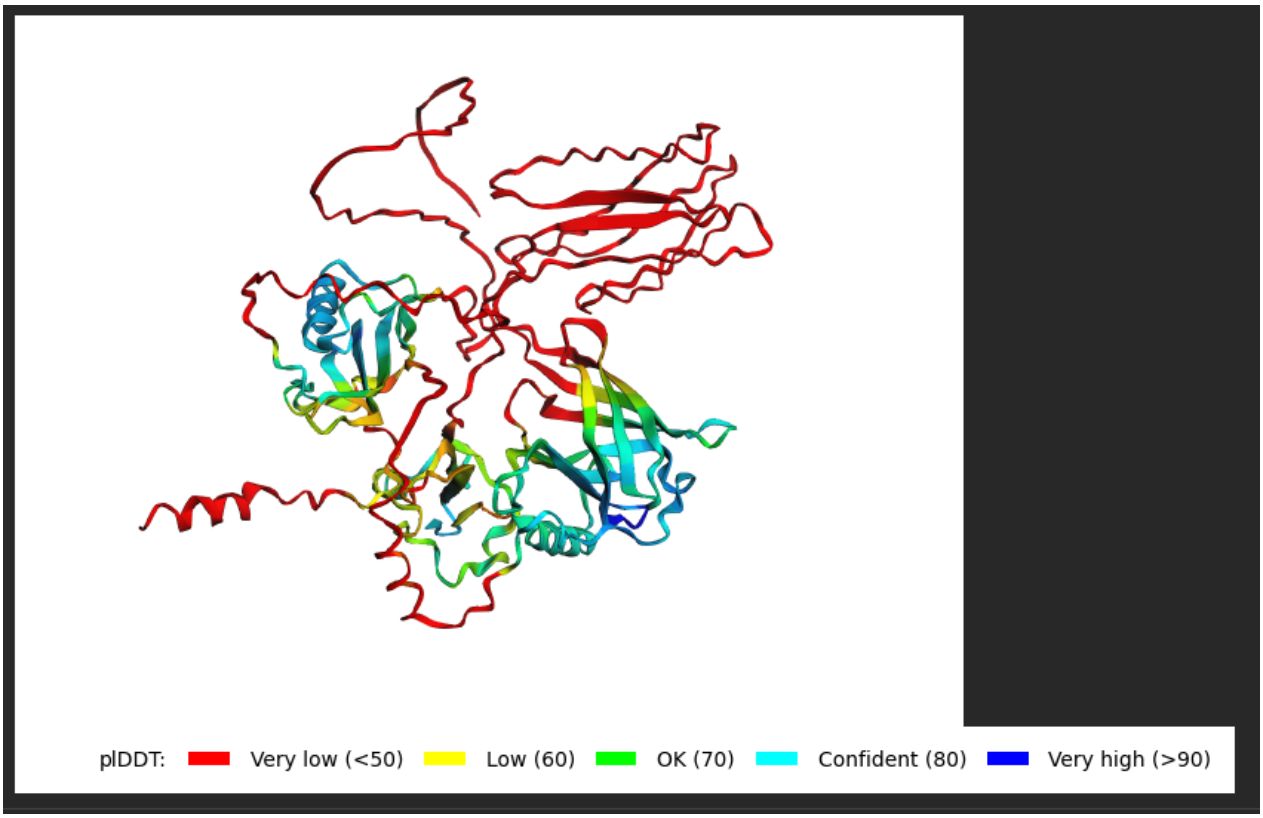
HINT: You can use a single sequence from your alignment or generate a consensus sequence from your alignment using the Bio3D function consensus(). The Bio3D functions blast.pdb(), plot.blast() and pdb.annotate() are likely to be of most relevance for completing this task. Note that the results of blast.pdb() contain the hits PDB identifier (or pdb.id) as well as Evalue and identity. The results of pdb.annotate() contain the other annotation terms noted above.

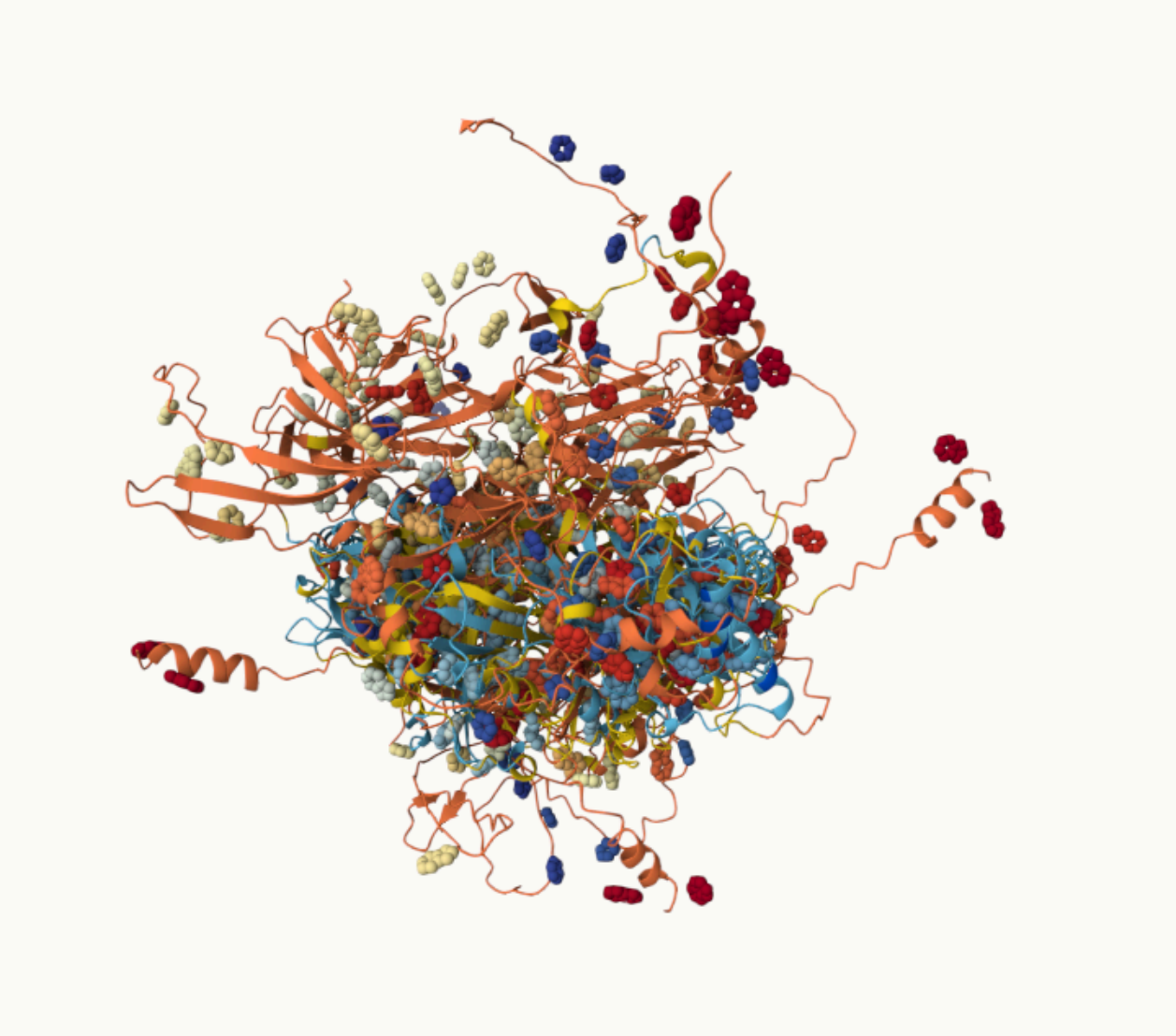
Note that if your consensus sequence has lots of gap positions then it will be better to use an original sequence from the alignment for your search of the PDB. In this case you could chose the sequence with the highest identity to all others in your alignment by calculating the row-wise maximum from your sequence identity matrix.

[**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 should highlight *conserved residues* that are likely to be functional as **spacefill** and the protein as **cartoon** colored by local alpha fold *pLDDT quality score*. You can determine conserved residues from the alignment generated by the AlphaFold server and use a conservation cutoff appropriate for the diversity of your protein alignment (e.g. between 60% and 99% conserved). Note that *pLDDT* 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.).


plDDT:
Very low (<50)
Low (60)
OK (70)
Confident (80)
Very high (>90)



[Q10] (i) Using your computed structure model (or your closest homologue of known structure from the PDB) predict and locate potential small molecule binding sites using the CASTpFold server (https://cfold.bme.uic.edu/castpfold/). Provide an image or screen-shot of your largest predicted pockets “negative volume” and provide it’s **area** and **volume**.

*Used 1AQB from PDB for binding site location using CASTpFold

Area (SA) ($Å^2$)	Volume (SA) ($Å^3$)
280.264	159.433

(ii) 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]”.

Not available as of 3/16/26

(iii) Briefly discuss (100 words max) the **druggability** of your novel protein based on:
- Presence of well-defined pockets (output of tools like CASTpFold), - Existence of known inhibitors for related proteins (your search of ChEMBEL), - Conservation of

binding sites across homologs (your conservation analysis in Q10),

- Potential therapeutic applications if this protein were targeted (you can use ChatGPT, Claude etc. backed up by your reading of the literature here).

I do not think that my novel protein is very druggable due to the pockets not being well defined, the largest pockets are only around ~280 A^2 in area and ~159 A^3 in volume.

Scoring Rubric: [60 total points available]

Q1 (4 points)

Protein name 1 Species 1 Accession number 1 Function known 1

Q2 (6 points)

Blast method 1 Database searched 1 Limits applied 1 Search output list (top hits) 1 Alignment of choice 1 Evalue and other alignment stats 1

Q3 (3 points)

Protein sequence of choice matches Subject above 1 Name in header 1 Species 1

Q4 (3 point)

Blastp output list with identities & Evalue 1 Top alignment shown with alignment statistics 1 Results indicates a “novel” gene found 1

Q5 (3 points)

MSA labeled with useful names 1 MSA trimmed appropriately (i.e. no gap overhangs) 1 Pasted MSA fits report page width (i.e. font, format) 1

Q6 (1 point)

Figure illustrates sequence clustering pattern 1

Q7 (10 points)

Heatmap figure included in report 5 Heatmap is legible (i.e. no labels obscured) 5

Q8 (10 points)

PDB identifiers from multiple species reported 5 Annotation of PDB source, resolution and technique 4 Annotation of Evalue and Sequence Identity 1

Q9 (10 points)

Structure figure provided 2 Uses white background for molecular figure 1 Figure of high resolution (i.e. not just snapshot) 1 Conserved residues as spacefill 3 Protein cartoon colored by pLDDT quality score 3

Q10 (10 points)

i) Binding site image, volume and area. 3 ii) Evidence of ChEMBEL searches 1 iii) Druggability discussion 6