BIMM-143: INTRODUCTION TO BIOINFORMATICS

The find-a-gene project assignment
https://bioboot.github.io/bimm143_S20/
Dr. Barry Grant

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.

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 the beginning of class **Tuesday May 5th** (05/05/20) at 12pm San Diego time. Note that these answers can be obtained very quickly (at best within 10 or 15 minutes), so if you don't succeed at first, just keep trying.

The complete assignment, including responses to all questions, is due **Friday June 5th** (06/05/20) at 12pm San Diego time.

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.

JTT009@ucsd.edu
A17197773

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. I will email you my decision; proceed with subsequent questions only after we are sure you have found a novel gene.

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 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: RBP4 (Retinol-binding protein 4)

Accession: P02753

Species: Homo Sapiens

Function Known: Retinol binding protein 4, also known as RBP4, is a transporter protein for retinol.

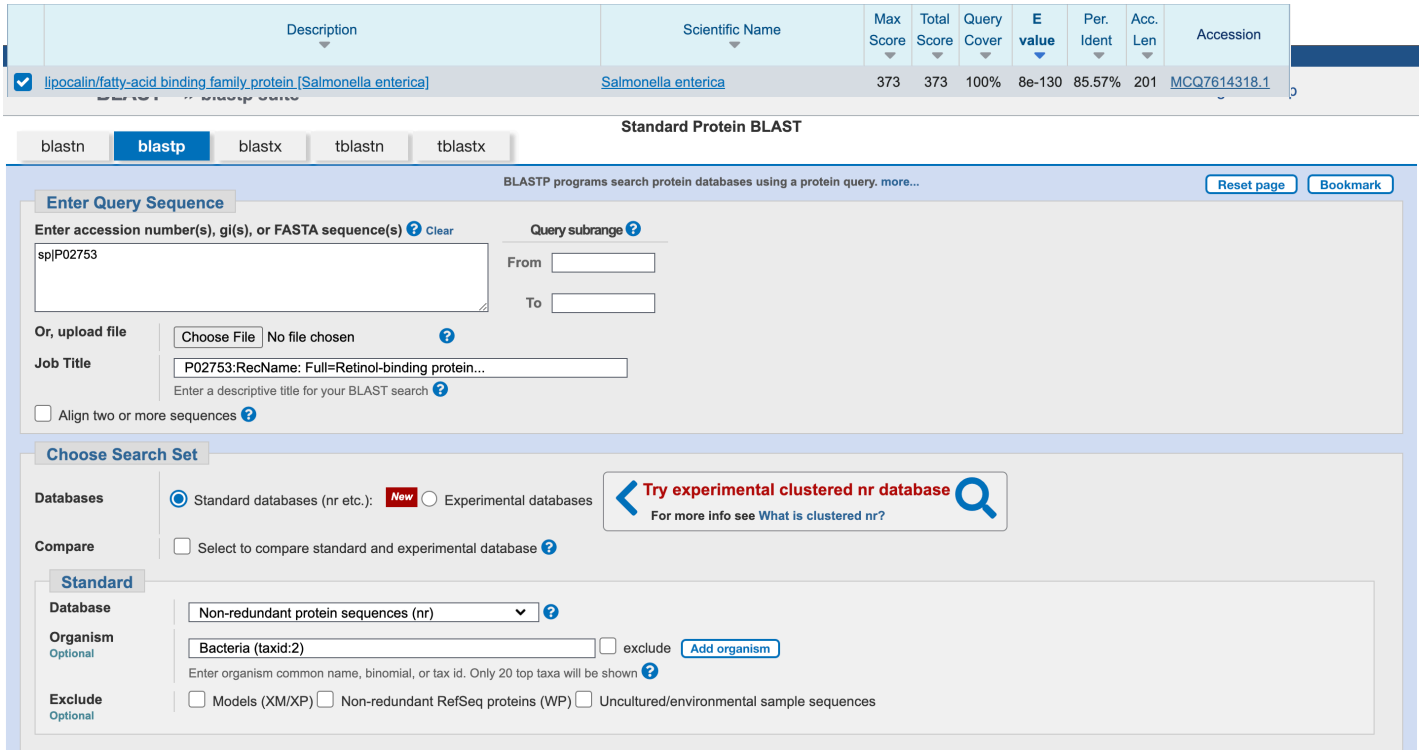
[Q2] Perform a BLAST search against a DNA database, such as a database consisting of genomic DNA or ESTs. The BLAST server can be at NCBI or elsewhere. Include details of the BLAST method used, database searched and any limits applied (e.g. Organism).

Method: BLAST search using the BLAST method "blastp" against the "nr" (non-redundant) protein database at NCBI. I set the organism limit to "Bacteria" to increase the chances of finding a novel gene.

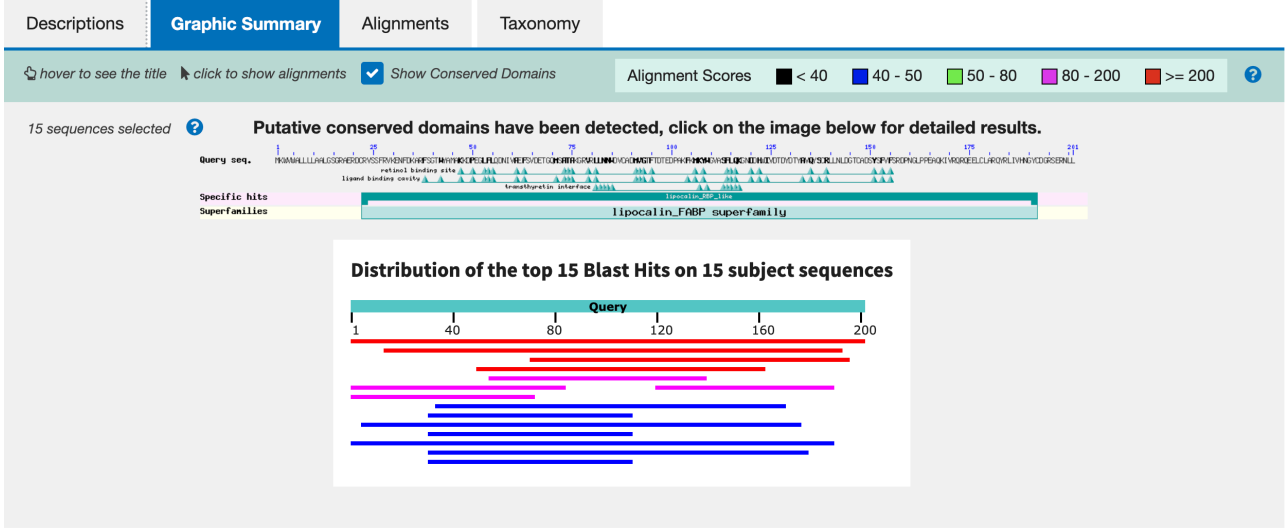
Database: Non-redundant protein sequences (nr)

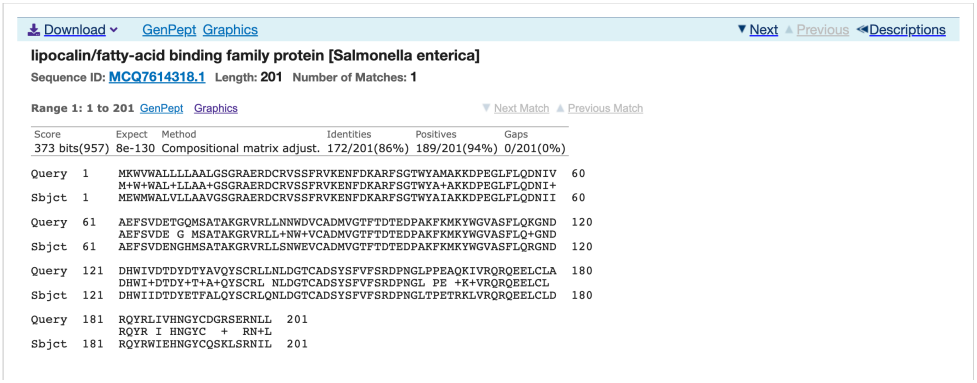
Organism: Bacteria (taxid:2)

Chosen Match:



Caption





Alignment statistics for match #1

Score	Expe	Method	Identities	Positives	Gaps
373	8E-1:	Compositional	172/201(	189/201(	0/201(

```
Query  1
MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV  60

M+W+WAL+LLAA+GSGRAERDCRVSSFRVKENFDKARFSGTWYA+AKKDPEGLFLQDNI+
Sbjct  1
MEWMWALVLLAAVGSGRAERDCRVSSFRVKENFDKARFSGTWYAIAKKDPEGLFLQDNII  60

Query  61
AEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGND  120
          AEFSVDE G
MSATAKGRVRLL+NW+VCADMVGTFTDTEDPAKFKMKYWGVASFLQ+GND
Sbjct  61
AEFSVDENGHMSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGND  120

Query  121
DHWIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLA  180
          DHWI+DTDY+T+A+QYSCRL NLDGTCADSYSFVFSRDPNGL PE
+K+VRQRQEELCL
Sbjct  121
DHWIIDTDYETFALQYSCRLQNLDGTCADSYSFVFSRDPNGLTPETRKLVRQRQEELCLD  180

Query  181  RQYRLIVHNGYCDGRSERNLL  201
             RQYR I HNGYC  +  RN+L
Sbjct  181  RQYRWIEHNGYCQSKLSRNIL  201
```

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.

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.

Chosen Sequence:

```
>MCQ7614318.1 lipocalin/fatty-acid binding family protein
[Salmonella enterica]
MEWMWALVLLAAVGSGRAERDCRVSSFRVKENFDKARFSGTWYAIAKKDPEGLFLQDNIIA
EFSVDENGH
MSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDHWIIDTDYET
FALQYSCRL
QNLDGTCADSYSFVFSRDPNGLTPETRKLVRQRQEELCLDRQYRWIEHNGYCQSKLSRNIL
```

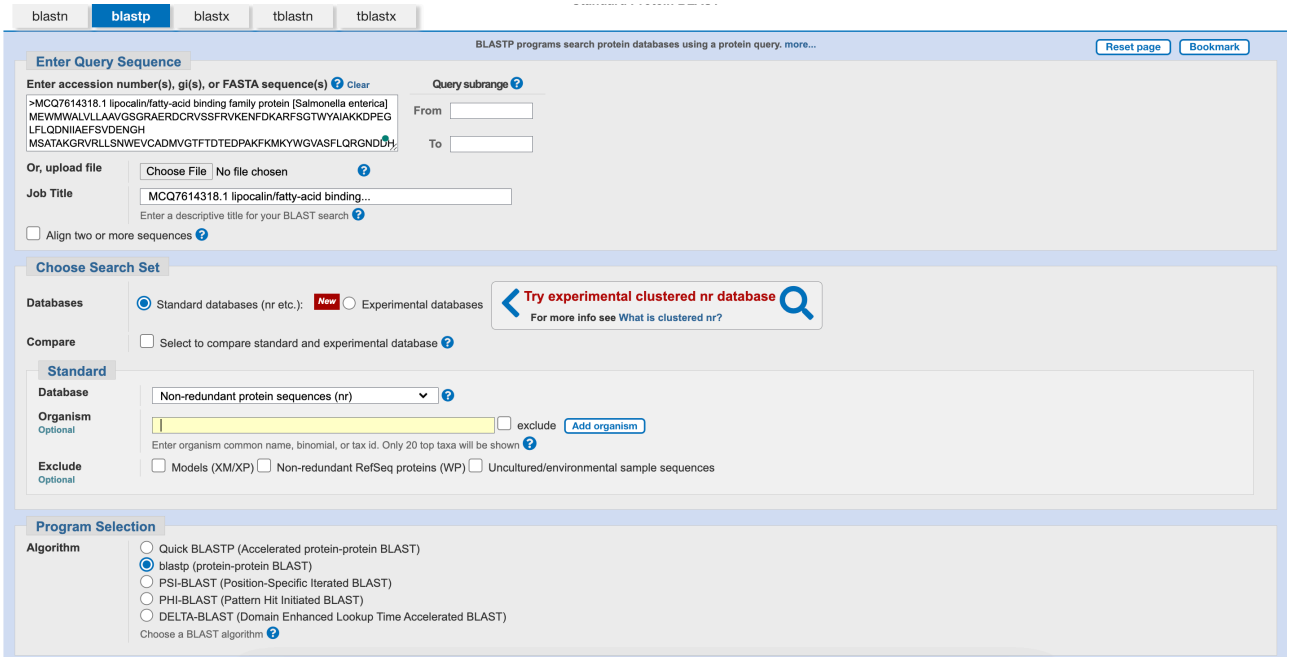
Here, tell me the name of the novel protein, and the species from which it derives. It is very unlikely (but still definitely possible) that you will find a novel gene from an organism such as *S. cerevisiae*, human or mouse, because those genomes have already been thoroughly annotated. It is more likely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

Name: lipocalin/fatty-acid binding family protein
Species: Salmonella enterica
Bacteria; Pseudomonadota; Gammaproteobacteria; Enterobacterales; Enterobacteriaceae; Salmonella

[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.

Details:
A BLASTP search against NR database (see setup in first screen-shot below) yielded a top hit result is to a novel protein from Jaculus Jaculus (Lesser Egyptian jerboa). See additional screen shots below for top hits and selected alignment details:



Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
lipocalin/fatty-acid binding family protein [Salmonella enterica]	Salmonella ente…	422	422	100%	4e-149	100.00%	201	MCQ7614318.1
retinol-binding protein 4 [Jaculus jaculus]	Jaculus jaculus	408	408	100%	2e-143	96.02%	201	XP_044986072.1
PREDICTED: retinol-binding protein 4 [Dipodomys ordii]	Dipodomys ordii	402	402	100%	3e-141	95.02%	201	XP_012889769.1
retinol-binding protein 4 isoform X1 [Mastomys coucha]	Mastomys coucha	400	400	100%	1e-140	94.53%	201	XP_031246073.1
retinol-binding protein 4 precursor [Rattus norvegicus]	Rattus norvegicus	400	400	100%	3e-140	94.03%	201	NP_037294.1
retinol-binding protein 4 [Arvicanthis niloticus]	Arvicanthis niloti…	400	400	100%	3e-140	94.53%	201	XP_034365668.1
plasma retinol binding protein 4 precursor [Castor fiber]	Castor fiber	399	399	100%	4e-140	94.53%	201	APD32941.1
retinol-binding protein 4 [Grammomys surdaster]	Grammomys sur…	399	399	100%	4e-140	94.03%	201	XP_028629352.1
retinol-binding protein 4 [Sciurus carolinensis]	Sciurus caroline…	399	399	100%	7e-140	93.53%	201	XP_047410179.1
retinol-binding protein 4 isoform X2 [Mastomys coucha]	Mastomys coucha	400	400	100%	9e-140	94.53%	246	XP_031246074.1
retinol-binding protein 4 isoform 2 precursor [Mus musculus]	Mus musculus	398	398	100%	1e-139	93.53%	201	NP_035385.1

The top result is to a protein from Jaculus Jaculus (Lesser Egyptian jerboa), see second screen shot below for alignment details:

Range 1: 1 to 201 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
422 bits(1085)	4e-149	Compositional matrix adjust.	201/201(100%)	201/201(100%)	0/201(0%)

```
Query  1    MEWMWALVLLAAVGSGRAERDCRVSSFRVKENFDKARFSGTWYAIAKKDPEGLFLQDNII  60
            MEWMWALVLLAAVGSGRAERDCRVSSFRVKENFDKARFSGTWYAIAKKDPEGLFLQDNII
Sbjct  1    MEWMWALVLLAAVGSGRAERDCRVSSFRVKENFDKARFSGTWYAIAKKDPEGLFLQDNII  60

Query  61   AEFSVDENGHMSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGND  120
            AEFSVDENGHMSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGND
Sbjct  61   AEFSVDENGHMSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGND  120

Query  121  DHWIIDTDYETFALQYSCRLQNLDGTCADSYSFVFSRDPNGLTPETRKLVRQRQEELCLD  180
            DHWIIDTDYETFALQYSCRLQNLDGTCADSYSFVFSRDPNGLTPETRKLVRQRQEELCLD
Sbjct  121  DHWIIDTDYETFALQYSCRLQNLDGTCADSYSFVFSRDPNGLTPETRKLVRQRQEELCLD  180

Query  181  RQYRWIEHNGYCQSKLSRNIL  201
            RQYRWIEHNGYCQSKLSRNIL
Sbjct  181  RQYRWIEHNGYCQSKLSRNIL  201
```

Download GenPept Graphics

retinol-binding protein 4 [Jaculus jaculus]

Sequence ID: XP_044986072.1 **Length:** 201 **Number of Matches:** 1

Range 1: 1 to 201 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
408 bits(1048)	2e-143	Compositional matrix adjust.	193/201(96%)	198/201(98%)	0/201(0%)

```
Query  1    MEWMWALVLLAAVGSGRAERDCRVSSFRVKENFDKARFSGTWYAIAKKDPEGLFLQDNII  60
            MEWMWALVLLAA+GSGRAERDCRVSSFRVKENFDKARFSGTWYAIAKKDPEGLFLQDNII
Sbjct  1    MEWMWALVLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAIAKKDPEGLFLQDNII  60

Query  61   AEFSVDENGHMSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGND  120
            AEF+VDENGHMSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQ+GND
Sbjct  61   AEFAVDENGHMSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGND  120

Query  121  DHWIIDTDYETFALQYSCRLQNLDGTCADSYSFVFSRDPNGLTPETRKLVRQRQEELCLD  180
            DHWIIDTDY+T+ALQYSCRL NLDGTCADSYSFVFSRDPNGL PETRKLVRQRQEELCLD
Sbjct  121  DHWIIDTDYDTYALQYSCRLLNLDGTCADSYSFVFSRDPNGLPPETRKLVRQRQEELCLD  180

Query  181  RQYRWIEHNGYCQSKLSRNIL  201
            RQYRWIEHNGYCQSKLS NIL
Sbjct  181  RQYRWIEHNGYCQSKLSGNIL  201
```

[**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.

[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] 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).

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] Generate a molecular figure of one of your identified PDB structures using the **NGL viewer** online (or **VMD/PyMol)**. 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?

[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?

Scoring Rubric:

[45 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 (4 points)	
Structure figure provided	2
Uses white background for molecular figure	1
Figure of high resolution (i.e. not just snapshot)	1
Q10 (1 point)	
Evidence of ChEMBEL searches	1