BIMM-143: INTRODUCTION TO BIOINFORMATICS

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

zashir@ucsd.edu

A16125522

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: Kinesin family member 11 (KIF11)

Accession: NP_004514

Species: Homo sapiens

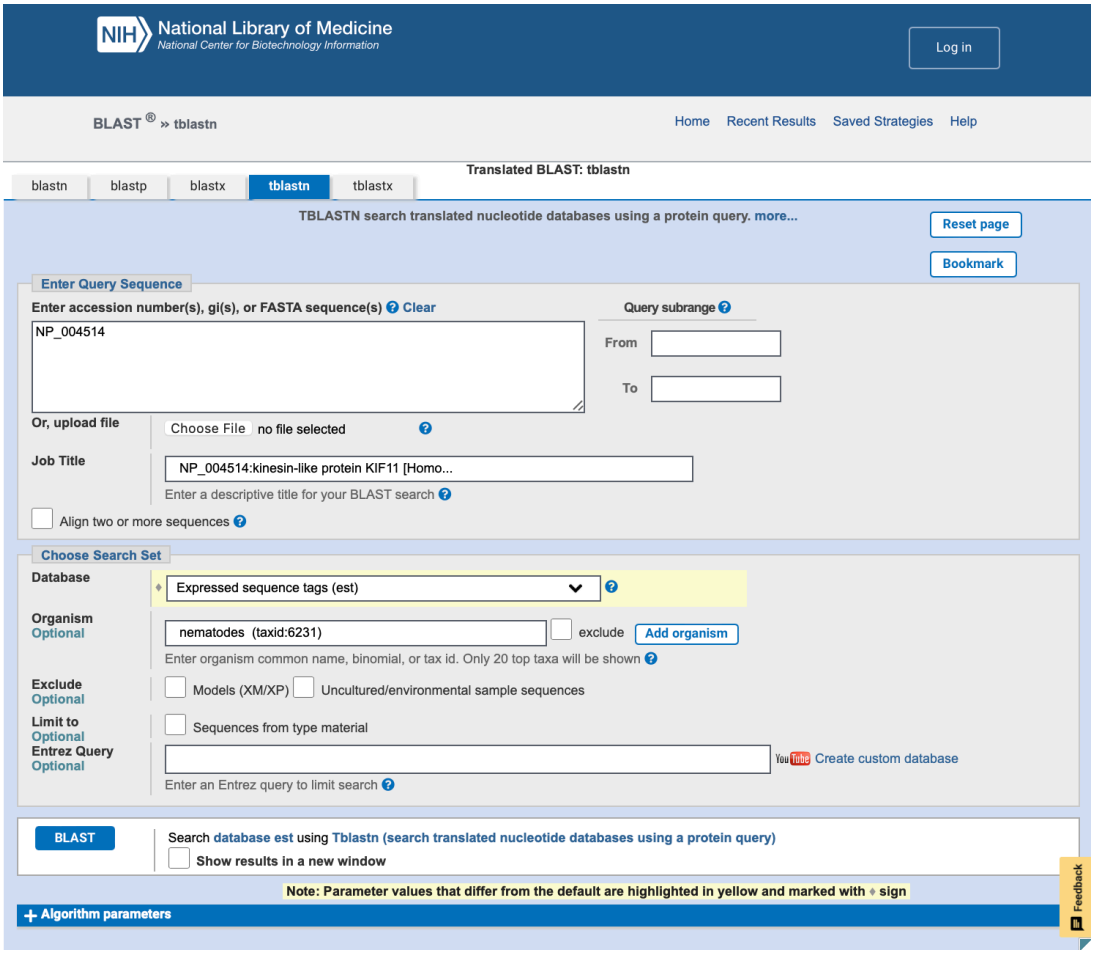
Function Known: Encoding a motor protein in which it belongs to the kinesin-like protein family. It is also known that it is involved in many kinds of the dynamics of spindle. Functions in chromosome positioning, centrosome separation and bipolar spindle establishment in mitosis of the cell.

[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) search against nematode ESTs

Database: Expressed Sequence Tags (est)

Organism: Nematodes (Taxid: 6231)



Chosen match: Accession JK291331.1, a 757 base pair clone from *Meloidogyne incognita*. See below for alignment details.

Query ID
NP_004514.2
Description
kinesin-like protein KIF11 [Homo sapiens]
Molecule type
amino acid
Query Length
1056
Other reports

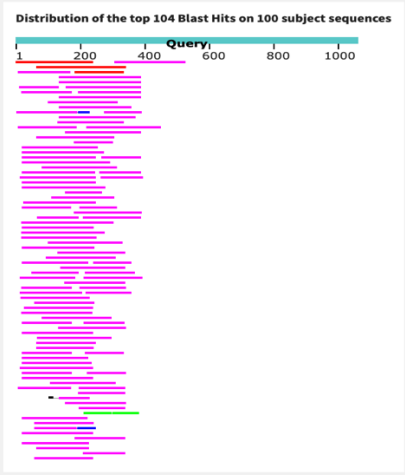
Filter Reset

Descriptions | **Graphic Summary** | Alignments | Taxonomy

hover to see the title *click to show alignments*

Alignment Scores < 40 | 40 - 50 | 50 - 80 | 80 - 200 | >= 200

100 sequences selected

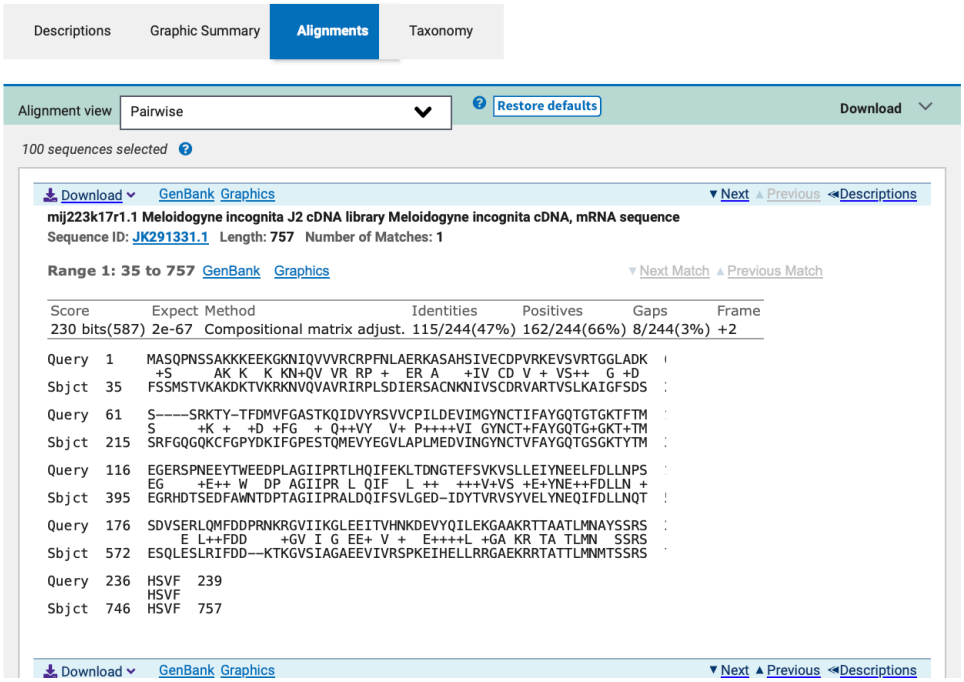


Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Select columns Show 100

select all *100 sequences selected* GenBank Graphics

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
mij223k17r1.1 Meloidogyne incognita J2 cDNA library Meloidogyne …	Meloidogy…	230	230	22%	2e-67	47.13%	757	JK291331.1
Pd_5pr_67L20 Panagrolaimus davidi 4 degree Panagrolaimus davi…	Panagrola…	206	206	26%	2e-58	45.36%	809	JZ645761.1
CJ983801 Bursaphelenchus xylophilus mixed-stage library Bursap…	Bursaphel…	200	200	14%	5e-58	64.47%	472	CJ983801.1
mij211a09r1.1 Meloidogyne incognita J2 cDNA library Meloidogyne …	Meloidogy…	196	196	23%	4e-55	41.57%	776	JK282786.1
mij216g17r1.1 Meloidogyne incognita J2 cDNA library Meloidogyne…	Meloidogy…	195	195	23%	5e-55	41.57%	760	JK287044.1
mij207m19r1.1 Meloidogyne incognita J2 cDNA library Meloidogyn…	Meloidogy…	186	186	21%	5e-52	43.53%	738	JK280452.1
kk03g09.y1 Ascaris suum female ovary Ascaris suum cDNA 5' simil…	Ascaris su…	173	173	18%	7e-48	46.67%	624	BQ095952.1
mij212p14r1.1 Meloidogyne incognita J2 cDNA library Meloidogyne…	Meloidogy…	173	173	24%	3e-47	43.02%	783	JK284117.1
Ls_af1_23b11_T7 Litomosoides sigmodontis adult female 1 (high …	Litomosoi…	170	170	20%	8e-47	44.95%	620	DN557730.1
mij201g24r1.1 Meloidogyne incognita J2 cDNA library Meloidogyne…	Meloidogy…	172	172	21%	1e-46	40.53%	839	JK275977.1
HTAB-aab12f01.b1 Heterorhabditis_bacteriophora_HTAB2_EST H…	Heterorha…	167	210	21%	6e-46	46.07%	696	ES742822.1
Pd_5pr_69H21 Panagrolaimus davidi 4 degree Panagrolaimus davi…	Panagrola…	167	167	22%	3e-45	43.98%	755	JZ646412.1
BJ101792 unpublished oligo-capped cDNA library, C. elegans L1 st…	Caenorha…	161	161	19%	9e-44	46.63%	606	BJ101792.1
Pd_5pr_87P16 Panagrolaimus davidi 4 degree Panagrolaimus davi…	Panagrola…	162	162	21%	3e-43	39.20%	770	JZ652958.1
HAF_01159 Heterodera avenae female adult Library Heterodera av…	Heteroder…	156	156	22%	1e-41	41.95%	731	JZ145890.1
HTAB-aae80a09.b1 Heterorhabditis_bacteriophora_HTAB2_EST H…	Heterorha…	156	156	22%	3e-41	42.21%	740	EX913685.1
CELK050FYF Yuji Kohara unpublished cDNA Caenorhabditis elega…	Caenorha…	150	150	11%	5e-41	63.33%	360	D67381.1
mij225p03r1.1 Meloidogyne incognita J2 cDNA library Meloidogyne…	Meloidogy…	154	154	22%	7e-41	35.02%	713	JK292508.1



Alignment details:

```
>gb|JK291331.1| mij223k17r1.1 Meloidogyne incognita J2 cDNA library
Meloidogyne incognita
cDNA, mRNA sequence.
Length=757

 Score = 230 bits (587),  Expect = 2e-67, Method: Compositional matrix
adjust.
 Identities = 115/244(47%, Positives = 162/244(66%, Gaps = 8/244(3%)
 Frame = +2

Query  1    MASQPNSSAKKKEEKGKNIQVVVRCRPFNLAERKASAHSIVECDPVRKEVSVRTGGLADK  60
             +S     AK K  K KN+QV VR RP +  ER A   +IV CD V + VS++  G +D
Sbjct  35   FSSMSTVKAKDKTVKRKNVQVAVRIRPLSDIERSACNKNIVSCDRVARTVSLKAIGFSDS  214

Query  61   S----SRKTY-TFDMVFGASTKQIDVYRSVVCPILDEVIMGYNCTIFAYGQTGTGKTFTM  115
             S     +K +  +D +FG  + Q++VY  V+ P++++VI GYNCT+FAYGQTG+GKT+TM
Sbjct  215  SRFGQGQKCFGPYDKIFGPESTQMEVYEGVLAPLMEDVINGYNCTVFAYGQTGSGKTYTM  394

Query  116  EGERSPNEEYTWEEDPLAGIIPRTLHQIFEKLTDNGTEFSVKVSLLEIYNEELFDLLNPS  175
             EG    +E++ W  DP AGIIPR L QIF  L ++  +++V+VS +E+YNE++FDLLN +
Sbjct  395  EGRHDTSEDFAWNTDPTAGIIPRALDQIFSVLGED-IDYTVRVSYVELYNEQIFDLLNQT  571

Query  176  SDVSERLQMFDDPRNKRGVIIKGLEEITVHNKDEVYQILEKGAAKRTTAATLMNAYSSRS  235
                E L++FDD    +GV I G EE+ V +  E++++L +GA KR TA TLMN  SSRS
Sbjct  572  ESQLESLRIFDD--KTKGVSIAGAEEVIVRSPKEIHELLRRGAEKRRTATTLMNMTSSRS  745

Query  236  HSVF  239
             HSVF
Sbjct  746  HSVF  757
```

[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:

>M. incognita protein (sequence taken from BLAST result)

FSSMSTVKAKDKTVKRKNVQVAVRIRPLSDIERSACNKNIVSCDRVARTVSLKAIGFSDS
SRFGQGQKCFGPYDKIFGPESTQMEVYEGVLAPLMEDVINGYNCTVFAYGQTGSGKTYTM
EGRHDTSEDFAWNTDPTAGIIPRALDQIFSVLGEDIDYTVRVSYVELYNEQIFDLLNQT
ESQLESLRIFDDKTKGVSIAGAEEVIVRSPKEIHELLRRGAEKRRTATTLMNMTSSRS
HSVF

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: Meloidogyne incognita

Species: Meloidogyne incognita

Eukaryota; Metazoa; Ecdysozoa; Nematoda; Chromadorea; Rhabditida; Tylenchina; Tylenchomorpha; Tylenchoidea; Meloidogynidae; Meloidogyninae; Meloidogyne; Meloidogyne incognita group.

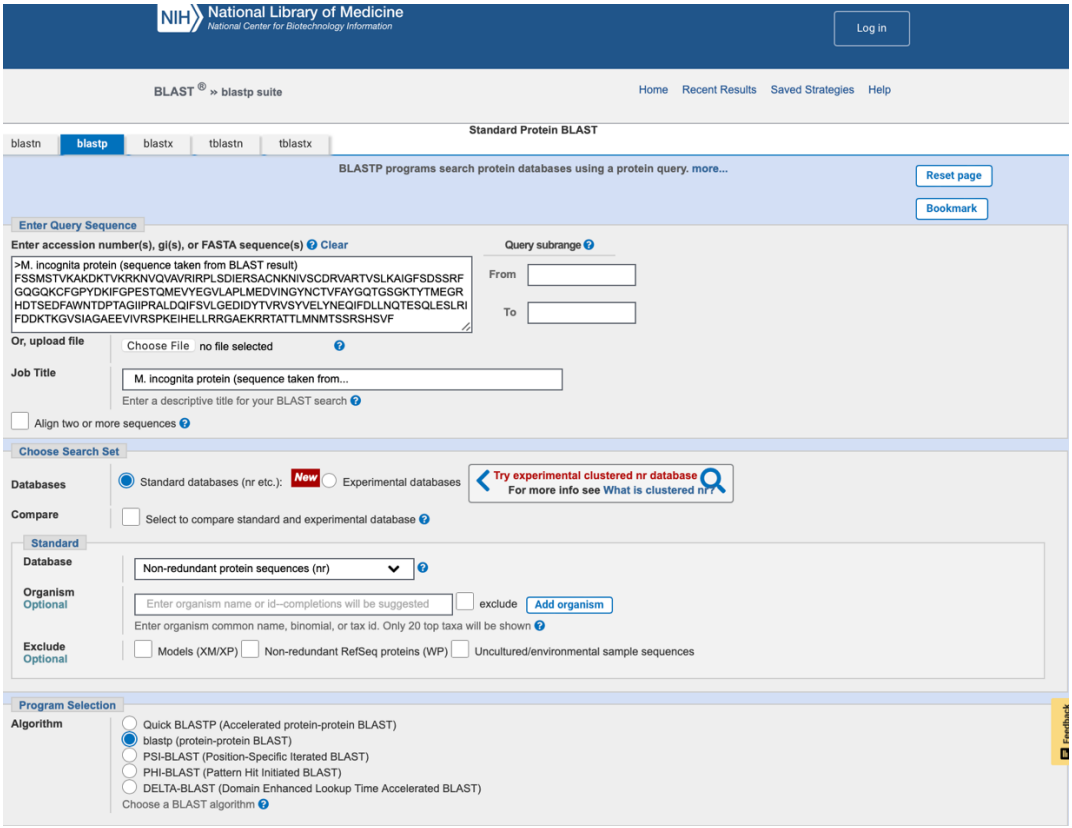
[Q4] Prove that this gene, and its corresponding protein, are novel. For the purposes of this project, “novel” is defined as follows. Take the protein sequence (your answer to [Q3]), and use it as a query in a blastp search of the nr database at NCBI.

- 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 protein from *Meloidogyne enterolobii* (nematodes).

See additional screen shots below for top hits and selected alignment details:



The top result is to a protein from *Meloidogyne enterolobii* (nematodes)see second screen shot below for alignment details:

Query Length

241

Other reports

Distance tree of results Multiple alignment MSA viewer

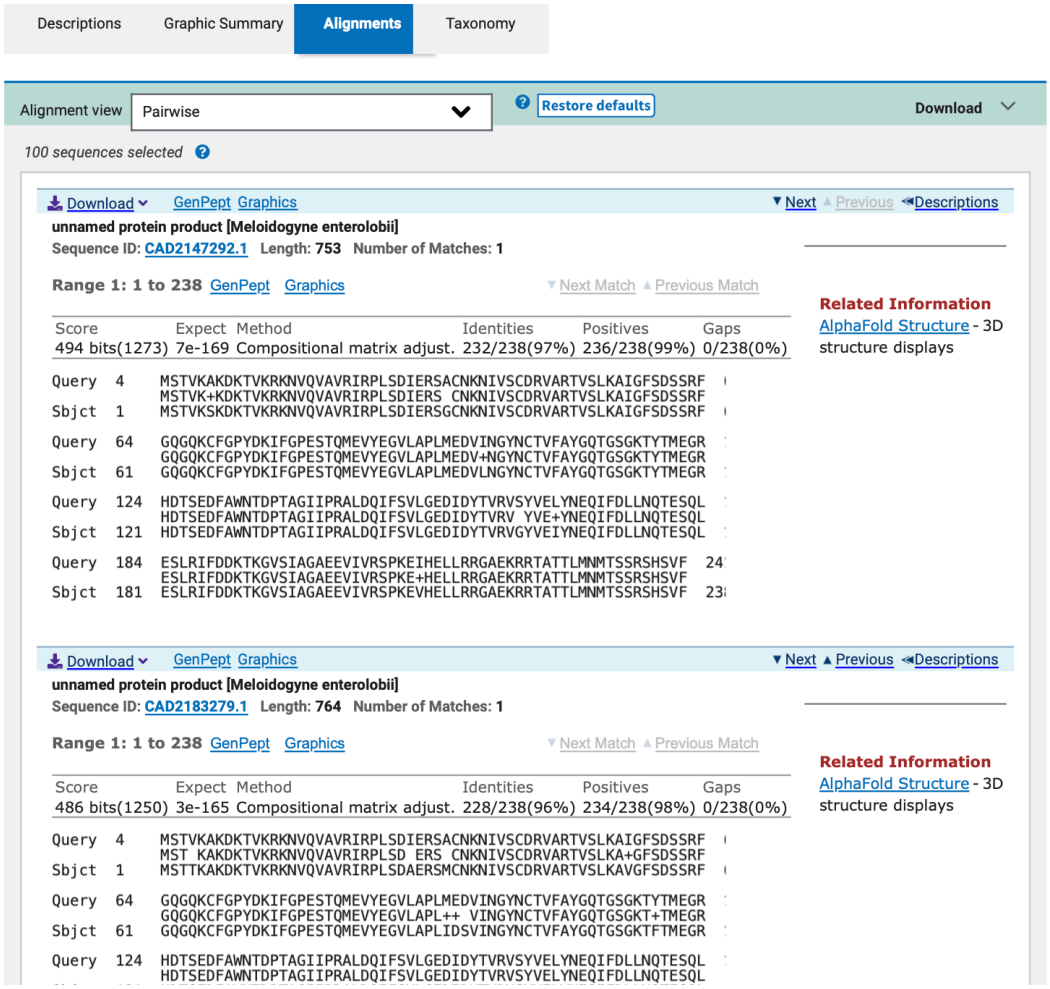
Compare these results against the new Clustered nr database BLAST

Descriptions Graphic Summary Alignments Taxonomy

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
unnamed protein product [Meloidogyne enterolobii]	Meloidog…	494	494	98%	7e-169	97.48%	753	CAD2147292.1
unnamed protein product [Meloidogyne enterolobii]	Meloidog…	486	486	98%	3e-165	95.80%	764	CAD2183279.1
unnamed protein product [Meloidogyne enterolobii]	Meloidog…	480	480	98%	5e-163	94.12%	764	CAD2169206.1
unnamed protein product [Meloidogyne enterolobii]	Meloidog…	463	463	98%	6e-162	92.44%	376	CAD2147351.1
unnamed protein product [Meloidogyne enterolobii]	Meloidog…	399	399	98%	3e-136	80.33%	390	CAD2124899.1
unnamed protein product [Meloidogyne enterolobii]	Meloidog…	391	391	98%	9e-135	79.08%	290	CAD2208809.1
unnamed protein product [Meloidogyne enterolobii]	Meloidog…	392	392	98%	5e-133	78.66%	439	CAD2177065.1
unnamed protein product [Meloidogyne enterolobii]	Meloidog…	391	391	98%	1e-122	78.24%	1504	CAD2156593.1
Kinesin-like protein [Meloidogyne graminicola]	Meloidog…	321	321	98%	2e-103	67.70%	570	KAF7629873.1
kinesin motor domain-containing protein [Ditylenchus destructor]	Ditylench…	278	278	89%	3e-85	61.26%	737	KAI1728869.1
kinesin motor domain-containing protein [Ditylenchus destructor]	Ditylench…	273	273	89%	3e-83	60.09%	739	KAI1711961.1
kinesin motor domain-containing protein [Ditylenchus destructor]	Ditylench…	273	273	89%	6e-83	60.55%	735	KAI1721682.1



[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