BIMM-143: INTRODUCTION TO BIOINFORMATICS

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

gag002@ucsd.edu
A16745338

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.

Protein name: RBP4
Species: Homo Sapien
Accession number: AF025334
Function known: transports vitamin A, protects retinol from oxidation, facilitates retinol uptake, regulates glucose metabolism, influences energy homeostasis

[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).

Blast method: tblastn search against homo sapiens
Database searched: Expressed Sequence Tags (est)
Organism Excluded: Homo Sapien (Taxid: 9606)

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.

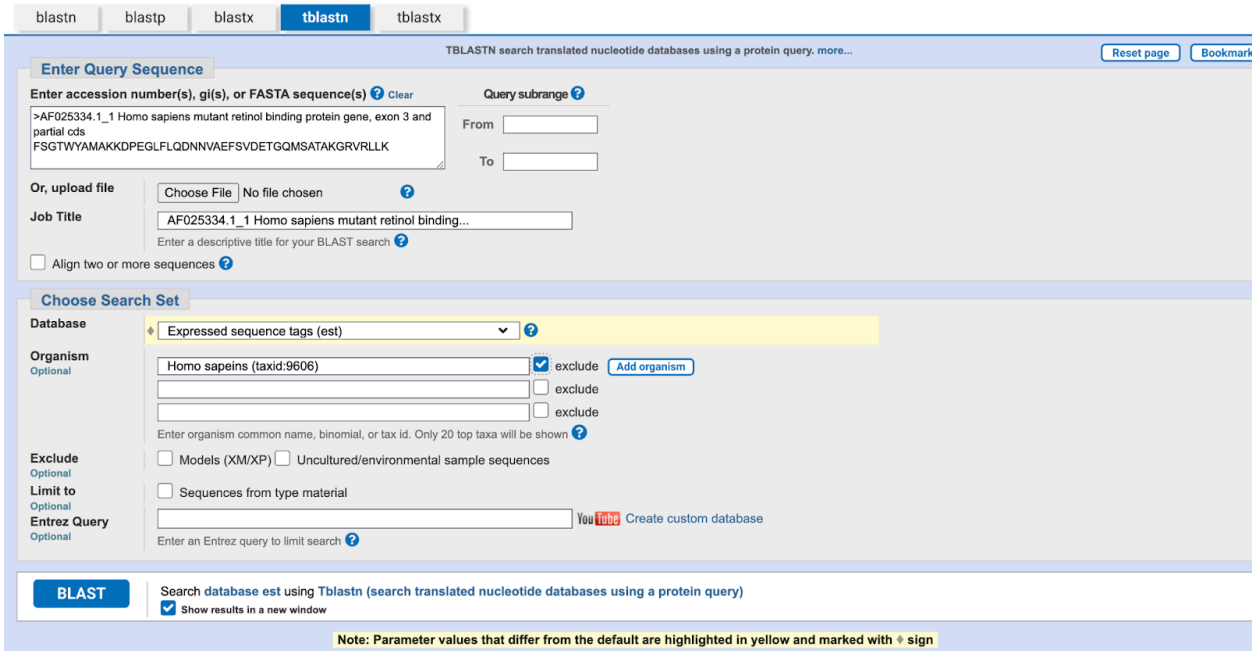


Image 1: output of tblastn search

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.

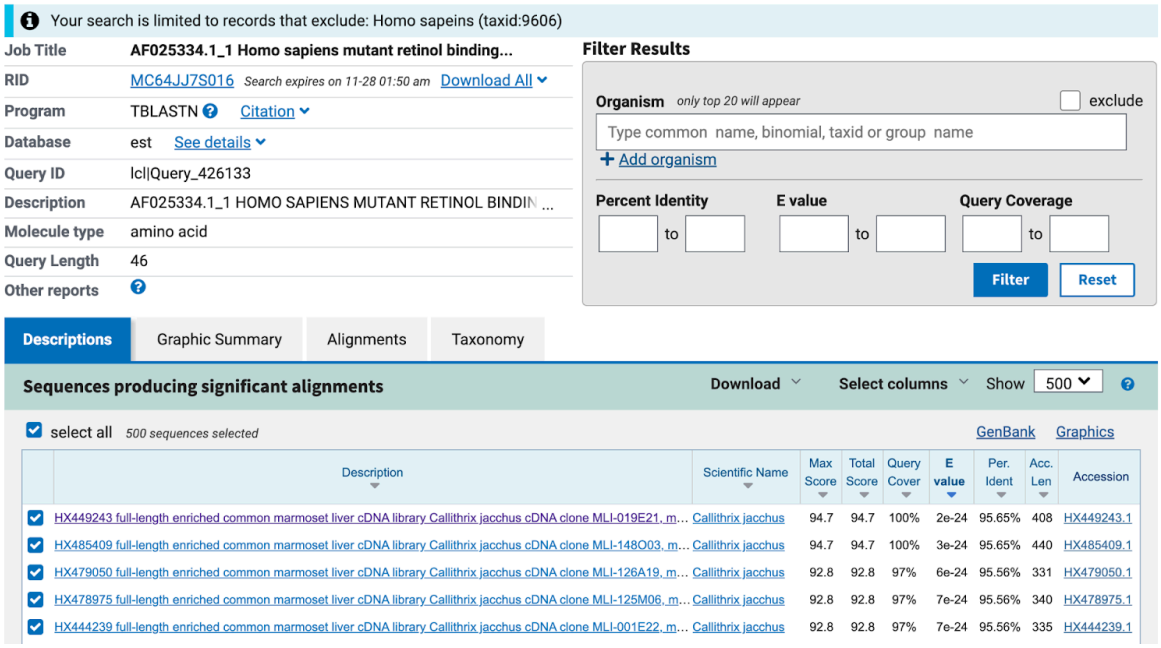


Image 2: Top 5 of tblastn search

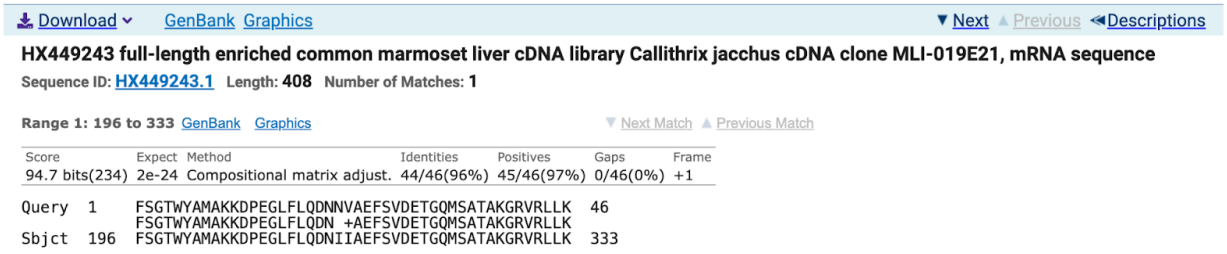


Image 3: Top match information

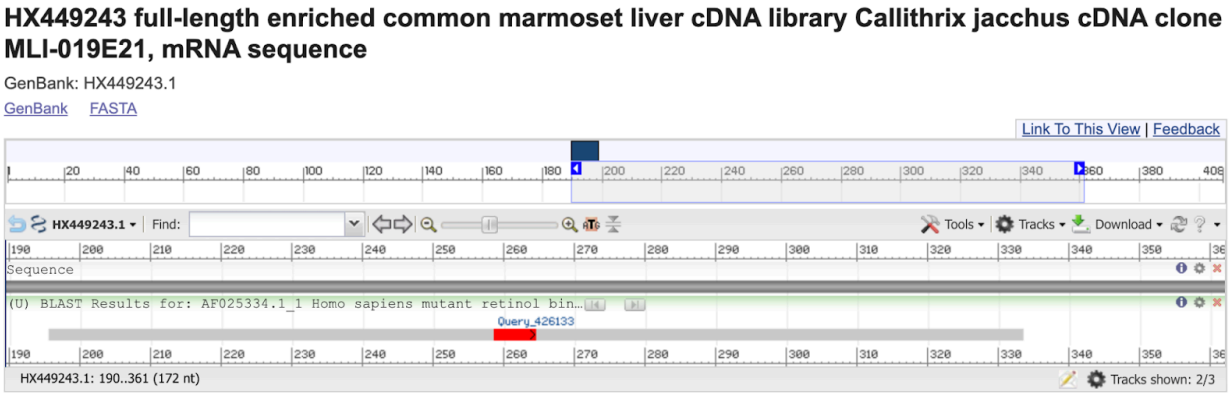


Image 4: Graphics of top match to protein

Alignment Details:

```
HX449243 full-length enriched common marmoset liver cDNA library
Callithrix jacchus cDNA clone MLI-019E21, mRNA sequence
Sequence ID: HX449243.1 Length: 408
```

Score	Expect	Method	Identities	Positives	Gaps	Frame
94.7 bits(234)	2e-24	Compositional matrix adjust.	44/46(96%)	45/46(97%)	0/46(0%)	+1

```
Query  1    FSGTWYAMAKKDPEGLFLQDNNVAEFSVDETGQMSATAKGRVRLLK  46
            FSGTWYAMAKKDPEGLFLQDN +AEFSVDETGQMSATAKGRVRLLK
Sbjct  196  FSGTWYAMAKKDPEGLFLQDNIIAEFSVDETGQMSATAKGRVRLLK  333
```

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:
>HX449243.1_1 full-length enriched common marmoset liver cDNA library Callithrix jacchus cDNA clone MLI-019E21, mRNA sequence
GLPSSTRARTLQPGLLAALLLVGVLLGKMKWVWALLLLAVLGISRAERDCRVSSFRVKEN
FDKARFSGTWYAMAKKDPEGLFLQDNIIAEFSVDETGQMSATAKGRVRLLKSVAARVAAL
FEFQGLPRALPADRHV

Name: Callithrix jacchus
Species: Callithrix jacchus
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
Platyrrhini; Cebidae; Callitrichinae; Callithrix; Callithrix.

[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.

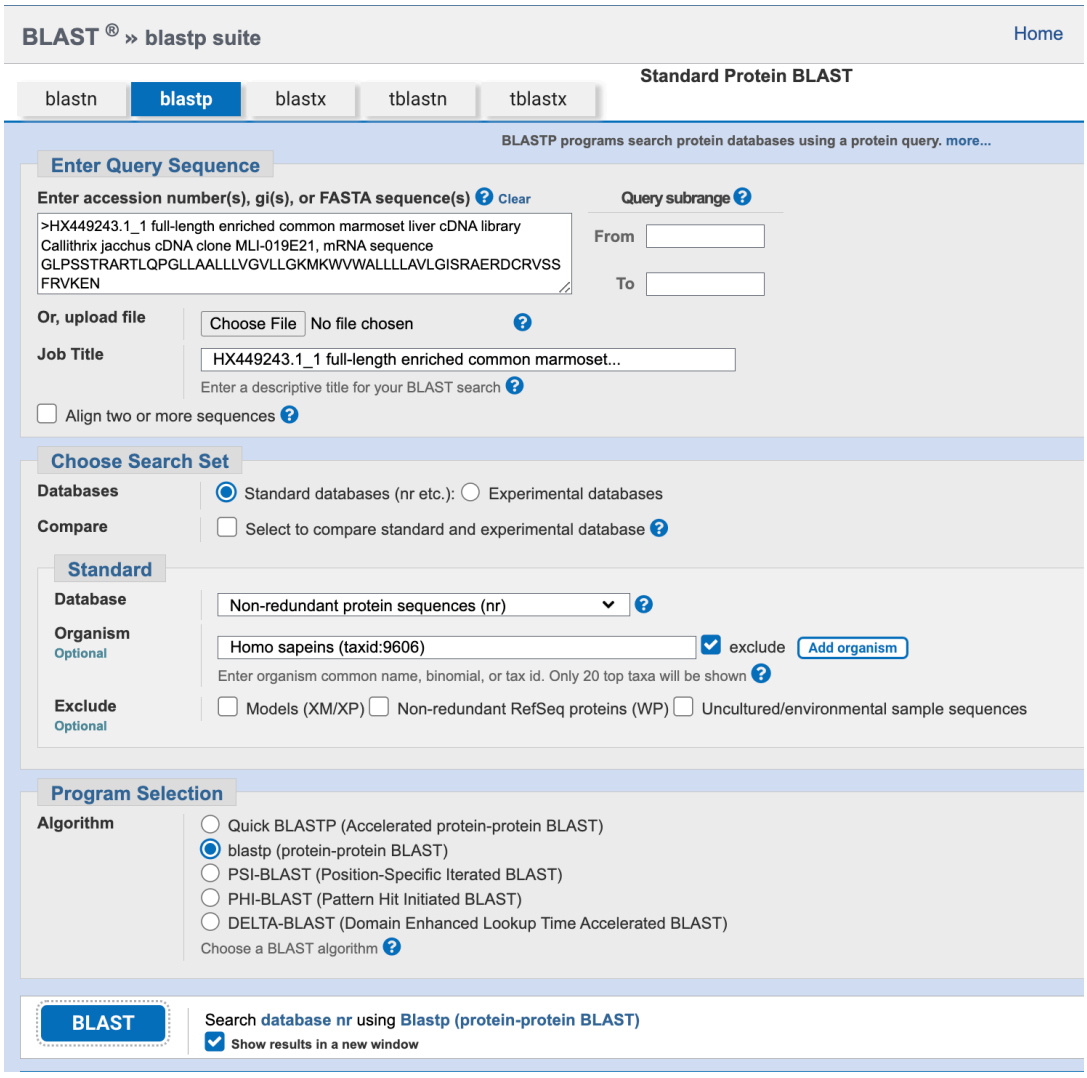


Image 5: blastp search

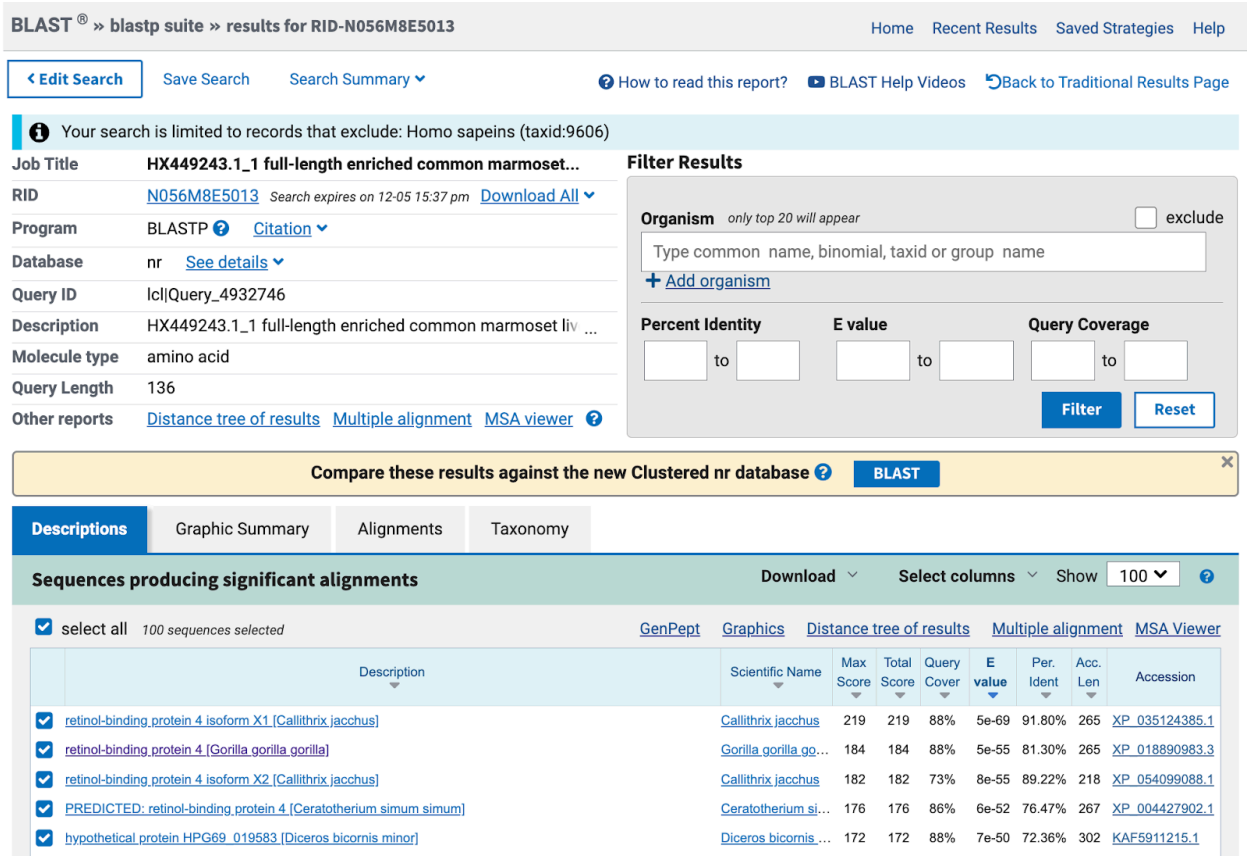


Image 6: Top 5 matches to blastp novel protein search results

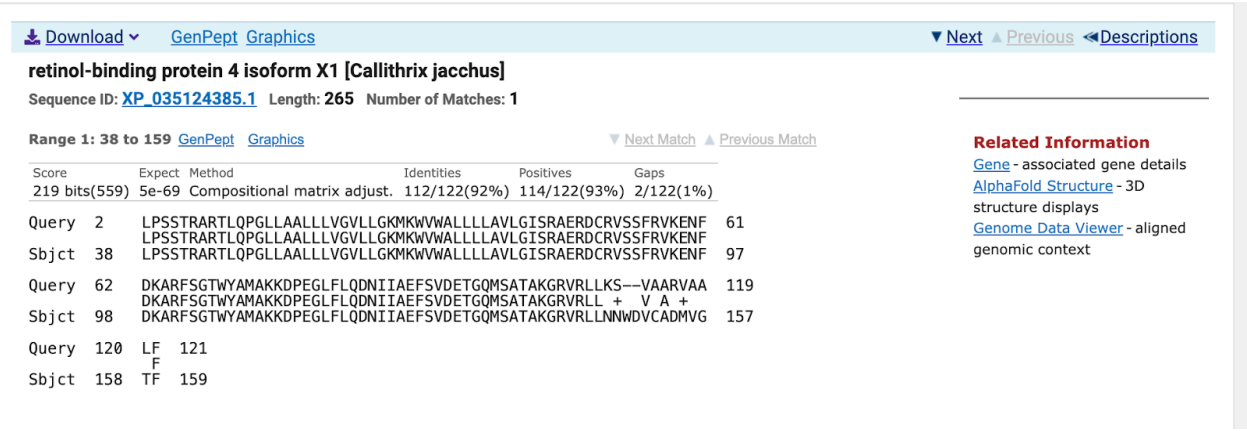


Image 4: First match to novel protein details

[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.

Re-labeled sequences for alignment:

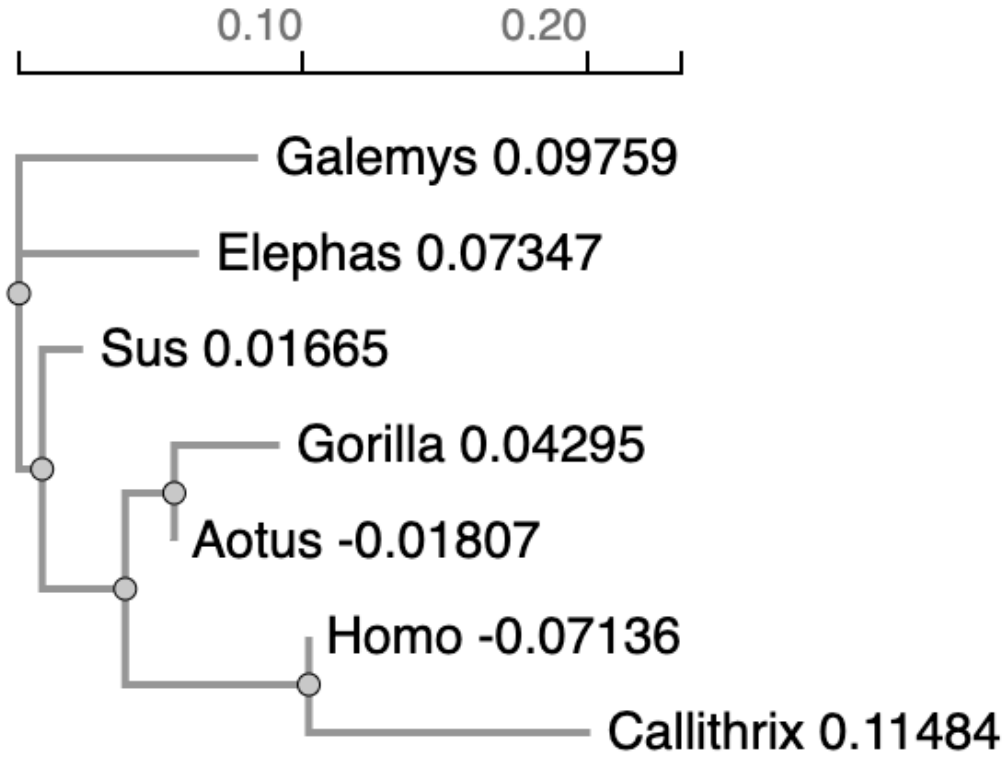
```
>Homo sapiens mutant retinol binding protein gene
FSGTWYAMAKKDPEGLFLQDNNVAEFSVDETGQMSATAKGRVRLLK

>Callithrix jacchus cDNA clone MLI-019E21, mRNA sequence
GLPSSTRARTLQPGLLAALLLVGVLLGKMKWVWALLLLAVLGISRAERDCRVSSFRVKEN
FDKARFSGTWYAMAKKDPEGLFLQDNIIAEFSVDETGQMSATAKGRVRLLKSVAARVAAL
FEFQGLPRALPADRHV

>Sus scrofa retinol binding protein 4, partial
RSKMEWVWALVLLAALGSAQAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDE
NGHMSATAKGRVRLLNNWDVCADMVGTFTDTEN

>Gorilla Gorilla Gorilla retinol-binding protein 4
MQAPPAPPLRSFTPRGYESATPSPRRYKAAERPRRAGLPRSTRARTRRPGLRAVPLPVGGFLGKMKWVWA
LLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAK
GRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRLLNLDGT
CADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL

>Aotus nancymaae retinol-binding protein 4 isoform X2
MKWVWALLLLAVLGSSRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQ
MSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRL
LNLDGTCADSYSFVFSRDPNGLPPEAQRIIRQRQEELCLARQYRLIVHNGYCDGKSERNLL

>Galemys pyrenaicus Retinol-binding protein 4
MQVLARGPRHLPLGLSPRAVTKARPPHPGAIKLPGGPRGALAQLLHARGDAGPGLRASRGGERRRAGCGS
RGRAVAQGRRPGAHGARFPQGGLLGRMEWVWALVLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYA
MAKKDPEGLFLQDNIITEFSVDQHGQMSATAKGRVRLLNSWDVCADMVGTFTDTEDPAKFKMKYWGVASF
LQKGNDDHWIIDTDYDTYAVQYSCRLQNLDGTCADSYSFIFSRDPNGLPPEAQRIVRRRQEELCLARQYR
LIAHNGECPGSGPRAGGQRGTFHKAVDR

>Elephas maximus indicus retinol-binding protein 4 isoform X1
MGKAALRWSGCQALIAARFPQGGLLGRMEWMWALVLLAALGSGRAERDCRVSSFRVKENFDKTRFSGTWY
AMAKKDPEGLFLQDNIIAEFSVDESGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVAS
```

FLQKGNDDHWIIDTDYDTYAVQYSCRLLNLDGTCADSYSFIFARDPYGLPPEVQKLVRQRQEELCLARQY
RMIVHNGYCDGKSEGHVL

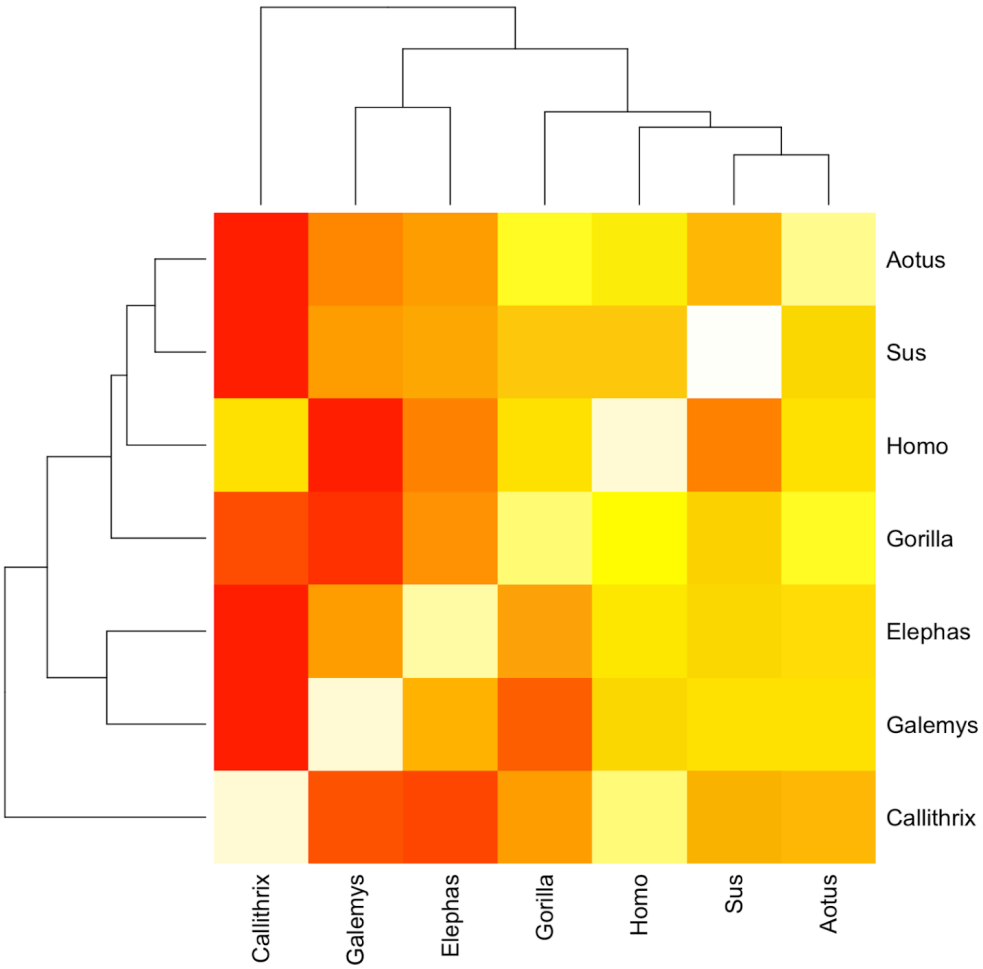
Alignment:

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


Galemys         MQVLARGPRHLPLGLSPRAVTKARPPHPGAIKLPGGPRGALAQLLHARGDAGPGLRASRG
Elephas         ------------------------------------------------------------
Sus             ------------------------------------------------------------
Gorilla         -------------------------------MQAPPAPPLRSFTPRGYESATPSPRRYK
Aotus           ------------------------------------------------------------
Homo            ------------------------------------------------------------
Callithrix      ------------------------------------------------------------


Galemys         GERRRAGCGSRGRAVAQGRRPGAHGARFPQGGLLGRMEWVWALVLLAALGSGRAERDCRV
Elephas         ---------MGKAALRWSGCQALIAARFPQGGLLGRMEWMWALVLLAALGSGRAERDCRV
Sus             ---------------------------------RSKMEWVWALVLLAALGSAQAERDCRV
Gorilla         AAERPRRAGLPRSTRARTRRPGLRAVPLPVGGFLGKMKWVWALLLLAALGSGRAERDCRV
Aotus           ------------------------------------MKWVWALLLLAVLGSSRAERDCRV
Homo            ------------------------------------------------------------
Callithrix      --------GLPSSTRARTLQPGLLAALLLVGVLLGKMKWVWALLLLAVLGISRAERDCRV


Galemys         SSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIITEFSVDQHGQMSATAKGRVRLLNS
Elephas         SSFRVKENFDKTRFSGTWYAMAKKDPEGLFLQDNIIAEFSVDESGQMSATAKGRVRLLNN
Sus             SSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDENGHMSATAKGRVRLLNN
Gorilla         SSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNN
Aotus           SSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNN
Homo            -------------FSGTWYAMAKKDPEGLFLQDNNVAEFSVDETGQMSATAKGRVRLLK-
Callithrix      SSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIIAEFSVDETGQMSATAKGRVRLLK-
                             ********************* ::*****: *:************:


Galemys         WDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIIDTDYDTYAVQYSCRLQNLD
Elephas         WDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIIDTDYDTYAVQYSCRLLNLD
Sus             WDVCADMVGTFTDTEN-------------------------------------------
Gorilla         WDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRLLNLD
Aotus           WDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRLLNLD
Homo            ------------------------------------------------------------
Callithrix      -SVAARVAALFE-------------------------------------------------


Galemys         GTCADSYSFIFSRDPNGLPPEAQRIVRRRQEELCLARQYRLIAHNGECPGSGPRAGGQRG
Elephas         GTCADSYSFIFARDPYGLPPEVQKLVRQRQEELCLARQYRMIVHNGYCDGKSEGHVL---
Sus             ------------------------------------------------------------
Gorilla         GTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL---
Aotus           GTCADSYSFVFSRDPNGLPPEAQRIIRQRQEELCLARQYRLIVHNGYCDGKSERNLL---
Homo            ------------------------------------------------------------
Callithrix      ----------FQGLPRALPADRHV-----------------------------------


Galemys         TFHKAVDR
Elephas         --------
Sus             --------
Gorilla         --------
Aotus           --------
Homo            --------
Callithrix      --------
```

[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.

ID	Technique	Resolution	Source	Evalue	Identity
4O9S_A	X-ray	2.30	Homo sapiens	5.01e-41	85.714
2WQ9_A	X-ray	1.65	Homo sapiens	5.46e-41	85.714
1JYJ_A	X-ray	2.00	Homo sapiens	3.27e-41	85.714

[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.).

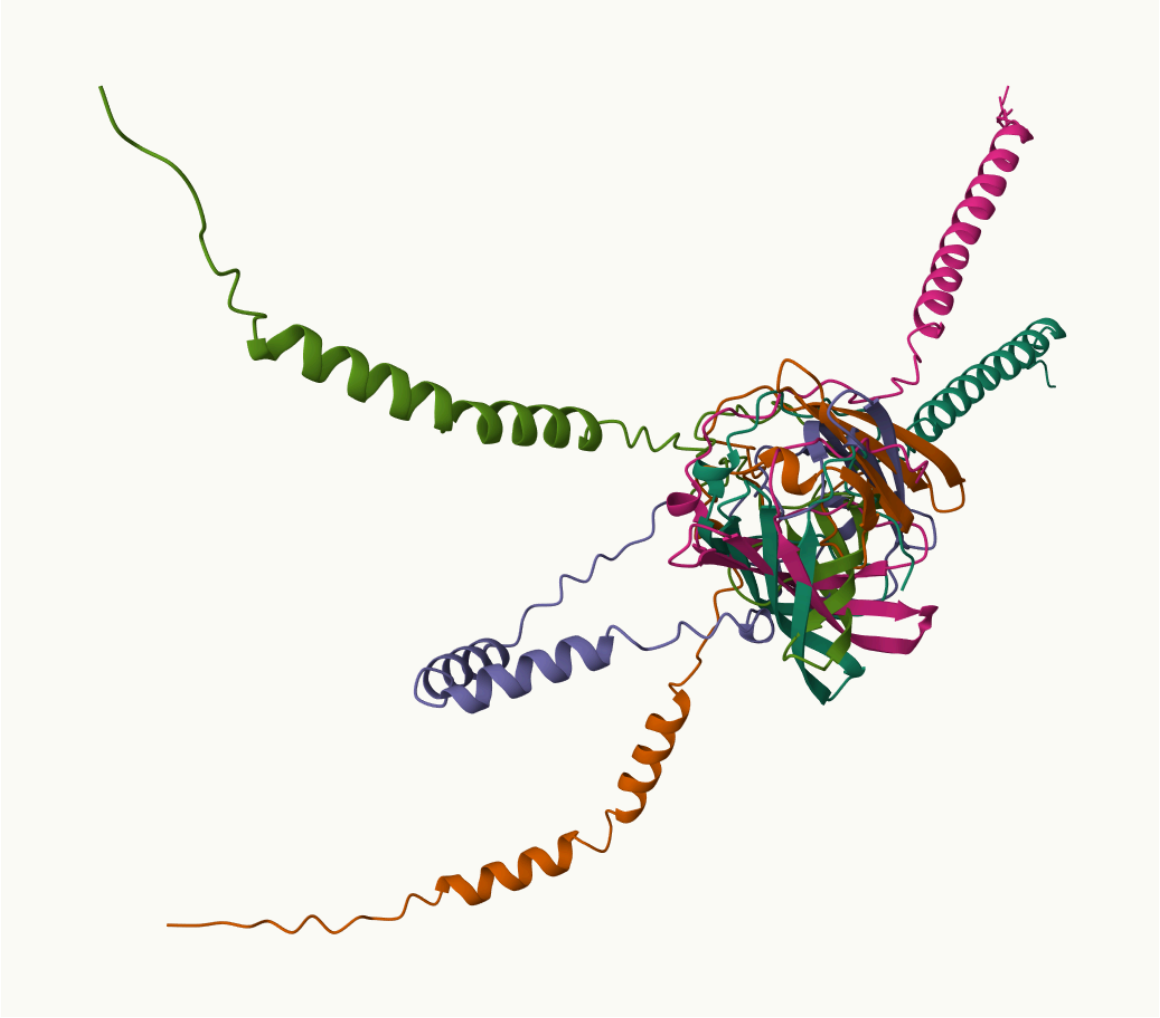


Figure above shows polymer for novel protein sequence

[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]”.

ChEMBEL search:
https://www.ebi.ac.uk/chembl/advanced_search/blast/eyJzZXF1ZW5jZSI6IkdMUFNTVFJBUlRMUVBHTExBQUxMTFZHVkxMR0tNS1dWV0FMTExMQVZMR0lTUkFFUkRDUlZTU0ZSVktFTlxuRkRLQVJGU0dUV1lBTUFLS0RQRUdMRkxRRE5JSUFFRlNWREVUR1FNU0FUQUtHUlZSTExLU1ZBQVJWQUFMXG5GRUZRR0xQUkFMUEFEUkhWXG4ifQ==

After searching using the novel protein sequence, there were 4 results and from those four only one had Retinol binding protein 4 as a target component.

ID: CHEMBL3100

Type: SINGLE PROTEIN

Preferred Name: Plasma retinol-binding protein

Link: https://www.ebi.ac.uk/chembl/explore/target/CHEMBL3100