

BIMM-143: INTRODUCTION TO BIOINFORMATICS

The find-a-gene project assignment

https://bioboot.github.io/bimm143_F21

Kaito Tanaka

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 its 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 (2.12.0) search with ESTs

Database: Expressed Sequence Tags (est)

Organism: None

Translated BLAST: tblastn

blastn | blastp | blastx | **tblastn** | tblastx

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)

P02753

Query subrange [?](#)

From

To

Or, upload file No file chosen [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

☐ Align two or more sequences [?](#)

Choose Search Set

Database [?](#)

Organism [Optional](#) ☐ exclude [Add organism](#)

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown [?](#)

Exclude [Optional](#) ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Limit to [Optional](#) ☐ Sequences from type material

Entrez Query [Optional](#) [YouTube](#) [Create custom database](#)

Enter an Entrez query to limit search [?](#)

BLAST Search database **est** using **Tblastn** (search translated nucleotide databases using a protein query)

☐ Show results in a new window

New columns added to the Description Table. Click 'Select Columns' or 'Manage Columns'.

Chosen match: Accession DR773728.1, a 830 base pair clone from *Macaca mulatta*. See below for alignment details.

<input checked="" type="checkbox"/>	DC627655 macaque liver cDNA library QlvC Macaca fascicularis cDNA clone QlvC-24313 5'. mRNA sequence	Macaca fascicularis	400	400	100%	5e-140	95.02%	937	DC627655.1
<input checked="" type="checkbox"/>	DC643063 macaque kidney cDNA library QreB Macaca fascicularis cDNA clone QreB-29179 5'. mRNA sequence	Macaca fascicularis	400	400	100%	6e-140	95.02%	941	DC643063.1
<input checked="" type="checkbox"/>	DC626124 macaque liver cDNA library QlvC Macaca fascicularis cDNA clone QlvC-19200 5'. mRNA sequence	Macaca fascicularis	400	400	100%	6e-140	95.02%	942	DC626124.1
<input checked="" type="checkbox"/>	DC627041 macaque liver cDNA library QlvC Macaca fascicularis cDNA clone QlvC-22254 5'. mRNA sequence	Macaca fascicularis	400	400	100%	6e-140	95.02%	945	DC627041.1
<input checked="" type="checkbox"/>	DC640901 macaque kidney cDNA library QreB Macaca fascicularis cDNA clone QreB-14114 5'. mRNA sequence	Macaca fascicularis	400	400	100%	6e-140	95.02%	946	DC640901.1
<input checked="" type="checkbox"/>	DC627217 macaque liver cDNA library QlvC Macaca fascicularis cDNA clone QlvC-23103 5'. mRNA sequence	Macaca fascicularis	400	400	100%	7e-140	95.02%	951	DC627217.1
<input checked="" type="checkbox"/>	ILLUMIGEN_MCQ_56576 Katze_MMLV Macaca mulatta cDNA clone IBLUW:30796 5' similar to Bases 4 to 805 hi...	Macaca mulatta	399	399	100%	7e-140	94.53%	830	DR773728.1
<input checked="" type="checkbox"/>	FS582264 macaque heart cDNA library QhtB Macaca fascicularis cDNA clone QhtB-19211 5'. mRNA sequence	Macaca fascicularis	400	400	100%	7e-140	95.02%	955	FS582264.1
<input checked="" type="checkbox"/>	DC629401 macaque liver cDNA library QlvC Macaca fascicularis cDNA clone QlvC-30263 5'. mRNA sequence	Macaca fascicularis	400	400	100%	7e-140	95.02%	955	DC629401.1
<input checked="" type="checkbox"/>	DC627267 macaque liver cDNA library QlvC Macaca fascicularis cDNA clone QlvC-23167 5'. mRNA sequence	Macaca fascicularis	400	400	100%	7e-140	95.02%	962	DC627267.1
<input checked="" type="checkbox"/>	DC628015 macaque liver cDNA library QlvC Macaca fascicularis cDNA clone QlvC-26021 5'. mRNA sequence	Macaca fascicularis	400	400	100%	8e-140	95.02%	972	DC628015.1
<input checked="" type="checkbox"/>	DC622200 macaque liver cDNA library QlvC Macaca fascicularis cDNA clone QlvC-06147 5'. mRNA sequence	Macaca fascicularis	400	400	100%	9e-140	95.02%	975	DC622200.1

[Download](#) [GenBank](#) [Graphics](#)

▼ [Next](#) ▲ [Previous](#) << [Descriptions](#)

ILLUMIGEN_MCQ_56576 Katze_MMLV Macaca mulatta cDNA clone IBIUW:30796 5' similar to Bases 4 to 805 highly similar to human RBP4 (Hs.50223), mRNA sequence

Sequence ID: [DR773728.1](#) Length: 830 Number of Matches: 1

Range 1: 58 to 660 [GenBank](#) [Graphics](#)

▼ [Next Match](#) ▲ [Previous Match](#)

Score	Expect	Method	Identities	Positives	Gaps	Frame
399 bits(1025)	7e-140	Compositional matrix adjust.	198/201(99%)	201/201(100%)	0/201(0%)	+1
Query	1	MKWVWAl1111aAlGSGRAERDCRVSSFRVKNFDKARFSGTWYAMAKKDPEGLFLQDNIV				60
		MKWVWAl1111aAlGSGRAERDCRVSSFRVKNFDKARFSGTWYAMAKKDPEGLFLQDNIV				
Sbjct	58	MKWVWAl1111aAlGSGRAERDCRVSSFRVKNFDKARFSGTWYAMAKKDPEGLFLQDNIV				237
Query	61	AEFSVDETGQMSATAKGRVRLNNWDVCDMVGTTDTEDPAFKFMKYVGVAASFQKQND				120
		AEFSVDETGQMSATAKGRVRLNNWDVCDMVGTTDTEDPAFKFMKYVGVAASFQKQND				
Sbjct	238	AEFSVDETGQMSATAKGRVRLNNWDVCDMVGTTDTEDPAFKFMKYVGVAASFQKQND				417
Query	121	DHWI1D1TDYD1YAVQYSCRLN1LDGTCADSYSFVFSRDPNGLPPEAQ1IVRQREELCLA				180
		DHWI1D1TDYD1YAVQYSCRLN1LDGTCADSYSFVFSRDPNGLPPEAQ1IVRQREELCLA				
Sbjct	418	DHWI1D1TDYD1YAVQYSCRLN1LDGTCADSYSFVFSRDPNGLPPEAQ1IVRQREELCLA				597
Query	181	ROYRLIVHNGYCDGRSERNLL 201				
		ROYRLIVHNGYCDGRS+RNLL				
Sbjct	598	ROYRLIVHNGYCDGRSERNLL 660				

```
>ILLUMIGEN_MCQ_56576_Katze_MMLV_Macaca_mulatta_cDNA_clone_IBIUW:30796 5' similar to Bases 4 to 805
highly similar to human RBP4 (Hs.50223), mRNA sequence
Sequence ID: DR773728.1 Length: 830
Range 1: 58 to 660
```

Score:399 bits(1025), Expect:7e-140,
Method:Compositional matrix adjust.,
Identities:198/201(99%), Positives:201/201(100%), Gaps:0/201(0%)

Query	1	MKWVWa11111a1GSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDP E GFLQDNIV	60
Sbjct	58	MKWVWALL1111AALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDP E GFLQDNIV	237
Query	61	AEFSVDETGQMSATAKGRVRLNNNDVDCADMVGTF TTDTPAKFKMKYWGVASFLQKGND	120
Sbjct	238	AEFSVDETGQMSATAKGRVRLNNNDVDCADMVGTF TTDTPAKFKMKYWGVASFLQKGND	417
Query	121	DHWIIVD TDYDTYAVQYSCRLNLDGTCADSYSFVFSRDPNGLPPEAQIVRQRQEELCLA	180
Sbjct	418	DHWIIVD TDYDTYAVQYSCRLNLDGTCADSYSFVFSRDPNGLPPEAQ+IVRQRQEELCLA	597
Query	181	RQYRLIVHNGYCDGRSERNLL	201
Sbjct	598	RQYRLIVHNGYCDGRS+RNLL	660

[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:

```
>ILLUMIGEN_MCQ_56576 Katze_MMLV Macaca mulatta cDNA clone IBIUW:30796 5' similar  
to Bases 4 to 805 highly similar to human RBP4 (Hs.50223), mRNA sequence (sequence taken  
from BLAST result)  
MKVWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAE  
FSVDETGQMSATAKGRVRLNNWDVCADMVGTFTDTEPAKFMMKYWGVASFLQKGNDHWH  
IIDTDYDTYAVQYSCRLNLDGTCADSYSFVFSRDPNGLPPEAQRIVRQRQEELCLARQYRLIVH  
NGYCDGRSKRNLL
```

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: Macaca retinol binding protein 4

Species: Macaca mulatta

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini;
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.

Details: A BLASTP search against NR database (see setup in first screen-shot below) yielded a top hit result to a protein from *Macaca fascicularis* (Crab-eating macaque). See additional screenshots below for top hits and selected alignment details:

blastnblasttblastntblastx

BLASTP programs search protein databases using a protein query. [more...](#)

Reset pageBookmark

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#)

Query subrange [?](#)

From
To

Or, upload file

Choose File

No file chosen [?](#)

Job Title

Protein Sequence

Enter a descriptive title for your BLAST search [?](#)

☐ Align two or more sequences [?](#)

Choose Search Set

Database

Non-redundant protein sequences (nr) [?](#)

Organism

Optional

Enter organism name or id—completions will be suggested [?](#)

☐ exclude [Add organism](#)

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. [?](#)

Exclude

Optional

☐ Models (XM/XP) ☐ Non-redundant RefSeq proteins (WP) ☐ Uncultured/environmental sample sequences

Program Selection

Algorithm

☐ Quick BLASTP (Accelerated protein-protein BLAST)
☒ blastp (protein-protein BLAST)
☐ PSI-BLAST (Position-Specific Iterated BLAST)
☐ PHI-BLAST (Pattern Hit Initiated BLAST)
☐ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm [?](#)

BLAST

Search database nr using Blastp (protein-protein BLAST)
☐ Show results in a new window

The top result is to a protein from *Macaca fascicularis* (Crab-eating macaque), see second screen shot below for alignment details:

BLAST® » blastp suite » results for RID-PZP5CNT3013

HomeRecent ResultsSaved StrategiesHelp

< Edit Search

Save Search

Search Summary ▾

[How to read this report?](#)

[BLAST Help Videos](#)

[Back to Traditional Results Page](#)

Job Title

unnamed protein product

RID

PZP5CNT3013

Search expires on 10-21 16:51 pm

[Download All ▾](#)

Program

BLASTP [?](#)

[Citation ▾](#)

Database

nr

[See details ▾](#)

Query ID

Ic|Query_40289

Description

unnamed protein product

Molecule type

amino acid

Query Length

201

Other reports

[Distance tree of results](#)

[Multiple alignment](#)

[MSA viewer](#) [?](#)

Filter Results

Organism

only top 20 will appear

☐ exclude

Type common name, binomial, taxid or group name

[+ Add organism](#)

Percent Identity

E value

Query Coverage

to

to

to

[Filter](#)

[Reset](#)

Descriptions

Graphic Summary

Alignments

Taxonomy

Sequences producing significant alignments

Download ▾

New Select columns ▾

Show 100 ▾ [?](#)

☐ select all 0 sequences selected

[GenPept](#)

[Graphics](#)

[Distance tree of results](#)

[Multiple alignment](#)

New [MSA Viewer](#)

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input type="checkbox"/>	PREDICTED: retinol-binding protein 4 [Macaca fascicularis]	Macaca fascicul...	419	419	100%	6e-148	99.50%	201	XP_005566031.1
<input type="checkbox"/>	retinol-binding protein 4 isoform X1 [Hylobates moloch]	Hylobates moloch	418	418	100%	2e-147	99.00%	214	XP_031992484.1
<input type="checkbox"/>	retinol-binding protein 4 [Papio anubis]	Papio anubis	418	418	100%	2e-147	99.00%	201	XP_003904062.1
<input type="checkbox"/>	retinol-binding protein 4 precursor [Pan troglodytes]	Pan troglodytes	417	417	100%	3e-147	98.51%	201	NP_001038

[Feedback](#)

PREDICTED: retinol-binding protein 4 [Macaca fascicularis]

Sequence ID: [XP_005566031.1](#) Length: **201** Number of Matches: **1**
[See 5 more title\(s\)](#) [See all Identical Proteins\(IPG\)](#)

Range 1: 1 to 201 [GenPept](#) [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Method	Identities	Positives	Gaps
419 bits(1077)	6e-148	Compositional matrix adjust.	200/201(99%)	201/201(100%)	0/201(0%)
Query 1	MKVWVALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV				60
Sbjct 1	MKVWVALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV				60
Query 61	AEFSVDETGQMSATAKGRVRLNNWVDCADMVCTFTDTEDPAKFKMKYWGVASFLQKGND				120
Sbjct 61	AEFSVDETGQMSATAKGRVRLNNWVDCADMVCTFTDTEDPAKFKMKYWGVASFLQKGND				120
Query 121	DHWIIDTDYDTYAVQYSCRLNLDGTCADSYSFVFSRDPNGLPPEAQIRIVRQREELCLA				180
Sbjct 121	DHWIIDTDYDTYAVQYSCRLNLDGTCADSYSFVFSRDPNGLPPEAQIRIVRQREELCLA				180
Query 181	RQYRLIVHNGYCDGRSKRNLL				201
Sbjct 181	RQYRLIVHNGYCDGRSERNLL				201

Related Information

[Gene](#) - associated gene details

[Genome Data Viewer](#) - aligned genomic context

[Identical Proteins](#) - Identical proteins to XP_005566031.1

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

Re-labeled sequences for alignment:

```
>sp|P02753|RET4_HUMAN Retinol-binding protein 4 OS=Homo sapiens OX=9606 GN=RBP4
PE=1 SV=3  RBP-4
Human_RBP4
MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV
AEFSVDETGQMSATAKGRVRLNNWDVCADMVGTFTDTEPAKFKMKYWGVASFLQKGND
DHWIVDTDYDTYAVQYSCRLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLA
RQYRLIVHNGYCDGRSERNLL
```

```
>ILLUMIGEN_MCQ_56576 Katze_MMLV Macaca mulatta cDNA clone IBIUW:30796 5' similar
to Bases 4 to 805 highly similar to human RBP4 (Hs.50223), mRNA sequence (sequence taken
from BLAST result)
Macaca_mulatta
MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAE
FSVDETGQMSATAKGRVRLNNWDVCADMVGTFTDTEPAKFKMKYWGVASFLQKGND DHW
IIDTDYDTYAVQYSCRLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVH
NGYCDGRSKRNLL
```

```
>ILLUMIGEN_MCQ_54653 Katze_MNLV Macaca nemestrina cDNA clone IBIUW:31443 5'
similar to Bases 5 to 854 highly similar to human RBP4 (Hs.50223), mRNA sequence
Macaca_nemestrina
MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV
AEFSVDETGQMSATAKGRVRLNNWDVCADMVGTFTDTEPAKFKMKYWGVASFLQKGND
DHWIIDTDYDTYAVQYSCRLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLA
RQYRLIVHNGYCDGRSERNLL
```

```
>DC629429 macaque liver cDNA library QlvC Macaca fascicularis cDNA clone QlvC-30299 5',
mRNA sequence
Macaca_fascicularis
MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV
AEFSVDETGQMSATAKGRVRLNNWDVCADMVGTFTDTEPAKFKMKYWGVASFLQKGND
DHWIIDTDYDTYAVQYSCRLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLA
```

RQYRLIVHNGYCDGRSERNLL

>HX496860 full-length enriched common marmoset liver cDNA library Callithrix jacchus cDNA clone MLI-189E17, mRNA sequence

Callithrix_jacchus

GISRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIIAEFSVDETGMQSA
TAKGRVRLNNWDVCADMVGTFDTEDPAKFKMKYWGVASFLQKGNDHWDIDTDYDTYA
VQYSCRLNLDGTCADSYSFVFSRDPNGLPPEAQRIIRQRQEELCLARQYRLIVHNGYCD
GKSERNLL

Alignment:

CLUSTAL multiple sequence alignment by MUSCLE (3.8)

```
Callithrix_jacchus      -----GISRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNII
Human_RBP4              MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV
Macaca_mulatta          MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV
Macaca_nemestrina       MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV
Macaca_fascicularis     MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV
                        * .*****:

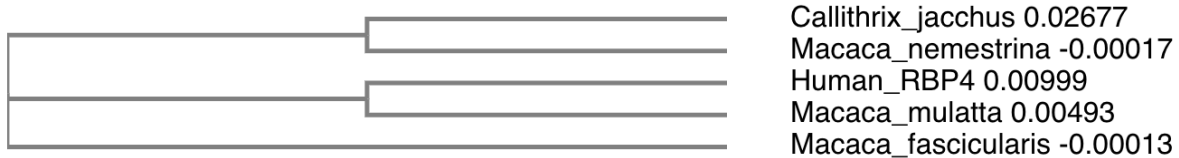
Callithrix_jacchus      AEFSVDETGMQMSATAKGRVRLNNWDVCADMVGTFDTEDPAKFKMKYWGVASFLQKGND
Human_RBP4              AEFSVDETGMQMSATAKGRVRLNNWDVCADMVGTFDTEDPAKFKMKYWGVASFLQKGND
Macaca_mulatta          AEFSVDETGMQMSATAKGRVRLNNWDVCADMVGTFDTEDPAKFKMKYWGVASFLQKGND
Macaca_nemestrina       AEFSVDETGMQMSATAKGRVRLNNWDVCADMVGTFDTEDPAKFKMKYWGVASFLQKGND
Macaca_fascicularis     AEFSVDETGMQMSATAKGRVRLNNWDVCADMVGTFDTEDPAKFKMKYWGVASFLQKGND
                        *****

Callithrix_jacchus      DHWIIDTDYDTYAVQYSCRLNLDGTCADSYSFVFSRDPNGLPPEAQRIIRQRQEELCLA
Human_RBP4              DHWIVDTDYDTYAVQYSCRLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLA
Macaca_mulatta          DHWIIDTDYDTYAVQYSCRLNLDGTCADSYSFVFSRDPNGLPPEAQRIVRQRQEELCLA
Macaca_nemestrina       DHWIIDTDYDTYAVQYSCRLNLDGTCADSYSFVFSRDPNGLPPEAQRIVRQRQEELCLA
Macaca_fascicularis     DHWIIDTDYDTYAVQYSCRLNLDGTCADSYSFVFSRDPNGLPPEAQRIVRQRQEELCLA
                        ****:*****.*:*****

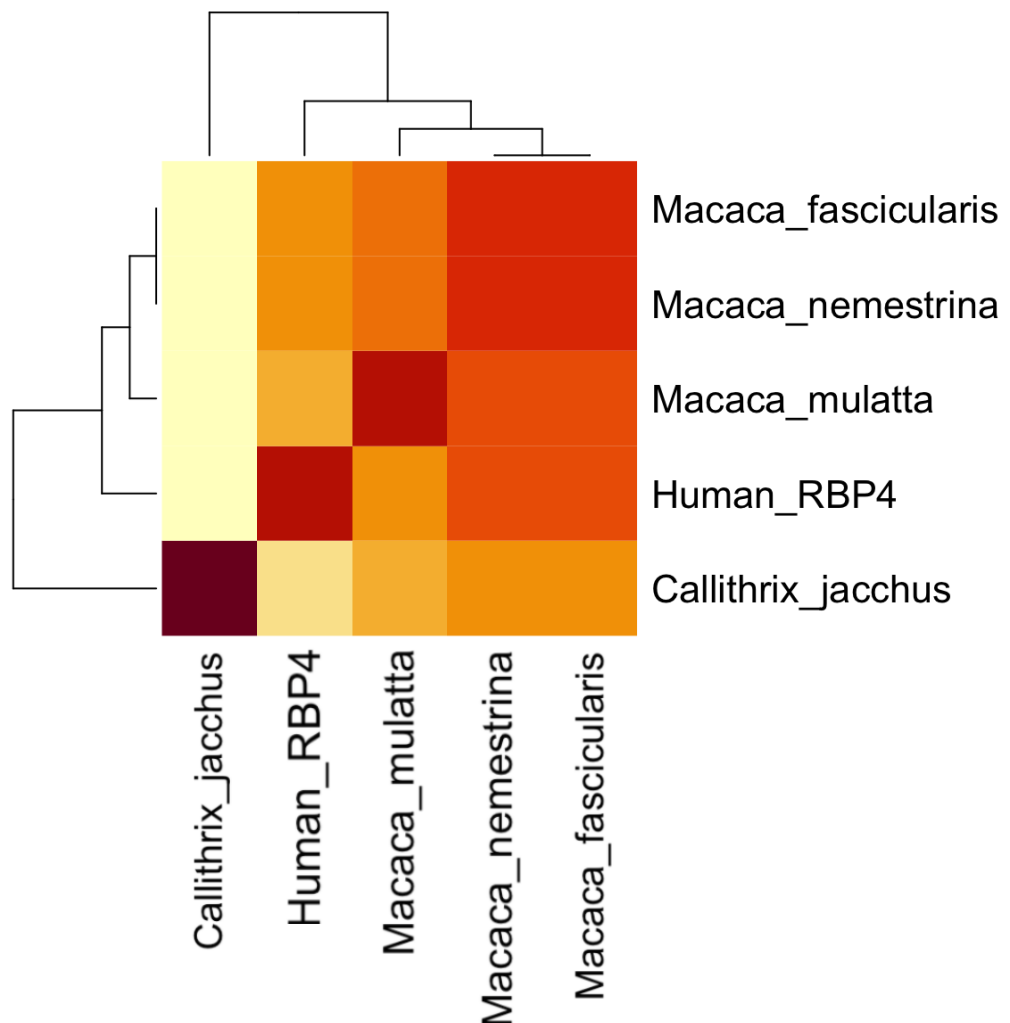
Callithrix_jacchus      RQYRLIVHNGYCDGKSERNLL
Human_RBP4              RQYRLIVHNGYCDGRSERNLL
Macaca_mulatta          RQYRLIVHNGYCDGRSKRNLL
Macaca_nemestrina       RQYRLIVHNGYCDGRSERNLL
Macaca_fascicularis     RQYRLIVHNGYCDGRSERNLL
                        *****.*:****
```

[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.

Use “simple phylogeny” and import the sequences into EBI, and create a neighbor-joining tree:



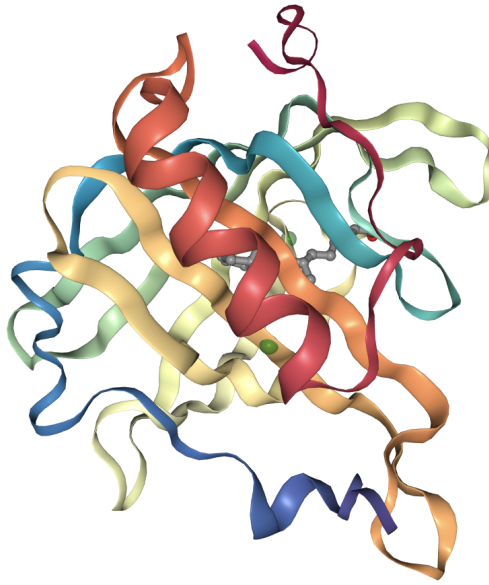
[Q7] Generate a sequence identity based heatmap of your aligned sequences using R. If necessary convert your sequence alignment to the ubiquitous FASTA format (Seaview can read in clustal format and “Save as” FASTA format for example). Read this FASTA format alignment into R with the help of functions 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).

ID	Technique	Resolution	Source	E-value	Identity
1AQB	X-ray diffraction	2.5	Sus scrofa domesticus	7e-131	92.35
1HBQ	X-ray diffraction	2.5	Bos taurus	3e-130	91.80
1RLB	X-ray diffraction	3.1	Gallus gallus	4e-123	91.38

[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?



It is very likely to be similar in structure to *Macaca* retinol binding protein given the high sequence similarity (>90%). In the figure above, the alpha globulin unit is colored in yellow and corresponds to the *Macaca* retinol binding protein subject of this report.

[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?

A Target search on ChEMBEL on my novel protein details 1 binding assay (ChEMBL3100), 5 Functional Assays, 1 unclassified protein, but no ligand efficiency data.

https://www.ebi.ac.uk/chembl/g/#blast_search_results/eyJzZXF1ZW5jZSI6Ik1LV1ZXQUxMTExBQUxHU0dSQUVSRENSVINTRIJS0VORKRLQVJGU0dUV1IBTUFLS0RQRUdMRkxRRE5JVkFFRINWREVUR1FNU0FUQUtHUIZSTExOTIdEVkNBRE1WR1RGVERURURQQUtGS01LWVdHVkFTRkxRS0dORERIV0JRFRFWURUWUFWUUVITQ1JMTE5MREdUQ0FEU1ITRIZGU1JEU E5HTFBQRUFUKIWUFSUUVFTENMQVJRWWJMSVZITkdZQ0RHUINLUk5MTCJ9

The gene that encodes this binding assay “catalyzes the conversion of prostaglandin H2 (PGH2) to prostaglandin D2 (PGD2)”. PGD2 functions as a neuromodulator as well as a trophic factor in the central nervous system. PGD2 is also involved in “smooth muscle contraction/relaxation and is a potent inhibitor of platelet aggregation”.

Wang, P. (2021, November 23). *PTGDS prostaglandin D2 synthase [Homo Sapiens (human)] - gene - NCBI*. National Center for Biotechnology Information. Retrieved December 6, 2021, from <https://www.ncbi.nlm.nih.gov/gene/5730>.

<https://www.ncbi.nlm.nih.gov/gene/5730>