[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:

Original query protein:

```
>Yellow Perch: 16-433 corticotropin-releasing factor receptor 1-like isoform X3
[Perca flavescens]
MMCVCLFLSGRVSPTQLTCETLMLLSTNLTARMLVFLNQTFGIRNSSGVFCDLSVDGIGTCWPLSAAGQLISRPCPEQFN
```

Novel protein:

```
>Pike Perch: 1-429 corticotropin-releasing factor receptor 1 isoform X1 [Sander
lucioperca]
MEKLLSQMMCVCLFLSGRVSPTQLTCETLILLSTNLTARTLVFLNQTFGVRNSSGVFCDLSVDGIGTCWPLSAAGQLISR
```

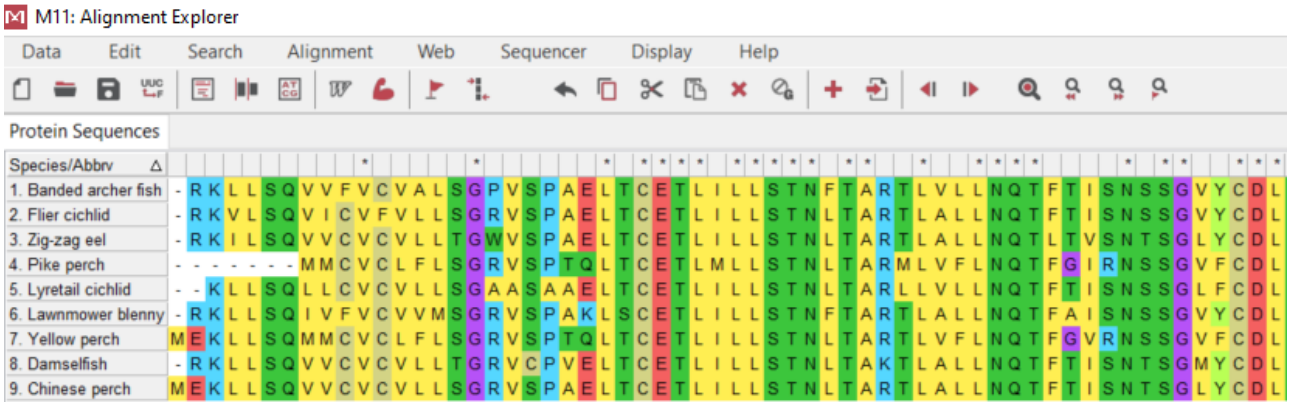
Other sequences for alignment:

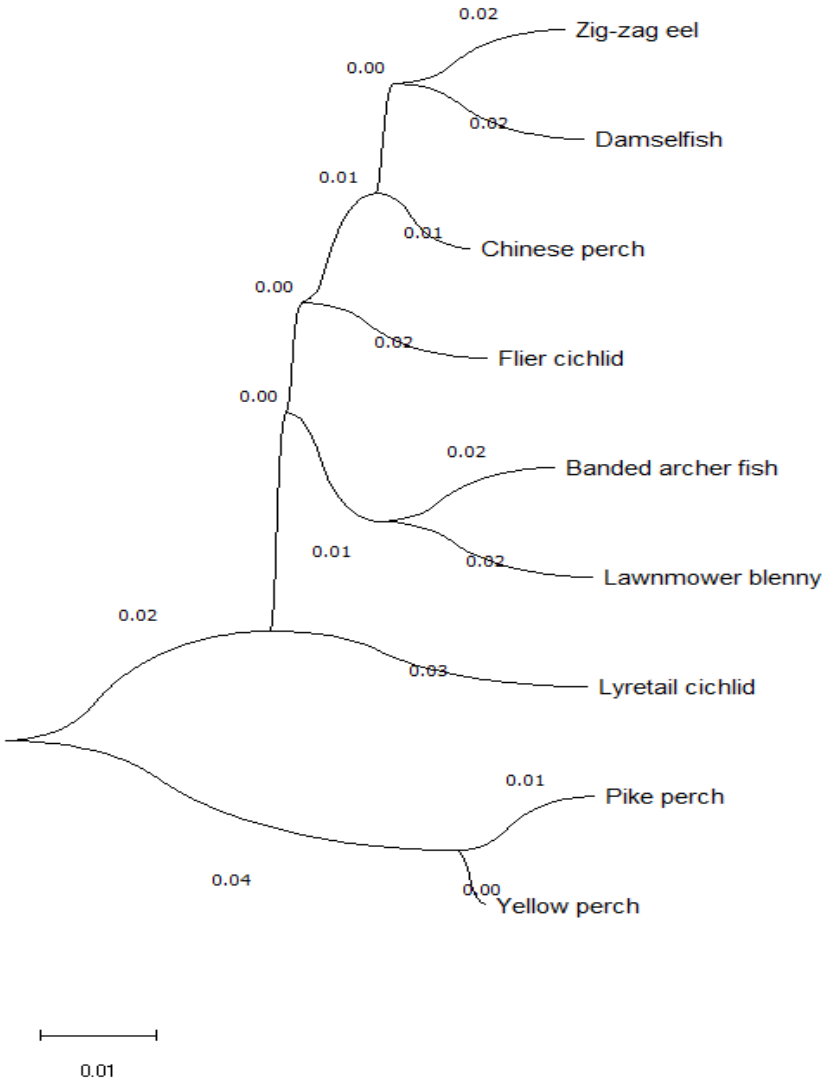
```
>Chinese Perch: 4-432 corticotropin-releasing factor receptor 1 isoform X1
[Siniperca chuatsi]
MEKLLSQVVCVCVLLSGRVSPAELTCETLILLSTNLTARTLALLNQTFTISNTSGLYCDLSVDGIGTCWPRSAAGELISR

>Damselfish: 3-430 corticotropin-releasing factor receptor 1-like isoform X2
[Acanthochromis polyacanthus]
RKVLSQVICVFVLLSGRVSPAELTCETLILLSTNLTARTLALLNQTFTISNSSGVYCDLSVDGIGTCWPRSAAGELISRP

>Flier cichlid: 5-432 corticotropin-releasing factor receptor 1-like
[Archocentrus centrarchus]
RKLLSQIVFVCVVMSGRVSPAKLSCETLILLSTNFTARTLALLNQTFAISNSSGVYCDLSVDGIGTCWPRSAAGELVSRP

>Zig-zag eel: 3-430 corticotropin-releasing factor receptor 1-like isoform X1
[Mastacembelus armatus]
RKILSQVVCVCVLLTGWVSPAELTCETLILLSTNLTARTLALLNQTLTVSNTSGLYCDLSVDGIGTCWPRSAAGELISRP

>Lawnmower blenny: 4-430 corticotropin-releasing factor receptor 1-like
[Salarias fasciatus]
KLLSQLLCVCVLLSGAASAAELTCETLILLSTNLTARLLVLLNQTFTISNSSGLFCDLSVDGIGTCWPRSAAGELVSRPC

>Banded archerfish:6-433 corticotropin-releasing factor receptor 1 isoform X1
[Toxotes jaculatrix]
RKLLSQVVCVCVLLTGRVCPVELTCETLILLSTNLTAKTLALLNQTFTISNTSGMYCDLSVDGIGTCWPRSAAGELISRP

>Lyretail cichlid: 3-430 corticotropin-releasing factor receptor 1
[Neolamprologus brichardi]
RKLLSQVVFVCVALSGPVSPAELTCETLILLSTNFTARTLVLLNQTFTISNSSGVYCDLSVDGIGTCWPRSAAGELVSRP
```

```
CLUSTAL format alignment by MAFFT (v7.503)

Yellow perch:     MEKLLSQMMCVCLFLSGRVSPTQLTCETLILLSTNLTARTLVFLNQTFGVRNSSGVFCDL
Pike perch:       -------MMCVCLFLSGRVSPTQLTCETLMLLSTNLTARMLVFLNQTFGIRNSSGVFCDL
Chinese perch:    MEKLLSQVVCVCVLLSGRVSPAELTCETLILLSTNLTARTLALLNQTFTISNTSGLYCDL
Damselfish:       -RKLLSQVVCVCVLLTGRVCPVELTCETLILLSTNLTAKTLALLNQTFTISNTSGMYCDL
Flier cichlid:    -RKVLSQVICVFVLLSGRVSPAELTCETLILLSTNLTARTLALLNQTFTISNSSGVYCDL
Zig-zag eel:      -RKILSQVVCVCVLLTGWVSPAELTCETLILLSTNLTARTLALLNQTLTVSNTSGLYCDL
Lawnmower:        -RKLLSQIVFVCVVMSGRVSPAKLSCETLILLSTNFTARTLALLNQTFAISNSSGVYCDL
Banded archer:    -RKLLSQVVFVCVALSGPVSPAELTCETLILLSTNFTARTLVLLNQTFTISNSSGVYCDL
Lyretail:         --KLLSQLLCVCVLLSGAASAAELTCETLILLSTNLTARLLVLLNQTFTISNSSGLFCDL
                         :: * : ::* ....:*:****:*****:**: *.:****: : *:**::***

XP_035852444.1: SVDGIGTCWPLSAAGQLISRPCPEQFNGIHYNTSNRVFRECQTNGSWAPRGNYSQCTEII
XP_028454502.1: SVDGIGTCWPLSAAGQLISRPCPEQFNGIHYNTSNRVFRECQTNGSWAPRGNYSQCTEII
XP_044037885.1: SVDGIGTCWPRSAAGELISRPCPEQFNGIHYNTTNRVYRECQSNGSWAPRGNYSQCTEII
XP_040917920.1: SVDGIGTCWPRSAAGELISRPCPEQFNGIHYNTTNRVYRECQSNGSWALRGNYSQCTEII
XP_022067791.1: SVDGIGTCWPRSAAGELISRPCPEQFNGIHYNTTNRVFRECLSNGSWAPRGNYSQCTEII
XP_026180043.1: SVDGIGTCWPRSAAGELISRPCPEQFNGIHYNTSNRVYRECQFNGSWAPRGNYSQCTEII
XP_030591515.1: SVDGIGTCWPRSAAGELVSRPCPEQFNGIHYNTTNRVYRECQVNGSWAPRGNYSQCTEII
XP_006805303.1: SVDGIGTCWPRSAAGELVSRPCPEQFNGIHYNTTNRVYRECQVNGSWAPRGNYSQCTEII
XP_029944739.1: SVDGIGTCWPRSAAGELVSRPCPEQFNGIHYNTTNRVYRDCQSNGSWAPRGNYSQCTEII
                ********** ****:*:***************:***:*:*  ***** ***********

XP_035852444.1: VMRKSKLHYQVAVIINYLGHCFSLGALLLAFTLFLRLRSIRCLRNIIHWNLISAFILRNA
XP_028454502.1: IMRKTKLHYQVAVIINYLGHCFSLGALLLAFTLFLRLRSIRCLRNIIHWNLISAFILRNA
XP_044037885.1: VLRKSKVHYQVAVIINYLGHCISLGALLLAFTLFMRLRSIRCLRNIIHWNLISAFILRNA
XP_040917920.1: VLRKSKVHYQVAVIINYLGHCISLGALLLAFTLFMRLRSIRCLRNIIHWNLISAFILRNA
XP_022067791.1: ILRKSKVHYHVAVIINYLGHCISLGALLLAFTLFMRLRSIRCLRNIIHWNLISAFILRNA
XP_026180043.1: VLRKSKVHYQVAVIINYLGHCISLGALLLAFTLFMRLRSIRCLRNIIHWNLISAFILRNA
XP_030591515.1: ILRKSKVHYQVAVIINYMGHCISLGALLLAFTLFMRLRSIRCLRNIIHWNLISAFILRNA
XP_006805303.1: VLRKSKVHYQVAVIINYLGHCISLGALLLAFTLFMRLRSIRCLRNIIHWNLISAFILRNA
XP_029944739.1: VMRKSKVHYHVAVIINYLGHCISLGALLLAFTLFMRLRSIRCLRNIIHWNLISAFILRNA
                ::**:*:**:*******:***:************:************************

XP_035852444.1: TWFIVQLTMNPAVTEGNQVWCRLVTAAYNYFHVTNFFWMFGEGCYLHTAVVLTYSTDKLR
XP_028454502.1: TWFIVQLTMNPTVTEGNQVWCRLVTAAYNYFHVTNFFWMFGEGCYLHTAVVLTYSTDKLR
XP_044037885.1: TWFIVQLTMTSAVTESNQVWCRLVTAGYNYFHVTNFFWMFGEGCYLHTAVVLTYSTDKLR
XP_040917920.1: TWFIVQLTMTPAVTESNQVWCRLVTAAYNYFHVTNFFWMFGEGCYLHTAVVLTYSTDKLR
XP_022067791.1: TWFIVQLTMNPAVTESNQVWCRLVTAGYNYFHVTNFFWMFGEGCYLHTAIVLTYSTDKLR
XP_026180043.1: TWFIVQLTMNPAVTESNQVWCRLVTAAYNYFHVTNFFWMFGEGCYLHTAVVLTYSTDKLR
XP_030591515.1: TWFIVQLTMNPAVTESNQVWCRLVTAGYNYFHVTNFFWMFGEGCYLHTAVVLTYSTDKLR
XP_006805303.1: TWFIVQLTMNPAVTERNQVWCRLVTAGYNYFHVTNFFWMFGEGCYLHTAVVLTYSTDKLR
XP_029944739.1: TWFIVQLTMNPAVTESNQVWCRLVTAGYNYFHVTNFFWMFGEGCYLHTAVVLTYSTDKLR
                *********..:*** **********.**********************:**********

XP_035852444.1: KWMFICIGWGIPFPIIVAWAFGKLYYDNEKCWFGKKAGVYTDYIYQGPMILVLLINFVFL
XP_028454502.1: KWMFICIGWGIPFPIIVAWAFGKLYYDNEKCWFGKKAGVYTDYIYQGPMILVLLINFVFL
XP_044037885.1: KWMFICIGWGIPFPIIVAWAFGKLYYDNEKCWFGKRAGVYTDYIYQGPMILVLLINFVFL
XP_040917920.1: KWMFICIGWGIPFPIIVAWAFGKLYYDNEKCWFGKRAGVYTDYIYQGPMILVLLINFVFL
XP_022067791.1: KWMFICIGWGIPFPIIVAWAFGKLYYDNEKCWFGKRAGVYTDYIYQGPMILVLLINFVFL
XP_026180043.1: KWMFICIGWGIPFPIIVAWAFGKLYYDNEKCWFGKRAGVYTDYIYQGPMILVLLINFVFL
XP_030591515.1: KWMFICIGWGIPFPIIVAWAFGKLYYDNEKCWFGKRAGVYTDYIYQGPMILVLLINFVFL
```

```
XP_006805303.1: KWMFICIGWGIPFPIIVAWAFGKLYYDNEKCWFGKRAGVYTDYIYQGPMILVLLINFVFL
XP_029944739.1: KWMFICIGWGIPFPIIVAWAFGKLYYDNEKCWFGKRAGVYTDYIYQGPMILVLLINFVFL
                ********************************:***********************

XP_035852444.1: FNIVRILMTKLRASTTSETIQYRKAVKATLVLLPLLGITYMLFFVNPGGEDELAQIVFIY
XP_028454502.1: FNIVRILMTKLRASTTSETIQYRKAVKATLVLLPLLGITYMLFFVNPGGEDELAQIVFIY
XP_044037885.1: FNIVRILMTKLRASTTSETIQYRKAVKATLVLLPLLGITYMLFFVNPGGEDEVAQIVFIY
XP_040917920.1: FNIVRILMTKLRASTTSETIQYRKAVKATLVLLPLLGITYMLFFVNPGGEDEVAQIVFIY
XP_022067791.1: FNIVRILMTKLRASTTSETIQYRKAVKATLVLLPLLGITYMLFFVNPGGEDEVAQIVFIY
XP_026180043.1: FNIVRILMTKLRASTTSETIQYRKAVKATLVLLPLLGITYMLFFVNPGGEDEVAQIVFIY
XP_030591515.1: FNIVRILMTKLRASTTSETIQYRKAVKATLVLLPLLGITYMLFFVNPGGEDEVAQIVFIY
XP_006805303.1: FNIVRILMTKLRASTTSETIQYRKAVKATLVLLPLLGITYMLFFVNPGGEDEVARIVFIY
XP_029944739.1: FNIVRILMTKLRASTTSETIQYRKAVKATLVLLPLLGITYMLFFVNPGGEDEVSQIVFIY
                **************************************************:::*****

XP_035852444.1: FNSILESFQGFFVSIFYCFL-NSEVRSAVRKRWIRWQDRHSLRSRAVRASSLPTSASRVS
XP_028454502.1: FNSILESFQGFFV-----FLNNSEVRSAVRKRWIRWQDRHSLRSRAVRASSLPTSASRVS
XP_044037885.1: FNSILESFQGFFVSVFYCFL-NSEVRSAVRKRWIRWQDRHSFRSRAVRATSLPTSPSRVS
XP_040917920.1: FNSILESFQGFFVSVFYCFL-NSEVRSAVRKRWIRWQDRHSIRSRTVRATSLPTSPSRVS
XP_022067791.1: FNSILESFQGFFVSVFYCFL-NSEVRSAVRKRWIRWQDRHSIRSRAVRATSLPTSPSRVS
XP_026180043.1: FNSILESFQGFFVSVFYCFL-NSEVRSAARKRWIRWQDRHSIRSRAVRATSLPTSPSRVS
XP_030591515.1: FNSILESFQGFFVSVFYCFL-NSEVRSAVRKRWIRWQDRHSIRSRAVRATSLPTSPSRVS
XP_006805303.1: FNSILESFQGFFVSVFYCFL-NSEVRSAVRKRWIRWQDRHSIRSRVIRATSLPTSPSRVS
XP_029944739.1: FNSILESFQGFFVSVFYCFL-NSEVRSAVRKRWIRWQDRHSIRSRTVRATSLPTSPSRVS
                *************     ** *******.************:***.:**:*****.****
```

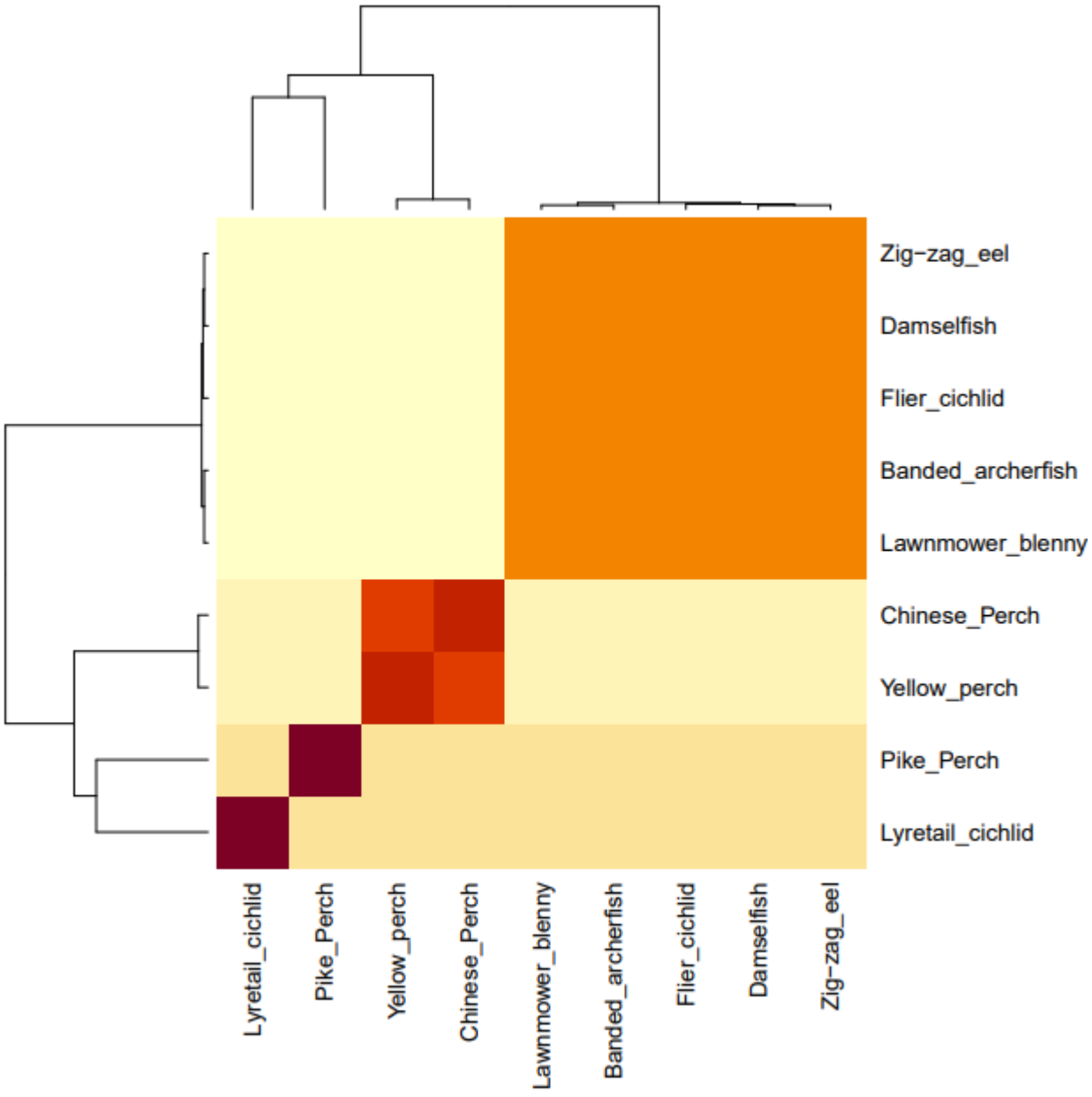
[Q6] Create a phylogenetic tree, using either a parsimony or distance-based approach. Bootstrapping and tree rooting are optional. Use “simple phylogeny” online from the EBI or any respected phylogeny program (such as MEGA, PAUP, or Phylip). Paste an image of your Cladogram or tree output in your report.




0.02
Zig-zag eel
0.00
0.02
Damselfish
0.01
0.01
Chinese perch
0.00
0.02
Flier cichlid
0.00
0.02
Banded archer fish
0.01
0.02
Lawnmower blenny
0.02
0.03
Lyretail cichlid
0.01
Pike perch
0.04
0.00
Yellow perch
0.01

[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	E-value	Identity
6P9X	Electron microscopy	2.910	Homo sapiens	4.9123 e-205	80%
6PB0	Electron microscopy	3.000	Homo sapiens	1.302 e-198	81%
4Z9G	X-ray diffraction	3.183	Entero-bacteria	7.816 e-121	53%

[Q9] Generate a molecular figure of one of your identified PDB structures using **VMD**. 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?

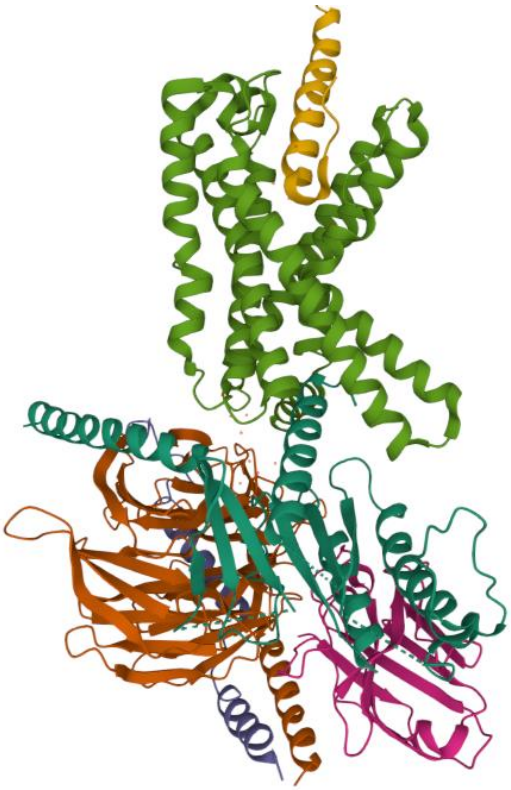


Figure 1: 6P9X

[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?

CHEMBL details 6 Functional Assays for CHEMBL613089; No ligand efficiency data.

https://www.ebi.ac.uk/chembl/target_report_card/CHEMBL613089/