BIMM-143: INTRODUCTION TO BIOINFORMATICS

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

Pamelina Lo
A16735368
palo@ucsd.edu

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 (note that the example report is from a previous quarter and the questions may differ).

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 12pm on the **Monday of Week 5** (see the Assignments and Grading section of our website for details). Note that these first set of answers can be obtained very quickly (at best within 15 or 20 minutes), so if you don't succeed at first, just keep trying.

The complete assignment, including responses to all questions, is due at 12pm on the **Monday of Week 10**.

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.

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 (and thus be successful in the later stages of the project).

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: transthyretin precursor

Accession: NP_000362

Species: Homo Sapiens

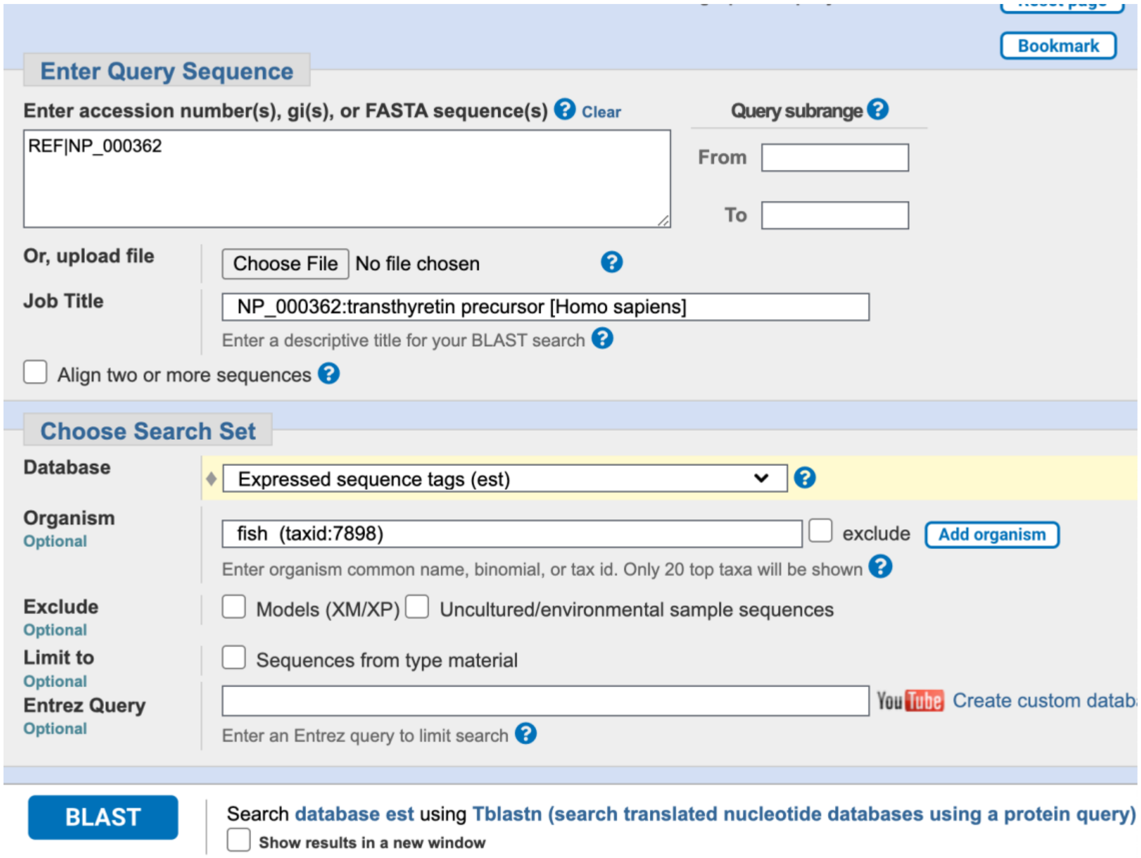
[**Q2**] Perform a BLAST search against a DNA database, such as a database consisting of genomic DNA or ESTs. The BLAST server can be at NCBI or elsewhere. Include details of the BLAST method used, database searched and any limits applied (e.g. Organism).

Method: TBLASTN search against fish ESTs.

Database: Expressed Sequence Tags (est)

Organism: fish (taxid:9263)

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.



Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s): REF|NP_000362

Query subrange: From ___ To ___

Or, upload file: No file chosen

Job Title: NP_000362:transthyretin precursor [Homo sapiens]

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Expressed sequence tags (est)

Organism (Optional): fish (taxid:7898) □ exclude

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): Create custom datab

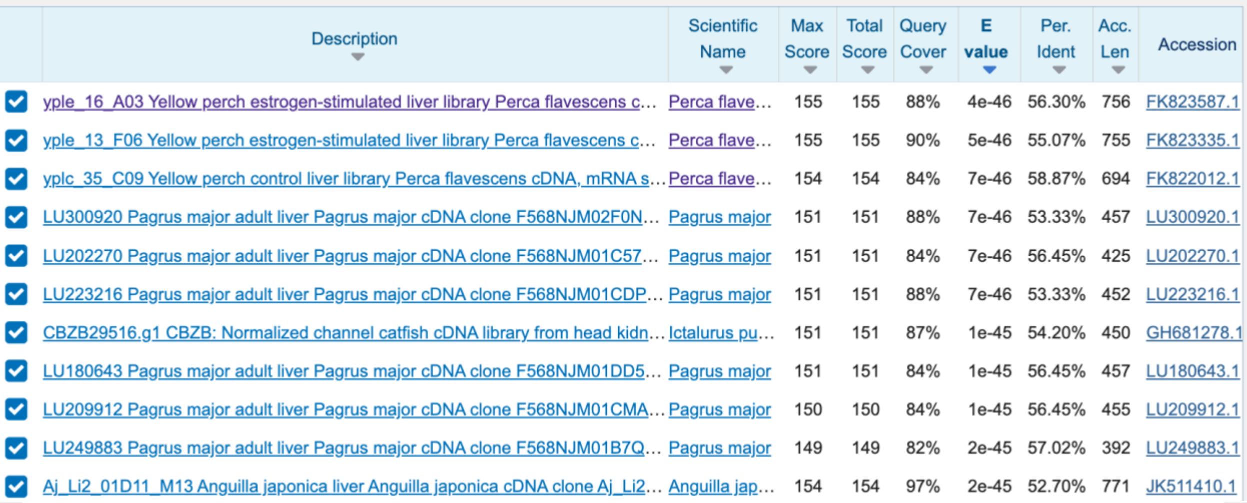
Enter an Entrez query to limit search

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

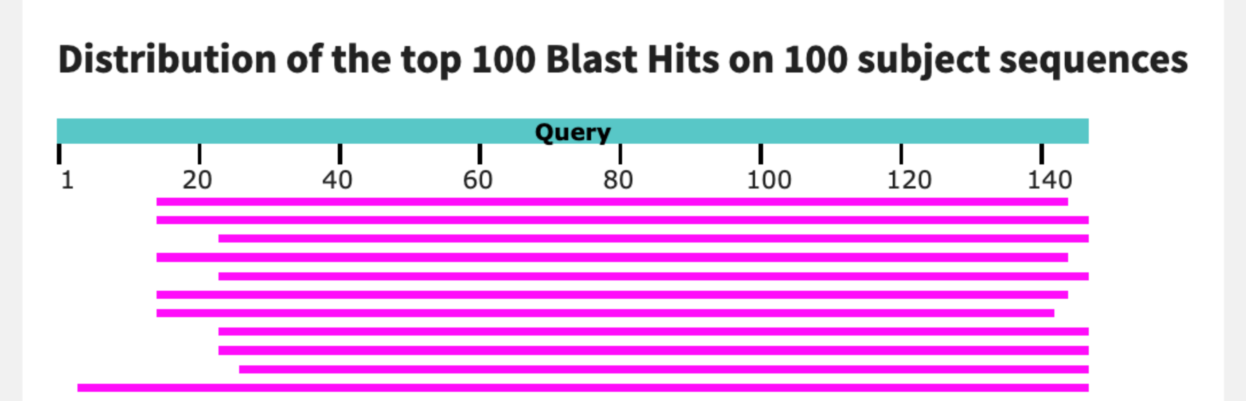
□ Show results in a new window

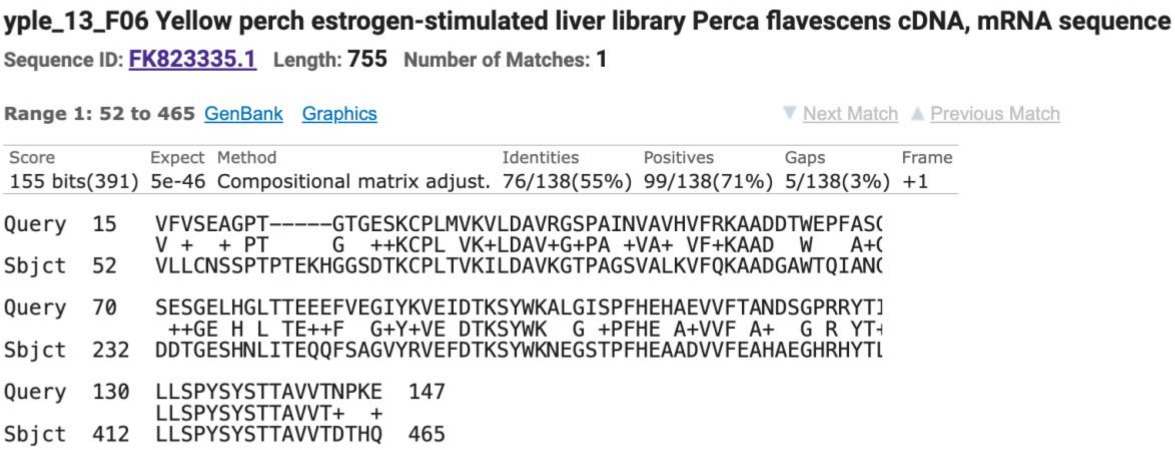
On the BLAST results, clearly indicate a match that represents a protein sequence, encoded from some DNA sequence, that is homologous to your query protein. I need to be able to inspect the pairwise alignment you have selected, including the E value and score. It should be labeled a "genomic clone" or "mRNA sequence", etc. - but include no functional annotation.

Chosen match: Accession FK823335.1, a 7**55** base pair clone from Perca flavescens



	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	yple_16_A03 Yellow perch estrogen-stimulated liver library Perca flavescens c...	Perca flave...	155	155	88%	4e-46	56.30%	756	FK823587.1
☑	yple_13_F06 Yellow perch estrogen-stimulated liver library Perca flavescens c...	Perca flave...	155	155	90%	5e-46	55.07%	755	FK823335.1
☑	yplc_35_C09 Yellow perch control liver library Perca flavescens cDNA, mRNA s...	Perca flave...	154	154	84%	7e-46	58.87%	694	FK822012.1
☑	LU300920 Pagrus major adult liver Pagrus major cDNA clone F568NJM02F0N...	Pagrus major	151	151	88%	7e-46	53.33%	457	LU300920.1
☑	LU202270 Pagrus major adult liver Pagrus major cDNA clone F568NJM01C57...	Pagrus major	151	151	84%	7e-46	56.45%	425	LU202270.1
☑	LU223216 Pagrus major adult liver Pagrus major cDNA clone F568NJM01CDP...	Pagrus major	151	151	88%	7e-46	53.33%	452	LU223216.1
☑	CBZB29516.g1 CBZB: Normalized channel catfish cDNA library from head kidn...	Ictalurus pu...	151	151	87%	1e-45	54.20%	450	GH681278.1
☑	LU180643 Pagrus major adult liver Pagrus major cDNA clone F568NJM01DD5...	Pagrus major	151	151	84%	1e-45	56.45%	457	LU180643.1
☑	LU209912 Pagrus major adult liver Pagrus major cDNA clone F568NJM01CMA...	Pagrus major	150	150	84%	1e-45	56.45%	455	LU209912.1
☑	LU249883 Pagrus major adult liver Pagrus major cDNA clone F568NJM01B7Q...	Pagrus major	149	149	82%	2e-45	57.02%	392	LU249883.1
☑	Aj_Li2_01D11_M13 Anguilla japonica liver Anguilla japonica cDNA clone Aj_Li2...	Anguilla jap...	154	154	97%	2e-45	52.70%	771	JK511410.1





yple_13_F06 Yellow perch estrogen-stimulated liver library Perca flavescens cDNA, mRNA sequence

Sequence ID: FK823335.1 Length: 755 Number of Matches: 1

Range 1: 52 to 465 GenBank Graphics

```
Score           Expect  Method                        Identities     Positives      Gaps        Frame
155 bits(391)   5e-46   Compositional matrix adjust.  76/138(55%)    99/138(71%)    5/138(3%)   +1

Query  15   VFVSEAGPT-----GTGESKCPLMVKVLDAVRGSPAINVAVHVFRKAADDTWEPFAS(
            V +  + PT     G  ++KCPL VK+LDAV+G+PA +VA+ VF+KAAD  W   A+(
Sbjct  52   VLLCNSSPTPTEKHGGSDTKCPLTVKILDAVKGTPAGSVALKVFQKAADGAWTQIAN(

Query  70   SESGELHGLTTEEEFVEGIYKVEIDTKSYWKALGISPFHEHAEVVFTANDSGPRRYTI
             ++GE H L TE++F  G+Y+VE DTKSYWK  G +PFHE A+VVF A+  G R YT+
Sbjct  232  DDTGESHNLITEQQFSAGVYRVEFDTKSYWKNEGSTPFHEAADVVFEAHAEGHRHYTL

Query  130  LLSPYSYSTTAVVTNPKE  147
            LLSPYSYSTTAVVT+  +
Sbjct  412  LLSPYSYSTTAVVTDTHQ  465
```

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:

```
>Transthyretin precursor (taken from BLAST result)
VLLCNSSPTPTEKHGGSDTKCPLTVKILDAVKGTPAGSVALKVFQKAADGAWTQIANGV
TDDTGESHNLITEQQFSAGVYRVEFDTKSYWKNEGSTPFHEAADVVFEAHAEGHRHY
TLALLLSPYSYSTTAVVTDTHQ
```

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: Perca flavescens transthyretin precursor, mRNA, partial cds.

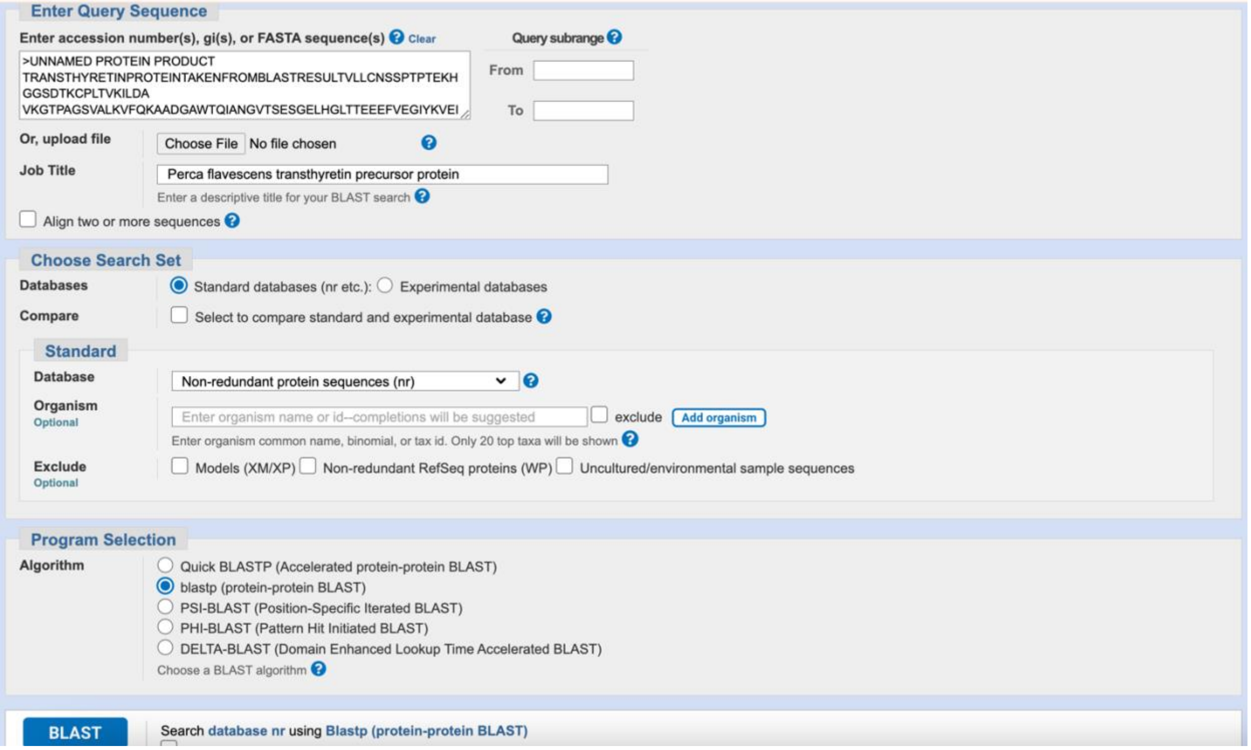
Species: Perca flavescens

> Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Actinopterygii; Neopterygii; Teleostei; Neoteleostei; Acanthomorphata; Eupercaria; Perciformes; Percoidei; Percidae; Percinae; Perca

[**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:
  BLASTP search against NR database to hit result of a protein from Perca flavescens.



Enter Query Sequence

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

```
>UNNAMED PROTEIN PRODUCT
TRANSTHYRETINPROTEINTAKENFROMBLASTRESULTVLLCNSSPTPTEKH
GGSDTKCPLTVKILDA
VKGTPAGSVALKVFQKAADGAWTQIANGVTSESGELHGLTTEEEFVEGIYKVEI
```

Query subrange

From

To

Or, upload file: Choose File No file chosen

Job Title: Perca flavescens transthyretin precursor protein

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Databases: ● Standard databases (nr etc.): ○ Experimental databases

Compare: □ Select to compare standard and experimental database

Standard

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

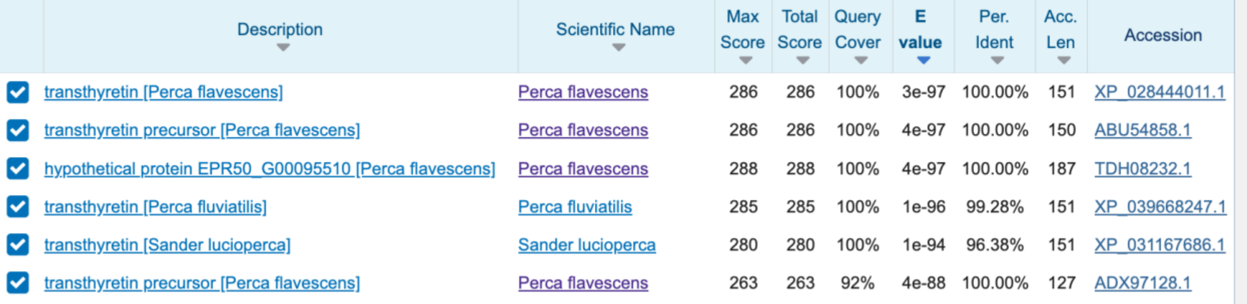
Program Selection

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)

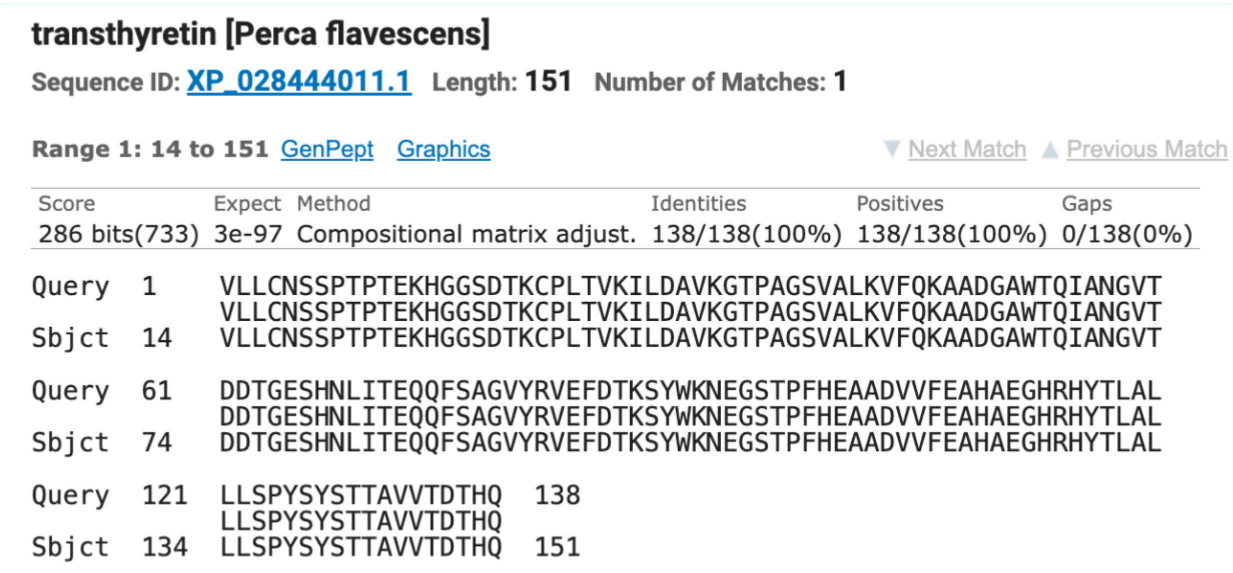
Choose a BLAST algorithm

BLAST Search database nr using Blastp (protein-protein BLAST)

Alignment details:



	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	transthyretin [Perca flavescens]	Perca flavescens	286	286	100%	3e-97	100.00%	151	XP_028444011.1
☑	transthyretin precursor [Perca flavescens]	Perca flavescens	286	286	100%	4e-97	100.00%	150	ABU54858.1
☑	hypothetical protein EPR50_G00095510 [Perca flavescens]	Perca flavescens	288	288	100%	4e-97	100.00%	187	TDH08232.1
☑	transthyretin [Perca fluviatilis]	Perca fluviatilis	285	285	100%	1e-96	99.28%	151	XP_039668247.1
☑	transthyretin [Sander lucioperca]	Sander lucioperca	280	280	100%	1e-94	96.38%	151	XP_031167686.1
☑	transthyretin precursor [Perca flavescens]	Perca flavescens	263	263	92%	4e-88	100.00%	127	ADX97128.1

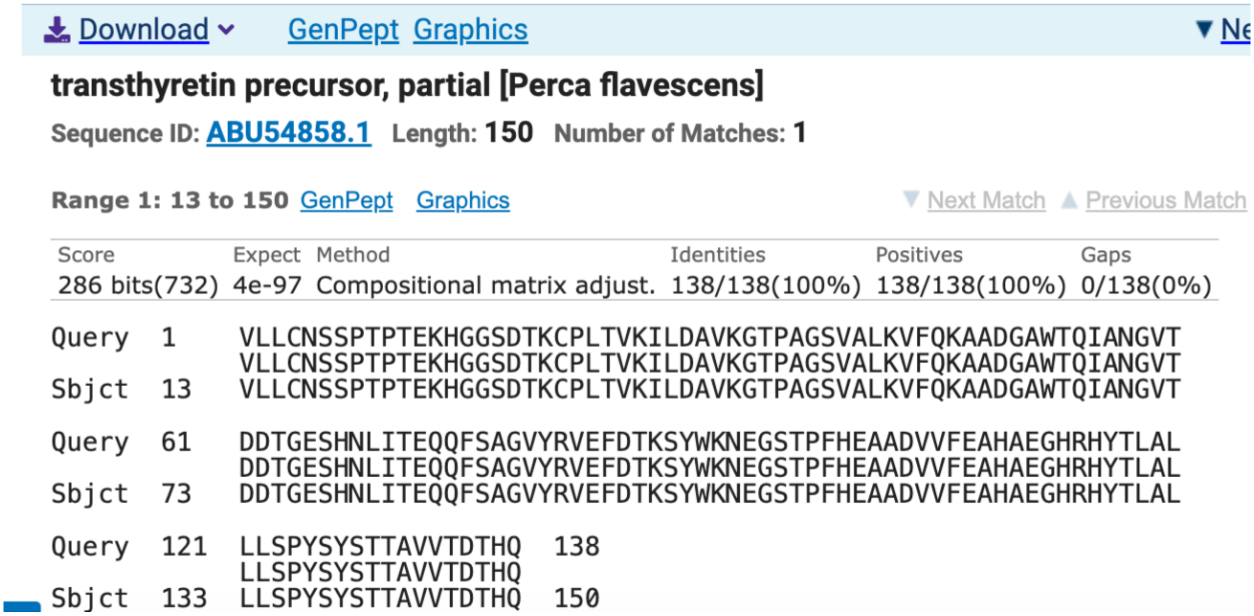


transthyretin [Perca flavescens]

Sequence ID: XP_028444011.1 Length: 151 Number of Matches: 1

Range 1: 14 to 151 GenPept Graphics

```
Score          Expect  Method                        Identities     Positives      Gaps
286 bits(733)  3e-97   Compositional matrix adjust.  138/138(100%)  138/138(100%)  0/138(0%)

Query  1    VLLCNSSPTPTEKHGGSDTKCPLTVKILDAVKGTPAGSVALKVFQKAADGAWTQIANGVT
            VLLCNSSPTPTEKHGGSDTKCPLTVKILDAVKGTPAGSVALKVFQKAADGAWTQIANGVT
Sbjct  14   VLLCNSSPTPTEKHGGSDTKCPLTVKILDAVKGTPAGSVALKVFQKAADGAWTQIANGVT

Query  61   DDTGESHNLITEQQFSAGVYRVEFDTKSYWKNEGSTPFHEAADVVFEAHAEGHRHYTLAL
            DDTGESHNLITEQQFSAGVYRVEFDTKSYWKNEGSTPFHEAADVVFEAHAEGHRHYTLAL
Sbjct  74   DDTGESHNLITEQQFSAGVYRVEFDTKSYWKNEGSTPFHEAADVVFEAHAEGHRHYTLAL

Query  121  LLSPYSYSTTAVVTDTHQ  138
            LLSPYSYSTTAVVTDTHQ
Sbjct  134  LLSPYSYSTTAVVTDTHQ  151
```



transthyretin precursor, partial [Perca flavescens]

Sequence ID: ABU54858.1 Length: 150 Number of Matches: 1

Range 1: 13 to 150 GenPept Graphics

```
Score          Expect  Method                        Identities     Positives      Gaps
286 bits(732)  4e-97   Compositional matrix adjust.  138/138(100%)  138/138(100%)  0/138(0%)

Query  1    VLLCNSSPTPTEKHGGSDTKCPLTVKILDAVKGTPAGSVALKVFQKAADGAWTQIANGVT
            VLLCNSSPTPTEKHGGSDTKCPLTVKILDAVKGTPAGSVALKVFQKAADGAWTQIANGVT
Sbjct  13   VLLCNSSPTPTEKHGGSDTKCPLTVKILDAVKGTPAGSVALKVFQKAADGAWTQIANGVT

Query  61   DDTGESHNLITEQQFSAGVYRVEFDTKSYWKNEGSTPFHEAADVVFEAHAEGHRHYTLAL
            DDTGESHNLITEQQFSAGVYRVEFDTKSYWKNEGSTPFHEAADVVFEAHAEGHRHYTLAL
Sbjct  73   DDTGESHNLITEQQFSAGVYRVEFDTKSYWKNEGSTPFHEAADVVFEAHAEGHRHYTLAL

Query  121  LLSPYSYSTTAVVTDTHQ  138
            LLSPYSYSTTAVVTDTHQ
Sbjct  133  LLSPYSYSTTAVVTDTHQ  150
```

[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.

Relabeled Sequences for alignment:

>|ref|NP_000362| transthyretin precursor [Homo Sapiens]

VFVSEAGPT-----GTGESKCPLMVKVLDAVRGSPAINVAVHVFRKAADDTWEPFASGKT

SESGELHGLTTEEEFVEGIYKVEIDTKSYWKALGISPFHEHAEVVFTANDSGPRRYTIAA
LLSPYSYSTTAVVTN

>transthyretin [Perca flavescens] (sequence taken from BLAST result)

VLLCNSSPTPTEKHGGSDTKCPLTVKILDAVKGTPAGSVALKVFQKAADGAWTQIANGV
TDDTGESHNLITEQQFSAGVYRVEFDTKSYWKNEGSTPFHEAADVVFEAHAEGHRHY
TLALLLSPYSYSTTAVVTDTHQ

>tiger_puffer gi|1698310325|ref|XP_011618262.2| transthyretin-like [Takifugu rubripes]

CHAAPILT-AHGGSDTKCPVTVKILDAVKGTPAGPMVLNLYQRTADGGWTQVANGMTD

ASGEIHNLITEQKFLPGVYRVDFDTKSYWKNEGSVPFHEVTNVVFEAHSEGHRHYTLA
MLLSPYSFTTTALVTD

>gorilla gi|2493370980|ref|XP_004059337.2| transthyretin [Gorilla gorilla gorilla]

VFVSEAGPT-----GTGESKCPLMVKVLDAVRGSPATNVAVHVFKKAADETWEPFASGKT

SESGELHGLTTEEEFVEGIYKVEIDTKSYWKALGISPFHEHAEVVFTANDSGPRRYTIAA

LLSPYSYSTTAVVTNPKE

>frogs gb|MEE6464125.1| **hypothetical protein FKM82_006185 [Ascaphus truei]**

LLICSAAPLVPRPHGAAVSKCPLMIKVLDAVRGSPAANVVVKVFKQEDDESWKMMSTG
KTTDQGEIHGLLTEEEFVEGLYKVEFATKPFWGKVGLSPFHEYVDVVFTANDAGHRHY
TIAVLLTPFSFSTTAVVSDPH

>snakes gi|2679162753|ref|XP_063156304.1| **transthyretin [Candoia aspera]**

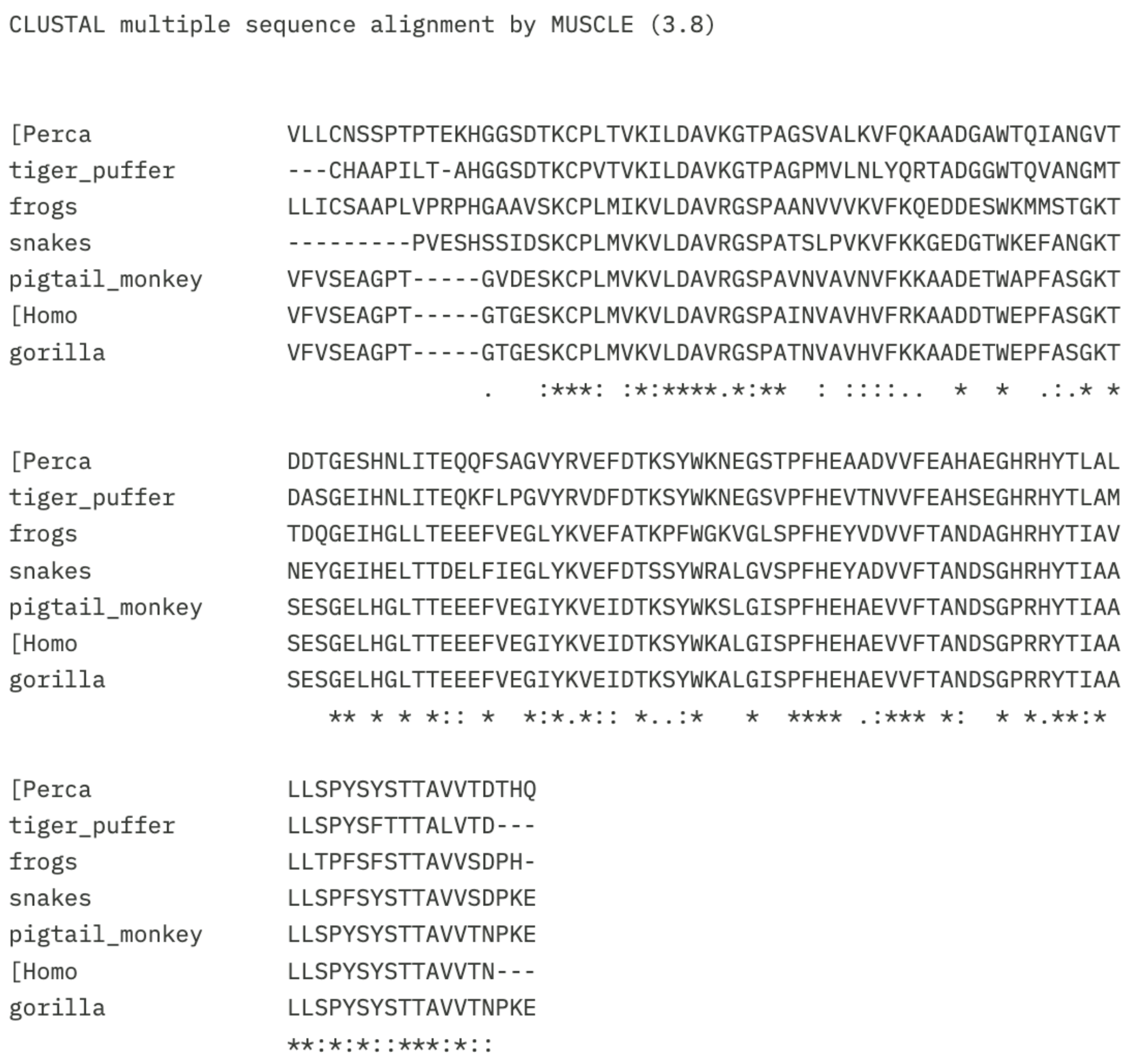
PVESHSSIDSKCPLMVKVLDAVRGSPATSLPVKVFKKGEDGTWKEFANGKTNEYGEIH
ELTTDELFIEGLYKVEFDTSSYWRALGVSPFHEYADVVFTANDSGHRHYTIAALLSPFSY
STTAVVSDPKE

>pigtail_monkey gi|795312522|ref|XP_011716303.1|**transthyretin [Macaca nemestrina]**

VFVSEAGPT-----GVDESKCPLMVKVLDAVRGSPAVNVAVNVFKKAADETWAPFASGKT
SESGELHGLTTEEEFVEGIYKVEIDTKSYWKSLGISPFHEHAEVVFTANDSGPRHYTIAA
LLSPYSYSTTAVVTNPKE

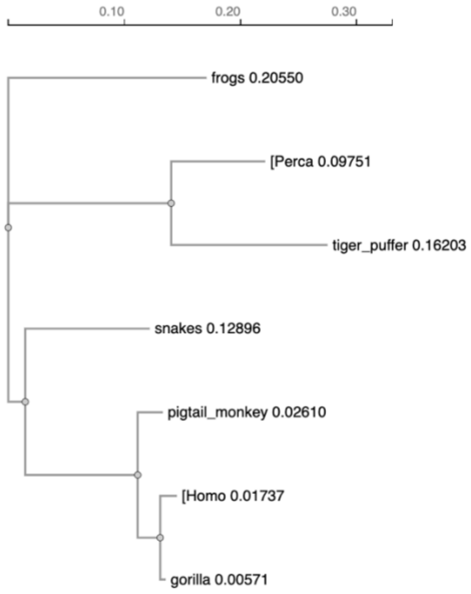
Alignment:

Obtained using MUSCLE (version 3.8) at EBI:



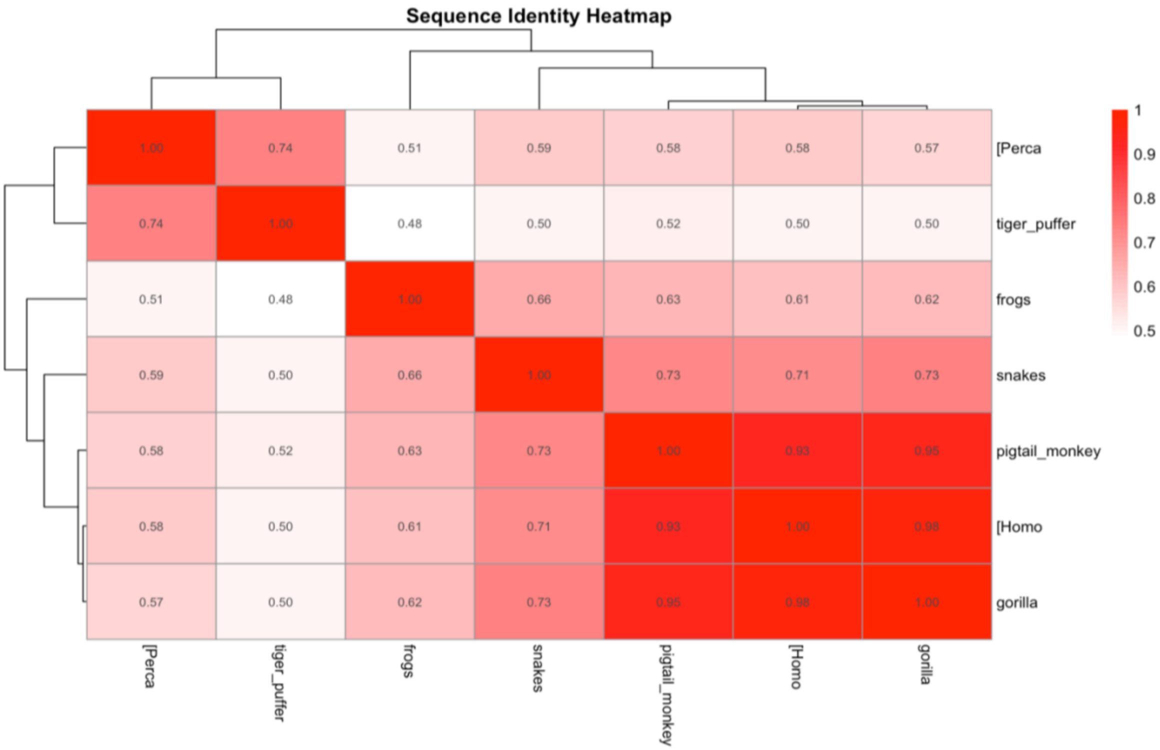
```
CLUSTAL multiple sequence alignment by MUSCLE (3.8)


[Perca              VLLCNSSPTPTEKHGGSDTKCPLTVKILDAVKGTPAGSVALKVFQKAADGAWTQIANGVT
tiger_puffer        ---CHAAPILT-AHGGSDTKCPVTVKILDAVKGTPAGPMVLNLYQRTADGGWTQVANGMT
frogs               LLICSAAPLVPRPHGAAVSKCPLMIKVLDAVRGSPAANVVVKVFKQEDDESWKMMSTGKT
snakes              ---------PVESHSSIDSKCPLMVKVLDAVRGSPATSLPVKVFKKGEDGTWKEFANGKT
pigtail_monkey      VFVSEAGPT-----GVDESKCPLMVKVLDAVRGSPAVNVAVNVFKKAADETWAPFASGKT
[Homo               VFVSEAGPT-----GTGESKCPLMVKVLDAVRGSPAINVAVHVFRKAADDTWEPFASGKT
gorilla             VFVSEAGPT-----GTGESKCPLMVKVLDAVRGSPATNVAVHVFKKAADETWEPFASGKT
                                  .   :***: :*:****.*:**  : ::::..  *  *  .:.* *

[Perca              DDTGESHNLITEQQFSAGVYRVEFDTKSYWKNEGSTPFHEAADVVFEAHAEGHRHYTLAL
tiger_puffer        DASGEIHNLITEQKFLPGVYRVDFDTKSYWKNEGSVPFHEVTNVVFEAHSEGHRHYTLAM
frogs               TDQGEIHGLLTEEEFVEGLYKVEFATKPFWGKVGLSPFHEYVDVVFTANDAGHRHYTIAV
snakes              NEYGEIHELTTDELFIEGLYKVEFDTSSYWRALGVSPFHEYADVVFTANDSGHRHYTIAA
pigtail_monkey      SESGELHGLTTEEEFVEGIYKVEIDTKSYWKSLGISPFHEHAEVVFTANDSGPRHYTIAA
[Homo               SESGELHGLTTEEEFVEGIYKVEIDTKSYWKALGISPFHEHAEVVFTANDSGPRRYTIAA
gorilla             SESGELHGLTTEEEFVEGIYKVEIDTKSYWKALGISPFHEHAEVVFTANDSGPRRYTIAA
                       ** * * *:: *  *:*.*:: *..:*   *  **** .:*** *:  * *.**:*

[Perca              LLSPYSYSTTAVVTDTHQ
tiger_puffer        LLSPYSFTTTALVTD---
frogs               LLTPFSFSTTAVVSDPH-
snakes              LLSPFSYSTTAVVSDPKE
pigtail_monkey      LLSPYSYSTTAVVTNPKE
[Homo               LLSPYSYSTTAVVTN---
gorilla             LLSPYSYSTTAVVTNPKE
                    **:*:*::***:*::
```

[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.



	[Perca	tiger_puffer	frogs	snakes	pigtail_monkey	[Homo	gorilla
[Perca	1.00	0.74	0.51	0.59	0.58	0.58	0.57
tiger_puffer	0.74	1.00	0.48	0.50	0.52	0.50	0.50
frogs	0.51	0.48	1.00	0.66	0.63	0.61	0.62
snakes	0.59	0.50	0.66	1.00	0.73	0.71	0.73
pigtail_monkey	0.58	0.52	0.63	0.73	1.00	0.93	0.95
[Homo	0.58	0.50	0.61	0.71	0.93	1.00	0.98
gorilla	0.57	0.50	0.62	0.73	0.95	0.98	1.00

[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.

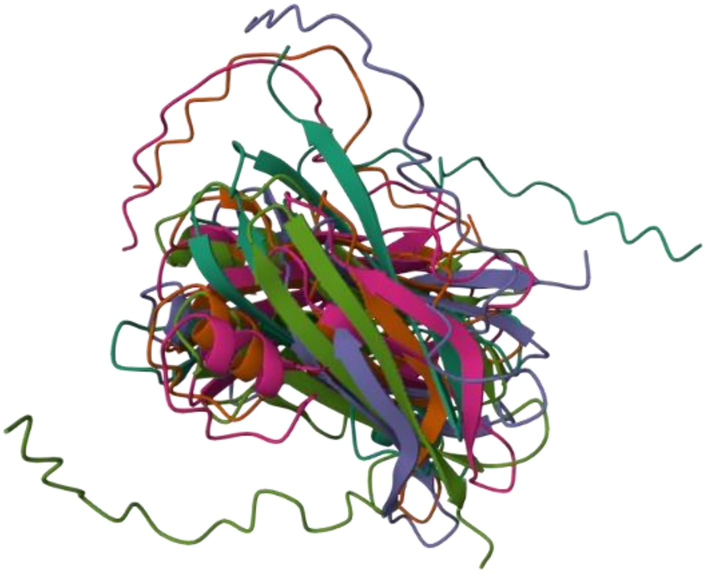
Note that if your consensus sequence has lots of gap positions then it will be better to use an original sequence from the alignment for your search of the PDB. In this case you could chose the sequence with the highest identity to all others in your alignment by calculating the row-wise maximum from your sequence identity matrix.

ID	Technique	Resolution	Source	Evalue	Identity
6GNM	X-ray Diffraction	2.24 Å	Sparus aurata	3e-79	81.8
1SN0	X-ray Diffraction	1.90 Å	Sparus aurata	6e-78	82.3
1OO2	X-ray Diffraction	1.56 Å	Sparus aurata	1e-68	80.7

[**Q9**] Using AlphaFold notebook generate a structural model using the default parameters for your novel protein sequence.

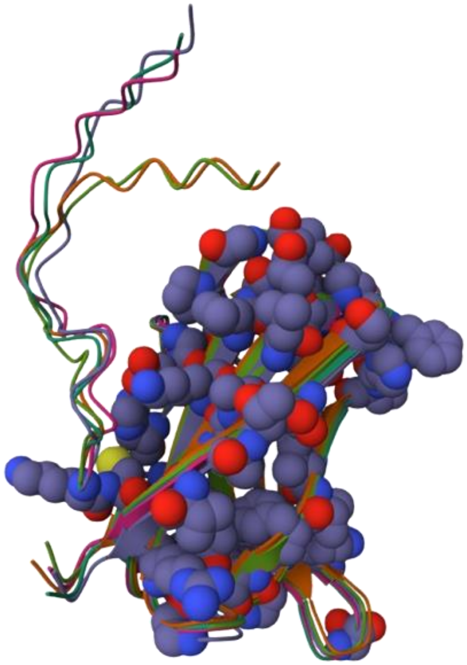
> Note that this can take some time depending upon your sequence length. If your model is taking many hours to generate or your input sequence yields a "too many amino acids" (i.e. length) error you can focus on a single domain from your sequence - identify region by searching for PFAM domain matches.

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.).

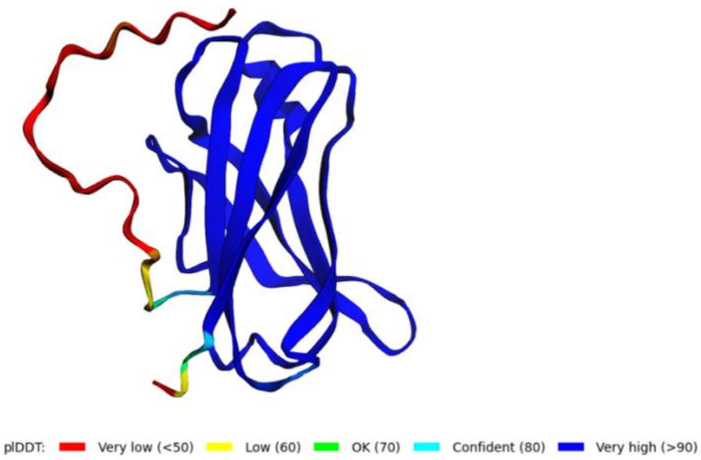


Molecular Figure of PDB structure in Mol* Viewer

Molecular Figure of PDB structure using Mol* Viewer (Superposed)



Molecular Figure of Protein Structure with Conserved Residues using Mol* Viewer in Spacefill



Protein structure colored by local AlphaFold2 pLDDT quality scores

Protein structure colored by Mol* Viewer pLDDT quality scores in Uncertainty/Disorder
Red for high confidence, blue for low confidence

[**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]”.

CHEMBL details 20 binding assay (CHEMBL3291833) and 3 functional assays. There is a graph that displays a visual representation of ligand efficiently data for the target protein (CHEMBL3100). The graph appears to have distribution of points clustered in the high-potency region that suggests the ligand has strong binding and efficient use of its molecular properties.

https://www.ebi.ac.uk/chembl/explore/assay/CHEMBL3291833

The binding assay is linked to a manuscript from Bioorganic and Medicinal Chemistry Letters that highlights the discovery of non-retinoid ligands for retinol-binding protein 4 (RBP4) to reduce renal excretion. The binding between the ligand and RBP4 disrupts the interaction between RBP4 and transthyretin to allow plasma protein to bind RBP4 and be protected.

Yingcai Wang, Richard Connors, Pingchen Fan, Xiaodong Wang, Zhongyu Wang, Jiwen Liu, Frank Kayser, Julio C. Medina, Sheree Johnstone, Haoda Xu, Stephen Thibault, Nigel Walker, Marion Conn, Ying Zhang, Qingxiang Liu, Mark P. Grillo, Alykhan Motani, Peter Coward, Zhulun Wang, Structure-assisted discovery of the first non-retinoid ligands for Retinol-Binding Protein 4, Bioorganic & Medicinal Chemistry Letters, Volume 24, Issue 13, 2014, Pages 2885-2891,ISSN 0960-894X, https://doi.org/10.1016/j.bmcl.2014.04.089.

https://www.sciencedirect.com/science/article/abs/pii/S0960894X14004466?via%3Dihub

Scoring Rubric: [50 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 (9 points)

PDB identifiers from multiple species reported	5
Annotation of PDB source, resolution and technique	4
Annotation of Evalue and Sequence Identity	1

Q9 (10 points)

Structure figure provided	2
Uses white background for molecular figure	1
Figure of high resolution (i.e. not just snapshot)	1
Conserved residues as spacefill	3

Protein cartoon colored by pLDDT quality score	3
Q10 (1 point)	
Evidence of ChEMBEL searches	1