**BIMM-143: INTRODUCTION TO BIOINFORMATICS**

The find-a-gene project assignment

pntran@ucsd.edu

PID: A16352716

**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:** Myoglobin isoform 1

**Accession:** NP_005359.1

**Species:** Homo Sapiens

**Function:** Captures oxygen that muscle cells use for energy

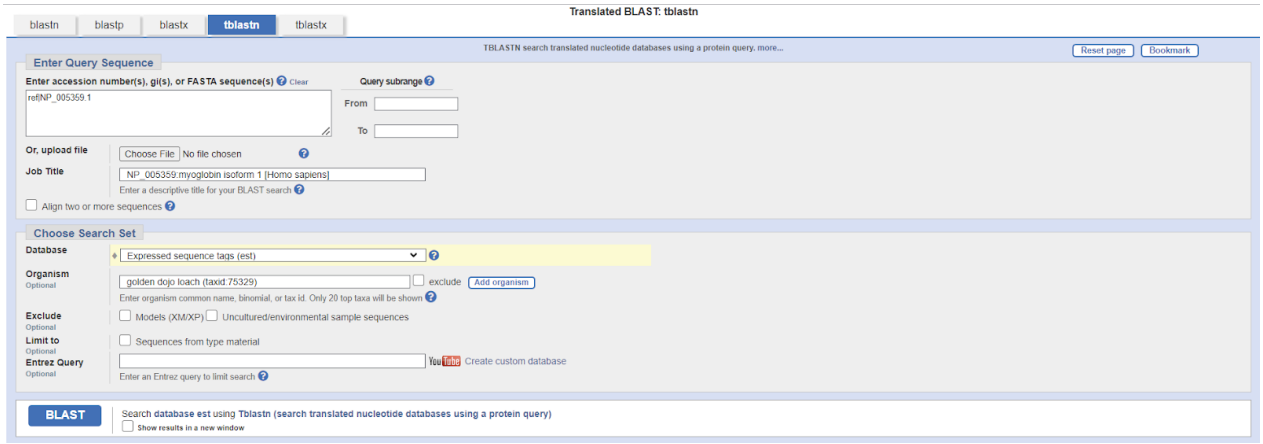
**[Q2]** Perform a BLAST search against a DNA database, such as a database consisting of genomic DNA or ESTs. The BLAST server can be at NCBI or elsewhere. Include details of the BLAST method used, database searched and any limits applied (e.g. Organism).

**Method:** TBLASTN Search against golden dojo loach ESTs

**Database:** Expressed Sequence Tags (est)

**Organism:** Golden Dojo Loach (Taxid: 75329)

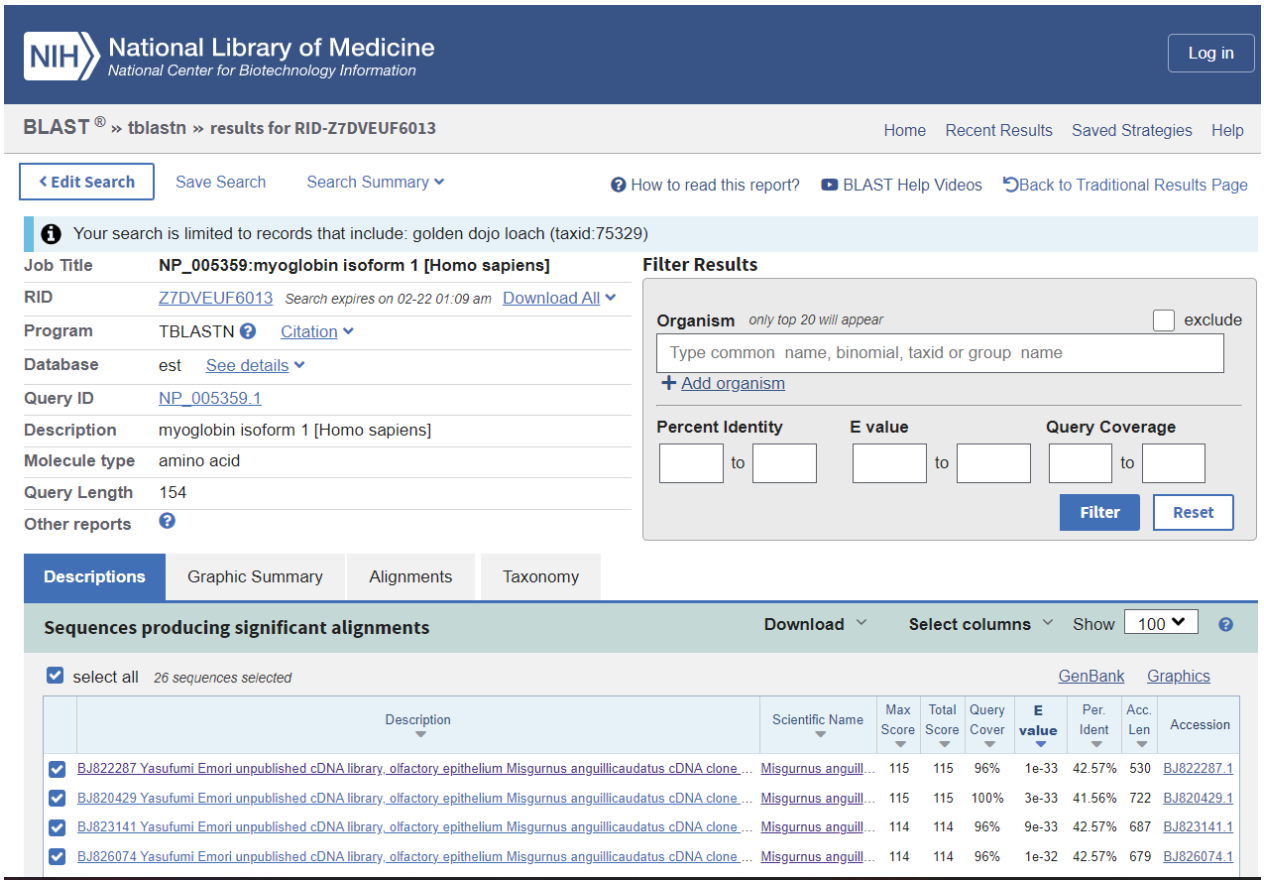
Also include the output of that BLAST search in your document. If appropriate, change the font to `Courier size 10` so that the results are displayed neatly. You can also screen capture a BLAST output (e.g. alt print screen on a PC or on a MAC press ⌘-shift-4. The pointer becomes a bulls eye. Select 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.




| | Description | Scientific Name | Max Score | Total Score | Query Cover | E value | Per. Ident | Acc. Len | Accession |
|---|---|---|---|---|---|---|---|---|---|
| ☑ | BJ822287 Yasufumi Emori unpublished cDNA library, olfactory epithelium Misgurnus anguillicaudatus cDNA clone ... | Misgurnus anguill... | 115 | 115 | 96% | 1e-33 | 42.57% | 530 | BJ822287.1 |
| ☑ | BJ820429 Yasufumi Emori unpublished cDNA library, olfactory epithelium Misgurnus anguillicaudatus cDNA clone ... | Misgurnus anguill... | 115 | 115 | 100% | 3e-33 | 41.56% | 722 | BJ820429.1 |
| ☑ | BJ823141 Yasufumi Emori unpublished cDNA library, olfactory epithelium Misgurnus anguillicaudatus cDNA clone ... | Misgurnus anguill... | 114 | 114 | 96% | 9e-33 | 42.57% | 687 | BJ823141.1 |
| ☑ | BJ826074 Yasufumi Emori unpublished cDNA library, olfactory epithelium Misgurnus anguillicaudatus cDNA clone ... | Misgurnus anguill... | 114 | 114 | 96% | 1e-32 | 42.57% | 679 | BJ826074.1 |
| ☑ | BJ818928 Yasufumi Emori unpublished cDNA library, olfactory epithelium Misgurnus anguillicaudatus cDNA clone ... | Misgurnus anguill... | 114 | 114 | 96% | 1e-32 | 42.57% | 665 | BJ818928.1 |
| ☑ | BJ828491 Yasufumi Emori unpublished cDNA library, olfactory epithelium Misgurnus anguillicaudatus cDNA clone ... | Misgurnus anguill... | 114 | 114 | 96% | 1e-32 | 42.57% | 679 | BJ828491.1 |
| ☑ | BJ820097 Yasufumi Emori unpublished cDNA library, olfactory epithelium Misgurnus anguillicaudatus cDNA clone ... | Misgurnus anguill... | 114 | 114 | 96% | 1e-32 | 42.57% | 683 | BJ820097.1 |
| ☑ | BJ820515 Yasufumi Emori unpublished cDNA library, olfactory epithelium Misgurnus anguillicaudatus cDNA clone ... | Misgurnus anguill... | 100 | 100 | 85% | 2e-28 | 43.51% | 414 | BJ820515.1 |
| ☑ | BJ829859 Yasufumi Emori unpublished cDNA library, olfactory epithelium Misgurnus anguillicaudatus cDNA clone ... | Misgurnus anguill... | 86.3 | 86.3 | 81% | 8e-22 | 38.89% | 693 | BJ829859.1 |
| ☑ | BJ823942 Yasufumi Emori unpublished cDNA library, olfactory epithelium Misgurnus anguillicaudatus cDNA clone ... | Misgurnus anguill... | 80.9 | 80.9 | 77% | 4e-20 | 37.50% | 612 | BJ823942.1 |
| ☑ | BJ831434 Yasufumi Emori unpublished cDNA library, olfactory epithelium Misgurnus anguillicaudatus cDNA clone ... | Misgurnus anguill... | 72.8 | 72.8 | 72% | 1e-16 | 36.04% | 788 | BJ831434.1 |
| ☑ | BJ834336 Yasufumi Emori unpublished cDNA library, olfactory epithelium Misgurnus anguillicaudatus cDNA clone ... | Misgurnus anguill... | 72.0 | 72.0 | 66% | 3e-16 | 36.27% | 762 | BJ834336.1 |
| ☑ | BJ839803 Yasufumi Emori unpublished cDNA library, olfactory epithelium Misgurnus anguillicaudatus cDNA clone ... | Misgurnus anguill... | 72.0 | 72.0 | 66% | 3e-16 | 36.27% | 752 | BJ839803.1 |
| ☑ | BJ837360 Yasufumi Emori unpublished cDNA library, olfactory epithelium Misgurnus anguillicaudatus cDNA clone ... | Misgurnus anguill... | 71.2 | 71.2 | 58% | 4e-16 | 41.11% | 759 | BJ837360.1 |
| ☑ | BJ835173 Yasufumi Emori unpublished cDNA library, olfactory epithelium Misgurnus anguillicaudatus cDNA clone ... | Misgurnus anguill... | 71.2 | 71.2 | 58% | 5e-16 | 41.11% | 761 | BJ835173.1 |
| ☑ | BJ833447 Yasufumi Emori unpublished cDNA library, olfactory epithelium Misgurnus anguillicaudatus cDNA clone ... | Misgurnus anguill... | 70.1 | 70.1 | 66% | 1e-15 | 35.29% | 740 | BJ833447.1 |
| ☑ | BJ833046 Yasufumi Emori unpublished cDNA library, olfactory epithelium Misgurnus anguillicaudatus cDNA clone ... | Misgurnus anguill... | 60.8 | 60.8 | 51% | 2e-12 | 37.97% | 571 | BJ833046.1 |
| ☑ | BJ821891 Yasufumi Emori unpublished cDNA library, olfactory epithelium Misgurnus anguillicaudatus cDNA clone ... | Misgurnus anguill... | 60.8 | 60.8 | 51% | 2e-12 | 37.97% | 629 | BJ821891.1 |
| ☑ | BJ829793 Yasufumi Emori unpublished cDNA library, olfactory epithelium Misgurnus anguillicaudatus cDNA clone ... | Misgurnus anguill... | 60.8 | 60.8 | 55% | 3e-12 | 37.65% | 700 | BJ829793.1 |
| ☑ | BJ831669 Yasufumi Emori unpublished cDNA library, olfactory epithelium Misgurnus anguillicaudatus cDNA clone ... | Misgurnus anguill... | 59.3 | 59.3 | 53% | 5e-12 | 36.59% | 539 | BJ831669.1 |
| ☑ | BJ820636 Yasufumi Emori unpublished cDNA library, olfactory epithelium Misgurnus anguillicaudatus cDNA clone ... | Misgurnus anguill... | 59.3 | 59.3 | 53% | 7e-12 | 36.59% | 588 | BJ820636.1 |
| ☑ | BJ829031 Yasufumi Emori unpublished cDNA library, olfactory epithelium Misgurnus anguillicaudatus cDNA clone ... | Misgurnus anguill... | 59.3 | 59.3 | 55% | 7e-12 | 36.05% | 623 | BJ829031.1 |
| ☑ | BJ818875 Yasufumi Emori unpublished cDNA library, olfactory epithelium Misgurnus anguillicaudatus cDNA clone ... | Misgurnus anguill... | 50.8 | 50.8 | 48% | 2e-09 | 34.67% | 267 | BJ818875.1 |
| ☑ | BJ830653 Yasufumi Emori unpublished cDNA library, olfactory epithelium Misgurnus anguillicaudatus cDNA clone ... | Misgurnus anguill... | 47.4 | 47.4 | 79% | 8e-08 | 25.41% | 451 | BJ830653.1 |
| ☑ | BJ819670 Yasufumi Emori unpublished cDNA library, olfactory epithelium Misgurnus anguillicaudatus cDNA clone ... | Misgurnus anguill... | 47.4 | 47.4 | 79% | 1e-07 | 25.41% | 499 | BJ819670.1 |
| ☑ | BJ831085 Yasufumi Emori unpublished cDNA library, olfactory epithelium Misgurnus anguillicaudatus cDNA clone ... | Misgurnus anguill... | 48.1 | 48.1 | 50% | 1e-07 | 31.17% | 679 | BJ831085.1 |

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 BJ822287.1, a 530 base pair clone from *Misgurnus anguillicaudatus*.



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

Alignment view: Pairwise | Restore defaults | Download

*26 sequences selected*

Download | GenBank | Graphics | ▼Next ▲Previous ◀◀Descriptions

**BJ822287 Yasufumi Emori unpublished cDNA library, olfactory epithelium Misgurnus anguillicaudatus cDNA clone dj18h20 5', mRNA sequence**

**Sequence ID:** BJ822287.1 **Length:** 530 **Number of Matches:** 1

**Range 1: 58 to 492** GenBank Graphics ▼Next Match ▲Previous Match

| Score | Expect | Method | Identities | Positives | Gaps | Frame |
|---|---|---|---|---|---|---|
| 115 bits(287) | 1e-33 | Compositional matrix adjust. | 63/148(43%) | 90/148(60%) | 3/148(2%) | +1 |

```
Query  7    EWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDLKKHG  66
            ++ LVL  WG VEAD  G G EVL RLFK HPETL+ F KF  +   D +  +  +  HG
Sbjct  58   DFDLVLKCWGPVEADYTGVGGEVLTRLFKDHPETLKLFPKFVGIGQGD-LAGNAAVAAHG  234

Query  67   ATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDFGA  126
            ATVL  L  +L+ KG H A +KPLA +HA  HKIP+   + I+E +++V+  +   D GA
Sbjct  235  ATVLKKLAELLRAKGEHAAVLKPLATTHANTHKIPLVNFKLITEALVKVMAERAGLD-GA  411

Query  127  DAQGAMNKALELFRKDMASNYKELGFQG  154
              Q A+ + +++   D+   YKE+G+ G
Sbjct  412  -GQAALKRVMDVVIGDIDKYYKEIGYAG  492
```

## Alignment details:

BJ822287 Yasufumi Emori unpublished cDNA library, olfactory epithelium Misgurnus anguillicaudatus cDNA clone dj18h20 5', mRNA sequence
Sequence ID: BJ822287.1 Length: 530 Number of Matches: 1

Score = 115 bits(287), Expect = 1e-33, Method: Compositional matrix adjust.
Identities = 63/148(43%), Positives = 90/148(60%), Gaps = 3/148(2%), Frame = +1

```
Query  7    EWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDLKKHG  66
            ++ LVL  WG VEAD  G G EVL RLFK HPETL+ F KF  +   D +  +  +  HG
Sbjct  58   DFDLVLKCWGPVEADYTGVGGEVLTRLFKDHPETLKLFPKFVGIGQGD-LAGNAAVAAHG  234

Query  67   ATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDFGA  126
            ATVL  L  +L+ KG H A +KPLA +HA  HKIP+   + I+E +++V+  +   D GA
Sbjct  235  ATVLKKLAELLRAKGEHAAVLKPLATTHANTHKIPLVNFKLITEALVKVMAERAGLD-GA  411

Query  127  DAQGAMNKALELFRKDMASNYKELGFQG  154
              Q A+ + +++   D+   YKE+G+ G
Sbjct  412  -GQAALKRVMDVVIGDIDKYYKEIGYAG  492
```

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:**

```
>Misgurnus anguillicaudatus protein (translated using EMBOSS Transeq)
FQTHEHLDSSEQPLITTMSDFDLVLKCWGPVEADYTGVGGEVLTRLFKDHPETLKLFPKFVGIGQGDLAGNAAVAAH
GATVLKKLAELLRAKGEHAAVLKPLATTHANTHKIPLVNFKLITEALVKVMAERAGLDGAGQAALKRVMDVVIGDID
KYYKEIGYAG*MRPNLSRV*YAG
```

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.

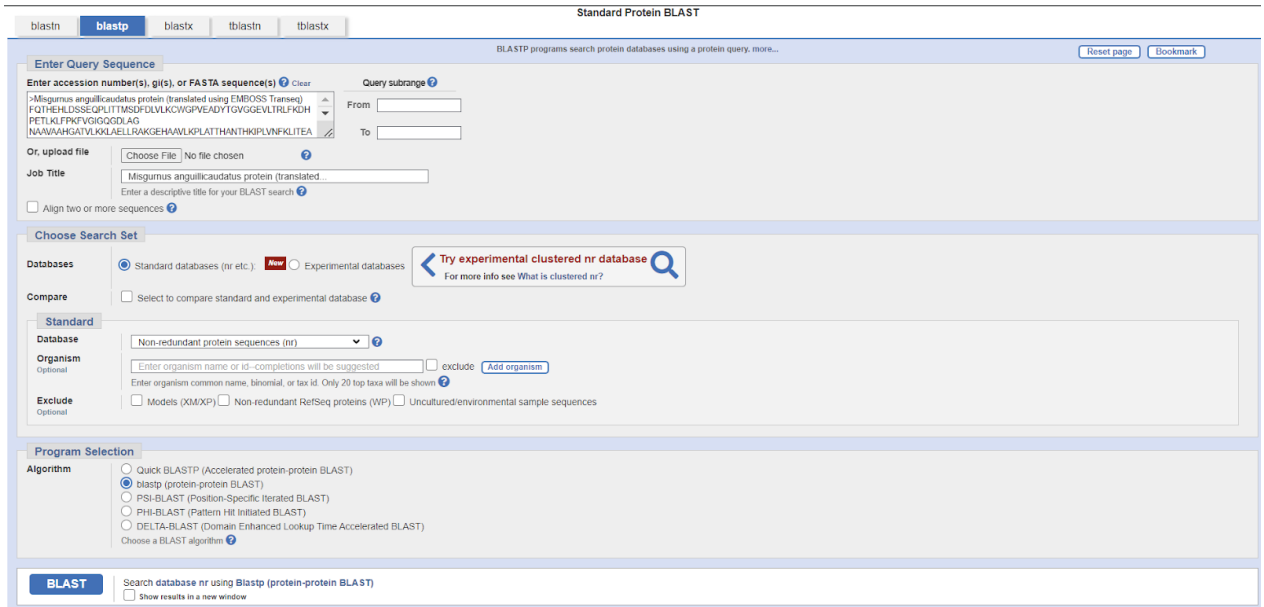
**Name:** *Misgurnus* myoglobin
**Species:** Misgurnus anguillicaudatus
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Actinopterygii; Neopterygii; Teleostei; Ostariophysi; Cypriniformes; Cobitidae; Cobitinae; Misgurnus.

**[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 yielded a top hit result to a protein from *Triplophysa rosa* (Chinese cavefish).




The top result is to a protein from *Triplophysa rosa* (Chinese cavefish)

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 |
|---|---|---|---|---|---|---|---|---|
| myoglobin [Triplophysa rosa] | Triplophysa rosa | 255 | 255 | 83% | 3e-84 | 83.67% | 147 | KAI7810808.1 |
| myoglobin [Puntigrus tetrazona] | Puntigrus tetra... | 249 | 249 | 83% | 5e-82 | 80.95% | 147 | XP_043080305.1 |
| PREDICTED: myoglobin [Sinocyclocheilus anshuiensis] | Sinocyclocheilu... | 246 | 246 | 83% | 6e-81 | 80.95% | 147 | XP_016358141.1 |
| myoglobin [Cyprinus carpio] | Cyprinus carpio | 246 | 246 | 83% | 7e-81 | 80.95% | 147 | XP_018966946.2 |
| myoglobin [Labeo rohita] | Labeo rohita | 246 | 246 | 88% | 5e-80 | 75.00% | 190 | XP_050965105.1 |
| Myoglobin [Anabarilius grahami] | Anabarilius gra... | 244 | 244 | 83% | 5e-80 | 80.27% | 147 | ROL52021.1 |
| hypothetical protein G5714_000293 [Onychostoma macrolepis] | Onychostoma ... | 244 | 244 | 88% | 2e-79 | 75.00% | 179 | KAF4118242.1 |
| myoglobin [Ctenopharyngodon idella] | Ctenopharyngo... | 242 | 242 | 83% | 3e-79 | 79.59% | 147 | XP_051751285.1 |
| myoglobin [Denticeps clupeoides] | Denticeps clup... | 242 | 242 | 83% | 4e-79 | 78.91% | 147 | XP_028830810.1 |
| myoglobin isoform X1 [Megalobrama amblycephala] | Megalobrama ... | 243 | 243 | 85% | 4e-79 | 77.48% | 188 | XP_048051875.1 |
| PREDICTED: myoglobin [Sinocyclocheilus grahami] | Sinocyclocheilu... | 242 | 242 | 83% | 4e-79 | 79.59% | 147 | XP_016145299.1 |

Alignment view Pairwise Restore defaults Download

*100 sequences selected*

Download GenPept Graphics ▼Next ▲Previous ◀◀Descriptions

**myoglobin [Triplophysa rosa]**

**Sequence ID:** KAI7810808.1 **Length:** 147 **Number of Matches:** 1

**Range 1: 1 to 147** GenPept Graphics ▼Next Match ▲Previous Match

| Score | Expect | Method | Identities | Positives | Gaps |
|---|---|---|---|---|---|
| 255 bits(652) | 3e-84 | Compositional matrix adjust. | 123/147(84%) | 135/147(91%) | 0/147(0%) |

```
Query  18   MSDFDLVLKCWGPVEADYTGVGGEVLTRLFKDHPETLKLFPKFVGIGQGDLAGNAAVAAH  77
            M+DFDLVLKCWG +EADYT  GGEVLTRLF++HPETLKLFPKFVGI QGDLAGNAAVAAH
Sbjct  1    MADFDLVLKCWGAMEADYTAHGGEVLTRLFQEHPETLKLFPKFVGIAQGDLAGNAAVAAH  60

Query  78   GATVLKKLAELLRAKGEHAAVLKPLATTHANTHKIPLVNFKLITEALVKVMAERAGLDGA  137
            GATVLKKL +LL+AKG+HA +LKPLA THAN HKIPL NFKLITE +V++MAERAGLDGA
Sbjct  61   GATVLKKLGDLLKAKGDHAGILKPLANTHANNHKIPLNNFKLITEIIVQLMAERAGLDGA  120

Query  138  GQAALKRVMDVVIGDIDKYYKEIGYAG  164
            GQAAL+RV DVVIGDID YYKEIGYAG
Sbjct  121  GQAALRRVFDVVIGDIDGYYKEIGYAG  147
```

**Related Information**
Genome Data Viewer - aligned genomic context

Download GenPept Graphics ▼Next ▲Previous ◀◀Descriptions

**myoglobin [Puntigrus tetrazona]**

**Sequence ID:** XP_043080305.1 **Length:** 147 **Number of Matches:** 1

**Range 1: 1 to 147** GenPept Graphics ▼Next Match ▲Previous Match

| Score | Expect | Method | Identities | Positives | Gaps |
|---|---|---|---|---|---|
| 249 bits(637) | 5e-82 | Compositional matrix adjust. | 119/147(81%) | 134/147(91%) | 0/147(0%) |

```
Query  18   MSDFDLVLKCWGPVEADYTGVGGEVLTRLFKDHPETLKLFPKFVGIGQGDLAGNAAVAAH  77
            M+DFD VLKCWG VEAD+ G GGEVLTRLFK+HPET KLFPKFVGI Q DLAGNAAVA+H
Sbjct  1    MADFDQVLKCWGAVEADFAGHGGEVLTRLFKEHPETQKLFPKFVGISQSDLAGNAAVASH  60

Query  78   GATVLKKLAELLRAKGEHAAVLKPLATTHANTHKIPLVNFKLITEALVKVMAERAGLDGA  137
            GATVLKKL ELL+A+G+HAA+LKPLAT+HAN HKI L NF+LITE LVKVMAE+AGLDGA
Sbjct  61   GATVLKKLGELLKARGDHAAILKPLATSHANIHKITLNNFRLITEVLVKVMAEKAGLDGA  120

Query  138  GQAALKRVMDVVIGDIDKYYKEIGYAG  164
            GQ+AL+RVM+VVIGDID YYKEIG+AG
Sbjct  121  GQSALRRVMEVVIGDIDAYYKEIGFAG  147
```

**Related Information**
Gene - associated gene details
Genome Data Viewer - aligned genomic context

**[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:

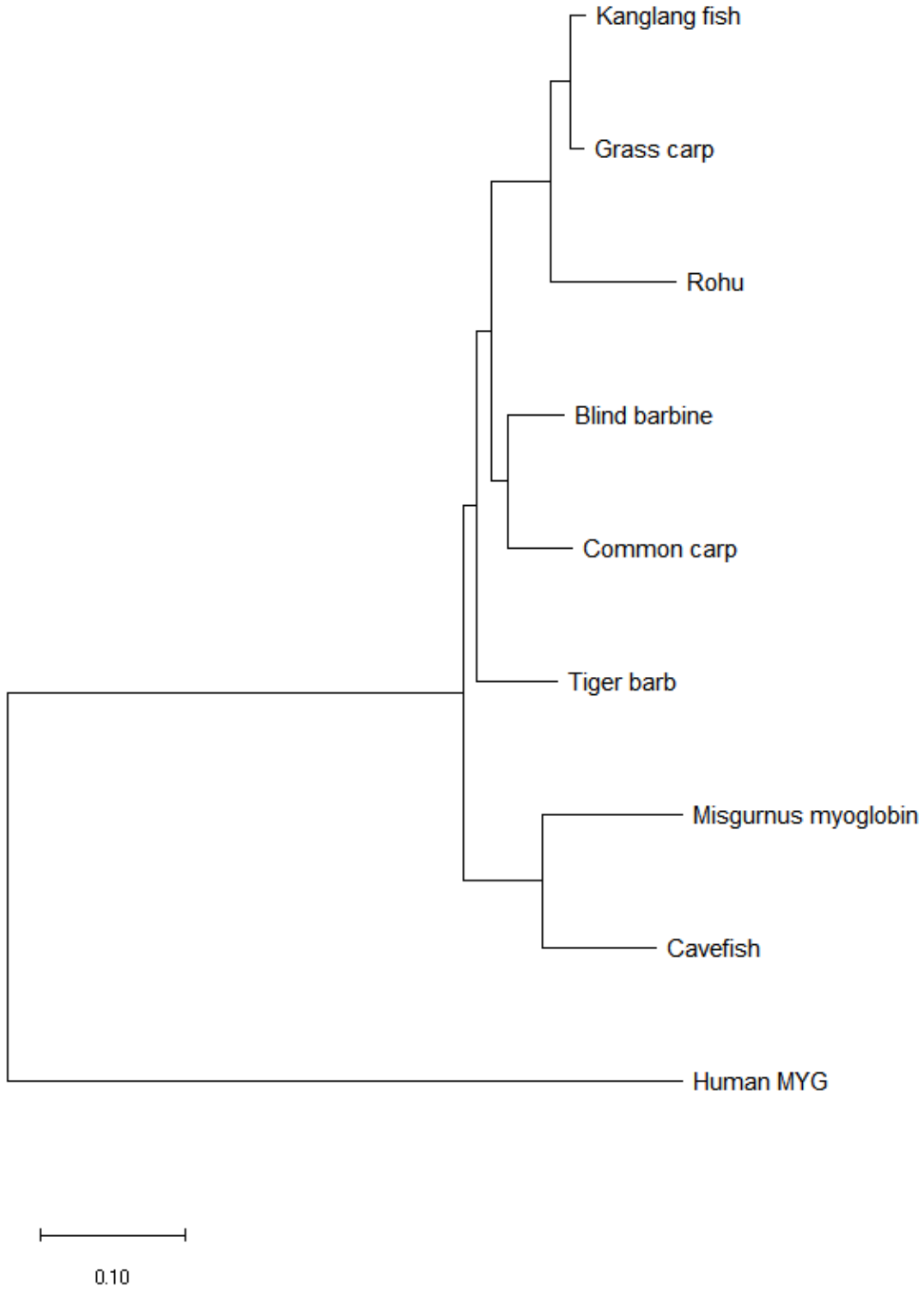
```
>Human_MYG gi|4885477|ref|NP_001349775.1|myoglobin isoform 1 [Homo sapiens]
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDLKKHGATVLTALGGILKKK
GHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG

>Misgurnus_myoglobin (translated using EMBOSS Transeq)
FQTHEHLDSSEQPLITTMSDFDLVLKCWGPVEADYTGVGGEVLTRLFKDHPETLKLFPKFVGIGQGDLAGNAAVAAHGATVLKKL
AELLRAKGEHAAVLKPLATTHANTHKIPLVNFKLITEALVKVMAERAGLDGAGQAALKRVMDVVIGDIDKYYKEIGYAGMRPNLS
RVYAG

>Cavefish gb|KAI7810808.1|myoglobin [Triplophysa rosa]
MADFDLVLKCWGAMEADYTAHGGEVLTRLFQEHPETLKLFPKFVGIAQGDLAGNAAVAAHGATVLKKLGDLLKAKGDHAG
ILKPLANTHANNHKIPLNNFKLITEIIVQLMAERAGLDGAGQAALRRVFDVVIGDIDGYYKEIGYAG

>Tiger_barb ref|XP_043080305.1| myoglobin [Puntigrus tetrazona]
MADFDQVLKCWGAVEADFAGHGGEVLTRLFKEHPETQKLFPKFVGISQSDLAGNAAVASHGATVLKKLGELLKARGDHAA
ILKPLATSHANIHKITLNNFRLITEVLVKVMAEKAGLDGAGQSALRRVMEVVIGDIDAYYKEIGFAG

>Blind_barbine ref|XP_016358141.1| PREDICTED: myoglobin [Sinocyclocheilus
anshuiensis]
MADHDLVLKCWGGVEADFEGHGGEVLTRLFKEHPETLKLFPKFVGIAQSDLVGNAAVAAHGATVLKKLGELLKARGDHAA
LLKPLATTHANTHKVALNNFRLITEVLVKVMAEKAGLDAAGQSALRRVMEAVIGDIDAYYKEIGFAG

>Common_carp ref|XP_018966946.2| myoglobin [Cyprinus carpio]
MADHELVLKCWGGVEADFEGTGGEVLTRLFKQHPETQKLFPKFVGIAQSDLAGNAAVKAHGATVLKKLGELLKARGDHAA
ILKPLATTHANTHKIALNNFRLITEVLVKVMAEKAGLDAGGQSALRRVMDVVIGDIDTYYKEIGFAG

>Rohu ref|XP_050965105.1|myoglobin [Labeo rohita]
MRGSDITWTLYKRRKLGKSDDLISFGEFSKPVTHSSERTPISTMAEHDQVLKYWGAIEADYTGNGGEVLTRLFKEYPDTQ
KLFPKFAGIAQSDLAGNAAVAAHGATVLKKLGELLKARGDHATILKPLANTHANTHKIALNNFRLITEVLVKVMAEKAGL
DAAGQAALRKIMDIVIGDIDRYYKEFGFAG

>Kanglang_fish gb|ROL52021.1| Myoglobin [Anabarilius grahami]
MADHELVLKCWGAVEADYTGHGGEVLTRLFKEYPDTLKLFPKFAGIAQSDLAGNAAVAAHGATVLKKLGELLKAKGDHAA
ILKPLANTHAKTHKIALNNFRLITEVLVKVMAEKAGLDAAGQSALRKVMDVVIGDIDGYYKEVGFAG

>Grass_carp ref|XP_051751285.1 myoglobin [Ctenopharyngodon idella]
MADHELVLKCWGAVEADYTGHGGEVLTRLFKEYPDTQKLFPKFVGIAQSDLAGNAAVAAHGATVLKKLGELLKAKGDHAA
ILKPLANSHAKTHKIALNNFRLITEVLVKVMAEKAGLDAAGQSALRKVMDVVIGDIDGYYKEVGFAG
```

## Alignment:

### Obtained using MUSCLE (version 3.8) at EBI:

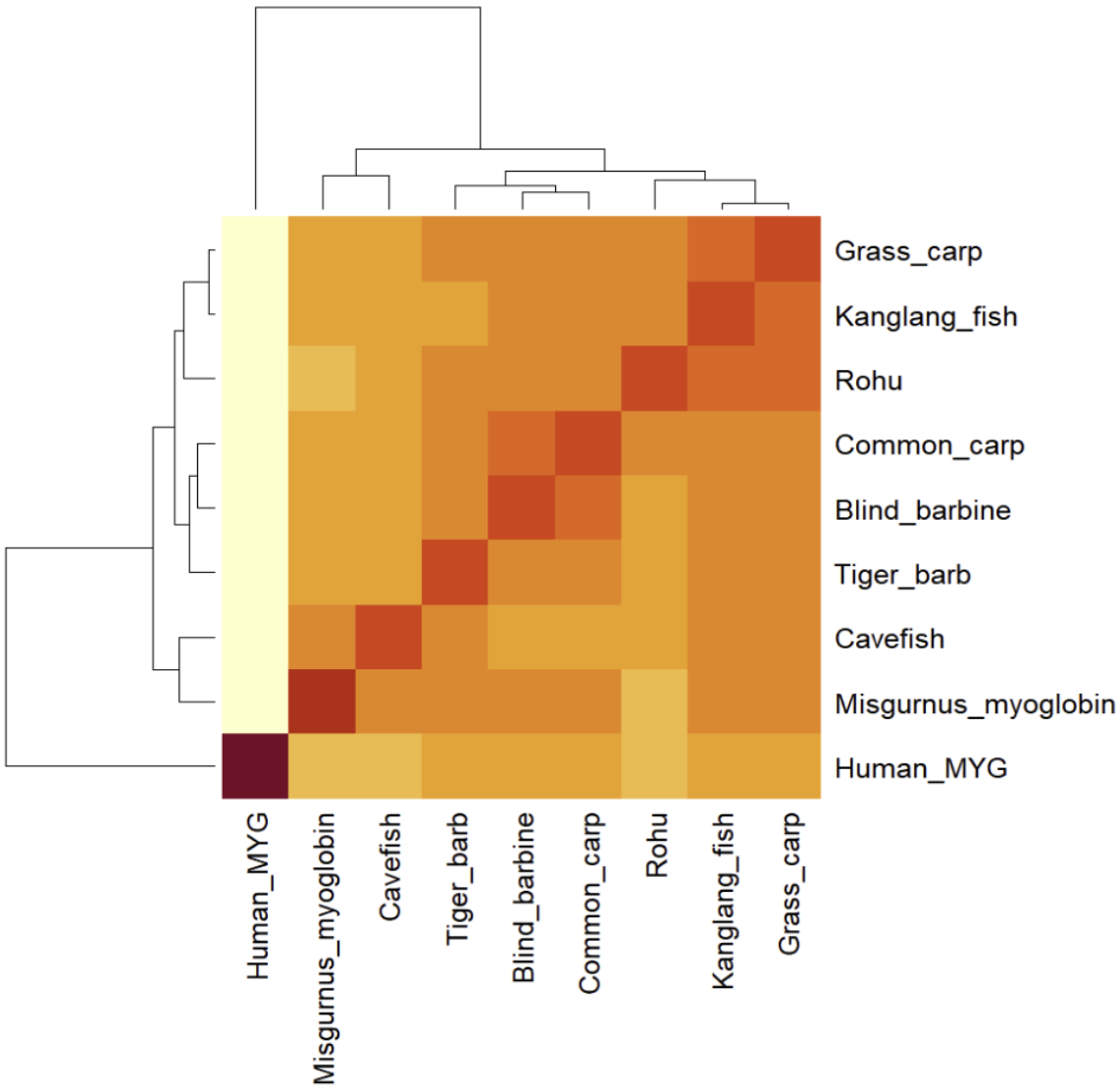
```
CLUSTAL multiple sequence alignment by MUSCLE (3.8)

Human_MYG                DGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDLKK
Tiger_barb               MADFDQVLKCWGAVEADFAGHGGEVLTRLFKEHPETQKLFPKFVGI-SQSDLAGNAAVAS
Blind_barbine            MADHDLVLKCWGGVEADFEGHGGEVLTRLFKEHPETLKLFPKFVGI-AQSDLVGNAAVAA
Common_carp              MADHELVLKCWGGVEADFEGTGGEVLTRLFKQHPETQKLFPKFVGI-AQSDLAGNAAVKA
Rohu                     MAEHDQVLKYWGAIEADYTGNGGEVLTRLFKEYPDTQKLFPKFAGI-AQSDLAGNAAVAA
Kanglang_fish            MADHELVLKCWGAVEADYTGHGGEVLTRLFKEYPDTLKLFPKFAGI-AQSDLAGNAAVAA
Grass_carp               MADHELVLKCWGAVEADYTGHGGEVLTRLFKEYPDTQKLFPKFVGI-AQSDLAGNAAVAA
Misgurnus_myoglobin      MSDFDLVLKCWGPVEADYTGVGGEVLTRLFKDHPETLKLFPKFVGI-GQGDLAGNAAVAA
Cavefish                 MADFDLVLKCWGAMEADYTAHGGEVLTRLFQEHPETLKLFPKFVGI-AQGDLAGNAAVAA

                          .: : **: ** :*** . * *** ***: :*:* : * **  : .:.:: ..

Human_MYG                HGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDF
Tiger_barb               HGATVLKKLGELLKARGDHAAILKPLATSHANIHKITLNNFRLITEVLVKVMAEK--AGL
Blind_barbine            HGATVLKKLGELLKARGDHAALLKPLATTHANTHKVALNNFRLITEVLVKVMAEK--AGL
Common_carp              HGATVLKKLGELLKARGDHAAILKPLATTHANTHKIALNNFRLITEVLVKVMAEK--AGL
Rohu                     HGATVLKKLGELLKARGDHATILKPLANTHANTHKIALNNFRLITEVLVKVMAEK--AGL
Kanglang_fish            HGATVLKKLGELLKAKGDHAAILKPLANTHAKTHKIALNNFRLITEVLVKVMAEK--AGL
Grass_carp               HGATVLKKLGELLKAKGDHAAILKPLANSHAKTHKIALNNFRLITEVLVKVMAEK--AGL
Misgurnus_myoglobin      HGATVLKKLAELLRAKGEHAAVLKPLATTHANTHKIPLVNFKLITEALVKVMAER--AGL
Cavefish                 HGATVLKKLGDLLKAKGDHAGILKPLANTHANNHKIPLNNFKLITEIIVQLMAER--AGL

                          :  ******. *. :*. .*  *   :**** :**. **:.:  : :*:* ::::: ..

Human_MYG                GADAQGAMNKALELFRKDMASNYKELGFQG
Tiger_barb               DGAGQSALRRVMEVVIGDIDAYYKEIGFAG
Blind_barbine            DAAGQSALRRVMEAVIGDIDAYYKEIGFAG
Common_carp              DAGGQSALRRVMDVVIGDIDTYYKEIGFAG
Rohu                     DAAGQAALRKIMDIVIGDIDRYYKEFGFAG
Kanglang_fish            DAAGQSALRKVMDVVIGDIDGYYKEVGFAG
Grass_carp               DAAGQSALRKVMDVVIGDIDGYYKEVGFAG
Misgurnus_myoglobin      DGAGQAALKRVMDVVIGDIDKYYKEIGYAG
Cavefish                 DGAGQAALRRVFDVVIGDIDGYYKEIGYAG

                         ..:.. .*.*:.. :: .  *:   **
```

[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).

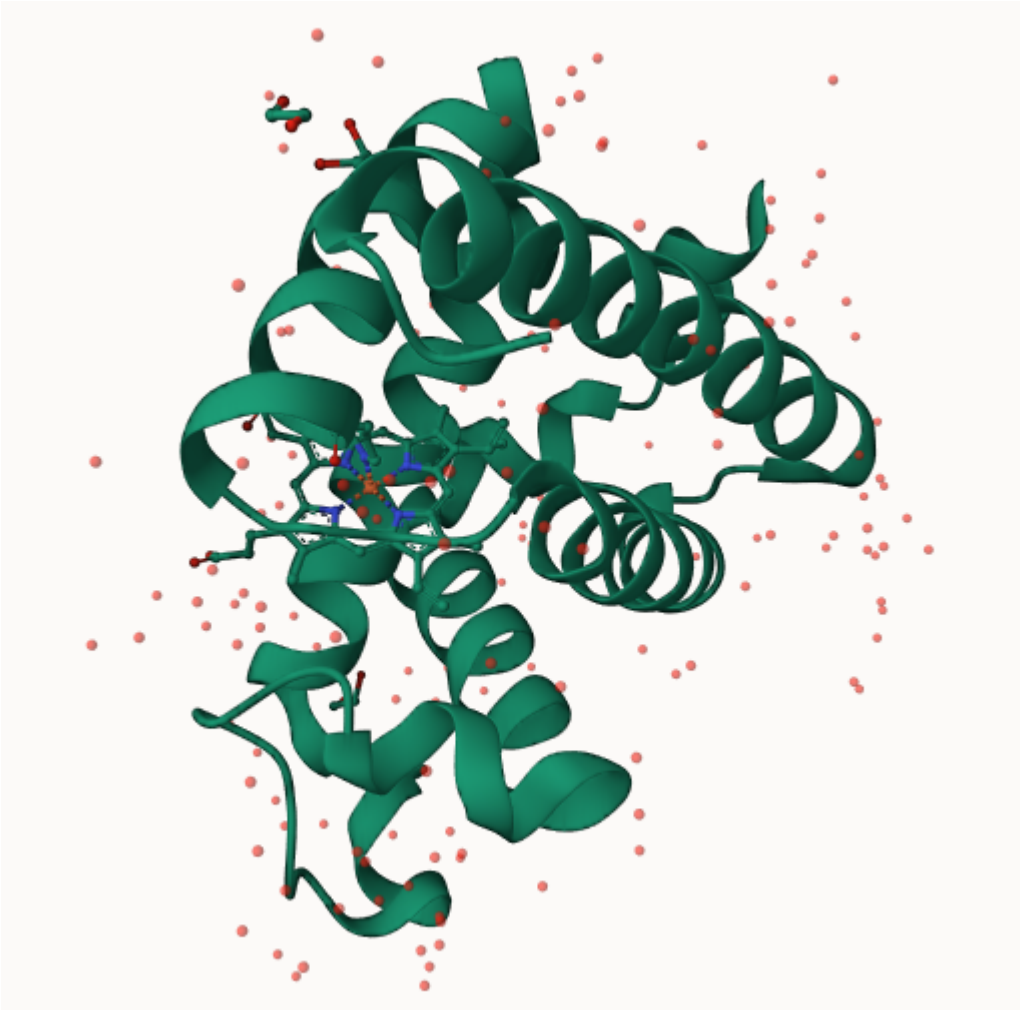
HINT: You can use a single sequence from your alignment or generate a consensus sequence from your alignment using the Bio3D function consensus(). The Bio3D functions blast.pdb(), plot.blast() and pdb.annotate() are likely to be of most relevance for completing this task. Note that the results of blast.pdb() 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 |
|---|---|---|---|---|---|
| 2NRL | X-RAY DIFFRACTION | 0.91 | Thunnus atlanticus | 4.15e-76 | 71.23 |
| 3QM5 | X-RAY DIFFRACTION | 0.91 | Thunnus atlanticus | 2.07e-75 | 71.034 |
| 7DDR | X-RAY DIFFRACTION | 1.50 | Escherichia coli | 2.39e-35 | 45.270 |

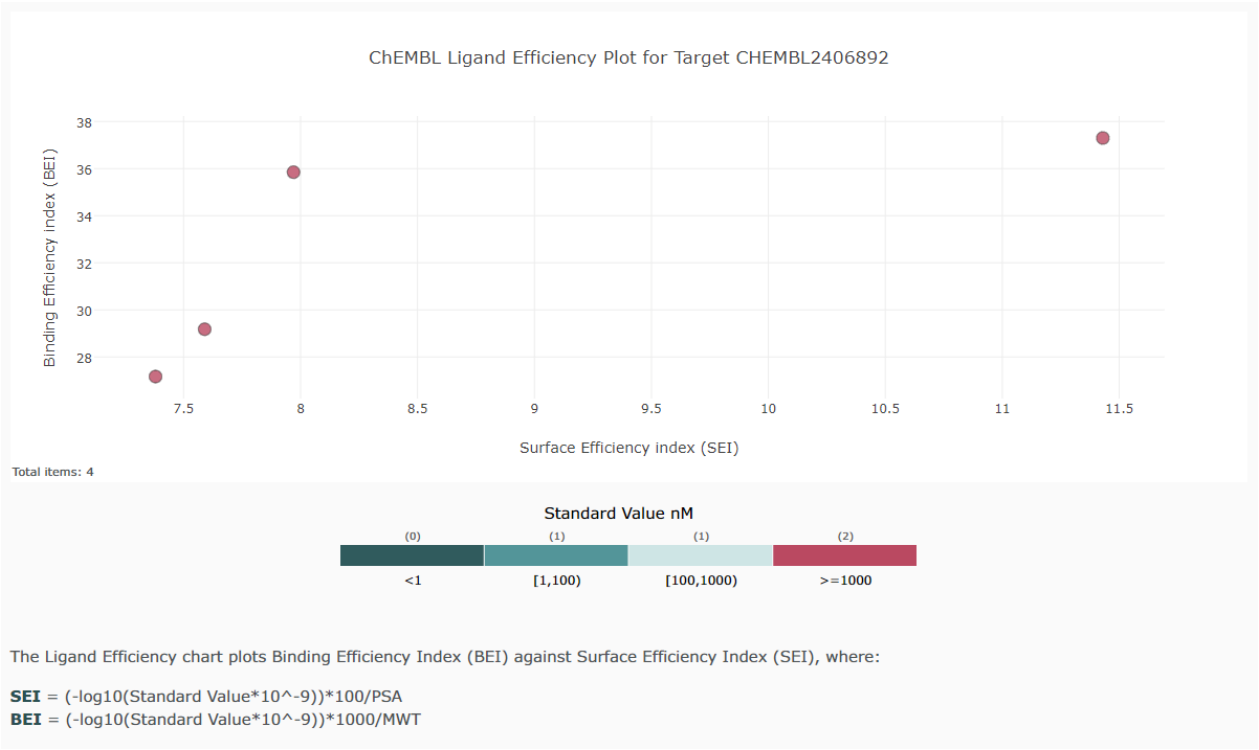
[Q9] Generate a molecular figure of one of your identified PDB structures using the **NGL viewer** online (or **VMD/PyMol)**. You can optionally highlight conserved residues that are likely to be functional. Please use a white or transparent background for your figure (i.e. not the default black).
Based on sequence similarity. How likely is this structure to be similar to your “novel” protein?

This structure is likely to be similar in structure to Misgurnus myoglobin given the high sequence similarity (>70%).



[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 1 Binding Assay (CHEMBL2406892) and 4 Functional Assays. 4 total items on ligand efficiency plot.




https://www.ebi.ac.uk/chembl/target_report_card/CHEMBL2406892/

Inhibition of myoglobin (unknown origin)-mediated arachidonic acid oxidation using [14C]AA as substrate after 3 hrs by GC/NICI/MS analysis

Shchepin, R. V., Liu, W., Yin, H., Zagol-Ikapitte, I., Amin, T., Jeong, B.-S., Roberts, L. J., Oates, J. A., Porter, N. A., & Boutaud, O. (2013). Rational design of novel pyridinol-fused ring acetaminophen analogues. ACS Medicinal Chemistry Letters, 4(8), 710–714. https://doi.org/10.1021/ml4000904

https://pubs.acs.org/doi/10.1021/ml4000904

**<u>Scoring Rubric</u>**:
[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