

BIMM-143: Introduction to Bioinformatics

The find-a-gene project assignment: Myocilin Protein*

Forrest Haolin Wang

PID: A15911047

*Restarted original project as initial gene wasn't novel

[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

Name: Myocilin Protein (from MYOC gene)
Accession: NP_000252.1
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 (2.13.0) search against
Database: Nucleotide Collection (nr/nt)
Organism: Brachydanio rerio frankei (taxid:7955)

Input:

The screenshot shows the NCBI BLAST search interface. The 'Enter Query Sequence' section has a text box containing 'NP_000252.1' and a 'Query subrange' section with 'From' and 'To' fields. Below this is an 'Or, upload file' section with a 'Choose File' button and 'No file chosen' text. The 'Job Title' field contains 'NP_000252:myocilin precursor [Homo sapiens]'. There is a checkbox for 'Align two or more sequences'. The 'Choose Search Set' section has a 'Database' dropdown set to 'Nucleotide collection (nr/nt)'. The 'Organism' section has a text box for 'Enter organism name or id--completions will be suggested' and an 'exclude' checkbox. The 'Exclude' section has checkboxes for 'Models (XM/XP)' and 'Uncultured/environmental sample sequences'. The 'Limit to' section has a checkbox for 'Sequences from type material'. The 'Entrez Query' section has a text box for 'Enter an Entrez query to limit search' and a 'YouTube Create custom database' link.

Enter Query Sequence

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

NP_000252.1

Query subrange [?](#)

From

To

Or, upload file

Choose File No file chosen [?](#)

Job Title

NP_000252:myocilin precursor [Homo sapiens]

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

☐ Align two or more sequences [?](#)

Choose Search Set

Database

Nucleotide collection (nr/nt) [?](#)

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) ☐ Uncultured/environmental sample sequences

Limit to

Optional

☐ Sequences from type material

Entrez Query

Optional

Enter an Entrez query to limit search [?](#) [YouTube](#) [Create custom database](#)

? Your search is limited to records that include: Brachydanio rerio krankei (taxid:7955)

Job Title	NP_000252:myocilin precursor [Homo sapiens]
RID	SJZGDS2F016 Search expires on 12-03 11:24 am Download All ▼
Program	TBLASTN ? Citation ▼
Database	nt See details ▼
Query ID	NP_000252.1
Description	myocilin precursor [Homo sapiens]
Molecule type	amino acid
Query Length	504
Other reports	?

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

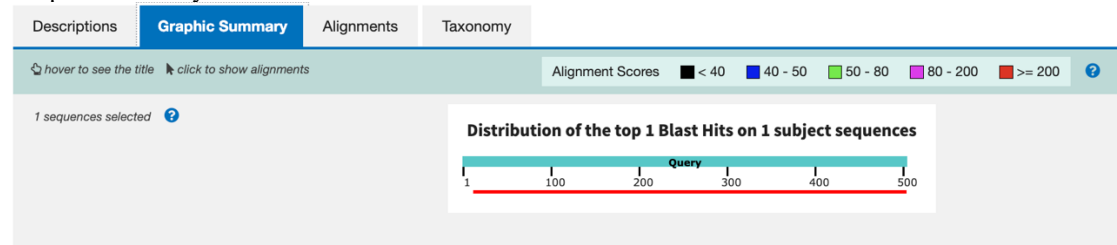
Download ▼ Select columns ▼ Show 100 ▼ ?

☒ select all 100 sequences selected

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

	Description ▼	Scientific Name ▼	Max Score	Total Score	Query Cover ▼	E value ▼	Per. Ident ▼	Acc. Len	Accession
<input checked="" type="checkbox"/>	Danio rerio myocilin mRNA complete cds	Danio rerio	414	414	93%	1e-138	43.78%	1425	AY551079.1
<input checked="" type="checkbox"/>	Danio rerio myocilin (myoc) mRNA	Danio rerio	415	415	97%	3e-134	42.72%	2524	NM_001015062.2
<input checked="" type="checkbox"/>	Danio rerio cDNA clone IMAGE:7267569 ***** WARNING: chimeric clone ****	Danio rerio	415	415	97%	3e-128	42.72%	4911	BC146643.1
<input checked="" type="checkbox"/>	Danio rerio strain Cooch Behar (CB) genome assembly chromosome: 20	Danio rerio	355	578	83%	4e-106	59.62%	51549904	LR812563.1
<input checked="" type="checkbox"/>	Danio rerio genome assembly chromosome: 20	Danio rerio	355	572	84%	4e-106	59.62%	52129346	LR812613.1
<input checked="" type="checkbox"/>	Danio rerio strain Nadia (NA) genome assembly chromosome: 20	Danio rerio	355	574	84%	4e-106	59.62%	54694995	LR812588.1
<input checked="" type="checkbox"/>	Danio rerio genome assembly chromosome: 20	Danio rerio	355	572	84%	4e-106	59.62%	55936343	LR812082.1

Chosen Match: Accession: NM_001015062.2



Alignment details:

>gb|NM_001015062.2| Danio rerio myocilin, mRNA (cDNA clone MGC:136645 IMAGE:8005695), complete cds.

Length = 2524 bp

Score = 415 bits (1066), Expect = 3e-134, Method: Compositional matrix adjust.

Identities = 223/522(43%), Positives = 320/522(61%), Gaps = 79/522(15%), Frame = +3

```
Query 13 PEMPAVQLLLLACLVWDVGAR---TAQLRKANDQSGRCQYTFSVASPNNESSCPEQSQAN
Sbjct 297 P M + +L ++ L+ +G++ +A LR+AN +GRCQYTF V SP E+SCP
PNMWFLAVLWISSLL--MGSQVQSSANLRRANAGNGRCQYTFMVDSPTASCP-----

Query 70 VIHNLQRDSSSTQRLDLEATKARLSSLESLLHQLTLDQAARPQETQEGQLQRELGLTLREF
S ++EA +RL LE+L+ +L +A
Sbjct 450 -----SPGSTPEMEALMSRLGLLEALVARLVGGEAMP-----

Query 130 QLETQTRELETAYSNLLRDKSVLEEEKKRLRQENENLARRLESSSQEVARLRRGQCPOI
++ L+ +Y+ ++ + + L+ EK+RL ++ ++L +R+E QE RLR C Q
Sbjct 549 SSQSSGSLQDSYNQVMGENAQLKREKQRLDRQVQDLQQRMEELRQEALRLSRPCMQ-

Query 190 DTARAVP-----PGSREVST-----WNLDTLAFQELKSELTE
T+ VP PGS V + W +QEL + +TE
Sbjct 723 QTSSRVPQKDNSFRPGSGHVPSNLASRPGNPQEDKSSLRDPAWQYSNPGYQELTAVVTE

Query 223 PASRILKESPSGYLRSGED-TGCGELVWVGEPLTLRTAETITGKYGVWMRDPKPTYPY
A + G D +GCG+LVWV P R A++I GKYGVWM+DP+ PY
Sbjct 903 TAPN-----QDGPADISGCGDLVWVENPEVHRKADSIA GKYGVWMQDPEAKEPY

Query 282 QETTWRIDTVGTDVRQVFEYDLISQFMQGYPSKVHILPRPLESTGAVVYSGSLYFQGA
+ WRID+VG++VRQ+F Y+ + Q +G+P+KV +LP +ESTGA +Y GSLY+Q
Sbjct 1053 PDMVWRIDSVGSEVRQLFGYENMDQLTRGFPTKVLLLPESESTGATMYKGSLEYQRR

Query 342 RTVIRYELNTETVKAKEIPGAGYHGQFPYSWG GYTDIDLAVDEAGLWVIYSTDEAKGA
RT+IRY+L+ E++ A +++P AG+HGQFPYSWG GYTDIDLA+DE GLW IYST++AKGA
Sbjct 1233 RTLIRYDLHAESIAARRDLPHAGFHGQFPYSWG GYTDIDLAIENGLWAIYSTNKAKGA

Query 402 VLSKLNPNENLELEQTWETNIRKQSVANAFIICGTLTVSSYTSADATVNFAYDTGTGIS
V+S+L+P NLE++ TWET IRK SVANAF+ICG LYTV+SYTS + TVN+ +DT T
Sbjct 1413 VISQLDPHNLEVKGTWETKIRKTSVANAFMICGKLYTVASYTSPNTTVNYMFDTATSQC

Query 462 TLTI PFKNRYKYSSMIDYNPLEKKLFAWDNLNMVTDIKLSK 503
+++PFKNRY+Y+SM+DYN ++KL+AWDN MV+Y ++L K
Sbjct 1593 AISV PFKNRYRYNSMVDYNSAKRKLYAWDNYYMVSYSVRLGK 1718
```

Comments: alignment looks a bit scarce in resemblance to Homo Sapiens with 15% gaps and only 61% positive identities. However, matching residues may imply key hydrophobic or polar folding sites. In my opinion, this query is within the range of a near match to non-homologous result, leaning more towards the latter. Being a near match indicates that this gene is novel.

[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.

Name: Dania rerio myocilin (myoc), mRNA

Species: Danio rerio

Animalia; Chordata; Teleostei; Cypriniformes; Cyprinidae; Danio; Danio rerio

Chosen Sequence ORF nucleotides:

```
ATGTATAAAGGCTCCTTATACTACCAGCGGAGGCTCAGCCGACCCTAATAAGATACGACCTACATGCTGAGAGCATTG
CTGCTCGTGTGATCTACCCCATGCTGGTTTCCATGGTCAGTTCCCCCTACTCATGGGGTGGCTACACAGACATTGACTT
GGCAATAGATGAAAATGGCTTATGGGCTATATACAGCACAAATAAAGCCAAGGGTGTATCGTGATCTCCAGCTGGAC
CCTCACAACTTGAGGTGAAGGGTACCTGGGAAACAAAATCCGCAAGACATCTGTGGCTAACGCTTTTATGATCTGTG
GCAAGCTCTACACAGTTGCTAGTTACACCTCACCACACACCGTTAATTACATGTTGACACTGCAACCAGCCAGGG
TAAAGCAATATCAGTGCCTTCAAAAACCGTTATCGCTACAACAGCATGGTAGACTACAACCTCCGCAAAAAGAAAGCTG
TATGCTTGGGATAATTATTACATGGTGTCTTACAGTGTGAGACTTGGCAAGCAGGAGTAA
```

Chosen Sequence ORF Translated (via EMBOSS Transeq)

```
MYKGSLEYQRRLSRTLIRYDLHAESIAARRDLPHAGFHGQFPYSWGGYTDIDLAI D ENGLWAIYSTNKAKGAIVISQLD
PHNLEVKG TWETKIRKTSVANAFMICGKLYTVASYTSPN TTVNMFDTATSQKAI SVPFKNRYRYNSMVDYNSAKRKL
YAWDNYYMVSYSVRLGKQE*
```

Chosen Sequence FASTA (sequence taken from BLAST result):

```
>NM_001015062.2 Danio rerio myocilin (myoc), mRNA
GGATCAACGACCAATCAAAGCAAGAAGCTCCAACGAGCCCTCGAAGGCAACGTTTAGCTTGACGAGAGAAAAGTCAGT
AGGAGCGAAAGATTGA AAAGACTCAGGTGA GAGGTCA TTAGTGGAAACGCCAGAGAAGTACGGGGCTCGGTTACACTTG
GAAGGTGA AAGGGAAGGAA TAGAT TAGTGC TAGAGCCGAAGAGGACCAGGAAGGACAAGTGA GCTGCAAGTCACCACTG
CATTA AACCTTGACAAAGGAGCTTCAGGCAACAGTAG AAAGAAGTTATCTGTGGTCCACCCCAACATGTGGTTTTTAGC
AGTGTGTGGATTTCTTCCCTGCTGATGGATCCCAGGTTTCAGAGCAGTGCTAA CCTTCGTCGAGCAAATGCTGGGAAT
GGTGCCTGTGTCAGTACACCTTTATGGTGGACAGCCCTACTGAGGCCAGCTGTCCATCACCAGGTTCACTCCATGATGG
AGGCTCTGATGTCTCGCTGGGGCTGCTAGAGGCACTGGTTGCTCGTCTCGTAGGAGGGGAAGCCATGCCAGAATCATC
ACAGAGCTCGGGATCTGGCCTTCAGGACTCTTACAACCAAGTGTATG GGGGAGAACGCGCAGCTCAAGAGAGAAAAGCAG
AGACTAGACAGACAAGTTCAGGACCTGCAGCAGAGGATGGAAGAGCTCCGCCAAGAGGCTGAGAGGCAGGAGGAGCAGAC
CCTGCATGCAGCAAATCTTCCAGAGTGCCGCAAAAAGACAACAGCTTCAGACCGGGATCAGGGCATGTACCCTCCAA
CTTGGCATCCAGACCTGGGAATCCACAAGAAGACAAAAGTAGTTTAA GAGACCCCGCATGGCAGTACTCAAATCCCGGA
TACCAAGAGTTGACGGCTGTGGTCACCGAGGTGACTGCCCCAAATCAGGACGGTCCAGCAGATATCTCAGGCTGTGGTGA
ATCTGGTGTGGGTGTA AAATCCTGAGGTGCATCGTAAGCTGATAGTATTGCTGGTAATATGGTGTGTGGATGCAAGA
TCCAGAAGCCAAGGAACCTTATGGTCCAGACATGGTATGGCGTATGATTCTGTCTGGTTCAGAGTGCCTCAACTCTTC
GGATAAGAAACATGGACAGCTGACAGCTGGCTTTCCACCAAGGTTCTACTCCTGCCAGAATCTGTAGAGCAGACAG
GTGCTACCATGTATAAAGGCTCCTTATACTACCAGCGGAGGCTCAGCCGCAACCCTAATAAGATACGACCTACATGCTGA
GAGCATTGCTGCTCGTGTGATCTACCCCATGCTGGTTTCCATGGTCAGTTCCCCCTACTCATGGGGTGGCTACACAGAC
ATTGACTTGGCAATAGATGAAAATGGCTTATGGGCTATATACAGCACAAATAAAGCCAAGGGTGTATCGTGATCTCCC
AGCTGGACCCCTCACAACTTGAGGTGAAGGGTACCTGGGAAACAAAATCCGCAAGACATCTGTGGCTAACGCTTTTAT
GATCTGTGGCAAGCTCTACACAGTTGCTAGTTACACCTCACCACACACCGTTAATTACATGTTTGACACTGCAACC
AGCCAGGGTAAAGCAATATCAGTGCCTTCAAAAACCGTTATCGCTACAACAGCATGGTAGACTACAACCTCCGCAAAA
GAAAGCTGTATGCTTGGGATAATTATTACATGGTGTCTTACAGTGTGAGACTTGGCAAGCAGGAGTAACATTGTAAAT
TTCCATATCATCAAAGCTTGCAATTTTTTTTTTTTACTTA AACATCCAGTAGGACGTCAATTGTTGGACTTACTGATTT
TCAAAGACAATGA ACTTTTTTTTCTGA AAAAAGAACTGGATAAGTCAACCAATCAGGCTTTTGTGGTATTTGTATAGG
AAACGATGTGTGTGTGTGTATACCAAAATAAACACTTTCCTGGTAACATAAAAACAGAGTTTTTTCAAGATTTTCCC
ATTATATACAGTAATACCTTACAATTAACCTGCATTGTATATCCTTCTCGATCATAGATTATCTTGTATAAAAAAGA
TGTGACCAACAAGTTCCAAAATGATTAATGTATCAATCATGTCGATATTCACATCCTTAATAATTTTATCCACTTAA
CATGTTAATAGAGAAATATTTTTTACCTATGTAGTAATGTCTTGTCTGTTTTTGTGTATATGCACTGAATAAATT
CAAGGAAGGGGGAAAAATATACAGCAAACAACGATTACAGCAAGCCTGCTTCTTAATTCATATATCATTGAGCAGTAC
GCAACTTTACGAACCATAACCTTGTGTTGTGTTGTGCTTAATACTTTATCATATAATCTTCACTCTGTATATGAG
TGATCATTTTCCCTTCTCTCTTTGATTTGCTGCAAACTCTCACATGCTGTGATGTAA GATTGGGATTTTGTGG
CTTTGGAAGCGTATTAATGATAAAGATTTAAGTCTGCCAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
```

[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 Input:

blastn

blastp

blastx

tblastn

tblastx

BLASTP programs search protein databases using a protein query. more...

Reset page

Bookmark

Enter Query Sequence

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

>NM_001015062.2_1 Danio rerio myocilin (myoc), mRNA
GSTTNQSKKLQPAPRRQRLA*ARESQ*ERKIEKTQVRGH*WNAREVRGSGYTW
KVKGKE*ISARAEEDQE
GVVSCKSPLH*TLDELQATVERSILWSTPTCGF*QCCGFLPC*WDPFRFVAVLT;
Query subrange ?
From
To

Or, upload file

Choose File No file chosen ?

Job Title

NM_001015062.2_1 Danio rerio myocilin (myoc),...

Enter a descriptive title for your BLAST search ?

☐ Align two or more sequences ?

Choose Search Set

Databases

☒ Standard databases (nr etc.): **Now** ☐ Experimental databases

[Try experimental clustered nr database](#)
For more info see [What is clustered nr?](#)

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 ?


BLASTP Output:


Descriptions	Graphic Summary	Alignments	Taxonomy					
Sequences producing significant alignments								
Download		Select columns	Show 100					
<input checked="" type="checkbox"/> 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
<input checked="" type="checkbox"/> myocilin-like [Carassius auratus]	Carassius ...	356	356	99%	3e-119	93.22%	476	XP_026147181.1
<input checked="" type="checkbox"/> hypothetical protein cypCar_00009366 [Cyprinus carpio]	Cyprinus c...	355	355	99%	4e-119	93.22%	476	KTF90649.1
<input checked="" type="checkbox"/> PREDICTED: myocilin-like [Sinocyclocheilus grahami]	Sinocycloc...	355	355	99%	5e-119	93.22%	476	XP_016122537.1
<input checked="" type="checkbox"/> PREDICTED: myocilin-like [Sinocyclocheilus grahami]	Sinocycloc...	350	350	99%	7e-119	92.09%	332	XP_016144475.1
<input checked="" type="checkbox"/> myocilin isoform X2 [Puntigrus tetrazona]	Puntigrus t...	353	353	99%	1e-118	92.09%	454	XP_043076193.1
<input checked="" type="checkbox"/> myocilin isoform X1 [Puntigrus tetrazona]	Puntigrus t...	354	354	99%	2e-118	92.09%	476	XP_043076192.1
<input checked="" type="checkbox"/> PREDICTED: myocilin-like [Sinocyclocheilus rhinocerosus]	Sinocycloc...	353	353	99%	2e-118	92.66%	476	XP_016373720.1
<input checked="" type="checkbox"/> Myocilin [Anabarrilius grahami]	Anabarriliu...	353	353	99%	3e-118	93.22%	476	ROL55559.1
<input checked="" type="checkbox"/> myocilin isoform X1 [Cyprinus carpio]	Cyprinus c...	349	349	99%	5e-117	90.96%	454	XP_042602651.1


Comment: Top result from Carassium auratus (Goldfish)


Alignments:



[Download](#)


[GenPept](#)


[Graphics](#)


[Next](#)


[Previous](#)


[Descriptions](#)

myocilin-like [Carassius auratus]

Sequence ID: [XP_026147181.1](#)
Length: 476
Number of Matches: 1

Range 1: 300 to 476
[GenPept](#)
[Graphics](#)

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

Score	Expect	Method	Identities	Positives	Gaps
356 bits(913)	3e-119	Compositional matrix adjust.	165/177(93%)	173/177(97%)	0/177(0%)
Query 1	MYKGSLEYQRRLSRTLIRYDLHAESIAARRDLPHAGFHGQFPYSWGGYTDIDLAI	DENGI			
Sbjct 300	MY+GSLYYQRRLSRTL+RYDL +ESIAARRDLPHAGFHGQFPYSWGGYTDIDLAI	DENGI			
Query 61	WAIYSTNKAKGAIVISQLDPHNLEVKG	TWETKIRKTSVANAFMICGKLYTVASYTSPNTI			
Sbjct 360	WAIYSTNKAKGAIVISQLDPHNLEVKG	TWETKIRKTSVANAFMICGKLYTVASYTSPNTI			
Query 121	VNYMFDTATSQGKAISVPFKNRYRYNSMVDYNSAKRRLYAWDNYYMVSYSVRLGKQE	1			
Sbjct 420	INMYDTATSQGTISVPFKNRYRYNSMVDYNPTQRKLYAWDNFYMVSYNVRLGKQE	4			

Related Information

[Gene](#) - associated gene details
 [Genome Data Viewer](#) - aligned genomic context

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_Myocilin gi|35014583|ref|NP_000252.1| myocilin precursor [Homo sapiens]EPARPPIQAPLSTAELSRGSLTKPLQ*GSSVHVAAALGLRCLSSCCFWPAWCGMWGPQLSSGRPMTRVADASIPSVVPVPM NPAAQSRARPCQSSITYRETAAPNA*TWRPPKLD SAPWRASSTN*PWTRLPGRRRPRRCGRGSWAP*GGSGTSWKPKPESWRLPTATSET SQFWRKRRSD*GKKMRIWPGGWKAAARR*QG*EGASVPRPETLLGLCHQAPEKFLRGIWTLWP SRN*SPS*LKFLLP EF*RRHLAISGVE RETPDVEN*FG*ESLSR*EQQKQLLASMVCGCETPSPTPTPRRPRGESTQLARMSARFLSMTSSASLCRATLLRFTYCLGHWKARVLWCT RGASISRALSPELS*DMS*IPRQ*RLRRKSLELATDSSRILGVATRTLTLWLMWKQASGSFTAPMRPKVPLSSPN*TQRIWNSNKPGRQTS VSSQSPMPSSSVAPCTPSAATPQQMLPSTLLMTQAQVSARP*PSHSRTAISTAA*LTTPWRRSSLP GTT*TWSLMTSSSPRCEKPPSCTG NGRRRCSGLLGAG*RESQPARAQALTAQVFINPEG*TWSPSNYSIGIVV*GRRQFHIINILYLLSAFMGCLMT*FKFSCDLGQKL*GII VSS*KPLLLHVWLPQATIKSITSKGSRIAPLASIEYK*DAFTTVGF*CFR*NTVGSHITLYIVK*NFLTQ
```

```
>Zebrafish gi|33573230|ref|NP_001015062|myocilin precursor [Danio rerio]* ATGTATAAAGGCTCCTTATACTACCAGCGGAGGCTCAGCCGCACCCTAATAAGATACGACCTACATGCTGAGAGCATTGCTGCTCGTCGTG ATCTACCCCATGCTGGTTTCCATGGTCAGTTCCTTACTCATGGGGTGGCTACACAGACATTGACTTGGCAATAGATGAAAATGGCTTATG GGCTATATACAGCACAAATAAAGCCAAGGGTGCTATCGTGATCTCCAGCTGGACCCCTACAACCTTGAGGTGAAGGGTACCTGGGAAACA AAAATCCGCAAGACATCTGTGGCTAACGCTTTATGATCTGTGGCAAGCTACACAGTTGCTAGTTACACCTCACCAACACCACCGGTTA ATTACATGTTTGACACTGCAACCGCCAGGGTAAAGCAATATCAGTGCCCTTCAAAAACCGTTATCGCTACAACAGCATGGTAGACTACAA CTCGCAAAAAGAAAGCTGTATGCTTGGGATAATTATTACATGGTGTCTTACAGTGTGAGACTTGGCAAGCAGGAGTAA *(sequence taken from BLAST result)
```

```
>Goldfish gi|113121143|ref|XP_026147181|myocilin-like [Carassius auratus] MYFLAMLWASCLLMGTHAQGSASFRRANAGSGRCQYTFMVDSPTEASCP SAGSSPEVEALKSRLGLLEALVARLAGGEAMSESSHGSGSQS GLQDAYNQAMGENARLQREKQRLDRQVQDLQQRMEELRQEAE RLRSRPCMQPPPRVPQNDNSFRPGSGPAVSQVLSRPGTTQGD KSSLRD PAWHYENPGYQEVTA VTEVSAPNQE GPADIPGCGDLVWVKEPEVHRKADS IAGKYGVWMQDPEAKEPYGAEMVWRIDSVGSEVRQLFGYE NMDQLSRGFPTKVLLLPE SMESTGATMYRGS LYYQRRLSRTLRLYD LLSESIAARRDLPHAGFHGQFPYSWGGYTDIDLAI DENGLWAIYS TNKAKGAIVISQLDPHNLEVKG TWETKIRKTSVANAFMICGKLYTVASYTSPNTTINMYDTAT SQGKTIISVPFKNRYRYNSMVDYNPTQR KLYAWDNFYMVSYNVRLGKQE
```

```
>Carp gi|27590662|ref|KTF90649| hypothetical protein cypCar_00009366 [Cyprinus carpio] MWFLVMLWASCLLMGTHAQSSASFRRANAGSGRCQYTFMVDSPTEASCP SAVSSPEIEALKSRLGLLEALVARLIGGEAMSKSTQSSGSQS GLQDAYSQVMGENAQLQREKQRLDRQVQDLQQRMEELRQEAE RLRSRPCMQPPPRVPQNDNSFRPGSGPALSHLVS RPNTQGD KSSLRD PAWHYENPGYQELTAVVTEVSAPNLEG PADISGCGDLVWVQEPEVHRKADS IAGKYGVWMQDPEAKEPYGPEMVWRIDAVGSEVRQLFGYE NMDQLSRGFPTKVLLLPE SMESTGATMYRGS LYYQRRLSRTLRLYD LLSESIAARRDLPHAGFHGQFPYSWGGYTDIDLAI DENGLWAIYS TNKAKGAIVISQLDPHNLEVKG TWETKIRKTSVANAFMICGKLYTVASYTSPNTTINMYDTAT SQGKTIISVPFKNRYRYNSMVDYNPAQR KLYAWDNFYMVSYNVRLGKQE
```

```
>Tiger Barb gi|1606681|ref|XP_043076193|myocilin isoform X2 [Puntigrus tetrazona]MWFLAVFWASCLLMGTHAQSSASFRRANTGSGRCQYTFMVDSPTEASCP SAGSSPEMEALKSRLGLLEALVARLVGGEAVP ELSQGSRPQSLQDAYNQMMGENTQLQREKQKLD RQVQDLQQRMEELRQEAE RLRSRPCMQPPPRVPQNDNSFRPGSDPSWHYENPGYQE LTAVVTEVSAPSPPEG PADISGCGDLVWVQEPEVHRKADNIAGKYGVWMQDPEAKEPYGTEMVWRIDAVGSEVRQLFGYENMDQLSRGFPTK VLLLPEFMESTGATMYRGS LYYQRRLSRTLRLYD LLSENIAARRDLPHAGFHGQFPYSWGGYTDIDLAI DENGLWAIYSTNKAKGAIVISQ LDPHNLEVKG TWETKIRKTSVAN SFMICGKLYTVASYTSPNTTINMYDTAT SQGKTIAVPFKNRYRYNSMVDYNPAQRKLYAWDNFYMVS YNVRLGKQE
```

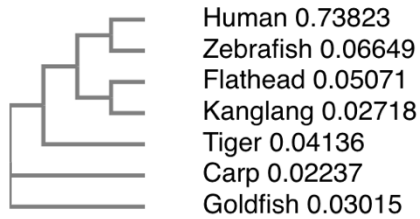
```
>Kanglang_Fish gi|495550|ref|ROL55559.1|Myocilin [Anabarrilius grahami] MWFLAVLCVSCLLMGTAQSSASFRRANAGNGRCQYSFTVDSPTEASCP SPSSPEMEALKSRLGLLEALVARLVGGEAVSESSQSGSQS GLQDAYNQMLGENAQLQREKQRLDRQVQDLQQRMEELRQEAE RLRSRPCVQPPPRVPQNDNSFRPGSGPALSHLVSSPGNTQGDTSSLRD PAWHFSNPEYQELTAVVTEVSAPNLEGPTDISGCGDLVWVQEP EHVHRKADS IAGKYGVWMQDPEAKEPYGPEMVWRIDAVGSEVRQLFGYE NMDQLSRGFPTKVLLLPE SVESTGATMYRGS LYYQRRLSRTLRLYD LLSESIAARRDLPHAGFHGQFPYSWGGYTDIDLAI DENGLWAIYST NKAKGAIVLSQLDPHNLEVKG TWETKIRKTSVANAFMICGKLYTVASYTSPNTTVNYMYDTAT SQSKTISVPFKNRYRYNSMVDYNPGQR KLYAWDNYIIVSYKVRLGKQE
```

```
>Flathead Minnow gi|120494087|ref|XP_39549125|myocilin [Pimephales promelas] MRFLAVLCVSCLLVGTQAQSSASFRRANAGNGRCQYSFTVDSPTEASCP SSGSSPEMEALKSRLGLLEALVARLVGGEAVSESSAGRSQS GLQDAYNQVMGEKAQIQREKQRLQVQDLQQRMEELRQEAE RLRSRPC TQKPPPRVSEIDSSFKPGSGPALSHLVS RPNTQGDTSSLRD PAWHSSSGYQELTAVVTEVSAPNLEG PADISGCGDLVWVDQPEMHRKADS IAGKYGVWMQDPEAKEPYGPEMVWRIDAVGSEVRQLFGYEN MDQLSRGFPTKVLLLPE SVESTGATMYRGS LYYQRRLSRTLRLYD LLSESIAARRDLPHAGFHGQFPYSWGGYTDIDLAI DENGLWAIYST NKVKGAIVLSQLDPHNLEVKG TWETKIRKTSVANAFMICGKLYTVASYI LPNTTINMYMYDTAT SQGKTIISVPFKNRYRYNSMVDYNPGQR KLYAWDNYIIVSYKVRLGKQL
```

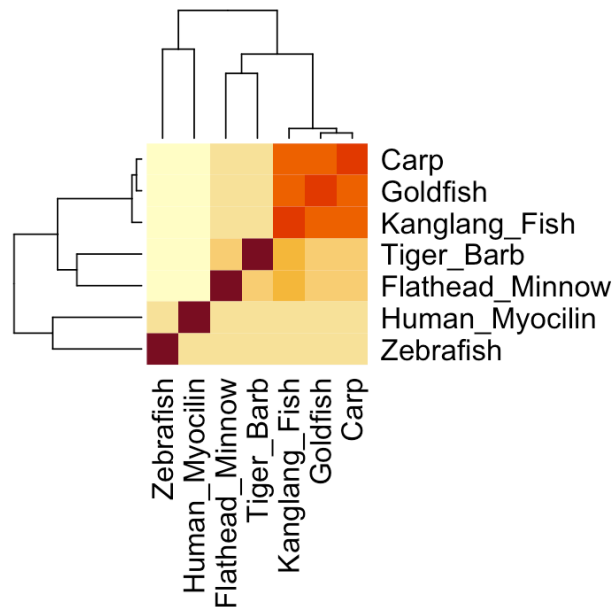
CLUSTAL multiple sequence alignment by MUSCLE (3.8)

Human	EPARPPIQAPLSTAELSRGSLTKPLQGSVVHVAALGLRCQLSSCCFWPAWCGMWGPQLSSGRPMTRVADASIPSVVPWF
Zebrafish	----MWFLAVLWISSLLMGSQVQS--SANLRANAGNGRCQYTFMVDSF-----TEASCPSPGSTP
Tiger	----MWFLAVFWASCLLMGTAAQS--STSFRRANTSGSRCQYFTTVDSP-----TEASCPSAGSSP
Carp	----MWFLVMLWASCLLMGTAAQS--SASFRRANAGSGRCQYTFMVDSF-----TEASCPSAVSSP
Goldfish	----MYFLAMLWASCLLMGTAAQS--SASFRRANAGSGRCQYTFMVDSF-----TEASCPSAGSSP
Flathead	----MRFLAVLCVSCLLLVGTQAQS--SASFRRANAGNGRCQYSFTVDSP-----TEASCPSGSSP
Kanglang	----MWFLAVLCVSCLLMGTAQS--SASFRRANAGNGRCQYSFTVDSP-----TEASCPSGSSP .: .: .: .: .: .: .: .: .: .
Human	MNPAAQSRARPQSSITYRETAAPNATWRPKLDAPWRASSTNPWTRLPGPRRPRRGCRGSWAPGGSGTSWKPKPESWRL
Zebrafish	EMEALMSRLGLLEALV-----ARLVGEAMPESSQSSG--SGLQDSY-----
Tiger	EMEALKSRLGLLEALV-----ARLVGGEAVPELSQGSRPQSGLODAY-----
Carp	EIEALKSRLGLLEALV-----ARLIGEAMSKSTQSSGSQSGLODAY-----
Goldfish	EVEALKSRLGLLEALV-----ARLAGEAMESSSHGSGSQSGLODAY-----
Flathead	EMEALKSRLGLLEALV-----ARLVGGEAVSESSAGRSQSGLODAY-----
Kanglang	EMEALKSRLGLLEALV-----ARLVGGEAVSESSQSGGSQSGLODAY----- .:**.* . * ** .: :
Human	PTATSSETSQFWRKRSDDRQVQDLQQ--RMEELRQEAEERLSRP-----CMQQTSSRVPKQDNSFRPGSGHVPSNLLPE
Zebrafish	-NQVMGENAQLKREKQRDLRQVQDLQQ--RMEELRQEAEERLSRP-----CMQQTSSRVPKQDNSFRPGSGHVPSNLASR
Tiger	-NQMMGENTLQREKQKLDRQVQDLQQ--RMEELRQEAEERLSRP-----CMQQPPrVPQNDSFRPGSDPS-----
Carp	-SQVMGENAQLQREKQRDLRQVQDLQQ--RMEELRQEAEERLSRP-----CMQQPPrVPQNDSFRPGSGPALSHLVS
Goldfish	-NQAMGENARLQREKQRDLRQVQDLQQ--RMEELRQEAEERLSRP-----CMQQPPrVPQNDSFRPGSGPAVSQLVSR
Flathead	-NQVMGEKAQIQREKQRLEKQVQDLQQ--RMEELRQEAEERLSRP-----CTQKPPrVSEIDSSFPGSGPALSHLVS
Kanglang	-NQLMGENAQLQREKQRDLRQVQDLQQ--RMEELRQEAEERLSRP-----CVOQPPrVPQNDSFRPGSGPALSHLVSS :**.* .* .: .: .: .: .: .: .: .
Human	FRAHLAISGVRETDPDENFGESLSREQQKQLLASMVCGCTPSPTPTTPRRPRGESTQLARMSARFLSMTSSASLCRAT
Zebrafish	PGNPQEDKSSLRDPAWQYSNPG-----YQELTAVVTEVTAPNQDGPADISGCCD-----
Tiger	-----WHYENPG-----YQELTAVVTEVSAPSPEGPADISGCCD-----
Carp	PGNTQGDKSSLRDPAWHYENPG-----YQELTAVVTEVSAPNLEGPADISGCCD-----
Goldfish	PGTTQGDKSSLRDPAWHYENPG-----YQEVTAIVTEVSAPNLEGPADIPGCCD-----
Flathead	PGNTQGDTSLLRDPAWH-SSSG-----YQELTAVVTEVSAPNLEGPADISGCCD-----
Kanglang	PGNTQGDTSLLRDPAWHFSNPE-----YQELTAVVTEVSAPNLEGPTDISGCCD----- .: .: .: .: .: .: .: .
Human	LLRFTYCLGHWKARVLWCTRGASISRALSPELSDMSIPRQLRRKSLELATTDSSRIILGVATRITLWLWMKQASGSFTAPM
Zebrafish	-----LVW-----VENPEVHRKA-----DS-----IAGKYGVWMQDPEAKEPYGPD
Tiger	-----LVW-----VQEPVHRKA-----DN-----IAGKYGVWMQDPEAKEPYGTEM
Carp	-----LVW-----VQEPVHRKA-----DS-----IAGKYGVWMQDPEAKEPYGPDM
Goldfish	-----LVW-----VKEPEVHRKA-----DS-----IAGKYGVWMQDPEAKEPYGAEM
Flathead	-----LVW-----VDQPEMHRKA-----DS-----IAGKYGVWMQDPEAKEPYGPDM
Kanglang	-----LVW-----VEGPEVHRKA-----DS-----IAGKYGVWMQDPEAKEPYGPDM .:**
Human	RPKVPLSSPNTQRIW NSNKPGRQTSVSSQSPMPSSSVAPCTPSAATPQOMLPSTLLMTQAQVSARPPSHSRTAITAALT
Zebrafish	VVRIDSVGSEVRQLF-----GYENMDQLTRGFPTKVLLLPESVESTGATMYKGSLEYQR-----RLSRTLIRYDLLSE
Tiger	VVRIDAVGSEVRQLF-----GYENMDQLSRGFPPTKVLLLPFMESTGATMYRGSLEYQR-----RLSRTLIRYDLLSE
Carp	VVRIDAVGSEVRQLF-----GYENMDQLSRGFPPTKVLLLPESMESTGATMYRGSLEYQR-----RLSRTLIRYDLLSE
Goldfish	VVRIDSVGSEVRQLF-----GYENMDQLSRGFPPTKVLLLPESMESTGATMYRGSLEYQR-----RLSRTLIRYDLLSE
Flathead	VVRIDAVGSEVRQLF-----GYENMDQLSRGFPPTKVLLLPESVESTGATMYRGSLEYQR-----RLSRTLIRYDLLSE
Kanglang	VVRIDAVGSEVRQLF*-GYENMDQLSRGFPPTKVLLLPESVESTGATMYKGSLEYQR-----RLSRTLIRYDLLSE .:**.* .* .: .: .: .: .: .: .
Human	TPWRRSSLPGTTT-----WSLMTSSSPRCEKPPSCCTGNRRRCSSGLLGAGRESQPARAQALTAFOVF-INPEGTWP
Zebrafish	SIAARRDLPHAGFHGQFPYSWGgyTDIDLAI-----NLGW-AIYSTNKAkGAIVISQLDPHNLEVKGtWET
Tiger	NIAARRDLPHAGFHGQFPYSWGgyTDIDLAI-----NLGW-AIYSTNKAkGAIVISQLDPHNLEVKGtWET
Carp	SIAARRDLPHAGFHGQFPYSWGgyTDIDLTI-----NLGW-AIYSTNKAkGAIVISQLDPHNLEVKGtWET
Goldfish	SIAARRDLPHAGFHGQFPYSWGgyTDIDLAI-----NLGW-AIYSTNKAkGAIVISQLDPHNLEVKGtWET
Flathead	SIAARRDLPHAGFHGQFPYSWGgyTDIDLAI-----NLGW-AIYSTNKAkGAIVISQLDPHNLEVKGtWET
Kanglang	SIAARRDLPHAGFHGQFPYSWGgyTDIDLsvDE-----NLGW-AIYSTNKAkGAIVISQLDPHNLEVKGtWET .: .* .**.* .* .: .: .: .: .
Human	SNYSgIVVGRrQHFIINILYLLSAfMGCLMTFFKfScDLGqKlGIIVSSKPlLLHvTWLPQAtIKSiTSKgSRIAPLASIEY
Zebrafish	KIRKTSVANA-FMICGKLYTVAsyTSpntTVNmYfdTatsQgKaIs-----VPFKnRYrNSmVDYNPaQRkLYAw--
Tiger	KIRKTSVANS-FMICGKLYTVAsyTSpntTiNmYdTatsQgKTIA-----VPFKnRYrNSmVDYNPaQRkLYAw--
Carp	KIRKTSVANA-FMICGKLYTVAsyTSpntTiNmYdTatsQgKTIA-----VPFKnRYrNSmVDYNPaQRkLYAw--
Goldfish	KIRKTSVANA-FMICGKLYTVAsyTSpntTiNmYdTatsQgKTIA-----VPFKnRYrNSmVDYNPaQRkLYAw--
Flathead	KIRKTSVANA-FMICGKLYTVAsyILpntTiNmYdTatsQgKTIA-----VPFKnRYrNSmVDYNPaQRkLYAw--
Kanglang	KIRKTSVANA-FMICGKLYTVAsyTSpntTVNmYdTatsQgKTIA-----VPFKnRYrNSmVDYNPaQRkLYAw-- .: .. .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .
Human	KDAFTTVGFCFR NTVGShITLyIVKNflTQ
Zebrafish	-DNYIMVSYSVR--LGKQE-----
Tiger	-DNFYMVSYNVR--LGKQE-----
Carp	-DNFYMVSYNVR--LGKQE-----
Goldfish	-DNFYMVSYNVR--LGKQE-----
Flathead	-DNYIVSYKVR--LGKQL-----
Kanglang	-DNYIVSYKVR--LGKQE----- *: :

[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.



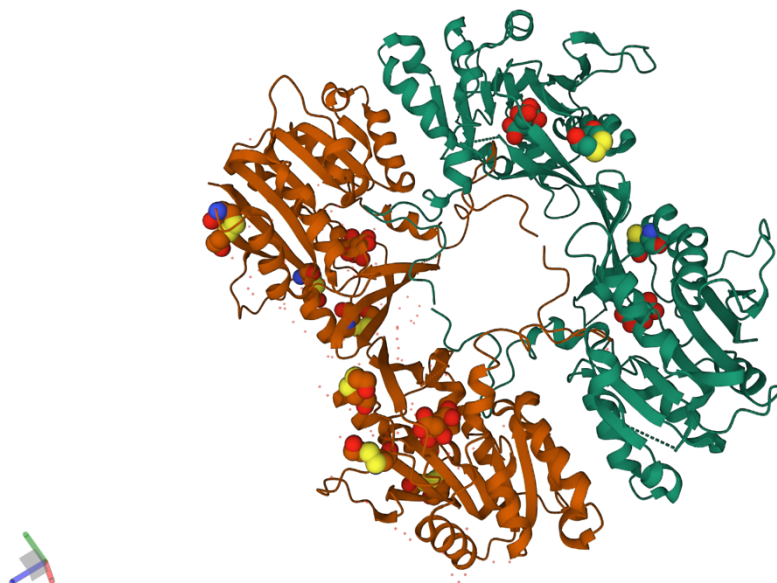
[Q7] Generate a sequence identity based heatmap of your aligned sequences using R. If necessary convert your sequence alignment to the ubiquitous FASTA format (Seaview can read in clustal format and “Save as” FASTA format for example). Read this FASTA format alignment into R with the help of functions 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
1R74	X-Ray Diffraction	2.55	Homo Sapiens	0.78	37.705
2AZT	X-Ray Diffraction	2.70	Homo Sapiens	0.81	37.705
1R8X	X-Ray Diffraction	2.95	Mus musculus	1.70	36.066

[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?



(High resolution snapshot)

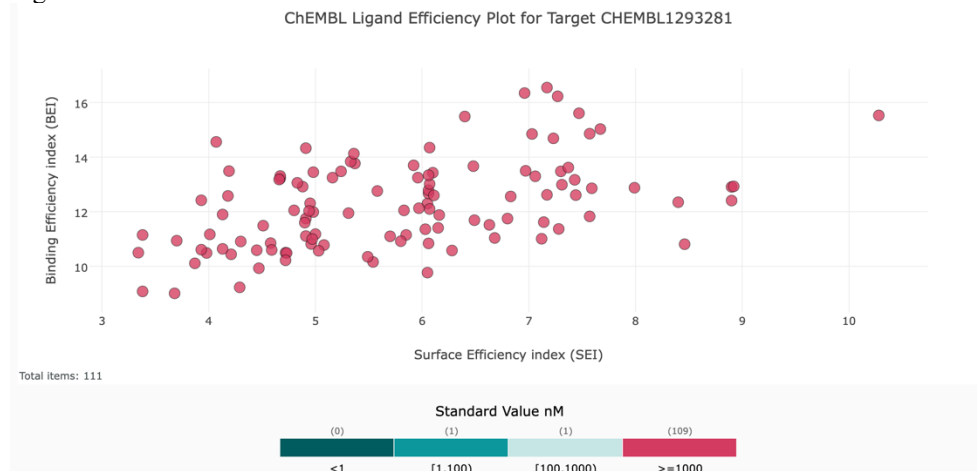
The above structure of 1R74, AKA the human “Chain A, Glycine N-methyltransferase” is not very likely to be similar in structure to the novel Zebrafish Myocilin Precursor as they have a sequence similarity of 37%, which is less than half in similarity. In the structure above, the colorful “space-filled” areas represent the ligand, and may correspond to the Zebrafish Myocilin Precursor of this report as although the structure is different, the active binding sites may be retained.

[Q10] Perform a “Target” search of ChEMBL (<https://www.ebi.ac.uk/chembl/>) with your novel sequence. Are there any Target Associated Assays and ligand efficiency ID Technique Resolution Source Eval Identity 3BOM X-RAY DIFFRACTION 1.35 Oncorhynchus mykiss 6.59E-63 81.4 1SPG X-RAY DIFFRACTION 1.95 Leiostomus xanthurus 3.16E-58 75.9 3BCQ X-RAY DIFFRACTION 2.4 Brycon cephalus 5.11E-57 77.2 data reported that may be useful starting points for exploring potential inhibition of your novel protein?

According to ChEMBL, of 371 records, the record with the highest “%identity illustrated 1 Functional Assay and 1 figure of ligand efficiencies (Fig.10). URL below:

https://www.ebi.ac.uk/chembl/target_report_card/CHEMBL1293281/

Fig.10



Carried out by the Scripps Research Institute Molecular Screening Center, the binding assay utilized “Fluorescence polarization-based biochemical high throughput response assay” to inhibit EBNA-1 (AKA the Epstein-Bar Virus Nucleation Antigen 1). Results suggested that more research is required via “two or more separate campaigns”. However, future research may be used to shape the treatment for EBV virus, a “orally-transmitted herpesvirus associated with [causing] numerous human neoplasms” which cells undergo mitosis.

Examining Fig.10, it is evident that the Binding Efficiency Index (BEI) at Standard Value nM ≥ 1000 shows no strong correlation to the Surface Efficiency Index (SEI). This suggests that current trials for EBNA-1 are not as effective, although the binding site is receptive and reacting to the response assay.

National Center for Biotechnology Information (2022). PubChem Bioassay Record for AID 2381, Source: The Scripps Research Institute Molecular Screening Center. Retrieved December 2, 2022 from <https://pubchem.ncbi.nlm.nih.gov/bioassay/2381>.