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:

Graphical user interface, text, application

Description automatically generated

Output:

Graphical user interface

Description automatically generated

Chosen Match: Accession: [NM\_001015062.2](https://www.ncbi.nlm.nih.gov/nucleotide/NM_001015062.2?report=genbank&log$=nucltop&blast_rank=2&RID=SJZGDS2F016)

Graphic Summary:

Application

Description automatically generated with medium confidence

Alignment details:

>gb|[NM\_001015062.2](https://www.ncbi.nlm.nih.gov/nucleotide/NM_001015062.2?report=genbank&log$=nucltop&blast_rank=2&RID=SJZGDS2F016)| 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



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:

ATGTATAAAGGCTCCTTATACTACCAGCGGAGGCTCAGCCGCACCCTAATAAGATACGACCTACATGCTGAGAGCATTGCTGCTCGTCGTGATCTACCCCATGCTGGTTTCCATGGTCAGTTCCCCTACTCATGGGGTGGCTACACAGACATTGACTTGGCAATAGATGAAAATGGCTTATGGGCTATATACAGCACAAATAAAGCCAAGGGTGCTATCGTGATCTCCCAGCTGGACCCTCACAACCTTGAGGTGAAGGGTACCTGGGAAACAAAAATCCGCAAGACATCTGTGGCTAACGCTTTTATGATCTGTGGCAAGCTCTACACAGTTGCTAGTTACACCTCACCAAACACCACGGTTAATTACATGTTTGACACTGCAACCAGCCAGGGTAAAGCAATATCAGTGCCCTTCAAAAACCGTTATCGCTACAACAGCATGGTAGACTACAACTCCGCAAAAAGAAAGCTGTATGCTTGGGATAATTATTACATGGTGTCTTACAGTGTGAGACTTGGCAAGCAGGAGTAA

Chosen Sequence ORF Translated (via EMBOSS Transeq)

MYKGSLYYQRRLSRTLIRYDLHAESIAARRDLPHAGFHGQFPYSWGGYTDIDLAIDENGLWAIYSTNKAKGAIVISQLDPHNLEVKGTWETKIRKTSVANAFMICGKLYTVASYTSPNTTVNYMFDTATSQGKAISVPFKNRYRYNSMVDYNSAKRKLYAWDNYYMVSYSVRLGKQE\*

Chosen Sequence FASTA (sequence taken from BLAST result):

>NM\_001015062.2 Danio rerio myocilin (myoc), mRNA

GGATCAACGACCAATCAAAGCAAGAAGCTCCAACCAGCCCCTCGAAGGCAACGTTTAGCTTGAGCGAGAGAAAGTCAGTAGGAGCGAAAGATTGAAAAGACTCAGGTGAGAGGTCATTAGTGGAACGCCAGAGAAGTACGGGGCTCCGGTTACACTTGGAAGGTGAAAGGGAAGGAATAGATTAGTGCTAGAGCCGAAGAGGACCAGGAAGGACAAGTGAGCTGCAAGTCACCACTGCATTAAACCTTGGACAAGGAGCTTCAGGCAACAGTAGAAAGAAGTTATCTGTGGTCCACCCCAACATGTGGTTTTTAGCAGTGTTGTGGATTTCTTCCCTGCTGATGGGATCCCAGGTTCAGAGCAGTGCTAACCTTCGTCGAGCAAATGCTGGGAATGGTCGCTGTCAGTACACCTTTATGGTGGACAGCCCTACTGAGGCCAGCTGTCCATCACCAGGTTCAACTCCTGAGATGGAGGCTCTGATGTCTCGCCTGGGGCTGCTAGAGGCACTGGTTGCTCGTCTCGTAGGAGGGGAAGCCATGCCAGAATCATCACAGAGCTCGGGATCTGGCCTTCAGGACTCTTACAACCAGGTGATGGGGGAGAACGCGCAGCTCAAGAGAGAAAAGCAGAGACTAGACAGACAAGTTCAGGACCTGCAGCAGAGGATGGAAGAGCTCCGCCAAGAGGCTGAGAGGCTGAGGAGCAGACCCTGCATGCAGCAAACTTCTTCCAGAGTGCCGCAAAAAGACAACAGCTTCAGACCGGGATCAGGGCATGTACCCTCCAACTTGGCATCCAGACCTGGGAATCCACAAGAAGACAAAAGTAGTTTAAGAGACCCCGCATGGCAGTACTCAAATCCCGGATACCAAGAGTTGACGGCTGTGGTCACCGAGGTGACTGCCCCAAATCAGGACGGTCCAGCAGATATCTCAGGCTGTGGTGATCTGGTGTGGGTTGAAAATCCTGAGGTGCATCGTAAAGCTGATAGTATTGCTGGTAAATATGGTGTGTGGATGCAAGATCCAGAAGCCAAGGAACCTTATGGTCCAGACATGGTATGGCGTATTGATTCTGTCGGTTCTGAAGTGCGTCAACTCTTCGGATATGAAAACATGGACCAGCTGACACGTGGCTTTCCCACCAAGGTTCTACTCCTGCCAGAATCTGTAGAGAGCACAGGTGCTACCATGTATAAAGGCTCCTTATACTACCAGCGGAGGCTCAGCCGCACCCTAATAAGATACGACCTACATGCTGAGAGCATTGCTGCTCGTCGTGATCTACCCCATGCTGGTTTCCATGGTCAGTTCCCCTACTCATGGGGTGGCTACACAGACATTGACTTGGCAATAGATGAAAATGGCTTATGGGCTATATACAGCACAAATAAAGCCAAGGGTGCTATCGTGATCTCCCAGCTGGACCCTCACAACCTTGAGGTGAAGGGTACCTGGGAAACAAAAATCCGCAAGACATCTGTGGCTAACGCTTTTATGATCTGTGGCAAGCTCTACACAGTTGCTAGTTACACCTCACCAAACACCACGGTTAATTACATGTTTGACACTGCAACCAGCCAGGGTAAAGCAATATCAGTGCCCTTCAAAAACCGTTATCGCTACAACAGCATGGTAGACTACAACTCCGCAAAAAGAAAGCTGTATGCTTGGGATAATTATTACATGGTGTCTTACAGTGTGAGACTTGGCAAGCAGGAGTAAACATTGTAAATTTCCATATCATCAAAGCTTGCATTTTTTTTTTTTACTTAAAACATCCAGTAGGACGTCAATTGTTGGACTTACTGATTTTCAAAGACAATGAACTTTTTTTTCTGAAAAAAGAACTGGATAAGTCAACCAATCAGGCTTTTGCTGGTATTTGTATAGGAAACTGAGTGTGTGTGTGTGTATACCAAATAAAACACTTTCCTGGTAACATAAAAACAGAGTTTTTTCAAGATTTTCCCATTATATACAGTAAATACCTTACAATTAACTGCATTGTATATCCTTCTCGATCATAGATTTATCTTTGATAAAAAAAGATGTTGACCAACAAGTTCCAAAATGGATAATGTATCAAATCATGTCGATATTCACATCCTTAATAATTTTATCCACTTAACATGTTAATAGAGAAATTATTTTTCTACCTATGTAAGTAATGTCTTGTCTGTTTTTGTGTATATGGCACTGAATAAATTCAAGGAAGGGGGAAAATATACTAGCAAACAACTGATTACAGCAAGCCTGCTTCTTAATTCATATATCATTGAGCAGTACGCAACTTTACGAACCTAAACCTTTGCTTGTTGTGTTTGTGCTCTAATACTTTATCATATAATCTTCACTCTGTATATGATGATCATTTTCCCTTCTCTCTTTTGAATTTGCTGCAAATCTCACATGCTGCTTGAGATGTTAAGATTGGGATTTTGTGGCTTTGGAAGCGTATTAACTTGAAAATAAAGATTTATAAAGTCTGCCAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

[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:

Graphical user interface, text, application, email

Description automatically generated

BLASTP Output: Graphical user interface, application

Description automatically generatedGraphical user interface, application

Description automatically generated

Comment: Top result from Carassium auratus (Goldfish)

Alignments:

Graphical user interface, text, application, email

Description automatically generated

[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\*GSSVHVAAALGLRCQLSSCCFWPAWCGMWGPGQLSSGRPMTRVADASIPSVWPVPMNPAAQSRARPCQSSITYRETAAPNA\*TWRPPKLDSAPWRASSTN\*PWTRLPGPRRPRRGCRGSWAP\*GGSGTSWKPKPESWRLPTATSSETSQFWRKRRSD\*GKKMRIWPGGWKAAARR\*QG\*EGASVPRPETLLGLCHQAPEKFLRGIWTLWPSRN\*SPS\*LKFLLPEF\*RRAHLAISGVERETPDVEN\*FG\*ESLSR\*EQQKQLLASMVCGCETPSPPTPTPRRPRGESTQLARMSARFLSMTSSASLCRATLLRFTYCLGHWKARVLWCTRGASISRALSPELS\*DMS\*IPRQ\*RLRRKSLELATTDSSRILGVATRTLTWLWMKQASGSFTAPMRPKVPLSSPN\*TQRIWNSNKPGRQTSVSSQSPMPSSSVAPCTPSAATPQQMLPSTLLMTQAQVSARP\*PSHSRTAISTAA\*LTTTPWRRSSLPGTT\*TWSLMTSSSPRCEKPPSCTGNGRRRCSGLLGGAG\*RESQPARAQAALTAFQVFINPEG\*TWSPSNYSGIVV\*GRRQFHIINILYLLSAFMGCLMT\*FKFSCDLGQKL\*GIIVSS\*KPLLLHVTWLPQATIKSITSKGSRIAPLASIEYK\*DAFTTVGF\*CFR\*NTVGSHITLYIVK\*NFLTQ

>Zebrafish gi|33573230|ref|NP\_001015062|myocilin precursor [Danio rerio]\*

ATGTATAAAGGCTCCTTATACTACCAGCGGAGGCTCAGCCGCACCCTAATAAGATACGACCTACATGCTGAGAGCATTGCTGCTCGTCGTGATCTACCCCATGCTGGTTTCCATGGTCAGTTCCCCTACTCATGGGGTGGCTACACAGACATTGACTTGGCAATAGATGAAAATGGCTTATGGGCTATATACAGCACAAATAAAGCCAAGGGTGCTATCGTGATCTCCCAGCTGGACCCTCACAACCTTGAGGTGAAGGGTACCTGGGAAACAAAAATCCGCAAGACATCTGTGGCTAACGCTTTTATGATCTGTGGCAAGCTCTACACAGTTGCTAGTTACACCTCACCAAACACCACGGTTAATTACATGTTTGACACTGCAACCAGCCAGGGTAAAGCAATATCAGTGCCCTTCAAAAACCGTTATCGCTACAACAGCATGGTAGACTACAACTCCGCAAAAAGAAAGCTGTATGCTTGGGATAATTATTACATGGTGTCTTACAGTGTGAGACTTGGCAAGCAGGAGTAA

\*(sequence taken from BLAST result)

>Goldfish gi|113121143|ref|XP\_026147181|myocilin-like [Carassius auratus]

MYFLAMLWASCLLMGTHAQGSASFRRANAGSGRCQYTFMVDSPTEASCPSAGSSPEVEALKSRLGLLEALVARLAGGEAMSESSHGSGSQSGLQDAYNQAMGENARLQREKQRLDRQVQDLQQRMEELRQEAERLRSRPCMQQPPPRVPQNDNSFRPGSGPAVSQLVSRPGTTQGDKSSLRDPAWHYENPGYQEVTAVITEVSAPNQEGPADIPGCGDLVWVKEPEVHRKADSIAGKYGVWMQDPEAKEPYGAEMVWRIDSVGSEVRQLFGYENMDQLSRGFPTKVLLLPESMESTGATMYRGSLYYQRRLSRTLLRYDLLSESIAARRDLPHAGFHGQFPYSWGGYTDIDLAIDENGLWAIYSTNKAKGAIVISQLDPHNLEVKGTWETKIRKTSVANAFMICGKLYTVASYTSPNTTINYMYDTATSQGKTISVPFKNRYRYNSMVDYNPTQRKLYAWDNFYMVSYNVRLGKQE

>Carp gi|27590662|ref|KTF90649| hypothetical protein cypCar\_00009366 [Cyprinus carpio]

MWFLVMLWASCLLMGTHAQSSASFRRANAGSGRCQYTFMVDSPTEASCPSAVSSPEIEALKSRLGLLEALVARLIGGEAMSKSTQSSGSQSGLQDAYSQVMGENAQLQREKQRLDRQVQDLQQRMEELRQEAERLRSRPCMQQPPPRVPQNDNSFRPGSGPALSHLVSRPGNTQGDKSSLRDPAWHYENPGYQELTAVVTEVSAPNLEGPADISGCGDLVWVQEPEVHRKADSIAGKYGVWMQDPEAKEPYGPEMVWRIDAVGSEVRQLFGYENMDQLSRGFPTKVLLLPESMESTGATMYRGSLYYQRRLSRTLLRYDLLSESIAARRDLPHAGFHGQFPYSWGGYTDIDLTIDENGLWAIYSTNKAKGAIVISQLDPHNLEVKGTWETKIRKTSVANAFMICGKLYTVASYTSPNTTINYMYDTATSQGKTISVPFKNRYRYNSMVDYNPAQRKLYAWDNFYMVSYNVRLGKQE

>Tiger\_Barb gi|1606681|ref|XP\_043076193|myocilin isoform X2 [Puntigrus tetrazona]MWFLAVFWASCLLMGTHAQSSTSFRRANTGSGRCQYTFTVDSPTEASCPSAGSSPEMEALKSRLGLLEALVARLVGGEAVPELSQGSRPQSGLQDAYNQMMGENTQLQREKQKLDRQVQDLQQRMEELRQEAERLRSRPCMQQPPPRVPQNDNSFRPGSDPSWHYENPGYQELTAVVTEVSAPSPEGPADISGCGDLVWVQEPEVHRKADNIAGKYGVWMQDPEAKEPYGTEMVWRIDAVGSEVRQLFGYENMDQLSRGFPTKVLLLPEFMESTGATMYRGSLYYQRRLSRTLVRYDLLSENIAARRDLPHAGFHGQFPYSWGGYTDIDLAIDENGLWAIYSTNKAKGAIVISQLDPHNLEVKGTWETKIRKTSVANSFMICGKLYTVASYTSPNTTINYMYDTATSQGKTIAVPFKNRYRYNSMVDYNPAQRKLYAWDNFYMVSYNVRLGKQE

>Kanglang\_Fish gi|495550|ref|ROL55559.1|Myocilin [Anabarilius grahami] MWFLAVLCVSCLLMGTQAQSSASFRRANAGNGRCQYSFTVDSPTEASCPSPGSSPEMEALKSRLGLLEALVARLVGGEAVSESSQGSGSQSGLQDAYNQLMGENAQLQREKQRLDRQVQDLQQRMEELRQEAERLRSRPCVQQPPPRVPQNDNSFRPGSGPALSHLVSSPGNTQGDTSSLRDPAWHFSNPEYQELTAVVTEVSAPNLEGPTDISGCGDLVWVEGPEVHRKADSIAGKYGVWMQDPEAKEPYGPEMVWRIDAVGSEVRQLFGYENMDQLSRGFPTKVLLLPESVESTGATMYKGSLYYQRRLSRTLIRYDLLSESIAARRDLPHAGFHGQFPYSWGGYTDIDLSVDENGLWAIYSTNKAKGAIVISQLDPHNLEVKGTWETKIRKTSVANAFMICGKLYTVASYTSPNTTVNYMYDTATSQSKTISVPFKNRYRYNSMVDYNPGQRKLYAWDNYYIVSYKVRLGKQE

>Flathead\_Minnow gi|120494087|ref|XP\_39549125|myocilin [Pimephales promelas]

MRFLAVLCVSCLLVGTQAQSSASFRRANAGNGRCQYSFTVDSPTEASCPSSGSSPEMEALKSRLGLLEALVARLVGGEAVSESSAGSRSQSGLQDAYNQVMGEKAQIQREKQRLEKQVQDLQQRMEELRQEAERLRSRPCTQKPPPRVSEIDSSFKPGSGPALSHLVSRPGNTQGDTSSLRDPAWHSSSGYQELTAVVTEVSAPNLEGPADISGCGDLVWVDQPEMHRKADSIAGKYGVWMQDPEAKEPYGPEMVWRIDAVGSEVRQLFGYENMDQLSRGFPTKVLLLPESVESTGATMYRGSLYYQRRLSRTLLRYDLLSESIAARRDLPHAGFHGQFPYSWGGYTDIDLAIDENGLWAIYSTNKVKGAIVLSQLDPHNLEVKGTWETKIRKTSVANAFMICGKLYTVASYILPNTTINYMYDTATSQGKTISVPFKNRYRYNSMVDYNPGQRKLQAWDNYYIVSYKVRLGKQL

**Alignment**: Obtained using MUSCLE (version 3.8) at EBI:

CLUSTAL multiple sequence alignment by MUSCLE (3.8)

Human EPARPPIQAPLSTAELSRGSLTKPLQGSSVHVAAALGLRCQLSSCCFWPAWCGMWGPGQLSSGRPMTRVADASIPSVWPVP

Zebrafish ----MWFLAVLWISSLLMGSQVQS--SANLRRANAGNGRCQYTFMVDSP--------------------TEASCPSPGSTP

Tiger ----MWFLAVFWASCLLMGTHAQS--STSFRRANTGSGRCQYTFTVDSP--------------------TEASCPSAGSSP

Carp ----MWFLVMLWASCLLMGTHAQS--SASFRRANAGSGRCQYTFMVDSP--------------------TEASCPSAVSSP

Goldfish ----MYFLAMLWASCLLMGTHAQG--SASFRRANAGSGRCQYTFMVDSP--------------------TEASCPSAGSSP

Flathead ----MRFLAVLCVSCLLVGTQAQS--SASFRRANAGNGRCQYSFTVDSP--------------------TEASCPSSGSSP

Kanglang ----MWFLAVLCVSCLLMGTQAQS--SASFRRANAGNGRCQYSFTVDSP--------------------TEASCPSPGSSP

: . : : \* \*: .: .:... \* : . \*\*\* : \*

Human MNPAAQSRARPCQSSITYRETAAPNATWRPPKLDSAPWRASSTNPWTRLPGPRRPRRGCRGSWAPGGSGTSWKPKPESWRL

Zebrafish EMEALMSRLGLLEALV------------------------------ARLVGGEAMPESSQSSG--SGLQDSY---------

Tiger EMEALKSRLGLLEALV------------------------------ARLVGGEAVPELSQGSRPQSGLQDAY---------

Carp EIEALKSRLGLLEALV------------------------------ARLIGGEAMSKSTQSSGSQSGLQDAY---------

Goldfish EVEALKSRLGLLEALV------------------------------ARLAGGEAMSESSHGSGSQSGLQDAY---------

Flathead EMEALKSRLGLLEALV------------------------------ARLVGGEAVSESSAGSRSQSGLQDAY---------

Kanglang EMEALKSRLGLLEALV------------------------------ARLVGGEAVSESSQGSGSQSGLQDAY---------

::\*\* \*\* . \*\_ \* \*\* :: :

Human PTATSSETSQFWRKRRSDDRQVQDLQQ--RMEELRQEAERLRSRP------CMQQTSSRVPQKDNSFRPGSGHVPSNLLPE

Zebrafish -NQVMGENAQLKREKQRLDRQVQDLQQ--RMEELRQEAERLRSRP------CMQQTSSRVPQKDNSFRPGSGHVPSNLASR

Tiger -NQMMGENTQLQREKQKLDRQVQDLQQ--RMEELRQEAERLRSRP------CMQQPPPRVPQNDNSFRPGSDPS-------

Carp -SQVMGENAQLQREKQRLDRQVQDLQQ--RMEELRQEAERLRSRP------CMQQPPPRVPQNDNSFRPGSGPALSHLVSR

Goldfish -NQAMGENARLQREKQRLDRQVQDLQQ--RMEELRQEAERLRSRP------CMQQPPPRVPQNDNSFRPGSGPAVSQLVSR

Flathead -NQVMGEKAQIQREKQRLEKQVQDLQQ--RMEELRQEAERLRSRP------CTQKPPPRVSEIDSSFKPGSGPALSHLVSR

Kanglang -NQLMGENAQLQREKQRLDRQVQDLQQ--RMEELRQEAERLRSRP------CVQQPPPRVPQNDNSFRPGSGPALSHLVSS

:\*\* \* .\* .\* :: . .\*.:.:.\*:..

Human FRRAHLAISGVERETPDVENFGESLSREQQKQLLASMVCGCETPSPPTPTPRRPRGESTQLARMSARFLSMTSSASLCRAT

Zebrafish PGNPQEDKSSLRDPAWQYSNPG--------YQELTAVVTEVTAPNQDGPADISGCGD------------------------

Tiger ---------------WHYENPG--------YQELTAVVTEVSAPSPEGPADISGCGD------------------------

Carp PGNTQGDKSSLRDPAWHYENPG--------YQELTAVVTEVSAPNLEGPADISGCGD------------------------

Goldfish PGTTQGDKSSLRDPAWHYENPG--------YQEVTAVITEVSAPNQEGPADIPGCGD------------------------

Flathead PGNTQGDTSSLRDPAWH-SSSG--------YQELTAVVTEVSAPNLEGPADISGCGD------------------------

Kanglang PGNTQGDTSSLRDPAWHFSNPE--------YQELTAVVTEVSAPNLEGPTDISGCGD------------------------

.. \* ::::: :\*. \*: \*:

Human LLRFTYCLGHWKARVLWCTRGASISRALSPELSDMSIPRQRLRRKSLELATTDSSRILGVATRTLTWLWMKQASGSFTAPM

Zebrafish --------------LVW-------------------VENPEVHRKA------DS-----IAGKYGVWMQDPEAKEPYGPDM

Tiger --------------LVW-------------------VQEPEVHRKA------DN-----IAGKYGVWMQDPEAKEPYGTEM

Carp --------------LVW-------------------VQEPEVHRKA------DS-----IAGKYGVWMQDPEAKEPYGPEM

Goldfish --------------LVW-------------------VKEPEVHRKA------DS-----IAGKYGVWMQDPEAKEPYGAEM

Flathead --------------LVW-------------------VDQPEMHRKA------DS-----IAGKYGVWMQDPEAKEPYGPEM

Kanglang --------------LVW-------------------VEGPEVHRKA------DS-----IAGKYGVWMQDPEAKEPYGPEM

::\*

Human RPKVPLSSPNTQRIW NSNKPGRQTSVSSQSPMPSSSVAPCTPSAATPQQMLPSTLLMTQAQVSARPPSHSRTAITAALTT

Zebrafish VWRIDSVGSEVRQLF-----GYENMDQLTRGFPTKVLLLPESVESTGATMYKGSLYYQR--------RLSRTLIRYDLHAE

Tiger VWRIDAVGSEVRQLF-----GYENMDQLSRGFPTKVLLLPEFMESTGATMYRGSLYYQR--------RLSRTLVRYDLLSE

Carp VWRIDAVGSEVRQLF-----GYENMDQLSRGFPTKVLLLPESMESTGATMYRGSLYYQR--------RLSRTLLRYDLLSE

Goldfish VWRIDSVGSEVRQLF-----GYENMDQLSRGFPTKVLLLPESMESTGATMYRGSLYYQR--------RLSRTLLRYDLLSE

Flathead VWRIDAVGSEVRQLF-----GYENMDQLSRGFPTKVLLLPESVESTGATMYRGSLYYQR--------RLSRTLLRYDLLSE

Kanglang VWRIDAVGSEVRQLF-----GYENMDQLSRGFPTKVLLLPESVESTGATMYKGSLYYQR--------RLSRTLIRYDLLSE

: :.\*\*: \*. :\* . .\*:. :\*. .: . \* .: ..:...::

Human TPWRRSSLPGTTT-------WSLMTSSSPRCEKPPSCTGNGRRRCSGLLGGAGRESQPARAQAALTAFQVF-INPEGTWSP

Zebrafish SIAARRDLPHAGFHGQFPYSWGGYTDIDLAIDE------------NGLW--AIYSTNKAKGAIVISQLDPHNLEVKGTWET

Tiger NIAARRDLPHAGFHGQFPYSWGGYTDIDLAIDE------------NGLW--AIYSTNKAKGAIVISQLDPHNLEVKGTWET

Carp SIAARRDLPHAGFHGQFPYSWGGYTDIDLTIDE------------NGLW--AIYSTNKAKGAIVISQLDPHNLEVKGTWET

Goldfish SIAARRDLPHAGFHGQFPYSWGGYTDIDLAIDE------------NGLW--AIYSTNKAKGAIVISQLDPHNLEVKGTWET

Flathead SIAARRDLPHAGFHGQFPYSWGGYTDIDLAIDE------------NGLW--AIYSTNKAKGAIVISQLDPHNLEVKGTWET

Kanglang SIAARRDLPHAGFHGQFPYSWGGYTDIDLSVDE------------NGLW--AIYSTNKAKGAIVISQLDPHNLEVKGTWET

: . \* .\*\* : \*. \*. . :: .\*\* \*

Human SNYSGIVVGRRQFHIINILYLLSAFMGCLMTFKFSCDLGQKLGIIVSSKPLLLHVTWLPQATIKSITSKGSRIAPLASIEY

Zebrafish KIRKTSVANA--FMICGKLYTVASYTSPNTTVNYMFDTATSQGKAIS-------VPFKNRYRYNSMVDYNSAKRKLYAW--

Tiger KIRKTSVANS--FMICGKLYTVASYTSPNTTINYMYDTATSQGKTIA-------VPFKNRYRYNSMVDYNPAQRKLYAW--

Carp KIRKTSVANA--FMICGKLYTVASYTSPNTTINYMYDTATSQGKTIA-------VPFKNRYRYNSMVDYNPAQRKLYAW--

Goldfish KIRKTSVANA--FMICGKLYTVASYTSPNTTINYMYDTATSQGKTIA-------VPFKNRYRYNSMVDYNPAQRKLYAW--

Flathead KIRKTSVANA--FMICGKLYTVASYILPNTTINYMYDTATSQGKTIA-------VPFKNRYRYNSMVDYNPAQRKLYAW--

Kanglang KIRKTSVANA--FMICGKLYTVASYTSPNTTVNYMYDTATSQGKTIA-------VPFKNRYRYNSMVDYNPAQRKLYAW--

.:: ... .:: :: . :: :\*\*\*... . \*.. \* \* . \*\* :::: \*.:

Human KDAFTTVGFCFR NTVGSHITLYIVKNFLTQ

Zebrafish -DNYYMVSYSVR--LGKQE-----------

Tiger -DNFYMVSYNVR--LGKQE-----------

Carp -DNFYMVSYNVR--LGKQE-----------

Goldfish -DNFYMVSYNVR--LGKQE-----------

Flathead -DNYYIVSYKVR--LGKQL-----------

Kanglang -DNYYIVSYKVR--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.

Text

Description automatically generated

[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.

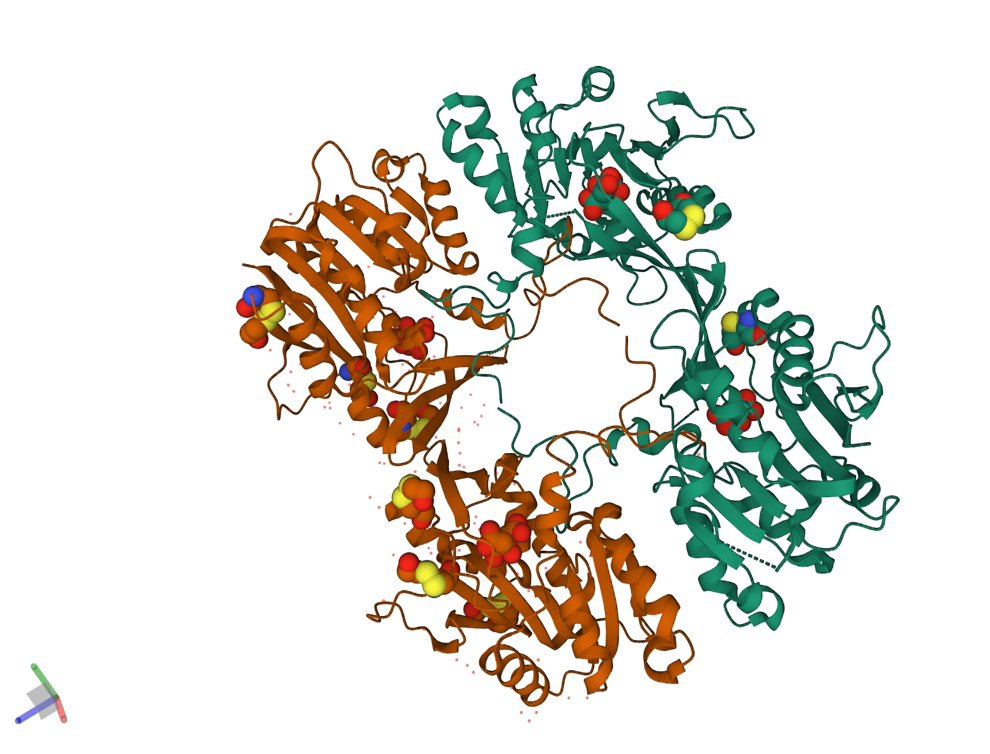
Chart

Description automatically generated

[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 ChEMBEL ( https://www.ebi.ac.uk/chembl/ ) with your novel sequence. Are there any Target Associated Assays and ligand efficiency ID Technique Resolution Source Evalue 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

Chart, scatter chart

Description automatically generated

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>.