

BIMM-143: INTRODUCTION TO BIOINFORMATICS

Professor Barry J. Grant

Find A Gene Final Project

Garrett Cole | A15988021
g1cole@ucsd.edu

BLAST » tblastn » results for RID-SDMVCPGG013

[Home](#)
[Recent Results](#)
[Saved Strategies](#)
[Help](#)

[Edit Search](#)
[Save Search](#)
[Search Summary](#)

[How to read this report?](#)
[BLAST Help Videos](#)
[Back to Traditional Results Page](#)

Your search is limited to records that include: nematodes (taxid:6231)

Warning: Searches from this IP address have consumed a large amount of server CPU time. Future searches may be penalized in fairness to other users. Please consider the BLAST+ binaries: <https://www.ncbi.nlm.nih.gov/books/NBK279690/>

Job Title

ref[NP_001269332]

RID

SDMVCPGG013 Search expires on 12-01 10:52 am [Download All](#)

Program

TBLASTN [Citation](#)

Database

est [See details](#)

Query ID

NP_001269332.1

Description

malate dehydrogenase, mitochondrial isoform 2 precursor ...

Molecule type

amino acid

Query Length

296

Other reports

Filter Results

Organism only top 20 will appear ☐ exclude

[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](#)

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

| | Description | Scientific Name | Max Score | Total Score | Query Cover | E value | Per. Ident | Acc. Len | Accession |
|-------------------------------------|--|--------------------|-----------|-------------|-------------|---------|------------|----------|----------------------------|
| <input checked="" type="checkbox"/> | L13 f PDT_30_053 Panagrolaimus davidi 20 degree Panagrolaimus davidi cDNA, mRNA sequence | Panagrolaimus d... | 261 | 261 | 83% | 2e-84 | 49.50% | 1090 | JZ673983.1 |
| <input checked="" type="checkbox"/> | Pd_3pr_63E16 Panagrolaimus davidi 4 degree Panagrolaimus davidi cDNA, mRNA sequence | Panagrolaimus d... | 259 | 259 | 83% | 1e-84 | 48.10% | 916 | JZ617320.1 |

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 JZ673983.1, a L13_f_PDT_30_053, 1090 base pair 20 degree cDNA from *Panagrolaimus davidi*

Alignment details:

| Download GenBank Graphics | | | | | | |
|--|---|--|--------------|--|--------------|-------|
| L13_f_PDT_30_053 Panagrolaimus davidi 20 degree Panagrolaimus davidi cDNA, mRNA sequence | | | | | | |
| Sequence ID: JZ673983.1 Length: 1090 Number of Matches: 1 | | | | | | |
| Range 1: 52 to 945 GenBank Graphics Next Match Previous Match | | | | | | |
| Score | Expect | Method | Identities | | Positives | Frame |
| 261 bits(667) | 2e-84 | Compositional matrix adjust. | 148/299(49%) | | 183/299(61%) | +1 |
| Query 20 | SAQNNA--- | KVAVLGASGGIGQPLSLLKNSPLVSRLTLYDIAHTPGVAADLSHIETKAA | | | | 76 |
| Sbjct 52 | SA+N + KVA+LGASGGIGQPL LLLK +P V+ L LYD+A+T GV ADLSHI+T A SARNTSSAPKVALLGASGGIGQPLGLLKTNPKVASLALYDVANTAGVGADLSHIDTHAQ | | | | | 231 |
| Query 77 | VKGYLGPEQLPDCLKGC | DVVVIPAGVPRKPGMTRDDLNFNTNATIVATLTAACAQHCPEAM | | | | 136 |
| Sbjct 232 | V + G L+G D+VVIPAGVPRKPGMTRDDLNFN NA IV L A A+ CP+A VTAHTGXXXXHSALEGADIVVIPAGVPRKPGMTRDDLNFVNAGIVRDLAEAAAKACPKAF | | | | | 411 |
| Query 137 | ICVIANP----- | GLDPA | | | | 154 |
| Sbjct 412 | + +I NP VAIITNPVNSTVP | IAAEVYKNNGVYDPKRIFGVTTLDVVR | | | | 591 |
| Query 155 | IGGHAGKTIIP | LISQCTPKVDFFQDQLTALTGRIQEAGTEVVKAKAGAGSATLSMAYAGA | | | | 214 |
| Sbjct 592 | IGGH+G TIIP | L+SQC P F ++ LT RIQ+AGTEVVKAKAGAGSATLSMA+AGA | | | | 771 |
| Query 215 | RFVFS | LVDMANGKEGVVECSFVKSQETE-CTYFSTPLLLGKKGIEK-----NLGIGKVS | | | | 267 |
| Sbjct 772 | RFV +L+ + GK+ V+C++V+S + YFSTPL L G+EK NL K+S | RFVDALISGLQGKK-TVQCAYVQSDVVVGVDYFSTPLEP | | | | 945 |

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:

```
>52-945_1 L13_f_PDT_30_053 Panagrolaimus davidi 20 degree  
Panagrolaimus davidi cDNA, mRNA sequence  
SARNTSSAPKVALLGASGGIGQPLGLLLKTNPKVASLALYDVANTAGVGADLSHIDTHAQ  
VTAHTGXXXXHSALEGADIVVIPAGVPRKPGMTRDDLNFVNAGIVRDLAEAAAKACPKAF  
VAIITNPVNSTVPIAAEVYKNNGVYDPKRIFGVTTLDVVRSAFIAELKKLDVSKTVIPV  
IGGHSQVTIIPLLSQCPQAQFSDSEIEKLTARIQDAGTEVVKAKAGAGSATLSMAFAGA  
RFVDALISGLQGKKTVCAYVQSDVVKGVDYFSTPLELEPNGVEKFLKTVNLXFMKIS
```

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: Malate Dehydrogenase

Species: *Panagrolaimus*

[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 the NR database produced a top hit from the *Halicephalobus* (*Panagrolaimidae*) species. Output details below:

Job Title 52-945_1 L13_f_PDT_30_053 Panagrolaimus davidi...

RID [SFS1UBZB013](#) Search expires on 12-02 05:59 am [Download All](#) ▼

Program BLASTP [Citation](#) ▼

Database nr [See details](#) ▼

Query ID lc|Query_8885

Description 52-945_1 L13_f_PDT_30_053 Panagrolaimus davidi 20 de ...

Molecule type amino acid

Query Length 298

Other reports [Distance tree of results](#) [Multiple alignment](#) [MSA viewer](#) [?](#)

Filter Results

Organism only top 20 will appear ☐ exclude

[+ Add organism](#)

Percent Identity

to

E value

to

Query Coverage

to

[Filter](#)

[Reset](#)



- Descriptions**
- [Graphic Summary](#)
- [Alignments](#)
- [Taxonomy](#)

Sequences producing significant alignments

[Download](#) ▼ [Select columns](#) ▼ [Show](#) [?](#)

☒ 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"/> | hypothetical protein FO519_002303 [Halicephalobus sp. NKZ332] | Halicephalobus sp. NKZ332 | 507 | 507 | 96% | 9e-179 | 86.76% | 338 | KAE9554492.1 |
| <input checked="" type="checkbox"/> | malate dehydrogenase [Aphelenchoides besseyi] | Aphelenchoides besseyi | 472 | 472 | 96% | 7e-160 | 79.44% | 697 | KAI6186967.1 |
| <input checked="" type="checkbox"/> | malate dehydrogenase [Aphelenchoides besseyi] | Aphelenchoides besseyi | 472 | 472 | 96% | 1e-159 | 80.14% | 715 | KAI6213918.1 |
| <input checked="" type="checkbox"/> | malate dehydrogenase [Aphelenchoides besseyi] | Aphelenchoides besseyi | 471 | 471 | 96% | 3e-163 | 80.14% | 442 | KAI6236646.1 |
| <input checked="" type="checkbox"/> | Malate dehydrogenase, mitochondrial [Strongyloides ratti] | Strongyloides ratti | 462 | 462 | 96% | 2e-161 | 79.44% | 338 | XP_024507190.1 |
| <input checked="" type="checkbox"/> | unnamed protein product [Caenorhabditis auriculariae] | Caenorhabditis auriculariae | 453 | 453 | 96% | 1e-157 | 76.66% | 337 | CAD6187635.1 |
| <input checked="" type="checkbox"/> | hypothetical protein GCK72_011031 [Caenorhabditis remanei] | Caenorhabditis remanei | 447 | 447 | 96% | 2e-155 | 75.96% | 341 | KAF1762768.1 |
| <input checked="" type="checkbox"/> | CBN-MDH-2 protein [Caenorhabditis brenneri] | Caenorhabditis brenneri | 447 | 447 | 96% | 2e-155 | 75.61% | 341 | EGT46353.1 |
| <input checked="" type="checkbox"/> | putative malate dehydrogenase, mitochondrial [Caenorhabditis elegans] | Caenorhabditis elegans | 447 | 447 | 96% | 3e-155 | 75.61% | 341 | NP_498457.1 |
| <input checked="" type="checkbox"/> | unnamed protein product [Caenorhabditis sp. 36 PRJEB53466] | Caenorhabditis sp. 36 PRJEB... | 447 | 447 | 96% | 6e-155 | 75.26% | 341 | CAI2349587.1 |
| <input checked="" type="checkbox"/> | Protein CBR-MDH-2 [Caenorhabditis briggsae] | Caenorhabditis briggsae | 445 | 445 | 96% | 3e-154 | 75.26% | 341 | XP_002642936.1 |
| <input checked="" type="checkbox"/> | lactate/malate dehydrogenase_NAD binding domain-containing protein [Ditylenchus destructor] | Ditylenchus destructor | 444 | 444 | 94% | 5e-154 | 75.62% | 343 | KAI1729498.1 |
| <input checked="" type="checkbox"/> | lactate/malate dehydrogenase_NAD binding domain-containing protein [Ditylenchus destructor] | Ditylenchus destructor | 444 | 444 | 94% | 9e-154 | 75.27% | 343 | KAI1721095.1 |

Job Title

52-945_1 L13_f_PDT_30_053 Panagrolaimus david...

RID

SFS1UBZB013 Search expires on 12-02 05:59 am [Download All](#) ▼

Program

BLASTP [Citation](#) ▼

Database

nr [See details](#) ▼

Query ID

lcl|Query_8885

Description

52-945_1 L13_f_PDT_30_053 Panagrolaimus david 20 de ...

Molecule type

amino acid

Query Length

298

Other reports

[Distance tree of results](#)
[Multiple alignment](#)
[MSA viewer](#)
[?](#)

Filter Results

Organism

only top 20 will appear

☐ exclude

Type common name, binomial, taxid or group name

[+ Add organism](#)

Percent Identity

to

E value

to

Query Coverage

to

Filter

Reset

Descriptions

Graphic Summary

Alignments

Taxonomy

Alignment view

Pairwise

[Restore defaults](#)

Download

100 sequences selected

Download

GenPept

Graphics

▼ Next

▲ Previous

◀ Descriptions

hypothetical protein FO519_002303 [Halicephalobus sp. NKZ332]

Sequence ID: [KAE9554492.1](#) Length: 338 Number of Matches: 1

Range 1: 19 to 305

GenPept

Graphics

▼ Next Match

▲ Previous Match

| Score | Expect | Method | Identities | Positives | Gaps |
|----------------|---|------------------------------|--------------|--------------|-----------|
| 507 bits(1305) | 9e-179 | Compositional matrix adjust. | 249/287(87%) | 265/287(92%) | 0/287(0%) |
| Query 1 | SARNTSSAPKVALLGASGGIGQPLGLLLKTNPKVASLALYDVANTAGVGADLSHIDTHAQ | 60 | | | |
| Sbjct 19 | +ARN+SSAPKVALLGASGGIGQPLGLLLKTNPKVASLALYDVANTAGVGADLSHID+ A+ TARNSSAPKVALLGASGGIGQPLGLLLKTNPKVASLALYDVANTAGVGADLSHIDSAAR | 78 | | | |
| Query 61 | VTHTGXXXXHSALEGADIVVIPAGVPRKPGMTRDDLFNVNAGIVRDLAEAAKACPKAF | 120 | | | |
| Sbjct 79 | VTHTGPNELHKALEGADVIVIPAGVPRKPGMTRDDLFNVNAGIVRDLSEAAKICPKAF | 138 | | | |
| Query 121 | VAIITNPNVNSTVPIAAEVYKNGVYDPKRIFGVTTLDVVRSAFIAELKKLDVSKTVIPV | 180 | | | |
| Sbjct 139 | VAIITNPNVNSTVPIAAEVYKNGVYDP+RIFGVTTLDVVR+QAF+AELK LDV+KTV+PV VAIITNPNVNSTVPIAAEVYKNGVYDPRRIFGVTTLDVVRQAFAELKGLDVNKTVPV | 198 | | | |
| Query 181 | IGGHSQVTIIPLLSQCPQSAQFSDSEIEKLTARIQDAGTEVVKAKAGAGSATLSMAFAGA | 240 | | | |
| Sbjct 199 | IGGHSQVTIIPLLSQ QP A+FS E EKLTAIRIQDAGTEVVKAKAG GSATLSMAFAGA IGGHSQVTIIPLLSQLPQGAQFSQDETEKLTARIQDAGTEVVKAGGGSATLSMAFAGA | 258 | | | |
| Query 241 | RFVDALISGLQKKTVQCAVYQSDVVKGVDFSTPLEPENGVEKFL 287 | | | | |
| Sbjct 259 | RFV LI LQKK VQC YVQSDVVKGVDFSTPVELGPNVEKIL 305 | | | | |

[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.

Labeled Sequences for Alignment:

Homo Sapiens (Humans):

```
> Human_MDH2 |NP_001269332.1| malate dehydrogenase, mitochondrial isoform
2 precursor [Homo sapiens]
SAQNNA---KVAVLGASGGIGQPLSLLLKNSPLVSRLTLYDIAHTPGVAADLSHIETKAAVKGYLGPEQLPDCL
KGCDVVVIPAGVPRKPGMTRDDLFNTNATIVATLTAACAQHCPEAMICVIANP-----
-----GLDPARVNVFVIGGHAGKTIIPVISQCTPKVDFFPQDQLTALTGRIQEAGTEVV
KAKAGAGSATLSMAYAGARFVFSLV DAMNGKEGVVECSFVKSQETE-CTYFSTPLLLGKKGIEK----NLGIG
KVS
```

Panagrolaimus davidi (Antarctic Nematode):

```
> Antartic_MDH2 | 52-945_1 L13_f_PDT_30_053 Panagrolaimus davidi 20 degree
Panagrolaimus davidi cDNA, mRNA sequence
SARNTSSAPKVALLGASGGIGQPLGLLLKTNPKVASLALYDVANTAGVGADLSHIDTHAQVTAHTGXXXXHSAL
EGADIVVIPAGVPRKPGMTRDDLFNVNAGIVRDLAEAAKACPKAFVAIITNPVNSTVPAAAEVYKNNGVYDPK
RIFGVTTLDVVRSAQFIAELKKLDVSKTVIPVIGGHSGVTIIPLLSQCPQSAQFSDSEIEKLTARIQDAGTEVV
KAKAGAGSATLSMAFAGARFVDALISGLQGKKTVCAYVQSDVVKGVDFSTPLELEPNNGVEKFLKTVNLXFMK
IS
```

Halicephalobus (Panagrolaimidae):

```
> Halicephalobus_MDH2 | KAE9554492.1:19-305 hypothetical protein
FO519_002303 [Halicephalobus sp. NKZ332]
TARNSSSAPKVALLGASGGIGQPLGLLLKTNPKVASLALYDVANTAGVGADLSHIDSAARVTSHTGPNELHKAL
EGADVIVIPAGVPRKPGMTRDDLFNVNAGIVRDLSEAAKICPKAFVAIITNPVNSTVPAAAEVYKNNGVYDPR
RIFGVTTLDVVRQAQFVAELKGLDVNKTVPVIGGHSGVTIIPLLSQLQPGAKFSQDETEKLTARIQDAGTEVV
KAKAGGGSATLSMAFAGARFVQGLIDALQGKKNVQCTYVQSDVVKGVDFSTPVELGPNNGVEKIL
```

Camelus Ferus (Camel):

```
> Camel_MDH2 | 165-1130_1 PREDICTED: Camelus ferus malate dehydrogenase 2
(MDH2), transcript variant X1, mRNA
FSTSAQNNAKVAVLGASGGIGQPLSLLLKNSPLVSRLTLYDIAHTPGVAADLSHIETRATVKGYLGPEQLPDCL
KGCDVVVIPAGVPRKPGMTRDDLFNTNATIVATLTAACAQHCPEAMICIIISNPVNSTIPITAEVFKKHGVYNPD
KIFGVTTLDIVRANTFVAELKGLDPARVNVFVIGGHAGKTIIPVISQCTPKVDFFPQDQLTTLTGRIQEAGTEVV
KAKAGAGSATLSMAYAGARFVFSLLDAMNGKEGVVECSFVKSQETDCPYFSTPLLLGKKGIEKNLGIGKISPFE
EKMIAEAIPELKASIKKGEEFVKSMK
```

Apodemus Sylvaticus (Wood Mouse):

```
> Mouse_MDH2 | 167-1132_1 PREDICTED: Apodemus sylvaticus malate
dehydrogenase 2 (LOC127672705), mRNA
FSTSAQNNAKVAVLGASGGIGQPLSLLLKNSPLVSRLTLYDIAHTPGVAADLSHIETRANVKGYLGPEQLPDCL
KGCDVVVIPAGVPRKPGMTRDDLFNTNATIVATLTAACAQHCPEAMICIIANPNVNSTIPITAEVFKKHGVYNPN
KIFGVTTLDIVRANTFVAELKGLDPARVNVFVIGGHAGKTIIPVISQCTPKVDFFPQDQLATLTGRIQEAGTEVV
KAKAGAGSATLSMAYAGARFVFSLV DAMNGKEGVVECSFVQSKETECTYFSTPLLLGKKGLEKNLGIGKITPFE
EKMIAEAIPELKASIKKGEDFVKNMK
```

Puma concolor (Cougar):

> Cougar_MDH2 | 138-977_1 PREDICTED: Puma concolor malate dehydrogenase 2 (MDH2), transcript variant X2, mRNA

FSTSAQNNNAKVAVLGASGGIGQPLSLLLKNSPLVSRLTLYDIAHTPGVAADLSHIETRAAVKGYLGPEQLPDCL
KGCDVVVIPAGVPRKPGMTRDDLFNTNASIVATLTAACAQHCPEAMICIIISNPGLDPARVNVFVIGGHAGKTII
PLISQCTPKVDLPQDQLTAVTGRIQEAGTEVVAKAGAGSATLSMAYAGARFVFSLVDAINGKEGVVECSFVKS
QETDCPYFSTPLLLGKKGIEKNLGIGKISPFEEKMIAEALPELKASIKKGEEFVKNMK

Erinaceus Europaëus (European Hedgehog):

> HedgeHog_MDH2 | 172-1011_1 PREDICTED: Erinaceus europaëus malate dehydrogenase 2 (MDH2), transcript variant X2, mRNA

FSTSTQNNNAKVAVLGASGGIGQPLSLLLKNSPLVSRLTLYDIAHTPGVAADLSHIETRANVKGYLEGPEQLPDCL
KGCDVVVVIPAGVPRKPGMTRDDLFNTNATIVATLAAACAQHCPEAMICIIANPGLDPARVNVFVIGGHAGKTII
PLISQCTPKVDLPQDKLTALTGRIQEAGTEVVQAKAGAGSATLSMAYAGARFVFSLV DAMNGKEGVVECSFVKS
QETDCTYFSTPLLLGRKGLEKNLGIGKVTPFEEKMISEAIPPELKASIKKGEEFVKNMK

Lipotes vexillifer (Yangtze River Dolphin):

> Dolphin_MDH2 | 49-888_1 PREDICTED: Lipotes vexillifer malate dehydrogenase 2, NAD (mitochondrial) (MDH2), transcript variant X2, mRNA

FSTSAQNNNAKVAVLGASGGIGQPLSLLLKNSPLVSRLTLYDIAHTPGVAADLSHIETRATVKGYLEGPEQLPDCL
KGCDVVVIPAGVPRKPGMTRDDLFNTNATIVATLTAACAQHCPEAMICIIISNPGLDPARVSVFVIGGHAGKTII
PLASQCTPKVDFPQDQLTTLIGRIQEAGTEVVAKAGAGSATLSMAYAGARFVFSLV DAMNGKEGVVECSFVKS
QETDCPFFSTPLLLGKKGIEKNLGIGKISPFEEKMIAEAIPELKASIKKGEEFVKNMK

Myotis lucifugus (Little Brown Bat):

> Bat_MDH2 | 171-1136_1 PREDICTED: Myotis lucifugus malate dehydrogenase 2 (MDH2), transcript variant X1, mRNA

FSTSAQNNNAKVAVLGASGGIGQPLSLLLKNSPLVSRLTLYDIAHTPGVAADLSHIETRASVKGYLEGPEQLPDCL
KGCDLVVIPAGVPRKPGMTRDDLFNTNATIVANLTAACAQNCPEAMICVIANPVNSTIPITSEVFKKHGVYNPN
KIFGVTTLDVVRANAFVAELKGLDPARVNVFVIGGHAGKTIIPLISQCTPKVEFPQDQLTTLTGRIQEAGTEVV
KAKAGAGSATLSMAYAGARFVFSLLDAINGKEGVVECSFVKSQETDCSYFSTPLLLGKKGIEKNLGIGKISSFE
EKMIAEAIPELKASIKKGEDFVKNMK

Carlito syrichta (Philippine tarsier):

> Tarsier_MDH2 | 194-1159_1 PREDICTED: Carlito syrichta malate dehydrogenase 2 (MDH2), transcript variant X1, mRNA

FGTSAQNNNAKVAVLGASGGIGQPLSLLLKNSPLVSRLTLYDIAHTPGVAADLSHIETRATVKGYLEGPEQLPDCL
KGCDVVVIPAGVPRKPGMTRDDLFNTNATIVATLAAACAQHCPEAMICIIANPVNSTIPITAEVFKKHGVYNPN
KVFGVTTLDIVRANTFVAELKGLDPARVNVFVIGGHAGKTIIPLISQCTPKVDFPQDQLTALTGRIQEAGTEVV
KAKAGAGSATLSMAYAGARFVFSLV DAMNGKEGVVECSFVKSQETDCTYFSTPLLLGKKGLEKNLGIGKVSSFE
EKMITEAMPELKASIKKGEEFVKNMK

Alignment (Obtained using MUSCLE via EBI):

CLUSTAL multiple sequence alignment by MUSCLE (3.8)

```
Bat_MDH2      FSTSAQNNAKVAVLGASGGIGQPLSLLLKNSPLVSRLTLYDIAHTPGVAADLSHIETRAS
HedgeHog_MDH2 FSTSTQNNAKVAVLGASGGIGQPLSLLLKNSPLVSRLTLYDIAHTPGVAADLSHIETRAN
Cougar_MDH2    FSTSAQNNAKVAVLGASGGIGQPLSLLLKNSPLVSRLTLYDIAHTPGVAADLSHIETRAA
Dolphin_MDH2   FSTSAQNNAKVAVLGASGGIGQPLSLLLKNSPLVSRLTLYDIAHTPGVAADLSHIETRAT
Camel_MDH2     FSTSAQNNAKVAVLGASGGIGQPLSLLLKNSPLVSRLTLYDIAHTPGVAADLSHIETRAT
Mouse_MDH2     FSTSAQNNAKVAVLGASGGIGQPLSLLLKNSPLVSRLTLYDIAHTPGVAADLSHIETRAN
Human_MDH2     ---SAQNNAKVAVLGASGGIGQPLSLLLKNSPLVSRLTLYDIAHTPGVAADLSHIETKAA
Tarsier_MDH2   FGTSAQNNAKVAVLGASGGIGQPLSLLLKNSPLVSRLTLYDIAHTPGVAADLSHIETRAT
Antartic_MDH2  SARNTSSAPKVALLGASGGIGQPLGLLLKTNPKVASLALYDVANTAGVGADLSHIDTHAQ
Halicephalobus_MDH2 TARNSSSAPKVALLGASGGIGQPLGLLLKTNPKVASLALYDVANTAGVGADLSHIDSAAR
               .:. .***:*****.***..* *: *:***:*.*.**.******: *

Bat_MDH2      VKGYLGPEQLPDCLKGCDLVVIPAGVPRKPGMTRDDLFTNTNATIVANLTAACAQNCPEAM
HedgeHog_MDH2 VKGYLGPEQLPDCLKGCDVVVVPAGVPRKPGMTRDDLFTNTNATIVATLAAACAQHCPEAM
Cougar_MDH2    VKGYLGPEQLPDCLKGCDVVVVPAGVPRKPGMTRDDLFTNTNASIVATLTAACAQHCPEAM
Dolphin_MDH2   VKGYLGPEQLPDCLKGCDVVVVPAGVPRKPGMTRDDLFTNTNATIVATLTAACAQHCPEAM
Camel_MDH2     VKGYLGPEQLPDCLKGCDVVVVPAGVPRKPGMTRDDLFTNTNATIVATLTAACAQHCPEAM
Mouse_MDH2     VKGYLGPEQLPDCLKGCDVVVVPAGVPRKPGMTRDDLFTNTNATIVATLTAACAQHCPEAM
Human_MDH2     VKGYLGPEQLPDCLKGCDVVVVPAGVPRKPGMTRDDLFTNTNATIVATLTAACAQHCPEAM
Tarsier_MDH2   VKGYLGPEQLPDCLKGCDVVVVPAGVPRKPGMTRDDLFTNTNATIVATLAAACAQHCPEAM
Antartic_MDH2  VTAHTGXXXHSALEGADIVVIPAGVPRKPGMTRDDLFTNVNAGIVRDLAEAAAKACPKAF
Halicephalobus_MDH2 VTSHTGPNELHKALEGADVIVIPAGVPRKPGMTRDDLFTNVNAGIVRDLSEAAAKICPKAF
               *..: *      .*:*.*:*:*****.*** ** *: *.*: ***:

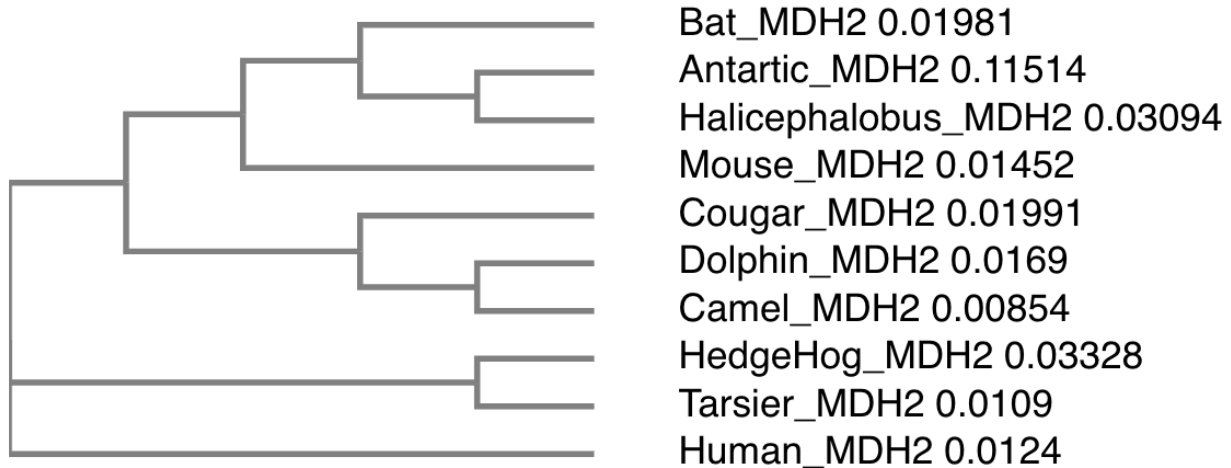
Bat_MDH2      ICVIANPVNSTIPITSEVFKKHGVYNPNKIFGVTTLDVVRANAFVAELKGLDPARVNVVPV
HedgeHog_MDH2 ICIIANP-----GLDPARVNVVPV
Cougar_MDH2    ICIISNP-----GLDPARVNVVPV
Dolphin_MDH2   ICIISNP-----GLDPARVSVVPV
Camel_MDH2     ICIISNPVNSTIPITAIEVFKKHGVYNPNKIFGVTTLDIVRANTFVAELKGLDPARVNVVPV
Mouse_MDH2     ICIIANPVNSTIPITAIEVFKKHGVYNPNKIFGVTTLDIVRANTFVAELKGLDPARVNVVPV
Human_MDH2     ICVIANP-----GLDPARVNVVPV
Tarsier_MDH2   ICIIANPVNSTIPITAIEVFKKHGVYNPNKVFGVTTLDIVRANTFVAELKGLDPARVNVVPV
Antartic_MDH2  VAIITNPVNSTVPPIAAEVYKNNGVYDPKRIFGVTTLDVVRSAFIAELKKLDVSKTVIPV
Halicephalobus_MDH2 VAIITNPVNSTVPPIAAEVYKNNGVYDPRRIFGVTTLDVVRQAQFVAELKGLDVNKTVPV
               :.:*:**                               ** ..:**

Bat_MDH2      IGGHAGKTIIP LISQCTPKVEFPQDQLTTLTGRIQEAGTEVVKAKAGAGSATLSMAYAGA
HedgeHog_MDH2 IGGHAGKTIIP LISQCTPKVDLPQDKLTALTGRIQEAGTEVVQAKAGAGSATLSMAYAGA
Cougar_MDH2    IGGHAGKTIIP LISQCTPKVDLPQDQLTAVTGRIQEAGTEVVKAKAGAGSATLSMAYAGA
Dolphin_MDH2   IGGHAGKTIIP LASQCTPKVDFPQDQLTTLTGRIQEAGTEVVKAKAGAGSATLSMAYAGA
Camel_MDH2     IGGHAGKTIIP VISQCTPKVDFPQDQLTTLTGRIQEAGTEVVKAKAGAGSATLSMAYAGA
Mouse_MDH2     IGGHAGKTIIP LISQCTPKVDFPQDQLATLTGRIQEAGTEVVKAKAGAGSATLSMAYAGA
Human_MDH2     IGGHAGKTIIP LISQCTPKVDFPQDQLTALTGRIQEAGTEVVKAKAGAGSATLSMAYAGA
Tarsier_MDH2   IGGHAGKTIIP LISQCTPKVDFPQDQLTALTGRIQEAGTEVVKAKAGAGSATLSMAYAGA
Antartic_MDH2  IGGHSGVTIIP LLSQCQPSAQFSDSEIEKLTARIQDAGTEVVKAKAGAGSATLSMAFAGA
Halicephalobus_MDH2 IGGHSGVTIIP LLSQLQPGAKFSQDETEKLTARIQDAGTEVVKAKAGGGSATLSMAFAGA
               *****: * ** * ..:..: : .***:*****:****.*****:***
```

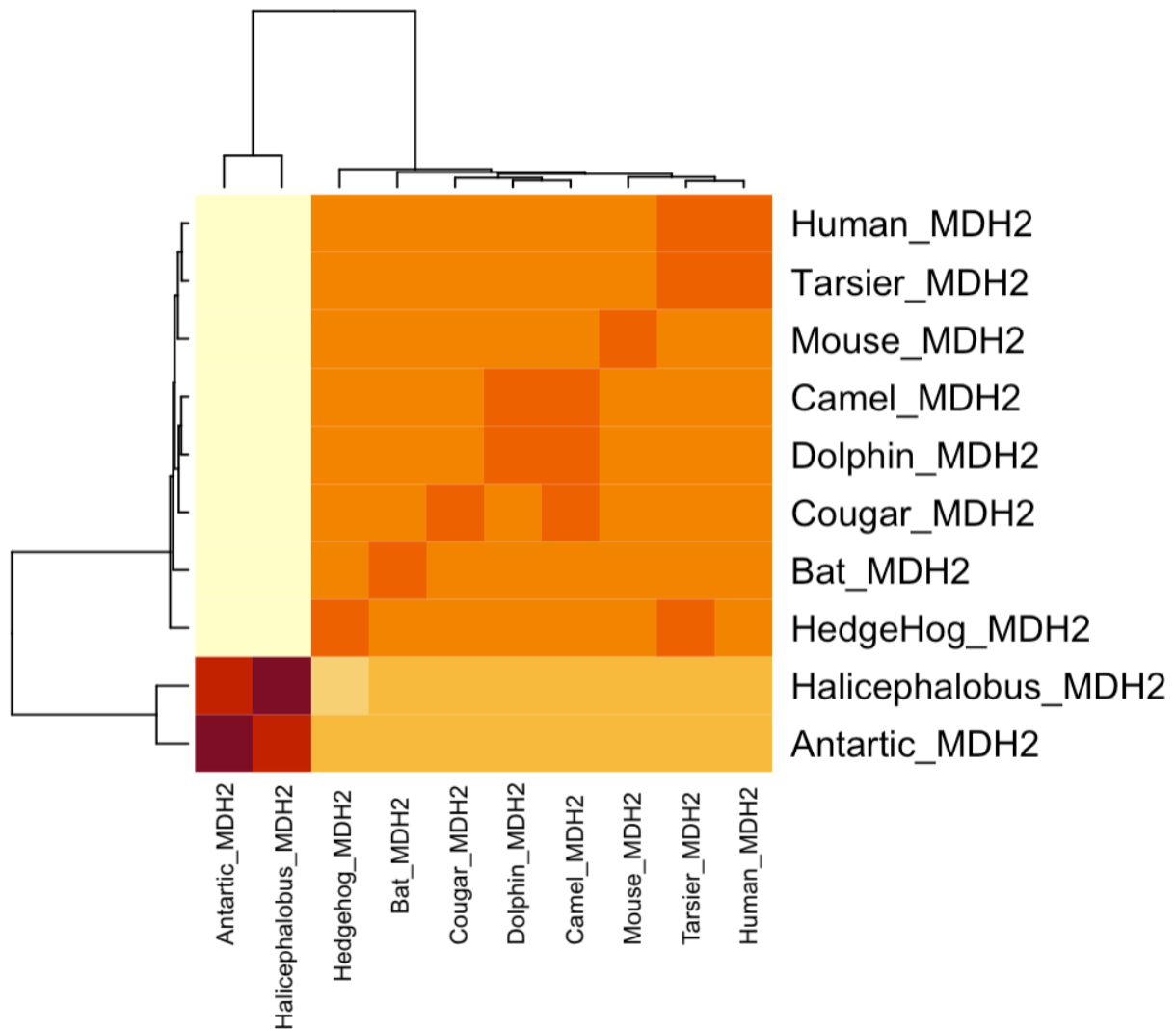
| | |
|---------------------|--|
| Bat_MDH2 | RFVFSLLDAINGKEGVVECSFVKSQETDCS-YFSTPLLLGKKGIEKNLGIGKISSFEEKM |
| HedgeHog_MDH2 | RFVFSLVDAMNGKEGVVECSFVKSQETDCT-YFSTPLLLGRKGLEKNLGIGKVTPFEEKM |
| Cougar_MDH2 | RFVFSLVDAMNGKEGVVECSFVKSQETDCP-YFSTPLLLGKKGIEKNLGIGKISPFEEM |
| Dolphin_MDH2 | RFVFSLVDAMNGKEGVVECSFVKSQETDCP-FFSTPLLLGKKGIEKNLGIGKISPFEEM |
| Camel_MDH2 | RFVFSLLDAMNGKEGVVECSFVKSQETDCP-YFSTPLLLGKKGIEKNLGIGKISPFEEM |
| Mouse_MDH2 | RFVFSLVDAMNGKEGVVECSFVKSQETECT-YFSTPLLLGKKGLEKNLGIGKITPFEEKM |
| Human_MDH2 | RFVFSLVDAMNGKEGVVECSFVKSQETECT-YFSTPLLLGKKGIEKNLGIGKVS----- |
| Tarsier_MDH2 | RFVFSLVDAMNGKEGVVECSFVKSQETDCT-YFSTPLLLGKKGLEKNLGIGKVSSFEEKM |
| Antartic_MDH2 | RFVDALISGLQGKKT-VQCAVQSDVVKGVDYFSTPLELEPNGVEKFLKTVNLXFMKIS- |
| Halicephalobus_MDH2 | RFVQGLIDALQGKKN-VQCTYVQSDVVKGVDFFSTPVELGPNGVEKIL----- |
| | *** .*:...:***: *:*:*:*. .. :****: * :*:** * |

| | |
|---------------------|-------------------------|
| Bat_MDH2 | IAEAIPELKASIKKGEDFVKNMK |
| HedgeHog_MDH2 | ISEAIPELKASIKKGEEFVKNMK |
| Cougar_MDH2 | IAEALPELKASIKKGEEFVKNMK |
| Dolphin_MDH2 | IAEAIPELKASIKKGEEFVKNMK |
| Camel_MDH2 | IAEAIPELKASIKKGEEFVKSMK |
| Mouse_MDH2 | IAEAIPELKASIKKGEDFVKNMK |
| Human_MDH2 | ----- |
| Tarsier_MDH2 | ITEAMPELKASIKKGEEFVKNMK |
| Antartic_MDH2 | ----- |
| Halicephalobus_MDH2 | ----- |

[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.

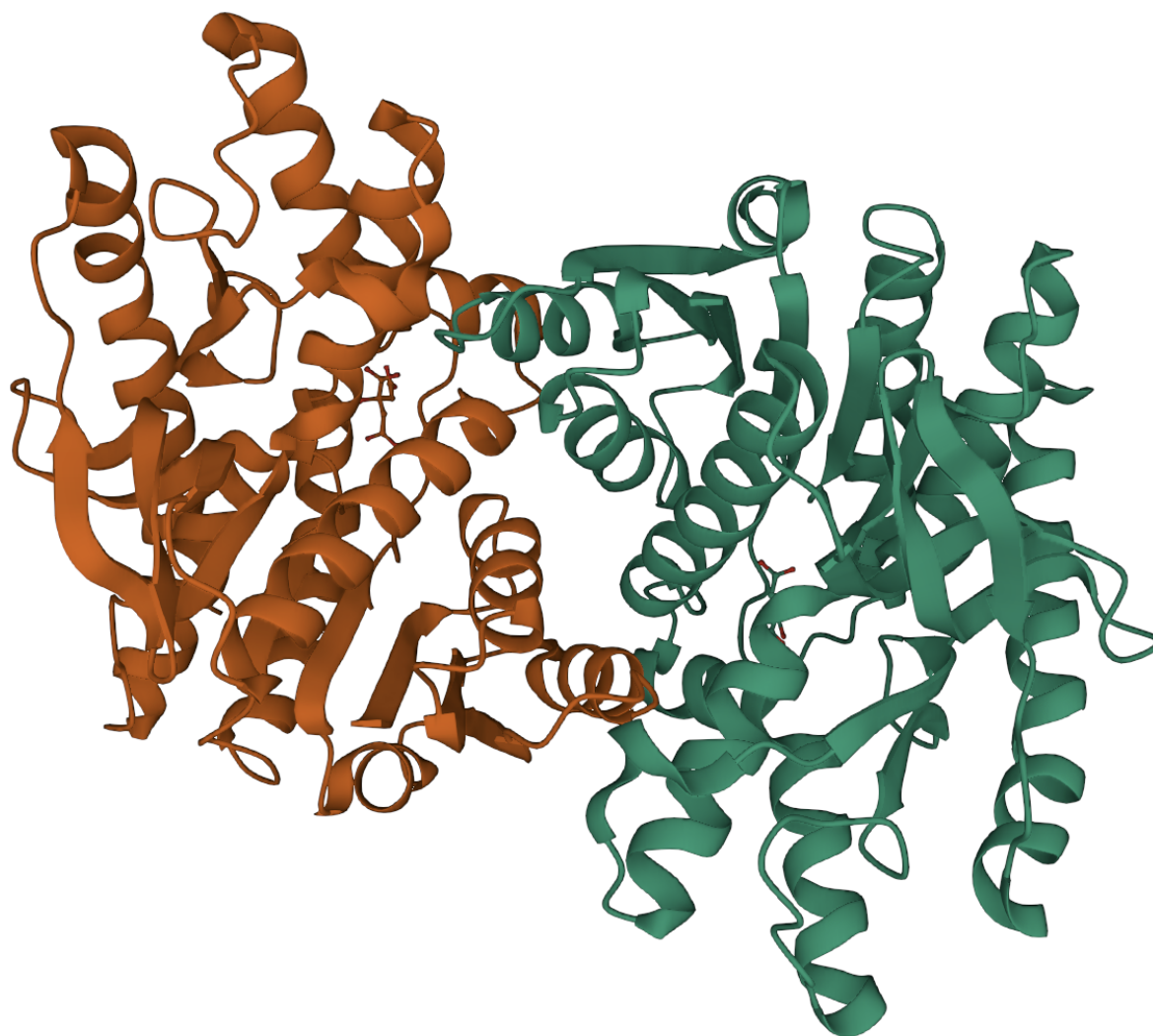
Note that if your consensus sequence has lots of gap positions then it will be better to use an original sequence from the alignment for your search of the PDB. In this case you could choose the sequence with the highest identity to all others in your alignment by calculating the row-wise maximum from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

| ID | Technique | Resolution | Source | Evalue | Identity |
|------|----------------------|------------|--------------------------------------|--------|----------|
| 1MLD | X-RAY DIFFRACTION | 1.83 Å | Sus scrofa (Wild Boar) | 0 | 63% |
| 4E0B | X-RAY DIFFRACTION | 2.17 Å | Vibrio vulnificus (Bacteria) | 0 | 60% |
| 1SMK | X-RAY DIFFRACTION | 2.50 Å | Citrullus lanatus (Watermelon) | 0 | 56% |

[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?

Based on the sequence similarity, I believe the structure to be somewhat similar to my “novel” protein since the sequence similarity is greater than 60%.



[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 data reported that may be useful starting points for exploring potential inhibition of your novel protein?

The “Target” search via ChEMBL outputted 2 Functional Assay (CHEMBL614281, CHEMBL2366649) and 5 Binding Assay (CHEMBL2095180, CHEMBL2189156, CHEMBL2242736, CHEMBL2326, CHEMBL2216)

Only three Binding Assay had ligand efficiency data, which is reported as follows:

| | | | | |
|---------------|--|------------|--|-----------|
| CHEMBL2095180 | | BEI: 27 | | SEI: 6.57 |
| CHEMBL2326 | | BEI: 27.69 | | SEI: 9.20 |
| CHEMBL2216 | | BEI: 35.08 | | SEI: 6.78 |

Scoring Rubric:
[45 total points available]

Q1 (4 points)

| | | |
|------------------|-----|---|
| Protein name | 1 | 1 |
| Species | 1 | 1 |
| Accession number | 1 | 1 |
| Function known | 0.5 | 1 |

Q2 (6 points)

| | | |
|------------------------------------|---|---|
| Blast method | 1 | 1 |
| Database searched | 1 | 1 |
| Limits applied | 1 | 1 |
| Search output list (top hits) | 1 | 1 |
| Alignment of choice | 1 | 1 |
| Evaluate and other alignment stats | 1 | 1 |

Q3 (3 points)

| | | |
|--|---|---|
| Protein sequence of choice matches Subject above | 1 | 1 |
| Name in header | 1 | 1 |
| Species | 1 | 1 |

Q4 (3 point)

| | | |
|---|---|---|
| Blastp output list with identities & Evaluate | 1 | 1 |
| Top alignment shown with alignment statistics | 1 | 1 |
| Results indicates a “novel” gene found | 1 | 1 |

Q5 (3 points)

| | | |
|---|---|---|
| MSA labeled with useful names | 1 | 1 |
| MSA trimmed appropriately (i.e. no gap overhangs) | 1 | 1 |
| Pasted MSA fits report page width (i.e. font, format) | 1 | 1 |

Q6 (1 point)

| | | |
|--|---|---|
| Figure illustrates sequence clustering pattern | 1 | 1 |
|--|---|---|

Q7 (10 points)

| | | |
|--|---|---|
| Heatmap figure included in report | 5 | 5 |
| Heatmap is legible (i.e. no labels obscured) | 5 | 5 |

Q8 (10 points)

| | | |
|--|---|---|
| PDB identifiers from multiple species reported | 5 | 5 |
| Annotation of PDB source, resolution and technique | 4 | 4 |

| | | |
|--|-------------|---------------------|
| Annotation of Evalute and Sequence Identity | 1 | 1 |
| Q9 (4 points) | | |
| Structure figure provided | 2 | 2 |
| Uses white background for molecular figure | 1 | 1 |
| Figure of high resolution (i.e. not just snapshot) | 1 | 1 |
| Q10 (1 point) | | |
| Evidence of ChEMBEL searches | 1 | 1 |
| Final Score: | 44.5 | / 45 = 98.8% |