San Luc

PID: A50910657

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.

Text

Description automatically generated

**Name**: Glutathione S-Transferase

**Species:** *C. elegans*

**Accession**:NP\_501848.1 (protein), NM\_069447.8 (mRNA)

**Function:** this protein enables glutathione transferase activity. It is involved in the glutathione metabolic process.

[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 (2.7.1) search against flatworms ESTs

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

**Organism**: flatworms (taxid:6157)

A picture containing graphical user interface

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Chosen sequence: FY942128 planarian head cDNA library Dugesia japonica cDNA clone Dj\_aH\_304\_P22, mRNA sequence.

Graphical user interface, text, application

Description automatically generated

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Score** | **Expect** | **Identities** | **Positives** | **Gaps** |
| 102 bits(255) | 6e-26 | 66/204(32%) | 103/204(50%) | 9/204(4%) |

Query 6 LLYFDARALAEPIRIMFAMLNVPYEDYRVSVEEWSKLKPTTPFGQLPILQVD-GEQFGQS 64

L YF+AR AE IR + + +V +ED R+ EEW +LKPT P GQLPI+Q+ G +S

Sbjct 7 LTYFNARGKAELIRFVLIVSDVEFEDKRIEFEEWPQLKPTIPTGQLPIVQLSCGGIINES 186

Query 65 MSITRYLARKFGLAGKTAEEEAYADSIVDQYRDFIFFFRQFTSSVFYGSDADHINKVRFE 124

M+I RY A+K+ L G EE D +V D F + VF+ D ++ E

Sbjct 187 MAIARYFAKKYHLTGSNENEEYKVDRVVCTLDD---LFNKVI-DVFHEKDEGKKETLKHE 354

Query 125 VVEPARDDFLAIINKFLAKSKSGFLVGDSLTWADIVIADNLTSLLKNGFLDFNKEKKLEE 184

+ E FL ++ +L F +GD + AD+ + + + ++ + KL

Sbjct 355 LNETHLPAFLDRLDYYLKDKNGDFFLGDHPSLADLQLVNVMDHFEES---QYQSHPKLVH 525

Query 185 FYNKI-HSIPEIKNYVATRKDSIV 207

Y K+ P++K+Y R+ SI+

Sbjct 526 CYQKVLEHYPKLKHYKDNRQKSII 597

[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.

Used EMBOSS transeq to translate the protein sequence above.

**>FY942128.1\_1 planarian head cDNA library Dugesia japonica cDNA clone Dj\_aH\_304\_P22, mRNA sequence**

IILTYFNARGKAELIRFVLIVSDVEFEDKRIEFEEWPQLKPTIPTGQLPIVQLSCGGIIN

ESMAIARYFAKKYHLTGSNENEEYKVDRVVCTLDDLFNKVIDVFHEKDEGKKETLKHELN

ETHLPAFLDRLDYYLKDKNGDFFLGDHPSLADLQLVNVMDHFEESQYQSHPKLVHCYQKV

LEHYPKLKHYKDNRQKSII\*KNSFTVSEYL\*KLMKLF\*LFQKLMIINLLLIVEKKKKKX

>FY942128.1\_2 planarian head cDNA library Dugesia japonica cDNA clone Dj\_aH\_304\_P22, mRNA sequence

LY\*HILMHEEKLN\*FDLF\*S\*AMLNLKIKELNSKNGHN\*NQQFQQVSCQLFNFLVEELSM

KAWQ\*RDILQRNTI\*PDRMKTKNIKLIELCVHSMICLIKLSTCSTRKMKGKRKH\*NMN\*M

KLICLHFLIDSITI\*KIKMAISSSAIILHLLIYNW\*MLWIILKNLNTRAIRN\*YIVIKRY

WNIIQNSSITKIIGKNQ\*SKKIHLLFQNIYKS\*\*NCFNYFKN\*\*\*\*IYY\*SLKKKKKKX

>FY942128.1\_3 planarian head cDNA library Dugesia japonica cDNA clone Dj\_aH\_304\_P22, mRNA sequence

YINIF\*CTRKS\*IDSICFDRKRC\*I\*R\*KN\*IRRMATIETNNSNRSVANCSTFLWRNYQ\*

KHGNSEIFCKEIPFNRIE\*KRRI\*S\*SSCVYTR\*FV\*\*SYRRVPRER\*REKGNIKT\*IK\*

NSFACIS\*\*TRLLFKR\*KWRFLPRRSSFTC\*FTIGKCYGSF\*RISIPEPSEISTLLSKGI

GTLSKTQALQR\*SAKINNLKKFIYCFRIFIKVDEIVLIISKINDNKFIINR\*KKKKKK

>FY942128.1\_4 planarian head cDNA library Dugesia japonica cDNA clone Dj\_aH\_304\_P22, mRNA sequence

FFFFFFND\*\*\*IYYH\*FLK\*LKQFHQLL\*IF\*NSK\*IFLDY\*FLPIIFVMLEFWIMFQYL

LITMY\*FRMALVLRFFKMIHNIYQL\*ISK\*RMIAEEEIAIFIF\*IVIESIKKCRQMSFI\*

FMF\*CFLFPFIFLVEHVDNFIKQIIECTHNSINFIFFVFIRSG\*MVFLCKISRYCHAFID

NSSTRKLNNWQLTCWNCWFQLWPFFEFNSFIFKFNIAYDQNKSNQFSFSSCIKIC\*Y

>FY942128.1\_5 planarian head cDNA library Dugesia japonica cDNA clone Dj\_aH\_304\_P22, mRNA sequence

FFFFFFQRLIINLLSLIFEIIKTISSTFINILKQ\*MNFFRLLIFADYLCNA\*VLDNVPIP

FDNNVLISDGSGIEILQNDP\*HLPIVNQQVKDDRRGRNRHFYLLNSNRVYQEMQANEFHL

IHVLMFPFSLHLSRGTRR\*LY\*TNHRVYTQLDQLYILRFHSIRLNGISLQNISLLPCFH\*

\*FLHKKVEQLATDLLELLVSIVAILRIQFFYLQIQHRLRSKQIESIQLFLVH\*NMLI\*X

>FY942128.1\_6 planarian head cDNA library Dugesia japonica cDNA clone Dj\_aH\_304\_P22, mRNA sequence

FFFFFFSTINNKFIIINF\*NN\*NNFINFYKYSETVNEFF\*IIDFCRLSL\*CLSFG\*CSNT

F\*\*QCTNFGWLWY\*DSSK\*SITFTNCKSASEG\*SPRKKSPFLSFK\*\*SSLSRNAGK\*VSF

NSCFNVSFFPSSFSWNTSITLLNKSSSVHTTRSTLYSSFSFDPVKWYFFAKYLAIAMLSL

IIPPQES\*TIGN\*PVGIVGFNCGHSSNSILLSSNSTSLTIKTNRINSAFPRALKYVNIX

Eukaryota; Metazoa; Spiralia; Lophotrochozoa; Platyhelminthes;

Rhabditophora; Seriata; Tricladida; Continenticola; Geoplanoidea;

Dugesiidae; Dugesia.

**Species*: Dugesia japonica***

[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.

Graphical user interface, text, application

Description automatically generated

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.

**>Dugesia japonica cDNA clone**

IILTYFNARGKAELIRFVLIVSDVEFEDKRIEFEEWPQLKPTIPTGQLPIVQLSCGGIIN ESMAIARYFAKKYHLTGSNENEEYKVDRVVCTLDDLFNKVIDVFHEKDEGKKETLKHELN

ETHLPAFLDRLDYYLKDKNGDFFLGDHPSLADLQLVNVMDHFEESQYQSHPKLVHCYQKV

LEHYPKLKHYKDNRQKSII\*KNSFTVSEYL\*KLMKLF\*LFQKLMIINLLLIVEKKKKKX

**>Glutathione S-transferase 4 [Caenorhabditis elegans]**

MPNYKLLYFDARALAEPIRIMFAMLNVPYEDYRVSVEEWSKLKPTTPFGQLPILQVDGEQFGQSMSITRY

LARKFGLAGKTAEEEAYADSIVDQYRDFIFFFRQFTSSVFYGSDADHINKVRFEVVEPARDDFLAIINKF

LAKSKSGFLVGDSLTWADIVIADNLTSLLKNGFLDFNKEKKLEEFYNKIHSIPEIKNYVATRKDSIV

**> glutathione S transferase-1 [Schmidtea mediterranea]**

MSTVKVTYFDARGRAELIRLVLKASKIEFEDVRITKDKWPEVKPTTPTGKLPVVEYEGKQLTQSMAIARV

VARKHGFMGEDDKEYYLVERAIGQMVDVLEGLYKIYFAPEEKKEELRAEYVATSGRDNLKALEGFIKETG

FFAGEKITLAELFFLVVSDYLVKLPQLYDDFPKLKELRERILKANTDVEEWVNTRPVTEM

**> glutathione S-transferase-like [Crassostrea virginica]**

MTKYTVHYFNVKGRGEIVRLILVAAGVDFEDNRVEREDWPKLKPTMPAGQMPVLEVDGKKYCQSIAIARY

LAREFGLGGSTNVEQLQVDQVVDTISDFLTEMYKPVFEQDATRKAEMNKKLNEETIPRVLGILQNFLEGN

GGDYFVGSKTSLADIYFMDVVSRLVEKDEKVLEKFPKLAASLQKTQALPKIEAYLAKRPKTEL

**>hypothetical protein P879\_00235 [Paragonimus westermani]**

LTYFNGRGRAEYIRMVLHAADLEFEDHRIEMNDWPTIKPTIAGGQLPVLDVTTCCGKSKQMNESMAIARW

FARKHHMMGSNDEEYYEVERVIGQCSDIYQDVYRIFRATGEEKQKLLKQFTEGNGPRLLKVISKHLEASP

TGLVVGDKPTLADFCILCAIDQVEVTVPGLSKDKFPIFERHRETVLKKHAKLAAYMETRPTT

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

Caenorhabditis MPNYKLLYFDARALAEPIRIMFAMLNVPYEDYRVSVEEWSKLKPTTPFGQLPILQV---D

Crassostrea MTKYTVHYFNVKGRGEIVRLILVAAGVDFEDNRVEREDWPKLKPTMPAGQMPVLEV---D

Dugesia ---IILTYFNARGKAELIRFVLIVSDVEFEDKRIEFEEWPQLKPTIPTGQLPIVQLS--C

Schmidtea MSTVKVTYFDARGRAELIRLVLKASKIEFEDVRITKDKWPEVKPTTPTGKLPVVEY---E

Paragonimus -----LTYFNGRGRAEYIRMVLHAADLEFEDHRIEMNDWPTIKPTIAGGQLPVLDVTTCC

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

Caenorhabditis G--EQFGQSMSITRYLARKFGLAGKTAEEEAYADSIVDQYRDFIFFFRQFTSSVFYGSDA

Crassostrea G--KKYCQSIAIARYLAREFGLGGSTNVEQLQVDQVVDTISDFL----TEMYKPVFEQDA

Dugesia G--GIINESMAIARYFAKKYHLTGSNENEEYKVDRVVCTLDDLF----NKVIDVFHEKDE

Schmidtea G--KQLTQSMAIARVVARKHGFMGEDDKEYYLVERAIGQMVDVL----EGLYKIYFAPEE

Paragonimus GKSKQMNESMAIARWFARKHHMMGSNDEEYYEVERVIGQCSDIY----QDVYRIFRATGE

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

Caenorhabditis DHINKVRFEVVEPARDDFLAIINKFLAKSKSGFLVGDSLTWADIVIADNLTSLLKNGFLD

Crassostrea TRKAEMNKKLNEETIPRVLGILQNFLEGNGGDYFVGSKTSLADIYFMDVVSRLVEKDEKV

Dugesia GKKETLKHELNETHLPAFLDRLDYYLKDKNGDFFLGDHPSLADLQLVNVMDHFE---ESQ

Schmidtea -KKEELRAEYVATSGRDNLKALEGFIKE--TGFFAGEKITLAELFFLVVSDYLV-KLPQL

Paragonimus -EKQKLLKQFTEGNGPRLLKVISKHLEASPTGLVVGDKPTLADFCILCAIDQVEVTVPGL

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

Caenorhabditis F-NKEKKLEEFYNKI-HSIPEIKNYVATRKDSIV

Crassostrea L-EKFPKLAASLQKT-QALPKIEAYLAKRPKTEL

Dugesia Y-QSHPKLVHCYQKVLEHYPKLKHYKDNRQKSII

Schmidtea Y-DDFPKLKELRERILKANTDVEEWVNTRPVTEM

Paragonimus SKDKFPIFERHRETVLKKHAKLAAYMETRPTT--

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

**NOTE: I added sequences from the planarian class too due to the lack sequences from the same family.**

[**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.

Graphical user interface, text, application, website

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.

Graphical user interface, text

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

Chart, waterfall 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).

In R, using the bio3d package, I calculated a consensus between all five sequences using consensus () function, however, there are too many gaps, so I calculated the rowSums of the sequence identities.

Graphical user interface, text, application

Description automatically generated

Since hypothetical protein P879\_00235 [Paragonimus westermani] has the highest sequence id calculation, it was chosen to blast for a structure on pdb. Blast using blast.pdb(). Here is the result.

A picture containing text

Description automatically generated

A picture containing text

Description automatically generated

Analyze the blast data using plot.blast() and annotate them using pdb.annotate()

A picture containing graphical user interface

Description automatically generated

Graphical user interface, application

Description automatically generated

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Technique** | **Resolution** | **Source** | **Evalue** | **Identity** |
| 1U3I | X-RAY DIFFRACTION | 1.89 | *Schistosoma mansoni* | 1.58e-62 | 47.030 |
| 1OE7 | X-RAY DIFFRACTION | 1.80 | *Schistosoma haematobium* | 9.26e-63 | 47.525 |
| 2WB9 | X-RAY DIFFRACTION | 1.59 | *Fasciola hepatica* | 6.24e-62 | 46.040 |

[**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?

A picture containing text, vector graphics

Description automatically generated

This structure is 1OE7 from *Schistosoma haematobium.* The ligand (glutathione) interacts with chain A (in red) at Lys45 (not shown on structure). Based on sequence similarity, the novel protein and this structure might not have very similar structure, since it only has an identity score of 47.5.

[**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?

Using chEMBEL target, I changed querystring with my novel sequence, but that do not yield any result.

Graphical user interface, text, application

Description automatically generated