Find A Gene Project

BIMM143

Hak010@ucsd.edu

A17238558

[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.

Name: *SLC6A4*

Ascension: EAW51223.1

Species: **Homo sapiens**

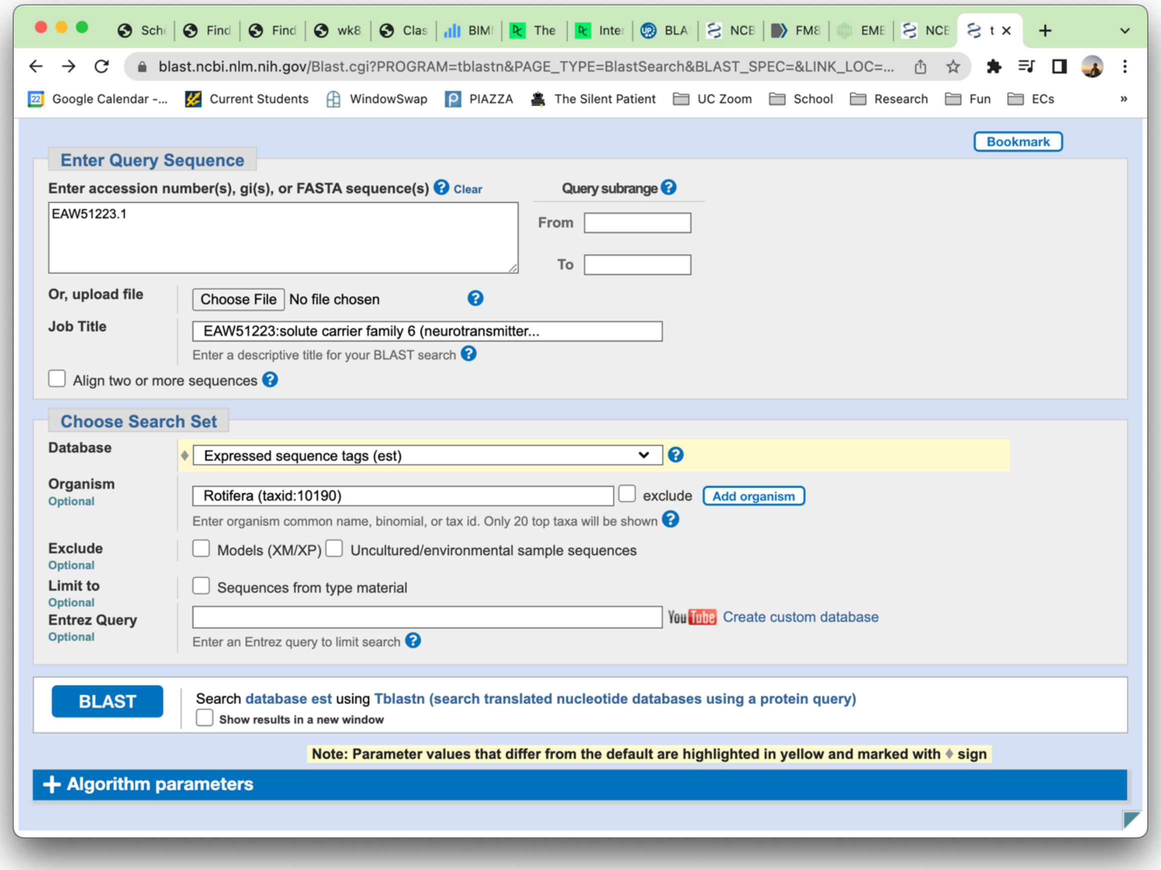
Function: This gene encodes an integral membrane protein that transports the neurotransmitter serotonin from synaptic spaces into presynaptic neurons. The encoded protein terminates the action of serotonin and recycles it in a sodium-dependent manner. This protein is a target of psychomotor stimulants, such as amphetamines and cocaine, and is a member of the sodium:neurotransmitter symporter family. A repeat length polymorphism in the promoter of this gene has been shown to affect the rate of serotonin uptake. There have been conflicting results in the literature about the possible effect, if any, that this polymorphism may play in behavior and depression. [provided by RefSeq, May 2019]

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

Method: tBLASTn search against rotifera ESTs

Database searched: Expressed tag sequence (EST)

Organism: Rotifera (Taxid: 10190)



Enter Query Sequence

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

EAW51223.1

Query subrange

From

To

Or, upload file: Choose File No file chosen

Job Title: EAW51223:solute carrier family 6 (neurotransmitter...

Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database: Expressed sequence tags (est)

Organism (Optional): Rotifera (taxid:10190) ☐ 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): Create custom database

Enter an Entrez query to limit search

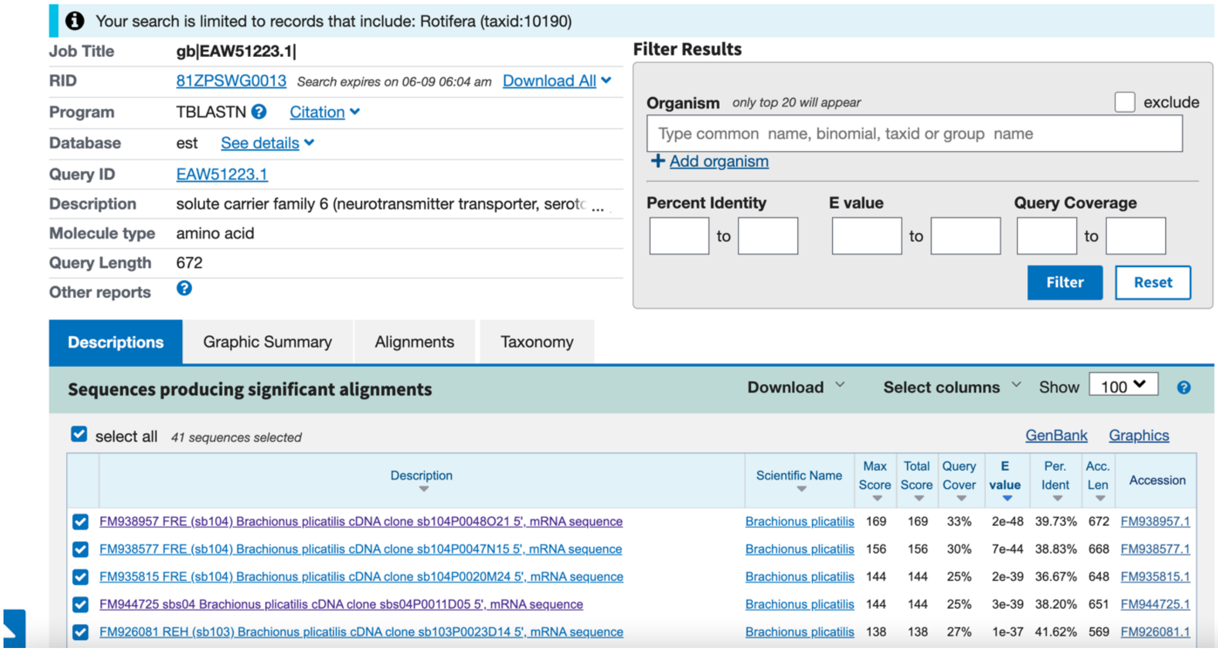
BLAST

Search database est using Tblastn (search translated nucleotide databases using a protein query)

☐ Show results in a new window

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

+ Algorithm parameters



Your search is limited to records that include: Rotifera (taxid:10190)

Job Title	gb|EAW51223.1|
RID	81ZPSWG0013 *Search expires on 06-09 06:04 am* Download All
Program	TBLASTN Citation
Database	est See details
Query ID	EAW51223.1
Description	solute carrier family 6 (neurotransmitter transporter, seroto ...
Molecule type	amino acid
Query Length	672
Other reports	

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 ___

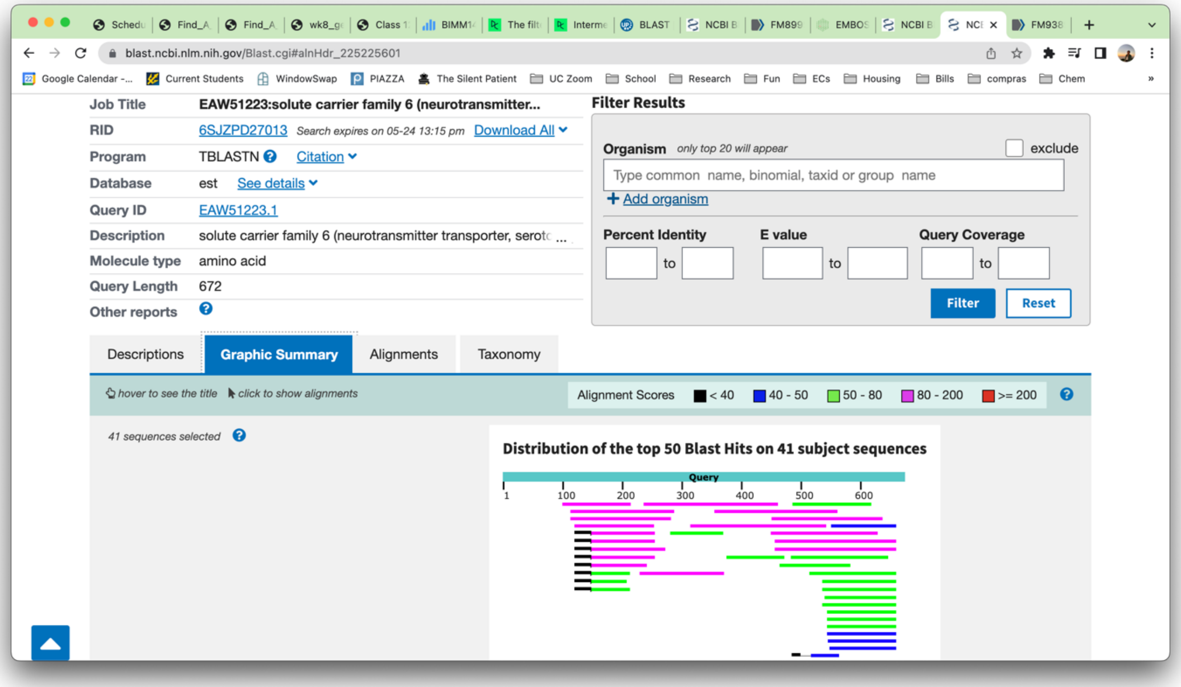
Filter Reset

Sequences producing significant alignments

Download Select columns Show 100

☑ select all *41 sequences selected* GenBank Graphics

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	FM938957 FRE (sb104) Brachionus plicatilis cDNA clone sb104P0048O21 5', mRNA sequence	Brachionus plicatilis	169	169	33%	2e-48	39.73%	672	FM938957.1
☑	FM938577 FRE (sb104) Brachionus plicatilis cDNA clone sb104P0047N15 5', mRNA sequence	Brachionus plicatilis	156	156	30%	7e-44	38.83%	668	FM938577.1
☑	FM935815 FRE (sb104) Brachionus plicatilis cDNA clone sb104P0020M24 5', mRNA sequence	Brachionus plicatilis	144	144	25%	2e-39	36.67%	648	FM935815.1
☑	FM944725 sbs04 Brachionus plicatilis cDNA clone sbs04P0011D05 5', mRNA sequence	Brachionus plicatilis	144	144	25%	3e-39	38.20%	651	FM944725.1
☑	FM926081 REH (sb103) Brachionus plicatilis cDNA clone sb103P0023D14 5', mRNA sequence	Brachionus plicatilis	138	138	27%	1e-37	41.62%	569	FM926081.1



Job Title	EAW51223:solute carrier family 6 (neurotransmitter...
RID	6SJZPD27013 *Search expires on 05-24 13:15 pm* Download All
Program	TBLASTN Citation
Database	est See details
Query ID	EAW51223.1
Description	solute carrier family 6 (neurotransmitter transporter, serotc ...
Molecule type	amino acid
Query Length	672
Other reports	

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

hover to see the title *click to show alignments*

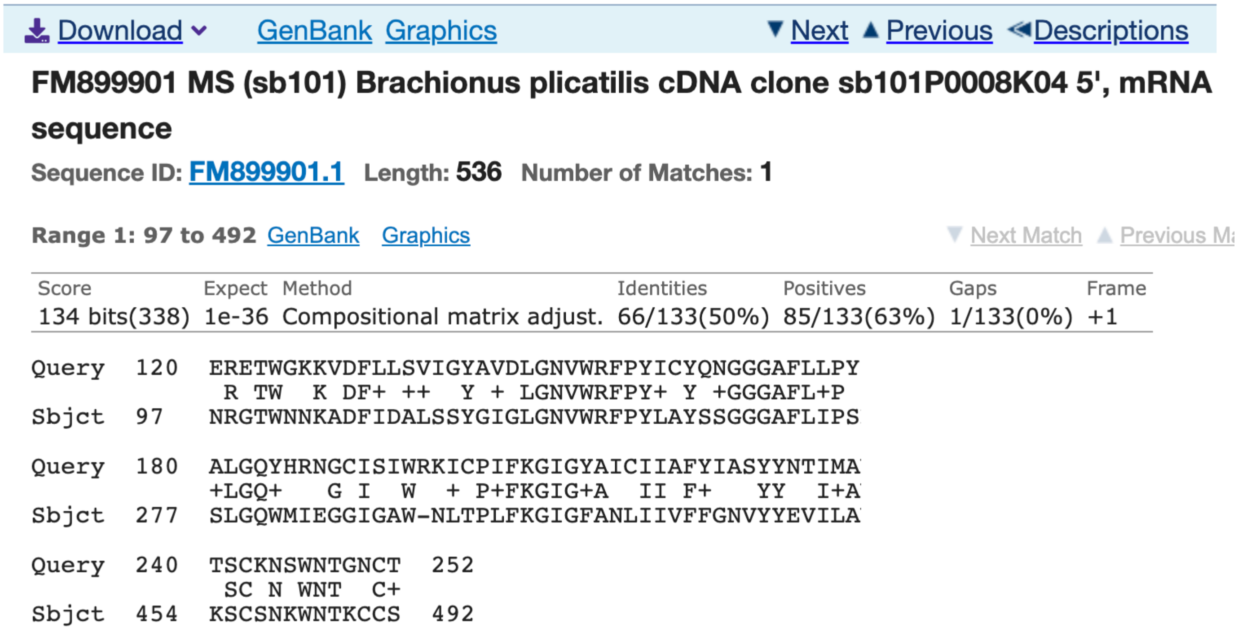
Alignment Scores ■ < 40 🟦 40 - 50 🟩 50 - 80 🟪 80 - 200 🟥 >= 200

41 sequences selected

Distribution of the top 50 Blast Hits on 41 subject sequences

On the BLAST results, clearly indicate a match that represents a protein sequence, encoded from some DNA sequence, that is homologous to your query protein. I need to be able to inspect the pairwise alignment you have selected, including the E value and score. It should be labeled a "genomic clone" or "mRNA sequence", etc. - but include no functional annotation.

Chosen match: Accession FM899901, Brachionus plicatilis cDNA clone sb101P0008K04 5', mRNA sequence



FM899901 MS (sb101) Brachionus plicatilis cDNA clone sb101P0008K04 5', mRNA sequence

Sequence ID: FM899901.1 Length: 536 Number of Matches: 1

Range 1: 97 to 492 GenBank Graphics

Score	Expect	Method	Identities	Positives	Gaps	Frame
134 bits(338)	1e-36	Compositional matrix adjust.	66/133(50%)	85/133(63%)	1/133(0%)	+1

```
Query  120  ERETWGKKVDFLLSVIGYAVDLGNVWRFPYICYQNGGGAFLLPY
             R TW  K DF+ ++   Y + LGNVWRFPY+ Y +GGGAFL+P
Sbjct  97   NRGTWNNKADFIDALSSYGIGLGNVWRFPYLAYSSGGGAFLIPS

Query  180  ALGQYHRNGCISIWRKICPIFKGIGYAICIIAFYIASYYNTIMA
            +LGQ+   G I  W  + P+FKGIG+A  II F+   YY  I+A
Sbjct  277  SLGQWMIEGGIGAW-NLTPLFKGIGFANLIIVFFGNVYYEVILA

Query  240  TSCKNSWNTGNCT  252
             SC N WNT  C+
Sbjct  454  KSCSNKWNTKCCS  492
```

In general, [Q2] is the most difficult for students because it requires you to have a "feel" for how to interpret BLAST results. You need to distinguish between a perfect match to your query (i.e. a sequence that is not "novel"), a near match (something that might be "novel", depending on the results of [Q4]), and a non-homologous result. If you are having trouble finding a novel gene try restricting your search to an organism that is poorly annotated.

[Q3] Gather information about this "novel" protein. At a minimum, show me the protein sequence of the "novel" protein as displayed in your BLAST results from [Q2] as FASTA format (you can copy and paste the aligned sequence subject lines from your BLAST result page if necessary) or translate your novel DNA sequence using a tool called EMBOSS Transeq at the EBI. Don't forget to translate all six reading frames; the ORF (open reading frame) is likely to be the longest sequence without a stop codon. It may not start with a methionine if you don't have the complete coding region. Make sure the sequence you provide includes a header/subject line and is in traditional FASTA format. 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.

Chosen Sequence:

>FM938426 FRE (sb104) Brachionus plicatilis cDNA clone sb104P0047G21 5', mRNA sequence

NRGTWNNKADFIDALSSYGIGLGNVWRFPYLAYSSGGGAFLIPSLISSIVVGIPYAILEV

SLGQWMIEGGIGAW-NLTPLFKGIGFANLIIVFFGNVYYEVILAWTLRYLYDSFSYGLPW

KSCSNKWNTKCCS

>B. plicatilis protein (sequence taken from BLAST result)

YTNF-SLINIHFF-FFP-KKSIQCLQVMTSNHNRGTWNNKADFIDALSSYGIGLGNVWRF
PYLAYSSGGGAFLIPSLISSIVVGIPYAILEVSLGQWMIEGGIGAWNLTPLFKGIGFANL
IIVFFGNVYYEVILAWTLRYLYDSFSYGLPWKSCSNKWNTKCCSEELLYGMSRDTAYTX

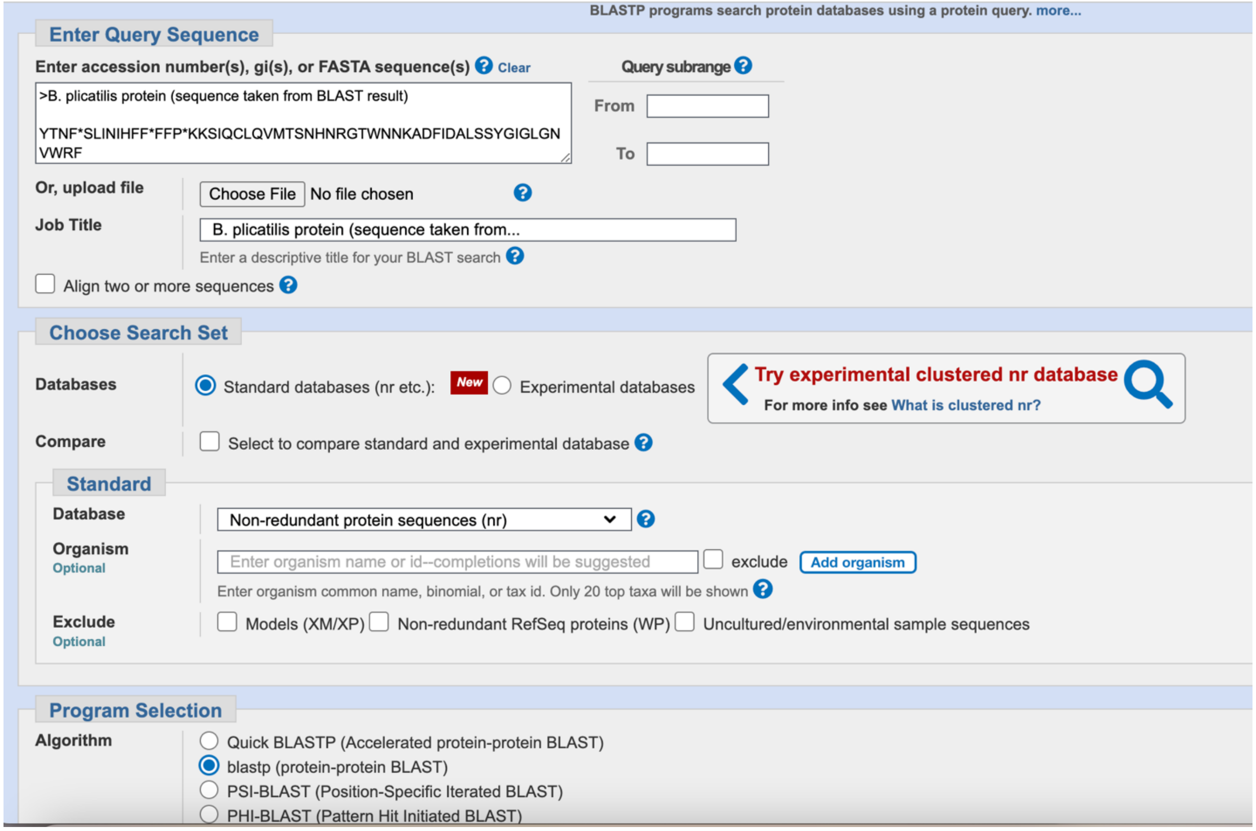
Name: Brachionus plicatilis

Species: Rotifera

Eukaryota; Metazoa; Spiralia; Gnathifera; Rotifera; Eurotatoria;

Monogononta; Pseudotrocha; Ploima; Brachionidae; Brachionus.

[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.



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

Enter Query Sequence

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

```
>B. plicatilis protein (sequence taken from BLAST result)

YTNF*SLINIHFF*FFP*KKSIQCLQVMTSNHNRGTWNNKADFIDALSSYGIGLGN
VWRF
```

Query subrange

From

To

Or, upload file: Choose File No file chosen

Job Title: B. plicatilis protein (sequence taken from...

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Databases: ● Standard databases (nr etc.): New ○ Experimental databases

Try experimental clustered nr database

For more info see What is clustered nr?

Compare: □ Select to compare standard and experimental database

Standard

Database: Non-redundant protein sequences (nr)

Organism (Optional): Enter organism name or id--completions will be suggested □ exclude Add organism

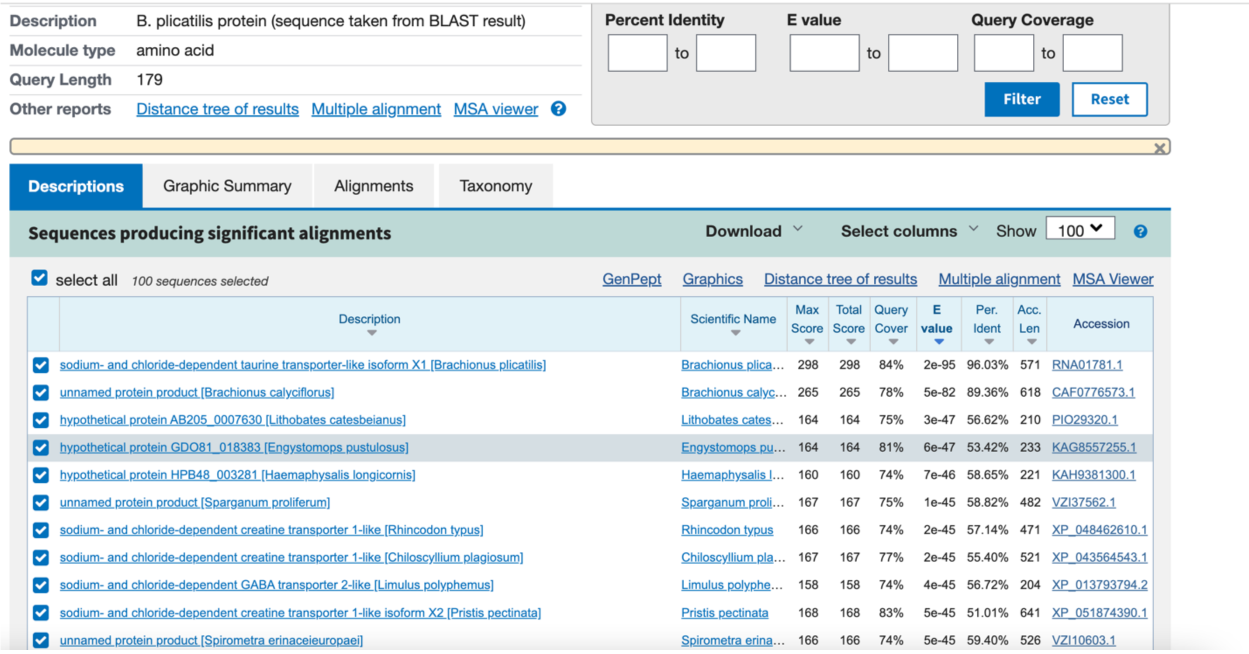
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

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)



Description	B. plicatilis protein (sequence taken from BLAST result)
Molecule type	amino acid
Query Length	179
Other reports	Distance tree of results Multiple alignment MSA viewer

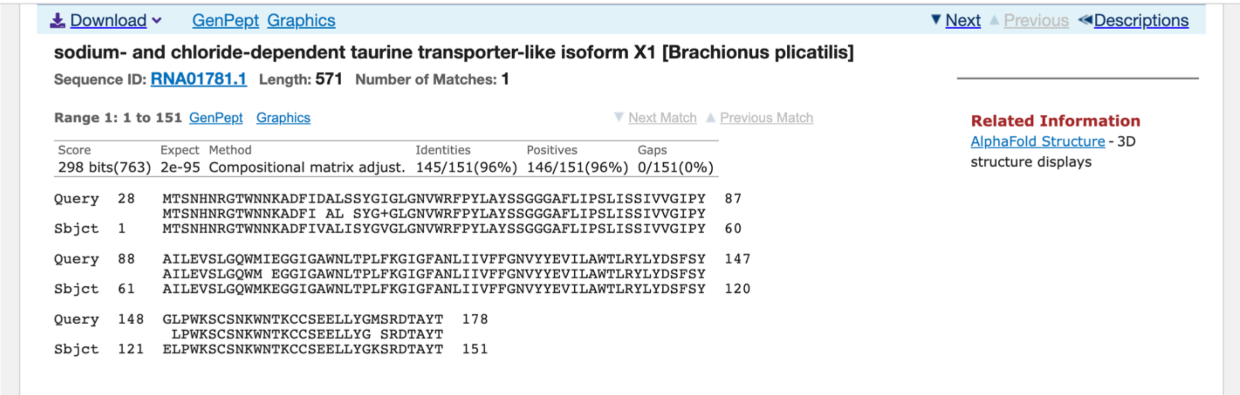
Percent Identity ___ to ___ E value ___ to ___ Query Coverage ___ to ___ Filter Reset

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Select columns Show 100

☑ select all *100 sequences selected* GenPept Graphics Distance tree of results Multiple alignment MSA Viewer

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	sodium- and chloride-dependent taurine transporter-like isoform X1 [Brachionus plicatilis]	Brachionus plica...	298	298	84%	2e-95	96.03%	571	RNA01781.1
☑	unnamed protein product [Brachionus calyciflorus]	Brachionus calyc...	265	265	78%	5e-82	89.36%	618	CAF0776573.1
☑	hypothetical protein AB205_0007630 [Lithobates catesbeianus]	Lithobates cates...	164	164	75%	3e-47	56.62%	210	PIO29320.1
☑	hypothetical protein GDO81_018383 [Engystomops pustulosus]	Engystomops pu...	164	164	81%	6e-47	53.42%	233	KAG8557255.1
☑	hypothetical protein HPB48_003281 [Haemaphysalis longicornis]	Haemaphysalis l...	160	160	74%	7e-46	58.65%	221	KAH9381300.1
☑	unnamed protein product [Sparganum proliferum]	Sparganum proli...	167	167	75%	1e-45	58.82%	482	VZI37562.1
☑	sodium- and chloride-dependent creatine transporter 1-like [Rhincodon typus]	Rhincodon typus	166	166	74%	2e-45	57.14%	471	XP_048462610.1
☑	sodium- and chloride-dependent creatine transporter 1-like [Chiloscyllium plagiosum]	Chiloscyllium pla...	167	167	77%	2e-45	55.40%	521	XP_043564543.1
☑	sodium- and chloride-dependent GABA transporter 2-like [Limulus polyphemus]	Limulus polyphe...	158	158	74%	4e-45	56.72%	204	XP_013793794.2
☑	sodium- and chloride-dependent creatine transporter 1-like isoform X2 [Pristis pectinata]	Pristis pectinata	168	168	83%	5e-45	51.01%	641	XP_051874390.1
☑	unnamed protein product [Spirometra erinaceieuropaei]	Spirometra erina...	166	166	74%	5e-45	59.40%	526	VZI10603.1



Download
GenPept Graphics
Next Previous Descriptions

sodium- and chloride-dependent taurine transporter-like isoform X1 [Brachionus plicatilis]

Sequence ID: RNA01781.1 Length: 571 Number of Matches: 1

Range 1: 1 to 151 GenPept Graphics

Next Match Previous Match
Related Information
AlphaFold Structure - 3D structure displays

Score	Expect	Method	Identities	Positives	Gaps
298 bits(763)	2e-95	Compositional matrix adjust.	145/151(96%)	146/151(96%)	0/151(0%)

```
Query  28   MTSNHNRGTWNNKADFIDALSSYGIGLGNVWRFPYLAYSSGGGAFLIPSLISSIVVGIPY  87
            MTSNHNRGTWNNKADFI AL SYG+GLGNVWRFPYLAYSSGGGAFLIPSLISSIVVGIPY
Sbjct  1    MTSNHNRGTWNNKADFIVALISYGVGLGNVWRFPYLAYSSGGGAFLIPSLISSIVVGIPY  60

Query  88   AILEVSLGQWMIEGGIGAWNLTPLFKGIGFANLIIVFFGNVYYEVILAWTLRYLYDSFSY  147
            AILEVSLGQWM EGGIGAWNLTPLFKGIGFANLIIVFFGNVYYEVILAWTLRYLYDSFSY
Sbjct  61   AILEVSLGQWMKEGGIGAWNLTPLFKGIGFANLIIVFFGNVYYEVILAWTLRYLYDSFSY  120

Query  148  GLPWKSCSNKWNTKCCSEELLYGMSRDTAYT  178
             LPWKSCSNKWNTKCCSEELLYG SRDTAYT
Sbjct  121  ELPWKSCSNKWNTKCCSEELLYGKSRDTAYT  151
```

[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 slc6a4 | solute carrier family 6 (neurotransmitter transporter, serotonin), member 4 [Homo sapiens]
MSQSRRVNPDDRELGGDPQIQAPRDQLGSLADGHQCHLLTSRMETTPLNSQKQLSACEDGEDCQENGVL
QKVVPTPGDKVESGQISNGYSAVPSPGAGDDTRHSIPATTTTLVAELHQGERETWGKKVDFLLSVIGYAVDL
GNVWRFPYICYQNGGGAFLLPYTIMAIFGGIPLFYMELALGQYHRNGCISIWRKICPIFKGIGYAICIIAFYIASY
YNTIMAWALYYLISSFTDQLPWTSCKNSWNTGNCTNYFSEDNITWTLHSTSPAEEFYTRHVLQIHRSKGLQD
LGGISWQLALCIMLIFTVIYFSIWKGVKTSGKVVWVTATFPYIILSVLLVRGATLPGAWRGVLFYLKPNWQKLL
ETGVWIDAAAQIFFSLGPGFGVLLAFASYNKFNNNCYQDALVTSVVNCMTSFVSGFVIFTVLGYMAEMRNE
DVSEVAKDAGPSLLFITYAEAIANMPASTFFAIIFFLMLITLGLDSTFAGLEGVITAVLDEFPHVWAKRRERFVL
AVVITCFFGSLVTLTFGGAYVVKLLEEYATGPAVLTVALIEAVAVSWFYGITQFCRDVKEMLGFSPGWFWRIC
WVAISPLFLLFIICSFLMSPPQLRLFQYNYPYWSIILGYCIGTSSFICIPTYIAYRLIITPGTFKERIIKSITPETPTEIPC
GDIRLNAV
> **Box-like Rotifer** | Brachionus plicatilis cDNA clone sb104P0047G21 5', mRNA sequence

NRGTWNNKADFIDALSSYGIGLGNVWRFPYLAYSSGGGAFLIPSLISSIVVGIPYAILEV

SLGQWMIEGGIGAW-NLTPLFKGIGFANLIIVFFGNVYYEVILAWTLRYLYDSFSYGLPW

KSCSNKWNTKCCS

>Tubeworm | unnamed protein product [Owenia fusiformis]

QEEERETWSKKLDFLLSVIGFAVDLGNVWRFPYICYKNGGGAFLIPYLVMLIFGGLPLFYLELAMGQFQRTGCI
TVWTRICPMFKGIGYGICICAFYVAIYYNTIIAWAVFYLGSCFQAQVPWATCNNEWNTENCTSLAFPDEN--
STVHSNFSESSAEEFFRRRVLQINLSTGINDIGGIRWPIMLCLMAVFLVVYFALWKGIKSVGKAVWVTATLPYI
VLFILLIRGVTLPGSADGILYYITPQWDKLQNRQVWTAAASQIFFSLGPGFGVLLALSSYNKFHNNCYRDALITS
SINCLTSFLAGFVVFSVLGYMAFKQGKDIEKVAE-
PGPGLVFIAYPEALATLDGAVFWSFIFFTMLIMLGLDTTFGGLEAIITAVHDEYPATL-
KRRELFVAVLIVFIFFGALPTTTYGGNYVIQLLDTHGAPIALLFIVFVEAVAVNWFYGVRRFSADIKTMLGAGPG
IFWKICWAGISPIFLFILFIMSCVDYNPDEMDK-
NYQYPRWAIAMGWLVTCSSIICIPIYLIYKFIATEGSIARRAYTIIQPE

> L. anatine | sodium-dependent serotonin transporter isoform X3 [Lingula anatina]

DTHLPPEGALAFPETEQGKVSDSHLAHVHRAHTDEDEEGKGVKNTEVNLKPFPEATTHEREGTVKRISSFDN
SCSETSSAPVADNIDTMSSQAVTQKVVLDGTVDGEKKILAETKDAKKAELP--
ERETWGKKVDFLLSVIGFAVDLGNVWRFPYVCYSNGGGAFLIPYIVMLIF
GGLPLFYMELALGQYQRSGCLTVWKRICPMFKGIGFGIIFIATWVSFYYNTIIAWAFYYLFSSMASEVPWATC
GNPWNTDNCTT-
FRDRSLNKTLAKNNYSKLASHEFFYRGVLELQGEGHVDDIGNIGPVKWQIALCLMAVFVLVYFALWKGVKTS
GKAVWFTATMPYIVLFILLIRGVTLEGSLSGILFYLRPEWDRLLVTQVWIDAAAQIFFSLGPGFGVLLALSSYNK
FHNNCYSDALLTSSINCATSFLAGFVVFSVLGHMAFMEGKDIKTVAQD-
GPGLVFVVYPEAIAALPGSVFWAIIFFLMLITLGLDSTFGGLEALITGICDEFPQTVGKRRELFVAGLMVYCFLG
ALSTTTEGGYNVFVLLDSHGVPISILFIVFIEAIAVNWFYGVNRFSGDIETMLGFQPGIYWKICWVAISPVFLLTL
FILSIVGYKPPVYTHDEPFPGWAIAIGWMITLSSLIPIPTYVVYLLLTSKGGL
KQRLLAMI

>Octopus | sodium-dependent serotonin transporter-like isoform X1 [Octopus sinensis]
VPVP-VGDGTMKIIERPK-------DEEERETWGKKLDFLLSVIGFAVDLGNVWRFPYIC
YRNGGGAFLIPYIIMLVFGGLPLFYMELALGQYQRCGCFTVWNRICPMFKGIGLSIFVIS
TYVAFYYNTIIAWSVYYLFSSFNYEVPWLSCNNSWNSDNCTTFEQRRNQSLPMNLSTSSAQEFFENNILEIQY
SKGIDDVGGVKWKIFLCLLGVFSIVYFSLWKGIKSSGKVVWVTATLPYIVLLVLLVRGCTLPGSYEGIIYYLKPN
WSMLLQPGVWIDAAAQIFFSLGPGFGVLLALSSYNKFNNNCYKDALITSAVNCCTSFFAGFAVFSVLGYMAH
VHKKSVADVSREDVG
IVYPEAIATLKGSVFWAIIFFVMLTTLGLDTTFGGLEAICTGILDEFPKLFLLVYCLLGGLATTTYGGIYVVQLLDT
YGAPISILFVVFLESVAVSWIYGVNRFSDDIESMIGTRPGIFWRGCWAVVSPVFLLMLFTLSVVSDSGPVYGNY
QYPSWSIGIGWIIVCSSLICIPLYIIYKFFTLEGSVCERLRKMIQPSELPKHV

>Polychaete worm | hypothetical protein CAPTEDRAFT_180018 [Capitella teleta]
QRETWGKKLDFLLSVIGFAVDLGNVWRFPYVCYNNGGGAFLVPYMIMYIFGGLPLFYMELALGQFQRCGCI
SVWKRICPMFKGIGFGICVIASYVAMYYNTIIAWSLYFLVSSFRSQVPWATCGNSWNTPNCYSAADLSNPNA
TILPRPNHSVSAANEFFDRSVLEIYKSTGIHDIGNVKWSIALCLIGVFVLVYFALWKGIKSSGKAVWITATLPYVV
LIILLIRGVTLPGSSSGIKYY
LKPEWKKLKDPQIWIAAAAQIFFSLGPGFGVLLALSSYNKFHNNCYKDALVTSTINCFTS
FLAGFVVFSVLGYMAEKQGTSIEKVAQE-GAGLVFVVYPEAIATLRGSSFWAIIFFLMLI
TLGLDSTFCGLEALITGVCDQWPWI-GRKRELFVAGLIVYCFFGALATTTYGGNYVLALL

```
DAHGAPIAILCICFLECIAISWFYGVRRFADDVEKMLGFRPGIFWQICWAGISPCFLFVL
FILSLVYYKP--IVLGSYVYPDWALGLGWVITASSLIWIPIYIVVRFFMTKGSLKDRWRS
MIQPEEMPSRPPDETMQMTPV
```

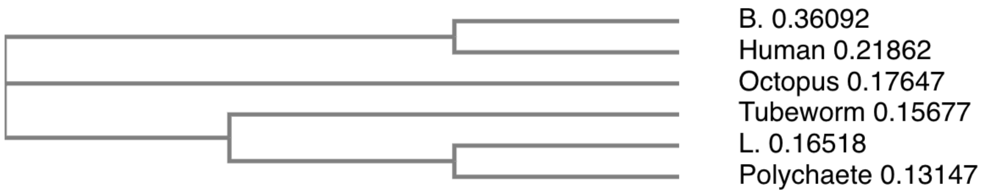
CLUSTAL multiple sequence alignment by MUSCLE (3.8)

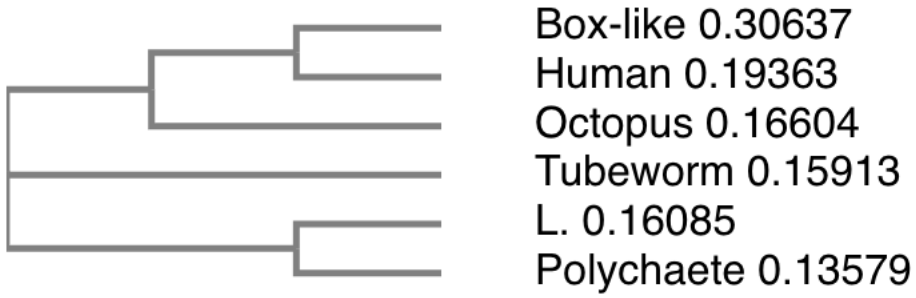
```
B.          --------YTNFSLINIHFFFFPKKSIQCLQVMTSNH-----------------------
Human       MSQSRRVNPDDRELGGDPQIQAPRDQLGSLADGHQCHLLTSRMETT--------------
Octopus     -------VPVPVGDGTMKIIERPKDEE---------------------------------
Tubeworm    ----------------------QEE-----------------------------------
L.          ----------DTHLPPEGALAFPETEQGKVSDSHLAHVHRAHTDEDEEGKGVKNTEVNLK
Polychaete  ------------------------------------------------------------


B.          ------------------------------------------------------------
Human       --PLNSQKQLSACEDGEDCQENGVLQKVVPTPGDKVESGQISNGYSAVPSPGAGDDTRHS
Octopus     ------------------------------------------------------------
Tubeworm    ------------------------------------------------------------
L.          PFPEATTHEREGTVKRISSFDNSCSETSSAPVADNIDTMSSQAVTQKVVLDGTVDGEKKI
Polychaete  ------------------------------------------------------------


B.          ---------------NRGTWNNKADFIDALSSYGIGLGNVWRFPYLAYSSGGGAFLIPSL
Human       IPATTTTLVAELHQGERETWGKKVDFLLSVIGYAVDLGNVWRFPYICYQNGGGAFLLPYT
Octopus     ---------------ERETWGKKLDFLLSVIGFAVDLGNVWRFPYICYRNGGGAFLIPYI
Tubeworm    ---------------ERETWSKKLDFLLSVIGFAVDLGNVWRFPYICYKNGGGAFLIPYL
L.          LAETKDAKKAELP--ERETWGKKVDFLLSVIGFAVDLGNVWRFPYVCYSNGGGAFLIPYI
Polychaete  ---------------QRETWGKKLDFLLSVIGFAVDLGNVWRFPYVCYNNGGGAFLVPYM
                           :* **.:* **: :: .:.:.*********:.* .******:*

B.          ISSIVVGIPYAILEVSLGQWMIEGGIGAWN-LTPLFKGIGFANLIIVFFGNVYYEVILAW
Human       IMAIFGGIPLFYMELALGQYHRNGCISIWRKICPIFKGIGYAICIIAFYIASYYNTIMAW
Octopus     IMLVFGGLPLFYMELALGQYQRCGCFTVWNRICPMFKGIGLSIFVISTYVAFYYNTIIAW
Tubeworm    VMLIFGGLPLFYLELAMGQFQRTGCITVWTRICPMFKGIGYGICICAFYVAIYYNTIIAW
L.          VMLIFGGLPLFYMELALGQYQRSGCLTVWKRICPMFKGIGFGIIFIATWVSFYYNTIIAW
Polychaete  IMYIFGGLPLFYMELALGQFQRCGCISVWKRICPMFKGIGFGICVIASYVAMYYNTIIAW
            : :. *:*   :*:::**:   * :  *  : *:***** .  .   :   **:.*:**

B.          TLRYLYDSFSYGLPWKSCSNKWNTKCCSE-------------------------------
Human       ALYYLISSFTDQLPWTSCKNSWNTGNCTNYFSEDN-----ITWTLHSTSPAEEFYTRHVL
Octopus     SVYYLFSSFNYEVPWLSCNNSWNSDNCTTFEQRRNQ----SLPMNLSTSSAQEFFENNIL
Tubeworm    AVFYLGSCFQAQVPWATCNNEWNTENCTSLAFPDENS---TVHSNFSESSAEEFFRRRVL
L.          AFYYLFSSMASEVPWATCGNPWNTDNCTTFRDRSLNKT--LAKNNYSKLASHEFFYRGVL
```

```
Polychaete     SLYFLVSSFRSQVPWATCGNSWNTPNCYSAADLSNPNATILPRPNHSVSAANEFFDRSVL
           :. :* ..:   :** :* * **:  *

B.          -------------------------------------------------------
Human         QIHRSKGLQDLGGI---SWQLALCIMLIFTVIYFSIWKGVKTSGKVVWVTATFPYIILSV
Octopus       EIQYSKGIDDVGGV---KWKIFLCLLGVFSIVYFSLWKGIKSSGKVVWVTATLPYIVLLV
Tubeworm        QINLSTGINDIGGI---RWPIMLCLMAVFLVVYFALWKGIKSVGKAVWVTATLPYIVLFI
L.          ELQGEGHVDDIGNIGPVKWQIALCLMAVFVLVYFALWKGVKTSGKAVWFTATMPYIVLFI
Polychaete     EIYKSTGIHDIGNV---KWSIALCLIGVFVLVYFALWKGIKSSGKAVWITATLPYVVLII


B.          -------------------------------------------------------
Human         LLVRGATLPGAWRGVLFYLKPNWQKLLETGVWIDAAAQIFFSLGPGFGVLLAFASYNKFN
Octopus       LLVRGCTLPGSYEGIIYYLKPNWSMLLQPGVWIDAAAQIFFSLGPGFGVLLALSSYNKFN
Tubeworm        LLIRGVTLPGSADGILYYITPQWDKLQNRQVWTAAASQIFFSLGPGFGVLLALSSYNKFH
L.          LLIRGVTLEGSLSGILFYLRPEWDRLLVTQVWIDAAAQIFFSLGPGFGVLLALSSYNKFH
Polychaete     LLIRGVTLPGSSSGIKYYLKPEWKKLKDPQIWIAAAAQIFFSLGPGFGVLLALSSYNKFH


B.          -------------------------------------------------------
Human         NNCYQDALVTSVVNCMTSFVSGFVIFTVLGYMAEMRNEDVSEVAKDAGPSLLFITYAEAI
Octopus       NNCYKDALITSAVNCCTSFFAGFAVFSVLGYMAHVHKKSVADVSRE-DVG---IVYPEAI
Tubeworm        NNCYRDALITSSINCLTSFLAGFVVFSVLGYMAFKQGKDIEKVAEP-GPGLVFIAYPEAL
L.          NNCYSDALLTSSINCATSFLAGFVVFSVLGHMAFMEGKDIKTVAQD-GPGLVFVVYPEAI
Polychaete     NNCYKDALVTSTINCFTSFLAGFVVFSVLGYMAEKQGTSIEKVAQE-GAGLVFVVYPEAI


B.          ---------------------------------------------ELL---------
Human         ANMPASTFFAIIFFLMLITLGLDSTFAGLEGVITAVLDEFPHVWAKRRERFVLAVVITCF
Octopus       ATLKGSVFWAIIFFVMLTTLGLDTTFGGLEAICTGILDEFP-------KLF---LLVYCL
Tubeworm        ATLDGAVFWSFIFFTMLIMLGLDTTFGGLEAIITAVHDEYPATL-KRRELFVAVLIVFIF
L.          AALPGSVFWAIIFFLMLITLGLDSTFGGLEALITGICDEFPQTVGKRRELFVAGLMVYCF
Polychaete     ATLRGSSFWAIIFFLMLITLGLDSTFCGLEALITGVCDQWP-WIGRKRELFVAGLIVYCF
                                       : :

B.          -------------------------------------YGMSR-------------
Human         FGSLVTLTFGGAYVVKLLEEYATGPAVLTVALIEAVAVSWFYGITQFCRDVKEMLGFSPG
Octopus       LGGLATTTYGGIYVVQLLDTYGAPISILFVVFLESVAVSWIYGVNRFSDDIESMIGTRPG
Tubeworm        FGALPTTTYGGNYVIQLLDTHGAPIALLFIVFVEAVAVNWFYGVRRFSADIKTMLGAGPG
L.          LGALSTTTEGGYNVFVLLDSHGVPISILFIVFIEAIAVNWFYGVNRFSGDIETMLGFQPG
Polychaete     FGALATTTYGGNYVLALLDAHGAPIAILCICFLECIAISWFYGVRRFADDVEKMLGFRPG
                                  **: .

B.          -------------------------------DTAYTX------------------
```

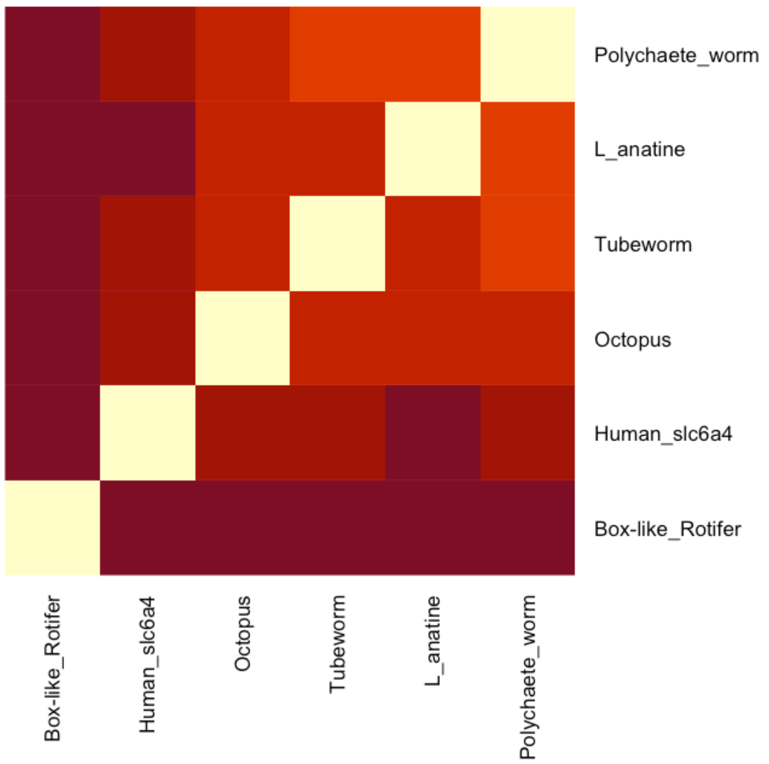
```
Human          WFWRICWVAISPLFL--LFIICSFLMSPPQLRLFQYNYPYWSIILGYCIGTSSFICIPTY
Octopus        IFWRGCWAVVSPVFLLMLFTLSVVSDSGPVYG--NYQYPSWSIGIGWIIVCSSLICIPLY
Tubeworm         IFWKICWAGISPIFLFILFIMSCVDYNPDEMD-KNYQYPRWAIAMGWLVTCSSIICIPIY
L.           IYWKICWVAISPVFLLTLFILSIVGYKPPVYT-HDEPFPGWAIAIGWMITLSSLIPIPTY
Polychaete      IFWQICWAGISPCFLFVLFILSLVYYKPIVLG--SYVYPDWALGLGWVITASSLIWIPIY
                           .  :.

B.           ---------------------------------------
Human          IAYRLIITPGTFKERIIKSITPE-TPTEIPCGDIRLNAV
Octopus        IIYKFFTLEGSVCERLRKMIQPSELPKHV----------
Tubeworm         LIYKFIATEGSIARRAYTIIQPE----------------
L.           VVYLLLTSKGGLKQRLLAMI-------------------
Polychaete      IVVRFFMTKGSLKDRWRSMIQPEEMPSRPPDETMQMTPV
```

[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).

ID	Technique	Resolution	Source	E-value	Identity
6YJJ	X-RAY DIFFRACTION	2.4	Rhincodon typus	6e-45	56.82
7FBK	X-RAY DIFFRACTION	1.9	Chiloscyllium plagiosum	9e-45	56.82
1QTJ	X-RAY DIFFRACTION	3.0	Limulus polyphemus	1e-44	57.25

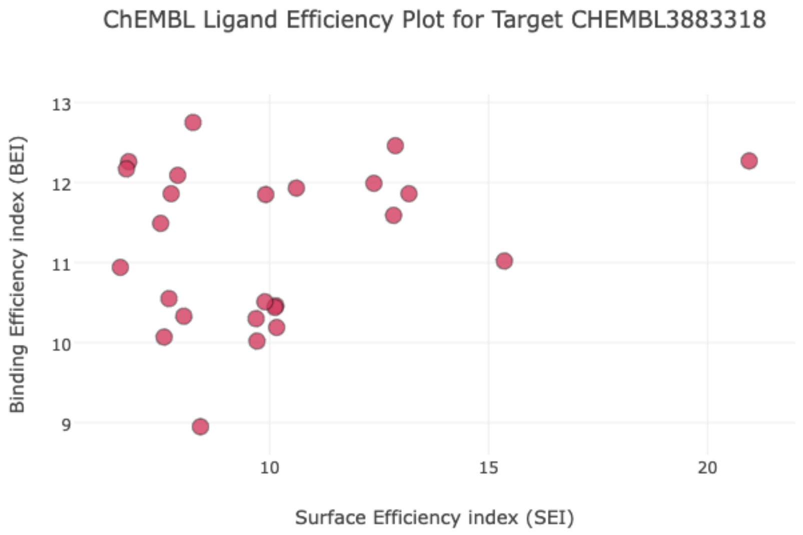
[Q9] Generate a molecular figure of one of your identified PDB structures using VMD. You can optionally highlight conserved residues that are likely to be functional. Please use a white or transparent background for your figure (i.e. not the default black). Based on sequence similarity. How likely is this structure to be similar to your "novel" protein? Very likely to be similar in structure to Anguillicola globin given the high sequence similarity (>80%). In the figure below the beta globin chain B is colored green and corresponds to the Anguillicola globin subject of this report.
ID: 6YJJ

A sequence identity of 56.82% indicates a moderate level of similarity between two protein sequences. While it suggests some commonality in their amino acid composition, it also implies a significant divergence in terms of sequence variation.

[Q10] Perform a “Target” search of ChEMBEL (https://www.ebi.ac.uk/chembl/) with your novel sequence. Are there any Target Associated Assays and ligand efficiency data reported that may be useful starting points for exploring potential inhibition of your novel protein?

CHEMBL details 1 Binding Assay (CHEMBL695842) and 3 Functional Assays; No ligand efficiency data.

https://www.ebi.ac.uk/chembl/target_report_card/CHEMBL3883318/



Chardin, P., Madaule, P., & Tavitian, A. (1988). Coding sequence of human rho cDNAs clone 6 and clone 9. *Nucleic acids research*, *16*(6), 2717. https://doi.org/10.1093/nar/16.6.2717
https://pubmed.ncbi.nlm.nih.gov/3283705/