BGGN-213: FOUNDATIONS OF BIOINFORMATICS
The find-a-gene project assignment

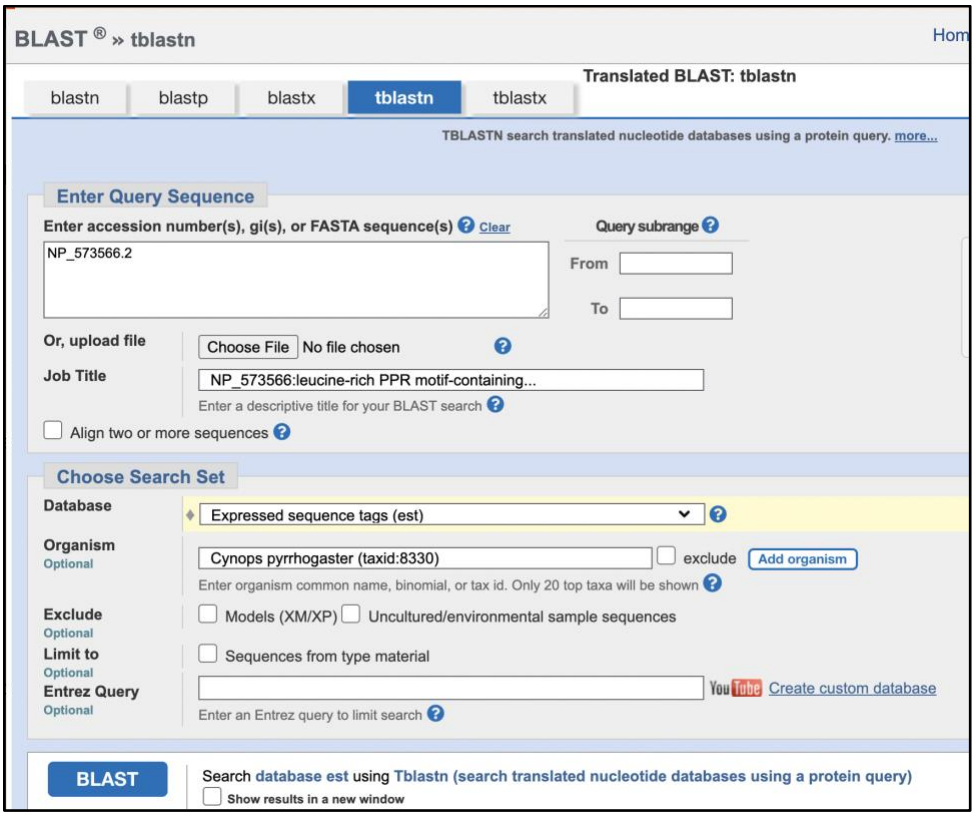
UCSD email: ptvo@ucsd.edu
PID: A59010610

[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. If you do not have a favorite protein, select human RBP4 or KIF11. Do not use beta globin as this is in the worked example report that I provide you with online.

Name: Leucine-rich PPR motif-containing protein (LRPPRC)
Accession: NP_573566
Species: Homo Sapiens

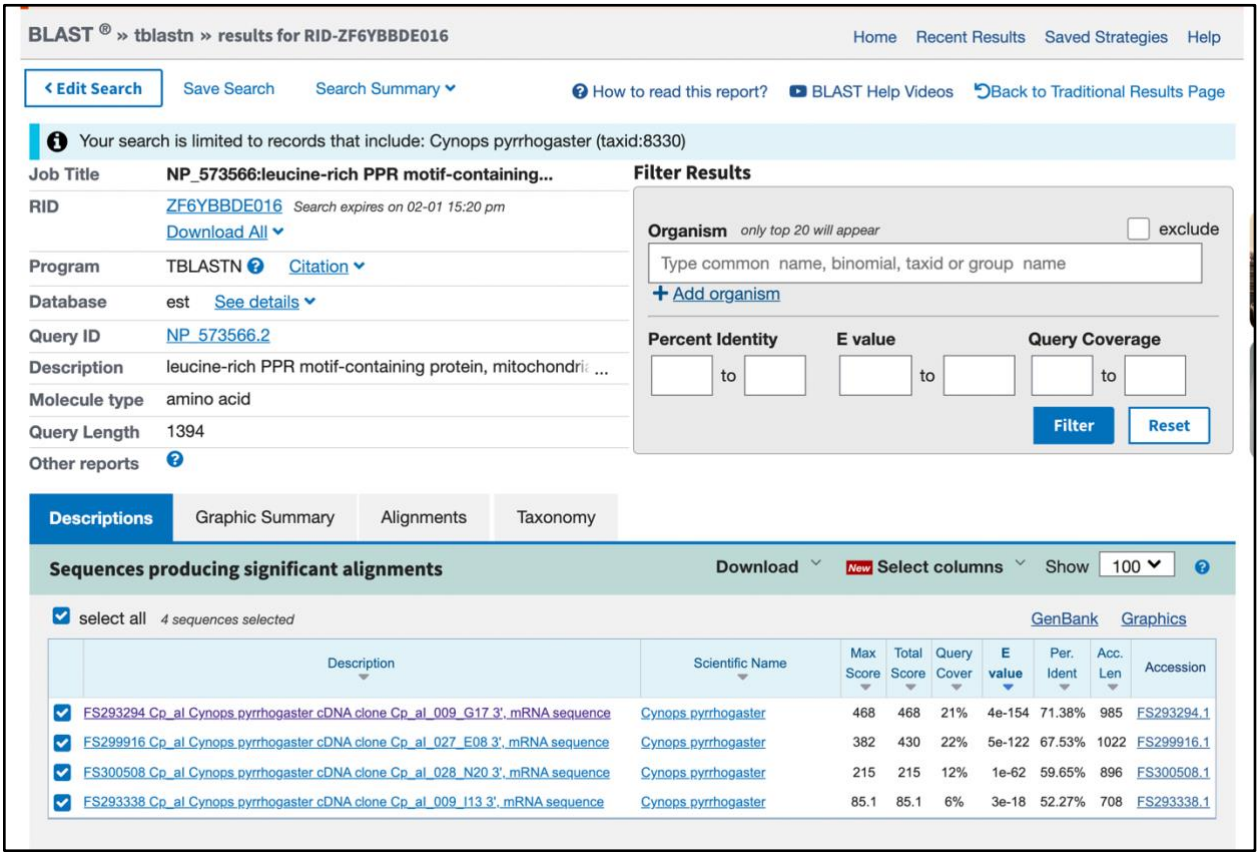
[Q2] Perform a BLAST search against a DNA database, such as a database consisting of genomic DNA or ESTs. The BLAST server can be at NCBI or elsewhere. Include details of the BLAST method used, database searched and any limits applied (e.g. Organism).

Method: TBLASTN against Cynops pyrrhogaster ESTs
Database: Expressed Sequence Tags (est)
Organism: Cynops pyrrhogaster (Taxid: 8330)



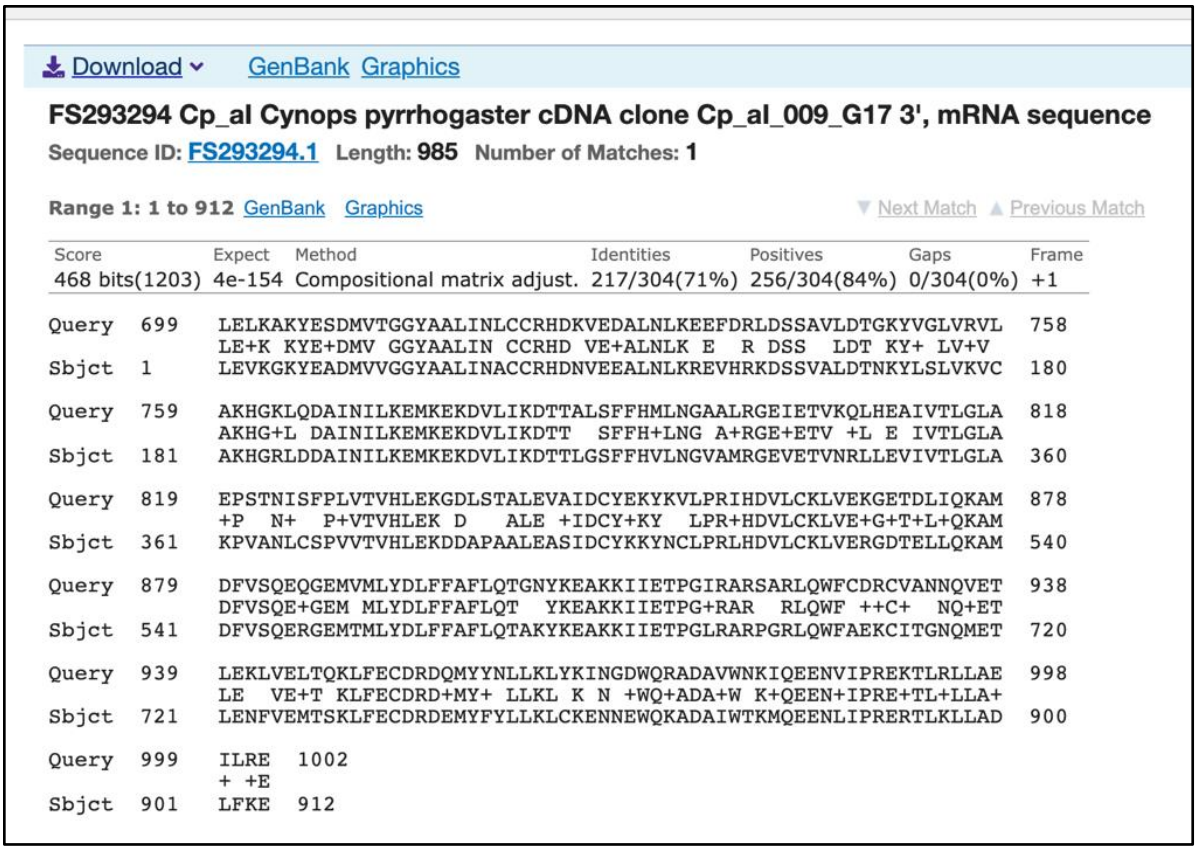
Also include the output of that BLAST search in your document. If appropriate, change the font to Courier size 10 so that the results are displayed neatly. You can also screen capture a BLAST output (e.g. alt print screen on a PC or on a MAC press ⌘-shift-4. The pointer becomes

a bulls eye. Select the area you wish to capture and release. The image is saved as a file called Screen Shot [].png in your Desktop directory). It is not necessary to print out all of the blast results if there are many pages.



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 FS293294.1, a 985 base pair clone from *Cynops pyrrhogaster*. See below for alignment details.



Download GenBank Graphics

FS293294 Cp_al Cynops pyrrhogaster cDNA clone Cp_al_009_G17 3', mRNA sequence

Sequence ID: FS293294.1 Length: 985 Number of Matches: 1

Range 1: 1 to 912 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
468 bits(1203)	4e-154	Compositional matrix adjust.	217/304(71%)	256/304(84%)	0/304(0%)	+1

```
Query  699   LELKAKYESDMVTGGYAALINLCCRHDKVEDALNLKEEFDRLDSSAVLDTGKYVGLVRVL  758
             LE+K KYE+DMV GGYAALIN CCRHD VE+ALNLK E  R DSS  LDT KY+ LV+V
Sbjct  1     LEVKGKYEADMVVGGYAALINACCRHDNVEEALNLKREVHRKDSSVALDTNKYLSLVKVC  180

Query  759   AKHGKLQDAINILKEMKEKDVLIKDTTALSFFHMLNGAALRGEIETVKQLHEAIVTLGLA  818
             AKHG+L DAINILKEMKEKDVLIKDTT  SFFH+LNG A+RGE+ETV +L E IVTLGLA
Sbjct  181   AKHGRLDDAINILKEMKEKDVLIKDTTLGSFFHVLNGVAMRGEVETVNRLLEVIVTLGLA  360

Query  819   EPSTNISFPLVTVHLEKGDLSTALEVAIDCYEKYKVLPRIHDVLCKLVEKGETDLIQKAM  878
             +P  N+  P+VTVHLEK D   ALE +IDCY+KY  LPR+HDVLCKLVE+G+T+L+QKAM
Sbjct  361   KPVANLCSPVVTVHLEKDDAPAALEASIDCYKKYNCLPRLHDVLCKLVERGDTELLQKAM  540

Query  879   DFVSQEQGEMVMLYDLFFAFLQTGNYKEAKKIIETPGIRARSARLQWFCDRCVANNQVET  938
             DFVSQE+GEM MLYDLFFAFLQT  YKEAKKIIETPG+RAR  RLQWF ++C+  NQ+ET
Sbjct  541   DFVSQERGEMTMLYDLFFAFLQTAKYKEAKKIIETPGLRARPGRLQWFAEKCITGNQMET  720

Query  939   LEKLVELTQKLFECDRDQMYYNLLKLYKINGDWQRADAVWNKIQEENVIPREKTLRLLAE  998
             LE  VE+T KLFECDRD+MY+ LLKL K N +WQ+ADA+W K+QEEN+IPRE+TL+LLA+
Sbjct  721   LENFVEMTSKLFECDRDEMYFYLLKLCKENNEWQKADAIWTKMQEENLIPRERTLKLLAD  900

Query  999   ILRE  1002
             + +E
Sbjct  901   LFKE  912
```

[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:

```
> Cynops pyrrhogaster protein (sequence taken from BLAST result)
LEVKGKYEADMVVGGYAALINACCRHDNVEEALNLKREVHRKDSSVALDTNKYLSLVKVCAKHGRLDDAINILKEMK
EKDVLIKDTTLGSFFHVLNGVAMRGEVETVNRLLEVIVTLGLAKPVANLCSPVVTVHLEKDDAPAALEASIDCYKKY
NCLPRLHDVLCKLVERGDTELLQKAMDFVSQERGEMTMLYDLFFAFLQTAKYKEAKKIIETPGLRARPGRLQWFAEK
CITGNQMETLENFVEMTSKLFECDRDEMYFYLLKLCKENNEWQKADAIWTKMQEENLIPRERTLKLLADLFKEMVRK
FHLMFLRIGMKKLQHQK*
```

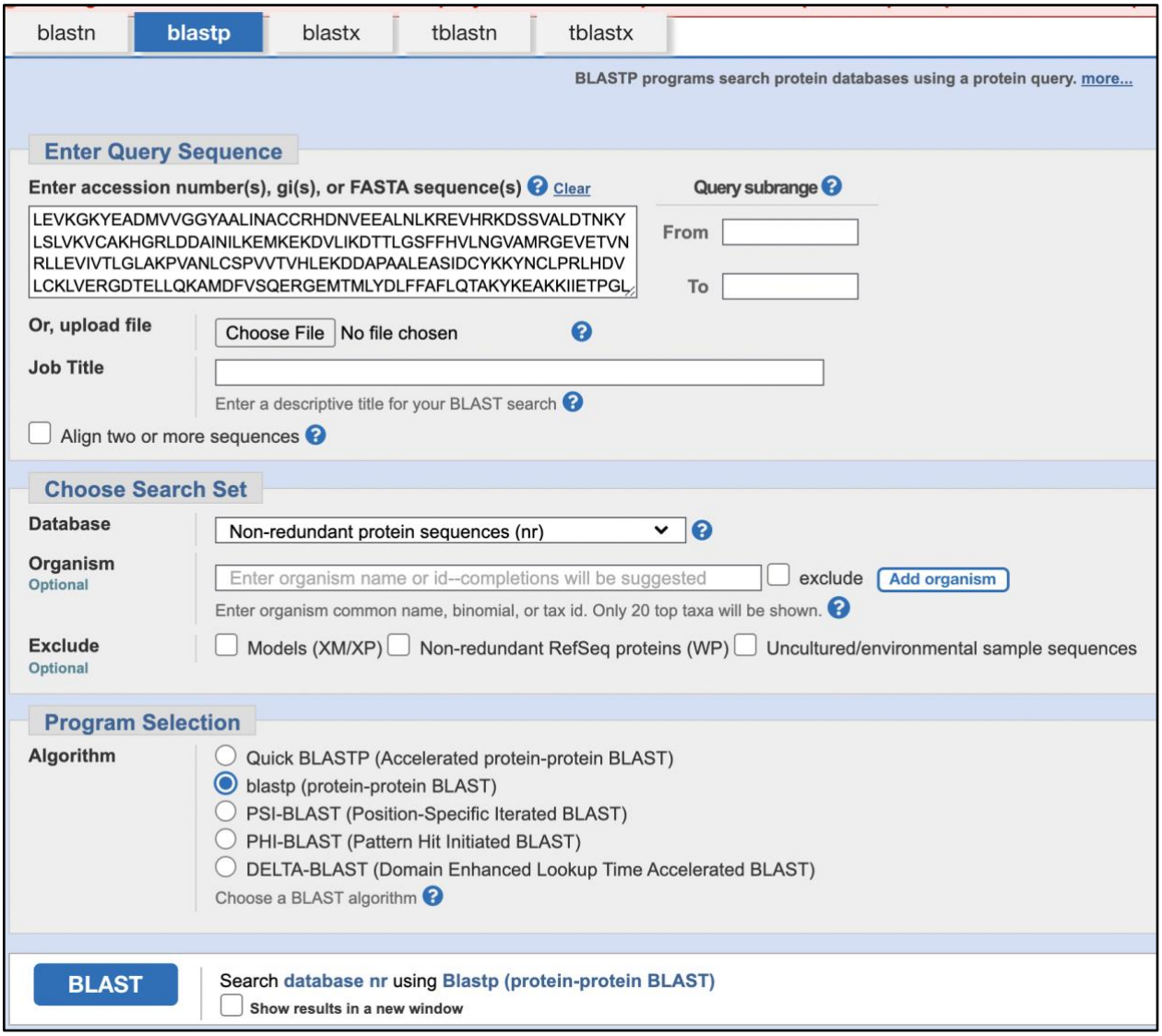
Name: *Cynops pyrrhogaster* LRPPRC
Species: *Cynops pyrrhogaster*

[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 NR database (see setup in first screen-shot below) yielded a top hit result is to a protein from C*ynops orientalis* (Chinese fire belly newt). See additional screen shots below for top hits and selected alignment details:



The top hit result is to a protein from C*ynops orientalis* (Chinese fire belly newt). See additional screen shots below for selected alignment details:

Download GenPept Graphics

leucine-rich PPR motif-containing protein LRPPRC [Cynops orientalis]

Sequence ID: QIS93427.1 **Length:** 1407 **Number of Matches:** 1

Range 1: 705 to 1007 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps
606 bits(1562)	0.0	301/303(99%)	301/303(99%)	0/303(0%)

```
Query  1     LEVKGKYEADMVVGGYAALINACCRHDNVEEALNLKREVHRKDSSVALDTNKYLSLVKVC  60
             LEVKGKYEADMVVGGYAALINACCRHDNVEEALNLKRE  RKDSSVALDTNKYLSLVKVC
Sbjct  705   LEVKGKYEADMVVGGYAALINACCRHDNVEEALNLKREAQRKDSSVALDTNKYLSLVKVC  764

Query  61    AKHGRLDDAINILKEMKEKDVLIKDTTLGSFFHVLNGVAMRGEVETVNRLLEVIVTLGLA  120
             AKHGRLDDAINILKEMKEKDVLIKDTTLGSFFHVLNGVAMRGEVETVNRLLEVIVTLGLA
Sbjct  765   AKHGRLDDAINILKEMKEKDVLIKDTTLGSFFHVLNGVAMRGEVETVNRLLEVIVTLGLA  824

Query  121   KPVANLCSPVVTVHLEKDDAPAALEASIDCYKKYNCLPRLHDVLCKLVERGDTELLQKAM  180
             KPVANLCSPVVTVHLEKDDAPAALEASIDCYKKYNCLPRLHDVLCKLVERGDTELLQKAM
Sbjct  825   KPVANLCSPVVTVHLEKDDAPAALEASIDCYKKYNCLPRLHDVLCKLVERGDTELLQKAM  884

Query  181   DFVSQERGEMTMLYDLFFAFLQTAKYKEAKKIIETPGLRARPGRLQWFAEKCITGNQMET  240
             DFVSQERGEMTMLYDLFFAFLQTAKYKEAKKIIETPGLRARPGRLQWFAEKCITGNQMET
Sbjct  885   DFVSQERGEMTMLYDLFFAFLQTAKYKEAKKIIETPGLRARPGRLQWFAEKCITGNQMET  944

Query  241   LENFVEMTSKLFECDRDEMYFYLLKLCKENNEWQKADAIWTKMQEENLIPRERTLKLLAD  300
             LENFVEMTSKLFECDRDEMYFYLLKLCKENNEWQKADAIWTKMQEENLIPRERTLKLLAD
Sbjct  945   LENFVEMTSKLFECDRDEMYFYLLKLCKENNEWQKADAIWTKMQEENLIPRERTLKLLAD  1004

Query  301   LFK  303
             LFK
Sbjct  1005  LFK  1007
```

Download GenPept Graphics

hypothetical protein KIL84_016365, partial [Mauremys mutica]

Sequence ID: KAH1172526.1 **Length:** 694 **Number of Matches:** 1

Range 1: 277 to 579 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Positives	Gaps
510 bits(1313)	2e-174	247/303(82%)	274/303(90%)	0/303(0%)

```
Query  1     LEVKGKYEADMVVGGYAALINACCRHDNVEEALNLKREVHRKDSSVALDTNKYLSLVKVC  60
             LEVK KYE DMVVGGYAALIN CCRHDNVE+A+NLK EV RKDSSVALDT+KYL+LVKV
Sbjct  277   LEVKAKYEPDMVVGGYAALINLCCRHDNVEDAMNLKEEVFRKDSSVALDTSKYLALVKVL  336

Query  61    AKHGRLDDAINILKEMKEKDVLIKDTTLGSFFHVLNGVAMRGEVETVNRLLEVIVTLGLA  120
              KHGRL+DAINILKEMKEKD+ IKDTT+ SFFH+LN  AMRGEVETVN+L E I+TLGLA
Sbjct  337   GKHGRLEDAINILKEMKEKDIPIKDTTVTSFFHILNAAAMRGEVETVNQLHESILTLGLA  396

Query  121   KPVANLCSPVVTVHLEKDDAPAALEASIDCYKKYNCLPRLHDVLCKLVERGDTELLQKAM  180
             KP ANLCSP++TVHLEKDD PAALEA+IDC KKY  +PRLHDVLC+L+E+G+T+LLQKAM
Sbjct  397   KPSANLCSPLITVHLEKDDVPAALEATIDCCKKYGKIPRLHDVLCRLIEQGNTDLLQKAM  456

Query  181   DFVSQERGEMTMLYDLFFAFLQTAKYKEAKKIIETPGLRARPGRLQWFAEKCITGNQMET  240
             DFVSQERGEMTMLYD+FFAFL T KYKEAKKIIETPGLRARPGRLQWFAEKCI  NQMET
Sbjct  457   DFVSQERGEMTMLYDIFFAFLNTGKYKEAKKIIETPGLRARPGRLQWFAEKCIATNQMET  516

Query  241   LENFVEMTSKLFECDRDEMYFYLLKLCKENNEWQKADAIWTKMQEENLIPRERTLKLLAD  300
             LEN VEMT KLFECDRD+MY+YLLKLCK +N+W+KADA WTKMQEEN+IPRE TL+LLAD
Sbjct  517   LENMVEMTQKLFECDRDQMYYYLLKLCKISNDWRKADATWTKMQEENVIPRETTLRLLAD  576

Query  301   LFK  303
               K
Sbjct  577   TLK  579
```

List of all hits:

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
leucine-rich PPR motif-containing protein LRPPRC [Cynops orientalis]	Cynops ori…	628	628	92%	0.0	99.34%	1407	QIS93427.1
hypothetical protein KIL84_016365 [Mauremys mutica]	Mauremys…	523	523	92%	1e-179	81.52%	694	KAH1172526.1
LOW QUALITY PROTEIN: leucine-rich PPR motif-containing protein, …	Rhinatrem…	538	538	93%	2e-177	82.24%	1407	XP_029449192.1
leucine-rich PPR motif-containing protein, mitochondrial-like [Pelodisc…	Pelodiscu…	523	523	92%	3e-177	79.21%	865	XP_006138247.1
leucine-rich PPR motif-containing protein, mitochondrial [Chelonoidis …	Chelonoidi…	523	523	92%	9e-174	80.53%	1188	XP_032630883.1
leucine-rich PPR motif-containing protein, mitochondrial [Chelonia my…	Chelonia …	527	527	92%	4e-173	81.19%	1399	XP_037751766.1
leucine-rich PPR motif-containing protein, mitochondrial isoform X2 […	Mauremys…	523	523	92%	8e-173	81.19%	1254	XP_039385622.1
leucine-rich PPR motif-containing protein, mitochondrial [Mauremys …	Mauremys…	526	526	92%	1e-172	81.52%	1398	XP_044867041.1
leucine-rich PPR motif-containing protein, mitochondrial [Dermochely…	Dermochel…	525	525	92%	5e-172	80.86%	1415	XP_038251504.2
leucine rich pentatricopeptide repeat containing [Chelydra serpentina]	Chelydra s…	523	523	92%	1e-171	80.20%	1397	KAG6933563.1
leucine-rich PPR motif-containing protein, mitochondrial [Gopherus e…	Gopherus …	523	523	92%	1e-171	80.53%	1399	XP_030413525.1
leucine-rich PPR motif-containing protein, mitochondrial [Terrapene c…	Terrapene …	523	523	92%	1e-171	80.20%	1397	XP_026508293.1
leucine-rich PPR motif-containing protein, mitochondrial isoform X1 […	Mauremys…	522	522	92%	3e-171	81.19%	1398	XP_039385621.1
leucine-rich PPR motif-containing protein, mitochondrial isoform X3 […	Sceloporu…	513	513	92%	1e-167	79.87%	1409	XP_042312263.1
leucine-rich PPR motif-containing protein, mitochondrial isoform X2 […	Sceloporu…	513	513	92%	1e-167	79.87%	1410	XP_042312254.1
leucine-rich PPR motif-containing protein, mitochondrial isoform X1 […	Sceloporu…	513	513	92%	1e-167	79.87%	1411	XP_042312243.1
leucine-rich PPR motif-containing protein, mitochondrial-like [Rhincod…	Rhincodon…	491	542	92%	1e-166	73.93%	719	XP_020390257.1
LRPPRC [Latimeria menadoensis]	Latimeria …	509	509	92%	7e-166	79.54%	1431	AWT24668.1
PREDICTED: leucine-rich PPR motif-containing protein, mitochondria…	Latimeria …	508	508	92%	1e-165	79.21%	1431	XP_005999623.1
PREDICTED: leucine-rich PPR motif-containing protein, mitochondria…	Anolis car…	500	500	92%	3e-165	76.90%	1147	XP_016850993.1
LRPPRC [Protopterus annectens]	Protopteru…	506	506	92%	1e-164	77.56%	1435	AWT24643.1
leucine-rich PPR motif-containing protein, mitochondrial [Trachemys …	Trachemy…	504	504	92%	3e-164	77.89%	1398	XP_034620325.1
leucine-rich PPR motif-containing protein, mitochondrial [Chrysemys …	Chrysemy…	503	503	92%	8e-164	77.89%	1397	XP_005296166.1
leucine-rich PPR motif-containing protein, mitochondrial [Polyodon sp…	Polyodon …	503	503	92%	8e-164	75.58%	1430	XP_041107358.1
PREDICTED: leucine-rich PPR motif-containing protein, mitochondria…	Anolis car…	502	502	92%	2e-163	76.90%	1409	XP_008113921.1
leucine-rich PPR motif-containing protein, mitochondrial isoform X1 […	Mus musc…	480	480	92%	1e-162	73.60%	699	XP_011244945.1
unnamed protein product [Mus musculus]	Mus musc…	481	481	92%	1e-162	73.60%	712	BAB29082.2
leucine-rich PPR motif-containing protein, mitochondrial isoform X2 […	Microcaeci…	498	498	94%	9e-162	77.92%	1407	XP_030051658.1
leucine-rich PPR motif-containing protein, mitochondrial isoform X1 […	Microcaeci…	498	498	94%	9e-162	77.92%	1408	XP_030051657.1
leucine-rich PPR motif-containing protein, mitochondrial [Amblyraja ra…	Amblyraja …	495	557	92%	1e-160	73.93%	1406	XP_032881396.1
leucine-rich PPR motif-containing protein, mitochondrial [Acipenser ru…	Acipenser …	495	495	92%	1e-160	74.92%	1432	XP_034777299.1
leucine-rich PPR motif-containing protein, mitochondrial isoform X2 […	Scyliorhin…	494	494	92%	1e-160	74.59%	1417	XP_038669480.1
leucine-rich PPR motif-containing protein, mitochondrial isoform X1 […	Scyliorhin…	494	494	92%	2e-160	74.59%	1420	XP_038669472.1
putative sodium-coupled neutral amino acid transporter 7 [Platysterno…	Platystern…	493	545	93%	3e-160	77.23%	1395	TFK01649.1
leucine-rich PPR motif-containing protein, mitochondrial [Podarcis mu…	Podarcis …	491	491	92%	2e-159	75.25%	1395	XP_028577588.1
PREDICTED: leucine-rich PPR motif-containing protein, mitochondria…	Lepisoste…	491	491	92%	2e-159	72.94%	1421	XP_015218688.1

[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.

Re-labeled sequences for alignment:

> Homo_sapiens | NP_573566.2

```
MAALLRSARWLLRAGAAPRLPLSLRLLPGGPGRLHAASYLPAARAGPVAGGLLSPARLYAIAAKEKDIQEESTFSSR
KISNQFDWALMRLDLSVRRTGRIPKKLLQKVFNDTCRSGGLGGSHALLLLRSCGSLLPELKLEERTEFAHRIWDTLQ
KLGAVYDVSHYNALLKVYLQNEYKFSPTDFLAKMEEANIQPNRVTYQRLIASYCNVGDIEGASKILGFMKTKDLPVT
EAVFSALVTGHARAGDMENAENILTVMRDAGIEPGPDTYLALLNAYAEKGDIDHVKQTLEKVEKSELHLMDRDLLQI
IFSFSKAGYPQYVSEILEKVTCERRYIPDAMNLILLLVTEKLEDVALQILLACPVSKEDGPSVFGSFFLQHCVTMNT
PVEKLTDYCKKLKEVQMHSFPLQFTLHCALLANKTDLAKALMKAVKEEGFPIRPHYFWPLLVGRRKEKNVQGIIEIL
KGMQELGVHPDQETYTDYVIPCFDSVNSARAILQENGCLSDSDMFSQAGLRSEAANGNLDFVLSFLKSNTLPISLQS
IRSSLLLGFRRSMNINLWSEITELLYKDGRYCQEPRGPTEAVGYFLYNLIDSMSDSEVQAKEEHLRQYFHQLEKMNV
KIPENIYRGIRNLLESYHVPELIKDAHLLVESKNLDFQKTVQLTSSELESTLETLKAENQPIRDVLKQLILVLCSEE
NMQKALELKAKYESDMVTGGYAALINLCCRHDKVEDALNLKEEFDRLDSSAVLDTGKYVGLVRVLAKHGKLQDAINI
LKEMKEKDVLIKDTTALSFFHMLNGAALRGEIETVKQLHEAIVTLGLAEPSTNISFPLVTVHLEKGDLSTALEVAID
CYEKYKVLPRIHDVLCKLVEKGETDLIQKAMDFVSQEQGEMVMLYDLFFAFLQTGNYKEAKKIIETPGIRARSARLQ
WFCDRCVANNQVETLEKLVELTQKLFECDRDQMYYNLLKLYKINGDWQRADAVWNKIQEENVIPREKTLRLLAEILR
EGNQEVPFDVPELWYEDEKHSLNSSSASTTEPDFQKDILIACRLNQKKGAYDIFLNAKEQNIVFNAETYSNLIKLLM
SEDYFTQAMEVKAFAETHIKGFTLNDAANSRLIITQVRRDYLKEAVTTLKTVLDQQQTPSRLAVTRVIQALAMKGDV
ENIEVVQKMLNGLEDSIGLSKMVFINNIALAQIKNNNIDAAIENIENMLTSENKVIEPQYFGLAYLFRKVIEEQLEP
AVEKISIMAERLANQFAIYKPVTDFFLQLVDAGKVDDARALLQRCGAIAEQTPILLLFLLRNSRKQGKASTVKSVLE
LIPELNEKEEAYNSLMKSYVSEKDVTSAKALYEHLTAKNTKLDDLFLKRYASLLKYAGEPVPFIEPPESFEFYAQQL
RKLRENSS
```

> Newt

```
LEVKGKYEADMVVGGYAALINACCRHDNVEEALNLKREVHRKDSSVALDTNKYLSLVKVCAKHGRLDDAINILKEMK
EKDVLIKDTTLGSFFHVLNGVAMRGEVETVNRLLEVIVTLGLAKPVANLCSPVVTVHLEKDDAPAALEASIDCYKKY
NCLPRLHDVLCKLVERGDTELLQKAMDFVSQERGEMTMLYDLFFAFLQTAKYKEAKKIIETPGLRARPGRLQWFAEK
CITGNQMETLENFVEMTSKLFECDRDEMYFYLLKLCKENNEWQKADAIWTKMQEENLIPRERTLKLLADLFKEMVRK
FHLMFLRIGMKKLQHQK
```

> Turtle | XP_006138247

```
MHSVDLFHFKGCLISGFRRSKNVDLWSKITELLYKDGRYCQTPPGPSEAVGYFLYNLIDSMSDSEVQAKEEHLRQYF
HQLKKMNIVIPPNIYRGIRNVLDSYHVPELIKDIMLVVDSEKLSAADIPKNTELEVLSLEEELEKLKAEKQPIGNVL
KQLIVALCAEENMQKALEVKAKYEPDMVVGGYAALINLCCRHDNAEDAMNLKEEVYRKDSSVALDTNKYLALVKVLG
KHGRLEDAINILKEMKEKDVPIRDTTVTSFFHILNAAAMQGEVETVNRLHESILTLGLVKPSTNLCSPLITVHLEKD
DVPAALEAAIDCFKKYGNIPRLHDILCRLIEKGNTDLLQKAMDFVSQERGEMTMLYDLFFAFLNTGKYKEAKKIIET
PGLRARPGRLQWFAEKCIANNQMDILENMVEMTEKLFECDRDQMYYYILQLCKISSDWRRADATWTKMQEENVIPRE
RTLRLLADILKNNGQEVPFDVPEIWYEDCVESAPVSENNPEKKILILCKKGNIQEAYNILLEEQKKDIMFPSFTYSI
LIKALLAEGCLEKAINVKNIAETHIKGFTLNDAASSLLIITQVRRDYLKDAISTLKAVLENDMVPTRLAVTRLIQAL
AMKGDVESVRTVEKMVENLAVSIGLSRMLFVNNTVLAHIKNNNLDTAVEYIESLIISGMQNPDSPITSISYVFRKVI
EEKLESALEKLSAMAERLVNQFGIYKPATDLFLQYVSEGRIDDARLLLQRCGGIAEQKKIMMAFIAKSSQKPGQGQK
IKMLLDLVPDFPEMEVVYSYLLKCHVLDNDIVSAKTLFEKAKAENIRTDEIFLKRLAFLLKSAGEPVPFTEPPESFH
FYVNKLRKERQELSSHDV
```

> Tortoise | XP_032630883

```
MAALLSWACRLRPLCSHLLLRLPRGVAAHRPALRAAGLVGAPGSTHLIYQARLFAIAPHQKGKVQEEPVLSVQNKQA
QHFDWALNKLDSSVRRTGRITRTLLLRIFHDMCRTGYPSSNQALLLLRSCGSLLPEVPLCERTELAHMIWGKLQDLG
AVYDASHYNALLKVYLQNEHKFSPTEFLSKMEEANVQPNRVTYQRLIAAYCNEGDIEGASKILGFMKMKELPITEAV
FSSLVTGHARTGDMENAENIFSVMRDAGIEPGPDTYLALLNAYAEKGDINSVKQILEKIEKMEGYLMDRDLLQVIYS
LAKAGYPQYVPDILSQMRYERGYIPDAMNLSLNLITQGLEDTAFQVFKSFPTLPSENHSETSMDYGNFFLQHCVHMD
KPLSKLKQFCDELKEANMHSSPLQFTLYCALESKKAALAINLMKTMKEEGLPLRPHYSWPLLVGYQKEKNVQGTIGV
```

LKAMHELGVEPDVETYTNYVLTNFDDIQTFRALLQENGCPFETEGLSVAALRHEAIYGTLENVLSLLSSPSMPSVNL
SHFKGCLIFGFKRSNDIDLWSKITELLYKDGRYCQTPPGPAEAVGYFLYNLIDSMSDSEVQAKEEHLRQYFHQLKKM
NVVIPTNIYRGIRNLLAAYHVPELIKDIILVTDREKLSVADIPKNTEVSALEEELEKLKAEKQPIGNVLKQLIVALC
AEENMQKALEVKAKYEPDMVVGGYAALINLCCRHDNVEDAMNLKEEVFRKDSSVALDTNKYLALVKVLGKHGRLEDA
INILKEIKEKDIPIKDTTVTSFFHILNAAAMRGEVETVNRLHESILTLGLAKPSTNLCTPLITVHLEKDDVPSALEA
TIDCYKKYGKMPRLHDILCRLIEQGNTELLQKAMDFVSQERGEMMMLYDIFFAFLNTGKYKEAKKIIETPGLRARPG
SLQWFAEKCIATNQMETLENMVEMTQKLFECDRDEMYYYLLKLCKISNDWRKAEATWTKMQEENVIPRARTLRLLAD
TLKNNGQEVPFDVPEIWYEDHVESASVLENNPEKKVLMLCKKGSVQEAYNILLEAQKKDIMFPASAYSTLIKALLAE
GCLEEALKVKNIAETHIKGFTLNDAASSLLIITQVRRDYLKDAISTLKKVLESDMVPMSLAVIRLIQALAMKGDLES
ICTVEKMVENLATSIGLSRMLFVNNTVLAHIKK

> Zebrafish | NP_001136064 (sequence from BLAST)
MAALLRSARLLKTSSASLIQVIGNSKHAPAAGRLYSGAIGTLRVGVCGRQTVPGRLNANSLFSYRPALPCVSCRQYA
VVPEQSGQVKDEASLAVRSKQAQQFDWALSKLDSSVRRTGRVTKTLLLHIFHDICRTGYPSGNQALLLLRSCGSLLP
EVPLAERTELVKRIWDKLLELGVSYDVSHYNALLKTYLQNEFRFSPTDFLAKMEAANVQPNRVTYQRLIAAYCEEGN
IEGASAILGFMKNKDLPITEAVFNSLVVGHARAGDITSSEGILSVMKSAGIEPGPDTYLSLLNMYAEKGDIDKIKQT
LDVVENADFFLMDRDLMQLVSSLARTGHEQHVPEIVSRMRHERGYVPDAINLCLNLITHGHEKTAFSVLKSLTGMLD
THTGDTPDFGNFFLRHCVNMDKSAEDIVGFCKDLKDLGLHSTPLQFTLQCALEGKKTSLSIGLMKRMKAESLPIKPH
YFLPLFAHHHKDKNIPAIIEVLRGMQEMSVSPDVDAFSFYILPSFPSLDNAKASLKEAGVDVNTDGLIVAELRVQAY
SGNLAKLLSLMSSPALPTIDLSVFRAGLIAGFKRFQNVENMAKITELLYKDARFAADGHDASENVGYFLYNLIDSMS
ETELQSNEEKIREYFGLLKSMNINISVNIFRGIRNILESHHVPELVKEALTLVDKTDDMTEVMMFRSSEGRISALVK
TLAEQKAEGKPAHLTLKKLINVLSIEEKLEQALDLKSKYEDDMTPAAYATLINLCCRHDNAEEALKLKIEMARKDSE
VALDAQKYIALVRVLSKHGKLEEALDILKEMKEKNIMIRDNLIGFLTFTMNSIAMKGDADAIRRLQETIFTLGLAKP
SNGLCSPLVTCYLESGDHAGAFDAVMECHKQYNQLPKIHDLMCSLVEKGDALLLQKVMEFLTLERGEMMMLYDLFFA
FLQTGRNREARKIIETPGLRARSNRLQWFAEKCIAAQQMEPLENLVDMTVKLFECDRDEMYHYLFRLCKETNDWRKA
EAIWMKMQEENLAPRERTLRLLAEILKSNNQEVPFEVPKVWFEDDKGQSEVVPEKEESSAVAKKTDDRSRLNATIRL
KLQSLCKKGEAQEAFDILKEVDSKGIVPGPAIYDAIIKALLAKGNIEDAISVKDIAVGHIPSFILSDVANNLLIISH
VKKCQMKDSVQVLRDMLKADQMPSQLAITRLVQGLADEGNLKDIQEVEAMTKAFGSFNLSNMLFVNNTALALLNNGD
VDSAVDMLQTFYTENTERQNNSIAHVFRKVLNANNDAAMDKLSAMAERLCNQFASYRAATDLFLVYVDTGRTEEAKF
LLQRCAAVGEQKDLLFSYVLRASQQPGQAAKVMSLMELIPDIREKEDIYSQLMKCHGLDQDLASAKALYERMQVEGV
RIDELTLKRLAKLYRDSGEPVPFEEPAESFRFYADKLKDQRTQSTASIDN

> Drosophila_mojavensis | XP_002002110
MASILRTGKLLRYFAGFTRNVVVNSVRDCESNNLLQSAPCMCGQFQNGFASNAAASKAEVTLDRQIRRLDQDARRMG
RISRRDLEEVLDEIRTHRTATSSQSLLVIRCCGNLVPEELPEVRTALVQEIWKTLNALNVPMDISHYNALLRVYLEN
EHQFAPTDFLAEIEAKGIEPNRVTYQRLIARYCQQGDIEGATRILEYMRGKKLPVNENVFNSLILGHSQANDLESAR
GILGVMKQAGLEPTADTYTTLLCAFARHGDIEALQSTLAECEPKEIILLDKDLLDVAYTLAVHGNGEHVDAVLSKLR
ISPGFNQDAVNIILRLVNKGQEDVALKLLRLMPRNSRVNGEPVDVGAFFIRQLVKANRPVEKILSICRTLQSEGLNP
KALTIATEAGLTNGVMNNALPLLHEMKNLGLPIRQHYFWPLFCSVDSNQVLDVVRRMQQDFALNPNSETVRDYVIPN
LKEKNWDRVVTVLRDAGVPNSTAVSSAVYAALTTHHIADAAKIMEQNRAYYMPLIFRQPLILALAHTNDFASFIRCV
RQVYEGLQLRAGKEATEEAAEAATDGEAPATPERQTPDVVGQIVYDAAIYFRRERVPTLEKILQGLVNQGLSLSSGK
ATALSELLGSEMTPKIAEQLGKLTSGELEPIPLPNSGKRSLDTLTIDELERFIKNVEAKGENSNNIKRQLLNACFRS
QNLEKTLQVIERLEKDTFQIPIGIYAQLVDLYTHHKRTSDALAQYNKLRTADASFKLDNFKTVRLADLLLQEERVDE
ALQLLKENQKEAAVEAGEGSFNYVSTVWRILNSIAETGNAERLKVVFDALVAGNYVVPTNVLLGPLIKVHLSRDDIP
KAIDAFEQICQQYKATPWKNELACRLIQKEDAANLQRLTDLSTSIHGEVNSLYDLVFSFVECGRVRQARKILETPGL
RTRPQRISNACDRYKNEGMLQPLEGLIEATKDLGHIDRNKIYYTLLLSYDKADETEKALGLWTKMQEEGVTPTDAFL
LKLAELLKRKNIDVPFVVPETQQPKSKRNKPAKTEANVAEQVVQSPPEPKVEPKAKAVPAAKPISNIAGFRRAIQAN
DPDAAISFKERVLSGDKFNVLDTSRLIELLVRADRLSEATKYVEELLAEKQHPQPKIFKFYLNKIAASGDLEVMQRI
GQQLNDEQKRLVSFDNRYCHAYIVAGKAEQFLKQLSTEIEAVKSSEEAGKLAEKFPRGGAVGILEKHPELIAQYETL
AEKYAAHNQLGPMNVLWMHLISNGQEVASKQIWDKHLSSAPRLMFQRVLQTAREQQDEKLASTVISQLRGSKISEGA
IGNAYSCLIDIQTTKGNTDKALEVLANAIKDVSLENINRTALLRLKQAVEEKSQQFPYTVPEKRAKAEDSSSSSSSS
SDDDVTPKRPETAPTRPERV

Alignment:
CLUSTAL multiple sequence alignment by MUSCLE (3.8)

```
Drosophila_mojavensis      MASILRTGKLLRYFA-------GFTRNVVVNSVRDCESNNLLQSAPCMCGQFQNGFASNA
Zebrafish                  MAALLRSARLLKTSSASLIQVIGNSKHAPAAGRLYSGAIGTLRVGVCGRQTVPGRLNANS
Homo_sapiens               MAALLRSARWLLRA--------GAAPRLPLSLRLLPGGPGRLHAASYLPAARAGPVAGGL
Newt                       ------------------------------------------------------------
Tortoise                   MAALLSWACRLR-------------PLCSHLLLRLPRGVAAHRPALRAAGLVGAPGSTHL
Turtle                     ------------------------------------------------------------

Drosophila_mojavensis      AA---------------------------------SKAEVTLDRQIRRLDQDARRMGRIS
Zebrafish                  LFSYRPALPCVSCRQYAVVPEQSGQVKDEASLAVRSKQAQQFDWALSKLDSSVRRTGRVT
Homo_sapiens               LS---------PARLYAIAAKEKDIQEESTFSS--RKISNQFDWALMRLDLSVRRTGRIP
Newt                       ------------------------------------------------------------
Tortoise                   IY---------QARLFAIAPHQKGKVQEEPVLSVQNKQAQHFDWALNKLDSSVRRTGRIT
Turtle                     ------------------------------------------------------------

Drosophila_mojavensis      RRDLEEVLDEIRTHRTATSSQSLLVIRCCGNLVPEELPEVRTALVQEIWKTLNALNVPMD
Zebrafish                  KTLLLHIFHDICRTGYPSGNQALLLLRSCGSLLPEVPLAERTELVKRIWDKLLELGVSYD
Homo_sapiens               KKLLQKVFNDTCRSGGLGGSHALLLLRSCGSLLPELKLEERTEFAHRIWDTLQKLGAVYD
Newt                       ------------------------------------------------------------
Tortoise                   RTLLLRIFHDMCRTGYPSSNQALLLLRSCGSLLPEVPLCERTELAHMIWGKLQDLGAVYD
Turtle                     ------------------------------------------------------------

Drosophila_mojavensis      ISHYNALLRVYLENEHQFAPTDFLAEIEAKGIEPNRVTYQRLIARYCQQGDIEGATRILE
Zebrafish                  VSHYNALLKTYLQNEFRFSPTDFLAKMEAANVQPNRVTYQRLIAAYCEEGNIEGASAILG
Homo_sapiens               VSHYNALLKVYLQNEYKFSPTDFLAKMEEANIQPNRVTYQRLIASYCNVGDIEGASKILG
Newt                       ------------------------------------------------------------
Tortoise                   ASHYNALLKVYLQNEHKFSPTEFLSKMEEANVQPNRVTYQRLIAAYCNEGDIEGASKILG
Turtle                     ------------------------------------------------------------

Drosophila_mojavensis      YMRGKKLPVNENVFNSLILGHSQANDLESARGILGVMKQAGLEPTADTYTTLLCAFARHG
Zebrafish                  FMKNKDLPITEAVFNSLVVGHARAGDITSSEGILSVMKSAGIEPGPDTYLSLLNMYAEKG
Homo_sapiens               FMKTKDLPVTEAVFSALVTGHARAGDMENAENILTVMRDAGIEPGPDTYLALLNAYAEKG
Newt                       ------------------------------------------------------------
Tortoise                   FMKMKELPITEAVFSSLVTGHARTGDMENAENIFSVMRDAGIEPGPDTYLALLNAYAEKG
Turtle                     ------------------------------------------------------------

Drosophila_mojavensis      DIEALQSTLAECEPKEIILLDKDLLDVAYTLAVHGNGEHVDAVLSKLRISPGFNQDAVNI
Zebrafish                  DIDKIKQTLDVVENADFFLMDRDLMQLVSSLARTGHEQHVPEIVSRMRHERGYVPDAINL
Homo_sapiens               DIDHVKQTLEKVEKSELHLMDRDLLQIIFSFSKAGYPQYVSEILEKVTCERRYIPDAMNL
Newt                       ------------------------------------------------------------
Tortoise                   DINSVKQILEKIEKMEGYLMDRDLLQVIYSLAKAGYPQYVPDILSQMRYERGYIPDAMNL
Turtle                     ------------------------------------------------------------

Drosophila_mojavensis      ILRLVNKGQEDVALKLLRLMP--RNSRVNGEPVDVGAFFIRQLVKANRPVEKILSICRTL
Zebrafish                  CLNLITHGHEKTAFSVLKSL-TGMLDTHTGDTPDFGNFFLRHCVNMDKSAEDIVGFCKDL
Homo_sapiens               ILLLVTEKLEDVALQILLACPVSKED----GPSVFGSFFLQHCVTMNTPVEKLTDYCKKL
Newt                       ------------------------------------------------------------
Tortoise                   SLNLITQGLEDTAFQVFKSFPTLPSENHSETSMDYGNFFLQHCVHMDKPLSKLKQFCDEL
Turtle                     ------------------------------------------------------------

Drosophila_mojavensis      QSEGLNPKALTIATEAGLTNGVMNNALPLLHEMKNLGLPIRQHYFWPLFCSVDSNQ----
Zebrafish                  KDLGLHSTPLQFTLQCALEGKKTSLSIGLMKRMKAESLPIKPHYFLPLFAHHHKDKNIPA
Homo_sapiens               KEVQMHSFPLQFTLHCALLANKTDLAKALMKAVKEEGFPIRPHYFWPLLVGRRKEKNVQG
Newt                       ------------------------------------------------------------
Tortoise                   KEANMHSSPLQFTLYCALESKKAALAINLMKTMKEEGLPLRPHYSWPLLVGYQKEKNVQG
Turtle                     ------------------------------------------------------------
```

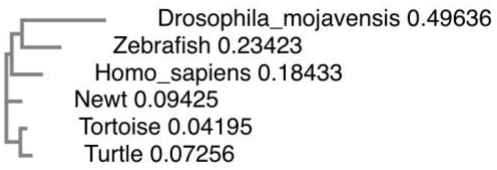
```
Drosophila_mojavensis      VLDVVRRMQQDFALNPNSETVRDYVIPNLKEKNWDRVVTVLRDAGVPNSTAVSSAVYAAL
Zebrafish                  IIEVLRGM-QEMSVSPDVDAFSFYILPSFPSLDNAKASLKEAGVDVNTDGLIVAELRVQA
Homo_sapiens               IIEILKGM-QELGVHPDQETYTDYVIPCFDSVNSARAILQENGCLSDSDMFSQAGLRSEA
Newt                       ------------------------------------------------------------
Tortoise                   TIGVLKAM-HELGVEPDVETYTNYVLTNFDDIQTFRALLQENGCPFETEGLSVAALRHEA
Turtle                     ------------------------------------------------------------

Drosophila_mojavensis      TTHHIADAAKIMEQNRAYYMPL-IFRQPLILALAHTNDFASFIRCVRQVYEGLQLRAGKE
Zebrafish                  YSGNLAKLLSLMSSPALPTIDLSVFRAGLIAGFKRFQNVENMAKITELLYK---------
Homo_sapiens               ANGNLDFVLSFLKSNTLP-ISLQSIRSSLLLGFRRSMNINLWSEITELLYK---------
Newt                       ------------------------------------------------------------
Tortoise                   IYGTLENVLSLLSSPSMPSVNLSHFKGCLIFGFKRSNDIDLWSKITELLYK---------
Turtle                     ----------------MHSVDLFHFKGCLISGFRRSKNVDLWSKITELLYK---------

Drosophila_mojavensis      ATEEAAEAATDGEAPATPERQTPDVVGQIVYDAAIYFRRERVPTLEKILQGLVNQ----G
Zebrafish                  ----------DARFAADG-HDASENVGYFLYNLIDSMSETELQSNEEKIREYFGLLKSMN
Homo_sapiens               ----------DGRYCQEP-RGPTEAVGYFLYNLIDSMSDSEVQAKEEHLRQYFHQLEKMN
Newt                       ------------------------------------------------------------
Tortoise                   ----------DGRYCQTP-PGPAEAVGYFLYNLIDSMSDSEVQAKEEHLRQYFHQLKKMN
Turtle                     ----------DGRYCQTP-PGPSEAVGYFLYNLIDSMSDSEVQAKEEHLRQYFHQLKKMN

Drosophila_mojavensis      LSLSSGKATALSELLGSEMTPKIAEQLGKLTSG--ELEPIPLPNSGKRSLDTLTIDELER
Zebrafish                  INISVNIFRGIRNILESHHVPELVKEALTLVDKTDDMTEVMMFRSSE-----GRISALVK
Homo_sapiens               VKIPENIYRGIRNLLESYHVPELIKDAHLLVES----KNLDFQKTVQ-----LTSSELES
Newt                       ------------------------------------------------------------
Tortoise                   VVIPTNIYRGIRNLLAAYHVPELIKDIILVTDR-EKLSVADIPKNTE-------VSALEE
Turtle                     IVIPPNIYRGIRNVLDSYHVPELIKDIMLVVDS-EKLSAADIPKNTE-----LEVLSLEE

Drosophila_mojavensis      FIKNVEAKGENSNNIKRQLLNACFRSQNLEKTLQVIERLEKDTFQIPIGIYAQLVDLYTH
Zebrafish                  TLAEQKAEGKPAHLTLKKLINVLSIEEKLEQALDLKSKYEDD---MTPAAYATLINLCCR
Homo_sapiens               TLETLKAENQPIRDVLKQLILVLCSEENMQKALELKAKYESD---MVTGGYAALINLCCR
Newt                       -------------------------------LEVKGKYEAD---MVVGGYAALINACCR
Tortoise                   ELEKLKAEKQPIGNVLKQLIVALCAEENMQKALEVKAKYEPD---MVVGGYAALINLCCR
Turtle                     ELEKLKAEKQPIGNVLKQLIVALCAEENMQKALEVKAKYEPD---MVVGGYAALINLCCR
                                                    *::  . * *   :  . ** *::   .

Drosophila_mojavensis      HKRTSDALAQYNKLRTADASFKLDNFKTVRLADLLLQEERVDEALQLLKE-NQKEAAVEA
Zebrafish                  HDNAEEALKLKIEMARKDSEVALDAQKYIALVRVLSKHGKLEEALDILKEMKEKNIMIRD
Homo_sapiens               HDKVEDALNLKEEFDRLDSSAVLDTGKYVGLVRVLAKHGKLQDAINILKEMKEKDVLIKD
Newt                       HDNVEEALNLKREVHRKDSSVALDTNKYLSLVKVCAKHGRLDDAINILKEMKEKDVLIKD
Tortoise                   HDNVEDAMNLKEEVFRKDSSVALDTNKYLALVKVLGKHGRLEDAINILKEIKEKDIPIKD
Turtle                     HDNAEDAMNLKEEVYRKDSSVALDTNKYLALVKVLGKHGRLEDAINILKEMKEKDVPIRD
                           *....:*:    :.   *:.  **  * : *. :  :  .:::*:::*** ::*:  :

Drosophila_mojavensis      GEGSFNYVSTVWRILNSIAETGNAERLKVVFDALVAGNYVVPTNVLLGPLIKVHLSRDDI
Zebrafish                  -----NLIGFLTFTMNSIAMKGDADAIRRLQETIFTLGLAKPSNGLCSPLVTCYLESGDH
Homo_sapiens               -----TTALSFFHMLNGAALRGEIETVKQLHEAIVTLGLAEPSTNISFPLVTVHLEKGDL
Newt                       -----TTLGSFFHVLNGVAMRGEVETVNRLLEVIVTLGLAKPVANLCSPVVTVHLEKDDA
Tortoise                   -----TTVTSFFHILNAAAMRGEVETVNRLHESILTLGLAKPSTNLCTPLITVHLEKDDV
Turtle                     -----TTVTSFFHILNAAAMQGEVETVNRLHESILTLGLVKPSTNLCSPLITVHLEKDDV
                                .    .   :*. *  *: : :. : : :.: . . *   :  *::. :*. .*

Drosophila_mojavensis      PKAIDAFEQICQQYKATPWKNELACRLIQKEDAANLQRLTDLSTSIHGEVNSLYDLVFSF
Zebrafish                  AGAFDAVMECHKQYNQLPKIHDLMCSLVEKGDALLLQKVMEFLTLERGEMMMLYDLFFAF
Homo_sapiens               STALEVAIDCYEKYKVLPRIHDVLCKLVEKGETDLIQKAMDFVSQEQGEMVMLYDLFFAF
Newt                       PAALEASIDCYKKYNCLPRLHDVLCKLVERGDTELLQKAMDFVSQERGEMTMLYDLFFAF
Tortoise                   PSALEATIDCYKKYGKMPRLHDILCRLIEQGNTELLQKAMDFVSQERGEMMMLYDIFFAF
Turtle                     PAALEAAIDCFKKYGNIPRLHDILCRLIEKGNTDLLQKAMDFVSQERGEMTMLYDLFFAF
                           . *::.  :  ::*   *. ::: * *::. ::  :*.  :: :  .**:  ***:.*:*
```

```
Drosophila_mojavensis      VECGRVRQARKILETPGLRTRPQRISNACDRYKNEGMLQPLEGLIEATKDLGHIDRNKIY
Zebrafish                  LQTGRNREARKIIETPGLRARSNRLQWFAEKCIAAQQMEPLENLVDMTVKLFECDRDEMY
Homo_sapiens               LQTGNYKEAKKIIETPGIRARSARLQWFCDRCVANNQVETLEKLVELTQKLFECDRDQMY
Newt                       LQTAKYKEAKKIIETPGLRARPGRLQWFAEKCITGNQMETLENFVEMTSKLFECDRDEMY
Tortoise                   LNTGKYKEAKKIIETPGLRARPGSLQWFAEKCIATNQMETLENMVEMTQKLFECDRDEMY
Turtle                     LNTGKYKEAKKIIETPGLRARPGRLQWFAEKCIANNQMDILENMVEMTEKLFECDRDQMY
                           :: .. .:*.**:****:*:*.  :.. .:.      :: ** ::: * .*   **:::*

Drosophila_mojavensis      YTLLLSYDKADETEKALGLWTKMQEEGVTPTDAFLLKLAELLKRKNIDVPFVVPET----
Zebrafish                  HYLFRLCKETNDWRKAEAIWMKMQEENLAPRERTLRLLAEILKSNNQEVPFEVPKVWFED
Homo_sapiens               YNLLKLYKINGDWQRADAVWNKIQEENVIPREKTLRLLAEILREGNQEVPFDVPELWYED
Newt                       FYLLKLCKENNEWQKADAIWTKMQEENLIPRERTLKLLADLFKE----------------
Tortoise                   YYLLKLCKISNDWRKAEATWTKMQEENVIPRARTLRLLADTLKNNGQEVPFDVPEIWYED
Turtle                     YYILQLCKISSDWRRADATWTKMQEENVIPRERTLRLLADILKNNGQEVPFDVPEIWYED
                           . ::   .  .:  .* . * *:***.: *    *  **: :.

Drosophila_mojavensis      ------------QQPKSKRNKPAK-----------------TEANVAEQVVQSPPEPKVE
Zebrafish                  DKGQSEVVPEKEESSAVAKKTDDRSRLNATIRLKLQSLCKKGEAQEAFDILKEVDSKGIV
Homo_sapiens               EKHSLNS-----SSASTTEPDFQK---------DILIACRLNQKKGAYDIFLNAKEQNIV
Newt                       ------------------------------------------------------------
Tortoise                   -----HV-----ESASVLENNPEK---------KVLMLCKKGSVQEAYNILLEAQKKDIM
Turtle                     -----CV-----ESAPVSENNPEK---------KILILCKKGNIQEAYNILLEEQKKDIM


Drosophila_mojavensis      PKAKAVPA-AKPISNIAGFRRAIQANDPDAAISFKERVLSGDKFNVLDTSRLIELLVRAD
Zebrafish                  PGPAIYDAIIKALLAKGNIEDAISVKD------IAVGHIPSFILSDVANNLLIISHVKKC
Homo_sapiens               FNAETYSNLIKLLMSEDYFTQAMEVKA------FAETHIKGFTLNDAANSRLIITQVRRD
Newt                       ------------------------------------MVRKFHL----------------
Tortoise                   FPASAYSTLIKALLAEGCLEEALKVKN------IAETHIKGFTLNDAASSLLIITQVRRD
Turtle                     FPSFTYSILIKALLAEGCLEKAINVKN------IAETHIKGFTLNDAASSLLIITQVRRD
                                                           :    :

Drosophila_mojavensis      RLSEATKYVEELLAEKQHPQPKIFKFYLNKIAASGDLEVMQRIGQQLNDEQKRL----VS
Zebrafish                  QMKDSVQVLRDMLKADQMPSQLAITRLVQGLADEGNLKDIQEVEAMTK-AFGSFNLSNML
Homo_sapiens               YLKEAVTTLKTVLDQQQTPSRLAVTRVIQALAMKGDVENIEVVQKMLNGLEDSIGLSKMV
Newt                       ------------------------------------------------MFLRIGMKK--
Tortoise                   YLKDAISTLKKVLESDMVPMSLAVIRLIQALAMKGDLESICTVEKMVENLATSIGLSRML
Turtle                     YLKDAISTLKAVLENDMVPTRLAVTRLIQALAMKGDVESVRTVEKMVENLAVSIGLSRML
                                                                              :

Drosophila_mojavensis      FDNRYCHAYIVAGKAEQFLKQLSTEIEAVKSSEEAG--KLAEKFPRGGAVGILEKHPELI
Zebrafish                  FVNNTALALLNNGDVDSAVDMLQTFYT---ENTERQNNSIAHVFRKVLNANNDAAMDKLS
Homo_sapiens               FINNIALAQIKNNNIDAAIENIENMLTSENKVIEPQYFGLAYLFRKVIEEQLEPAVEKIS
Newt                       ------LQHQK----------------------------------------------------
Tortoise                   FVNNTVLAHIKK---------------------------------------------------
Turtle                     FVNNTVLAHIKNNNLDTAVEYIESLIISGMQNPDSPITSISYVFRKVIEEKLESALEKLS


Drosophila_mojavensis      AQYETLAEKYAAHNQLGPMNVLWMHLISNGQEVASKQIWDK--HLSSAPRLMFQRVLQTA
Zebrafish                  AMAERLCNQFASYR---AATDLFLVYVDTGRTEEAKFLLQRCAAVGEQKDLLFSYVLRAS
Homo_sapiens               IMAERLANQFAIYK---PVTDFFLQLVDAGKVDDARALLQRCGAIAEQTPILLLFLLRNS
Newt                       ------------------------------------------------------------
Tortoise                   ------------------------------------------------------------
Turtle                     AMAERLVNQFGIYK---PATDLFLQYVSEGRIDDARLLLQRCGGIAEQKKIMMAFIAKSS


Drosophila_mojavensis      REQQDEKLASTVISQLRGSKISEGAIGNAYSCLIDIQTTKGNTDKALEVLANAIKDVSLE
Zebrafish                  QQPGQAAKVMSLMELIPDIREKE----DIYSQLMKCHGLDQDLASAKALYERMQVEGV--
Homo_sapiens               RKQGKASTVKSVLELIPELNEKE----EAYNSLMKSYVSEKDVTSAKALYEHLTAKNT--
Newt                       ------------------------------------------------------------
Tortoise                   ------------------------------------------------------------
Turtle                     QKPGQGQKIKMLLDLVPDFPEME----VVYSYLLKCHVLDNDIVSAKTLFEKAKAENI--
```

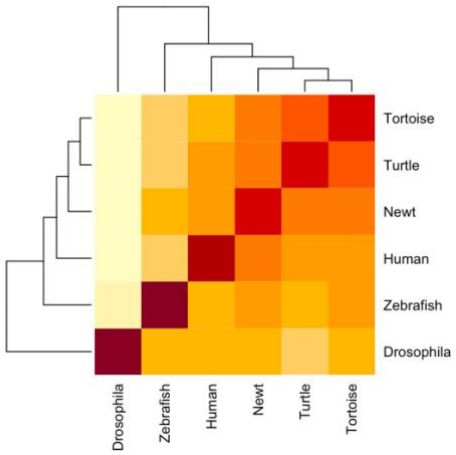
```
Drosophila_mojavensis      NINRTALLRLKQAVEEKSQQFPYTVPEKRAKAEDSSSSSSSSSDDDVTPKRPETAPTRPE
Zebrafish                  RIDELTLKRLAKLYRDSGEPVPFEEPAE----------SFRFYADKLKDQRTQSTASIDN
Homo_sapiens               KLDDLFLKRYASLLKYAGEPVPFIEPPE----------SFEFYAQQLRKLRENSS-----
Newt                       ------------------------------------------------------------
Tortoise                   ------------------------------------------------------------
Turtle                     RTDEIFLKRLAFLLKSAGEPVPFTEPPE----------SFHFYVNKLRKERQELSSHDV-

Drosophila_mojavensis      RV
Zebrafish                  --
Homo_sapiens               --
Newt                       --
Tortoise                   --
Turtle                     --
```

[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.

Obtained from EBI Simple Phylogeny website:



[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.

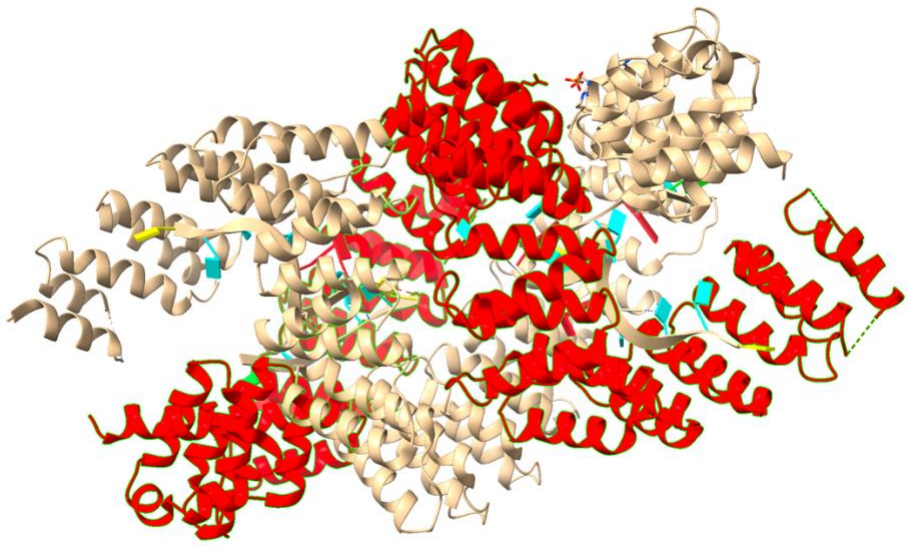


[Q8] Using R/Bio3D (or an online blast server if you prefer), search the main protein structure database for the most similar atomic resolution structures to your aligned sequences.
List the top 3 unique hits (i.e. not hits representing different chains from the same structure) along with their Evalue 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	Evalue	Identity
6EEN	X-RAY DIFFRACTION	2.01	Zea mays	2.12e-14	25.338
5I9H	X-RAY DIFFRACTION	2.50	Escherichia coli BL21(DE3)	8.80e-13	22.321
4M59	X-RAY DIFFRACTION	2.46	Zea mays	2.53e-11	25.188

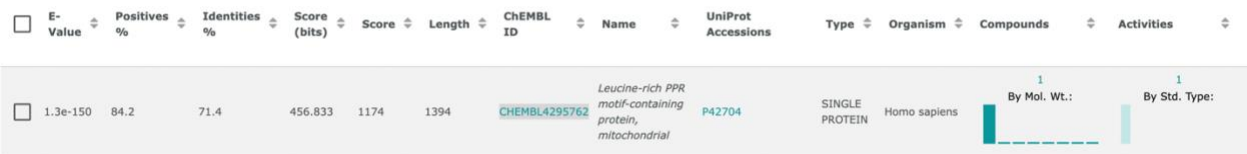
[Q9] Generate a molecular figure of one of your identified PDB structures using VMD. You can optionally highlight conserved residues that are likely to be functional. Please use a white or transparent background for your figure (i.e. not the default black).
Based on sequence similarity. How likely is this structure to be similar to your “novel” protein?

It is unlikely to be similar in structure to *Zea mays* LRPPRC protein given the low sequence similarity (~25%). In the figure below LRPPRC chain A is colored red.

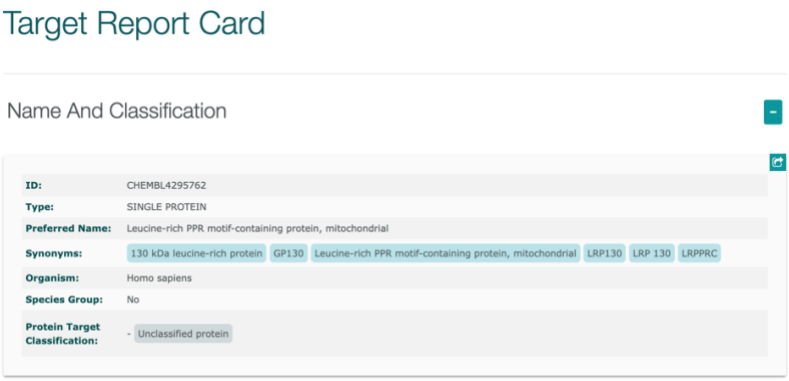


[Q10] Perform a “Target” search of ChEMBEL (https://www.ebi.ac.uk/chembl/) with

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?



E-Value	Positives %	Identities %	Score (bits)	Score	Length	ChEMBL ID	Name	UniProt Accessions	Type	Organism	Compounds	Activities
1.3e-150	84.2	71.4	456.833	1174	1394	CHEMBL4295762	*Leucine-rich PPR motif-containing protein, mitochondrial*	P42704	SINGLE PROTEIN	Homo sapiens	1 By Mol. Wt.:	1 By Std. Type:



ChEMBL ID: CHEMBL4295762 (https://www.ebi.ac.uk/chembl/target_report_card/CHEMBL4295762/)

No ligand efficiency data.

Binding assay linked to a small molecule compound (CHEMBL4129274) that was showed to act as a multi-targeting TAK1-centered inhibitors for cancer and other diseases.

DOI: http://dx.doi.org/10.1016%2Fj.bmc.2016.11.034