

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 its 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)

BLAST® » tblastn

Translated BLAST: tblastn

blastn blastp blastx **tblastn** tblastx

TBLASTN search translated nucleotide databases using a protein query. [more...](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)

NP_573566.2

Query subrange [?](#)

From

To

Or, upload file No file chosen [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

☐ Align two or more sequences [?](#)

Choose Search Set

Database [?](#)

Organism ☐ exclude

Optional Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown [?](#)

Exclude ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Optional Limit to ☐ Sequences from type material

Optional Entrez Query

Optional Enter an Entrez query to limit search [?](#)

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

☐ Show results in a new window

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.

BLAST ® » tblastn » results for RID-ZF6YBBDE016 [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](#)

i Your search is limited to records that include: *Cynops pyrrhogaster* (taxid:8330)

Job Title	NP_573566:leucine-rich PPR motif-containing...	Filter Results
RID	ZF6YBBDE016 <small>Search expires on 02-01 15:20 pm</small> Download All ▼	Organism <small>only top 20 will appear</small> <input type="checkbox"/> exclude Type common name, binomial, taxid or group name + Add organism
Program	TBLASTN Citation ▼	Percent Identity <input type="text"/> to <input type="text"/> E value <input type="text"/> to <input type="text"/> Query Coverage <input type="text"/> to <input type="text"/> Filter Reset
Database	est See details ▼	
Query ID	NP_573566.2	
Description	leucine-rich PPR motif-containing protein, mitochondri...	
Molecule type	amino acid	
Query Length	1394	
Other reports	?	

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

Sequences producing significant alignments [Download](#) [New](#) [Select columns](#) [Show](#) 100 [?](#)

☒ select all 4 sequences selected [GenBank](#) [Graphics](#)

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	FS293294 Cp_al Cynops pyrrhogaster cDNA clone Cp_al_009_G17 3' mRNA sequence	Cynops pyrrhogaster	468	468	21%	4e-154	71.38%	985	FS293294.1
<input checked="" type="checkbox"/>	FS299916 Cp_al Cynops pyrrhogaster cDNA clone Cp_al_027_E08 3' mRNA sequence	Cynops pyrrhogaster	382	430	22%	5e-122	67.53%	1022	FS299916.1
<input checked="" type="checkbox"/>	FS300508 Cp_al Cynops pyrrhogaster cDNA clone Cp_al_028_N20 3' mRNA sequence	Cynops pyrrhogaster	215	215	12%	1e-62	59.65%	896	FS300508.1
<input checked="" type="checkbox"/>	FS293338 Cp_al Cynops pyrrhogaster cDNA clone Cp_al_009_I13 3' mRNA sequence	Cynops pyrrhogaster	85.1	85.1	6%	3e-18	52.27%	708	FS293338.1

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
Sbjct 1	LE+K KYE+DMV GGYAALIN CCRHD VE+ALNLK E R DSS LDT KY+ LV+V					180
Query 759	LEVKGKYEADMVVGGYAALINACCRHDNVEEALNLKREVHRKDSVALDTNKYLSLVKVC					180
Sbjct 181	LEVKGKYEADMVVGGYAALINACCRHDNVEEALNLKREVHRKDSVALDTNKYLSLVKVC					360
Query 819	AKHGKLQDAINILKEMKEKDVLIKDTTALSFFHMLNGAALRGEIETVKQLHEAIVTLGLA					818
Sbjct 181	AKHG+L DAINILKEMKEKDVLIKDTT SFFH+LNG A+RGE+ETV +L E IVTLGLA					360
Query 819	AKHGRLLDDAINILKEMKEKDVLIKDTTALSGFFHVLNGVAMRGEVETVNRLLEIVITLGLA					360
Sbjct 181	AKHGRLLDDAINILKEMKEKDVLIKDTTALSGFFHVLNGVAMRGEVETVNRLLEIVITLGLA					360
Query 819	EPSTNISFPLVTVHLEKGDLSALEVAIDCYEKYKVLPRHDLVCKLVEKGETDLIQKAM					878
Sbjct 361	+P N+ P+VTVHLEK D ALE +IDCY+KY LPR+HDVLCCLVE+G+T+L+QKAM					540
Query 879	KPVANLCSPPVTVHLEKDDAPAALASIDCYKKYNCLPRLHDLVCKLVERGDTELLQKAM					540
Sbjct 361	KPVANLCSPPVTVHLEKDDAPAALASIDCYKKYNCLPRLHDLVCKLVERGDTELLQKAM					540
Query 879	DFVSQEQGEMVMLYDLFFAFLQTGNYKEAKKIIETPGIRARSARLQWFCDCRVANNQVET					938
Sbjct 541	DFVSQE+GEM MLYDLFFAFLQT YKEAKKIIETPG+RAR RLQWF ++C+ NQ+ET					720
Query 939	DFVSQERGEMTMYDLFFAFLQTAKYKEAKKIIETPGLRARPGRQLQWFAEKCITGNQMET					720
Sbjct 541	DFVSQERGEMTMYDLFFAFLQTAKYKEAKKIIETPGLRARPGRQLQWFAEKCITGNQMET					720
Query 939	LEKLVLTQKLFECDRDQMYNLLKLYKINGDWQRADAVWNKIQEENVIPREKTLRLLA					998
Sbjct 721	LE VE+T KLFECDRD+MY+ LLKL K N +WQ+ADA+W K+QEEN+IPRE+TL+LLA+					900
Query 999	LENFVEMTSKLFECDRDEMYFYLLKLCKENNEWQKADAIWTKMQEENLIPRERTLKLLAD					900
Sbjct 721	LENFVEMTSKLFECDRDEMYFYLLKLCKENNEWQKADAIWTKMQEENLIPRERTLKLLAD					900
Query 999	ILRE 1002					
Sbjct 901	+ +E					
	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)
LEVKGKYEADMVVGGYAALINACCRHDNVEEALNLKREVHRKDSVALDTNKYLSLVKVC
AKHGRLLDDAINILKEMKEKDVLIKDTTALSGFFHVLNGVAMRGEVETVNRLLEIVITLGLA
KPVANLCSPPVTVHLEKDDAPAALASIDCYKKYNCLPRLHDLVCKLVERGDTELLQKAM
DFVSQERGEMTMYDLFFAFLQTAKYKEAKKIIETPGLRARPGRQLQWFAEKCITGNQMET
LENFVEMTSKLFECDRDEMYFYLLKLCKENNEWQKADAIWTKMQEENLIPRERTLKLLAD
LFKEMVRKFHLMFLRIGMKKLQHQK*
```

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 *Cynops orientalis* (Chinese fire belly newt). See additional screen shots below for top hits and selected alignment details:

The screenshot shows the NCBI BLAST search interface. At the top, there are tabs for different BLAST programs: blastn, **blastp**, blastx, tblastn, and tblastx. Below the tabs, a message states: "BLASTP programs search protein databases using a protein query. [more...](#)".

The main section is titled "Enter Query Sequence". It contains a text area for the query sequence, which is populated with a FASTA sequence: LEVKGKYEADMVGGYAALINACCRHDNVEEALNLKREVRKRDSSVALDTNKY LSLVKVCAKHGRLLDDAINILKEMKEKDVLIKDTTLGSFFHVLNGVAMRGEVETVN RLLEIVITLGLAKPVANLCSPPVTVHLEKDDAPAALEASIDCYKKYNCLPRLHDV LCKLVERGDTTELLQKAMDFVSQERGETMPLYDLFFAFLQAKYKEAKKIETPGL. To the right of the text area is a "Query subrange" section with "From" and "To" input fields. Below the text area is a section for "Or, upload file" with a "Choose File" button and a "No file chosen" status. Below that is a "Job Title" input field with a placeholder text: "Enter a descriptive title for your BLAST search". There is also a checkbox for "Align two or more sequences".

The next section is titled "Choose Search Set". It contains a "Database" dropdown menu set to "Non-redundant protein sequences (nr)". Below this is an "Organism" section with a text input field for "Enter organism name or id--completions will be suggested", an "exclude" checkbox, and an "Add organism" button. Below the organism input field is a text input field for "Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown." There are also checkboxes for "Exclude" options: "Models (XM/XP)", "Non-redundant RefSeq proteins (WP)", and "Uncultured/environmental sample sequences".

The next section is titled "Program Selection". It contains an "Algorithm" section with radio buttons for: "Quick BLASTP (Accelerated protein-protein BLAST)", **blastp (protein-protein BLAST)**, "PSI-BLAST (Position-Specific Iterated BLAST)", "PHI-BLAST (Pattern Hit Initiated BLAST)", and "DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)". Below the radio buttons is a text input field for "Choose a BLAST algorithm".

At the bottom, there is a "BLAST" button and a section for "Search database nr using Blastp (protein-protein BLAST)". There is also a checkbox for "Show results in a new window".

The top hit result is to a protein from *Cynops 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	LEVKGKYEADMVVGGAALINACCRHDNVEEALNLKREVHRKDSSVALDTNKYLSLVKVC	60		
Sbjct 705	LEVKGKYEADMVVGGAALINACCRHDNVEEALNLKRE RKDSSVALDTNKYLSLVKVC	764		
Query 61	AKHGRLDDAINILKEMKEKDVLIKDITLGSFFHVLNGVAMRGEVETVNRLLLEVIVTLGLA	120		
Sbjct 765	AKHGRLDDAINILKEMKEKDVLIKDITLGSFFHVLNGVAMRGEVETVNRLLLEVIVTLGLA	824		
Query 121	KPVANLCSPPVTVHLEKDDAPAALASIDCYKKYNCLPRLHDVLCCLVERGDTELLQKAM	180		
Sbjct 825	KPVANLCSPPVTVHLEKDDAPAALASIDCYKKYNCLPRLHDVLCCLVERGDTELLQKAM	884		
Query 181	DFVSQERGETMLYDLFFAFLLQTAKYKEAKKIETPGLRARPGRLQWFAEKICITGNQMET	240		
Sbjct 885	DFVSQERGETMLYDLFFAFLLQTAKYKEAKKIETPGLRARPGRLQWFAEKICITGNQMET	944		
Query 241	LENFVEMTSKLFECDRDEMYFYLLKLCKENNEWQKADAIWTKMQEENLIPRERTLKLAD	300		
Sbjct 945	LENFVEMTSKLFECDRDEMYFYLLKLCKENNEWQKADAIWTKMQEENLIPRERTLKLAD	1004		
Query 301	LFK 303			
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	LEVKGKYEADMVVGGAALINACCRHDNVEEALNLKREVHRKDSSVALDTNKYLSLVKVC	60		
Sbjct 277	LEVK KYE DMVVGGAALIN CCRHDNVE+A+NLK EV RKDSSVALDT+KYL+LVKV	336		
Query 61	AKHGRLDDAINILKEMKEKDVLIKDITLGSFFHVLNGVAMRGEVETVNRLLLEVIVTLGLA	120		
Sbjct 337	KHGRL+DAINILKEMKEKD+ IKDIT+ SFFH+LN AMRGEVETVN+L E I+TLGLA	396		
Query 121	KPVANLCSPPVTVHLEKDDAPAALASIDCYKKYNCLPRLHDVLCCLVERGDTELLQKAM	180		
Sbjct 397	KP ANLCSP++TVHLEKDD PAALAA+IDC KKY +PRLHDVLC+L+E+G+T+LLQKAM	456		
Query 181	DFVSQERGETMLYDLFFAFLLQTAKYKEAKKIETPGLRARPGRLQWFAEKICITGNQMET	240		
Sbjct 457	DFVSQERGETMLYD+FFAFLL T KYKEAKKIETPGLRARPGRLQWFAEKIATNQMET	516		
Query 241	LENFVEMTSKLFECDRDEMYFYLLKLCKENNEWQKADAIWTKMQEENLIPRERTLKLAD	300		
Sbjct 517	LENVEMTQKLFECDRDQMYYYLLKLCKISNDWRKADATWTKMQEENVIPRETTLRLAD	576		
Query 301	LFK 303			
Sbjct 577	TLK 579			