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

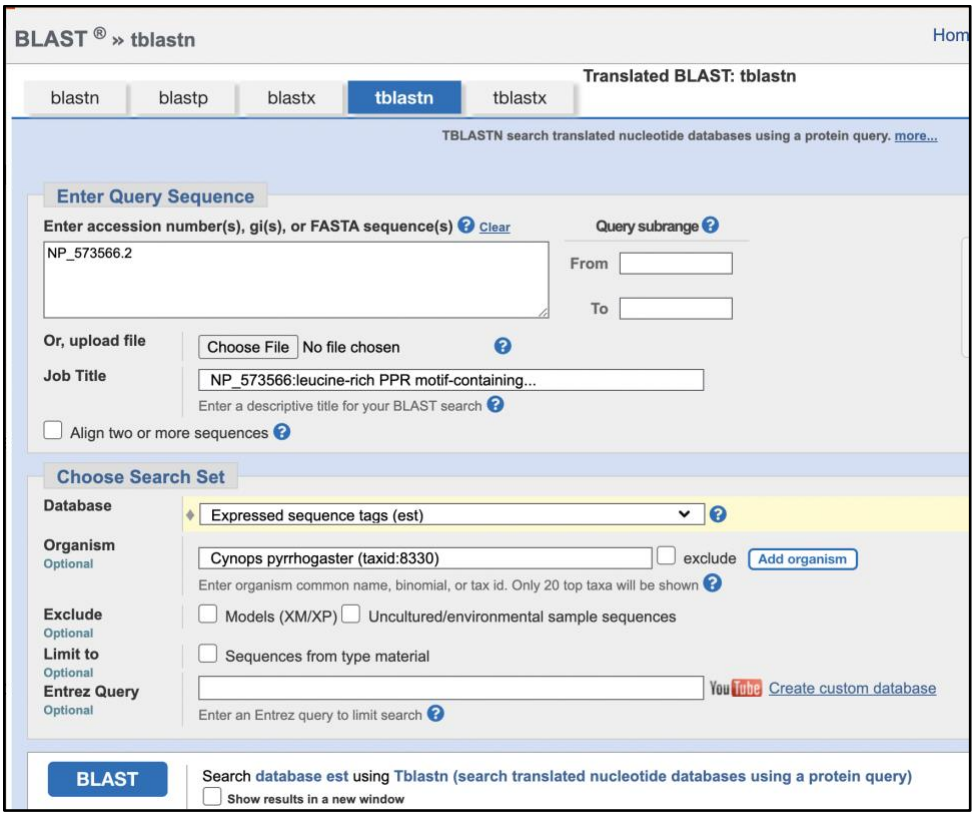
UCSD email: ptvo@ucsd.edu
PID: A59010610

[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

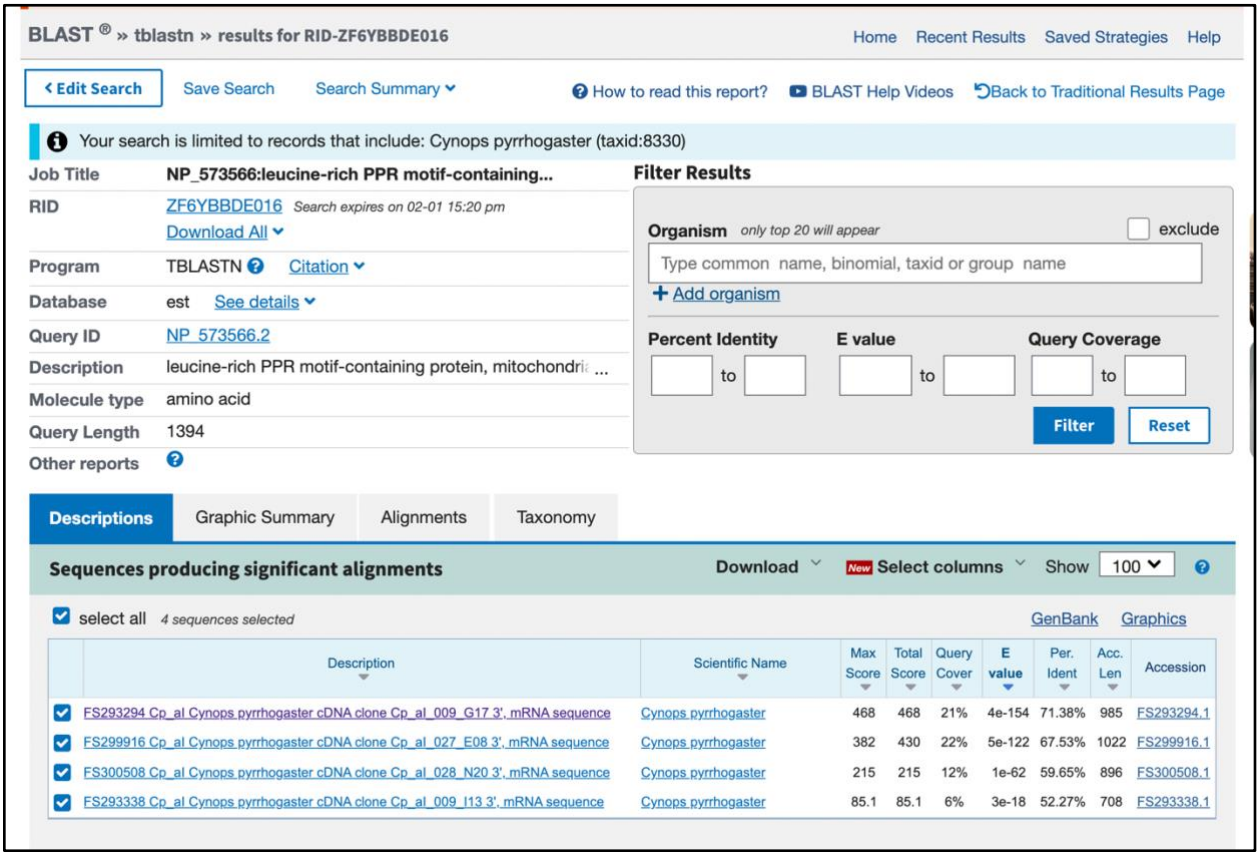
[Q2] Perform a BLAST search against a DNA database, such as a database consisting of genomic DNA or ESTs. The BLAST server can be at NCBI or elsewhere. Include details of the BLAST method used, database searched and any limits applied (e.g. Organism).

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



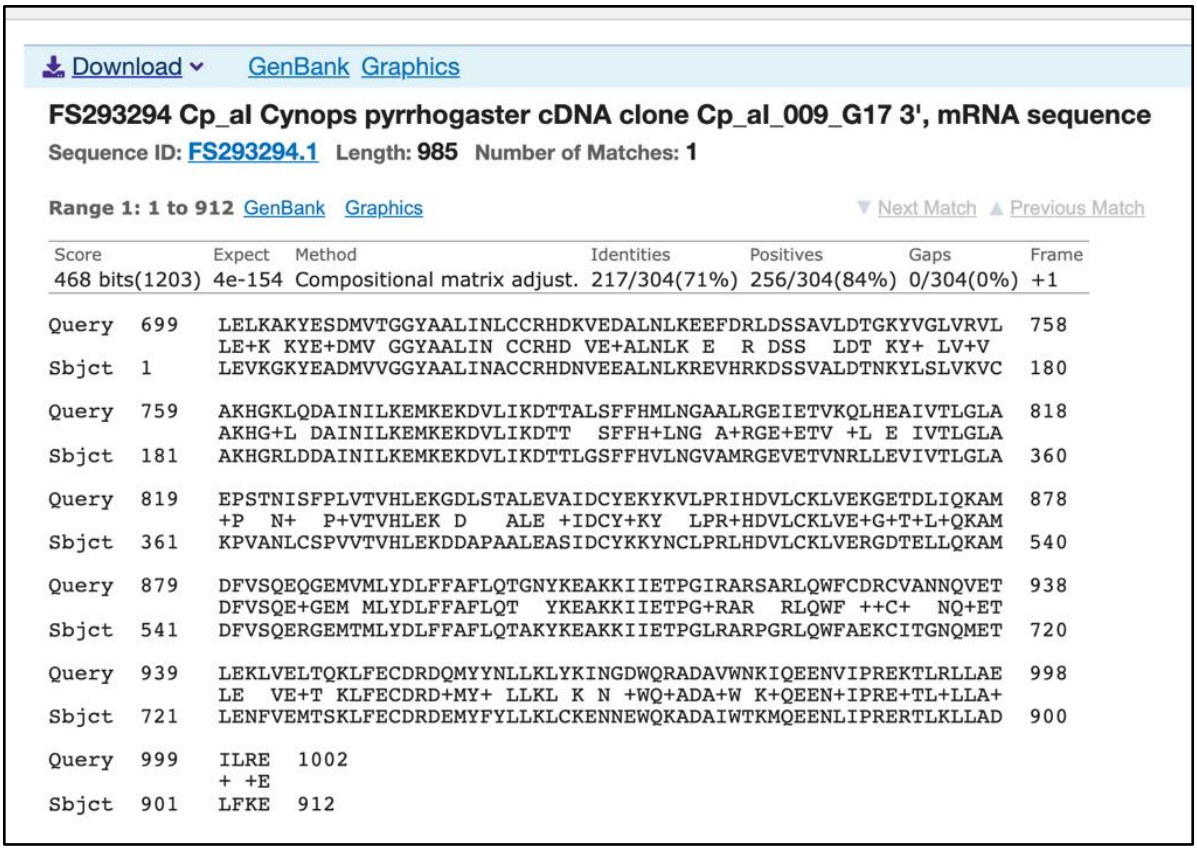
Also include the output of that BLAST search in your document. If appropriate, change the font to Courier size 10 so that the results are displayed neatly. You can also screen capture a BLAST output (e.g. alt print screen on a PC or on a MAC press ⌘-shift-4. The pointer becomes

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*
```

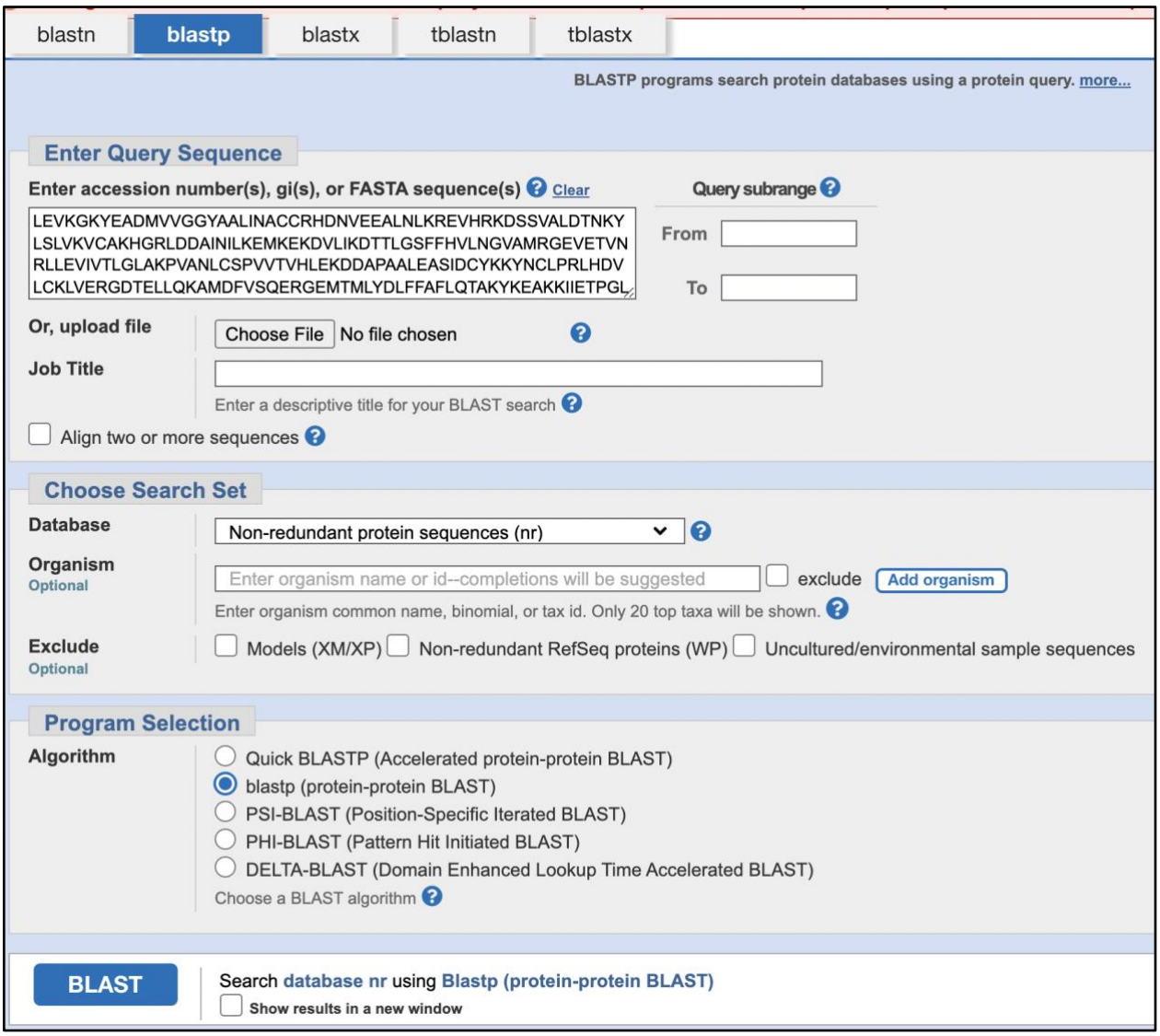
Name: *Cynops pyrrhogaster* LRPPRC
Species: *Cynops pyrrhogaster*

[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
```