**BIMM-143: INTRODUCTION TO BIOINFORMATICS**
**The find-a-gene project assignment** https://bioboot.github.io/bimm143_S20/
**Dr. Barry Grant**

**Name:** Hyeseung (Frankie) Son
**UCSD email:** hson@ucsd.edu
**PID:** A16025601

Overview:

The find-a-gene project is a required assignment for BIMM-143. You should prepare a written report in PDF format that has responses to each question labeled [Q1] - [Q10] below. You may wish to consult the scoring rubric at the end of this document and the example report provided online. The objective of this assignment is for you to demonstrate your grasp of database searching, sequence analysis, structure analysis, and the R environment that we have covered in class.

Submission instructions:

Submit this preliminary report as one document with screenshots of the results inserted appropriately. See the demonstration report linked to on the course website for an example of format. I will email you my decision; proceed with subsequent questions only after we are sure you have found a novel gene.

**[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.

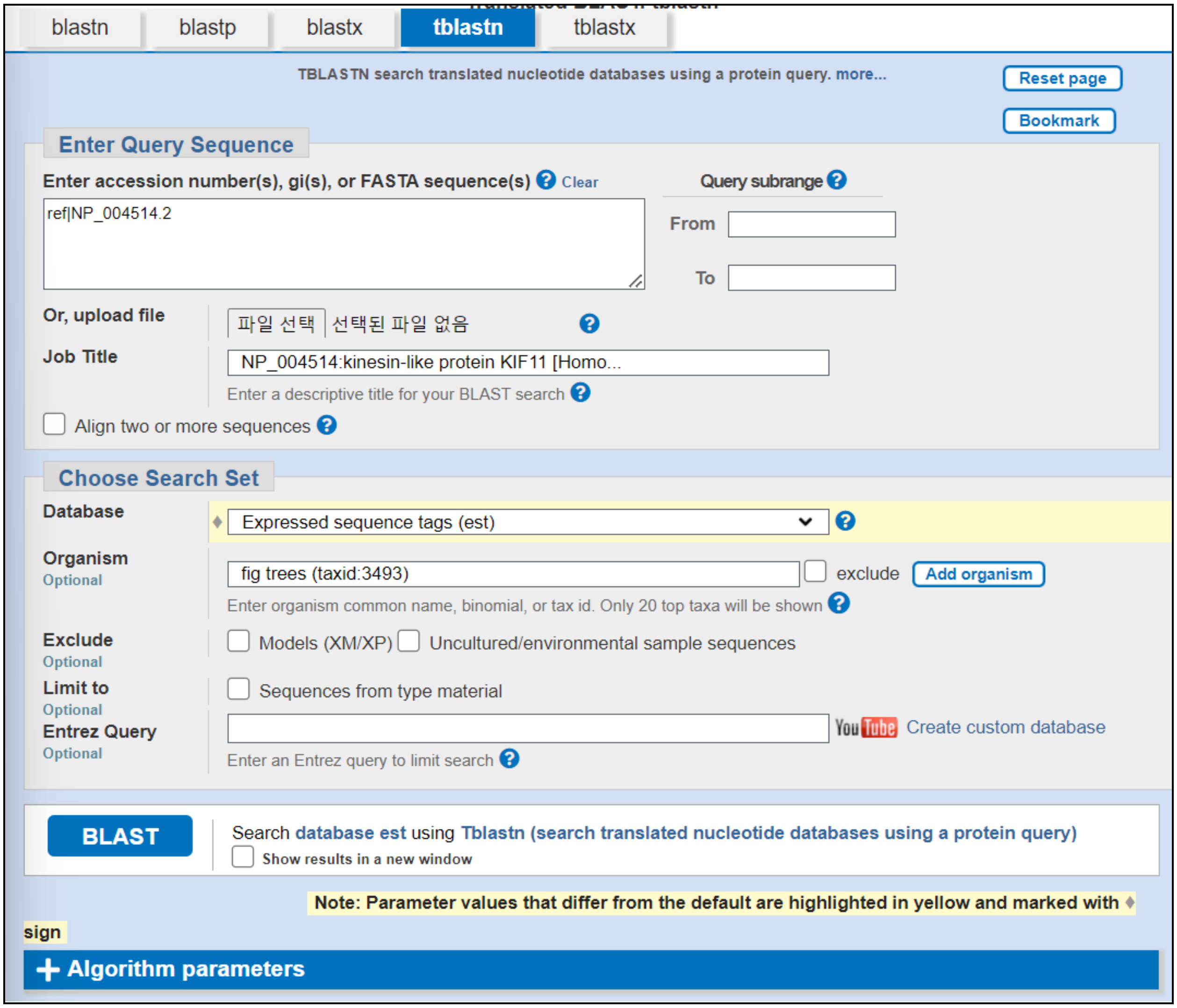
Name: KIF11
Acession: NP_004514.2
Species: Homo Sapiens

**[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
Database: expressed sequence tags (est)
Organism: Fig trees (taxid: 3493)

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.

Search Input: tblastn



TBLASTN search translated nucleotide databases using a protein query. more...

Enter Query Sequence

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

ref|NP_004514.2

Query subrange

From

To

Or, upload file 파일 선택 선택된 파일 없음

Job Title NP_004514:kinesin-like protein KIF11 [Homo...

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database ♦ Expressed sequence tags (est)

Organism Optional fig trees (taxid:3493) □ 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

Enter an Entrez query to limit search

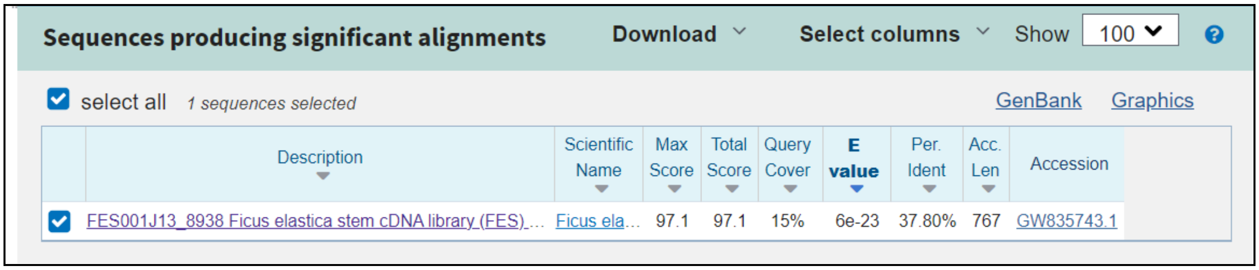
BLAST Search database est using Tblastn (search translated nucleotide databases using a protein query)

□ Show results in a new window

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

+ Algorithm parameters

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.



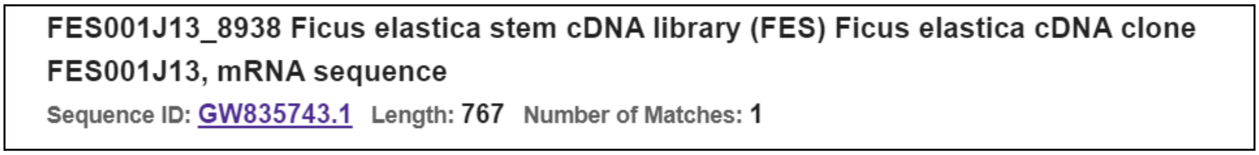
Sequences producing significant alignments

☑ select all 1 sequences selected

| | Description | Scientific Name | Max Score | Total Score | Query Cover | E value | Per. Ident | Acc. Len | Accession |
|---|---|---|---|---|---|---|---|---|---|
| ☑ | FES001J13_8938 Ficus elastica stem cDNA library (FES)... | Ficus ela... | 97.1 | 97.1 | 15% | 6e-23 | 37.80% | 767 | GW835743.1 |

Chosen Match: Accession number GW835743.1, acDNA sequence from Ficus elastica.

Alignment details:



**FES001J13_8938 Ficus elastica stem cDNA library (FES) Ficus elastica cDNA clone FES001J13, mRNA sequence**

Sequence ID: GW835743.1 Length: 767 Number of Matches: 1

```
Query  13   EEKGKNIQVVVRCRPFNLAERKASAHSIVECDPVRKEVSVRTGGLADKSSRKTYTFDMVF  72
            E KG NI+V  R RP  L +  +S    V   P   E   R   L+    + ++ FD VF
Sbjct  183  ELKG-NIRVFCRVRPL-LPDDGSSGEGKVISYPTSMETLGRGIDLSQIGQKHSFMFDKVF  356

Query  73   GASTKQIDVYRSVVCPILDEVIMGYNCTIFAYGQTGTGKTFTMEGERSPNEEYTWEEDPL  132
                 Q DV+  +   ++   + GY   IFAYGQTG+GKT+TM G+    E        L
Sbjct  357  MPDASQEDVFEEI-SQLVQSALDGYKVCIFAYGQTGSGKTYTMMGKPGQPE--------L  509

Query  133  AGIIPRTLHQIF---EKLTDNGTEFSVKVSLLEIYNEELFDLLN  173
             G+IPR+L QIF   + L   G ++ ++VS+LEIYNE + DLL+
Sbjct  510  KGLIPRSLEQIFRTRQSLLPQGWKYEMQVSMLEIYNETVRDLLS  641
```

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.

FASTA Sequence, translated from DNA sequence:

```
>GW835743.1_1 FES001J13_8938 Ficus elastica stem cDNA library (FES) Ficus
elastica cDNA clone FES001J13, mRNA sequence
RTC**MHKRNFRYPTYPSWRQKQNMKNRRKS*VNYKIAWRMPNLKLLKERCCAKSYIIRF
WN*RGTFGCSVECDHYCLMMVLLVKGRLSPIPHQWKLLDEALICHKLGKNILSCLTKFSC
LMHRKKMSLKKSHSLFKVRLTVIRSAFSPMGKRVQAKPIP*WVNQDSPS*KG*FLVP*NK
YFELDNLFCHKVGNMKCRYLCWRYITKLFGTCYLQIDHLLICCERKTVLVKHTQSNMT*M
GIHMYRI*QLWMFIVL
```

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.

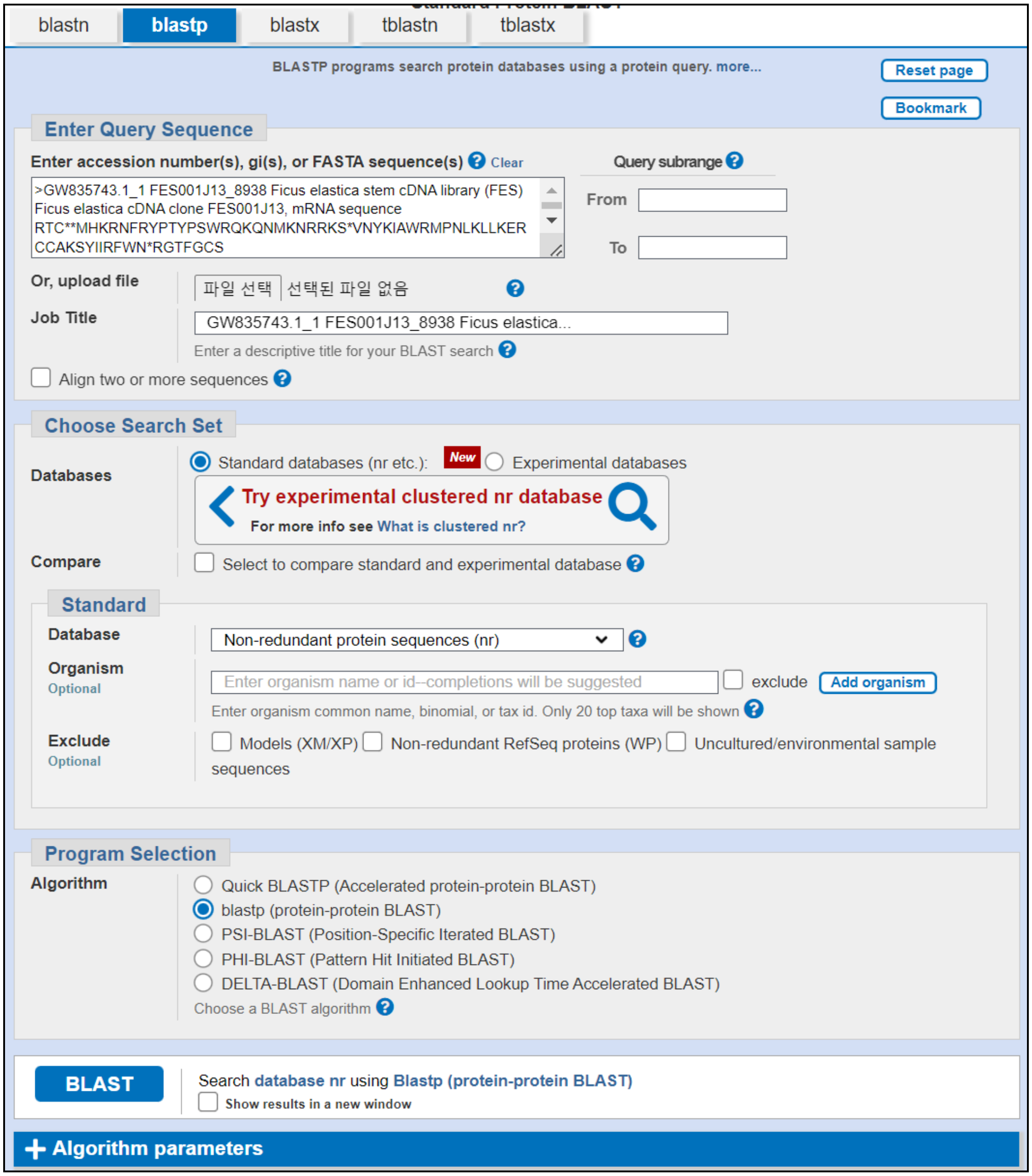
Name of novel protein: Kinesin-1 [Striga hermonthica]
Species: *Striga hermonthica*

> *Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae; Pentapetalae; asterids; lamiids; Lamiales; Orobanchaceae; Buchnereae; Striga.*

**[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 Search: Using FASTA sequence from Q3



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

**Enter Query Sequence**

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

```
>GW835743.1_1 FES001J13_8938 Ficus elastica stem cDNA library (FES)
Ficus elastica cDNA clone FES001J13, mRNA sequence
RTC**MHKRNFRYPTYPSWRQKQNMKNRRKS*VNYKIAWRMPNLKLLKER
CCAKSYIIRFWN*RGTFGCS
```

Query subrange

From

To

Or, upload file: 파일 선택 | 선택된 파일 없음

Job Title: GW835743.1_1 FES001J13_8938 Ficus elastica...

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)
○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

BLAST

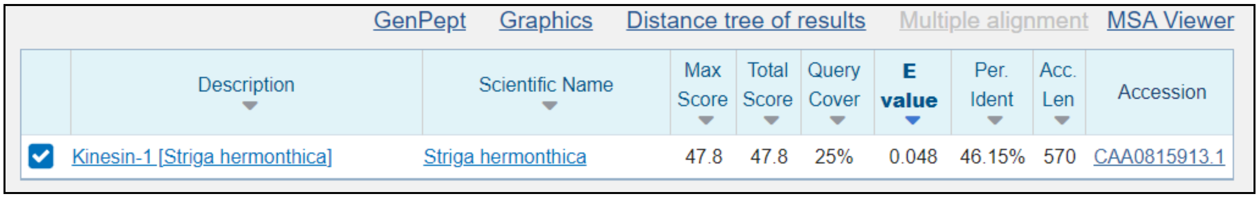
Search database nr using Blastp (protein-protein BLAST)

□ Show results in a new window

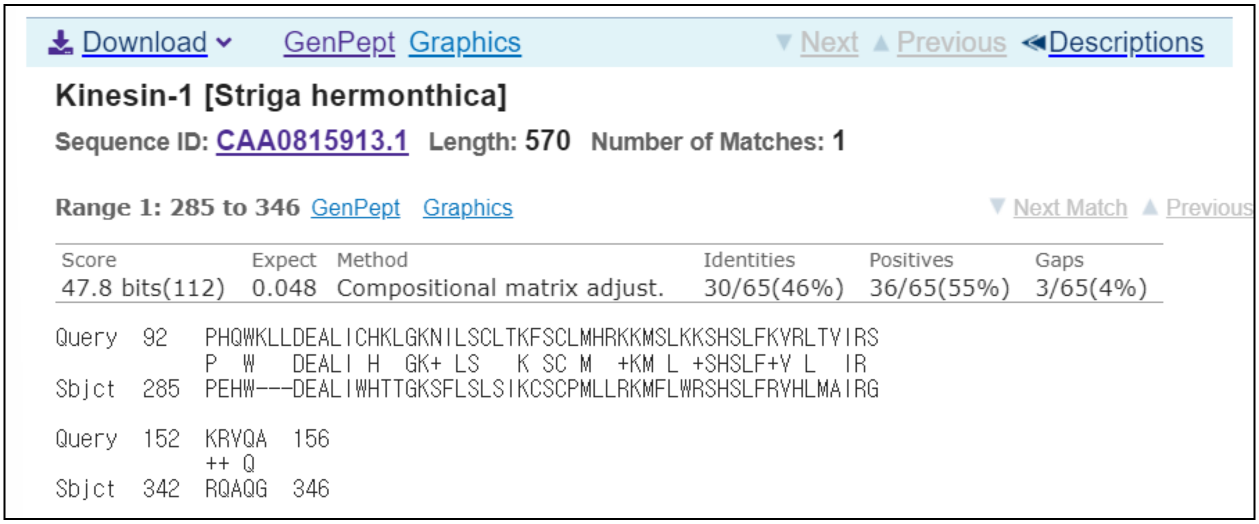
+ Algorithm parameters

The chosen protein is:
Kinesin-1 [Striga hermonthica], the search query loads a match with a low e- value, and low percent identity, indicating that this is likely a novel protein.

Sequence ID: CAA0815913.1



| | Description | Scientific Name | Max Score | Total Score | Query Cover | E value | Per. Ident | Acc. Len | Accession |
|---|---|---|---|---|---|---|---|---|---|
| ☑ | Kinesin-1 [Striga hermonthica] | Striga hermonthica | 47.8 | 47.8 | 25% | 0.048 | 46.15% | 570 | CAA0815913.1 |



Kinesin-1 [Striga hermonthica]

Sequence ID: CAA0815913.1 Length: 570 Number of Matches: 1

Range 1: 285 to 346 GenPept Graphics

| Score | Expect | Method | Identities | Positives | Gaps |
|---|---|---|---|---|---|
| 47.8 bits(112) | 0.048 | Compositional matrix adjust. | 30/65(46%) | 36/65(55%) | 3/65(4%) |

```
Query  92   PHQWKLLDEALICHKLGKNILSCLTKFSCLMHRKKMSLKKSHSLFKVRLTVIRS
            P  W   DEALI H  GK+ LS   K SC M  +KM L +SHSLF+V L  IR
Sbjct  285  PEHW---DEALIWHTTGKSFLSLSIKCSCPMLLRKMFLWRSHSLFRVHLMAIRG

Query  152  KRVQA  156
            ++ Q
Sbjct  342  RQAQG  346
```

Here is a side by side alignment of the search query and the top search hit (Kinesin-1)

```
Query  92   PHQWKLLDEALICHKLGKNILSCLTKFSCLMHRKKMSLKKSHSLFKVRLTVIRSAFSPMG  151
            P  W   DEALI H  GK+ LS   K SC M  +KM L +SHSLF+V L  IR     M
Sbjct  285  PEHW---DEALIWHTTGKSFLSLSIKCSCPMLLRKMFLWRSHSLFRVHLMAIRGLHGYMV  341

Query  152  KRVQA  156
            ++ Q
Sbjct  342  RQAQG  346
```

Here is a full length sequence of the isolated protein (top search hit) in FASTA format:

```
>CAA0815913.1 Kinesin-1 [Striga hermonthica]
MRSAGRIYTRLSKFSVLPPVVLDFSQSEKLEIVANVKKYLQFFHLEYYPKLELEAYTNLTFEFYTTFKFV
KNGTDVVCRLGDKRKTIDTALMHQIFGFVSTGAEAPTNGLIVASIQTSDRFSPSFGMLVAALTRHFKSPM
REEDVVEAQRLVIKYFCAERNGSTEAEDTDLRGMVKEIVARMEFLMEALGGEVATLRLDLQQVQDEVATY
KEWIGKSIPELHSWQTKATESTCLSQSEQIRRLQKQLAVSKELKGNIRVFCRVRPFLSDDGVGNNAKVVS
FPTSPEHWDEALIWHTTGKSFLSLSIKCSCPMLLRKMFLWRSHSLFRVHLMAIRGLHGYMVRQAQGWKYD
MRISMLEIYNETIRDLLAPNRTCSDASRAENAGKQYAIKHDANGNTQVFDLTVVDVQSSKEVSYLLERAA
QSRSVGKTQMNEQSSRSHFVFTLRIMGFNENTDQQVCVLNLIDLAGSERLSKSGSTGNQLKETQAINKSL
SSLSDVIFALAKKEEHVPYRNSKLTYLLQPCLGGDSKTLMFVNVSPDHSLEGESLCSLRFAARVNACEIG
VPRRQTNLRS
```