

**BIMM-143: INTRODUCTION TO BIOINFORMATICS**  
**The find-a-gene project assignment [https://bioboot.github.io/bimm143\\_S20/](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 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: KIF11

Accession: 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

blastn

blastp

blastx

**tblastn**

tblastx

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

Reset page

Bookmark

Enter Query Sequence

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

Query subrange [?](#)

From

To

Or, upload file

파일 선택

선택된 파일 없음 [?](#)

Job Title

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

☐ Align two or more sequences [?](#)

Choose Search Set

Database

Expressed sequence tags (est) [?](#)

Organism

Optional

☐ 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

[YouTube](#) [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

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

Download

Select columns

Show

100

☒ select all 1 sequences selected

[GenBank](#)

[Graphics](#)

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	<a href="#">FES001J13_8938 Ficus elastica stem cDNA library (FES)...</a>	<a href="#">Ficus ela...</a>	97.1	97.1	15%	6e-23	37.80%	767	<a href="#">GW835743.1</a>

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

Query 133 AGIIPRTLHQIF---EKLTDNNGTEFSVKVSLLEIYNEELFDLLN 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*VNYKIAWRMPNLKLLKERCCAHSYIIRF
WN*RGTFGCSVECDHYCLMMVLLVKGRLSPIPHQWKLLEALICHKLGNILSCLTKFSC
LMHRKKMSLKKSHSLFKVRLTVIRSAFSPMGKRVQAKPIP*WVNQDSPS*KG*FLVP*NK
YFELDNLFCHKVGNMKCRYLCWRYITKLFGTCYLGIDHLLICCEKRTVLVKHTQSNMT*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

blastn

**blastp**

blastx

tblastn

tblastx

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

Reset page

Bookmark

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\*\*MHKRNFRYPYTPSWRQKQNMKNRRKS\*VNYKIAWRMPNLKLLKER  
CCAKSYIIRFWN\*RGTFGC

Query subrange ?

From

To

Or, upload file

파일 선택

선택된 파일 없음 ?

Job Title

GW835743.1\_1 FES001J13\_8938 Ficus elastica...

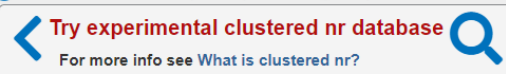
Enter a descriptive title for your BLAST search ?

☐ 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

<a href="#">GenPept</a> <a href="#">Graphics</a> <a href="#">Distance tree of results</a> <a href="#">Multiple alignment</a> <a href="#">MSA Viewer</a>									
	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	<a href="#">Kinesin-1 [Striga hermonthica]</a>	<a href="#">Striga hermonthica</a>	47.8	47.8	25%	0.048	46.15%	570	<a href="#">CAA0815913.1</a>

[Download](#)
[GenPept](#)
[Graphics](#)
[Next](#)
[Previous](#)
[Descriptions](#)

### Kinesin-1 [Striga hermonthica]

Sequence ID: [CAA0815913.1](#) Length: 570 Number of Matches: 1

Range 1: 285 to 346 [GenPept](#) [Graphics](#) [Next Match](#) [Previous](#)

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	PHQWKLLDEALICHKLGKNILSCLTKFSCLMHRKKMSLKKSHSLFKYRLTVIR	
		P W DEALI H GK+ LS K SC M +KM L +SHSLF+V L IR M	
Sbjct	285	PEHW---DEALIWHHTGKSFLSLSIKCSCPMLLRKMFWRSHSLFRVHLMAIRG	

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 PHQWKLLDEALICHKLGKNILSCLTKFSCLMHRKKMSLKKSHSLFKYRLTVIRSAFSPMG 151
          P W  DEALI H  GK+ LS  K SC M  +KM L +SHSLF+V L  IR      M
Sbjct 285 PEHW---DEALIWHHTGKSFLSLSIKCSCPMLLRKMFWRSHSLFRVHLMAIRGLHGYMV 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]
MRSAGRIYTRLSKFSVLPPVLDIFSQSEKLEIVANVKYQLQFFHLEYYPKLELEAYTNLTFFEFYTTFKFV
KNGTDVVCRLGDKRKTIDTALMHQIFGFVSTGAEAPTNGLIVASIQTSDRFSPSFGMLVAALTRHFKSPM
REEDVVEAQRLVIKYFCAERNGSTEAEDTDLRGMVKEIVARMEFLMEALGGEVATLRDLQQVQDEVATY
KEWIGKSIPELHSWQTKATESTCLSQSEQIRRLQKQLAVSKELKGNIRVFCRVRPFLSDDGVGNNAKVVS
FPTSPEHWDEALIWHHTGKSFLSLSIKCSCPMLLRKMFWRSHSLFRVHLMAIRGLHGYMVRQAQGWKYD
MRISMLEIYNETIRDLLAPNRTCSDASRAENAGKQYAIKHDANGNTQVFDLTVVDVQSSKEVSYLLERAA
QSRSVGKTQMNEQSSRSHFVFTLRIMGFNENTDQQVCVLNLIIDLAGSERLSKSGSTGNQLKETQAINKSL
SSLSDVIFALAKKEEHVPYRNSKLTYLLQPCLGGDSKTLMFVNVSPDHSLEGESLCSLRFAARVNACEIG
VPRRQTNLRS
  
```

**[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 alignment for building a phylogenetic tree that illustrates species divergence.

### BLASTP Search: Using FASTA sequence from Q3

```
>NP_004514.2 kinesin-like protein KIF11 [Homo sapiens]
MASQPNSSAKKKEEKGNIQVVVRCRPFNLAERKASAHSIVECDPVRKEVSVRTGGLADKSSRKTYTFDMVFGASTKQI
DVYRSVVCPIDEVIMGYNCTIFAYGQTGTGKTFTMEGERSPNEEYTWEEEDPLAGIIPRTLHQIFEKLTNGTEFSVKV
SLLEIYNEELFDLLNPSSDVSERLQMFDDPRNKRGVIIKGLEEITVHNKDEVYQILEKGAAKRTTAATLMNAYSSRSHS
VFSVTIHKETTIDGEELVKIGKLNLDLAGSENIGRSGAVDKRAREAGNINQSLTLGRVITALVERTPHVPYRESKL
TRILQDSLGGRTSIIATISPASLNLEETLSTLEYAHRANKILNKPEVNQKLTKKALIKKEYTEEIERLKRDLAAAREK
NGVYI SEENFRVMSGKLTVQEEQIVELIEKIGAVEEELNRVTELFMDNKNELDQCKSDLQNKQTQELETTQKHLQETKLQ
LVKEYIITSALESTEEKLHDAASKLLNTVEETTKDVSGLHSKLDRKKAVDQHNAAEQDIFGKNLNSLFNNMEELIKDGS
SKQKAMLEVHKTLFGNLLSSVSALDTITTVALGSLTIPENVSTHVSQIFNMILKEQSLAAESKTVLQELINVLKTDL
LSSLEMILSPTVVSILKINSQKHFKTSLTVADKIEDQKKELDGFLSILCNNLHELQENTICSLVESQKQCGNLTEDL
KTIKQTHSQELCKLMNLWTERFCALEEKCENIQKPLSSVQENIQQKSKDIVNKMFTFHSQKFCADSDGFSQELRNFNQEG
TKLVEESVKHSDKLNGLNLEKISQETEQRCESLNTRTVYFSEQWVSSLNEREQELHNLLEVVSQCCEASSSDITEKSDGR
KAAHEKQHNI FLDQMTIDEDKLI AQNLELNETIKIGLTKLNCFLEQDLKLDIPTGTTPQRKSYLPSTLVRTEPREHLL
DQLKRKQPELLMMLNCSENNKEETIPDVDVEEAVLGQYTEEPLSQEPSVDAGVDCSSIGGVPPFQHKKSHGKDKENRGI
NTLERSKVEETTEHLVTKSRLPLRAQINL
```

```
>CAA0815913.1 Kinesin-1 [Striga hermonthica]
MRSAGRIYTRLSKFSVLPPVLDLFSQSEKLEIVANVKYLQFFHLEYYPKLELEAYTNLTFFEFYTTFKFVKNGTDVVC
RGDGRKTIDTALMHQIFGFVSTGAEAPTNGLIVASIQTSDRFSPSFGMLVAALTRHFKSPMREEDVVEAQRLLVIKYFCA
ERNGSTAEADTDLRGMVKEIVARMEFLMEALGGEVATLRLDLQVQDEVATYKEWIGKSIPELHSHWQTKATESTCLSQS
EQIRRLQKQLAVSKELKGNIRVFCRVRPFLSDDGVGNNAKVVSFPTSPEHWDEALIWHHTGKSFLSLSIKCSCPMLLRK
MFLWRSHSLFRVHLMAIRGLHGYMVRQAQGWKYDMRISMLEIYNETIRDLLAPNRTCSASRAENAGKQYAIKHDANGN
TQVFDLTVVDVQSSKEVSYLLERAAQSRVSGKTQMNEQSSRSHFVFTLRIMGFNENTDQQVCVLNLDLAGSERLSKSG
STGNQLKETQAINKSLSSLSDVIFALAKKEEHVPYRNSKLTYLLQPCLGDSKTLMFVNVSPDHSLEGESLCSLRFAAR
VNACEIGVPRRQTNLRS
```

```
>KAI3457004.1 hypothetical protein Pfo_013667 [Paulownia fortunei]
MASRNQNKPPSSPSHYSVDEVSVDKRRRIGNTKMPPNTGTRMQTRQAFSVVNGGQDLPPISGPPSNSGSDSGVIEFT
KEDVEALLNEKLRIKNKFNYKEKSEQMAECIKRLKQCIKWQQLEGNYVTEQEKLKNLLELAEKRSNDMKLLMKAKEDE
LNSIIMELRNLEALQEKFAKEELDKLEALDSLAREKDSRLAAERVQASISEELKRTQQDNASGIQKIQSLNDMYKRLQ
EYNTSLQQYNSRLQSELHATNETLKRVEKEKAAVVENLSTLRGHYTSLQEQLTSSRALQDEAMKQKEALGSEVTCRGLD
LQQVRDDRDRQLLQVQALSAEVVYKECTGKSIAELDSLTTKTNELESTCLSQSEQIRRLQEQLAFAEKKLKLSDMSAM
ETRSEFEEQKTFILELQNRLADAELKIVEGEKLRKKLHNTILELKGNIRVFCRVRPLLSDDGVGIDAKVVSFPTSMEAL
GRGIDLTQNGQKLSFTFDKVFPLPDASQEDVFVEISQLVQSALDGYKVCIFAYGQTGSGKTYTMMGKPGPPDQKGLIPRS
LEQVFETRQILQAQGWKYEMQVSMLEIYNETIRDLLAPNRSFGDASRAETGGKQYAIKHDANGNTHVSDLTIVDVRSSK
EVSYLLDRAAQSRVSGKTQMNEQSSRSHFVFTMRIMGFNESTDQQVQGVNLDLAGSERLSKSGSTGDRLKETQAINK
SLSSLSDVIFALAKKEEHVPYRNSKLTYLLQPCLGDSKTLMFVNVSPDPSSVGESLCSLRFAARVNACEIGVPRRQTN
LRSSDSRLSIG
```



>GER43185.1 kinesin [Striga asiatica]

MSSSKQLDIQFEDHRQNQNKWLCPLREDVFAAFISKDNPTVHNIFGAASSLFSPFLFGKFFDPSDAFPLWEFDPQALLP  
NNFNSSEHETVDWFRMENGYYVLAQQLPNGTSQNTIQVCIANGKILEIYGQWKQQRESKTKDWKSSHWHEHGFVRRLLELP  
EQADWRKLEAHVKNELVLEIKVPDITTEGDVAQMVERSLSMRENKDSVDKVSVDRRRKMPPLNTGIRVRKAFSVVNGGQN  
LPQVSGPPSSSGSECGVSEFTKEEVEALLNGKLQIKNKNFYKEKSEQMAECIKKLKQCIKWQQLEGNVYVTEQGKLDL  
LGVAEKRSNDMELLMKAKEDELNSIIMDLRQKLEDLQEKFVGEEREKLDALDSLEKEKFYRLAAEKLQFSISEELKRVQ  
EDNAAGIQKIQTLDNDMYKRLQEYNTSLQQYNSRLQSELQATNETLKRVEKEKAAVVENLSTLRGHYTSLQEQLTSSRAL  
QDEAMKQKEALGSEVTSLRGDLQQVRDDRDRQLLQVQALSAEVVYKECTGKSIAELDSLSTKTTELESTCLSQSEQIR  
RLQEQLAFAEKKLKSISAMETRSEFEEQKTTILELQNCLADAESKIVEGEKLRKKLHNTILELKGNIIRVFCRVRPLL  
SDDGVGNDAKVVSFPISTETLGRGIDLAQNGQKHSFTFDKVFMPDASQEDVFVEISQLVQSALDGYKVCIFAYGQTGSG  
KTHMTMMGKPLPDQKGLIPRSLEQVFETRQILQAQGWKYDMQVSMLEIYNETIRDLLAPNRTGLDASRAENAGKQYAIK  
HDANGNTHVSELTVVDVRSKEVSYLLDRAAQSRVSGKTQMNEQSSRSHFVFTLRIMGFNESTDQQVQGVNLNIDLGS  
ERLSKSGSTGDRLETQAINKSLSSLSVDVIFALAKKEEHVPYRNSKLTLYLLQPCLGGSKTLMFVNVSPDPSSVGESLC  
SLRFAARVNACEIGVPRRQTNLRSSSSSSSDSRLSIG

>XP\_011095312.1 kinesin-like protein KIN-14N [Sesamum indicum]

MASKNQNKPPSSPSHSKYSVDDVSDKRRRIGNTKMPPNSGTRVQTRQAFSVVNGGQDPPPTSGPPSNSG  
SDSGVTEFTREDVEALLIEKLRIKNKNFYKEKSEQMAEYIKRLKQCIKWQQCEGNVYVTEQEKLNLELAEKKCNDE  
LLMKAKEDELNSIIMELRNLEALQEKFSKEELDKLEALDSLAKEDSRLAERLNASLSEELKRSQEDNASNVQKIQS  
LNDMYKRLHEYNTSLQQYNSKLQSEIHAIKETLKHVEQEKAIVENLSTLRGHSTSLQEQLASSRASQDEALKQKEALG  
SEVTCRLGELQQVRDDRDRQLVQVQALSAEVVYKECTGKSIAELDSLTTKTNELESTCLSQSEQIRRLHEQLAFAEKK  
LKLSDMSAMETRSEFEEQKTIISQLQNRLADAESKIVEGEQLRKKLHNTILELKGNIIRVFCRVRPLLSDDGVGADTKVV  
SFPTSMEAQGRGIDLTQNGQKLSFTFDKVFVPDASQEDVFVEISQLVQSALDGYKVCIFAYGQTGSGKTYTMMGKPAPI  
DQKGLIPRSLEQVFETRQILQAQGWKYGMQVSMLEIYNETIRDLLAPNRSGFDASRAENAGKQYSIKHDANGNTHVSDL  
TIVDVHSSKEVSYLLDRAAQSRVSGKTQMNEQSSRSHFVFTLRITGFNESTDQQVQGVNLNIDLASERLSKSGSTGDR  
LKETQAINKSLSSLSVDVIFALAKKEEHVPYRNSKLTLYLLQPCLGGSKTLMFVNVSPDPSSVGESLCSLRFAARVNACE  
IGVPRRQTNLRSLSRLSIG

## Alignment: Obtained using MUSCLE at EBI (3.8): CLUSTAL multiple sequence alignment

Human_kinesin	MASQPNSSAKKKEEKGNIQVVRCRPFNLAERKASAHSIVECDPVRKEVSVRTGGLADK
Striga_hermonthica	-----
Striga_asiatica	MSSSKQLDIQFEDHRQNQNKWLCPLREDVFAAFISKDNPTVHNIFGAASSLFSPFLFGKF
Paulownia_fortunei	MAS-----RNQNK-----
Sesamum_indicum	MAS-----KNQNK-----

Human_kinesin	SSRKTYTFDMVFGASTKQIDVYRSVVCPIDEVIMGYNCTIFAYGQTGTGKTFTM----
Striga_hermonthica	-----
Striga_asiatica	FDPSDAFPLWEFDPQALLPNNFNSSEHETVDWFRMENGYYVLAQQLPNGTSQNTIQVCIAN
Paulownia_fortunei	-PPSS-----
Sesamum_indicum	-PPSS-----

Human_kinesin	-----EGERSPNEEYTWEEDPLAGIIPRTLHQIFEKLTDNNGTEFSVKVSL
Striga_hermonthica	-----
Striga_asiatica	GKILEIYGQWKQQRESKTKDWKSSHWHEHGFVRRLLEPEQADWRKLEAHVKNELVLEIKV
Paulownia_fortunei	-----PSHSKY-----
Sesamum_indicum	-----PSHSKY-----

Human_kinesin	LEIYNEELF-----DLLNPSSDVSERLQMFDDPRNKRGVIIKGLEEITVH
Striga_hermonthica	-----MRSA-GRIYTRLKFSV-

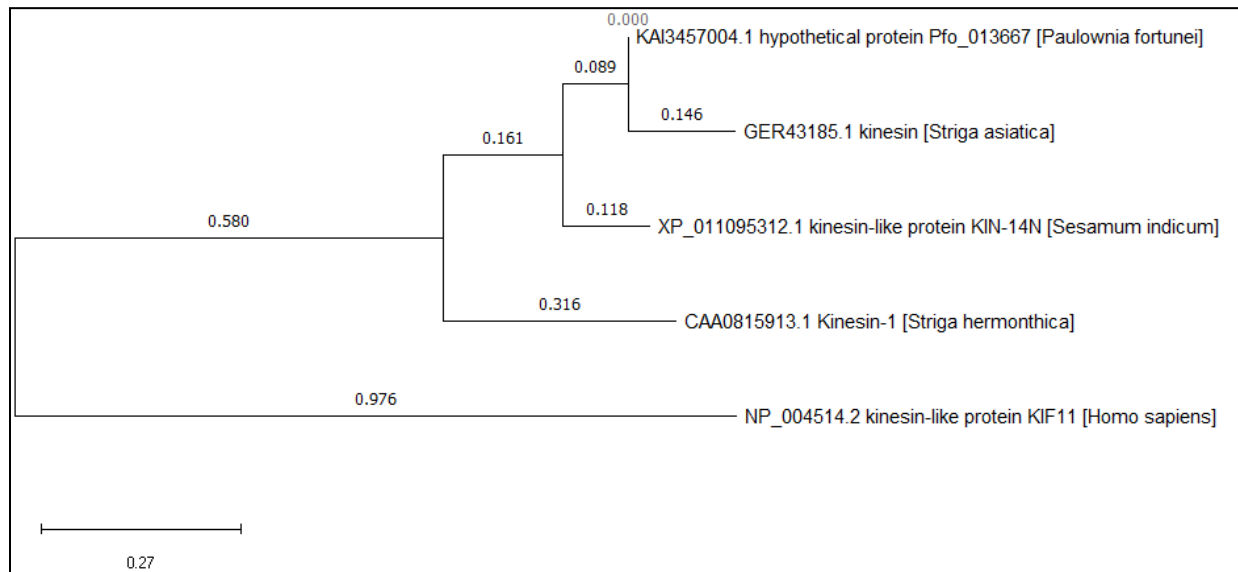
Striga_asiatica	PDITTEGDVAQMVERSLSMRENKDSVDKVSVDRRRK-----PLNT-GIRV--RKAFSVV
Paulownia_fortunei	-----SVDEVSVDKRRRIGNTKMPPNT-GTRMQTRQAFSVV
Sesamum_indicum	-----SVDDVSVDKRRRIGNTKMPPNS-GTRVQTRQAFSVV
	. * . : *
Human_kinesin	NKDEVYQILEKGAAKRRTAATLMNAYSSRSHSVFSVTIHMKETTIDGEELVKIGKLNLD
Striga_hermonthica	-----LPPV-----VL-----DFSQSE-----KLEI--
Striga_asiatica	NGGQNLPQVSGPPSSSGSECGVS-----EFTKEEVEALLNGKLQIKN
Paulownia_fortunei	NGGQDLPPISGPPSNGSDSGVI-----EFTKEDVEALLNEKLRIKN
Sesamum_indicum	NGGQDPPPTSGPPSNGSDSGVT-----EFTREDVEALLIEKLRIKN
	: : . ** :
Human_kinesin	LAGSENIGRSGAVDKRAREAGNINQSLTLGRVITALVERTPHVPYRESKLTRILQDSL
Striga_hermonthica	-----VANVKKYLQFFH-----
Striga_asiatica	KFNYKE-----KSEQMAECIKKLKQCIKWFOQLEGNYVTEQGKLDLLGVAEK
Paulownia_fortunei	KFNYKE-----KSEQMAECIKRLKQCIKWFOQLEGNYVTEQEKLNLELAEK
Sesamum_indicum	KFNYKE-----KSEQMAEYIKRLKQCIKWFOQCEGNYVTEQEKLNLELAEK
	: : . : :
Human_kinesin	GRTRTSIIATISPASLNLEETLSTLEYAHRANKNILNKPEVNQKLTKKALIKYEETEEIERL
Striga_hermonthica	----LEYYPKLELEAYT-----NLTFEFYTF----KFVKNGTDVVCRL
Striga_asiatica	RSNDMELLMKAKADELN-----SIIMDLRQKLEDLQEFVGEEREKLDAL
Paulownia_fortunei	RSNDMKLLMKAKADELN-----SIIMELRKNLEALQEFKFAKEELDKLEAL
Sesamum_indicum	KCNMELLMKAKADELN-----SIIMELRNNLEALQEFKSKEELDKLEAL
	. . . . : . . : : : *
Human_kinesin	KRDLAAAREKNGVYI SEENFRVMSGKLTVQEEQIVELIEKIGAVEEELNRVTELFMDNKN
Striga_hermonthica	G---DKRKTIDTALMHQIFGFVSTGAEAPTNGLIVASIQTSDRFSPSFGMLVAA----LT
Striga_asiatica	D---SLEKEKFYRLAAEKLQFSISEELKRVQEDNAAGIQKIQTLDNMYKRLQEY----NT
Paulownia_fortunei	D---SLAREKDSRLAAERVQASISEELKRTQQDNASGIQKIQSLNDMYKRLQEY----NT
Sesamum_indicum	D---SLAKEKDSRLAAERLNASLSEELKRSQEDNASNVQKIQSLNDMYKRLHEY----NT
	. : : : . : . : .
Human_kinesin	ELDQCKSDLQNKTOELETQKHLQETKLQLVKEEYITSALESTEEKLHDAASKLLNTVEE
Striga_hermonthica	RHFKSPMREEDVVEAQLVIKYFCAER-----NGSTEAEEDTLRG-----
Striga_asiatica	SLQQYNSRLQSELQATNETLKRVEKEK-----AAVVENLST-LRG-----
Paulownia_fortunei	SLQQYNSRLQSELHATNETLKRVEKEK-----AAVVENLST-LRG-----
Sesamum_indicum	SLQQYNSKLQSEIHAIKETLKHVEQEK-----SAIVENLST-LRG-----
	: : . * . . . * . * .
Human_kinesin	TTKDVSGLSKLDLRKKAVDQHNAEAQDIFGKNLNSLFNNMEELIKDGSSK--QKAMLEVH
Striga_hermonthica	---MVKEIVARMEF-----LMEALGGEVATLRDLQVQVD-----E
Striga_asiatica	---HYTSLQEQLTSSRALQDEAMKQKEALGSEVTSLRGDLQVVRDDRDRQLLVQVQALSAE
Paulownia_fortunei	---HYTSLQEQLTSSRALQDEAMKQKEALGSEVTCLRGDLQVVRDDRDRQLLVQVQALSAE
Sesamum_indicum	---HSTSLQEQLASSRASQDEALKQKEALGSEVTCLRGELQVVRDDRDRQLVQVQALSAE
	. : . : : * : : * : : : .
Human_kinesin	KTLFGNLLSSSVSALDTITTVALGSLTIPENVSTHVSQIFNMILKEQSLAAESKTVLQE
Striga_hermonthica	VATYKEWIGKSIPELHSWQTKA----T--ESTCLSQSEQIRR--LQKQLAVSK-----
Striga_asiatica	VVKYKECTGKSIAELDSLTKT----TELESTCLSQSEQIRR--LQEQLAFAEKKKLKLS
Paulownia_fortunei	VVKYKECTGKSIAELDSLTKT----NELESTCLSQSEQIRR--LQEQLAFAEKKKLKLS
Sesamum_indicum	VVKYKECTGKSIAELDSLTKT----NELESTCLSQSEQIRR--LHEQLAFAEKKKLKLS
	. : : . * . * : * : . . : : . * . * : : :
Human_kinesin	LINVLKTDLLSSLEMILSPTVVSILKINSQLKHIFKTSLTVADKIEDQKKELDGFLSILC

Species	Sequence
Striga_hermonthica	-----ELKGNIRVFC
Striga_asiatica	ISAMETRSEFEEQKTILELQNCLADAESKIVE---GEKLRKKLHNTILELKGNIRVFC
Paulownia_fortunei	MSAMETRSEFEEQKTFILELQNRLADAELKIVE---GEKLRKKLHNTILELKGNIRVFC
Sesamum_indicum	MSAMETRSEFEEQKTIISQLQNRLADAESKIVE---GEQLRKKLHNTILELKGNIRVFC **.* : ::*
Human_kinesin	NNLHELQENTICSLVESQKQCGNLTEDLKTIKQTHSQELCKLMNLWTERFCALEEEKCENI
Striga_hermonthica	RVRPFLSDDGVGN-----NAKVVSFPPTSPEHWDEALIWHTTGKSFSLSLIKC
Striga_asiatica	RVRPLLSDDGVGN-----DAKVVSFPITSTETLGRGIDLAQNGQKHSFTFDKV
Paulownia_fortunei	RVRPLLSDDGVGI-----DAKVVSFPPTSMEALGRGIDLTQNGQKLSFTFDKV
Sesamum_indicum	RVRPLLSDDGVGA-----DTKVVSFPPTSMEAQGRGIDLTQNGQKLSFTFDKV . *.: : : *.: . * * :
Human_kinesin	QKPLSSVQENIQQKSKDIVNKMTFHSQKFCADSDGFSQELRNFNQEGTKLVEESVKHSDK
Striga_hermonthica	SCP-----MLLRKMFLW-----RSHSLFRV-----HLMA
Striga_asiatica	FMPDASQEDVFVEISQLVQSALDGY--KVCIFAYGQTGSGKTHMMGK-----PGLPD
Paulownia_fortunei	FLPDASQEDVFVEISQLVQSALDGY--KVCIFAYGQTGSGKTYTMMGK-----PGPPD
Sesamum_indicum	FVPDASQEDVFVEISQLVQSALDGY--KVCIFAYGQTGSGKTYTMMGK-----PAPID * : : ....
Human_kinesin	LNGNLEKISQETEQRCESLNTRTVYFSEQW----VSSLNEREQELHNLLEVVSQCCEAS
Striga_hermonthica	IRG-----LHGYMVRQAQGWKYDMRISMLEIYNETIRDLLAPNRTCSDAS
Striga_asiatica	QKGLIPRSLEQV-----FETRQILQAQGWKYDMQVSMLEIYNETIRDLLAP-----
Paulownia_fortunei	QKGLIPRSLEQV-----FETRQILQAQGWKYEMQVSMLEIYNETIRDLLAPNRSGFDAS
Sesamum_indicum	QKGLIPRSLEQV-----FETRQILQAQGWKYGMQVSMLEIYNETIRDLLAPNRSGFDAS . * : : : * : * * : : : *
Human_kinesin	SSDITEKSDGRKAAHEKQHNIF-LDQMTIDEDKLIAQNLE-LNETIKIGLTKLN----C
Striga_hermonthica	RAENAGKQYAIKHDANGNTQVFDLTVDVQSSKEVSYLLERAAQSRSVGKTQMNEQSSRS
Striga_asiatica	-----
Paulownia_fortunei	RAETGGKQYAIKHDANGNTHVSDLTIVDVRSSKEVSYLLDRAAQSRSVGKTQMNEQSSRS
Sesamum_indicum	RAENAGKQYSIKHDANGNTHVSDLTIVDVHSSKEVSYLLDRAAQSRSVGKTQMNEQSSRS
Human_kinesin	FLEQDLKLDIPTGTTPQRKSYLPSTLVRTEPREHLLDQLKRKQPELLMMLNCSENNKEE
Striga_hermonthica	HFVFTLRIMGFNENTDQQVC-VLNLIDLAGSERLSKSGSTGNQLKETQAINKSLSSLSDV
Striga_asiatica	-----
Paulownia_fortunei	HFVFTMRIMGFNESTDQQVQGVNLIDLAGSERLSKSGSTGDRLKETQAINKSLSSLSDV
Sesamum_indicum	HFVFTLRITGFNESTDQQVQGVNLIDLAGSERLSKSGSTGDRLKETQAINKSLSSLSDV
Human_kinesin	TIPDVVVEEAVLGQYTEEPLSQEPSVDAGVDCSSIGGVPFFQHKKSHGKDKENRGINTLE
Striga_hermonthica	IFALAKKEEHVPYRNSKLTYYLLQPCL--GGDSKTLMFVNVSPDHSLEGE-----SLCSLR
Striga_asiatica	-----
Paulownia_fortunei	IFALAKKEEHVPYRNSKLTYYLLQPCL--GGDSKTLMFVNVSPDPSSVGE-----SLCSLR
Sesamum_indicum	IFALAKKEEHVPYRNSKLTYYLLQPCL--GGDSKTLMFVNVSPDPSSVGE-----SLCSLR
Human_kinesin	RSKVEETTEHLVTKSRLPLRAQINL----
Striga_hermonthica	FAARVNACEIGVPRRQTNLRS-----
Striga_asiatica	-----
Paulownia_fortunei	FAARVNACEIGVPRRQTNLRSSDSRLSIG
Sesamum_indicum	FAARVNACEIGVPRRQTNLRSLSRLSIG

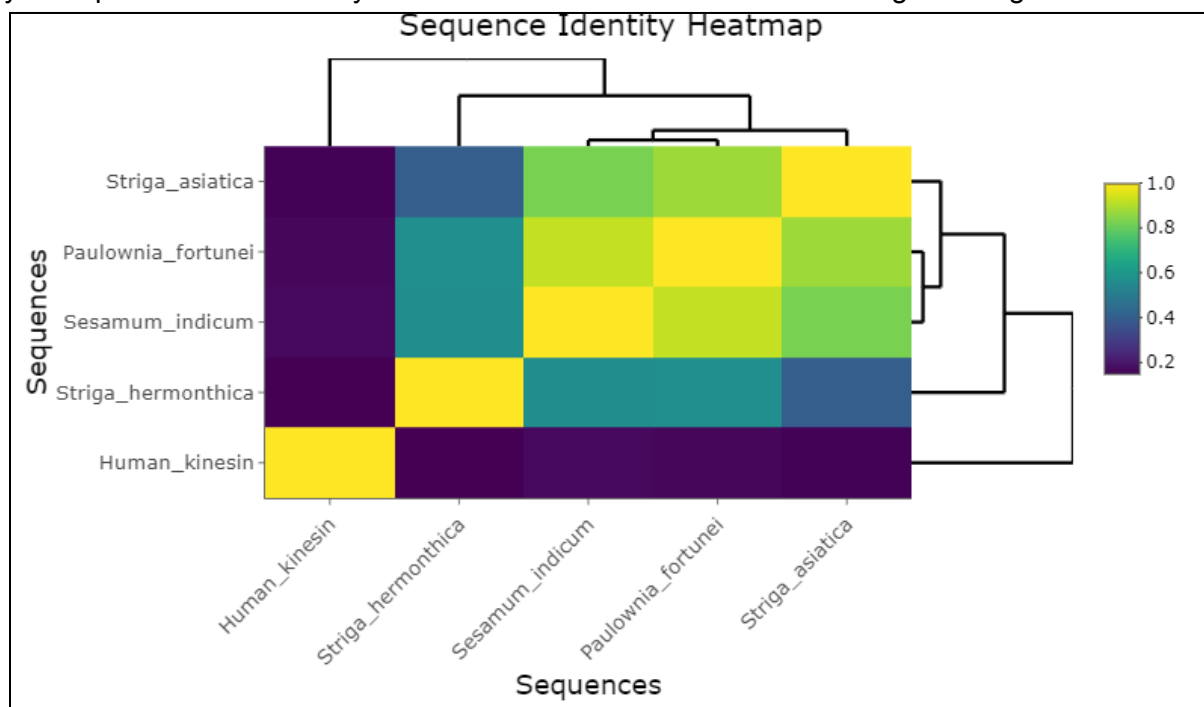
**[Q6]** Create a phylogenetic tree, using either a parsimony or distance-based approach. Bootstrapping and tree rooting are optional. Use “simple phylogeny” online from the EBI or any

respected phylogeny program (such as MEGA, PAUP, or Phylip). Paste an image of your Cladogram or tree output in your report.

Imported sequences into MEGA, aligned with MUSCLE, and created neighbor-joining tree:



**[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 it to your report. Do make sure your labels are visible and not cut at the figure margins.



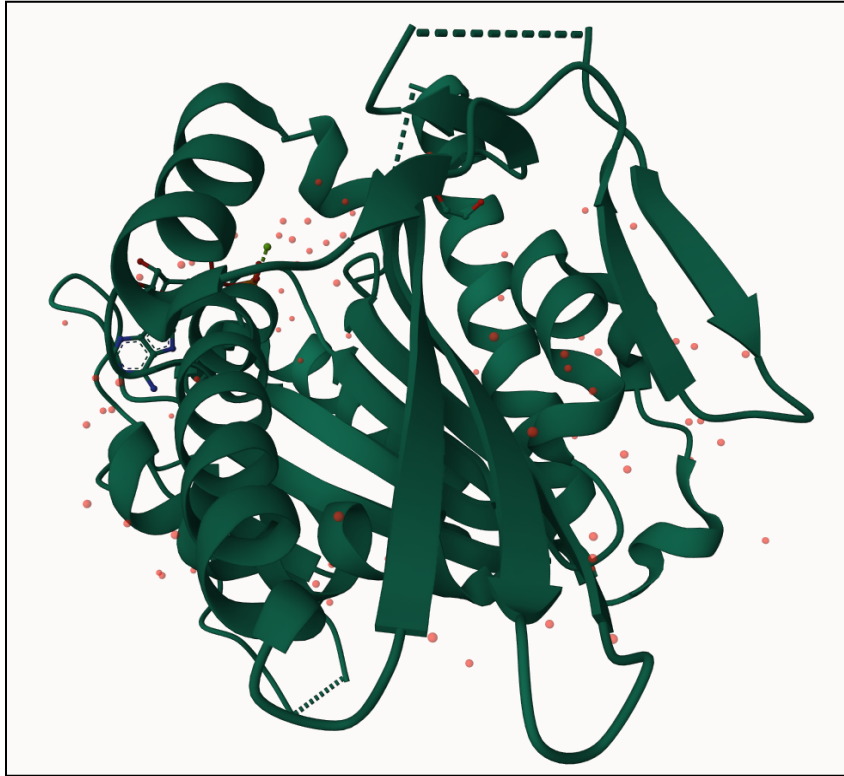
**[Q8]** Using R/Bio3D (or an online blast server if you prefer), search the main protein structure database for the most similar atomic resolution structures to your aligned sequences. List the top 3 unique hits (i.e. not hits representing different chains from the same structure) along with their Evalue 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 (experimental Technique), resolution (resolution), and source organism (source)

Paulownia fortunei has the highest percent matches across all species searched.

PDB-ID	Technique	Resolution	Source	E-value	%Identity
3T0Q	X-ray diffraction	2.35	Eremothecium gossypii	1.25e-97	47.6
4GKR	X-ray diffraction	2.69	Candida glabrata	6.91e-99	46.6
2NCD	X-ray diffraction	2.50	Drosophila melanogaster	2.19e-93	44.2

**[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?

3-D Molecular Structure of PDB structure with highest % identity: 3T0Q using Molstar:



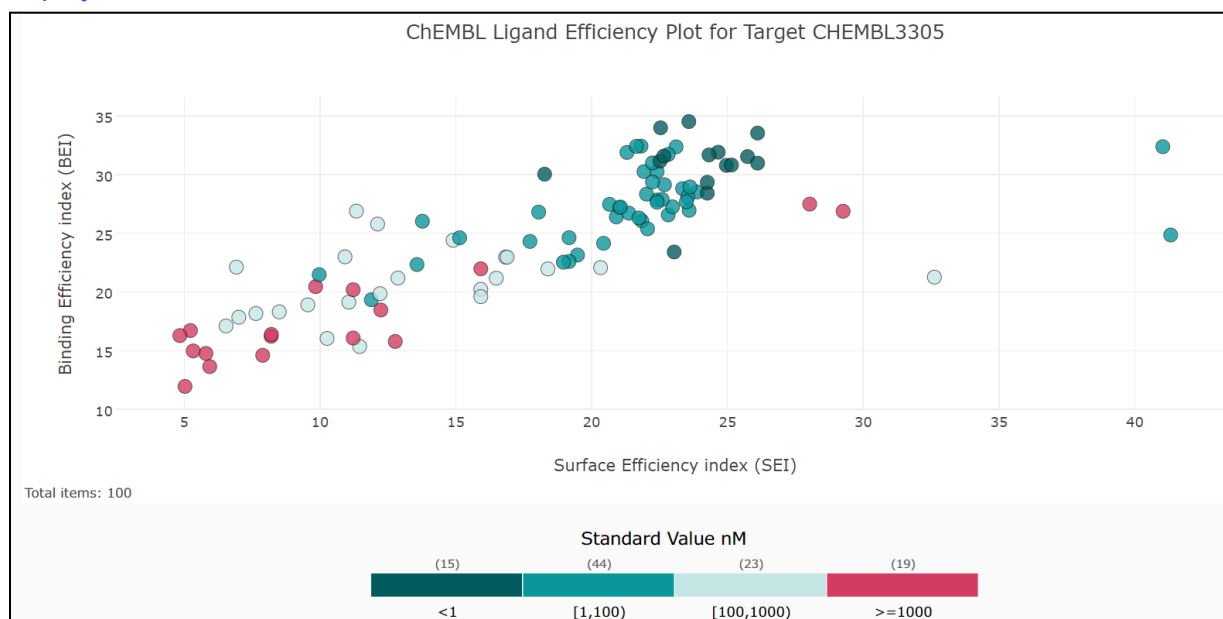
The above structure is the alpha chain of the “Motor Domain Structure of the Kar3-like kinesin from *Eremothecium gossypii*”. It is not likely to be similar in structure to the *Striga hermonthica*’s Kinesin-like protein, given its low sequence similarity (47.6%).

**[Q10]** Perform a “Target” search of ChEMBL ( <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?

Based on a ChEMBL Target search of novel protein sequence for striga hermonthica’s Kinesin-1 protein, the top search yielded:

[https://www.ebi.ac.uk/chembl/target\\_report\\_card/CHEMBL3305/](https://www.ebi.ac.uk/chembl/target_report_card/CHEMBL3305/)

CHEMBL3305, which details 26 Binding assays, 1 Functional Assay, and 7 ADME (absorption, distribution, metabolism, and excretion) Assays. There was ligand efficiency data available and displayed:



The ligand efficiency data is shown in this list:

[https://www.ebi.ac.uk/chembl/g/#browse/activities/filter/target\\_chembl\\_id%3ACHEMBL3305%20AND%20standard\\_type%3A\(IC50%20OR%20Ki%20OR%20EC50%20OR%20Kd\)%20AND%200\\_exists\\_%3Astandard\\_value%20AND%20\\_exists\\_%3Aligand\\_efficiency](https://www.ebi.ac.uk/chembl/g/#browse/activities/filter/target_chembl_id%3ACHEMBL3305%20AND%20standard_type%3A(IC50%20OR%20Ki%20OR%20EC50%20OR%20Kd)%20AND%200_exists_%3Astandard_value%20AND%20_exists_%3Aligand_efficiency)

One of the binding assays linked a journal article investigating the synthesis and biological component analysis of radiolabeled ligands for better identification and imaging of emerging androgen receptor-positive cancers and better targeting for therapy of the affected areas.

“Radiolabeled 5-Iodo-3'-O-(17 $\beta$ -succinyl-5 $\alpha$ -androstan-3-one)-2'-deoxyuridine and Its 5'-Monophosphate for Imaging and Therapy of Androgen Receptor-Positive Cancers: Synthesis and Biological Evaluation”

Zbigniew P. Kortylewicz\*, Jessica Nearman, and Janina Baranowska-Kortylewicz\*

<https://pubs.acs.org/doi/10.1021/jm9005803>