BGGN-213: FOUNDATIONS OF BIOINFORMATICS – FIND A GENE PROJECT
http://thegrantlab.org/bggn213

Name: Torrey Rhyne
PID: A14397504

Questions:

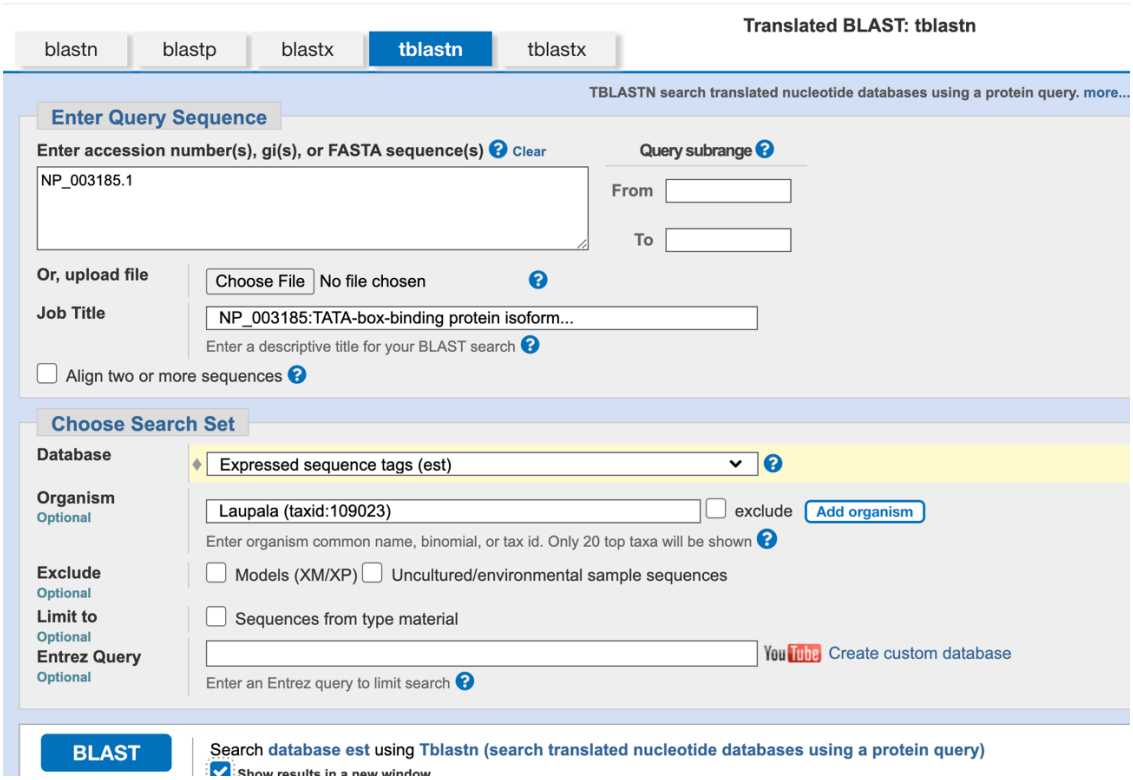
[Q1] Tell me the name of a protein you are interested in. Include the species and the accession number. This can be a human protein or a protein from any other species as long as it's function is known.
If you do not have a favorite protein, select human RBP4 or KIF11. Do not use beta globin as this is in the worked example report that I provide you with online.

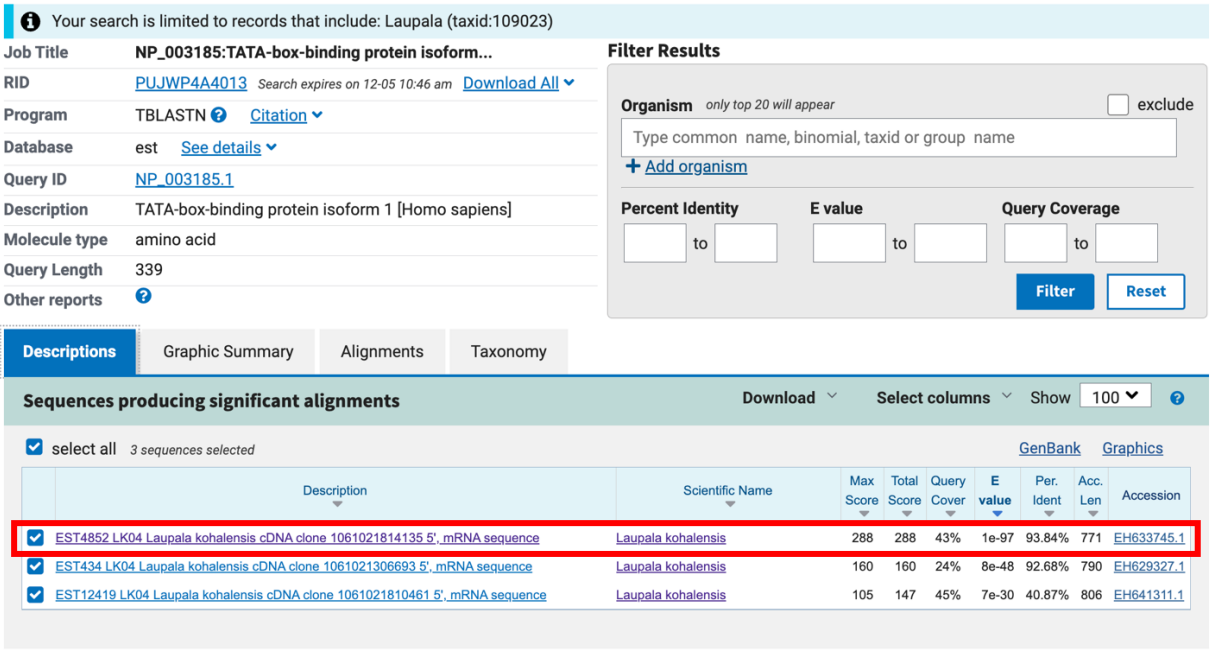
Name: TATA-box binding protein (TBP)
Species: Homo sapiens
Accession: NP_003185.1
Function: subunit of the basal transcription factor TFIID; binds promoter DNA at the TATA box; plays a major role in the initiation of transcription by RNA polymerase II.

[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 (2.14.1) search against Laupala ESTs
Database: Expressed sequence tags (est)
Organism: Laupala (taxid:109023)

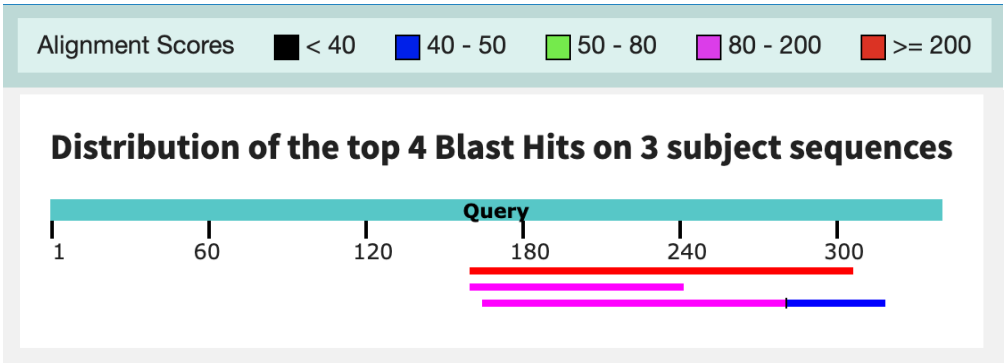
Also include the output of that BLAST search in your document. If appropriate, change the font to Courier size 10 so that the results are displayed neatly. You can also screen capture a BLAST output (e.g. alt print screen on a PC or on a MAC press ⌘-shift-4. The pointer becomes a bulls eye. Select the 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 EH633745.1 – a 771 base pair cDNA clone from Laupala kohalensis. See below for alignment.



Download GenBank Graphics

EST4852 LK04 Laupala kohalensis cDNA clone 1061021814135 5', mRNA sequence

Sequence ID: EH633745.1 **Length:** 771 **Number of Matches:** 1

Range 1: 304 to 741 GenBank Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
288 bits(736)	1e-97	Compositional matrix adjust.	137/146(94%)	140/146(95%)	0/146(0%)	+1

```
Query  160  GIVPQLQNIVSTVNLGCKLDLKTIALRARNAEYNPKRFAAVIMRIREPRTTALIFSSGKM  219
            IVPQLQNIVSTVNLGCKLDLK IAL ARNAEYNPKRFAAVIMRIREPRTTALIFSSGKM
Sbjct  304  AIVPQLQNIVSTVNLGCKLDLKKIALHARNAEYNPKRFAAVIMRIREPRTTALIFSSGKM  483

Query  220  VCTGAKSEEQSRLAARKYARVVQKLGFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHQQ  279
            VCTGAKSEE SRLAARKYAR++QKLGFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLTH Q
Sbjct  484  VCTGAKSEEDSRLAARKYARIIQKLGFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHGQ  663

Query  280  FSSYEPELFPGLIYRMIKPRIVLLIF  305
            FSSYEPELFPGLIYRM+KPRIVLLI
Sbjct  664  FSSYEPELFPGLIYRMVKPRIVLLIL  741
```

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. 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.

> EH633745.1 Laupala kohalensis protein sequence (used EBI EMBOSS Transeq)
GTPLHQPEEDQQILPQAQQQQQQNQQGTPLLQMPLLATPQKIMHTYAPSGFTTPQSL
MQPQTPQNMMSPMVASSSQIIPPSSLGPATPSPMTPMTPSSADPAIVPQLQNIVSTVN
LGCKLDLKKIALHARNAEYNPKRFAAVIMRIREPRTTALIFSSGKMVCTGAKSEEDSRLA
ARKYARIIQKLGFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHGQFSSYEPELFPGLIY
RMVKPRIVLLILYLESGLTVQS

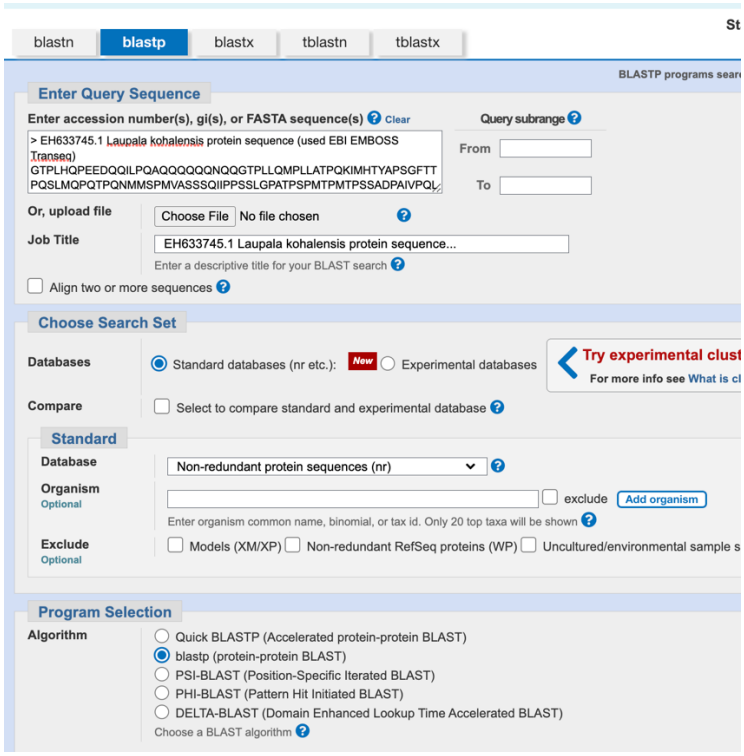
Here, tell me the name of the novel protein, and the species from which it derives. It is very unlikely (but still definitely possible) that you will find a novel gene from an organism such as *S. cerevisiae*, human or mouse, because those genomes have already been thoroughly annotated. It is more likely that you will discover a new gene in a genome that is currently being sequenced, such as bacteria or plants or protozoa.

Name: *Laupala* TBP
Species: Laupala kohalensis (a cricket)
Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Polyneoptera; Orthoptera; Ensifera; Gryllidea; Grylloidea; Gryllidae; Trigonidiinae; Laupala

[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.

The top hits from a BLASTP search against the NR database are TBP from many other species, but not Laupala. The top hit is TBP from Gryllus bimaculatus (another cricket). See below for setup, top hits, and alignment.



Sequences producing significant alignments Download Select columns Show 100

select all *0 sequences selected* GenPept Graphics Distance tree of results Multiple alignment MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
TATA-box-binding protein [Gryllus bimaculatus]	Gryllus bimaculatus	453	453	96%	6e-159	93.52%	294	GLH06850.1
hypothetical protein Cfor_03815 [Coptotermes formosanus]	Coptotermes formosanus	434	434	96%	2e-151	90.32%	270	GFG38212.1
TATA-box-binding protein [Cryptotermes secundus]	Cryptotermes secundus	434	434	96%	3e-151	89.68%	299	XP_023703955.1
TATA-box-binding protein [Athalia rosae]	Athalia rosae	417	417	96%	1e-144	82.95%	304	XP_012257098.1
TATA-box-binding protein [Neodiprion lecontei]	Neodiprion lecontei	417	417	96%	2e-144	83.20%	302	XP_015524603.1
TATA-box-binding protein [Diprion similis]	Diprion similis	416	416	96%	3e-144	82.88%	303	XP_046734297.1
unnamed protein product [Timema bartmani]	Timema bartmani	414	414	96%	3e-143	87.85%	292	CAD7449976.1
TATA-box-binding protein [Venturia canescens]	Venturia canescens	412	412	96%	2e-142	81.04%	316	XP_043276984.1
TATA-box-binding protein [Orussus abietinus]	Orussus abietinus	412	412	96%	2e-142	79.10%	311	XP_012281073.1
PREDICTED: TATA-box-binding protein [Fopius arisanus]	Fopius arisanus	411	411	96%	3e-142	83.27%	304	XP_011307885.1
TATA-box-binding protein [Ooceraea biroi]	Ooceraea biroi	411	411	96%	4e-142	83.33%	306	XP_011330203.1
TATA-box-binding protein isoform X1 [Belonocnema kinseyi]	Belonocnema kinseyi	410	410	96%	9e-142	82.76%	308	XP_033224405.1

TATA-box-binding protein [Gryllus bimaculatus]

Sequence ID: GLH06850.1 Length: 294 Number of Matches: 1

Range 1: 16 to 261 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
453 bits(1166)	6e-159	Compositional matrix adjust.	231/247(94%)	235/247(95%)	1/247(0%)

```
Query  1    GTPLHQPEEDQQILPQAQQQQQQNQQGTPLLQMPLLATPQKIMHTYAPSGFTTPQSLMQP  60
            GTPLHQPEEDQQILP AQQQQQ  QQGTPL QMPLL +PQK MHTYAP+GFTTPQSLMQP
Sbjct  16   GTPLHQPEEDQQILPMAQQQQQN-QQGTPLQQMPLLGSPQKTMHTYAPAGFTTPQSLMQP  74

Query  61   QTPQNMMSPMVASSSQIIPPSSLGPATPSPMTPMTPSSADPAIVPQLQNIVSTVNLGCKL  120
            QTP NMMSPMVAS+SQI+ PSSLGPATP PMTPMTPSSADP IVPQLQNIVSTVNL CKL
Sbjct  75   QTPHNMMSPMVASASQIVAPSSLGPATPGPMTPMTPSSADPGIVPQLQNIVSTVNLCCKL  134

Query  121  DLKKIALHARNAEYNPKRFAAVIMRIREPRTTALIFSSGKMVCTGAKSEEDSRLAARKYA  180
            DLKKIALHARNAEYNPKRFAAVIMRIREPRTTALIFSSGKMVCTGAKSEEDSRLAARKYA
Sbjct  135  DLKKIALHARNAEYNPKRFAAVIMRIREPRTTALIFSSGKMVCTGAKSEEDSRLAARKYA  194

Query  181  RIIQKLGFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHGQFSSYEPELFPGLIYRMVKP  240
            RIIQKLGFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHGQFSSYEPELFPGLIYRMVKP
Sbjct  195  RIIQKLGFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHGQFSSYEPELFPGLIYRMVKP  254

Query  241  RIVLLIL  247
            RIVLLI
Sbjct  255  RIVLLIF  261
```

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

Re-labeled sequences for alignment:

```
> Human_TBP | NP_003185.1 | TATA-box-binding protein isoform 1 [Homo
sapiens]
MDQNNSLPPYAQGLASPQGAMTPGIPIFSPMMPYGTGLTPQPIQNTNSLSILEEQQRQQQQQQQQQQQQ
QQQQQQQQQQQQQQQQQQQQQQQQQQQQAVAAAAVQQSTSQQATQGTSGQAPQLFHSQTLTTAPLPGTTPLYP
SPMTPMTPITPATPASESSGIVPQLQNIVSTVNLGCKLDLKTIALRARNAEYNPKRFAAVIMRIREPRTT
ALIFSSGKMVCTGAKSEEQSRLAARKYARVVQKLGFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHQQF
SSYEPELFPGLIYRMIKPRIVLLIFVSGKVVLTGAKVRAEIYEAFENIYPILKGFRKTT

> Cricket_(Laupala) | EH633745.1 | translated protein sequence from
EBI EMBOSS Transeq [Laupala kohalensis]
GTPLHQPEEDQQILPQAQQQQQQNQQGTPLLQMPLLATPQKIMHTYAPSGFTTPQSLMQPQTPQNMMSPM
VASSSQIIPPSSLGPATPSPMTPMTPSSADPAIVPQLQNIVSTVNLGCKLDLKKIALHARNAEYNPKRFA
AVIMRIREPRTTALIFSSGKMVCTGAKSEEDSRLAARKYARIIQKLGFPAKFLDFKIQNMVGSCDVKFPI
RLEGLVLTHGQFSSYEPELFPGLIYRMVKPRIVLLILYLESGLTVQS

> Cricket_(Gryllus) | GLH06850.1 | TATA-box-binding protein [Gryllus
bimaculatus]
```

MDQMLPSPGFSIPSIGTPLHQPEEDQQILPMAQQQQQNQQGTPLQQMPLLGSPQKTMHTYAPAGFTTPQS
LMQPQTPHNMMSPMVASASQIVAPSSLGPATPGPMTPMTPSSADPGIVPQLQNIVSTVNLCCKLDLKKIA
LHARNAEYNPKRFAAVIMRIREPRTTALIFSSGKMVCTGAKSEEDSRLAARKYARIIQKLGFPAKFLDFK
IQNMVGSCDVKFPIRLEGLVLTHGQFSSYEPELFPGLIYRMVKPRIVLLIFVSGKVVLTGAKVRQEIYEA
FDNIYPILKSFKKQ

> Termite | XP_023703955.1 | TATA-box-binding protein [Cryptotermes secundus]
MDQMLPSPGFSIPSIGTPLHQPEEDQQILPHAQQQQQQQQQQLPPPPTSLQQMPVLSTPQKTMHTYAPSG
FTPQSLMQPQTPQNLMSPMVTPSSQMATLSNMGPATPSPMTPMTPSSADPGIIPQLQNIVSTVNLGCKLD
LKKIALHARNAEYNPKRFAAVIMRIREPRTTALIFSSGKMVCTGAKSEEDSRLAARKYARIIQKLGFPAK
FLDFKIQNMVGSCDVKFPIRLEGLVLTHGQFSSYEPELFPGLIYRMVKPRIVLLIFVSGKVVLTGAKVRQ
EIYEAFDNIYPILKSFKKQ

> Sawfly_(Athalia) | XP_012257098.1 | TATA-box-binding protein [Athalia rosae]
MDQMLPSPGFSIPSIGTPLHQPEEDQQILPNALQQQQQQVQQSHQIQQMHPMGSNIHGMPMLATPHKTMH
TYTPAFATPQSLMQPQTPQNMMSPMVQPASQIAPPSIGPSTPGPMTPMTPASADPGILPQLQNIVSTVNL
NCKLELKKIALHARNAEYNPKRFAAVIMRIREPRTTALIFSSGKMVCTGAKSEEDSRLAARKYARIIQKL
GFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHGQFSSYEPELFPGLIYRMVKPRIVLLIFVSGKVVLTG
AKVRQEIYEAFDNIYPILKSFKKQ

> Sawfly_(Neodiprion) | XP_015524603.1 | TATA-box-binding protein [Neodiprion lecontei]
MDQMLPSPGFSIPSIGTPLHQPEEDQQILPNALQQQQQVQQNQLQQMHPMGSNVHGMPMLGTPHKTMHTY
APAFATPQSLMQPQTPQNMMSPMVQPASQIAPPSIGPSTPGPMTPMTPASADPGILPQLQNIVSTVNLNC
KLNLKEIALHARNAEYNPKRFAAVIMRIREPRTTALIFSSGKMVCTGAKSEEDSRLAARKYARIIQKLGF
PAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHGQFSSYEPELFPGLIYRMVKPRIVLLIFVSGKVVLTGAK
VRQEIYEAFDNIYPILKSFKKQ

> Wasp_(Venturia) | XP_043276984.1 | TATA-box-binding protein [Venturia canescens]
MDQMLPSPGFSIPSIGTPLHQPEEDQQILPNALQQQQQQQQQQQQQQTSQQYQMQQLHSMSPTMQANSMM
MLSTPQKTMHTYAPTPTFATPQSLMQPQTPQNMMSPMVQPTSQIAPSSIGPSTPGPMTPMTPASADPGIL
PQLQNIVSTVNLTCKLDLKKIALHARNAEYNPKRFAAVIMRIREPRTTALIFSSGKMVCTGAKSEEDSRL
AARKYARIIQKLGFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHGQFSSYEPELFPGLIYRMVKPRIVL
LIFVSGKVVLTGAKVRQEIYEAFDNIYPILKSFKKQ

> Wasp_(Orussus) | XP_012281073.1 | TATA-box-binding protein [Orussus abietinus]
MDQMLPSPGFSIPSIGTPLHQPEEDQQILPNALQQQQQQQQQQQSQQYQLQQLHSMSPNMQGSMLMISTP
QKTMHTYAPTPAFATPQSLMQPQTPQNMMSPIVQPSSQIAPSSIGPSTPGPMTPMTPASADPGILPQLQN
IVSTVNLNCKLDLKKIALHARNAEYNPKRFAAVIMRIREPRTTALIFSSGKMVCTGAKSEEDSRLAARKY
ARIIQKLGFAAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHGQFSSYEPELFPGLIYRMVKPRIVLLIFVS
GKVVLTGAKVRQEIYEAFDNIYPILKSFKKQ

> Ant | XP_011330203.1 | TATA-box-binding protein [Ooceraea biroi]
MDQMLPSPGGFSIPSIGTPLHQPEEDQQILPNALQQQQQSQQYQLQQLHSMSPNMQSGMLMITTPQKTMH
TYAPTPTFATPQSLMQPQTPQNMMSPIVPQSNQIAPPSIGPATPGPMTPMTPASADPGILPQLQNIVSTV
NLNCKLDLKKIALHARNAEYNPKRFAAVIMRIREPRTTALIFSSGKMVCTGAKSEADSRLAARKYARIIQ

KLGFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLSHGQFSSYEPELFPGLIYRMVKPRIVLLIFVSGKVVL
TGAKVRKEIYEAFDNIYPILKSFKKQ

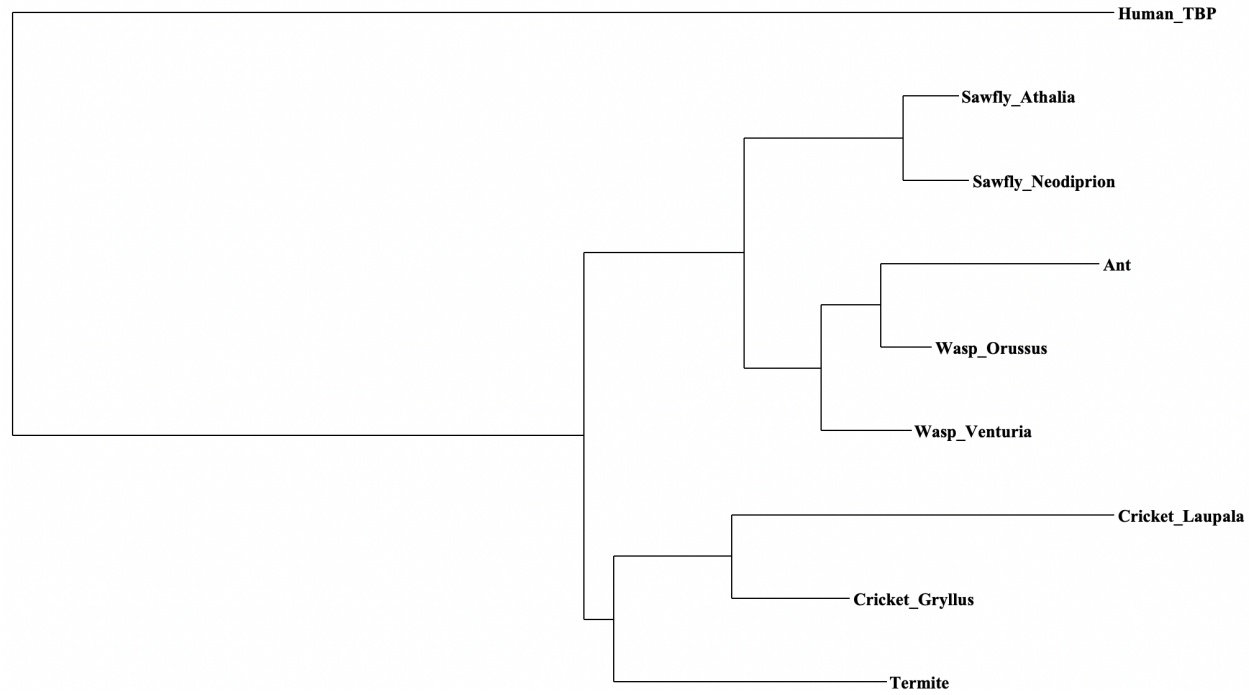
Alignment:

```
CLUSTAL multiple sequence alignment by MUSCLE (3.8)


Human_TBP               MDQNNSLPPYAQGLASPQGAMTPGIPIFSPMMPYGTGLTPQPIQNTNSLSILEEQQRQQQ
Cricket_(Laupala)       ----------------------GTPLHQP----------------------EEDQQILP
Sawfly_(Athalia)        MDQ---MLPSP-GFSIP----SIGTPLHQP----------------------EEDQQILP
Sawfly_(Neodiprion)     MDQ---MLPSP-GFSIP----SIGTPLHQP----------------------EEDQQILP
Ant                     MDQ---MLPSPGGFSIP----SIGTPLHQP----------------------EEDQQILP
Wasp_(Venturia)         MDQ---MLPSP-GFSIP----SIGTPLHQP----------------------EEDQQILP
Wasp_(Orussus)          MDQ---MLPSP-GFSIP----SIGTPLHQP----------------------EEDQQILP
Cricket_(Gryllus)       MDQ---MLPSP-GFSIP----SIGTPLHQP----------------------EEDQQILP
Termite                 MDQ---MLPSP-GFSIP----SIGTPLHQP----------------------EEDQQILP
                                             * *:..*                      **:*.


Human_TBP               QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQA--VAAAAVQQSTSQQATQGTSGQA
Cricket_(Laupala)       QAQQQQQQ-------------------NQQGTPLLQ--MPLLATPQKIMHTYAPS-GFTT
Sawfly_(Athalia)        NALQQQQQ--------QVQQSHQIQQMHPMGSNIHG--MPMLATPHKTMHTYTPA--FAT
Sawfly_(Neodiprion)     NALQQQQQ----------VQQNQLQQMHPMGSNVHG--MPMLGTPHKTMHTYAPA--FAT
Ant                     NALQQQQQ----------SQQYQLQQLHSMSPNMQS-GMLMITTPQKTMHTYAPTPTFAT
Wasp_(Venturia)         NALQQQQQQQQQQQQQQTSQQYQMQQLHSMSPTMQANSMMMLSTPQKTMHTYAPTPTFAT
Wasp_(Orussus)          NALQQQQQ----QQQQQQQSQQYQLQQLHSMSPNMQG-SMLMISTPQKTMHTYAPTPAFAT
Cricket_(Gryllus)       MAQQQQQN-------------------QQGTPLQQ--MPLLGSPQKTMHTYAPA-GFTT
Termite                 HAQQQQQQ--------------QQQQLPPPPTSLQQ--MPVLSTPQKTMHTYAPS-GF-T
                           ****:                              :      :.  :  :      :


Human_TBP               PQ-LFHSQT---LTTAPLPGTTPL-YPSPMTPMT--PITPATPASESSGIVPQLQNIVST
Cricket_(Laupala)       PQSLMQPQTPQNMMSPMVASSSQIIPPSSLGPATPSPMTPMTPSSADPAIVPQLQNIVST
Sawfly_(Athalia)        PQSLMQPQTPQNMMSPMVQPASQI-APPSIGPSTPGPMTPMTPASADPGILPQLQNIVST
Sawfly_(Neodiprion)     PQSLMQPQTPQNMMSPMVQPASQI-APPSIGPSTPGPMTPMTPASADPGILPQLQNIVST
Ant                     PQSLMQPQTPQNMMSPIVPQSNQI-APPSIGPATPGPMTPMTPASADPGILPQLQNIVST
Wasp_(Venturia)         PQSLMQPQTPQNMMSPMVQPTSQI-APSSIGPSTPGPMTPMTPASADPGILPQLQNIVST
Wasp_(Orussus)          PQSLMQPQTPQNMMSPIVQPSSQI-APSSIGPSTPGPMTPMTPASADPGILPQLQNIVST
Cricket_(Gryllus)       PQSLMQPQTPHNMMSPMVASASQIVAPSSLGPATPGPMTPMTPSSADPGIVPQLQNIVST
Termite                 PQSLMQPQTPQNLMSPMVTPSSQMATLSNMGPATPSPMTPMTPSSADPGIIPQLQNIVST
                        ** *::.**   : :. :  :. :   . : * *  *:** **:* ...*:*********


Human_TBP               VNLGCKLDLKTIALRARNAEYNPKRFAAVIMRIREPRTTALIFSSGKMVCTGAKSEEQSR
Cricket_(Laupala)       VNLGCKLDLKKIALHARNAEYNPKRFAAVIMRIREPRTTALIFSSGKMVCTGAKSEEDSR
Sawfly_(Athalia)        VNLNCKLELKKIALHARNAEYNPKRFAAVIMRIREPRTTALIFSSGKMVCTGAKSEEDSR
Sawfly_(Neodiprion)     VNLNCKLNLKEIALHARNAEYNPKRFAAVIMRIREPRTTALIFSSGKMVCTGAKSEEDSR
Ant                     VNLNCKLDLKKIALHARNAEYNPKRFAAVIMRIREPRTTALIFSSGKMVCTGAKSEADSR
Wasp_(Venturia)         VNLTCKLDLKKIALHARNAEYNPKRFAAVIMRIREPRTTALIFSSGKMVCTGAKSEEDSR
Wasp_(Orussus)          VNLNCKLDLKKIALHARNAEYNPKRFAAVIMRIREPRTTALIFSSGKMVCTGAKSEEDSR
Cricket_(Gryllus)       VNLCCKLDLKKIALHARNAEYNPKRFAAVIMRIREPRTTALIFSSGKMVCTGAKSEEDSR
Termite                 VNLGCKLDLKKIALHARNAEYNPKRFAAVIMRIREPRTTALIFSSGKMVCTGAKSEEDSR
                        *** ***:** ***.****************************************** :**


Human_TBP               LAARKYARVVQKLGFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHQQFSSYEPELFPGL
Cricket_(Laupala)       LAARKYARIIQKLGFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHGQFSSYEPELFPGL
Sawfly_(Athalia)        LAARKYARIIQKLGFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHGQFSSYEPELFPGL
Sawfly_(Neodiprion)     LAARKYARIIQKLGFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHGQFSSYEPELFPGL
Ant                     LAARKYARIIQKLGFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLSHGQFSSYEPELFPGL
Wasp_(Venturia)         LAARKYARIIQKLGFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHGQFSSYEPELFPGL
Wasp_(Orussus)          LAARKYARIIQKLGFAAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHGQFSSYEPELFPGL
Cricket_(Gryllus)       LAARKYARIIQKLGFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHGQFSSYEPELFPGL
Termite                 LAARKYARIIQKLGFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHGQFSSYEPELFPGL
                        ********::*****.****************************:* *************


Human_TBP               IYRMIKPRIVLLIFVSGKVVLTGAKVRAEIYEAFENIYPILKGFRKT
Cricket_(Laupala)       IYRMVKPRIVLLILYLES----GLTVQS-------------------
```

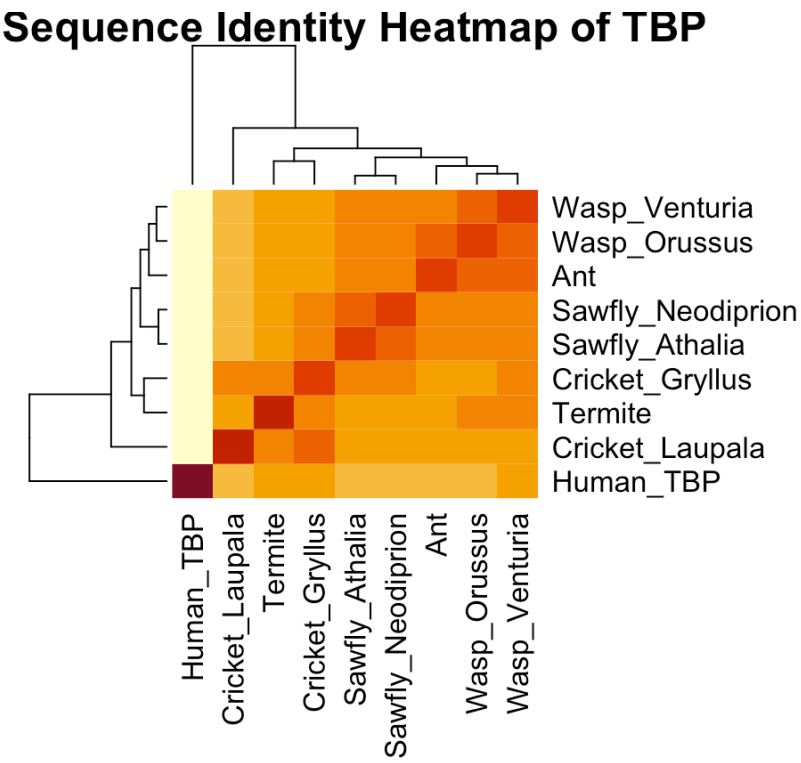
```
Sawfly_(Athalia)          IYRMVKPRIVLLIFVSGKVVLTGAKVRQEIYEAFDNIYPILKSFKKQ
Sawfly_(Neodiprion)       IYRMVKPRIVLLIFVSGKVVLTGAKVRQEIYEAFDNIYPILKSFKKQ
Ant                       IYRMVKPRIVLLIFVSGKVVLTGAKVRKEIYEAFDNIYPILKSFKKQ
Wasp_(Venturia)           IYRMVKPRIVLLIFVSGKVVLTGAKVRQEIYEAFDNIYPILKSFKKQ
Wasp_(Orussus)            IYRMVKPRIVLLIFVSGKVVLTGAKVRQEIYEAFDNIYPILKSFKKQ
Cricket_(Gryllus)         IYRMVKPRIVLLIFVSGKVVLTGAKVRQEIYEAFDNIYPILKSFKKQ
Termite                   IYRMVKPRIVLLIFVSGKVVLTGAKVRQEIYEAFDNIYPILKSFKKQ
                          ****:********:   .    * .*.
```

[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.

Aligned with MUSCLE and created a distance-based phylogenetic 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 to your report. Do make sure your labels are visible and not cut at the figure margins.



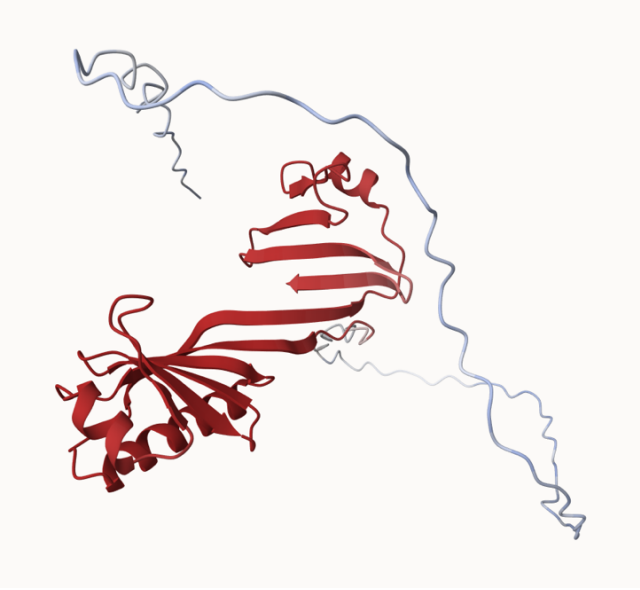
[Q8] Using R/Bio3D (or an online blast server if you prefer), search the main protein structure database for the most similar atomic resolution structures to your aligned sequences.
List the top 3 *unique* hits (i.e. not hits representing different chains from the same structure) along with their Evalue 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).
HINT: You can use a single sequence from your alignment or generate a consensus sequence from your alignment using the Bio3D function consensus(). The Bio3D functions blast.pdb(), plot.blast() and pdb.annotate() are likely to be of most relevance for completing this task. Note that the results of blast.pdb() contain the hits PDB identifier (or pdb.id) as well as Evalue and identity. The results of pdb.annotate() contain the other annotation terms noted above.
Note that if your consensus sequence has lots of gap positions then it will be better to use an original sequence from the alignment for your search of the PDB. In this case you could chose the sequence with the highest identity to all others in your alignment by calculating the row-wise maximum from your sequence identity matrix.

The consensus sequence has lots of gaps (mostly caused by human TBP), so I chose the sequence most similar to all the others in the alignment (Wasp_Venturia).
Below are the top 3 unique hits from different species:

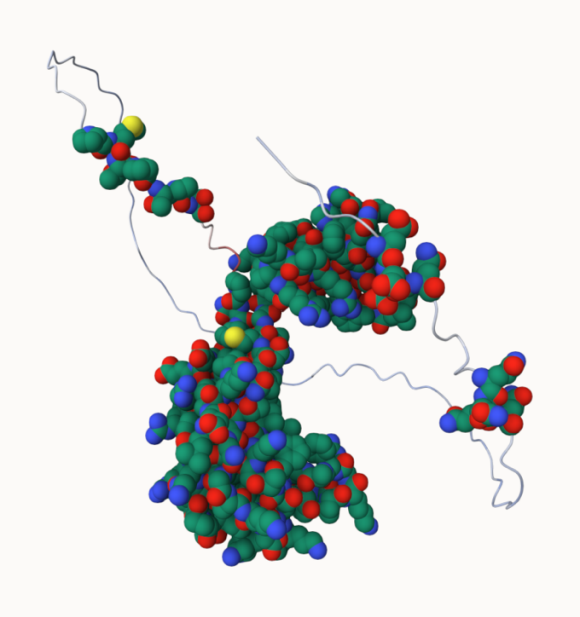
ID	Technique	Resolution	Source	E-value	Identity
5FUR_A	electron microscopy	8.5	Homo sapiens	2.14 e-125	86.34
1RM1_A	X-ray diffraction	2.5	Saccharomyces cerevisiae	9.24 e-111	80.77
7Z7N_D	electron microscopy	5.1	Thermochaetoides thermophila	1.78e-110	74.52

[Q9] Using AlphaFold notebook generate a structural model using the default parameters for your novel protein sequence.
Note that this can take some time depending upon your sequence length. If your model is taking many hours to generate or your input sequence yields a “too many amino acids” (i.e. length) error you can focus on a single domain from your sequence - identify region by searching for PFAM domain matches.
Once complete save the resulting PDB format file for your records. Finally, generate a molecular figure of your generated PDB structure using the Mol* viewer online (or VMD/PyMol/Chimera if you prefer). To complete your analysis you can optionally highlight *conserved residues* that are likely to be functional as spacefill and the protein as cartoon colored by local alpha fold *pLDDT quality score*. This score is contained in the B-factor column of your PDB downloaded file. Please use a white or transparent background for your figure (i.e. not the default black in PyMol/VMD/Chimera etc.).

Predicted structure of *Laupala kohalensis* TBP, colored by AlphaFold pLDDT quality score.

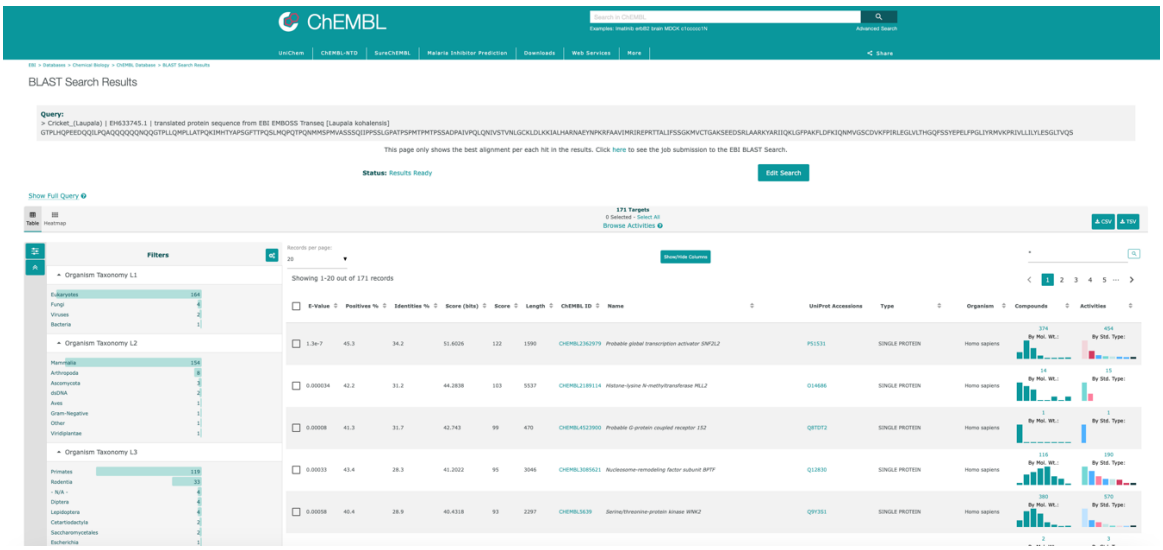


Conserved residues (based on my MSA) highlighted with spacefill. There are a few more conserved residues in the disordered region, but it was taking a long time to add them all as “components” in Mol* viewer (TBP is very conserved!).



[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? If there are no assays listed here simply list “non available as of [date]”.

My search (novel Laupala TBP sequence, default parameters, 12/6/23) generated 171 hits.



However, none of the hits are targeting TBP, even when I search the human sequence or just “TBP”.

The top hit is targeting SNF2L2, a chromatin remodeling complex subunit (CHEMBL2362979).

There are 80 associated assays, 79 binding and 1 functional.
Description of the top binding assay (CHEMBL4276258): Binding affinity to N-terminal MBP-fused human BRM RecA domain E852Q mutant expressed in Escherichia coli BL21 Star (DE3) by ITC method.

Ligand efficiency data: