BIMM-143: INTRODUCTION TO BIOINFORMATICS

The find-a-gene project assignment

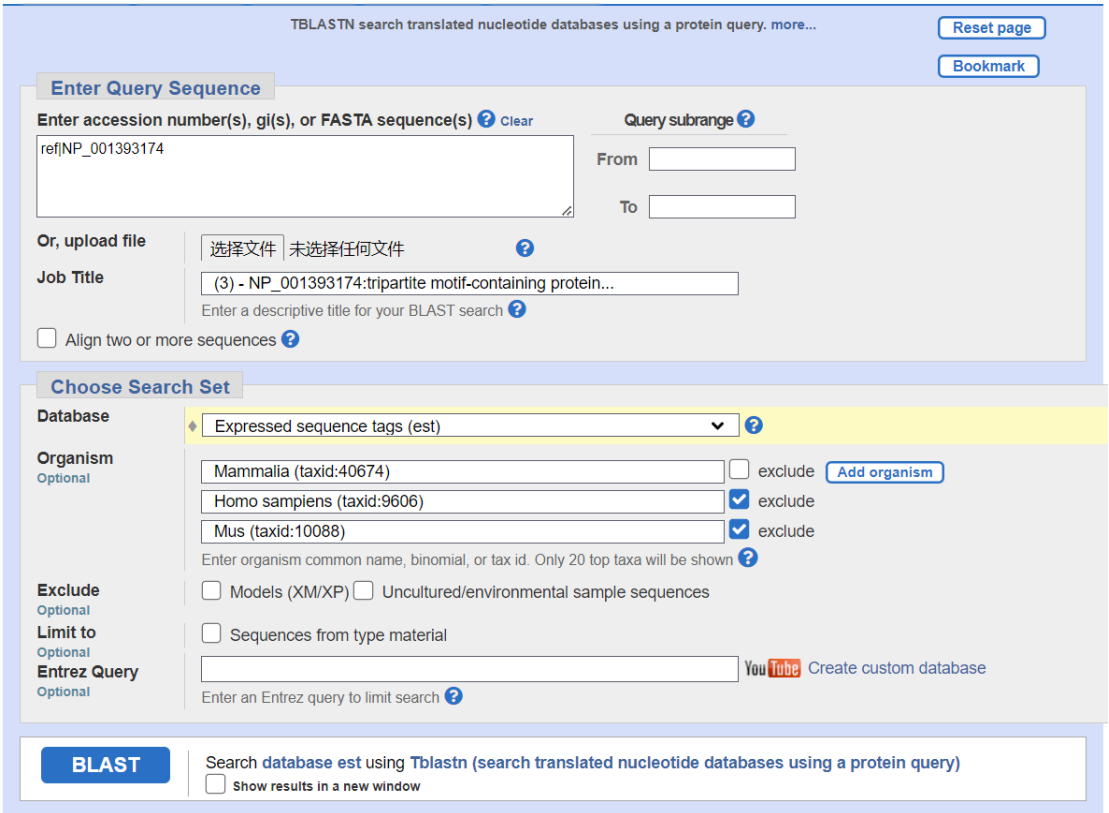
Yufei Liu (A16222438)

1. 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: tripartite motif-containing protein 46 isoform 3 (TRIM46 protein)
   Accession: NP_001393174
   Species: Homo Sapiens

2. 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 Mammalia (excluding Human and Mouse) ESTs
   Database: EST
   Organism: Mammalia (Taxid: 40674); excluding Homo sapiens (Taxid: 9606) and Mus (Taxid: 10088)



Chosen match: Accession FY641334.1, an 892 base pair clone from *Notamacropus eugenii*. See below for alignment details.

Your search is limited to records that include: Mammalia (taxid:40674) ; and exclude: Homo sampiens (taxid:9606), Mus (taxid:10088)

Edit and Resubmit | Save Search Strategies | ▸Formatting options | ▸Download

Job title: | How to read this page | Blast report description | Click here to use the new BLAST results page

NP_001393174:tripartite motif-containing protein...

RID K372MZTE013 (Expires on 10-21 05:15 am)

Query ID NP_001393174.1
Description tripartite motif-containing protein 46 isoform 3 [Homo sapiens]
Molecule type amino acid
Query Length 788

Database Name est
Description Database of GenBank+EMBL+DDBJ sequences from EST Divisions
Program TBLASTN 2.14.1+ ▸Citation

Other reports: ▸Search Summary [Taxonomy reports]

Graphic Summary

Distribution of the top 106 Blast Hits on 100 subject sequences

Mouse over to see the title, click to show alignments

Questions/comments

Descriptions

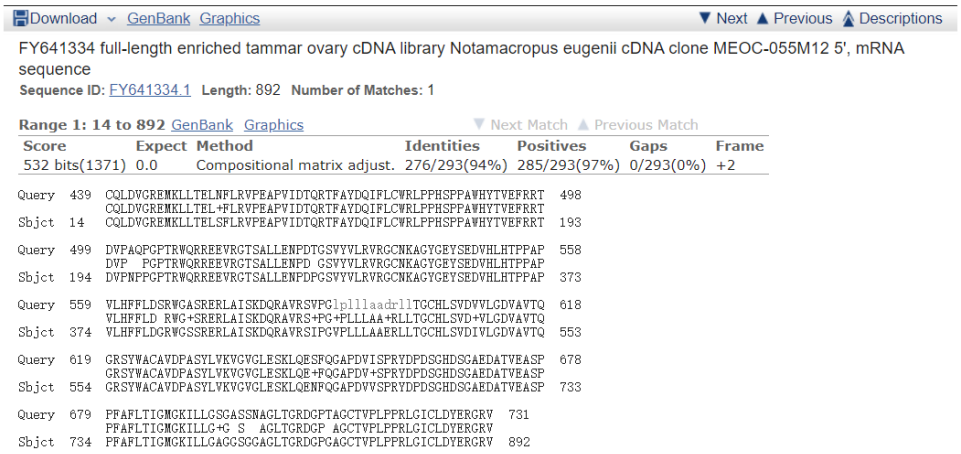
Sequences producing significant alignments:

Select: All None Selected:0

Alignments | Download | GenBank Graphics

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☐	LB02738.CR.1_L10 GC_BGC-27 Bos taurus cDNA clone IMAGE:8624220 5', mRNA sequenc	577	577	37%	0.0	94.88%	EH207649.1
☐	FY641334 full-length enriched tammar ovary cDNA library Notamacropus eugenii cDNA clone	532	532	37%	0.0	91.13%	FY641334.1
☐	FY683205 full-length enriched tammar testis cDNA library Notamacropus eugenii cDNA clone	511	511	35%	4e-176	90.46%	FY683205.1
☐	FY541103 full-length enriched tammar gravid uterus cDNA library Notamacropus eugenii cDN	492	492	32%	9e-169	90.73%	FY541103.1
☐	FY679234 full-length enriched tammar testis cDNA library Notamacropus eugenii cDNA clone	486	486	32%	1e-166	90.62%	FY679234.1
☐	FY732654 full-length enriched tammar testis cDNA library Notamacropus eugenii cDNA clone	479	479	31%	9e-164	91.20%	FY732654.1
☐	LB02734.CR.2_H17 GC_BGC-27 Bos taurus cDNA clone IMAGE:8622595 5', mRNA sequenc	474	548	35%	3e-161	91.63%	EH206451.1
☐	FY731673 full-length enriched tammar testis cDNA library Notamacropus eugenii cDNA clone	449	449	30%	2e-152	90.72%	FY731673.1
☐	FS640286 full-length enriched swine cDNA library, adult brain (frontal lobe) Sus scrofa cDNA	439	439	35%	5e-148	79.42%	FS640286.1
☐	FY732421 full-length enriched tammar testis cDNA library Notamacropus eugenii cDNA clone	404	404	28%	2e-134	86.61%	FY732421.1
☐	FY710744 full-length enriched tammar testis cDNA library Notamacropus eugenii cDNA clone	400	400	27%	4e-133	86.10%	FY710744.1
☐	130996 MARC 1PIG Sus scrofa cDNA 5', mRNA sequence	398	398	25%	6e-133	98.50%	BE032156.1
☐	FY662790 full-length enriched tammar testis cDNA library Notamacropus eugenii cDNA clone	394	394	26%	2e-131	89.47%	FY662790.1
☐	FQ087797 Rattus norvegicus brain Sprague-Dawley Rattus norvegicus cDNA clone TL0AAA7	395	395	32%	3e-131	76.65%	FQ087797.1
☐	HX201890 full-length enriched swine cDNA library, adult brain (frontal lobe) Sus scrofa cDNA	388	388	32%	1e-128	77.47%	HX201890.1
☐	FY654415 full-length enriched tammar testis cDNA library Notamacropus eugenii cDNA clone	387	387	25%	1e-128	90.64%	FY654415.1
☐	FQ097952 Rattus norvegicus brain Sprague-Dawley Rattus norvegicus cDNA clone TL0AAA5	377	377	31%	3e-124	75.60%	FQ097952.1

Alignment Details:



Download GenBank Graphics ▼ Next ▲ Previous Descriptions

FY641334 full-length enriched tammar ovary cDNA library Notamacropus eugenii cDNA clone MEOC-055M12 5', mRNA sequence

Sequence ID: FY641334.1 **Length:** 892 **Number of Matches:** 1

Range 1: 14 to 892 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
532 bits(1371)	0.0	Compositional matrix adjust.	276/293(94%)	285/293(97%)	0/293(0%)	+2

```
Query  439  CQLDVGREMKLLTELNFLRVPEAPVIDTQRTFAYDQIFLCWRLPPHSPPAWHYTVEFRRT  498
            CQLDVGREMKLLTEL+FLRVPEAPVIDTQRTFAYDQIFLCWRLPPHSPPAWHYTVEFRRT
Sbjct  14   CQLDVGREMKLLTELSFLRVPEAPVIDTQRTFAYDQIFLCWRLPPHSPPAWHYTVEFRRT  193

Query  499  DVPAQPGPTRWQRREEVRGTSALLENPDTGSVYVLRVRGCNKAGYGEYSEDVHLHTPPAP  558
            DVP  PGPTRWQRREEVRGTSALLENPD GSVYVLRVRGCNKAGYGEYSEDVHLHTPPAP
Sbjct  194  DVPNPPGPTRWQRREEVRGTSALLENPDPGSVYVLRVRGCNKAGYGEYSEDVHLHTPPAP  373

Query  559  VLHFFLDSRWGASRERLAISKDQRAVRSVPGlplllaadrllTGCHLSVDVVLGDVAVTQ  618
            VLHFFLD RWG+SRERLAISKDQRAVRS+PG+PLLLAA+RLLTGCHLSVD+VLGDVAVTQ
Sbjct  374  VLHFFLDGRWGSSRERLAISKDQRAVRSIPGVPLLLAAERLLTGCHLSVDIVLGDVAVTQ  553

Query  619  GRSYWACAVDPASYLVKVGVGLESKLQESFQGAPDVISPRYDPDSGHDSGAEDATVEASP  678
            GRSYWACAVDPASYLVKVGVGLESKLQE+FQGAPDV+SPRYDPDSGHDSGAEDATVEASP
Sbjct  554  GRSYWACAVDPASYLVKVGVGLESKLQENFQGAPDVVSPRYDPDSGHDSGAEDATVEASP  733

Query  679  PFAFLTIGMGKILLGSGASSNAGLTGRDGPTAGCTVPLPPRLGICLDYERGRV  731
            PFAFLTIGMGKILLG+G S  AGLTGRDGP AGCTVPLPPRLGICLDYERGRV
Sbjct  734  PFAFLTIGMGKILLGAGGSGGAGLTGRDGPGAGCTVPLPPRLGICLDYERGRV  892
```

3. 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:

   ```
   >Notamacropus eugenii protein (sequence translated by EMBOSS Transeq)
   GLLDCQLDVGREMKLLTELSFLRVPEAPVIDTQRTFAYDQIFLCWRLPPHSPPAWHYTVE
   FRRTDVPNPPGPTRWQRREEVRGTSALLENPDPGSVYVLRVRGCNKAGYGEYSEDVHLHT
   PPAPVLHFFLDGRWGSSRERLAISKDQRAVRSIPGVPLLLAAERLLTGCHLSVDIVLGDV
   AVTQGRSYWACAVDPASYLVKVGVGLESKLQENFQGAPDVVSPRYDPDSGHDSGAEDATV
   EASPPFAFLTIGMGKILLGAGGSGGAGLTGRDGPGAGCTVPLPPRLGICLDYERGRV
   ```

   Name: Notamacropus eugenii TRIM46

   Species: Notamacropus eugenii (tammar wallaby).

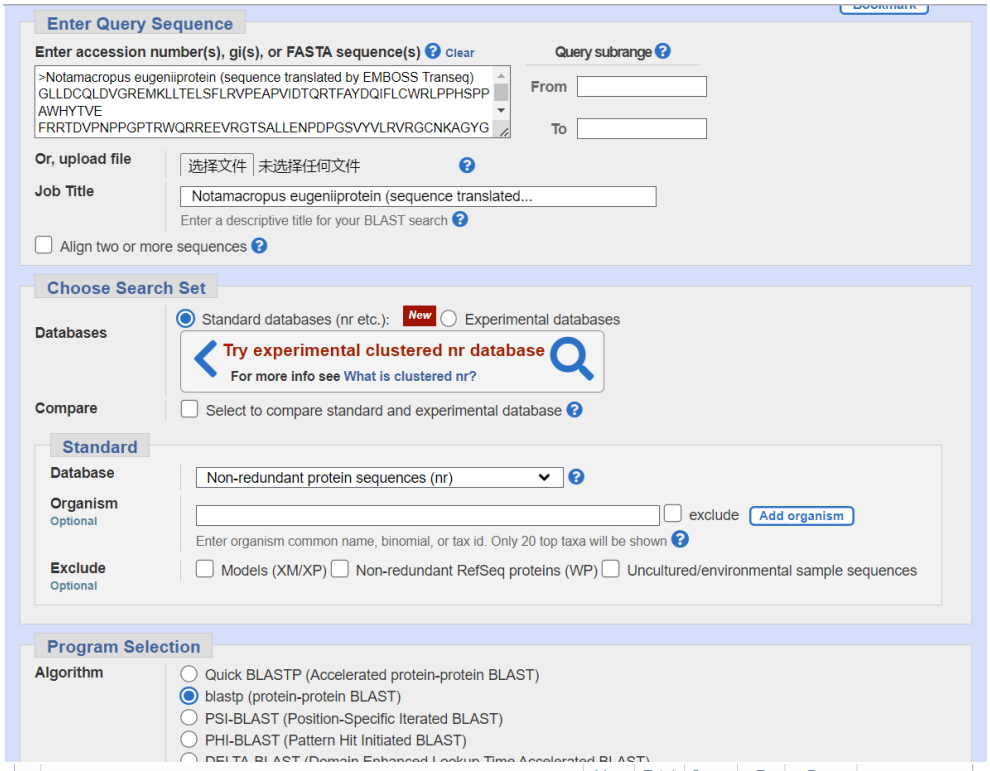
   cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amniota; Mammalia; Theria; Metatheria; Diprotodontia; Macropodidae; Notamacropus

4. 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.

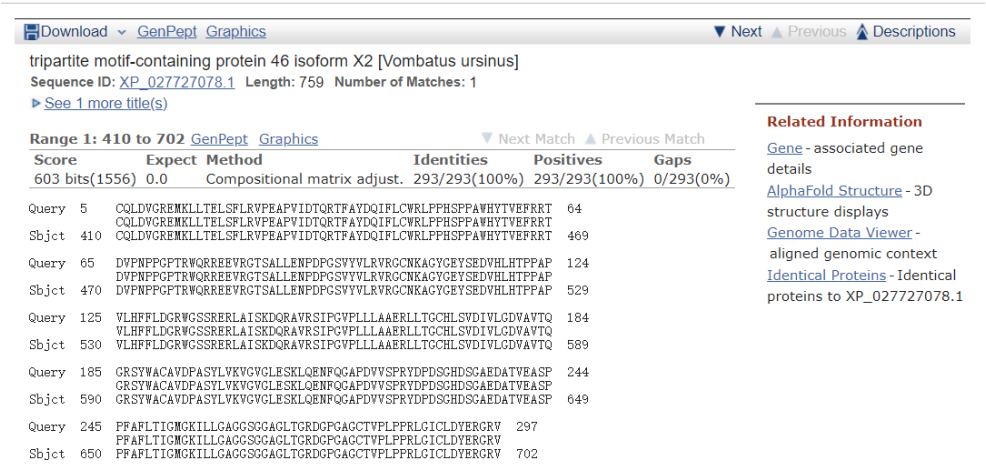
Details: A BLASTP search against NR database (see setup in first screen-shot below) yielded a top hit result is to a protein from different species, Vombatus ursinus (common wombat).

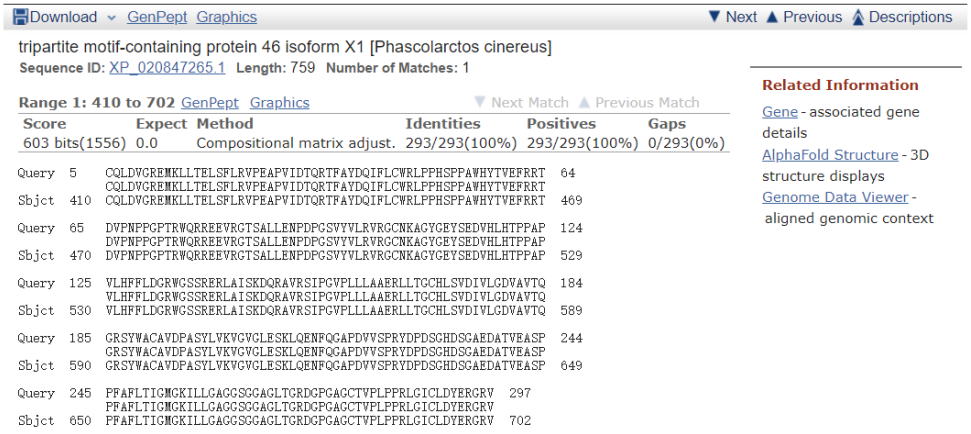
See additional screen shots below for top hits and selected alignment details:



Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
tripartite motif-containing protein 46 isoform X2 [Vombatus ursinus]	603	603	98%	0.0	100.00%	XP_027727078.1
tripartite motif-containing protein 46 isoform X1 [Phascolarctos cinereus]	603	603	98%	0.0	100.00%	XP_020847265.1
tripartite motif-containing protein 46 isoform X1 [Vombatus ursinus]	603	603	98%	0.0	100.00%	XP_027727077.1
tripartite motif-containing protein 46 isoform X2 [Dromiciops gliroides]	603	603	98%	0.0	100.00%	XP_043818602.1
tripartite motif-containing protein 46 isoform X1 [Dromiciops gliroides]	603	603	98%	0.0	100.00%	XP_043818601.1
tripartite motif-containing protein 46 [Trichosurus vulpecula]	603	603	98%	0.0	100.00%	XP_036612290.1
tripartite motif-containing protein 46 isoform X1 [Monodelphis domestica]	603	603	98%	0.0	100.00%	XP_056672280.1
tripartite motif-containing protein 46 isoform X2 [Phascolarctos cinereus]	603	603	98%	0.0	100.00%	XP_020847266.1
tripartite motif-containing protein 46 isoform X4 [Dromiciops gliroides]	602	602	98%	0.0	100.00%	XP_043818604.1
tripartite motif-containing protein 46 isoform X3 [Dromiciops gliroides]	602	602	98%	0.0	100.00%	XP_043818603.1
tripartite motif-containing protein 46 isoform X2 [Monodelphis domestica]	602	602	98%	0.0	100.00%	XP_056672281.1
tripartite motif-containing protein 46 [Gracilinanus agilis]	602	602	98%	0.0	99.66%	XP_044530451.1
tripartite motif-containing protein 46 isoform X3 [Phascolarctos cinereus]	602	602	98%	0.0	100.00%	XP_020847267.1
tripartite motif-containing protein 46 isoform X5 [Dromiciops gliroides]	602	602	98%	0.0	100.00%	XP_043818605.1
tripartite motif-containing protein 46 isoform X4 [Antechinus flavipes]	601	601	98%	0.0	99.66%	XP_051849843.1

The top result is to a protein from Vombatus ursinus (common wombat), see second screen shot below for alignment details





5. 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.

Re-labeled sequences for alignment:

```
>Human | NP_001393174.1 tripartite motif-containing protein 46 isoform 3
[Homo sapiens]
MAEGEDMQTFTSIMDALVRISLCSGEREARDRGLGRSVNQPKAGALEKLQTSMKNMEKELLCPVCQEMYKQPLVL
PCTHNVCQACAREVLGQQGYIGHGGDPSSEPTSPASTPSTRSPRLSRRTLPKPDRLDRLLKSGFGTYPGRKRGAL
HPQVIMFPCPACQGDVELGERGLAGLFRNLTLERVVERYRQSVSVGGAILCQLCKPPPLEATKGCTECRATFCNE
CFKLFHPWGTQKAQHEPTLPTLSFRPKGLMCPDHKEEVTHYCKTCQRLVCQLCRVRRTHSGHKITPVLSAYQALK
DKLTKSLTYILGNQDTVQTQICELEEAVRHTEVSGQQAKEEVSQLVRGLGAVLEEKRASLLQAIEECQQERLARL
SAQIQEHRSLLDGSGLVGYAQEVLKETDQPCFVQAAKQLHNRIARATEALQTFRPAASSSFRHCQLDVGREMKLL
TELNFLRVPEAPVIDTQRTFAYDQIFLCWRLPPHSPPAWHYTVEFRRTDVPAQPGPTRWQRREEVRGTSALLENP
```

DTGSVYVLRVRGCNKAGYGEYSEDVHLHTPPAPVLHFFLDSRWGASRERLAISKDQRAVRSVPGLPLLLAADRLL
TGCHLSVDVVLGDVAVTQGRSYWACAVDPASYLVKVGVGLESKLQESFQGAPDVISPRYDPDSGHDSGAEDATVE
ASPPFAFLTIGMGKILLGSGASSNAGLTGRDGPTAGCTVPLPPRLGICLDYERGRVSFLDAVSFRGLLECPLDCS
GPVCPAFCFIGGGAVQLQEPVGTKPERKVTIGGFAKLD

>Tammar_wallaby | Notamacropus eugenii TRIM46 (sequence from blast result)
GLLDCQLDVGREMKLLTELSFLRVPEAPVIDTQRTFAYDQIFLCWRLPPHSPPAWHYTVEFRRTDVPNPPGPTRW
QRREEVRGTSALLENPDPGSVYVLRVRGCNKAGYGEYSEDVHLHTPPAPVLHFFLDGRWGSSRERLAISKDQRAV
RSIPGVPLLLAAERLLTGCHLSVDIVLGDVAVTQGRSYWACAVDPASYLVKVGVGLESKLQENFQGAPDVVSPRY
DPDSGHDSGAEDATVEASPPFAFLTIGMGKILLGAGGSGGAGLTGRDGPGAGCTVPLPPRLGICLDYERGRV

>Mouse | NP_898858.1 tripartite motif-containing protein 46 isoform 2 [Mus
musculus]
MAEGEDMQTFTSIMDALVRISTSMKNMEKELLCPVCQEMYKQPLVLPCTHNVCQACAREVLGQQGYIGHGGDPSS
EPTSPASTPSTRSPRLSRRTLPKPDRLDRLLKSGFGTYPGRKRGALHPQTILFPCPACQGDVELGERGLSGLFRN
LTLERVVERYRQSVSVGGAILCQLCKPPPLEATKGCTECRATFCNECFKLFHPWGTQKAQHEPTLPTLSFRPKGL
MCPDHKEEVTHYCKTCQRLVCQLCRVRRTHSGHKITPVLSAYQALKDKLTKSLAYILGNQDTVQTQICELEETIR
HTEVSGQQAKEEVSQLVRGLGAVLEEKRASLLQAIEECQQERLSRLSAQIHEHQSLLDGSGLVGYAQEVLKETDQ
PCFVQAAKQLHNRIARATEALQTFRPAASSSFRHCQLDVGREMKLLTELSFLRVPEAPVIDTQRTFAYDQIFLCW
RLPPHSPPAWHYTVEFRRTDVPAQPGPTRWQRREEVRGTSALLENPDTGSVYVLRVRGCNKAGYGEYSEDVHLHT
PPAPVLHFFLDGRWGASRERLAISKDQRAVRSIPGLPLLLAAERLLTGCHLSVDVVLGDVAVTQGRSYWACAVDP
ASYLVKVGVGLESKLQESFQGAPDVISPRYDPDSGHDSGAEDAAVEALPPFAFLTIGMGKILLGSGASSNAGLTG
RDGPTASCTVPLPPRLGICLDYERGRVSFLDAVSFRGLLECPLDCSGPVCPAFCFIGGGAVQLQEPVGTKPERKV
TIGGFAKLD

>Central_European_red_deer | OWK05312.1 TRIM46 [Cervus elaphus hippelaphus]
MLVPGACAVTSHPHPSPHPHRAPALSPGFAAAAGIGHPGAGGHARAMAEGEDMQTFTSIMDALVRISTSMKNMEK
ELLCPVCQEMYKQPLVLPCTHSVCQACAREVLGQQGYIGHGGDPSSEPTSPASTPSTRSPRLSRRTLPKPDRLDR
LLKSGFGTYPGRKRGALHPQVIMFPCPACQGDVELGERGLAGLFRNLTLERVVERYRQSVSVGGAILCQLCKPPP
LEATKGCTECRATFCNECFKLFHPWGTQKAQHEPTLPTLSFRPKGLMCPDHKEEVTHYCKTCQRLVCQLCRVRRT
HSGHKITPVLSAYQALKDKLTKSLTYILGNQDTVQTQICELEETVRHTEVSGQQAKEEVSQLVRGLGAVLEEKRA
SLLQAIEECQQERLARLSAQIQEHRSLLDGSGLVGYAQEVLKETDQPCFVQAAKQLHNSSSFRHCQLDVGREMKL
LTELNFLRVPEAPVIDTQRTFAYDQIFLCWRLPPHSPPAWHYTIEFRRTDVPAQPGPTRWQRREEVRGTSALLEN
PDTGSVYVLRVRGCNKAGYGEYSEDVHLHTPPAPVLHFFLDGRWGTSRERLAISKDQRAVRSVPGLPLLLAAERL
LTGCHLSVDVVLGDVAVTQGRSYWACAVDPASYLVKVGVGLESKLQESFQGAPDVISPRYDPDSGHDSGAEDATV
EASPPFAFLTIGMGKILLGAGASSNAGLTGRDGPAASCTVPLPPRLGICLDYERGRVSFLDAVSFRGLLECPLDC
SGPVCPAFCFIGGGAVQLQEPVGTKPERKVTIGGFAKLD

>Mexican_tetra | KAG9279374.1 tripartite motif-containing protein 46
isoform X1 [Astyanax mexicanus]
MDVLARLSSNMKSMERELQCPVCKDIVKQPVVLPCLHSVCLLCASEVLVQSGYPQPELPPEPNSPASTPNTRSPR
QARRPMPKADRALRPGFGTYPGRRRKEGHTQLMLFPCVPCGRDVELGERGLVDCMRNLTLERIVERYRHTVSLGS
VAVMCQFCKPPQTLEATKGCADCRASFCNECFKLYHPWGTPRAQHEHVLPTLNFRPKVLTCPEHDQEKLQFYCKS
CQRLLCSLCKLRRVHGGHKIVPVTQAYQTLKDKITKELGYILSNQGTVLTQITQLENAITQTEVNSVAAREQLSQ
CVRELMALLSERQAMLAQGLESSRQKRSEALANQVAERRSLLEHAGLMAFTQELLKETDSACFVHAARQTHNRLA
QSIESLQSFSLSADPSFRHFQLDVSRELKLLTDLNFIQAPLAPVIDTQRTLAYDQLFLCWRLPPDSSPAWHFSVE
YRRRGVVPGGGARGGMRGGLSGARWGWQRLEEVSGTSAVIDRLEMDSVYVLRVRGCNKAGFGEYSEEVYLHTPPA
PVLNFYLDSRWGLHADRLVVSKEQRCARSVPGLSLLQAADRALTSCHLTADLLVGDVAVTQGRHYWACSVEPGSY
LVKVGVGLESKLQEWFHLPQDMASPRYDPDSGHDSGAEDSSDSPPSFTFLTMGMGKIYLPSSANSHHGTANYRDG
GIANGNGPSSPTGVTYPLPPRLGVCLDVEKGRVTFYDAHSLRPLWEGAVDCSAPVCPAFCFIGGGALQLQELVAN
RNADQTPVRRVTIQSRVTKLN

>Blue_mussel | CAG2244917.1 TRIM46 [Mytilus edulis]
MKKVKSSYKKDNLNSFLTSAKECCSLSEQLINRNSKRSFLNVHQTVEAHMKRYLNTPVEKSTCGEKDSEENMIDF
DDHLRLFERNVEILENDVGDGFNEPLIEPAKVSGLKFETDLESPFACRSADNQSVSTKPFANMLCGSRRTLYKYE
GVFANTSFVFGKESSIEILIRFQLVQQQNEQKYDKLTVFEFGLREDSISSELLFPSFLSVTAFPCSNTFGVCLLS
GNGILLANKDILERKPNSSIVEGQFSINYQPSDSLFSIRAKYPKSNTTTELHRAQVFDFRIPVWAVFAAYNSDKF
NVSMTITTHEVGFNRSFGSKCFQKCCICGNTSDITYNCQHCNINLCEICRCPHEKENKTHNLIVNKPTTFREEDD
NLDVCQNLHHERAKLRYFCNSSNCQIVLCPLCVIAEHRDISKHELEDIDEAFEKEKRIGK

>Rock_dove | KAK2540081.1 Trim46 [Columba livia]

MCQFCKPPQLEATKGCTECKSSFCNECFKLYHPWGTQKAQHEPTPPTLTFRPKGLMCPEHKEEVTHYCKTCQRLV
CQLCRVRRTHTSHKITPVLSAYQALREKLSKSIAYILSSQDTVQTQIAELEETVKHTEANGSQAKEEVSQLIGAL
GAMLEEKRAALLQAIEECQQQRLASLHGQIQEHQAMLENSGMVGYAQEVLKETDHPCFVQAAKQLHNRILRATDS
LQSFRPAATASFSHFQLDVSRELKLLTDLAFIRGNGSAAGAPAAPRAELQPPCPSRSSAAGSRGSAGPPIPVRDG
PRPSSPQNDPCPRARTGRAAGKGRRSCRRVSTEPRLPHRRQEPRGPAGPAGDAFQTCFYLSLFIGCGRC

>Mosquito | KFB48111.1 Trim46 protein [Anopheles sinensis]
MLRLGFIRTERSRLADPRQLERFNSLPVMAVQTSAPRGRGGVGVGDLRRRPERTSARGSIESYCGEAGEIASEKN
PSPVQSPGECTCNFSLQLHKHGNILPTTSAVFFSSSVSFSATIPSRTKAPGFYSTIPELLARAGTYWRKFTPWAS
ARFREKKRWLSRALKLLIRKTQQHPPKTLDGCPPTPSRRSSGWPV

Alignment (Obtained using MUSCLE (version 3.8) at EBI):

```
CLUSTAL multiple sequence alignment by MUSCLE (3.8) default setting:

Blue_mussel                 ------------------------------------------------------------
Mosquito                    ------------------------------------------------------------
Rock_dove                   ------------------------------------------------------------
Mexican_tetra               ------------------------------------------------------------M
Tammar_wallaby              ------------------------------------------------------------
Mouse                       ---------------------------------------------MAEGEDMQTFTSIM
Human                       ---------------------------------------------MAEGEDMQTFTSIM
Central_European_red_deer   MLVPGACAVTSHPHPSPHPHRAPALSPGFAAAAGIGHPGAGGHARAMAEGEDMQTFTSIM



Blue_mussel                 ------------------------------------------------------------
Mosquito                    ------------------------------------------------------------
Rock_dove                   ------------------------------------------------------------
Mexican_tetra               DVLARLS----------------------------SNMKSMERELQCPVCKDIVKQPVV
Tammar_wallaby              ------------------------------------------------------------
Mouse                       DALVRIS----------------------------TSMKNMEKELLCPVCQEMYKQPLV
Human                       DALVRISLCSGEREARDRGLGRSVNQPKAGALEKLQTSMKNMEKELLCPVCQEMYKQPLV
Central_European_red_deer   DALVRIS----------------------------TSMKNMEKELLCPVCQEMYKQPLV



Blue_mussel                 ------------------------------------------------------------
Mosquito                    --------------MLRLGFI----------------------RTERSRLADPRQLERF
Rock_dove                   ------------------------------------------------------------
Mexican_tetra               LPCLHSVCLLCASEVLVQSGY--PQPELPPEPNSPASTPNTRSPRQARRPMPKA---DRA
Tammar_wallaby              ------------------------------------------------------------
Mouse                       LPCTHNVCQACAREVLGQQGYIGHGGDPSSEPTSPASTPSTRSPRLSRRTLPKPDRLDRL
Human                       LPCTHNVCQACAREVLGQQGYIGHGGDPSSEPTSPASTPSTRSPRLSRRTLPKPDRLDRL
Central_European_red_deer   LPCTHSVCQACAREVLGQQGYIGHGGDPSSEPTSPASTPSTRSPRLSRRTLPKPDRLDRL
```

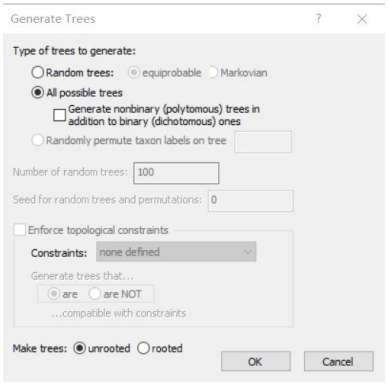
```
Blue_mussel                 ------------------------------------------------------------
Mosquito                    -------------NSLPVMAVQTSAPRGRGGVGVGD-----------------------
Rock_dove                   ------------------------------------------------------------
Mexican_tetra               LRPGFGTYPGRRRKEGHTQLMLFPCVPCGRDVELGERGLVDCMRNLTLERIVERYRHTVS
Tammar_wallaby              ------------------------------------------------------------
Mouse                       LKSGFGTYPGRKRGALHPQTILFPCPACQGDVELGERGLSGLFRNLTLERVVERYRQSVS
Human                       LKSGFGTYPGRKRGALHPQVIMFPCPACQGDVELGERGLAGLFRNLTLERVVERYRQSVS
Central_European_red_deer   LKSGFGTYPGRKRGALHPQVIMFPCPACQGDVELGERGLAGLFRNLTLERVVERYRQSVS

Blue_mussel                 ----------------------MKKVKSSYKKDNLNSF-----------------------
Mosquito                    ---------LRRRPE-RTSARGSIE---SYCGEA--------------------------
Rock_dove                   ------MCQFCKPPQ-LEATKGCTECKSSFCNECFKLYHPWGTQKAQHEPTPPTLTFRPK
Mexican_tetra               LGSVAVMCQFCKPPQTLEATKGCADCRASFCNECFKLYHPWGTPRAQHEHVLPTLNFRPK
Tammar_wallaby              ------------------------------------------------------------
Mouse                       VGG-AILCQLCKPPP-LEATKGCTECRATFCNECFKLFHPWGTQKAQHEPTLPTLSFRPK
Human                       VGG-AILCQLCKPPP-LEATKGCTECRATFCNECFKLFHPWGTQKAQHEPTLPTLSFRPK
Central_European_red_deer   VGG-AILCQLCKPPP-LEATKGCTECRATFCNECFKLFHPWGTQKAQHEPTLPTLSFRPK

Blue_mussel                 --------LTSAKECCSLSEQLINRNSKRSFLNVHQTVEAHMKRY--LNTPVEKSTCGEK
Mosquito                    ------------------------------------------------------------
Rock_dove                   GLMCPEH-KEEVTHYCKTCQRLVCQLCRVRRTHTSHKITPVLSAYQALREKLSKSIAYIL
Mexican_tetra               VLTCPEHDQEKLQFYCKSCQRLLCSLCKLRRVHGGHKIVPVTQAYQTLKDKITKELGYIL
Tammar_wallaby              ------------------------------------------------------------
Mouse                       GLMCPDH-KEEVTHYCKTCQRLVCQLCRVRRTHSGHKITPVLSAYQALKDKLTKSLAYIL
Human                       GLMCPDH-KEEVTHYCKTCQRLVCQLCRVRRTHSGHKITPVLSAYQALKDKLTKSLTYIL
Central_European_red_deer   GLMCPDH-KEEVTHYCKTCQRLVCQLCRVRRTHSGHKITPVLSAYQALKDKLTKSLTYIL

Blue_mussel                 DSEENMIDFDDHLRLFERNVEILENDVGDGFNEPLIEPAKVSGLKFETDLESPFACRSAD
Mosquito                    ---------------------------------------GEIASEKNPSPVQSPGECTCNF
Rock_dove                   SSQDTVQTQIAELEETVKHTEANGSQAKEEVSQLIGALGAMLEEKRAALLQAIEECQQQR
Mexican_tetra               SNQGTVLTQITQLENAITQTEVNSVAAREQLSQCVRELMALLSERQAMLAQGLESSRQKR
Tammar_wallaby              ------------------------------------------------------------
Mouse                       GNQDTVQTQICELEETIRHTEVSGQQAKEEVSQLVRGLGAVLEEKRASLLQAIEECQQER
Human                       GNQDTVQTQICELEEAVRHTEVSGQQAKEEVSQLVRGLGAVLEEKRASLLQAIEECQQER
Central_European_red_deer   GNQDTVQTQICELEETVRHTEVSGQQAKEEVSQLVRGLGAVLEEKRASLLQAIEECQQER
```

```
Blue_mussel                    NQSVSTKPFANMLCGSRRTLYKYEGVFANTSFVFGKESSIEILIRFQLVQQQNEQKYDKL
Mosquito                       SLQLHKHGN----ILPTTSAVFFSSSVSFSATIPSRTKAPGFYSTIPEL---LARAGTYW
Rock_dove                      LASLHGQ------IQEHQAMLENSGMVGYAQEVLKETDHPCFVQAAKQLHNRILRATDSL
Mexican_tetra                  SEALANQ------VAERRSLLEHAGLMAFTQELLKETDSACFVHAARQTHNRLAQSIESL
Tammar_wallaby                 -----------------GLLD---------------------------------------
Mouse                          LSRLSAQ------IHEHQSLLDGSGLVGYAQEVLKETDQPCFVQAAKQLHNRIARATEAL
Human                          LARLSAQ------IQEHRSLLDGSGLVGYAQEVLKETDQPCFVQAAKQLHNRIARATEAL
Central_European_red_deer      LARLSAQ------IQEHRSLLDGSGLVGYAQEVLKETDQPCFVQAAKQLHN---------


Blue_mussel                    TVF------EFGLREDSISSELLFP---SFLS----------------------------
Mosquito                       RKFTPWASARFREKKRWLSRALKLL-----IR----------------------------
Rock_dove                      QSFRPAATASFSHFQLDVSRELKLLTDLAFIRGNGSAAGAPAAPRAELQ-----------
Mexican_tetra                  QSFSLSADPSFRHFQLDVSRELKLLTDLNFIQ-------APLAPVIDTQRTLAYDQLFLC
Tammar_wallaby                 -------------CQLDVGREMKLLTELSFLR-------VPEAPVIDTQRTFAYDQIFLC
Mouse                          QTFRPAASSSFRHCQLDVGREMKLLTELSFLR-------VPEAPVIDTQRTFAYDQIFLC
Human                          QTFRPAASSSFRHCQLDVGREMKLLTELNFLR-------VPEAPVIDTQRTFAYDQIFLC
Central_European_red_deer      -------SSSFRHCQLDVGREMKLLTELNFLR-------VPEAPVIDTQRTFAYDQIFLC
                                             :  :.  : :      :


Blue_mussel                    ---------------------VTAFPCSNT--------------FGVCLLSGNGILLANK
Mosquito                       ----------------------KTQQHPPK------------------------------
Rock_dove                      ------------------------PPCPSR--------------SSAAGSRGSAG-----
Mexican_tetra                  WRLPPDSSPAWHFSVEYRRRGV-VPGGGARGGMRGGLSGARWGWQRLEEVSGTSAVIDRL
Tammar_wallaby                 WRLPPHSPPAWHYTVEFRRTDVPNPPGPTR-------------WQRREEVRGTSALLENP
Mouse                          WRLPPHSPPAWHYTVEFRRTDVPAQPGPTR-------------WQRREEVRGTSALLENP
Human                          WRLPPHSPPAWHYTVEFRRTDVPAQPGPTR-------------WQRREEVRGTSALLENP
Central_European_red_deer      WRLPPHSPPAWHYTIEFRRTDVPAQPGPTR-------------WQRREEVRGTSALLENP


Blue_mussel                    D-----ILERKPNSSIVEGQFSINYQ----PSDSL-FSIRAKYPKSNTTTELHRAQVFDF
Mosquito                       --------TLDGC----------------PPTPS-------------------------R
Rock_dove                      -----------------------------PPIPV----------RDGPRPSSPQNDPCPR
Mexican_tetra                  EMDSVYVVLRVRGCNKAGFGEYSEEVYLHTPPAPVLNFYLDSRWGLHADRLVVSKEQRCAR
Tammar_wallaby                 DPGSVYVLRVRGCNKAGYGEYSEDVHLHTPPAPVLHFFLDGRWGSSRERLAISKDQRAVR
Mouse                          DTGSVYVLRVRGCNKAGYGEYSEDVHLHTPPAPVLHFFLDGRWGASRERLAISKDQRAVR
Human                          DTGSVYVLRVRGCNKAGYGEYSEDVHLHTPPAPVLHFFLDSRWGASRERLAISKDQRAVR
Central_European_red_deer      DTGSVYVLRVRGCNKAGYGEYSEDVHLHTPPAPVLHFFLDGRWGTSRERLAISKDQRAVR
                                                             *
```

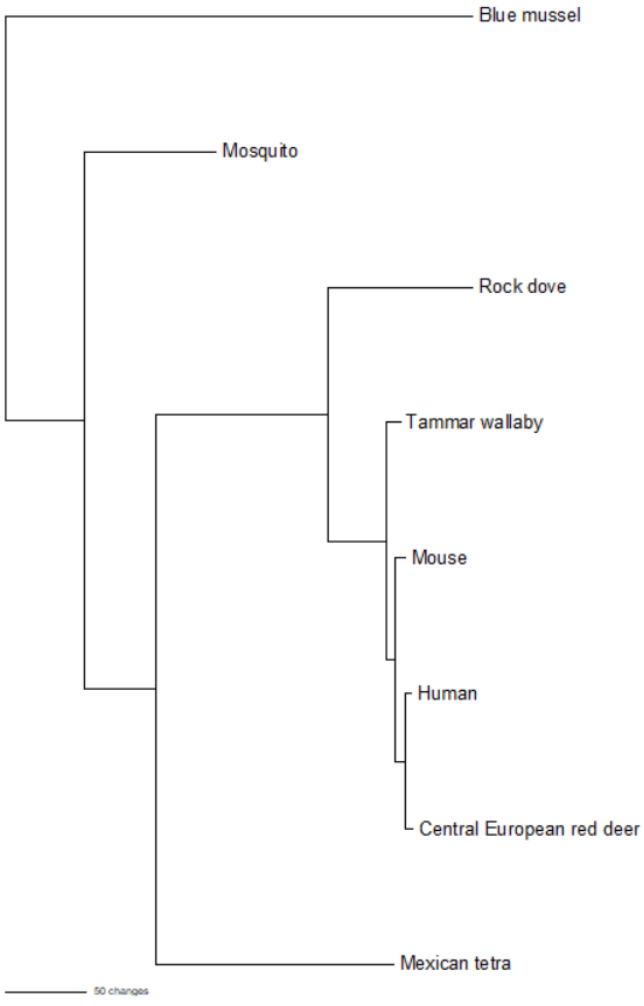
```
Blue_mussel                 RIPVWAVFAA----YNSDKFNVSMTITTHEVGFNRSFGSKCFQKCCI-----CGNTSDIT
Mosquito                    RSSGWPV-------------------------------------------------------
Rock_dove                   ARTGR--------------------------AAGKGRRSC--------------------
Mexican_tetra               SVPGLSLLQAADRALTSCHLTADLLVGDVAVTQGRHYWACSVEPGSYLVKVGVGLESKLQ
Tammar_wallaby              SIPGVPLLLAAERLLTGCHLSVDIVLGDVAVTQGRSYWACAVDPASYLVKVGVGLESKLQ
Mouse                       SIPGLPLLLAAERLLTGCHLSVDVVLGDVAVTQGRSYWACAVDPASYLVKVGVGLESKLQ
Human                       SVPGLPLLLAADRLLTGCHLSVDVVLGDVAVTQGRSYWACAVDPASYLVKVGVGLESKLQ
Central_European_red_deer   SVPGLPLLLAAERLLTGCHLSVDVVLGDVAVTQGRSYWACAVDPASYLVKVGVGLESKLQ
                                  .

Blue_mussel                 YNCQHCNINLCEICRCPHEKENKTHNLIVNKPTTF--------------------------
Mosquito                    -------------------------------------------------------------
Rock_dove                   RRVSTEPRLPHRRQEP-----RGPAGPAGDAFQTCFYLSLFI-------------------
Mexican_tetra               EWFHLPQDMASPRYDPDSGHDSGAEDSS-DSPPSFTFLTMGMGKIYLPSSANSHHGTANY
Tammar_wallaby              ENFQGAPDVVSPRYDPDSGHDSGAEDATVEASPPFAFLTIGMGKILLGAGGSG-------
Mouse                       ESFQGAPDVISPRYDPDSGHDSGAEDAAVEALPPFAFLTIGMGKILLGSGASS-------
Human                       ESFQGAPDVISPRYDPDSGHDSGAEDATVEASPPFAFLTIGMGKILLGSGASS-------
Central_European_red_deer   ESFQGAPDVISPRYDPDSGHDSGAEDATVEASPPFAFLTIGMGKILLGAGASS-------

Blue_mussel                 -------REEDDN-----------LDVCQNLHHERAKLRYFCNSS---------NCQIVL
Mosquito                    -------------------------------------------------------------
Rock_dove                   ---------------GC--------GRC---------------------------------
Mexican_tetra               RDGGIANGNGPSSPTGVTYPLPPRLGVC--LDVEKGRVTFYDAHSLRPLWEGAVDCSAPV
Tammar_wallaby              -GAGLTGRDGPGA--GCTVPLPPRLGIC--LDYERGRV----------------------
Mouse                       -NAGLTGRDGPTA--SCTVPLPPRLGIC--LDYERGRVSFLDAVSFRGLLECPLDCSGPV
Human                       -NAGLTGRDGPTA--GCTVPLPPRLGIC--LDYERGRVSFLDAVSFRGLLECPLDCSGPV
Central_European_red_deer   -NAGLTGRDGPAA--SCTVPLPPRLGIC--LDYERGRVSFLDAVSFRGLLECPLDCSGPV

Blue_mussel                 CPLCVIAEHRDISKHELEDIDEA---FEKEKRIGK------
Mosquito                    ----------------------------------------
Rock_dove                   ----------------------------------------
Mexican_tetra               CPAFCFIGGGALQLQELVANRNADQTPVRRVTIQSRVTKLN
Tammar_wallaby              ----------------------------------------
Mouse                       CPAFCFIGGGAVQLQEPVGTK-----PERKVTIGG-FAKLD
Human                       CPAFCFIGGGAVQLQEPVGTK-----PERKVTIGG-FAKLD
Central_European_red_deer   CPAFCFIGGGAVQLQEPVGTK-----PERKVTIGG-FAKLD
```

6. 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.

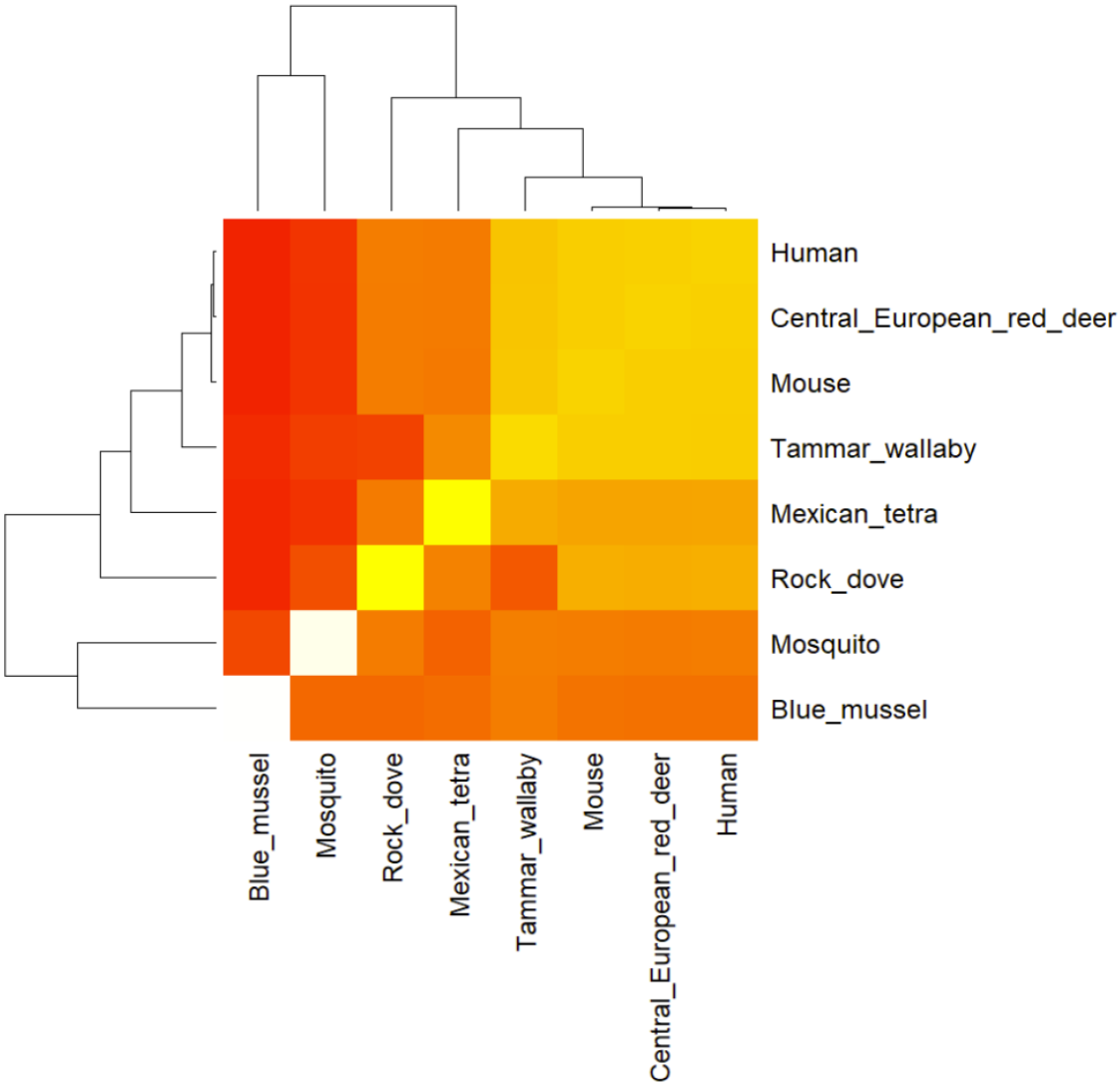
The MUSCLE result above is converted to .nxs file by Seaview, and subjected to phylogeny analysis (Parsimony method), phylogeny options as shown below:



The best-scored tree of all the 10395 possible trees is shown below:



7. 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.

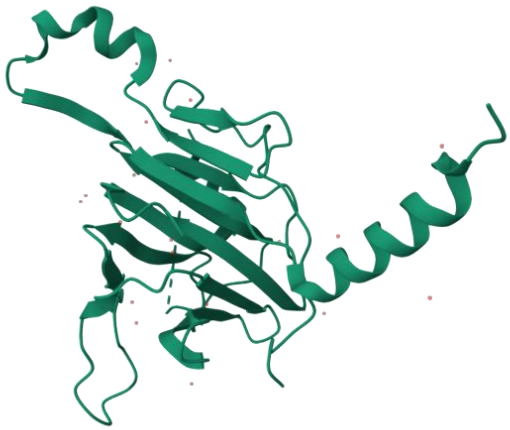


8. Using R/Bio3D (or an online blast server if you prefer), search the main protein structure database for the most similar atomic resolution structures to your aligned sequences. List the top 3 unique hits (i.e. not hits representing different chains from the same structure) along with their Evalue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

PDB Blast of Tammar wallaby TRIM46 sequence:

ID	Technique	Resolution	Source	E value	Identity
7QS4	X-RAY DIFFRACTION	2.25 Å	Homo sapiens	7.27e-44	43.258
2DJS	SOLUTION NMR	NA	Homo sapiens	2.40	37.838
7S7K	X-RAY DIFFRACTION	3.15 Å	Mus musculus	3.30	24.510

9. 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



Structure of 7QS4. It is unlikely to be similar to my novel protein for the identity is only 43%.

10. 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?

    The search using novel protein sequence yields 51 targets, only 2 of which have significant E value, but the identities are as low as 26%:

    https://www.ebi.ac.uk/chembl/target_report_card/CHEMBL4602/
    https://www.ebi.ac.uk/chembl/target_report_card/CHEMBL4739689/

    Both are assays on Ephrin type-A receptor 7, the former one in human and the latter one in mouse. Since our novel protein should be TRIM46, a microtubule-binding protein which also have ubiquitin ligase activity, this is unlikely useful for studying function of our novel protein.

    I also tried to search for TRIM46, and there is only an assay on TRIM24 and TRIM33 (other proteins of TRIM family), but not TRIM46:
    https://www.ebi.ac.uk/chembl/assay_report_card/CHEMBL4418857/
    This is a binding assay on the IC50 of TRIM24 and TRIM33 of a TRIM33 inhibitor. The researchers used this assay in their patent "Inhibitors of trim33 and methods of use".