Norman Lee A18086849
nol007@ucsd.edu

[Q1] Tell me the name of a protein you are interested in. Include the species, accession number and known function. 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: GTP-binding nuclear protein Ran isoform 1 [Homo sapiens]
Accession: NP_006316.1
Species: Homo sapiens
GTP-binding protein involved in nucleocytoplasmic transport. Required for the import of protein into the nucleus and also for RNA export. Involved in chromatin condensation and control of cell cycle..

[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).

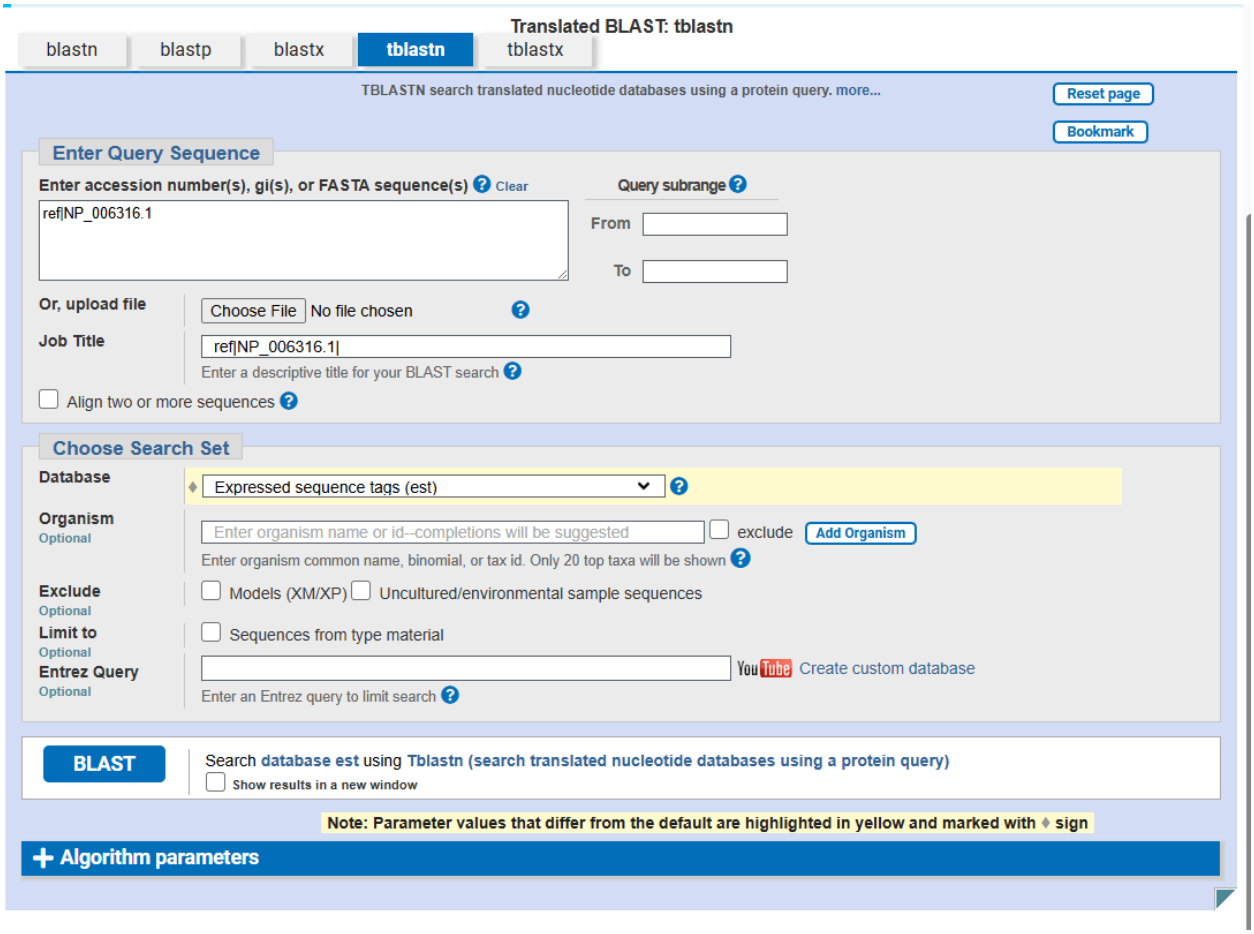
Also include the output of that BLAST search in your document. If appropriate, change the font to `Courier size 10` so that the results are displayed neatly. You can also screen capture a BLAST output (e.g. alt print screen on a PC or on a MAC press ⌘-shift-4. The pointer becomes a bulls eye. Select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

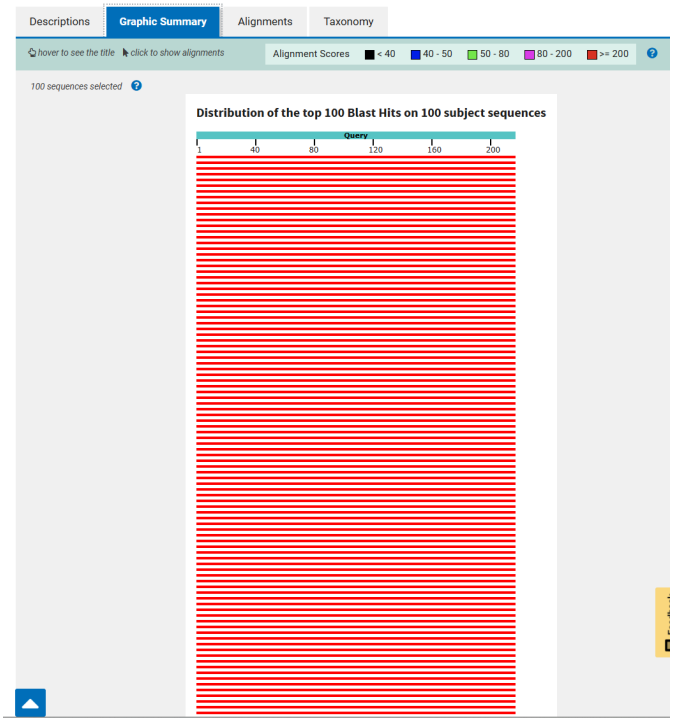
On the BLAST results, clearly indicate a match that represents a protein sequence, encoded from some DNA sequence, that is homologous to your query protein. I need to be able to inspect the pairwise alignment you have selected, including the E value and score. It should be labeled a "genomic clone" or "mRNA sequence", etc. - but include no functional annotation.

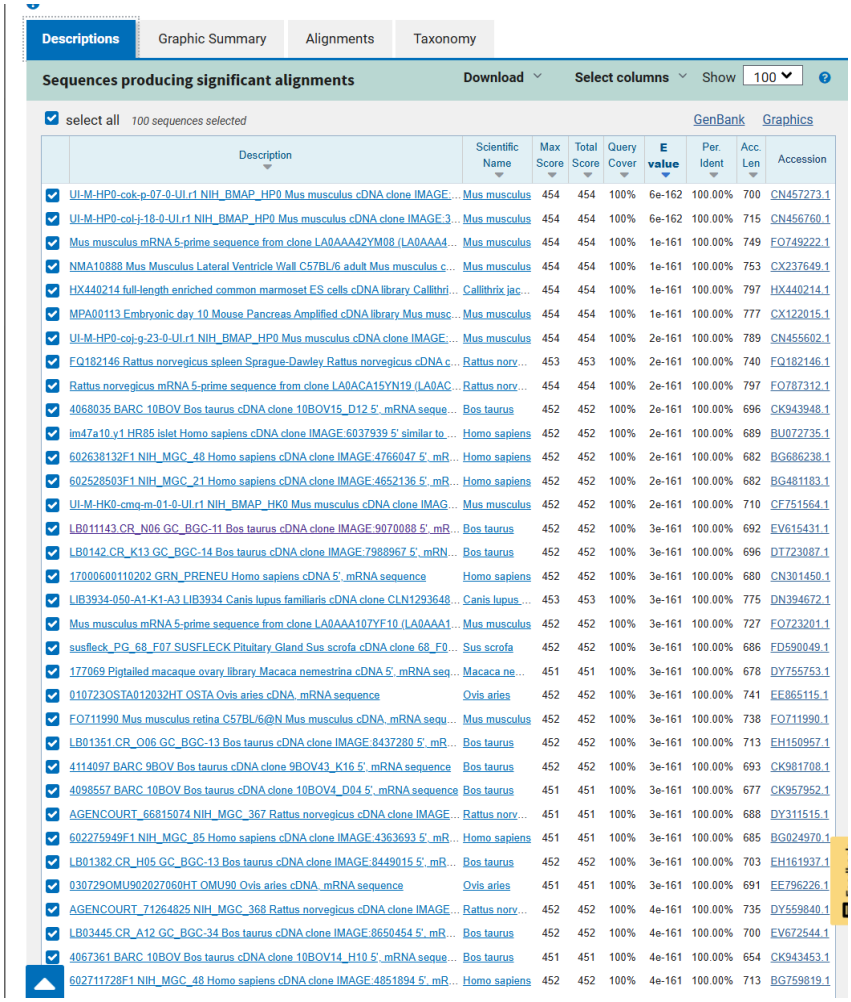
In general, [Q2] is the most difficult for students because it requires you to have a “feel” for how to interpret BLAST results. You need to distinguish between a perfect match to your query (i.e. a sequence that is not “novel”), a near match (something that might be “novel”, depending on the results 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.

Method: tblastn on ncbi
Database: expressed sequence tags
Organism: not specified


Translated BLAST: 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
ref|NP_006316.1
Query subrange
From
To
Or, upload file
Choose File
No file chosen
Job Title
ref|NP_006316.1|
Enter a descriptive title for your BLAST search
Align two or more sequences
Choose Search Set
Database
Expressed sequence tags (est)
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)
Uncultured/environmental sample sequences
Limit to
Optional
Sequences from type material
Entrez Query
Optional
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


Descriptions
Graphic Summary
Alignments
Taxonomy
hover to see the title
click to show alignments
Alignment Scores
< 40
40 - 50
50 - 80
80 - 200
>= 200
100 sequences selected
Distribution of the top 100 Blast Hits on 100 subject sequences
Query
1
40
80
120
160
200



Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments — Download — Select columns — Show 100

select all 100 sequences selected — GenBank — Graphics

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
UI-M-HP0-cok-p-07-0-UI.r1 NIH_BMAP_HP0 Mus musculus cDNA clone IMAGE:...	Mus musculus	454	454	100%	6e-162	100.00%	700	CN457273.1
UI-M-HP0-col-j-18-0-UI.r1 NIH_BMAP_HP0 Mus musculus cDNA clone IMAGE:3...	Mus musculus	454	454	100%	6e-162	100.00%	715	CN456760.1
Mus musculus mRNA 5-prime sequence from clone LA0AAA42YM08 (LA0AAA4...	Mus musculus	454	454	100%	1e-161	100.00%	749	FO749222.1
NMA10888 Mus Musculus Lateral Ventricle Wall C57BL/6 adult Mus musculus c...	Mus musculus	454	454	100%	1e-161	100.00%	753	CX237649.1
HX440214 full-length enriched common marmoset ES cells cDNA library Callithri...	Callithrix jac...	454	454	100%	1e-161	100.00%	797	HX440214.1
MPA00113 Embryonic day 10 Mouse Pancreas Amplified cDNA library Mus musc...	Mus musculus	454	454	100%	1e-161	100.00%	777	CX122015.1
UI-M-HP0-coj-g-23-0-UI.r1 NIH_BMAP_HP0 Mus musculus cDNA clone IMAGE:...	Mus musculus	454	454	100%	2e-161	100.00%	789	CN455602.1
FQ182146 Rattus norvegicus spleen Sprague-Dawley Rattus norvegicus cDNA c...	Rattus norv...	453	453	100%	2e-161	100.00%	740	FQ182146.1
Rattus norvegicus mRNA 5-prime sequence from clone LA0ACA15YN19 (LA0AC...	Rattus norv...	454	454	100%	2e-161	100.00%	797	FO787312.1
4068035 BARC 10BOV Bos taurus cDNA clone 10BOV15_D12 5', mRNA seque...	Bos taurus	452	452	100%	2e-161	100.00%	696	CK943948.1
im47a10.y1 HR85 islet Homo sapiens cDNA clone IMAGE:6037939 5' similar to ...	Homo sapiens	452	452	100%	2e-161	100.00%	689	BU072735.1
602638132F1 NIH_MGC_48 Homo sapiens cDNA clone IMAGE:4766047 5', mR...	Homo sapiens	452	452	100%	2e-161	100.00%	682	BG686238.1
602528503F1 NIH_MGC_21 Homo sapiens cDNA clone IMAGE:4652136 5', mR...	Homo sapiens	452	452	100%	2e-161	100.00%	682	BG481183.1
UI-M-HK0-cmg-m-01-0-UI.r1 NIH_BMAP_HK0 Mus musculus cDNA clone IMAG...	Mus musculus	452	452	100%	2e-161	100.00%	710	CF751564.1
LB011143.CR_N06 GC_BGC-11 Bos taurus cDNA clone IMAGE:9070088 5', mR...	Bos taurus	452	452	100%	3e-161	100.00%	692	EV615431.1
LB0142.CR_K13 GC_BGC-14 Bos taurus cDNA clone IMAGE:7988967 5', mRN...	Bos taurus	452	452	100%	3e-161	100.00%	696	DT723087.1
17000600110202 GRN_PRENEU Homo sapiens cDNA 5', mRNA sequence	Homo sapiens	452	452	100%	3e-161	100.00%	680	CN301450.1
LIB3934-050-A1-K1-A3 LIB3934 Canis lupus familiaris cDNA clone CLN1293648...	Canis lupus ...	453	453	100%	3e-161	100.00%	775	DN394672.1
Mus musculus mRNA 5-prime sequence from clone LA0AAA107YF10 (LA0AAA1...	Mus musculus	452	452	100%	3e-161	100.00%	727	FO723201.1
susfleck_PG_68_F07 SUSFLECK Pituitary Gland Sus scrofa cDNA clone 68_F0...	Sus scrofa	452	452	100%	3e-161	100.00%	686	FD590049.1
177069 Pigtailed macaque ovary library Macaca nemestrina cDNA 5', mRNA seq...	Macaca ne...	451	451	100%	3e-161	100.00%	678	DY755753.1
010723OSTA012032HT OSTA Ovis aries cDNA, mRNA sequence	Ovis aries	452	452	100%	3e-161	100.00%	741	EE865115.1
FO711990 Mus musculus retina C57BL/6@N Mus musculus cDNA, mRNA sequ...	Mus musculus	452	452	100%	3e-161	100.00%	738	FO711990.1
LB01351.CR_O06 GC_BGC-13 Bos taurus cDNA clone IMAGE:8437280 5', mR...	Bos taurus	452	452	100%	3e-161	100.00%	713	EH150957.1
4114097 BARC 9BOV Bos taurus cDNA clone 9BOV43_K16 5', mRNA sequence	Bos taurus	452	452	100%	3e-161	100.00%	693	CK981708.1
4098557 BARC 10BOV Bos taurus cDNA clone 10BOV4_D04 5', mRNA sequence	Bos taurus	451	451	100%	3e-161	100.00%	677	CK957952.1
AGENCOURT_66815074 NIH_MGC_367 Rattus norvegicus cDNA clone IMAGE...	Rattus norv...	451	451	100%	3e-161	100.00%	688	DY311515.1
602275949F1 NIH_MGC_85 Homo sapiens cDNA clone IMAGE:4363693 5', mR...	Homo sapiens	451	451	100%	3e-161	100.00%	685	BG024970.1
LB01382.CR_H05 GC_BGC-13 Bos taurus cDNA clone IMAGE:8449015 5', mR...	Bos taurus	452	452	100%	3e-161	100.00%	703	EH161937.1
030729OMU902027060HT OMU90 Ovis aries cDNA, mRNA sequence	Ovis aries	451	451	100%	3e-161	100.00%	691	EE796226.1
AGENCOURT_71264825 NIH_MGC_368 Rattus norvegicus cDNA clone IMAGE...	Rattus norv...	452	452	100%	4e-161	100.00%	735	DY559840.1
LB03445.CR_A12 GC_BGC-34 Bos taurus cDNA clone IMAGE:8650454 5', mR...	Bos taurus	452	452	100%	4e-161	100.00%	700	EV672544.1
4067361 BARC 10BOV Bos taurus cDNA clone 10BOV14_H10 5', mRNA seque...	Bos taurus	451	451	100%	4e-161	100.00%	654	CK943453.1
602711728F1 NIH_MGC_48 Homo sapiens cDNA clone IMAGE:4851894 5', mR...	Homo sapiens	452	452	100%	4e-161	100.00%	713	BG759819.1

Chosen match:

HX208769 full-length enriched swine cDNA library, adult bone marrow Sus scrofa cDNA clone BMWN10053A03, mRNA sequence

Alignment details:

HX208769 full-length enriched swine cDNA library, adult bone marrow Sus scrofa cDNA clone BMWN10053A03, mRNA sequence

Sequence ID: HX208769.1 Length: 725 Number of Matches: 1

Range 1: 43 to 690 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
452 bits(1162)	4e-161	Compositional matrix adjust.	216/216(100%)	216/216(100%)	0/216(0%)	+1

```
Query  1    MAAQGEPQVQFKLVLVGDGGTGKTTFVKRHLTGEFEKKYVATLGVEVHPLVFH
            MAAQGEPQVQFKLVLVGDGGTGKTTFVKRHLTGEFEKKYVATLGVEVHPLVFH
Sbjct  43   MAAQGEPQVQFKLVLVGDGGTGKTTFVKRHLTGEFEKKYVATLGVEVHPLVFH

Query  61   FNVWDTAGQEKFGGLRDGYYIQAQCAIIMFDVTSRVTYKNVPNWHRDLVRVCE
            FNVWDTAGQEKFGGLRDGYYIQAQCAIIMFDVTSRVTYKNVPNWHRDLVRVCE
Sbjct  223  FNVWDTAGQEKFGGLRDGYYIQAQCAIIMFDVTSRVTYKNVPNWHRDLVRVCE

Query  121  GNKVDIKDRKVKAKSIVFHRKKNLQYYDISAKSNYNFEKPFLWLARKLIGDPN
            GNKVDIKDRKVKAKSIVFHRKKNLQYYDISAKSNYNFEKPFLWLARKLIGDPN
Sbjct  403  GNKVDIKDRKVKAKSIVFHRKKNLQYYDISAKSNYNFEKPFLWLARKLIGDPN

Query  181  ALAPPEVVMDPALAAQYEHDLEVAQTTALPDEDDDL  216
            ALAPPEVVMDPALAAQYEHDLEVAQTTALPDEDDDL
Sbjct  583  ALAPPEVVMDPALAAQYEHDLEVAQTTALPDEDDDL  690
```

Q3] Gather information about this “novel” protein. At a minimum, show me the protein sequence of the “novel” protein as displayed in your BLAST results from [Q2] as FASTA format (you can copy and paste the aligned sequence subject lines from your BLAST result page if necessary) or translate your novel DNA sequence using a tool called EMBOSS Transeq at the EBI. Don’t forget to translate all six reading frames; the ORF (open reading frame) is likely to be the longest sequence without a stop codon. It may not start with a methionine if you don’t have the complete coding region. Make sure the sequence you provide includes a header/subject line and is in traditional FASTA format.

Chosen sequence:

>HX208769.1 HX208769 full-length enriched swine cDNA library, adult bone marrow Sus scrofa cDNA clone BMWN10053A03, mRNA sequence

GAGTCAGACGGGCGCGGAGACGCTTCTGGAAGTAACATCACGATGGCTGCCCAAGGAGAG
CCCCAAGTTCAGTTCAAACT
TGTATTGGTTGGTGATGGTGGTACTGGGAAAACTACATTCGTGAAACGTCATCTGACTGGTG
AATTTGAGAAGAAGTATG
TAGCTACCTTGGGTGTTGAGGTCCATCCCCTTGTGTTCCATACCAACAGAGGACCTATTAAG
TTCAATGTATGGGATACG
GCTGGTCAGGAGAAATTTGGTGGACTGAGAGATGGCTATTATATCCAAGCTCAGTGTGCCAT
TATAATGTTTGATGTAAC
ATCGAGAGTTACTTACAAGAACGTACCTAACTGGCATAGAGATCTGGTACGAGTGTGTGAAA
ACATCCCCATTGTGTTGT
GTGGCAACAAAGTGGATATTAAGGACAGAAAGGTTAAGGCAAAATCGATTGTCTTCCACCGA
AAGAAGAACCTTCAGTAC

TACGACATTTCTGCAAAAAGTAACTACAACTTTGAAAAGCCCTTCCTCTGGCTTGCTAGGAA
ACTGATCGGAGACCCTAA
CTTGGAGTTTGTCGCCATGCCTGCTCTTGCCCCGCCAGAGGTGGTCATGGACCCAGCCTT
GGCAGCACAGTATGAGCATG
ATCTAGAGGTTGCTCAGACAACTGCTCTCCCGGATGAAGATGATGACCTGTGAGAAAACAA
AGCTGGAGCCCAGCGTCAG
AAGTC

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: full-length enriched swine cDNA library, adult bone marrow
Sus scrofa cDNA clone BMWN10053A03, mRNA sequence.
Species: Sus scrofa (pig)
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Laurasiatheria; Artiodactyla; Suina; Suidae;
Sus.

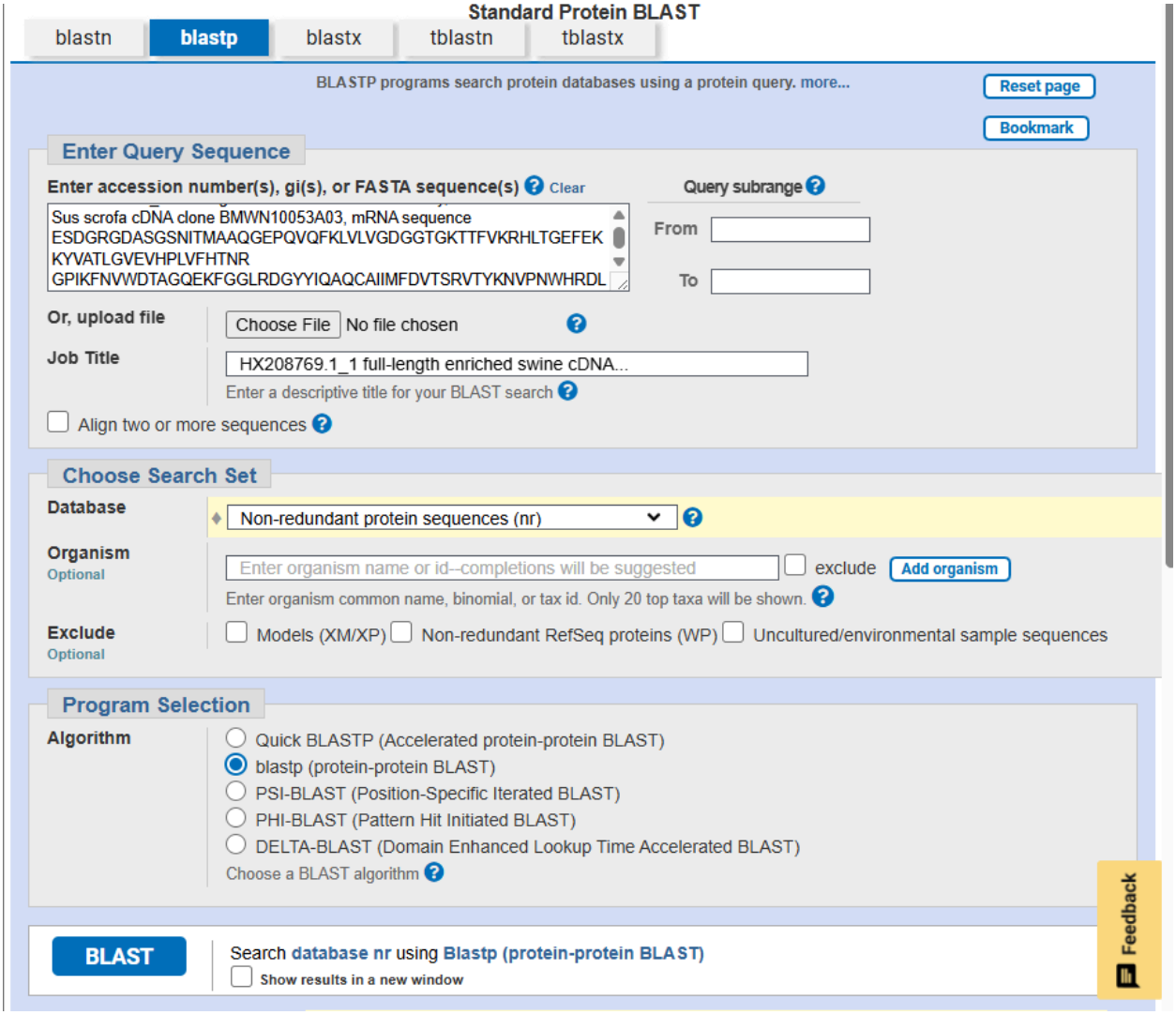
[Q4] Prove that this gene, and its corresponding protein, are novel. For the purposes of this project, “novel” is defined as follows. Take the protein sequence (your answer to [Q3]), and use it as a query in a blastp search of the nr database at NCBI.

• 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

Search details: blastp against nr database



The top hit is the GTP-binding nuclear protein Ran from Camelus dromedariu. The translated Sus scrofa cDNA sequence HX208769 shows 100% amino acid identity to Ran proteins from other mammals, including Camelus dromedarius. But, no identical sequence from Sus scrofa itself was present in the NR database.

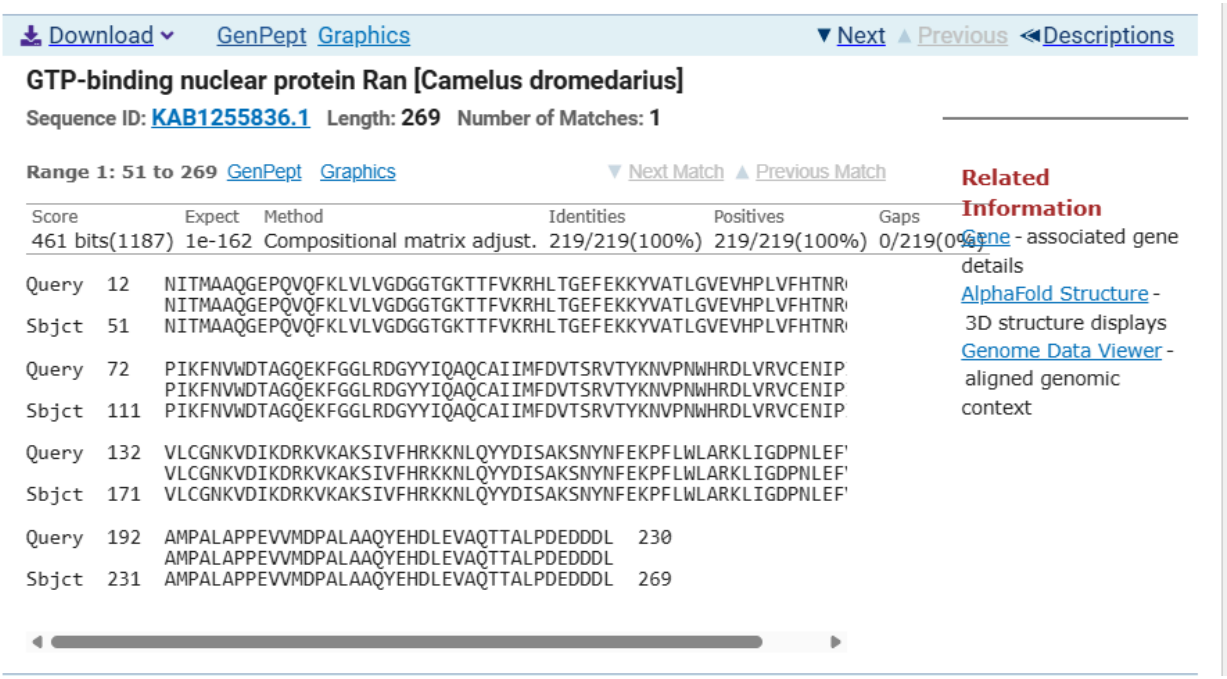
Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Select columns Show 100

select all *100 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
GTP-binding nuclear protein Ran [Camelus dromedarius]	Camelus ...	461	461	90%	1e-162	100.00%	269	KAB1255836.1
Chain A, GTP-binding nuclear protein Ran [Homo sapiens]	Homo sa...	457	457	90%	5e-161	100.00%	261	2N1B_A
GTP-binding nuclear protein Ran [Tupaia chinensis]	Tupaia ch...	456	456	89%	1e-160	100.00%	270	ELW66979.1
Chain A, GTP-binding nuclear protein Ran [Homo sapiens]	Homo sa...	454	454	89%	3e-160	100.00%	237	5DH9_A
GTP-binding nuclear protein Ran [Bos taurus]	Bos taurus	453	453	89%	3e-160	100.00%	216	NP_001029877.1
Chain A, GTP-binding nuclear protein Ran [Homo sapiens]	Homo sa...	454	454	89%	3e-160	100.00%	235	6LQ9_A
Homo sapiens RAN, member RAS oncogene family, partial [sy...	synthetic ...	453	453	89%	3e-160	100.00%	217	AAP36765.1
GTP-binding nuclear protein Ran, partial [Galemys pyrenaicus]	Galemys ...	451	451	89%	2e-159	100.00%	215	KAG8515028.1
PREDICTED: GTP-binding nuclear protein Ran [Elephantulus ...	Elephant...	458	458	90%	3e-161	99.54%	271	XP_006901057.1
GTP-binding nuclear protein Ran [Trichechus manatus latirostris]	Trichechu...	456	456	90%	3e-161	99.54%	235	XP_004385066.2
Chain A, GTP-binding nuclear protein Ran [Homo sapiens]	Homo sa...	454	454	90%	1e-160	99.54%	217	7MNW_A
GTP-binding nuclear protein Ran isoform X2 [Balaenoptera ricei]	Balaenop...	452	452	89%	5e-160	99.54%	216	XP_059750727.1
PREDICTED: GTP-binding nuclear protein Ran-like [Capra hirc...	Capra hir...	452	452	89%	5e-160	99.54%	216	XP_017913784.1
GTP-binding nuclear protein Ran-like [Peromyscus californicus ...	Peromysc...	452	452	89%	5e-160	99.54%	216	XP_052573848.1
Chain A, GTP-binding nuclear protein Ran [Homo sapiens]	Homo sa...	452	452	89%	6e-160	99.54%	216	7MNZ_A
GTP-binding nuclear protein Ran-like [Ovis canadensis]	Ovis cana...	452	452	89%	7e-160	99.54%	216	XP_069426077.1
Chain A, PROTEIN (RAN) [Canis lupus familiaris]	Canis lup...	452	452	89%	8e-160	99.54%	216	1QG4_A
GTP-binding nuclear protein Ran-like [Lagenorhynchus albirost...	Lagenorh...	452	452	89%	9e-160	99.54%	216	XP_060003415.1
hypothetical protein H8959_004262 [Pygathrix nigripes]	Pygathrix ...	452	452	89%	1e-159	99.54%	216	KAL4671553.1
RAN, member RAS oncogene family [Homo sapiens]	Homo sa...	452	452	89%	1e-159	99.54%	216	AAH72000.1
hypothetical protein G4228_008568 [Cervus hanglu yarkanden...	Cervus h...	452	452	89%	1e-159	99.54%	216	KAF4017502.1
GTP-binding nuclear protein Ran-like [Dama dama]	Dama dama	452	452	89%	1e-159	99.54%	216	XP_060985313.1
GTP-binding nuclear protein Ran [Aotus nancymaae]	Aotus na...	452	452	89%	1e-159	99.54%	216	XP_064235613.1
GTP-binding nuclear protein Ran-like [Odocoileus virginianus]	Odocoile...	452	452	89%	1e-159	99.54%	216	XP_020764906.2
GTP-binding nuclear protein Ran-like [Rhinopithecus roxellana]	Rhinopith...	452	452	89%	1e-159	99.54%	216	XP_030786084.1
RAN member RAS oncogene family, partial [synthetic construct]	synthetic ...	452	452	89%	1e-159	99.54%	217	AAX42876.1
GTP-binding nuclear protein Ran-like [Cervus canadensis]	Cervus c...	451	451	89%	1e-159	99.54%	216	XP_043330371.1
GTP-binding nuclear protein Ran-like [Pan paniscus]	Pan panis...	451	451	89%	1e-159	99.54%	216	XP_034792664.1

Feedback



Download GenPept Graphics ▼ Next ▲ Previous ◀◀ Descriptions

GTP-binding nuclear protein Ran [Camelus dromedarius]

Sequence ID: KAB1255836.1 Length: 269 Number of Matches: 1

Range 1: 51 to 269 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
461 bits(1187)	1e-162	Compositional matrix adjust.	219/219(100%)	219/219(100%)	0/219(0%)

```
Query  12   NITMAAQGEPQVQFKLVLVGDGGTGKTTFVKRHLTGEFEKKYVATLGVEVHPLVFHTNR
            NITMAAQGEPQVQFKLVLVGDGGTGKTTFVKRHLTGEFEKKYVATLGVEVHPLVFHTNR
Sbjct  51   NITMAAQGEPQVQFKLVLVGDGGTGKTTFVKRHLTGEFEKKYVATLGVEVHPLVFHTNR

Query  72   PIKFNVWDTAGQEKFGGLRDGYYIQAQCAIIMFDVTSRVTYKNVPNWHRDLVRVCENIP
            PIKFNVWDTAGQEKFGGLRDGYYIQAQCAIIMFDVTSRVTYKNVPNWHRDLVRVCENIP
Sbjct  111  PIKFNVWDTAGQEKFGGLRDGYYIQAQCAIIMFDVTSRVTYKNVPNWHRDLVRVCENIP

Query  132  VLCGNKVDIKDRKVKAKSIVFHRKKNLQYYDISAKSNYNFEKPFLWLARKLIGDPNLEF
            VLCGNKVDIKDRKVKAKSIVFHRKKNLQYYDISAKSNYNFEKPFLWLARKLIGDPNLEF
Sbjct  171  VLCGNKVDIKDRKVKAKSIVFHRKKNLQYYDISAKSNYNFEKPFLWLARKLIGDPNLEF

Query  192  AMPALAPPEVVMDPALAAQYEHDLEVAQTTALPDEDDDL  230
            AMPALAPPEVVMDPALAAQYEHDLEVAQTTALPDEDDDL
Sbjct  231  AMPALAPPEVVMDPALAAQYEHDLEVAQTTALPDEDDDL  269
```

Related Information

Gene - associated gene details

AlphaFold Structure - 3D structure displays

Genome Data Viewer - aligned genomic context

PART 2

[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.

Q5 (3 points) MSA labeled with useful names 1 MSA trimmed appropriately (i.e. no gap overhangs) 1 Pasted MSA fits report page width (i.e. font, format) 1

Re-labeled sequence for alignment:

>Human|4504349|NP_006316.1| GTP-binding nuclear protein Ran isoform 1 [Homo sapiens]

MAAQGEPQVQFKLVLVGDGGTGKTTFVKRHLTGEFEKKYVATLGVEVHPLVFHTNRGPIKFN VWDTAGQEKFGGLRDGYYIQAQCAIIMFDVTSRVTYKNVPNWHRDLVRVCENIPIVLCGNKVD IKDRKVKAKSIVFHRKKNLQYYDISAKSNYNFEKPFLWLARKLIGDPNLEFVAMPALAPPEVV MDPALAAQYEHDLEAQTTALPDEDDDLL

>WildBoar(novel)
ESDGRGDASGSNITMAAQGEPQVQFKLVLVGDGGTGKTTFVKRHLTGEFEKKYVATLGVEVH
PLVFHTNRGPIKFNVWDTAGQEKFGGLRDGYYIQAQCAIIMFDVTSRVTYKNVPNWHRDLVRV
CENIPIVLCGNKVDIKDRKVKAKSIVFHRKKNLQYYDISAKSNYNFEKPFLWLARKLIGDPNLE
FVAMPALAPPEVVMDPALAAQYEHDLEVAQTTALPDEDDDL*ENKAGAQRQKS

>Camel|9838|ref|KAB1255836.1|GTP-binding nuclear protein Ran [Camelus dromedarius]
MYSSPTLGDAERRHPKENVSSECTALSGPLTGLSSPDPKYSMIASLFTTRNITMAAQGEPQV
QFKLVLVGDGGTGKTTFVKRHLTGEFEKKYVATLGVEVHPLVFHTNRGPIKFNVWDTAGQEKF
GGLRDGYYIQAQCAIIMFDVTSRVTYKNVPNWHRDLVRVCENIPIVLCGNKVDIKDRKVKAKSI
VFHRKKNLQYYDISAKSNYNFEKPFLWLARKLIGDPNLEFVAMPALAPPEV
VMDPALAAQYEHDLEVAQTTALPDEDDDL

>Chinesetreeshrew|37347|ref|ELW66979.1|GTP-binding nuclear protein Ran [Tupaia chinensis]
MRTEGVASSAASCPADEPTRRCTAGATSKPRKASQSAPWAGPTRRQVSSDWSDAMAAQGE
PQVQFKLVLVGDGGTGKTTFVKRHLTGEFEKKYVATLGVEVHPLVFHTNRGPIKFNVWDTAG
QEKFGGLRDGYYIQAQCAIIMFDVTSRVTYKNVPNWHRDLVRVCENIPIVLCGNKVDIKDRKVK
AKSIVFHRKKNLQYYDISAKSNYNFEKPFLWLARKLIDPNLEFVAMPALAPPEVVMDPALAAQ
YEHDLEVAQTTALPDEDDDL

>FloridaManatee|127582|ref|XP_004385066.2|GTP-binding nuclear protein Ran [Trichechus manatus latirostris]
MWRPPAASRSPPFLCRTITMAAQGEPQVQFKLVLVGDGGTGKTTFVKRHLTGEFEKKYVATL
GVEVHPLVFHTNRGPIKFNVWDTAGQEKFGGLRDGYYIQAQCAIIMFDVTSRVTYKNVPNWH
RDLVRVCENIPIVLCGNKVDIKDRKVKAKSIVFHRKKNLQYYDISAKSNYNFEKPFLWLARKLI
GDPNLEFVAMPALAPPEVVMDPALAAQYEHDLEVAQTTALPDEDDDL

>RicesWhale|2661301|ref|XP_059750727.1|GTP-binding nuclear protein Ran isoform X2 [Balaenoptera ricei]
MWRPPAASRSPPFLCRTITMAAQGEPQVQFKLVLVGDGGTGKTTFVKRHLTGEFEKKYVATL
GVEVHPLVFHTNRGPIKFNVWDTAGQEKFGGLRDGYYIQAQCAIIMFDVTSRVTYKNVPNWH
RDLVRVCENIPIVLCGNKVDIKDRKVKAKSIVFHRKKNLQYYDISAKSNYNFEKPFLWLARKLI
GDPNLEFVAMPALAPPEVVMDPALAAQYEHDLEVAQTTALPDEDDDL

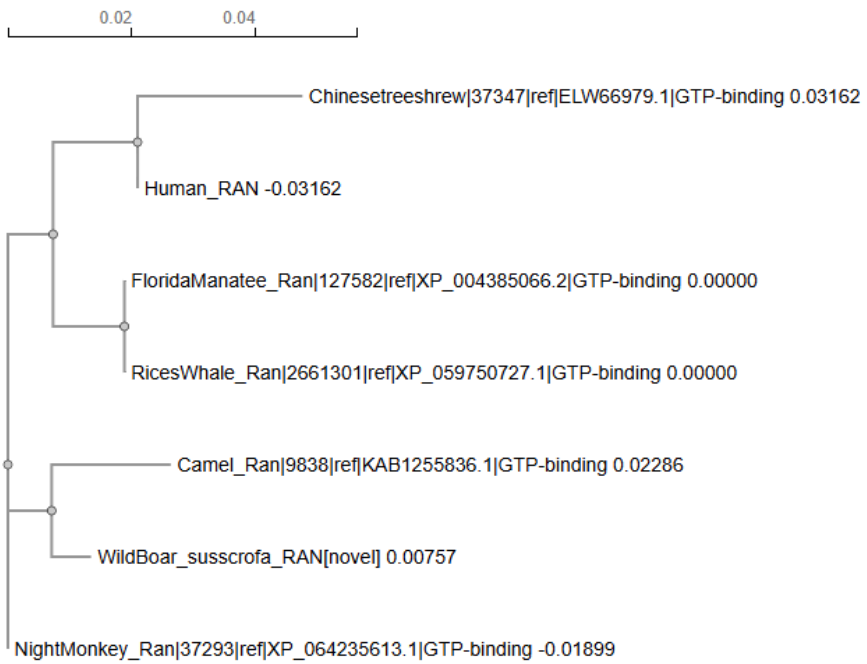
>NightMonkey|37293|ref|XP_064235613.1|GTP-binding nuclear protein Ran [Aotus nancymaae]
MATQGEPQVQFKLVLVGDGGTGKTTFVKRHLTGEFEKKYVATLGVEVHPLVFHTNRGPIKFNV
WDTAGQEKFGGLRDGYYIQAQCAIIMFDVTSRVTYKNVPNWHRDLVRVCENIPIVLCGNKVDI
KDRKVKAKSIVFHRKKNLQYYDISAKSNYNFEKPFLWLARKLIGDPNLEFVAMPALAPPEVVM
DPALAAQYEHDLEVAQTTALPDEDDDL

Results

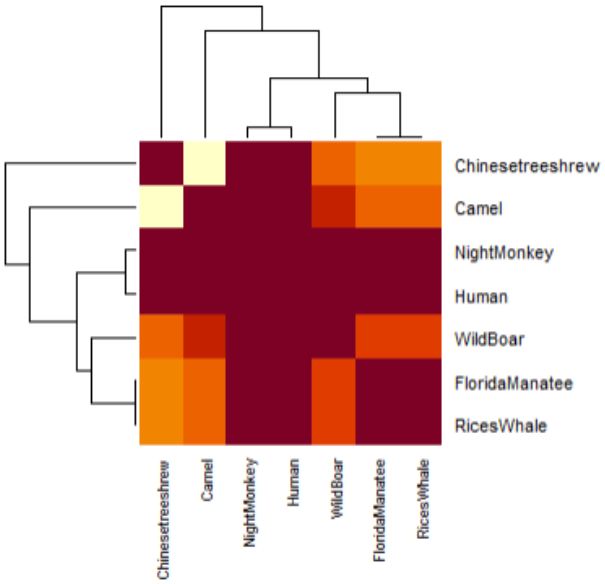
```
clustalw.aln

CLUSTAL 2.1 multiple sequence alignment


Human|4504349|NP_006316.1|          ------------------------------------------------
Camel|9838|ref|KAB1255836.1|GT      -MYSSPTLGDAERRHPKENVSSECTALSGPLTGLSSPDPKYSMIASLFTT
FloridaManatee|127582|ref|XP_0      ----------------------------------MWRPPAASRSPPFLC
RicesWhale|2661301|ref|XP_0597      ----------------------------------MWRPPAASRSPPFLC
WildBoar_novel_                     ---------------------------------------ESDGRGDASG
NightMonkey|37293|ref|XP_06423      ------------------------------------------------
Chinesetreeshrew|37347|ref|ELW      MRTEGVASSAASCPADEPTRRCTAGATSKPRKASQSAPWAGPTRRQVSSD


Human|4504349|NP_006316.1|          ----MAAQGEPQVQFKLVLVGDGGTGKTTFVKRHLTGEFEKKYVATLGVE
Camel|9838|ref|KAB1255836.1|GT      RNITMAAQGEPQVQFKLVLVGDGGTGKTTFVKRHLTGEFEKKYVATLGVE
FloridaManatee|127582|ref|XP_0      RTITMAAQGEPQVQFKLVLVGDGGTGKTTFVKRHLTGEFEKKYVATLGVE
RicesWhale|2661301|ref|XP_0597      RTITMAAQGEPQVQFKLVLVGDGGTGKTTFVKRHLTGEFEKKYVATLGVE
WildBoar_novel_                     SNITMAAQGEPQVQFKLVLVGDGGTGKTTFVKRHLTGEFEKKYVATLGVE
NightMonkey|37293|ref|XP_06423      ----MATQGEPQVQFKLVLVGDGGTGKTTFVKRHLTGEFEKKYVATLGVE
Chinesetreeshrew|37347|ref|ELW      WSDAMAAQGEPQVQFKLVLVGDGGTGKTTFVKRHLTGEFEKKYVATLGVE
                                        **:****************************************


Human|4504349|NP_006316.1|          VHPLVFHTNRGPIKFNVWDTAGQEKFGGLRDGYYIQAQCAIIMFDVTSRV
Camel|9838|ref|KAB1255836.1|GT      VHPLVFHTNRGPIKFNVWDTAGQEKFGGLRDGYYIQAQCAIIMFDVTSRV
FloridaManatee|127582|ref|XP_0      VHPLVFHTNRGPIKFNVWDTAGQEKFGGLRDGYYIQAQCAIIMFDVTSRV
RicesWhale|2661301|ref|XP_0597      VHPLVFHTNRGPIKFNVWDTAGQEKFGGLRDGYYIQAQCAIIMFDVTSRV
WildBoar_novel_                     VHPLVFHTNRGPIKFNVWDTAGQEKFGGLRDGYYIQAQCAIIMFDVTSRV
NightMonkey|37293|ref|XP_06423      VHPLVFHTNRGPIKFNVWDTAGQEKFGGLRDGYYIQAQCAIIMFDVTSRV
Chinesetreeshrew|37347|ref|ELW      VHPLVFHTNRGPIKFNVWDTAGQEKFGGLRDGYYIQAQCAIIMFDVTSRV
                                    ************************************************


Human|4504349|NP_006316.1|          TYKNVPNWHRDLVRVCENIPIVLCGNKVDIKDRKVKAKSIVFHRKKNLQY
Camel|9838|ref|KAB1255836.1|GT      TYKNVPNWHRDLVRVCENIPIVLCGNKVDIKDRKVKAKSIVFHRKKNLQY
FloridaManatee|127582|ref|XP_0      TYKNVPNWHRDLVRVCENIPIVLCGNKVDIKDRKVKAKSIVFHRKKNLQY
RicesWhale|2661301|ref|XP_0597      TYKNVPNWHRDLVRVCENIPIVLCGNKVDIKDRKVKAKSIVFHRKKNLQY
WildBoar_novel_                     TYKNVPNWHRDLVRVCENIPIVLCGNKVDIKDRKVKAKSIVFHRKKNLQY
NightMonkey|37293|ref|XP_06423      TYKNVPNWHRDLVRVCENIPIVLCGNKVDIKDRKVKAKSIVFHRKKNLQY
Chinesetreeshrew|37347|ref|ELW      TYKNVPNWHRDLVRVCENIPIVLCGNKVDIKDRKVKAKSIVFHRKKNLQY
                                    ************************************************


Human|4504349|NP_006316.1|          YDISAKSNYNFEKPFLWLARKLIGDPNLEFVAMPALAPPEVVMDPALAAQ
Camel|9838|ref|KAB1255836.1|GT      YDISAKSNYNFEKPFLWLARKLIGDPNLEFVAMPALAPPEVVMDPALAAQ
FloridaManatee|127582|ref|XP_0      YDISAKSNYNFEKPFLWLARKLIGDPNLEFVAMPALAPPEVVMDPALAAQ
RicesWhale|2661301|ref|XP_0597      YDISAKSNYNFEKPFLWLARKLIGDPNLEFVAMPALAPPEVVMDPALAAQ
WildBoar_novel_                     YDISAKSNYNFEKPFLWLARKLIGDPNLEFVAMPALAPPEVVMDPALAAQ
NightMonkey|37293|ref|XP_06423      YDISAKSNYNFEKPFLWLARKLIGDPNLEFVAMPALAPPEVVMDPALAAQ
Chinesetreeshrew|37347|ref|ELW      YDISAKSNYNFEKPFLWLARKLID-PNLEFVAMPALAPPEVVMDPALAAQ
                                    **********************. ************************


Human|4504349|NP_006316.1|          YEHDLEAQTTALPDEDDDLL-----------
Camel|9838|ref|KAB1255836.1|GT      YEHDLEVAQTTALPDEDDDL-----------
FloridaManatee|127582|ref|XP_0      YEHDLEVAQTTALPDEDDDL-----------
RicesWhale|2661301|ref|XP_0597      YEHDLEVAQTTALPDEDDDL-----------
WildBoar_novel_                     YEHDLEVAQTTALPDEDDDLENKAGAQRQKS
NightMonkey|37293|ref|XP_06423      YEHDLEVAQTTALPDEDDDL-----------
Chinesetreeshrew|37347|ref|ELW      YEHDLEVAQTTALPDEDDDL-----------
                                    ******.  *:   ::** *
```

[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.

Q6 (1 point) Figure illustrates sequence clustering pattern 1



[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 (experimentalTechnique), 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.

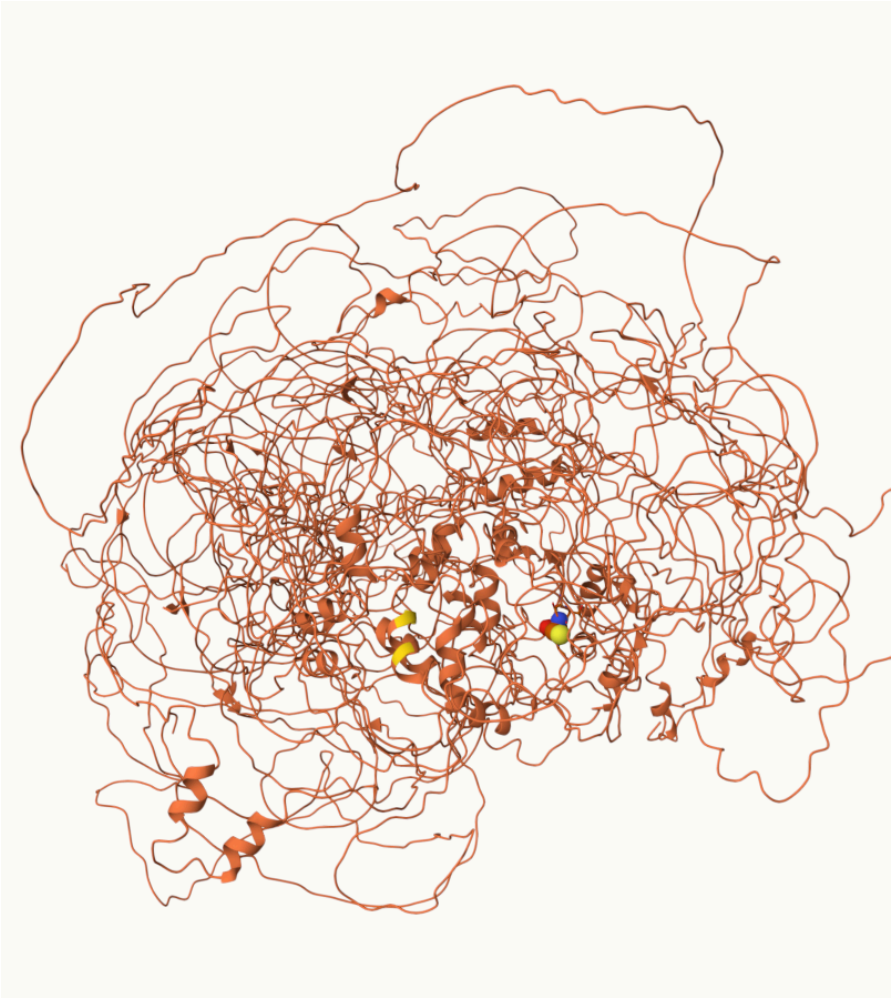
ID	Technique	Resolution	Source	Evalue	Identity
1A2K_C	X-ray Diffraction	2Å	Canis lupus familiaris	3e-161	99.54%
3ICQ_B	X-ray Diffraction	3.2 Å	Saccharomyces cerevisiae (brewer's yeast)	6e-113	89.82%
4DJT_A	X-ray Diffraction	1.8 Å	Encephalitozoon cuniculi GB-M1	2e-55	46.19%

[Q9] Using AlphaFold notebook generate a structural model using the default parameters for your novel protein sequence.

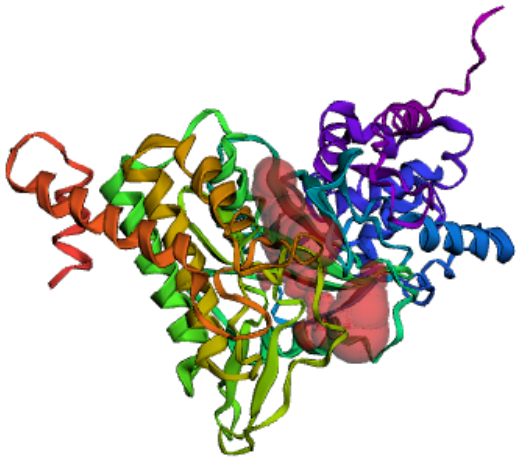
Note that this can take some time depending upon your sequence length. If your model is taking many hours to generate or your input sequence yields a “too many amino acids” (i.e. length) error you can focus on a single domain from your sequence - identify region by searching for PFAM domain matches.

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 should highlight conserved residues that are likely to be functional as spacefill and the protein as

cartoon colored by local alpha fold pLDDT quality score. You can determine conserved residues from the alignment generated by the AlphaFold server and use a conservation cutoff appropriate for the diversity of your protein alignment (e.g. between 60% and 99% conserved). Note that pLDDT 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.).



[Q10] (i) Using your computed structure model (or your closest homologue of known structure from the PDB) predict and locate potential small molecule binding sites using the CASTpFold server (https://cfold.bme.uic.edu/castpfold/). Provide an image or screen-shot of your largest predicted pockets “negative volume” and provide it’s area and volume.



Pocket Info

	Pocket ID	Area (SA) (Å^2)	Volume (SA) (Å^3)
-	1	660.826	381.704

Show negative volume: Negative volume color: Representation style: Cartoon

> Atom Info

(ii)

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]”.

non available as of 12/7

(iii) Briefly discuss (100 words max) the druggability of your novel protein based on:

- Presence of well-defined pockets (output of tools like CASTpFold),
- Existence of known inhibitors for related proteins (your search of ChEMBEL),
- Conservation of binding sites across homologs (your conservation analysis in Q10),
- Potential therapeutic applications if this protein were targeted (you can use ChatGPT, Claude etc. backed up by your reading of the literature here).

My predicted structure contains a large, well-defined pocket (Area: 660.8 Å^2; Volume: 381.7 Å^3), suggesting a ligand-accessible cavity suitable for small-molecule binding.ChEMBL shows no existing inhibitors for this protein. GTP-binding proteins in the Ran family often share conserved nucleotide-binding motifs, implying potential cross-reactivity with known GTP-competitive scaffolds. Conservation across homologs indicates that the binding site is structurally stable, increasing druggability. If targeted, this protein could be inhibited by compounds that disrupt GTP binding or nucleotide-cycling, which may alter nuclear transport or cell-cycle regulation in relevant disease contexts.