BGGN-213: FOUNDATIONS OF BIOINFORMATICS

The find-a-gene project assignment

http://thegrantlab.org/bggn213/

Dr. Barry Grant

Overview:

The find-a-gene project is a required assignment for BGGN-213. You should prepare a written report in **PDF** format that has responses to each question labeled **[Q1] - [Q10]** below. You may wish to consult the scoring rubric at the end of this document and the example report provided online.

The objective with this assignment is for you to demonstrate your grasp of database searching, sequence analysis, structure analysis and the R environment that we have covered in class.

Due Date:

Your responses to questions Q1-Q4 are due at the beginning of **Week 5**. Note that these answers can be obtained very quickly (at best within 10 or 15 minutes), so if you don't succeed at first, just keep trying.

The complete assignment, including responses to all questions, is due at the beginning of **Week 10**. Late responses will not be accepted under any circumstances.

Submission instructions:

Submit your PDF document to GradeScope as directed on our class website. Please do make sure your document is in PDF format and named something like BGGN213_F20_[yourUCSDname].pdf for example, my document would be named BGGN213_F20_bjgrant.pdf

Be sure to include your UCSD email and PID number on the first page of your report.

Submit your preliminary report with answers to Q1-Q4 at the beginning of **week 5** so we can determine if you have found a novel gene. Submit this preliminary report as one document with screen shots of the results inserted appropriately.

See the demonstration report linked to on the course website for an example of format. I will email you my decision; proceed with subsequent questions only after we are sure you have found a novel gene.

For the final report add your results for Q5-Q10 to the preliminary report and submit a final document containing the results for all questions. Please do not submit only Q5-Q10 answers as the final report.

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 its function is known.

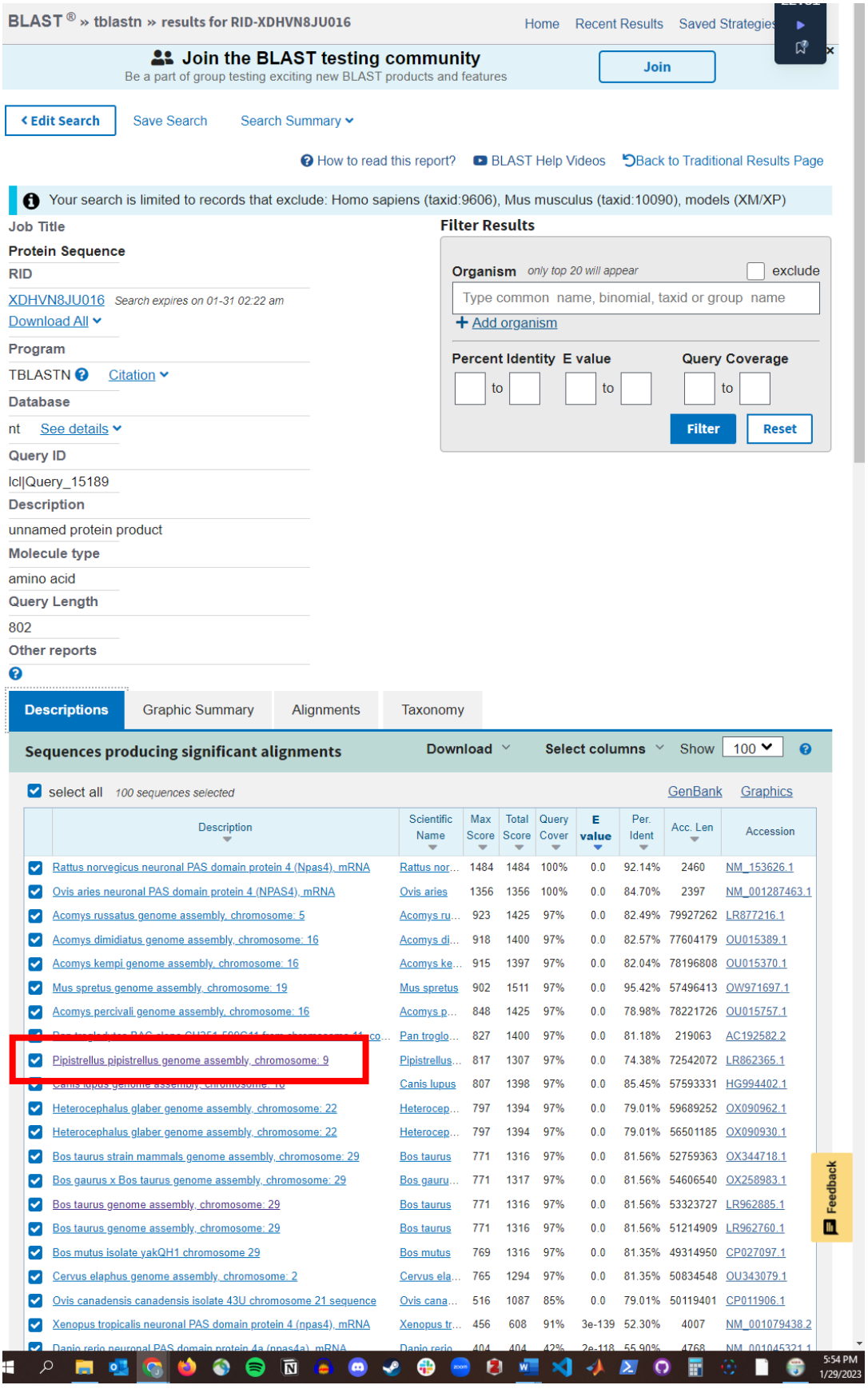
If you do not have a favorite protein, select human RBP4 or KIF11. Do not use beta globin as this is in the worked example report that I provide you with online.

- Protein: NPAS4, accession Q8BGD7
- Species: *Mus musculus*, taxid:10088
- Function: Neuron-specific bHLH-PAS transcription factor. Activity-inducible transcription factor regulating circuit excitation/inhibition balance.

[**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).

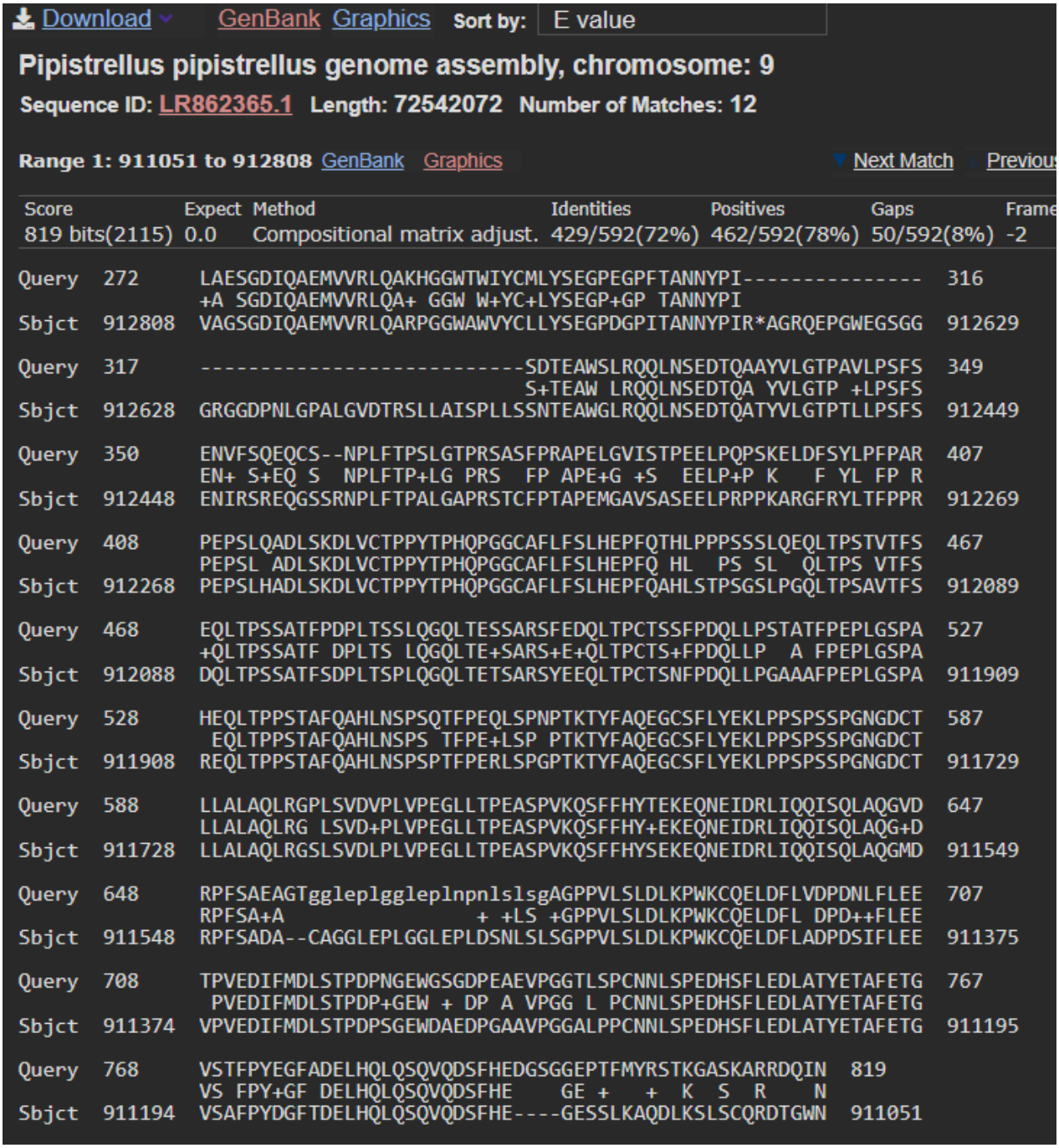
- tblastn search against the NCBI nucleotide (nt) collection. Excluded *Homo sapiens* (taxid 9606), *Mus musculus* (taxid 10088), and predicted transcripts (XM/XP models). Default tblastn parameters.

- **Result:** *Pipistrellus pipistrellus* genome assembly, chromosome 9.
    - Accession: LR862365.1
    - E-value: 0.0
    - Percent identity: 74.4%
    - Coverage: 97%


BLAST® » tblastn » results for RID-XDHVN8JU016
Home
Recent Results
Saved Strategies
Join the BLAST testing community
Be a part of group testing exciting new BLAST products and features
Join
Edit Search
Save Search
Search Summary
How to read this report?
BLAST Help Videos
Back to Traditional Results Page
Your search is limited to records that exclude: Homo sapiens (taxid:9606), Mus musculus (taxid:10090), models (XM/XP)
Job Title
Protein Sequence
RID
XDHVN8JU016 Search expires on 01-31 02:22 am
Download All
Program
TBLASTN Citation
Database
nt See details
Query ID
lcl|Query_15189
Description
unnamed protein product
Molecule type
amino acid
Query Length
802
Other reports
Filter Results
Organism only top 20 will appear
exclude
Type common name, binomial, taxid or group name
Add organism
Percent Identity
E value
Query Coverage
to
Filter
Reset
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
Rattus norvegicus neuronal PAS domain protein 4 (Npas4), mRNA Rattus nor... 1484 1484 100% 0.0 92.14% 2460 NM_153626.1
Ovis aries neuronal PAS domain protein 4 (NPAS4), mRNA Ovis aries 1356 1356 100% 0.0 84.70% 2397 NM_001287463.1
Acomys russatus genome assembly, chromosome: 5 Acomys ru... 923 1425 97% 0.0 82.49% 79927262 LR877216.1
Acomys dimidiatus genome assembly, chromosome: 16 Acomys di... 918 1400 97% 0.0 82.57% 77604179 OU015389.1
Acomys kempi genome assembly, chromosome: 16 Acomys ke... 915 1397 97% 0.0 82.04% 78196808 OU015370.1
Mus spretus genome assembly, chromosome: 19 Mus spretus 902 1511 97% 0.0 95.42% 57496413 OW971697.1
Acomys percivali genome assembly, chromosome: 16 Acomys p... 848 1425 97% 0.0 78.98% 78221726 OU015757.1
Pan troglo... 827 1400 97% 0.0 81.18% 219063 AC192582.2
Pipistrellus pipistrellus genome assembly, chromosome: 9 Pipistrellus... 817 1307 97% 0.0 74.38% 72542072 LR862365.1
Canis lupus 807 1398 97% 0.0 85.45% 57593331 HG994402.1
Heterocephalus glaber genome assembly, chromosome: 22 Heterocep... 797 1394 97% 0.0 79.01% 59689252 OX090962.1
Heterocephalus glaber genome assembly, chromosome: 22 Heterocep... 797 1394 97% 0.0 79.01% 56501185 OX090930.1
Bos taurus strain mammals genome assembly, chromosome: 29 Bos taurus 771 1316 97% 0.0 81.56% 52759363 OX344718.1
Bos gaurus x Bos taurus genome assembly, chromosome: 29 Bos gauru... 771 1317 97% 0.0 81.56% 54606540 OX258983.1
Bos taurus genome assembly, chromosome: 29 Bos taurus 771 1316 97% 0.0 81.56% 53323727 LR962885.1
Bos taurus genome assembly, chromosome: 29 Bos taurus 771 1316 97% 0.0 81.56% 51214909 LR962760.1
Bos mutus isolate yakQH1 chromosome 29 Bos mutus 769 1316 97% 0.0 81.35% 49314950 CP027097.1
Cervus elaphus genome assembly, chromosome: 2 Cervus ela... 765 1294 97% 0.0 81.35% 50834548 OU343079.1
Ovis canadensis canadensis isolate 43U chromosome 21 sequence Ovis cana... 516 1087 85% 0.0 79.01% 50119401 CP011906.1
Xenopus tropicalis neuronal PAS domain protein 4 (npas4), mRNA Xenopus tr... 456 608 91% 3e-139 52.30% 4007 NM_001079438.2
Feedback
5:54 PM
1/29/2023

[**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.

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.



- *Pipistrellus pipistrellus*, chromosome 9. Unknown protein.
- Full subject sequence:
    - RPPGMYRSTKGASKARRDQINAEIRNLKELLPLAEADKVRLSYLHIMSLA CIYTRKGVFFAGGAPLEGPTGLLSTQELEDIVAALPGFLLVFTAEGKLLYL SESVSEHLGHSMVDLVAQGDSIYDIIDPADHLTVRQQLALPSALDTDRLF RCRFNTSKSLRRQSAGNKLVLIRGRFHAHPPGAYWAGNPVFTAFCAPL EPRPRLGPGAGPASLFLAMFQSRHAKDLTLLDISESVLIYLGFERSELLC KSWYGLLHPEDLGHASVQHYRLLAGSGDIQAEMVVRLQAKPGGWAWV YCLLYSEGPDGPITANNYPISYPEAWSLRQQLNSEDTQATYVLGTPTLLP

SFSENIRSREQGSSRNPLFTPALGAPRSTRLPTAPEMGAVSTSEELPRP
PKGRGFRYLTFPPRPEPSLQADLSKDLVCTPPYTPHQPGGCAFLFSLHE
PFQAHLSTPSGSLPGQLTPSAVTFSDQLTPSSATFSDPLTSPLQGQLTE
TSARSYEEQLTPCTSNFPDQLLPGTAAFPEPLGSPACEQLTPPSTAFQA
HLNSPSPTFPEQLSPSPTKTYFAQEGCSFLYEKLPPSPSSPGNGDCTLL
ALAQLRGSLSVDLPLVPEGLLTPEASPVKQSFFHYSEKEQNEIDRLIQQIS
QLAQGMDRPFSADACAGGLEPLGGLEPLDSNLSLSGPPVLSLDLKPWK
CQELDFLADPDNIFLEEVPVEDIFMDLSTPDPSGEWGSEDPGAAVPGGA
LSPCNNLSPEDHSFLEDLATYETAFETGVSAFPYDGFTDELHQLQSQVQ
DSFHEDGSGGEPTF

[**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.~~

Importantly, the first hit is from a different species in the same genus as the query protein sequence!

Job Title

predicted-NPAS4

RID

Y14HRH2T016 *Search expires on 02-07 12:37 pm*

Download All

Program

BLASTP Citation

Database

nr See details

Query ID

lcl|Query_35492

Description

unnamed protein product

Molecule type

amino acid

Query Length

802

Other reports

Distance tree of results Multiple alignment MSA viewer

Filter Results

Organism *only top 20 will appear* exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity to

E value to

Query Coverage to

Filter Reset

Compare these results against the new Clustered nr database BLAST

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
neuronal PAS domain-containing protein 4 [Pipistrellus kuhlii]	Pipistrellus kuhlii	1610	1610	99%	0.0	100.00%	798	XP_036272589.1
neuronal PAS domain-containing protein 4 [Eptesicus fuscus]	Eptesicus fuscus	1513	1513	99%	0.0	95.74%	798	XP_008145138.1
PREDICTED: neuronal PAS domain-containing protein 4 isoform X1 [Myotis davidii]	Myotis davidii	1481	1481	100%	0.0	93.53%	859	XP_006771618.1
PREDICTED: neuronal PAS domain-containing protein 4 [Miniopterus natalensis]	Miniopterus natalensis	1481	1481	99%	0.0	92.50%	800	XP_016063789.1
PREDICTED: neuronal PAS domain-containing protein 4 [Myotis brandtii]	Myotis brandtii	1478	1478	99%	0.0	94.00%	800	XP_005864341.1
neuronal PAS domain-containing protein 4 [Myotis lucifugus]	Myotis lucifugus	1477	1477	99%	0.0	94.00%	800	XP_006101660.1
neuronal PAS domain-containing protein 4 isoform X2 [Myotis myotis]	Myotis myotis	1470	1470	99%	0.0	93.25%	800	XP_036179825.1
neuronal PAS domain-containing protein 4 [Artibeus jamaicensis]	Artibeus jamaicensis	1469	1469	99%	0.0	91.50%	800	XP_036991264.1
neuronal PAS domain-containing protein 4 isoform X1 [Sturnira hondurensis]	Sturnira hondurensis	1462	1462	99%	0.0	91.00%	800	XP_036909295.1
neuronal PAS domain-containing protein 4 isoform X1 [Desmodus rotundus]	Desmodus rotundus	1462	1462	99%	0.0	91.62%	800	XP_024430096.1
Neuronal PAS domain-containing protein 4 [Myotis davidii]	Myotis davidii	1457	1457	98%	0.0	93.32%	811	ELK24301.1
neuronal PAS domain-containing protein 4 isoform X1 [Molossus molossus]	Molossus molossus	1455	1455	99%	0.0	92.38%	800	XP_036115100.1
neuronal PAS domain-containing protein 4 isoform X1 [Myotis myotis]	Myotis myotis	1452	1452	99%	0.0	92.72%	813	XP_036179824.1
neuronal PAS domain-containing protein 4 [Phyllostomus hastatus]	Phyllostomus hastatus	1451	1451	99%	0.0	90.88%	800	XP_045711343.1
neuronal PAS domain-containing protein 4 [Phyllostomus discolor]	Phyllostomus discolor	1449	1449	100%	0.0	90.42%	915	XP_028373096.2
PREDICTED: neuronal PAS domain-containing protein 4 [Hipposideros armiger]	Hipposideros armiger	1445	1445	99%	0.0	91.00%	797	XP_019500743.1
neuronal PAS domain protein 4 [Phyllostomus discolor]	Phyllostomus discolor	1445	1445	99%	0.0	90.50%	800	KAF6105141.1
neuronal PAS domain-containing protein 4 isoform X1 [Pteropus alecto]	Pteropus alecto	1441	1441	99%	0.0	91.25%	800	XP_006911123.1
neuronal PAS domain-containing protein 4 isoform X1 [Panthera tigris]	Panthera tigris	1441	1441	99%	0.0	91.25%	800	XP_007091910.1
neuronal PAS domain-containing protein 4 isoform X1 [Lontra canadensis]	Lontra canadensis	1440	1440	99%	0.0	91.25%	800	XP_032735443.1
neuronal PAS domain-containing protein 4 isoform X2 [Meles meles]	Meles meles	1440	1440	99%	0.0	91.12%	800	XP_045872367.1
neuronal PAS domain-containing protein 4 isoform X1 [Canis lupus dingo]	Canis lupus dingo	1439	1439	99%	0.0	91.25%	800	XP_025302112.3
neuronal PAS domain-containing protein 4 isoform X1 [Prionailurus bengalensis]	Prionailurus bengale...	1439	1439	99%	0.0	91.12%	800	XP_043438548.1
neuronal PAS domain-containing protein 4 isoform X1 [Felis catus]	Felis catus	1439	1439	99%	0.0	91.12%	800	XP_003993721.1
neuronal PAS domain-containing protein 4 isoform X1 [Mustela putorius furo]	Mustela putorius furo	1439	1439	99%	0.0	91.25%	800	XP_004759482.1

Feedback

[**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.

[**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.

[**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.

[Q9] Generate a molecular figure of one of your identified PDB structures using **VMD**. You can optionally highlight conserved residues that are likely to be functional. Please use a white or transparent background for your figure (i.e. not the default black).

Based on sequence similarity. How likely is this structure to be similar to your "novel" protein?

[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?

Scoring Rubric:
[45 total points available]

Q1 (4 points)
Protein name 1 Species 1 Accession number 1 Function known 1

Q2 (6 points)
Blast method 1 Database searched 1 Limits applied 1 Search output list (top hits) 1 Alignment of choice 1 Evalue and other alignment stats 1

Q3 (3 points)
Protein sequence of choice matches Subject above 1 Name in header 1 Species 1

Q4 (3 point)
Blastp output list with identities & Evalue 1 Top alignment shown with alignment statistics 1

Results indicates a "novel" gene found 1

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

Q6 (1 point)
Figure illustrates sequence clustering pattern 1

Q7 (10 points)
Heatmap figure included in report 5 Heatmap is legible (i.e. no labels obscured) 5

Q8 (10 points)
PDB identifiers from multiple species reported 5 Annotation of PDB source, resolution and technique 4 Annotation of Evalue and Sequence Identity 1

Q9 (4 points)
Structure figure provided 2 Uses white background for molecular figure 1 Figure of high resolution (i.e. not just snapshot) 1

Q10 (1 point)
Evidence of ChEMBEL searches 1