Taylor Forman

PID: A5910460

BGGN-213: FOUNDATIONS OF BIOINFORMATICS

The find-a-gene project assignment

http://thegrantlab.org/bggn213/

Dr. Barry Grant

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 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: Retinol binding protein isoform B

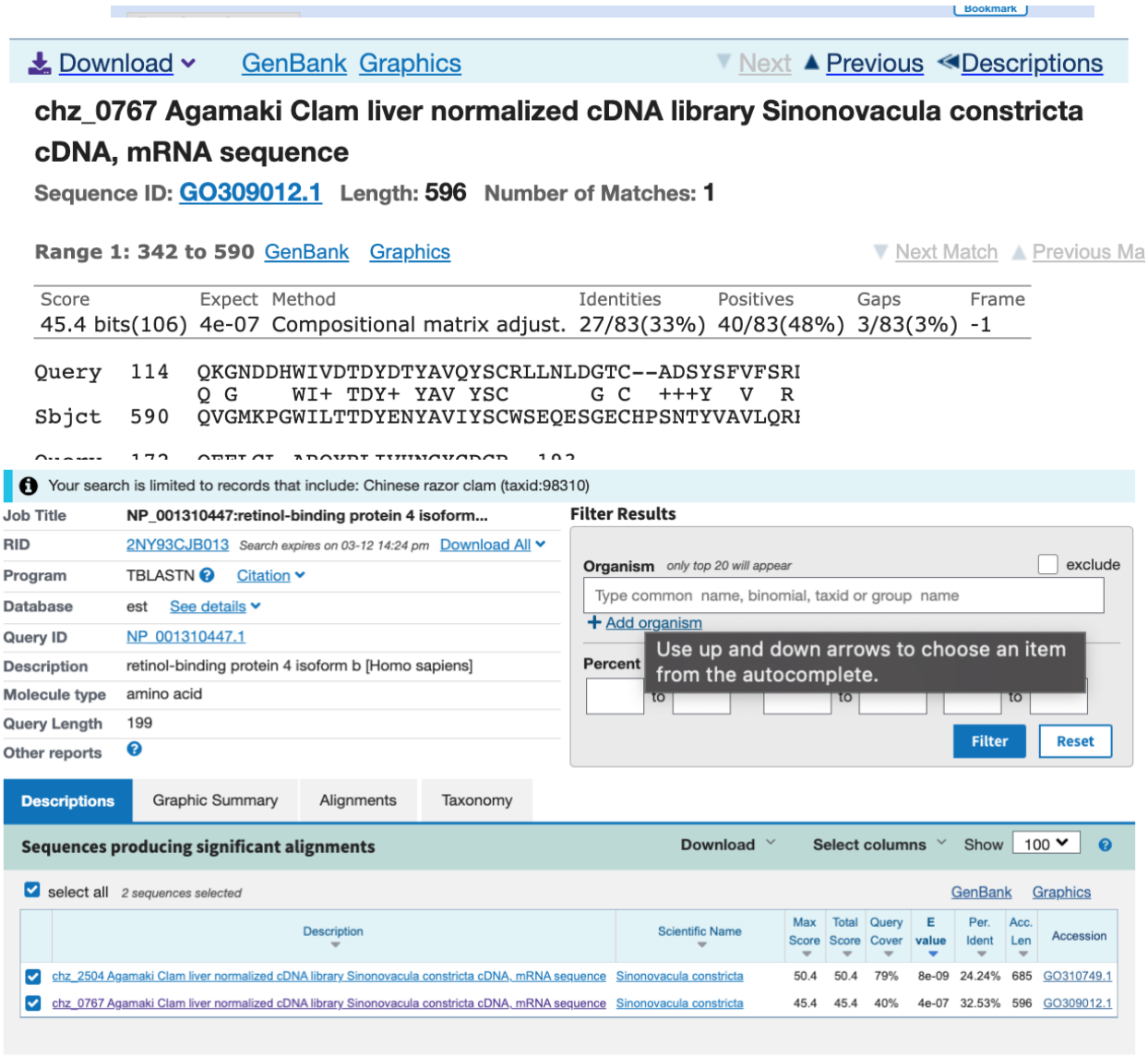
Accession: NP_001310447

Species: Homo sapiens

Known Function: It delivers retinol from the liver stores to the peripheral tissues. In plasma, the RBP-retinol complex interacts with transthyretin which prevents its loss by filtration through the kidney glomeruli.

[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 (2.7.1) search against Chinese razor clam ESTs
Database: Expressed Sequence Tags (est)
Organism: Chinese razor clam (taxid: 98310)



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.

[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:

> chz_0767 Agamaki Clam liver normalized cDNA library Sinonovacula constricta cDNA, mRNA sequence

QVGMKPGWILTTDYENYAVIYSCWSEQESGECHPSNTYVAVLQRKTDDISPSHRVEID
RALRRACVEPKKLSKITHYGYCLGR

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: chz_0767 Agamaki Clam liver normalized cDNA library Sinonovacula constricta cDNA, mRNA sequence

Species: Sinonovacula constricta

Eukaryota; Metazoa; Spiralia; Lophotrochozoa; Mollusca; Bivalvia;

Autobranchia; Heteroconchia; Euheterodonta; Imparidentia;

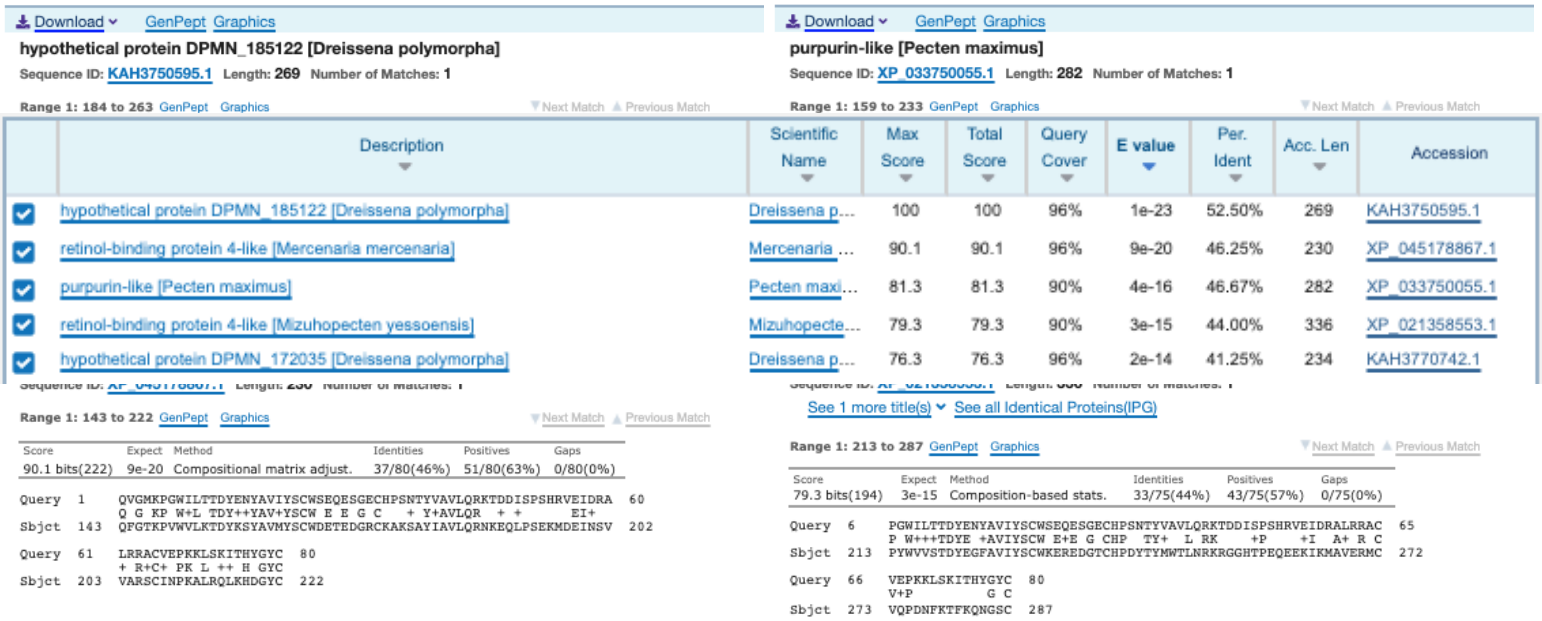
Neoheterodontei; Cardiida; Tellinoidea; Solecurtidae; Sinonovacula.

[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.



[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.

Sequences for alignment:

>NP_001310447.1 retinol-binding protein 4 isoform b [Homo sapiens]
MNYSKIPAQVDLRRQTERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDET
GQMS
ATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAV
QYSCRLLN
LDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL

> chz_0767 Agamaki Clam liver normalized cDNA library Sinonovacula constricta cDNA, mRNA sequence (taken from BLAST)
QVGMKPGWILTTDYENYAVIYSCWSEQESGECHPSNTYVAVLQRKTDDISPSHRVEIDRALRRACV
EPKKLSKITHYGYCLGR

>FQ660773 Crassostrea gigas library (Genoscope - CEA) Crassostrea gigas cDNA clone WY0AAA57YO16FM1, mRNA sequence
RAQTGTCPVVSSISVQPGFDYESLANESRWNVVLYS--
KIIPIEGLDMQVVKSDVSLMFSRDADNNKTVTLAGRVRLASSFAFCLKLQGSVEDSSEVPAKLRQSFY
NPLTN-QFLDFNFWILHTDYTNLAVVYACEKIMAADGTCDPGSSYMWTLSRGTSHTAAERERI-
QEIFQSMCLDMTSLRQIEHSDVC

>ai_D001.69 Bay Scallop planktonic veliger larvae ZAP Express Library Argopecten irradians cDNA 5' similar to PURP_CHICK PURPURIN PRECURSOR gi|86420|pir|A26969 re, mRNA sequence
DAMWVVSTDYDGYAVTYGCDKVLPESGYCDPSKEAVYTLNRRQDGHTKQQLIKIENALNSVCVSA
RTLRPMQQIGEC

>HCINT08C09 hard clam SMART cDNA library from intestine Meretrix meretrix cDNA 5', mRNA sequence
KVQENFDLSKYLGRWYENRRYSNLFSLFSNCVTAEYTALSETSVRVNNTGWKYLSNYYDNAIGEAIV
MSDGKLGVRF-------SEF--
QPYEDYWVLSTDYTSYSIVWSCMETPKGPIQFNSQYLWILSRSPDGVPDEQLQKIM

>AM877207 mge013 Ruditapes philippinarum cDNA clone mge013P0007L12 5', mRNA sequence
KCKNFTTQPNFDVKRYAGGWYDIEKTFFAGQMNKSCVKAEYSLRDDGKIDVLNQDYTAELHEENTT
G--IAFYKDPNVKSKLTVKL-GTSP-----EANYWIVETDYDTYALIWSCA--ELEGIAHADIGWILGRKQR-
--LDENLITRLKQKLTSLGLNYE

>pmaximaP0017C17_654 Adult silver lipped oyster (Pinctada maxima) Pinctada maxima cDNA 5', mRNA sequence

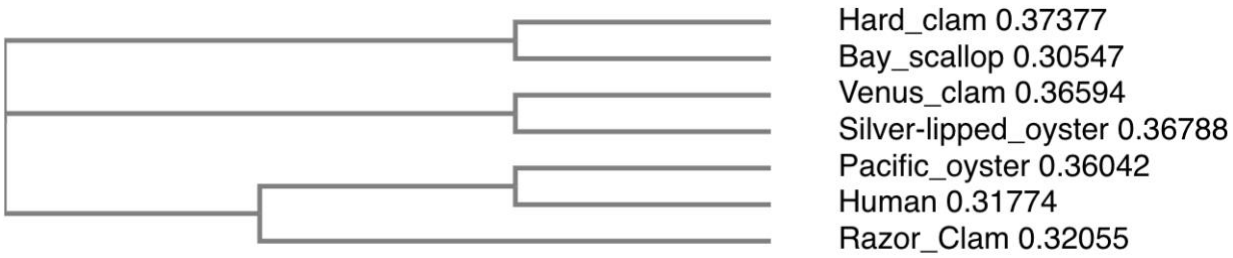
KDCVISNFQTQSNFEADKFVGKWYEIEWMTHQAENPNDFW--
DDYVTNYTLNDDGSFSLFTAFRSN--PNKTICSLQNAVMYRTSN-AKYDV---
AVSSCRQIRHSPQWIISTDYIRYAIIYSCHVQNIDGTCKTWVAKTFSR-
KRTLDDRYISLAHDTYKDLCLNRH

Alignment:

```
CLUSTAL O(1.2.4) multiple sequence alignment


Hard_clam                -------------------------KVQENFDLSKY--LGRWYENRRYSNL--------F
        25
Venus_clam               --------------------KCKNFTTQPNFDVKRY--AGGWYDIEKTFFA--------G
        30
Silver-lipped_oyster     -----------------KDCVISNFQTQSNFEADKF--VGKWYEIEWMTHQA---ENPND
        38
Pacific_oyster           ------------RAQTGTCPVVSSISVQPGFDYESLANESRWNVVLYS--KIIPIEGLDM
        46
Human                    MNYSKIPAQVDLRRQTERDCRVSSFRVKENFDKARF--SGTW----YAMAKKDP-EGL--
        51
Razor_Clam               ------------------------------------------------------------
        0
Bay_scallop              ------------------------------------------------------------
        0


Hard_clam                SLFSNCVTAEYTALSETSVRVNNTGWKYLSNYYDNAIG-EAIVMSDGKLGVRF-------
        77
Venus_clam               QMNKSCVKAEYSLRDDGKIDVLNQDYTAELHEENTTG---IAFYKDPNVKSKLTVKL-GT
        86
Silver-lipped_oyster     FW--DDYVTNYTLNDDGSFSLFTAFRSN--PNKTICSL-QNAVMYRTSN-AKYDV---AV
        89
Pacific_oyster           QVVKSDVSLMFSRDADNNKTVTLAGRVRLASSFAFCLKLQGSVEDSSEVPAKLRQSFYNP
        106
Human                    -FLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDT-EDPAKFKMKYWGV
        109
Razor_Clam               ------------------------------------------------------------
        0
Bay_scallop              ------------------------------------------------------------
        0


Hard_clam                SEF--QPYEDYWVLSTDYTSYSIVWSCMET--PKGPIQFNSQYLWILSRSPDGVPDE--Q
        131
Venus_clam               SP-----EANYWIVETDYDTYALIWSCA---ELEGIAHADIGW--ILGRKQR---LDENL
        133
Silver-lipped_oyster     SSCRQIRHSPQWIISTDYIRYAIIYSCHVQ-NIDGTCKTWVAK--TFSR-KRTLDDR--Y
        143
Pacific_oyster           LTN-QFLDFNFWILHTDYTNLAVVYACEKIMAADGTCDPGSSYMWTLSRGTSHTAAE--R
        163
Human                    ASFLQKGNDDHWIVDTDYDTYAVQYSCRLL-NLDGTCADSYSF--VFSRDPNGLPPE--A
        164
Razor_Clam               ----QVGMKPGWILTTDYENYAVIYSCWSE-QESGECHPSNTYVAVLQRKTDDISPS--H
        53
Bay_scallop              --------DAMWVVSTDYDGYAVTYGCDKVLPESGYCDPSKEAVYTLNRRQDGHTKQ--Q
        50
                                    *:: ***   :: :.*      .*           : *

Hard_clam                LQKIM--------------------------------      136
Venus_clam               ITRLKQKLTSLGLNYE--------------------      149
```

```
Silver-lipped_oyster      ISLAHDTYKDLCLNRH--------------------      159
Pacific_oyster            ERI-QEIFQSMCLDMTSLRQIEHSDVC---------      189
Human                     QKIVRQRQEELC-LARQYRLIVHNGYCDGRSERNLL      199
Razor_Clam                RVEIDRALRRACVEPKKLSKITHYGYCLGR------      83
Bay_scallop               LIKIENALNSVCVSARTLRPMQQIGEC---------      77
```

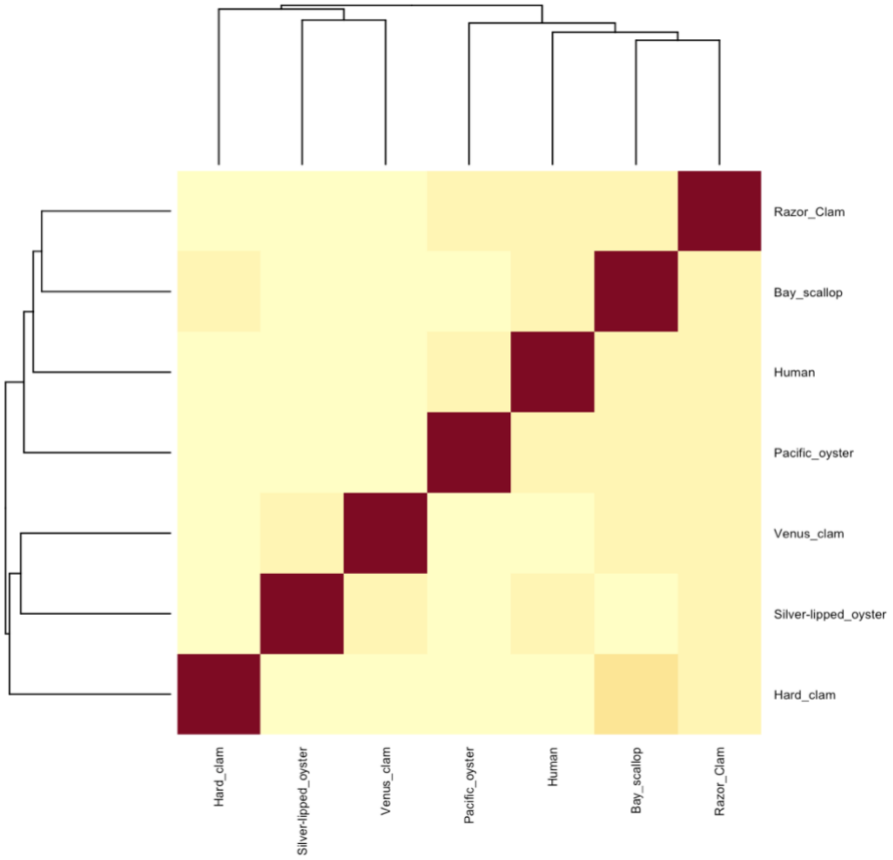
[Q6] Create a phylogenetic tree, using either a parsimony or distance-based approach. Bootstrapping and tree rooting are optional. Use “simple phylogeny” online from the EBI or any respected phylogeny program (such as MEGA, PAUP, or Phylip). Paste an image of your Cladogram or tree output in your report.



[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 into R with the help of the functions in **Bio3D package**.



Calculate a sequence identity (again within the Bio3D package). Then generate a heatmap 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

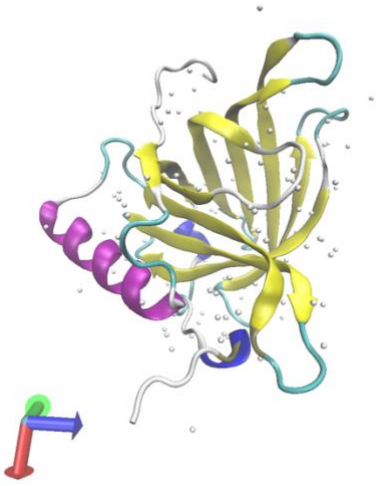
you could chose the sequence with the highest identity to all others in your alignment by calculating the row-wise maximum from your sequence identity matrix.

Description: df [3 × 7]

structureId <chr>	experimentalTechnique <chr>	resolution <dbl>	scopDomain <chr>	source <chr>	identity <dbl>	evalue <dbl>
1AQB	X-ray	1.65	Retinol binding protein	Sus scrofa domesticus	55.556	2.4
1JYJ	X-ray	2.00	Retinol binding protein	Homo sapiens	21.569	3.9
6QBA	X-ray	1.80	automated matches	Homo sapiens	23.636	8.4

3 rows

[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?

Based on the sequence similarity (~55%), I believe this structure is similar to my novel protein. It is also a RBP4 protein, though from a *Sus scrofa domesticus* (pig). Because of the similarity in protein function to my original query, I suspect it is similar to the novel protein, too. It should be noted that the Evalues of the top three results are relatively high, which suggests that this is a good starting point for similarity but may not actually indicate true homology, particularly as the list descends from top hit to third hit.

[**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?

ChEMBL indicated 22 potential assays based on the top hit for the novel sequence, 19 of which are binding assays and 3 that are functional (https://www.ebi.ac.uk/chembl/target_report_card/CHEMBL3100/). An antagonist was developed with activity at maltose binding protein-tagged RBP4 expressed in E. coli that was able to inhibit retinol-induced protein activity.

Design, Synthesis, and Evaluation of Nonretinoid Retinol Binding Protein 4 Antagonists for the Potential Treatment of Atrophic Age-Related Macular Degeneration and Stargardt Disease

Christopher L. Cioffi, Nicoleta Dobri, Emily E. Freeman, Michael P. Conlon, Ping Chen, Douglas G. Stafford, Daniel M. C. Schwarz, Kathy C. Golden, Lei Zhu, Douglas B. Kitchen, Keith D. Barnes, Boglarka Racz, Qiong Qin, Enrique Michelotti, Charles L. Cywin, William H. Martin, Paul G. Pearson, Graham Johnson, and Konstantin Petrukhin

Journal of Medicinal Chemistry 2014 57 (18), 7731-7757

DOI: 10.1021/jm5010013

Associated Assays