**BIFX-550  
S. Ravichandran, Ph.D.  
Biology, Hood College, Frederick, MD 21702**

Ongoing Assignment-bifx550c

Your Name: James Jedediah Smith

There will be three ongoing assignments, bifx550a-c. Assignments will be released as we progress through the class. These assignments are set up to help the students complete BIFX550’s find-a-gene project. You will use your find-a-gene query (gene/protein) to answer these questions. If you are working on bifx550a, please download the file, rename it as your firstinitiallastname\_bifx550a.docx (example, SRavichandran\_bifx550a.doc). Complete the assignment and upload the document to Hood College BIFX550 class BlackBoard site before the deadline

*Please note when I say, “this gene” or “this protein”, it means your query gene/protein respectively.*

**Assignment bifx550c (for due date lookup Blackboard announcement pages)**

At this point, you should have identified your novel gene. This assignment is all about your final results.

1. Remind me the name of a gene that you are working on. Include the species and the accession number. (You may have answered this question, but please provide them again here for the sake of completion. Thanks.)  
     
   Midline 1 (MID1). The NCBI ID is 4281. It is ‘from’ humans.
2. Prove that this gene, and its corresponding protein, are novel. For this project, “novel” is defined as follows. Take the protein sequence and use it as a query in a blastp search of the nr database at NCBI. Please make sure you satisfy the novel-gene conditions that are described in the Find-a-gene-project document (available from BlackBoard). Please provide the protein sequence in fasta format below.

Novel gene was identified in *Eubalaena glacialis*.

>gnl|SRA|SRR10251454.3998192.2

CTGTTTGTCTTCAACTTCCCAGGCTCACTGCTGCGGCTGCCCGCCTGGTTGATGGCCTTGACAATGAAGATGTACTTGGTGCCGCTCTGCAGACCGTGGACGGTGTAGTGGTTCTGCTTGATGTTGGGCACGATCATCCAGCTATCGGCCG

>lcl|ORF5|Novel\_MID1\_Eubalaena\_glacialis

ADSWMIVPNIKQNHYTVHGLQSGTKYIFIVKAINQAGSRSSEPGKLKTN

Graphical user interface, application

Description automatically generated

1. Graphical user interface, text, application, email

   Description automatically generatedGenerate a multiple sequence alignment with your novel protein, your original query protein (novel protein identified in step b), and a group of other members (homologous) of this family. A typical number of proteins to use in a multiple sequence alignment is a minimum of 5 or 10 and a maximum of 30, although the exact number depends on your novel gene.

1. Graphical user interface, text, application

   Description automatically generatedCreate a phylogenetic tree, using either parsimony or distance-based approach. Bootstrapping and tree rooting is optional. Use any program like MEGA X or Jalview.  
     
   I attempted to create a parsimony tree in MEGA X, but it appeared to be randomized. Every time I generated the tree, it was something different. I double checked the certainty for most of them, and it labeled most of the connections as “ambiguious.” I then generated a maximum likelihood tree with bootstrapping to get a hopefully more confident and less random tree. The left tree is the parsimony tree, the right is the maximum likelihood one.

Graphical user interface, text, application

Description automatically generated

1. You can use the online protein prediction servers (see below for more information) for creating a 3D structure for your novel protein. Compare the predicted structure of your novel protein to that of a known 3D structure from your gene family. For example, if I am modeling HBB family and my starting query protein is human HBB, I will compare my novel HBB protein (say, from the organism sea worm) with human HBB 3D structure. Hint: Usually the protein structures are available from Protein Data Bank database ([www.rcsb.org](http://www.rcsb.org)); you can also get this information from UniProt.

Phyre2 is a free, user-friendly protein modeling server. Here is the link: <http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index> How does this work:

Very easy to use. After you have discovered your novel protein, you can take the fasta sequence and paste it into the box from the server to get a model (3D).

Map

Description automatically generatedGraphical user interface, application, Word

Description automatically generatedLeft image is the predicted structure of my novel sequence from Phyre2. Right image is the [AlphaFold](https://alphafold.ebi.ac.uk/entry/O15344) entry for MID1. The PDB did not have a model of MID1 that included the region my novel sequence was found inside. The big yellow Alpha Helix in the right image is represted in the curly blue bit of the left image. As you can see, it appears that my novel sequence on the left seems to have a less clearly defined helix structure. However, the subsequent Beta Sheets seem to be very similar to the ones found in the regular human MID1.

1. Discuss the significance of your novel gene. What have you learned about this gene/protein family?

The gene as a whole plays an important role in microtubual formation. When mutations occur that disrupt this function, it often leads to Opitz G/BBB Syndrome. This typically happens around the the C-terminus of MID1. As the AlphaFold or a simple alignment will easily reveal, the section of novel protein that I have found is right in the middle of MID1, and thus unlikely to play a role in causing Opitz G/BBB Syndrome.

My novel orthlolog was located in a whale species on the verge of extinction. While I doubt that this new information will do anything to save them, the fact it can be found in such a diverse array of organisms certainly speaks to how important MID1 is to the development of complex organisms. It would be interesting to see what type of organisms lack MID1 entirely or whether Opitz G/BBB Syndrome is a uniquely human disease.