BIMM 104: **Genome Informatics**

Course Instructor: Dr. Barry J. Grant (bjgrant@umich.edu)
Class Website: https://bioboot.github.io/bimm104_w18/

Find A Gene Mid-Term Assignment (Draft)

Steps 1-4 due at the beginning of class, Thursday, February 20, 2018.
Final report due at the beginning of class, Thursday, March 6, 2018.
(The following is taken from the web documents that accompany Jonathan Pevsner's text, *Bioinformatics and Functional Genomics*, 2nd Ed., 2008. It has been adapted for this course.)

The purpose of this assignment is for you to grasp the principles of database searching and sequence analysis that we cover in the course. You should prepare a written report in a Word document that has the following components (1 – 9 below). Email your document as an attachment named *BIMM104_W18_[yourUniqueName].docx* to me (bjgrant@umich.edu). For example, my document would be named *BIMM104_W18_bjgrant.docx*

This assignment accounts for 10% of your final course grade.

Email me a preliminary report on steps 1-4 by February 20 so I can determine if you have found a novel gene. Steps [1] to [4] can be accomplished very quickly, so if you don't succeed at first, just keep trying. Submit the preliminary report as **one** document with screen shots of the results inserted appropriately. See the Example answer linked on the web page for an example of format. I will email you my decision; proceed with steps 5-7 **only** after we are sure you have found a novel gene.

For the final report add your results for steps 5-9 to the preliminary report and send a final document containing the results for **all** steps. Please do not send only steps 5-9 as the final report.

Step [1] Tell me the name of a protein you are interested in. Include the species, the accession number and the function of the protein. This can be a human protein or a protein from any other species as long as it's function is known.

Step [2] Perform a BLAST type search against a DNA database, such as a database consisting of genomic DNA or ESTs. The BLAST server can be at NCBI or elsewhere; you can use **PSI-BLAST** if you want. You may want to use **tblastn** to insure that all reading frames are searched. Include the alignment output of that BLAST search in your document (abbreviated, if necessary, to the relevant part). 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. 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 the *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 must just be labeled a "genomic clone" or "mRNA sequence", etc. - but no functional annotation.

In general, step [2] 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 step [4]), and a non-homologous result.

If you are having trouble finding a novel gene try looking at an organism that is poorly annotated.

Step [3] Gather information about this “novel” protein. Show me the *protein* sequence of the “novel” protein. If the protein sequence is not displayed on any of the linked web pages you may have to translate your novel DNA sequence. Use a Python script or on on-line service (e.g. www.ebi.ac.uk/Tools/st/emboss_transeq/ or http://web.expasy.org/translate/). 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 translation you choose includes the same sequence as in the Sbjct line of the alignment included from Step 2. In some cases, you will be able to do further BLAST searches to obtain even more sequence of your novel gene.

In this step, 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.

Step [4] Prove that this gene, and its corresponding protein, are novel. For the purposes of this project, “novel” is defined by the following criteria:

Take the protein sequence (step [3]), and use it as a query in a **blastp** search of the nr database at NCBI. Show the top hits.

> a. If there is a match with 100% amino acid identity to a protein in the database, from the *same species*, then this new gene sequence is not novel. Sometimes you may obtain a match with ~98% identity from the same species and this could be due to sequences errors; it will be a judgment call depending on other factors.

b. If there is a match with 100% identity, but to a *different species* than the one you started with, then you may have succeeded in finding a novel gene.

c. If there is no match with 100% identity, then it is likely that your protein is novel.

In general, if you determine the name or function of your candidate gene from the header information returned from a sequence database you do not have a novel gene.

Step [5] (30%) Generate a multiple sequence alignment with your novel protein, your original query protein, and a group of protein of this family from different species. A typical number of proteins to use in a multiple sequence alignment is a minimum of 5 or 10 and a maximum 30, although the exact number is up to you. You may want to use PSI-BLAST to obtain enough related sequences. Include the MSA in your report. Use Courier font with a size appropriate to fit page width.

Indicate your sequence in the alignment by choosing an appropriate name for each sequence in the sequence file (i.e. **edit the sequence files so that the species, or short common, names (rather than accession numbers) display in the MSA and in the phylogenetic tree plot below**). Trim your sequences so that the ends don't include gaps. This is most easily accomplished by editing the separate sequences in the file you input to the MSA program. Use a constant width font such as Courier to display your MSA. The goal in this step is to create an interesting an MSA for building a phylogenetic tree that illustrates species divergence.

Step [6] (20%) Create a phylogenetic tree, using the CLUSTAL program of your choice (command line clustalww2, ClustalX, Clustal Omega). Evaluate your tree with bootstrapping, display the bootstrap values on the tree using NJPlot (or other app of your choice), save the tree as a PDF file and insert the PDF into your Word document. Use appropriate names in the sequence files so that comparable names are in both the MSA and the tree. Use short, common names (or abbreviate *G. sp.*) for the species so that the branches in your tree will be easily identified. If your novel gene species is an outlier you need to go back to step 5 and choose a variety of species that are closer to your novel sequence. The goal is to create a tree that illustrates how your novel gene species fits into phylogenetic classifications.

Step [7] (20%) Discuss the relationships shown in the phylogenetic tree plot. Use NCBI Taxonomy Identifier and give the hierarchical classification for at least one major clade in your tree (e.g. KPCOFGS or similar). Discuss whether or not the tree branches agree with the taxonomic classification.

~~**Step [8]** (10%) Compare the predicted structure of your protein to that of a known structure. Use Swiss-Model to create a homology model. If you are successful in obtaining a homology model show the evaluation scores from QMEAN. Include a molecular graphics image of your model protein superpositioned on the template.~~ (**Please**

~~**use white background in PyMOL for printing**~~). ~~Discuss the correlation of residue evaluation scores (the low scorers in the Local Quality Estimate plot) with regard to the model superposition on the template structure.~~

~~If no template can be found and Swiss-Model fails indicate that in your report. If you have a transmembrane protein show the results of the TMHMM server analysis.~~

Enter your protein sequence in **predictprotein.org** and report on any significant motifs or functional attributes of your protein.

Step [8] (30%) Discuss the significance of your novel gene. What have you learned about this gene/protein family?

Scoring Rubric

Step 1 (4 points)	
Protein name	1
Species	1
Accession number	1
Function	1
Step 2 (2 points)	
Search output list (top hits) with choice indicated	1
Alignment with header of choice	1
Step 3 (3 points)	
Protein sequence of choice matches Sbjct above	1
Name in header	1
Species	1
Step 4 (1 point)	
Blastp output list with identities	1
Step 5 (6 points)	
MSA organism/species/common names	2
MSA trimmed (no gaps, overhangs)	2
MSA font fits width	2
Step 6 (4 points)	
Illustrates phylogenetic distribution	2
Evaluation shown	2
Step 7 (2 points)	
Taxonomy listed for novel gene	1
Discussion of tree versus taxonomy	1
Step 8 (14 points)	
If model,	
Show Model-Template Alignment with fit color key	2
Show molecular graphic alignment (poor side chains as sticks)	2
Show QMEAN4 scores, Local Quality, Normalized Score	2
Discussion of poor segments	6
If transmembrane,	
Show TMHMM plot	6
Discuss TM predictions	6
Predictprotein results	2
Step 9 (4 points)	4