(Preliminary document. A word document will be updated in the next few days)

BIFX-550: Assignment

Your Name:

This will be an ongoing assignment. You will use your find-a-gene project (gene/protein) as your query to answer these questions. When you download this file, rename it as firstinitiallastname_bifx550a.docx (example, SRavichandran_bifx550a.doc). I believe, this ongoing assignment can help you with your final project.

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

I. Part I: This section is due on October 11, 2018

1. What is the gene/protein that you have chosen for your final project? Please provide NCBI and Ensembl IDs.

2. Are there any alternate names for this gene?

3. On what chromosome is this gene located and is it a positive or negatively stranded gene? What NCBI database do you go to get this information?

4. Compare and comment about the transcripts between Ensembl and NCBI?

5. Please provide the protein IDs (UniProt, NCBI and Ensembl) of the query protein that you will be using for your find-a-gene search and provide its sequence length (i.e. how many amino acids?)

6. Briefly explain what the functions are for the gene? *For example, Caspase-9 gene is an enzyme that is critical for apoptotic pathway in many tissues*.

7. Are there any known homologs for your gene/protein? *We will cover Homology in the Oct 4 class.*

8. For a normal (healthy) individual, at what sites (lung, heart etc.) is this protein expressed? Where in NCBI is this information stored? Do you know whether there is any association of the query protein expression with any disease (or disorder) states (i.e. overexpression for cancer patients (compared to healthy individual) is noted on liver? If so, please provide them.

9. Is there a **TATA-[A/T]-A** motif in your corresponding transcript? If so, give the transcript ID (NCBI/Ensembl) and the position range (ex. 130,000-130,005) of this motif. Please indicate the Assembly (Example: GRCh38) that you are using for this question. For example, if my mRNA is NM_001042594 then my corresponding will be NP_001036059 and my position could be 130,000-130,005. Note the position number shown here is an example range and not the true motif location.

10. Is there a major pathway you can associate your gene to?

11. Do you have any information about what protein(s) your query protein might interact with? Where do you go to get this information (NCBI or Ensembl)?

12. What conserved domains are known for your protein? Domains are conserved part of a protein sequence. Its 3D structure can independently exist, evolve and function. For example, kinase domain (function: phosphorylation) has

been found in many protein families. There are roughly 53 million domains classified into some 2000 super families. We covered this concept in class 3 and will go over again in later classes.
Link for kinase domain: https://en.wikipedia.org/wiki/Protein_kinase_domain

13. Are there any disease(s) associated with this gene/protein? Can you cite 1/2 relevant review/paper (PubMed links would be OK) that show associations of this gene with the disease(s)?

14. We will cover the NextGen or Micro-array technologies later in the class. But, we briefly talked about where the NextGen or Microarray data is stored in NCBI? Please provide the database names and visit the databases to explore what information is available for your protein or gene. Please provide a one/two-line summary on the search results for your gene/protein in these databases.

15. Are there any SNPs known for your protein? What database(s) do you go to get this information?

16. What organism(s) do you think you will avoid when you search for novel gene?
    a. We haven't learned about how to do the search but at this point you can make an educated guess on what organisms (databases) you will avoid?

(copy this to your ongoing assignment I and submit as one document)