**Homework Assignment 4**

*Due date and time: 11 AM, Wednesday, February 3rd, 2016*

*You may work with other members of your group on homework assignments but what you write-up and turn in must be solely your individual intellectual contribution.*

*Please submit homework assignments in hard copy and electronic formats. Python scripts must be uploaded as \*.py files, so they could be tested. Please also submit your python scripts in hard copy format.*

#### You have just discovered that the mouse (*Mus musculus*) protein CRTC2 that you have been investigating is a coactivator of CREB. You now wish to learn about the patterns of sequence conservation within this protein family for follow-up studies.

#### **TASK 1**: Perform a BLAST search with the goal of assembling the sequences of CRTC2 paralogs in mouse and putative orthologs in *Gallus gallus*, *Canis lupus familiaris*, *Alligator sinensis*, *Danio rerio*, *Bos taurus*, and *Apis mellifera*. Describe your search strategy including the BLAST algorithm, database, and any other parameter that you changed from default values to find homologues in the aforementioned species. How many paralogs (*ignoring any products from alternatively-spliced isoforms*) of CRTC2 are there in mouse? Is there a CRTC2 homologue/ortholog in (a) *Drosophila melanogaster*, (b) *Caenorhabditis elegans*, (c) *Saccharomyces cerevisiae*,and (d) *Nematostella vectensis*? Comment on what portion of the mouse CRTC2 sequence is conserved in the various species where a homologue was discovered. Download the sequences of CRTC2 homologues in FASTA format from as many species above as possible and compile them into a single file and upload along with your assignment. (**6 points**)

**TASK 2**: In this task, you will need to write a series of scripts. Make sure your scripts are interactive and intuitive for end users:

1. Write a function that reads a FASTA file and returns the corresponding protein or nucleotide sequence. Use the file contig.txt included with the assignment to test whether your script functions as expected. (**2 points**)
2. Write a function that would read a file with the data for the coordinates of various exons in the transcript; use the file exon\_coords\_1.txt to develop/test your script. Now combine this script along with the one in T2i to produce a mature mRNA transcript (for the purposes of this task produce a cDNA which is the same as mRNA except all Us are Ts) for the nucleotide sequence in contig.txt based on the exon coordinates in the file exon\_coords\_1.txt. Output the cDNA sequence to a file in FASTA format but be sure to produce a file with 70 characters of sequence per line. (**4 points**)
3. Now use the [ExPaSy Translate](http://web.expasy.org/translate/) tool to translate the cDNA sequence from T2ii above to generate the corresponding protein sequence. The longest polypeptide sequence (among all possible sequences generated in the six reading frames) is the sequence of interest. Be sure to set output format to “Compact”. The sequence should be familiar; what is it? (**2 points**)
4. Now use the script in T2ii to generate the cDNA for a variant which has the exon coordinates exon\_coords\_2.txt. Compare the encoded sequence with the one in T2iii using BLAST (for 2 protein sequences). What is different between the two sequences? (**2 points**)
5. Write a function that would take a protein sequence from a file in FASTA format and output the sequence composition into another file. The output should be formatted so it looks something like:

Ala (A) 24 7.4%

where the first number is the number of instances of that amino acid in the sequence and the second is the % content. To produce neatly-formatted (and therefore, easily readable) output, use the following notation in the write function of the last two columns:

… + “%3d" % number + “ %4.1f” % floating + …

where number and floating are variables that are of type int and float, respectively. Use the protein sequences you used above (but be sure to create the FASTA files first!) to test your script and compare your output to the [ExPaSy ProtParam](http://web.expasy.org/protparam/) tool. (**4 points**)