**Homework Assignment 5**

*Due date and time: 11 AM, Wednesday, February 10th, 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.*

**TASK 1**: Compile orthologous CRTC2 sequences from the following species: *Mus musculus*, *Rattus norvegicus*, *Homo sapiens*, *Gallus gallus*, *Canis lupus familiaris*, *Bos taurus*, *Danio rerio*, *Xenopus (Silurana) tropicalis*, and *Apis mellifera.* Use the CRTC2 sequence with Uniprot accession Q53ET0. Assume that the sequence with the highest score from each species in your BLAST searches corresponds to the CRTC2 ortholog.

1. Now perform a multiple sequence alignment of the respective full-length proteins using (a) [CLUSTAL Omega](http://www.ebi.ac.uk/Tools/msa/clustalo/) (it is a newer version of CLUSTAL W) and (b) [MUSCLE](http://www.ebi.ac.uk/Tools/msa/muscle/) (use default parameters for both). Compare and contrast the results produced by the two programs. Which MSA is better in your opinion and why? Relative to the human sequence, how are the mouse, rat, and cow sequences contributing to the delineation of structurally/functionally important regions of the protein. Compare and contrast the CRTC2 MSA with that of hemoglobin that we discussed in class (Topic 6: Multiple Sequence Alignments). In what way(s) do the conservation patterns for the two proteins differ? (**4 points**)
2. Now use the [MEME](http://meme.sdsc.edu/meme/cgi-bin/meme.cgi) web server to look for motifs in the aforementioned CRTC2 orthologs. Use default parameters but set the maximum number of motifs to find to 8 and under the *Advanced options* section change the maximum width to 75. Compare and contrast the results you obtained from MUSCLE/CLUSTAL  and MEME. Do you think MEME provides any new insight over and above what MUSCLE or CLUSTAL  did? If so, how; if not, why not? In what way does the alignment provided by MEME differ from those returned by MUSCLE and CLUSTAL . (**4 points**)

**TASK 2**: In this task, you will need to write a series of python scripts or regular expressions as indicated.

1. Construct regular expressions to test for the presence of the following patterns in an input sequence (see Chapter 7 of the text or lecture slides in Python Topic 4). **There is no need to provide the entire script**, although you should write it to test if your expression works as desired:
2. Presence of CCG or CCC
3. Presence of GGX (where X is any nucleotide)
4. One or more occurrences of G or C following TT
5. See the last slide in Topic 5 (BLAST) for the syntax for Prosite patterns and construct an expression for the patterns: A-x-[ST](2)-x(0,1)-V> and <[AC]-x-V-x(4)-{ED}. (**4 points**)
6. Your studies have shown that CRTC2 shuttles between the nucleus and the cytoplasm upon receipt of appropriate stimuli. Write a python script to test the existence of a nuclear export signal (NES) in the human CRTC2 protein above. Capture the salient features of two distinct NES sequence motifs to construct regular expressions based on this [figure](http://nar.oxfordjournals.org/content/early/2012/10/22/nar.gks936/F2.large.jpg) from a paper (**note that prolines are not allowed between conserved residues within the motif**). Your script should return the location(s) (i.e. start and end points) of the NES as well as the actual sequence corresponding to the motif(s). (**2 points**)
7. Taking inspiration from the function described in page 202 of the python textbook, write a script that would translate DNA in all six reading frames (just like the [ExPaSy Translate](http://web.expasy.org/translate/) tool) and output the results to a file. (**4 points**)
8. Using the same dictionary for the genetic code as in T2iii above, write a python script to calculate the number of codons that code for each of the 20 amino acids. Output the results to standard output. (**2 points**)