**NAME:_____**

Bioinformatics CS300 /Bio300
Date: 31 Oct 2017

Copy the files orf_ii.py and the included unknown.fasta files into another directory, outside of the class's shared repository.

1) Run the python orf_ii.py code using the unknown.fasta. What do you see?

2) Use your favorite editor to edit the orf_ii.py file. Locate the fifth line where the variable "threshold" has been set to 10. Change this variable to 0 and then run the program using the same fasta file as before. What do you notice this parameter does?

3) Try changing the threshold parameter to other values to determine it maximum value before no results are returned for the fasta file. When you have determined this max value, reset the threshold value to 5 and rerun your results.

4) Go to NCBI's ORF finder at https://www.ncbi.nlm.nih.gov/orffinder/

5) Copy and paste the code from unknown.fasta file. Determine what the open reading frames are.

6) As you wait for your results to appear, name Compare and contrast the facility of using script based tools as opposed to using web-based tools.

7) How many open reading frames did you find using NCBI?

8) Check and compare the results of the python tool and the online tool. Do any of the outputs match?

9) Why are there any differences between the tool and the web tool?

10) Find at least three open reading frames in which both tools agree. List these open reading frames using the labels from NCBI. List a few others that you have found if available.

11) Can you find the exact names of these proteins? Use Blast to determine what each of these proteins does in terms of function. Name two or three organisms in which this same protein is present.

12) For each of your proteins, determine their functions. Pay special attention to the E-values when you decide upon the functions of these proteins from your open reading frames.

Note: The unknown sequence is the following:

ATGTTCACTACCAAGGTAAATATGTACCCAGAGGTGCCCAGCTCATCCCAGGTGTCAGACG
ACATAGACAATGACACGCACATCGACGAGGTCGCTGCATTTGTGAGAAAGTGGTCGGCAGC
CGGACTATCTCCCCCCATCACCCTTGCGAAGAACCTCAGAGCATGGATATCAAGCAACACC
AGCCCTGGAAGCCCCCTAGTGTTGGATGACAGAATGCTGAGCCTTACAACCATGATATGGA
ACACAGCAGCAGAGCACTACACAATGATAGGCAAATCCCAGGTCAATCGTATGTCATCACT
CATAGATCAGCTGGGGGAGATTTCCGGCCGCAAACCGCCGCAGGGCCCAGCATTCGACATG
CCTCCTCCCCCTCCTAAGAGAAAACATCCGGATTCACTAGACACTAATCCAATATTAGGCTT
AATAGGTCAAGATTGGGACGACAATAAAGACAAGCACTGGAGAGAGAAACCAGCAGACA
AGAAGCTCCTCGTGCTCAACTGGGTGTTGCATGAGTATCTGGGGGTCCTCACAAAACCTGT
CACCATCAAGTGGATAACGGATAACCCCGCGTCTTTAGAGTTGGGAGCAGTGTCAGCTTAT
GCCCTGAAACATCAGGCCAGCTTATCCGACTGCGACAAGGAAGCCCTCAGAGCGTTGGTG
GTTCAAACAGTGAAAAACACCCCCAAAAGGCCATGCCTGGACTAG