Assignment #1 due Wed, Jan 16

Write code to identify all open reading frames (ORFs) in an inputted sequence string. Remember to consider both the "forward" direction and the reverse complement. In order to eliminate very short ORFs, use the random shuffle function from class to simulate 1,000 random sequences (with the same length and GC content as the actual sequence), compute the longest ORF for each simulation, and the longest ORF for all of the simulations. Only consider those ORFs in the actual sequence longer than this threshold. Your code should output (use the print function) the value of this threshold, and for each ORF longer than this threshold,

1. The start and end positions
2. Whether it is in the forward direction or the reverse complement
3. The DNA sequence

Turn in your code (printed on paper, we will use electronic submissions in later weeks, I recommend writing multiple functions).

I have posted two FASTA files on Blackboard (X73525.fasta and salDNA.fa). Also **turn in the outputs from your code** for these two files.

Use (protein) BLAST to find the best matches for at least one ORF from each FASTA file https://blast.ncbi.nlm.nih.gov/Blast.cgi **Write a few sentences** explaining what you found.

GeneMark is a more complicated gene prediction program that we will discuss next week http://exon.gatech.edu/GeneMark/ Use this program to study the two FASTA files (the data are from Salmonella so use the bacteria option GeneMark.hmm with Heuristic models). **Write a few sentences** comparing the results of running GeneMark to our simpler ORF strategy.