BCH 571 Bioinformatics for Life Scientists

Lab 4

This lab will use a set of yeast open reading frames (ORFs). Scer_100genes.nt.txt contains 100 open reading frames (out of about 6,000 total in the yeast genome). Each ORF is on its own line. The gene name appears first on the line, followed by a ":", and then the ORF (from start to stop codons).

Program 1- (call this lab4.1.py)

Program 1 is designed to give experience with dictionaries. Use a dictionary to count the number of occurrences of codon pairs across the set of 100 ORFs. Codon pairs are overlapping 2-codon sequences, that do not include the start or stop. The ORF ATG CGT CGG ACT TAA contains the codon pairs CGTCGG and CGGACT.

Count the number of each codon pair in the ORFs and print the dictionary to the screen. (Do not print 0 for codon pairs that do not occur.)

Program 2- (call this lab 4.2.py)

Program 2 is designed to give experience with the Python regular expression library. CGACGA is a codon pair that is read slowly by ribosomes. In each ORF, find all instances of CGACGA and determine if the CGACGA is in-frame or out-of-frame. For this purpose, it is helpful to know that % is the modulus operator. X%Y returns the "remainder" of X divided by Y: e.g. 3%3 = 0, 4%3=1, 5%3=2, and 6%3=0.

Print the output to the screen. Each gene should have one line of output, with the gene name, the position of CGACGA, and whether the instance is in-frame or out-of-frame.