

Bioinformatics: Life Sciences on Your Computer

Week 4 Peer Assignment

Due July 8, 2014

1. At NCBI, find the nucleotide sequence with accession number KJ668270. The gag, pol and env genes should be present in some sense, but there is overlap between the genes. Notice that the CDS locations overlap. The DNA positions corresponding to the cleaved proteins are labeled "mat_peptide."
 - a. At which numerical position can you find a start codon that is used in two of the CDS regions.
 - b. At which numerical position can you find a stop codon that is used in two of the CDS locations?
 - c. Which is the only CDS to contain the entire protease region?
 - d. Which CDS does NOT use AUG as its start codon?
 - e. Which codon is used by that CDS that uses the alternate start codon and what amino acid does that codon usually code for?
 - f. If you look at the note for the 639..5843 CDS, you will notice that the stop codon at 2250-2252 can be suppressed some of the time, and another amino acid is substituted at that position. Which amino acid is substituted?
2. Please refer to the nucleotide sequence at NCBI with accession number M74164.
 - a. How many CDS locations are found on this operon?
 - b. What is the exact numerical location (beginning and end) of the 5' UTR of the mRNA?
 - c. The -35 signal is shown at 303..308 and the first CDS begins at 382. Why is the -35 signal more than 35 bases upstream of the start codon?
 - d. Click the link for the 1991 PubMed reference from *Biochemistry*. The gene for the *E. coli* fadB multifunctional enzyme and the gene for what rat enzyme are believed to be homologous?
3. Use the DNA sequence at <https://d396qusza40orc.cloudfront.net/bioinform/assignments/predict.txt>. Run Glimmer at NCBI (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi) selecting Linear for Topology. Then run FGENESB at Softberry (<http://www.softberry.com/berry.phtml?topic=fgenesb&group=programs&subgroup=gfindb>) and choose BACTERIAL Generic as the closest organism.
 - a. How many CDS regions are predicted by each prediction program?
 - b. How different total mRNA molecules are needed to code for all of the CDS regions predicted by FGENESB?
 - c. The programs differ at the last CDS location. Is the difference in location at the start codon or at the stop codon?
 - d. How many CDS regions are on the minus strand operon predicted by FGENESB?
 - e. Use the Gene 4 protein sequence predicted by FGENESB. Run a Blast search of the Reference Proteins database. What is the likely enzyme name of this predicted protein?