Homework 6: ORF detection in python

Source: « Introduction to Computational Genomics, a case studies approach », Nello Cristianini and Matthew W. Hahn, Cambridge University Press,2007

Open Reading frames: A (protein-coding) gene consists of streches of (non-stop) codons that begins with a start codon and ends in a stop codon. However, only prokaryotic genes consist of single, continuous open-reading-frames (eukaryotic genes are interrupted by transcribed, but not translated, sequences called introns - the translated portions of the genes are referred to as exons). Because prokaryotic genes do not contain introns, gene finding can be simplified into a search for ORFs.

For this homework, we ask you to write a small program in python that finds ORFs in prokaryotic sequences. This program will have, at least, the sequence (in FASTA format) and a minimum length (in number of nucleotides) as input and will return the different ORFs as well as their position and their frame.

Algorithm: Given a DNA sequence ***s***, a positive integer *k*, for each reading frame decompose the sequence into triplets, and find all stretches of triplets starting with a start-codon and ending with a stop codon. Repeat also for the reverse complement of the sequence. Output all ORFs longer than the prefixed threshold *k*.

Execute your program on the mtDNA from the human (NC_001807) and the mouse (NC_005089). Note that the genetic code from mitochondria is slightly different from the standart one. In particular the one for those vertebrates has different start and stop codons, resulting in different ORFs: the codon TGA means stop in the universal code, but code for tryptophan in mtDNA; AGA and AGG code for arginine in the universal code and the stop codon in mtDNA; and ATA represents isolencine in the universal code and methionine mtDNA.

- What fraction of the sequence represents (candidate) protein coding genes ?
- Try various choices of length threshold. How does this affect the number of ORFs you find ?