Biomedical Software Engineering

Mount Sinai School of Medicine
Graduate School of Biomedical Sciences
Prof. Arthur Goldberg
Spring II, 2024

Comments on Assignment 2: Write a Python program that uses the genetic code

HW 2 is trickier than it looks, because the specified requirements contain many details. It's tricky to scan the mini-chromosome and find all proteins, and to treat ATG as a special case because it encodes for both START and methionine.

Also, details like provide units (amino acids, or nucleotides) on the summary data, and don't report mean, minimum length, and max length aggregate stats for a mini-chromosome that contains no proteins require that one read the assignment carefully after you think the program is done.

Many different approaches can solve HW 2. E.g., one could scan a mini-chromosome and obtain all data about proteins, failures, etc. Or, one could scan the mini-chromosome, determine the boundaries of proteins, and then rescan it using those boundaries. One common incorrect approach was to divide the sequence into amino acids based on the open reading frame and then scan for a start codon. This does not take into account the fact that the length of untranslated regions before the start codon might not be divisible by three.

The software engineering techniques we're studying, especially OO programming and unit testing, make it easier to design and implement tricky code like this.

I have read your code and annotated it in comments or strings with my initials with feedback. Please read them. I provide feedback on the specifics of HW 2 and general coding advice, which I hope you find helpful.

The "Please include the variables' units too." was asking for units for "the length in amino acids of the shortest, and longest protein; the mean protein length in amino acids; ... and the total amount of non-coding DNA in nucleotides". It can be really helpful to track units in scientific software.

A note: HW 2 says "Coding DNA is any DNA that encodes for a protein, plus the START and STOP codons that enclose it." but I believe that this is not standard genetics practice, and the START and STOP codons are not considered coding DNA.