GENOME-SCALE ALGORITHM DESIGN
by Veli Mäkinen, Djamal Belazzougui, Fabio Cunial and Alexandru I. Tomescu
Cambridge University Press, 2015
www.genome-scale.info

Exercises for Chapter 1. Molecular biology and high-throughput sequencing

1.1 Write a program that lists all the DNA sequences that encode a given protein sequence.

1.2 In a given organism some codons are used more frequently than others to encode the same amino acid. Given the observed frequency of every codon in a species, normalize it into probabilities and write a program that, given a protein sequence, samples a random DNA sequence that encodes that protein under such codon usage probabilities.

1.3 In a given organism, some *codons pairs* occur less frequently than others.

(a) Given the set of all exons of an organism, write a program that computes the ratio $z(XY)$ between the observed and the *expected* number of occurrences of every pair of consecutive codons XY. Note that the expected number of occurrences of pair XY depends on the frequency with which X and Y are used to encode their corresponding amino acids.

(b) Given a DNA sequence S that encodes a protein P, let $f(S)$ be the average of $z(XY)$ over all consecutive pairs of codons XY in S. Write a program that computes, if it exists, a permutation S' of the codons of S, such that S' still encodes P, but $f(S') < f(S)$. Such an *artificially attenuated* version of S has been shown to decrease the rate of protein translation: for more details, see [1].

References

[1] Coleman, J. R., Papamichail, D., Skiena, S., Futcher, B., Wimmer, E. and Mueller, S. (2008), 'Virus attenuation by genome-scale changes in codon pair bias', *Science* **320**(5884), 1784–1787.