Marburg University

Department of Computer Science
Prof. Dr. Sebastian Wild

Date: 2025-10-31
Version: 2025-10-31 08:19

Exercise Sheet 3 for Algorithms of Bioinformatics (Winter 2025/26)

Hand In: *Until 2025-11-07 18:00, on ILIAS.*

Problem 1

10 + 10 + 10 + 10 + 30 points

Here we consider a revised null model for random open reading frames.

a) Consider the hypothetical DNA sequence

`AUGAAACCCAUAAGGGGG...`

What would our stop codon automaton from class do on this example?
Could it represent it a valid (tiny) gene?

b) Consider now instead a model that generates one of the 64 possible base triples (codons) in each step uniformly at random. What is the expected length (in bases) of a random open reading frame? Compare this (again) to the result from class.

c) For the model from b), compute the variance and standard deviation of the length of a random open reading frame.

d) Use *Chebyshev's inequality* to bound the probability of seeing an open reading frame of more than 400 bases in a random RNA in both models.

Alternatively: Use computer algebra to compute this probability exactly.

e) Suppose a hypothetical alien life form uses the stop codons `UAA`, `UAG`, and `UUU` instead of the Earthly `UAA`, `UAG`, and `UGA`.

How does this affect the expected lengths of random open reading frames in both null models? Adapt the calculations for the expected lengths accordingly.