ECE 598OM: Homework 2 - Due end of March 2023

Issued by: Olgica Milenkovic

Problem 1 (Computational): You will be provided access to new, in-house PacBion kinetics data for native and methylated DNA bases. Your task will be to come up with a ”good” compression algorithm for the same.

Problem 2 (Analytical): Read and report on the analysis of the longest common substring problem of two random strings as described in Arratia+Wterman, for the case of general distributions (not necessarily uniform) and Markov chain models.

Problem 3 (Computational): Implement an algorithm of your own choice for constructing suffix trees (one example is Ukkonen's algorithm, you can find good descriptions of the methods in Dan Gusfield's book *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*.

a) Test the algorithms on two real DNA strings retrieved from the NCBI repository;
b) Test the algorithm on randomly constructed strings described in Arratia+Tavare; try to verify computationally the results of the longest common substring analysis.

Problem 4 (Analytical, open ended): Try to perform the DNA synthesis scheduling analysis from Makarychev et al. for balanced collection of strings. Here, assume that you have sequences of length $16m$ and that each block of length 16 (non-overlapping) has a balanced GC content.