CSCI 4192 Assignment 2

posted: 04.11.2020

Instructor: Travis Gagie

due: midnight 15.11.2020

Please work together for this assignment! The 7zipped file `https://www.dropbox.com/s/q99sdmq9qcuteu0/4192_assignment_2_data.7z?dl=0` contains a 1000000-character text file `ref_genome` and 10 text files `reads_0`, ..., `reads_9` with roughly 20000000 characters in each, divided into 100-characters strings with one string per line.

FM-index `ref_genome` and align the strings in each file `read_X` against it like a readset — considering that there are both differences between `ref_genome` and the "genome" they came from, and sequencing errors, so you may need the techniques you've learned to reduce approximate matching to exact matching. Return the 10 resulting consensus strings (i.e., your best guess as to the 10 "genomes" the readsets came from). To keep things simple, the genomes differ only by substitutions and the sequencing errors are also all substitutions.

Name your 10 guesses at the genomes `assembly_0`, ..., `assembly_9`, 7zip them together and submit them by email.