CS548
Assignment 2 (Due March 26, 2020)
Suffix Array, BWT, Read Alignment

Computer Science Department
Colorado State University

March 3, 2020

1. Implementing suffix array and BWT [30 pts].
Write a program to compute the suffix array (start locations of alphabetically sorted suffixes) of an input DNA sequence. Your algorithm does not need to be time- or memory-optimal; the input sequence is assumed to be no longer than 200,000 bp. From the suffix array, compute the Burrows-Wheeler transform (BWT) of the input sequence, and output it in a text file. Try your algorithm on the Acinetobacter phage 133 genome from `http://www.ncbi.nlm.nih.gov/nuccore/326536151?report=fasta` (click "Send" in the top right corner and select file in "Choose Destination").

2. Implementing FM index [30 pts].
Suppose the Burrows-Wheeler transform of a sequence is given as input. In fact, you can use your code above to obtain the BWT. Write a program to compute the Ferragina-Manzini (FM) index from the input BWT. Use your FM index to reconstruct the inverse BWT. Test your program on the Acinetobacter phage 133 genome.

3. Hands-on experience [40 pts].
For this assignment, use "bwa" on the department Linux machines. Download *E. coli* K12 reference genome from `http://www.ncbi.nlm.nih.gov/nuccore/556503834?report=fasta` (click "Send" in the top right corner and select file in "Choose Destination") and use "bwa index" to index it. Download Illumina sequencing reads from the course website and map them to the *E. coli* K12 reference genome using "bwa aln" and then "bwa samse". Finally, use IGV (`https://www.broadinstitute.org/igv/`) to visualize coverage (depth of sequencing) for the entire genome and two specific windows: [1,030,000 - 1,030,500] and [3,700,000 - 3,700,500]. Include those three coverage plots in your submission.

Submit your source codes and a report including the results of each task in a zip file.