Name: _____

CSE 4502/5717 Big Data Analytics
Exam II; April 11, 2019

Note: You are supposed to give proofs to the time and processor bounds of your algorithms. Read the questions carefully before attempting to solve them.

1. (25 points) Input is a sequence X with n elements that is residing in D disks. The problem is to sort X. It is known that each element in X is an integer in the range $[1, B]$, where B is the block size. It is known that $M = \Theta(B^2 D)$, M being the main memory size. Show how to sort X in two (read) passes through the data.

2. (25 points) The input for this problem is a string X of length n. The goal is to find the longest repeated substring, i.e., a substring of maximum length that appears in at least two different positions in X. Present an $O(n)$ time algorithm to solve this problem.

3. (25 points) Input are a text T of length m and a pattern P of length n. T and P are strings from an alphabet Σ, with $\sigma = |\Sigma|$. The problem is to find all the occurrences of P in T within a Hamming distance of 1. If s_1 and s_2 are strings of the same length, then the Hamming distance between them is defined to be the number of places in which they differ. For example, the Hamming distance between *aacgtt* and *agcgat* is 2. Present an algorithm for this problem that runs in $O(\sigma n^2 + m)$ time.

4. (25 points) In this problem we are given a text T, a pattern P, and the suffix array S for T. The problem is to identify all the occurrences of P in T. Let $|T| = m$ and $|P| = n$. Present an algorithm to solve this problem in $O(1)$ time using $n\sqrt{m}$ CRCW PRAM processors. Specifically, the output should be an array $A[1 : m]$ such that $A[i] = 1$ if $P = T_i$; (If $T = t_1 t_2 \cdots t_m$ then $T_i = t_i t_{i+1} \cdots t_{i+n-1}$); Also, $A[i] = 0$ if $P \neq T_i$, for $1 \leq i \leq m$.