CAP 6515 HOMEWORK ASSIGNMENT 2
DUE ON 10-25-2022

Note: Any solution to an algorithm design question MUST contain the following four sections:

(1) **Problem statement.** A clear unambiguous statement of the problem to be solved, which includes the input, the output, and the object function with the constraints.
(2) **Algorithm description.** A clear, unambiguous description of the algorithm.
(3) **Correctness proof.** A convincing mathematical argument that the algorithm described solves the computational problem described.
(4) **Time analysis.** A time analysis of the algorithm, up to order, in terms of all relevant parameters.

You may use any algorithms and data structures from class.

1. Ukkonen's algorithm

(I) Formalize the pseudocode for the Ukkonen's algorithm for constructing the suffix tree of a given string in linear time. (II) Draw the implicit suffix tree and show the list of rules used for each phase $(i+1)$ and each extension (j) to construct the suffix tree for string "xabxababxba" by using the Ukkonen's algorithm. (50%)

2. Suffix tree for large alphabet

When introducing the Ukkonen's algorithm for suffix tree constructing, we assume a constant size of the alphabet. If we assume the alphabet size $|a|$ is comparable to the length of the input string n, there is a trivial low bound $O(n \log n)$ for applying Ukkonen's algorithm. Describe a simple algorithm to achieve this lower bound. (25%)

3. Peptide vaccine design

The activation of helper T-cells is essential to initiate a protective immune response. To mimic pathogen invasion, biologists synthesize peptide vaccines, i.e. small peptides of the essential proteins from a pathogen (bacterium or virus) that can be recognized by the major histocompatibility complex (MHC) and presented to the helper T-cells. A simple version of the *peptide vaccine design problem* can be formulated as the *shortest unique substring problem*, which attempts to find the shortest peptide in the proteins of the pathogen (called pathogen proteins) that are not a part of any protein from the host (human) (called host proteins). (25%)