CMSC 423 – Midterm 1 – September 27, 2018

Name _____

Honor Pledge
The University of Maryland Code of Academic Integrity requests that you write by hand and sign the following statement pledging your commitment to academic integrity. Please do so in the blank space below the text of the honor pledge.

I pledge on my honor that I have not given or received any unauthorized assistance on this examination.

Signature _____

To help with grading, please write your answers within the box provided on the exam. If you need to use additional pages, please clearly write your name at the top of each page.

IMPORTANT: The in-class exam only counts for 90% of your exam grade. To earn the last 10%, you need to work together with your teammates and return a fully correct exam.

Please budget your time carefully! <u>You only have 75 minutes available for this exam.</u> There is a very good chance you will not be able to answer all the questions in the allotted time (though it is definitely possible to do that).

1. The basics (25 points). Please be brief in your answers.
    a. (5 points) Why is Laplace's rule necessary in motif finding?

    b. (5 points) What is the longest open reading frame (ORF) in this string. Assume there is only one possible start codon (ATG) and one possible stop codon (TAA). For full credit indicate the number of amino-acids encoded by this ORF.

ATTACGATCGATAAAT ATGAGGGTAATGCATTAATA

    c. (5 points) True or false: clumps of frequent k-mers always indicate the location of the origin of replication in bacterial genomes? Please explain your answer for full credit.

d. (5 points) Name the biological process that creates protein from RNA molecules.

e. (5 points) Assume you are given a string of length n. What is the worst-case runtime (in big-O notation) of a naïve algorithm for computing Z values?

2. Patterns in DNA (10 points)

a. (5 points) The book proposes an encoding that converts each k-mer into an integer. What integer value would be assigned to the 4-mer: TACC

b. (5 points) What is the G/C skew at each location in the following string? For full credit indicate the formula you are using:

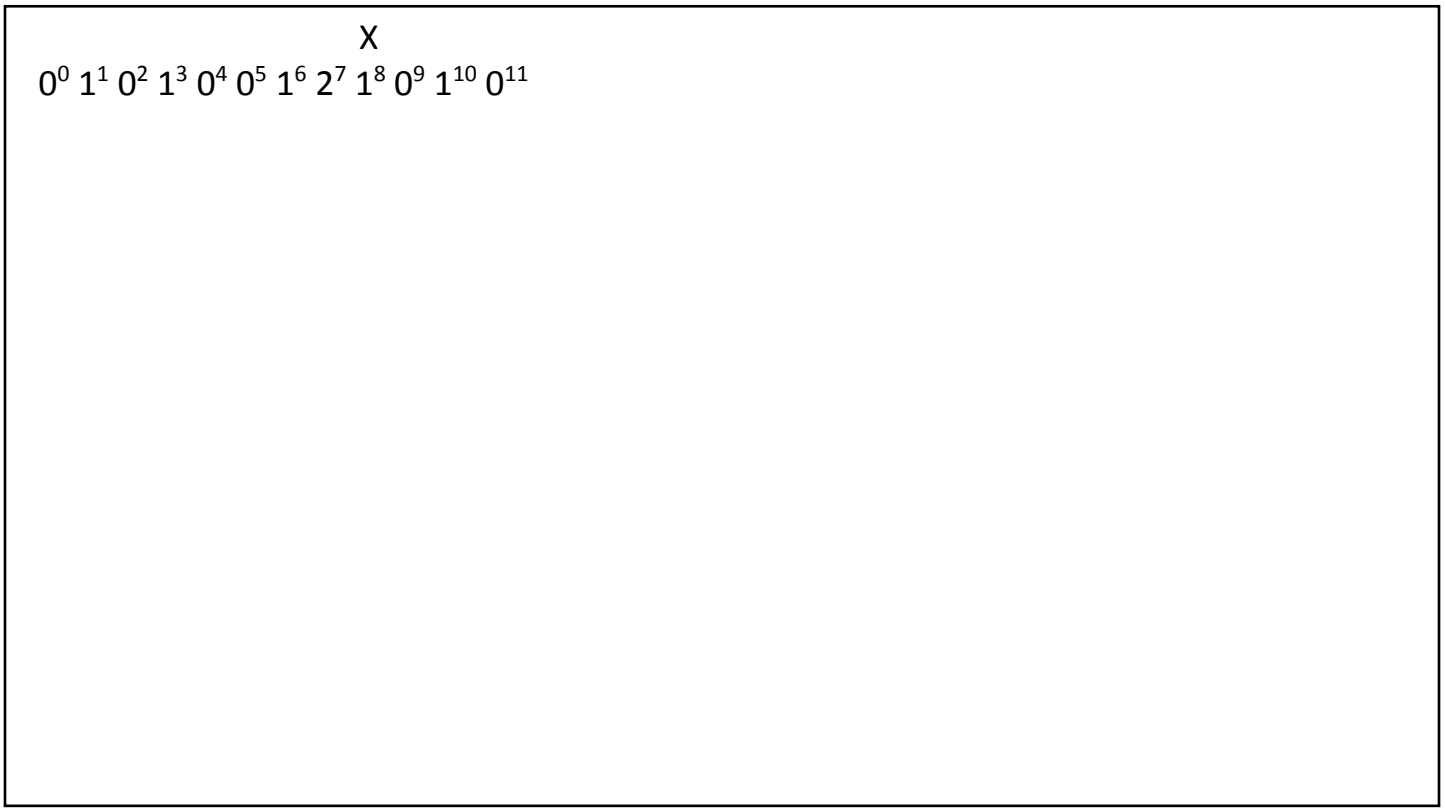
G A C A G A T T A G T C A T T A G A T C T C C T C A C G

3. KMP & Z algorithm (25 points)
    a. (10 points) In the following text I have filled the Z values up to position 14:

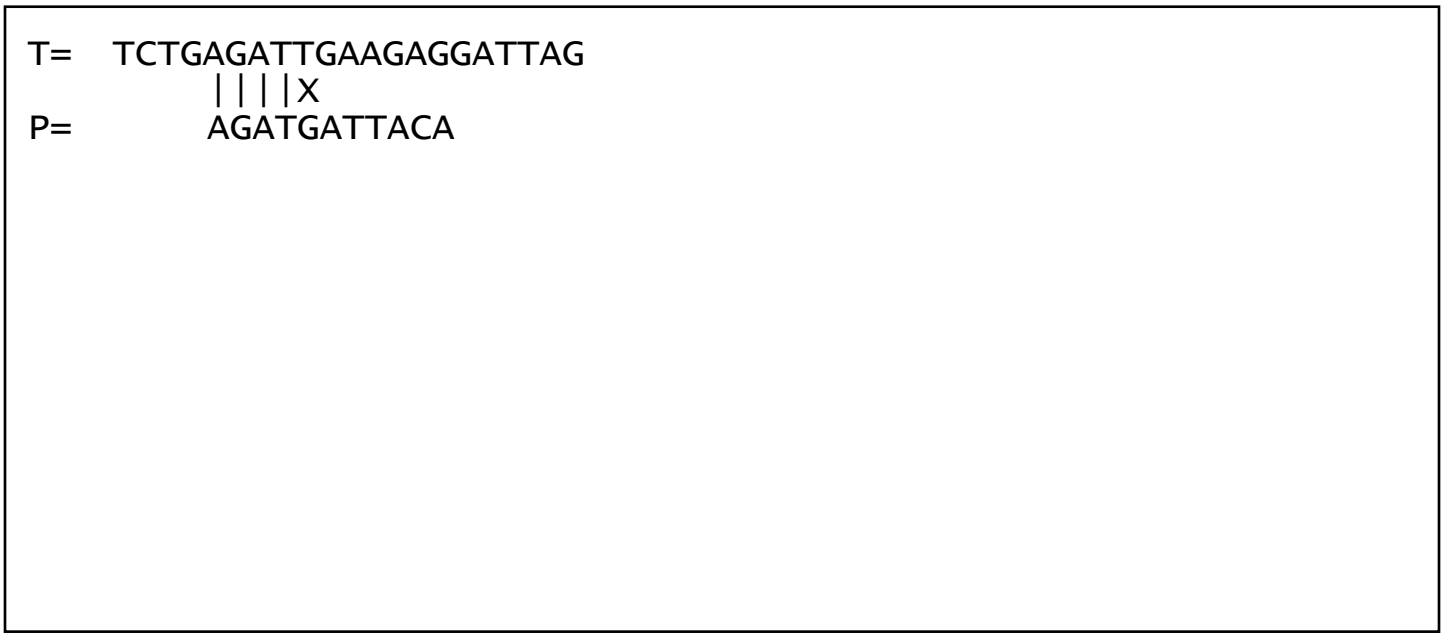
$$A^0C^1C^2A^3G^4A^5C^6C^7A^8T^9A^{10}G^{11}A^{12}C^{13}C^{14}A^{15}C^{16}C^{17}T^{18}$$
$$Z = 0\ 0\ 0\ 1\ 0\ 4\ 0\ 0\ 1\ 0\ 1\ 0\ 4\ 0\ 0$$

Please describe the logic used by the efficient Z algorithm computation to determine the value of Z[15].
What is this value?
How many text characters did you have to match?

b. (10 points) Below you are given just the sp(i) values created by the KMP algorithm for a pattern whose sequence you don't know. Assume the KMP algorithm has found a mismatch between the pattern and the text at position 8 (the position is marked in superscript). How far would the algorithm shift the pattern after this mismatch? You can simply draw the pattern at its new location if you prefer.



c. (5 points) Given the pattern and the text shown below, with vertical bars indicating that KMP algorithm found a match. How far would KMP shift the pattern after the mismatch indicated with an X?



4. Motif Search (20 points)

TTACCTT**AAC**
GATG**TCT**GTC
C**CGC**CGTCGT
CA**CTA**ACGAG
CGTCAGA**GGT**

Assume, we are using Gibbs Sampler to find a 3-mer from the above set of DNA fragments. At the i-th iteration, the candidate-motifs from each fragment are underlined.

a. (5 points) The Gibbs sampler selected the last sequence to seed the next iteration. What is the probability matrix of the profile (after Laplace's rule) and what is the consensus sequence?

b. (10 points) Given the profile matrix shown below, what is the profile most probable 3-mer in the last sequence of the motif shown above?

A	0.6	0.1	0.4
C	0.2	0.3	0.1
G	0.1	0.4	0.3
T	0.1	0.2	0.2

c. (5 points) What is the entropy of the following column in a motif profile matrix?

A 2/16
C 4/16
G 2/16
T 8/16

5. Think outside the box problem (10 points).

Assume you implemented the KMP algorithm and are now applying it to different types of texts. Do you expect it to be faster or slower when applied to DNA or to English text? Briefly explain why. While the runtime is O(n) irrespective of the alphabet used, the actual number of comparisons made by the algorithm depends on the actual string it is applied to.