Spring 2017 **CMSC 423: Final** H. Corrada Bravo
Time: 2 Hours

WAIT FOR INSTRUCTIONS BEFORE BEGINNING

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

Signature and UID: _____

Print name: _____

- ***Write your answers with enough detail about your approach and concepts used, so that the grader will be able to understand it easily.***
- ***The sum of the grades is 105, but your grades would be out of 100 (thus you get 5 bonus points by solving all the problems).***
- ***Select the best choice for the first 8 problems and mark it by X in the table below.***

Problem	1	2	3	4	5	6	7	8
A								
B								
C								
D								
E								

DO NOT WRITE BELOW THIS LINE

Problems 1-8:	/30	Problem 12:	/12
Problem 9:	/8	Problem 13:	/15
Problem 10:	/8	Problem 14:	/20
Problem 11:	/12	**Total:**	

Multiple-choice Problems (Answer THE BEST CHOICE in the Table of the First Page and NOT HERE):

Problem 1. **(2 points**) What is an *open reading frame* (ORF)?

a) any translatable sequence of nucleotides
b) any sequence of codons
c) a long enough sequence of aminoacids
d) a long enough sequence of codons without an intervening stop codon
e) None of the above

Problem 2. **(3 points)** Which of the following statements are true of relative entropy and entropy. You should consider profiles with a single column, and non-uniform background nucleotide distributions.

a) The relative entropy of a profile is always equal to its entropy.
b) The relative entropy of a profile is never equal to its entropy.
c) In motif finding, we search for profiles that **maximize** relative entropy.
d) In motif finding, we search for profiles that **minimize** entropy.
e) All of the above, except (a).

Problem 3. **(6 points)** Consider the following set *S* of reads `ATG`, `GGG`, `GGT`, `GTA`, `GTG`, `TAT`, `TGG`. Which of the following are true of the DeBruijn graph for S. Make sure to write down the DeBruijn graph below and show all your work.

a) Nodes are labeled by 3-mers

b) Nodes are labeled by 2-mers

c) There are more than 1 Eulerian **path** in the graph

d) An Eulerian **path** in the graph corresponds to string TATGTGGGTA

e) Both (b) and (c)

Problem 4. **(2 points**) Which of the following best describes the distinction between exonic and intronic regions in DNA:

a) exonic regions are (usually) transcribed into RNA

b) intronic regions are (usually) not transcribed into RNA

c) different transcripts (isoforms) for a gene are given, in part, by differences in exon splicing

d) all of (a), (b) and (c)

e) none of (a), (b), and (c)

Problem 5. **(3 points)** Which of the following statements are true of the Burrows-Wheeler transform BWT(S) of string S

a) there is an algorithm using BWT(S) with time complexity O(|P|) to determine if string P occurs in S

b) the ith occurrence of character c in BWT(S) corresponds to the ith occurrence of c in S

c) BWT(S) is sufficient to reconstruct the first column of the rotation matrix of S

d) (a) and (c)

e) None of the above

Problem 6. (6 points) Assuming Z-values $Z_2, \ldots, Z_8$ are already computed, how many character comparisons are required to compute Z_9 for string S=ACCACTACCAG by the linear time Z-algorithm discussed in class? (Note that the table of Z-values is 1-indexed, Z_1 is the Z-value of the first character in S.

a) 0 b) 3 c) 1 d) 9 e) 4

Problem 7. (4 points) Which of the following statements best represents complexity of using suffix arrays for exact string matching of query string P to target string T

a) preprocessing time: O(|P|), space: O(|P|), search time: O(|T|)

b) preprocessing time: O(|T|), space: O(|T|), search time: O(|P| * $\log_2$ |T|)

c) preprocessing time: O(|T| * log |T|), space: O(|T|), search time: O(|P| * $\log_2$ |T|)

d) preprocessing time: O(|T| + |P|), space: O(|T|), search time: O(|P| * |T|)

e) None of the above

Problem 8. (4 points) In which of the following situations would you use the KMP algorithm instead of a suffix tree?

a) There is no situation in which KMP is preferred

b) Matching many short patterns to a single long target

c) Matching a single pattern to multiple distinct targets

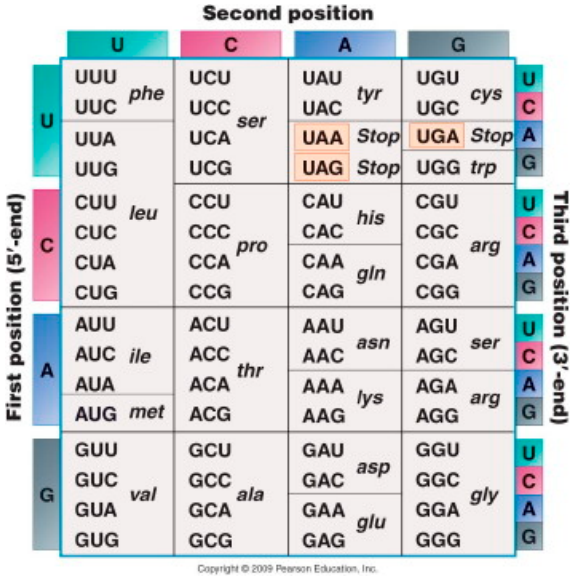
d) (b) and (c)

e) None of the above

Short Questions (show all derivations as appropriate for full credit):

Problem 9. (8 points) Identify the longest open reading frame in the following DNA sequence and translate it into an aminoacid sequence (start codon is ATG):

```
TGCGTATGTATGTCAGACGGTGAGACGCTTGCGGGCTAAGCGACG
```



Problem 10. (8 points) The BWT of a string is "ipssm$pissii" using LF mapping find the number of times pattern "iss" appeared in the original string (show your work, including the LF mapping table, and how top and bottom pointers are updated in each iteration of the search algorithm).

Problem 11 (12 points) We saw in class that the worst-case space complexity of suffix *tries* is $O(n^2)$, where *n* is the size of the string. Discuss the two methods used to turn a suffix trie into a suffix tree so that only linear space is used in the worst case. Sketch a proof that only linear space is required.

Problem 12. **(12 points)** Consider pattern P=AGTCGA and target T= AGCAGTCGAGTC. Show how the Z-algorithm would be used to find *all* occurrences of P in T in linear time. Calculate all Z-scores required. List those positions in T that need any explicit character comparisons to compute their Z-scores.

Long Questions (you should always PROVE THE CORRECTNESS of your solutions)

Problem 13 (20 points) Consider the motif finding problem. In this question you will extend the algorithms you learned about to search for *maximum relative entropy* profiles.

(a) Argue why we should be looking for *maximum* relative entropy profiles, instead of *minimum* relative entropy profiles. Two or three sentences are sufficient.

(b) In the randomized motif search algorithms, we used the concept of *most probable* k-mer in a DNA string to determine next states to explore in the search. Given a profile *Profile,* and a DNA string, how is the *most probable* k-mer in a DNA string defined?

(c) Given a *Profile* and a set of *background frequencies,* define a score that you would use in these search algorithms in order to find *maximum relative entropy* profiles. You should write a mathematical expression that computes this score given: *Profile* (defined by nucleotide probabilities p_{rj} for nucleotide *r* and position *j,* background probabilities b_r and *k*-mer Pattern). Write a sentence or two describing this scoring function, make sure to mention if this score is a probability (or not).

(d) In the Gibbs sampler a move was made by randomly choosing a k-mer within a DNA string with probability depending on the *Profile-probability* of each k-mer given the current *Profile.* How would you modify the Gibbs sampler to use your new score? Write out the pseudo-code for your modified Gibbs sampler.

Problem 14 (15 points) Consider the suffix tree of some arbitrary string S. Recall that leaf *i* corresponds to the suffix of S starting at position *i.* The least common ancestor of two nodes in a tree is the lowest node shared by the paths from the two nodes to the root. Assume *u* is the least common ancestor of leaves *i* and *j* in the suffix tree of string S.

a) What does this node represent? Give an example.

b) Describe an algorithm that will compute the Z values for S using the suffix tree.