Fall 2014 **CMSC 423: MidTerm2** H. Corrada Bravo
Time: 1 Hour, 15 Minutes

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 10 problems and mark it by X in the table below.***

Problem	1	2	3	4	5
A			X		
B					
C	X		X		
D					X
E		X		X	

DO NOT WRITE BELOW THIS LINE

Problems 1-5:	/30	Problem 9:	/15
Problem 6:	/10	Problem 10:	/25
Problem 7:	/10		
Problem 8:	/15	**Total:**	/100

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

Problem 1. **(3 points**) Given a directed graph G=(V,E), the Hamiltonian Path problem is:

a) Find a path that visits all edges in E exactly once
b) Find the path that visits the most vertices in V, while visiting every edge in E exactly once
c) Find a path that visits all vertices in V exactly once
d) Find the shortest path between a pair of specific nodes u and v in V
e) None of the above

Problem 2. **(6 points**) Which of these statements are accurate for genome assembly

a) The Hamiltonian approach is problematic due to the complexity of the Hamiltonian path problem

b) The time complexity of constructing a read overlap graph is the same as the time complexity of constructing a DeBruijn graph

c) The Eulerian approach is problematic due to the large number of Eulerian paths in a DeBruijn graph

d) All of the above

e) Only (a) and (c)

Problem 3. (8 points) Which of these is the time complexity of the farthest-first traversal algorithm for the k-center clustering problem of n m-dimensional points. (Remember to chose the **best** answer, provide an explanation)

a) $O(n^2m + k^2n)$

b) $O(kn)$

c) $O(k^2nm)$

d) $O(n)$

e) $O(km^2 + n)$

Problem 4. (5 points) Which of these are reasons to use inexact string matching methods to compare biological sequences:

a) Exact matching misses string overlaps required for genome assembly assuming sequencing errors
b) Exact matching would not sensitively identify protein sequences from different species with potentially the same molecular function
c) Genomic variants in sequences from an individual may not match any position of a reference genome when using exact matching
d) Only a) and c)
e) All of a), b) and c)

Problem 5. (8 points) Consider the incorrect recurrence relations for global alignment with affine gap penalties below. How many mistakes are there? (Assume the cost of a gap of length g is open + (g-1) * extend). Circle and explain/correct the mistakes.

$$M(i,j) = max \begin{cases} M(i-1,j-1) + SCORE(x_i, y_j) \\ X(i,j-1) \\ Y(i-1,j) \end{cases}$$

$$X(i,j) = max \begin{cases} open + M(i,j-1) \\ extend + X(i,j-1) \\ extend + Y(i,j-1) \end{cases}$$

$$Y(i,j) = max \begin{cases} open + M(i-1,j) \\ extend + X(i-1,j) \\ extend + Y(i-1,j) \end{cases}$$

a) None
b) One
c) Two
d) Four
e)Nine

No ‘extend + Y’ on the X recurrence relation, no ‘extend + X’ on the Y recurrence relation, X(i,j) and Y(i,j) for gap closing on the M recurrence relation. So, four

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

Problem 6. (10 points) You are a clinical genomicist and have sequenced a patient's genome. To find possible disease causing mutations you are going to compare the millions of reads generated by the sequencing instrument to a reference human genome: you want to find the placement of sequence queries (reads) of **length 100** along the human genome (3 Gbp), allowing for at most **4 mis-matches** in a 100 bp read.

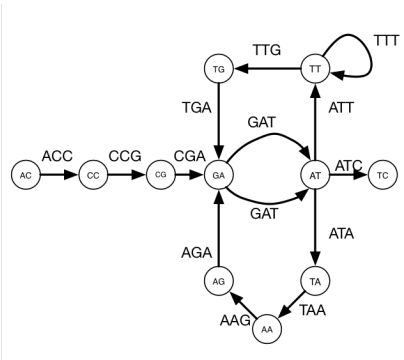
However, as you might have guessed, even using a dynamic programming solution to the fitting alignment problem is not efficient. Instead you will use a much more efficient exact matching algorithm to find candidate matching positions in the genome and then use the Smith-Waterman algorithm only in these candidate positions in the genome.

The proposed algorithm is to divide each **100 bp** read into non-overlapping k-mers and match each k-mer exactly to the reference genome. What is the maximum k-mer length possible that guarantees that no correct inexact matches (as defined above) are missed? Explain.

Divide the read into 5 equal sized k-mers, at least one of them matches exactly, so largest k =20

Problem 7. (10 points) Given the DeBruijn graph below, solve the contig generation problem. List each resulting path, and it's corresponding sequence. For example, this is one of the returned contigs:

AT->TC, ATC



AC->CC->CG->GA, ACCGA

GA->AT, GAT

GA->AT, GAT

AT->TA->AA->AG->GA, ATAAGA

AT->TT, ATT

TT->TT, TTT

TT->TG->GA, TTGA

Problem 8. (15 points) Given the pairwise distance matrix below for 10 genes, use farthest-first traversal to find 4 cluster centers, starting with gene A. Resolve all ties between genes by their lexicographical order. List the four genes selected as cluster centers and the resulting cluster assignments. Show your work for every iteration of the algorithm.

	A	B	C	D	E	F	G	H	I
B	3								
C	3	4							
D	3	6	4						
E	1	2	4	5					
F	5	3	5	9	5				
G	5	5	2	6	5	5			
H	2	5	3	2	4	7	5		
I	2	2	5	5	1	5	5	4	
J	3	1	3	6	2	3	4	5	3

It 1:

dist to nearest center B:3, C:3, D:3, E:1, F:5, G:5, H:2, I:2, J:3 / choose F as next center

It 2:

dist to nearest center B:3, C:3, D:3, E:1, G:5, H:2, I:2, J:3 / choose G as next center

It 3:

Dists B:3, C:2, D:3, E:1, H:2, I:2, J:3 / choose B as next center

Centers: A,F,G,B

Assignment: A:A, B:B, C:G, D:A, E:A, F:F, G:G, H:A, I:A, J:B

Problem 5. (15 points) (a) Define the concept of “coverage” as used in genome assembly. (b) The Lander-Waterman statistic provides a mathematical model of the relationship between “coverage” and the number of contiguous pieces (contigs/islands) of sequence that can be assembled from a given genome. Describe roughly the relationship between the two (a sketch illustrating the function given by the Lander-Waterman statistic is sufficient). Mention the rate (linearly, quadratically, exponentially, etc.) at which the expected number of contigs changes as coverage increases.

The average number of times a position of the genome is sequenced. It decreases exponentially with coverage (after it’s high enough).

Long Question

Problem 10. (25 points) We've seen in class two algorithms that use probability estimates as part of an optimization problem: (a) in the Gibbs sampling algorithm for motif finding, we used the 'profile probability' of a k-mer to sample positions in DNA sequences containing a protein binding site, and (b) in the EM algorithm used in fuzzy k-means, we used 'assignment probability' to calculate cluster centers using weighted averages. Design an EM algorithm to solve the motif finding problem.

1. In fuzzy k-means, the parameters of interest were the k centers. What is the estimate of interest in motif finding?
2. In fuzzy k-means, *HiddenMatrix* was a matrix with a row for each point (e.g., expression from one gene across multiple timepoints) and a column for each center. What does $HiddenMatrix_{ij}$ correspond to in fuzzy k-means?
3. What should the dimensions of *HiddenMatrix* be for your motif finding EM algorithm? What does $HiddenMatrix_{ij}$ correspond to in this case?
4. How did you compute $HiddenMatrix_{ij}$ in fuzzy k-means? Write the mathematical expression. How would you compute $HiddenMatrix_{ij}$ for motif finding? Write a mathematical expression. Note: this is the *E-step* in your algorithm.
5. How was *HiddenMatrix* used to calculate centers in fuzzy k-means? Write the mathematical expression to calculate the *jth* cluster center. Note: this is the *M-step* in fuzzy k-means.
6. Given *HiddenMatrix,* how would you use it to calculate a motif profile? Write a mathematical expression for entry p_{cl} corresponding to nucleotide *c* and position *l* of the profile. Note: this is the *M-step* of your algorithm.

1. the profile, 2. The probability point i belongs to cluster j, 3. Num sequences by num. starting positions, the probability that motif starts in position j of string i,

4. $HiddenMatrix_{ij} = e^{-\beta d(x_i, center_j)} \Big/ \sum_l e^{-\beta d(x_i, center_l)}$

For motif finding we use profile probability of a given k-mer and normalize to make it a probability. Sij is the k-mer at position j in the ith string in dataset, p is current profile:

$$HiddenMatrix_{ij} = ProfileProb(s_{ij}, p) \Big/ \sum_l^{n-k+1} ProfileProb(s_{il}, p)$$

5. $center_j = \sum_i HiddenMatrix_{ij} x_i \Big/ \sum_i HiddenMatrix_{ij}$

6. For motif finding we count the number of times a nucleotide occurs at a given position of the motif, now we do 'weighted counts'

$$p_{cl} = \frac{\sum_{j=1}^{n-k+1} \sum_i HiddenMatrix_{ij} I(s_{i(j+l-1)} = c)}{\sum_{j=1}^{n-k+1} \sum_i HiddenMatrix_{ij}}$$

where $I(s_{i(j+l-1)} = c)$ is 1 if the (j+l-1)'th character of the i-th string in the dataset is c and 0 otherwise.