|  |  |  |
| --- | --- | --- |
| 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** |  |  |  |  |  |
| **B** |  |  |  |  |  |
| **C** |  |  |  |  |  |
| **D** |  |  |  |  |  |
| **E** |  |  |  |  |  |

DO NOT WRITE BELOW THIS LINE

|  |  |  |  |
| --- | --- | --- | --- |
| Problems 1-5: |  | Problem 10: |  |
| Problem 6-8: |  |  |  |
| Problem 9: |  | **Total:** |  |

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

1. **(2 points**) Which of these best represents the relationship between genotype and phenotype?

a) there is **no** relationship between genotype and phenotype

b) an individual’s phenotype **completely** determines their genotype

c) an individual’s genotype **completely** determines their phenotype

d) an individual’s phenotype **partially** determines their genotype

e) none of the above

1. **(3 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

1. **(5 points)** Which of the following statements best represents complexity of using suffix arrays for exact string matching of query string Q to target string T

a) preprocessing: O(|Q|), space: O(|Q|), search: O(|T|)

b) preprocessing: O(|T|), space: O(|T|), search: O(|Q| \* log |T|)

c) preprocessing: O(|T| \* log |T|), space O(|T|), search O(|Q| \* log |T|)

d) preprocessing: O(|T| + |Q|), space O(|T|), search O(|Q| \* |T|)

e) None of the above

**Problem 4. (5 points)** Assume I want to use a Bayesian formulation to compute the *posterior* probability P(M1|x,y) that sequences x (query) and y (target in a large database D) are related by, say, evolution (model M1), rather than chance (model M2). It requires ratio P(M1)/P(M2) is given. How should this ratio be set:

a) proportional to |D|

b) independent of |D|

c) proportional to 1/|D|

d) proportional to |x| \* |y|

e) none of the above

**Problem 5. (5 points)** Consider the 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 \* extend).

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

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

**Problem 6. (10 points)** You are trying to find the placement of sequence queries (reads) of length 100 along the human genome (3 Gbp), allowing for at most 4 mis-matches, however you do not want to use Smith-Waterman directly. Instead you will use an exact matching algorithm to find good candidate matches then extend these with Smith-Waterman. What is the length of the exact match seeds (substrings of the query) you will use in order to guarantee you that no correct alignments are missed? Explain.

**Problem 7. (10 points)** Construct the suffix array for string RATATATRA and trace how binary search is used to find all matches of query TAT.

**Problem 8. (10 points)** The table below shows the optimal pairwise alignments between strings ATCTTAG, ATTAG, AGCTG and ATG. Show how to construct a multiple sequence alignment (MSA) using the STAR approximation. Make sure to state 1) which sequence is the “center”, and why?, 2) the order in which sequences are added to the MSA, and 3) each partial MSA as sequences are added (i.e., you should have 3 MSAs in all). The pairwise alignments are calculated using global alignment with linear gap penalty with the following parameters: match: -2, mismatch: 1, gap: 5

|  |  |  |
| --- | --- | --- |
| ATCTTAG  A--TTAG cost: 0 |  |  |
| ATCTTAG  AGC-T-G cost: 3 | ATTAG  AGCTG cost: -1 |  |
| ATCTTAG  A---T-G cost: 14 | ATTAG  A-T-G cost: 4 | AGCTG  A--TG cost: 4 |

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

**Problem 9.** **(20 points)** In class we discussed two approaches to sequence alignment – global and local alignment. A global alignment requires the two sequences to be aligned end-to-end while a local alignment allows one to ignore any mismatches occurring at the end of the sequences. There is, however, a middle ground - the **semiglobal** alignment. In a semi-global alignment all characters in the two sequences must be aligned, however only gaps internal to the alignment are counted, while gaps at either end of the alignment are "free".   
For example, the following sequences aligned optimally using global alignment:  
  
CAGCACTTGGATTCTCCGG  
CAGC-----G-T-----GG  
  
can be aligned optimally in a semiglobal fashion as follows:  
  
CAGCA-CTTGGATTCTCGG  
---CAGCGTGG--------  
  
Describe a dynamic programming algorithm that computes the semiglobal alignment of two strings in time O(mn). I.e., describe recurrence relations, starting conditions and how to use table(s) to implement the DP algorithm.

**Problem 10. (15 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, 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 for motif finding: what is the estimate of interest? (the parallel in fuzzy k-means are the k centers), what does the ‘HiddenMatrix’ correspond to in this case? How do you use ‘profile probability in this case’?