|  |  |  |
| --- | --- | --- |
| Fall 2013 | **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** |  |  |  | **X** |  |  |  |  |
| **B** |  |  |  |  | **X** |  | **X** |  |
| **C** |  | **X** |  |  |  |  |  |  |
| **D** |  |  | **X** |  |  | **X** |  | **X** |
| **E** | **X** |  |  |  |  |  |  |  |

DO NOT WRITE BELOW THIS LINE

|  |  |  |  |
| --- | --- | --- | --- |
| Problems 1-8: |  | Problem 13: |  |
| Problem 9-11: |  | Problem 14: |  |
| Problem 12: |  | **Total:** |  |

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

1. **(2 points**) Which of the following are examples of sequence mutations?

a) Single Nucleotide Polymorphism (SNP) b) insertion c) inversion

d) deletion e) all of the above

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

1. **(6 points)** Consider the Sequence Profile (Position Specific Scoring Matrix, or PSSM) for some DNA motif shown below. In which position of string ATTCAGGA is the highest probability substring of length 3 found? (Note: assume 1-indexing so the first character is in position 1. Show your work.)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Pos 1 | Pos 2 | Pos 3 |
| A | .8 | .6 | .4 |
| C | .2 | .3 | .5 |
| G | 0 | .3 | .1 |
| T | 0 | 0 | 0 |

a) 1 b) 2 c) 4 d) 5 e) 6

1. **(2 points**) Which of the following resources *does not* contain high-throughput sequencing data from population experiments:

a) KEGG Pathway Database b) 1000 genomes project

c) Short Read Archive d) (a) and (b) e) all of the above

1. **(3 points)** Why is *overfitting* the training data a problem in the classification setting?

a) classifiers that overfit the training data will have high error rate on the training data

b) classifiers that overfit the training data will perform poorly on new data

c) *overfitting* is not a problem in classification

d) (a) and (c) e) None of the above

1. **(6 points**) I have learned a nearest centroids classifier to obtain the two centroids given below (Table 1). What will be the output of the prediction procedure for this classifier for the three samples given below (Table 2)? Show your work (a drawing is sufficient).

Table 1. Centroids

|  |  |  |
| --- | --- | --- |
|  | Normal | Severe |
| Gene 1 | 1 | -1 |
| Gene 2 | 0 | 0 |
| Gene 3 | 1 | -1 |

Table . Samples

|  |  |  |  |
| --- | --- | --- | --- |
|  | Samp. A | Samp. B | Samp. C |
| Gene 1 | 2 | -1 | 2 |
| Gene 2 | 0 | 0 | 0 |
| Gene 3 | 2 | 0 | -1 |

a) A: normal, B: normal, C: severe

b) A: severe, B: severe, C:normal

c) A: normal, B: severe, C: severe

d) A: normal, B: severe, C:normal

e) A: severe, B: normal, C:severe

1. **(4 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

1. **(4 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, and that we are minimizing the cost of the alignment).

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

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

1. **(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

****

TGCGTATGTATG TCA GAC GGT GAG ACG CTT GCG GGC TAAGCGACG

Met Ser Asp Gly Glu Thr Leu Ala Gly

1. **(8 points)** Assume you are given a query string of 60 bp and you wish to find whether it matches somewhere in the human genome with at most 2 mismatches. A simple heuristic (to avoid using the Smith-Waterman algorithm for inexact matching) is to use an exact matching algorithm to find candidate places in the genome where your query string might match. What length of query substring exact matches would you search for so that you can guarantee you will not miss any matches of the full query within the genome? Briefly explain why you will not miss any matches.

A string of 60bp matches with at most 2 mismatches if and only if a substring of 20bp matches exactly. Therefore I would look for exact matches for substrings of 20bp (by dividing the query string into three equal sized substrings for example).

1. **(12 points)** (a) Please define the concept of “coverage” as used in genome assembly. (b) The Lander-Waterman statistics allows to study the correlation between “coverage” and the number of contiguous pieces (contigs/islands) of sequence that can be assembled from a given genome. Describe roughly what is the relationship between the two. Be sure to mention at what rate (linearly, quadratically, exponentially, etc.) does the number of expected number of contigs decrease as coverage increases.

Coverage is the oversampling of the genome in sequencing (how many full copies of the genome are sequenced). The number of contigs decreases exponentially as coverage increases.

1. **(12 points) Prefix-Suffix matching.** Give an algorithm that takes two strings and , and finds the longest suffix of that matches a prefix of . The algorithm should run in time O(|| + ||).

Construct string and compute Z-values for S. Identify positions *i* in such that which corresponds to suffixes of that matches a prefix of . Find the smallest *i* satisfying this condition.

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

1. **(15 points)** In the Hamiltonian path approach to genome assembly, we constructed a graph where sequences are nodes and edges are drawn between them if we find that they *overlap.* Describe how you would modify the inexact matching dynamic programming algorithm to compute overlaps between pairs of sequences provided as input to a genome assembler. Please specifically address the following points:
2. How do you define an overlap (drawing is sufficient)?
3. What are the initial conditions in the DP table?
4. In which cell of the DP table will the score of the optimal alignment be found?
5. What recurrence relations will you use: global alignment or local alignment (adding 0 to “start over” an alignment)? No need to actually write out the relations.

Note: assume all overlaps are proper, i.e., none of the sequences is contained within another one. Also, assume we only care about overlaps between “forward” versions of the sequences (no need for reverse complements). You can assume linear (not affine) gap penalties.

First assume that the overlap is between a suffix of x and a prefix of y, so (a) the overlap alignment starts with a series of gaps in y and ends with a series of gaps in x. Note: We can run the algorithm twice (x-y and y-x) and choose the best score.

(b) the first row is initialized to 0 since alignments that start with gaps in y don’t cost anything.

(c) the answer will be the maximum value in the last column since that is be optimal alignment ending with a gap in x.

(d) use the global alignment recurrences since all characters in each string have to be aligned.

1. **(20 points)** **Clustering and Classification.** We discussed in class, and you implemented in project IV, a resampling method to calculate the *prediction error* of the nearest centroid classifier on a given dataset. Recall that the procedure was: (1) randomly divide the samples in the dataset into two groups, (2) train the nearest centroid classifier on the first group, (4) test on the second group by predicting using the trained classifier and compute error rate, and (5) repeat a number of times (50) and compute average error rate over the 50 replications. In this question, you will propose a similar method to analyze the performance of the k-means clustering algorithm.

**(1).** Suppose I have two datasets D1 and D2 and have used the k-means algorithm on dataset D1 to obtain k cluster centroids. Now, I ask you to use the k-means algorithm on D2 to obtain another set of k cluster centroids. *Propose a method to measure how “similar” the two sets of cluster centroids are.* Your answer should include the definition of a measure (a single number) that summarizes the similarity of the two sets of cluster centroids. Feel free to use pictures to motivate your solution

**(2).** Describe how to use your method from part (1) to define a resampling method that measures the performance of the k-means algorithm (for a fixed value of k) on a given dataset.

**(3).** Describe how you would use your method from part (2) to choose k for a given dataset. Be sure to describe your expectation of how your method will behave as k=2,3,…

1. Here are two related methods: a) For each centroid in the D2 clustering, find the nearest centroid in the D1 clustering and compute the squared Euclidean Distance between them. Add these distances and use that as a statistic of similarity. Similar clusterings will have small distances. b) Find the nearest centroid from the D1 clustering for each sample in D2. Now, for each pair of samples in D2 that are in the same cluster (according to the D2 clustering) find the proportion of times they are assigned to different clusters in the D1 clustering. Use this percentage as a measurement of similarity (similar to error rate). Similar clusterings will have small error rate.
2. Randomly split the given dataset into two groups calculate the statistic from (1) and report the average.
3. For each value of k=2,… use procedure in (2) and choose the value that minimizes it. For small k it will be high, it will decrease as k increases until k is too large, when it will increase again.