Midterm 2015

Reminder: This is a take-home exam. Students are expected to develop, write up, and hand in their own individual solutions and, in doing so, develop a sufficient understanding of the problem and solution so as to be able to explain it adequately to the instructor. Under **no circumstances** should a student discuss the exam with others while taking it, copy or consult the solution of another student, or copy a solution from any other source, including the Internet. You must show all work.

1. Provide the pseudo-code for an algorithm that will determine the length of the shortest subsequence that occurs exactly once in a genomic sequence of length N. You must provide all assumptions made for your algorithm.

2. As reported in the HG38 version of the human genome, please answer the following questions about the gene, CD8B.

a. What chromosome is this gene on?

b. What is the start and end coordinate of this gene?

c. What strand of the DNA contains this gene?

d. How many transcripts are in this gene as reported by the RefSeq Consortium?

e. How many distinct exons are reported in the RefSeq transcripts?

f. How many distinct introns are reported in the RefSeq transcripts?

g. List the genomic coordinates of each intron (start-end) from above which contains a canonical splice sites.

h. List the genomic coordinates of each intron (start-end) from above which contains a non-canonical splice site.

i. What is the official gene symbol of the nearest gene to CD8B? What strand of the DNA contains this nearest neighboring gene? How would the location of these genes be described in relation to one another?

3. Describe the minimum information needed to unambiguously define the location of a gene? Provide an example using a gene of your choice.

4A. Write a program which will randomly (each run of the program will have distinct outputs) generate two sequences which will result in a better local alignment as compared to a global alignment (using blosum62 scoring matrix), both 10 amino acids longs. Discuss your criteria and any limitations in generating your sequences. Align the 2 sequences using dynamic programming using the blosum62 scoring matrix. Show the DP matrix and the resultant alignment. Discuss the alignment.

4B. Write a program which will randomly (each run of the program will have distinct outputs) generate two sequences which will result in a better global alignment as compared to a local alignment (using blosum62 scoring matrix), both 10 amino acids longs. Discuss your criteria and any limitations in generating your sequences. Align the 2 sequences using dynamic programming using the blosum62 scoring matrix. Show the DP matrix and the resultant alignment. Discuss the alignment.

5. Describe a scenario where a researcher would be interested in investigating both the local and global alignment of two sequences.

6. Bacterial genomes are often circular. To transform to a linear form, some genome assembly programs will pick a random location in the genome to break the circle. Thus, it is possible that running the same program multiple times we would get different answers, corresponding to different circular rotations of the same string. Provide the psuedo-code that will determine if two DNA strings are circular rotations of each other. For example TTGATC is a circular rotation of ATCTTG. You must state all assumptions.

7. We can define a set of distinct substrings of a string S that includes all substrings. However, each repeat is only represented once. For example, for the string S = AATATT, this set is:

{A, T, AA, AT, TA, TT, AAT, ATA, TAT, ATT, AATA, ATAT, TATT, AATAT, ATATT, AATATT}

You are given a suffix tree of S. Provide the pseudo-code for an algorithm that counts the number of distinct substrings of S. For full credit, this should run in O(n) time.