Homework 1

The goal of this homework is to examine the topics related to pairwise and multiple sequence alignments.

Problem 1. Dynamic Programming Practice (10)

The longest increasing subsequence problem for a set of numbers is to find a subsequence of a given set of numbers, in which elements are in sorted order, from lowest to highest. The subsequence of numbers is not necessarily contiguous. For example, in sequence (9, 5, 8, 7, 15), the longest increasing subsequence is (5, 8, 15). Provide a dynamic algorithm to find the longest subsequence for a sequence of integers. Assume that all integers are unique.

Problem 2. Substitution Matrices (10)

Assume that we have the following table about the frequencies of DNA base pair occurrences based on a collection of aligned sequences. Calculate the joint and marginal distributions for each pair of bases and individual bases. Use these probability values to derive a score matrix of base pair substitutions. Round each number to the closest integer less than or equal to the score value (floor).

	A	T	G	C
A	20	14	7	10
T	14	7	12	9
G	7	12	10	7
C	10	9	7	5

Problem 3. Global and Local Sequence Alignment. (20)

(3a). Convert the last four digits of your student ID (学号) and birth date to the base-four (i.e., quaternary) numbers. Then convert these base-four numbers to DNA letters, using the mappings: A->0, T->1, G->2 and C->3. For example, suppose your student ID is 2008011238, and your birth date is 19890721 (i.e., July 21, 1989). Then for your student ID, the base-four sequence of "1238" is 103112, and the corresponding DNA sequence is TACTTG. For your birth date, the base-four sequence of "0721" is 23101, and the corresponding DNA sequence is GCTAT. You can use some web tools (e.g., http://www.mathsisfun.com/numbers/convert-base.php or http://www.kaagaard.dk/service/convert.htm) to convert decimal numbers to quaternary numbers.

Based on the above two sequences generated from your student ID and birth date, compute their global alignment. Assume a match score of +1, a gap penalty of -3 and a substitution score of -1. Show your work by filling in the dynamic programming table that stores the alignment scores and trace the path corresponding to the alignment. If there are multiple possible alignments, show all of them along with their traceback paths.

(3b) Use the same scheme to convert the last four digits of your student ID and the last **Three** digits of your birth into DNA sequences. Write the recurrence for a local alignment. Using the same scores above to perform a local alignment for these two sequences.

Problem 4. Multiple Sequence Alignment (20)

(5a). Ask another four DNA sequences from your classmates who are taking the same course (using the last digits of student ID, as in 3a). Combine them with your own sequence, you will obtain five sequences in total. Use the Star algorithm to identify a multiple sequence alignment for these five sequences. To find a center, pick the string that has the maximal average alignment score. To compute distance between two sequences use the score in 3a.

(5b). Write pseudo code that uses the guide tree to identify a multiple sequence alignment for n sequences. Use your code to identify a multiple sequence alignment for the above sequences. Show your work by drawing the tree and the intermediate sequences you get.

Problem 5. BLAST Exercise (5 points)

Use the BLAST program to search over a nucleotide database and find the DNA sequence that have the greatest similarity with the sequence that you generate in 3a (i.e., using the last four digits of your student ID). The BLAST program is available here: http://blast.ncbi.nlm.nih.gov/. Write down the name of this sequence found from the database and the letter correspondence in the alignment.

Problem 6. A critique of the Compressive Genomics Paper. (Extra credit). (10)

Read the following Nature Biotechnology paper from Berger's group and its Supplementary Information.

Compressive Genomics. Po-Ru Loh, Michael Baym, and Bonnie Berger. Nature Biotechnology, 2012.

Write a one-page critical review of the core scientific problem addressed in the paper. Describe one major methodological innovation, and one major biological innovation of this method. If there are any potential limitations, name them and propose ideas to address these limitations.