B363: Bioinformatics algorithms

HW4 (Due: **Nov. 11 Friday 5pm**)

http://darwin.informatics.indiana.edu/col/courses/B363-16

(You do not need to write computer programs for the following questions.)

1. (10 pts) Does an optimal multiple alignment always induce optimal pairwise alignments? If yes, explain your answer; otherwise, give a counterexample.

2. (15 pts) Construct the alignment graph of ACTACA and ACAA (using the score = 2 for matches, = -1 for mismatches, -7 for gap opening penalty and -1 for gap extension penalty). Show the optimal alignment, and its corresponding path in the alignment graph.

3. (10 pts) The divide-and-conquer algorithm presents a simple approach to parallelize the pairwise alignment of very long DNA sequences (e.g., genomic sequence of millions of nucleotides), because 1) the FromSource(i) and ToSink(i) algorithms can be carried out independently on two separate CPUs without exchanging intermediate results; and 2) the alignment of the prefix and suffix strings broken at the midNode can be carried out separately on separate CPUs. However, the two subproblems after the divide-and-conquer may take much different time to complete (note that the areas under these two subproblems may not be the same if the midNode locates far below n/2 in rows even though it always locates on m/2 in columns). Modify the algorithm of finding the midNode so that the run time of the two subproblems after the divide-and-conquer are approximately equal, assuming the two input sequences have about the same length (m ~ n). (Hint: you can try to find a midNode not on the column of m/2, but on a main diagonal of the alignment grid).

4. (20 pts) Given the permutation of (+1, -3, -4, -2, +6, +7, -5, +8), 1) using the greedy approach (p304) to find a sequence of reversals to sort it into a identify permutation (+1, +2, +3, +4, +5, +6, +7, +8); 2) What are the breakpoints in this permutation? 3) Build the breakpoint graph from the permutation; 4) How many red-blue alternating cycles in the breakpoint graph? 5) what is the minimum number of reversals to sort the permutation to the identity permutation?

5. (10 pts). Dr. Smart claims that if the fragile regions are randomly distributed in the genome, then the distance between to consecutive fragile regions should follow an exponential distribution. Is he right? Explain your answer.

6. (10 pts) Consider the translocation from two linear chromosome (1, -2, 3, 4) and (-5, -6, -7, 8) into two linear chromosome (1, -2, -7, 8) and (-5, -6, 3, 4). Illustrate it by using a 2-break in circular chromosomes. (Hint: see Figure 6.13 as an example).

7. (15 pts) Devise an algorithm based on k-mer sorting to find all shared k-mers between two given genomic sequences of length m and n, respectively. Sketch your algorithm in pseudocode. What is the run time of your algorithm in terms of big-O notation?

8. (10 pts) Is the following distance matrix additive? If so, provide the corresponding tree; otherwise, explain why?

	i	***j***	***k***	***l***
i	0	3	4	3
j	3	0	4	5
k	4	4	0	2
l	3	5	2	0