B363: Bioinformatics algorithms

HW2 (Due: **Sep. 23 Wed, 11:59pm**)

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

1. (10 pts) Dr. Smart argues that because the randomized algorithms (such as Gibbs sampling algorithm) can identify the motifs consisting one k-mer in each input sequence, the consensus of these k-mers should be always the median string of the same set of input sequences. Is he correct? If yes, explain why. If not, can you give a counter example (i.e., where the consensus of the identified motifs is NOT the median string of the input sequences)?

2. (10 pts) Mr. Fuzzy devises the following algorithm to find the median string of length *k* for a given set of *t* DNA sequences *Dna*={Dna_1, Dna_2,…, Dna_t}. Is the algorithm correct? Why?

```
Input: Dna, integer k
FuzzyMedianString(Dna, k)
for each string Text in Dna
    d_min ← k × t;
    for each k-mer P in Text
        d ← 0
        for each string Text' in Dna such that Text' ≠ Text
            d_min' ← k
            for each k-mer P' in Text'
                d' ← HammingDiance(P, P')
                if d' < d_min'
                    d_min' ← d'
            d ← d + d_min'
        if d < d_min
            d_min ← d
            median ← P
Output median
```

3. (10 pts) In practice, a motif can occur in either of the two strands of a DNA sequence. Present the revised Gibbs sampling algorithm to find the motifs in a set of given DNA sequences such that the motif can be in the forward or reverse complement of each sequence. Please express your algorithm in pseudocode.

4. (10 pts) The set of genes involved in a biological process (e.g., circadian clock) are sometimes modulated by two complementary transcription factors, i.e. one transcription factor regulates (activates or represses) a subset of genes, and the other factor regulates the remaining genes. To address this problem, we consider the following *twin motifs finding problem*. Given a set *t* DNA sequences *Dna*={Dna_1, Dna_2,…, Dna_t}, two sets of *k*-mers, $motif_1$ and $motif_2$, are called a pair of *twin motifs* in *Dna*, if *Dna* can be partitioned into two subsets Dna_1 and Dna_2, and $motif_1$ and $motif_2$ each contains one k-mer from each sequence in Dna_1

and Dna_2, respectively. The *twin motifs finding problem* is defined as:
Input: a set *t* DNA sequences *Dna*={Dna_1, Dna_2,…, Dna_t} and integer k.
Output: a pair of twin motif $motif_1$ and $motif_2$, minimizing score($motif_1$) + score($motif_2$).
Note: the score of the motif is defined as the sum of the hamming distances between each k-mer and the consensus k-mer (slides 27-30 of Chapter 2). Devise an algorithm to solve the twin motif finding problem.

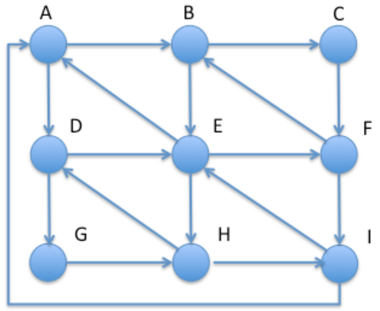
5. (5 pts) Ms. Curious is a biology graduate student. She is interested in a transcriptional factor NobX. She conducted two experiments, and obtained the following two sets of DNA fragments as potential binding sites of NobX, respectively. As the results are so different, she suspected one experiment was contaminated. Can you tell which one is more likely wrong? Why?

```
A:
TTACCTTAAG
GTTCCATAAC
TAACCTTGAC
TTACCTCAAC
TTAGATCAAC

B:
ACACAGGCAC
TGAACCGGAC
GATCCGGCGC
CACGGATCTC
ACACCGGTAC
```

6. (10 pts) Given the following k-mer (k=3) composition: S={ATG, GGG, GGT, GTA, GTG, TAT, TGG}, show the *de Bruijn Graph* of S, and reconstruct all possible sequences *s* whose k-mer compositions are equal to *S.*

7. (5 pts) Consider the graph below. Does it have a Eulerian circle? If yes, find the path. If not, why?



8. (10 pts) In practical DNA sequencing, the sequence fragments (referred to as the *reads*) can be sampled from either of the two strands of the DNA, and it is unknown which strand each read is from. Briefly describe the Eulerian path

approach to assemble the reads into a genome that takes this practical issue into account.

9. (10 pts) Reconstruct the DNA sequence spelled by the following Eulerian path of (2, 1)-mers: (AG|AG) → (GC|GC) → (CA|CT) → (AG|TG) → (GC|GC) → (CT|CT) → (TG|TG) → (GC|GC) → (CT|CA). Note: (2, 1)-mer is a pair 2-mer with fixed distance (i.e., the number of nucleotides between two 2-mers) of 1.

10. (10 pts) To find a Eulerian cycle in a balanced, strongly connected graph, Dr. Smart suggested to start from a vertex with high indegree (and high outdegree). What is his rationale? Assuming you start from the vertex with the highest indegree in the graph, is it always true that the first cycle you get is a Eulerian cycle? Justify your answer.

11. (10 pts) A *bubble* in a *de Bruijn* graph consists of two *parallel* paths connecting from the same vertex A to the same vertex B. 1) Devise an algorithm for detecting bubbles in a given de Bruijn Graph. 2) After a bubble is detected, you must decide which of the two paths in the bubble to remove. How should you make this decision? Explain your answer.