B363: Bioinformatics algorithms

HW2 (Due: **Sep. 23 Friday BEFORE** Lab session)

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

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

1. (10 pts) Show that the *consensus* (string) P of t given motifs (s_1, s_2, …, s_t) of length k is the median string of the motifs, i.e., among all string of length k, P has the minimum sum of Hamming distances with the motifs: $\sum_{i=1}^{t} d(P, s_i)$.
2. (10 pts) Dr. Smart argues that because the randomized algorithms (such as Gibbs sampling algorithm) can identify the motifs including one in each input sequence, the consensus of these motifs 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)?

3. (10 pts) Mr. Fuzzy devises the following algorithm can 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
```

4. (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.

5. (10 pts) Ms. Curious is a biology graduate student. She is interested in a

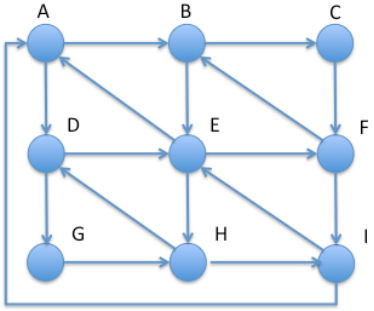
transcriptional factor NobX. She conducted two experiments, and obtained the following two sets of DNA fragments as potential binding sites of NobX, respectively. As the results are so different, she suspected one experiment was contaminated. Can you tell which one is more likely wrong? Why?

```
A:
TTACCTTAAG
GTTCCATAGC
TAAGCTGGAC
TTACATCAAC
GTAGATCAAC

B:
ACACAGGCAC
TGAACCGGAC
GATCCGGCGC
CACGGATCTC
ACACCGGTAC
```

6. (10 pts) Given the following k-mer 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. (10 pts) Consider the graph below. Does it have a Eulerian path? If yes, find the path. If not, why?



8. (10 pts) In practice of DNA sequencing, the sequence fragment (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. Brief describe the Eulerian path approach to assemble reads into a genome that takes into this practical issue into account.

9. (10 pts) Mr. Fuzzy proposes the following greedy algorithm to reconstruct a DNA sequence from short fragment. 1) Start from an arbitrary fragment *S,* and remove it from the pool of fragments; 2) find in the remaining fragments the one *S'* with the longest overlap, i.e., $S=S_1 \odot S_2$ and $S'=S_2 \odot S_3$, where $\odot$ means the concatenation, such that S_2 is the overlap between S and S', that is maximum; 3) replace *S* by $S_1 \odot S_2 \odot S_3$, and remove S' from the pool; 4) iterate to step 2, until no

more fragment in the pool overlaps with S (i.e., $|S_2|=0$); 5) output *S*. Is his algorithm correct? If not, give a counter example.

10. (10 pts) Show the complete theoretical mass spectrum generated from the cyclopeptide NLYV.