B363: Bioinformatics algorithms

HW1 (Due: **Sep. 9 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) Illustrate the output of the frequent word algorithm FREQUENTWORDS(Text, k) on the following DNA sequence for k=3: GAGTTAACGAACGCTTAAC. Are there any 3-mers significantly frequent? Justify your answer.

2. (10 pts) Estimate the running time of COMPUTINGFREQUENCIES (that is based on the hash table of the size 4^k) and characterize the values of |Text| and k when it is indeed faster than FREQUENTWORDS. Does the Replication Origin Finding Problem satisfy this condition?

3. (15 pts) Mr. Fuzzy devises the following algorithm to generate the set of *1*-neighbors of a string *P*. Is the algorithm correct? If not, can you modify it to make it correct?

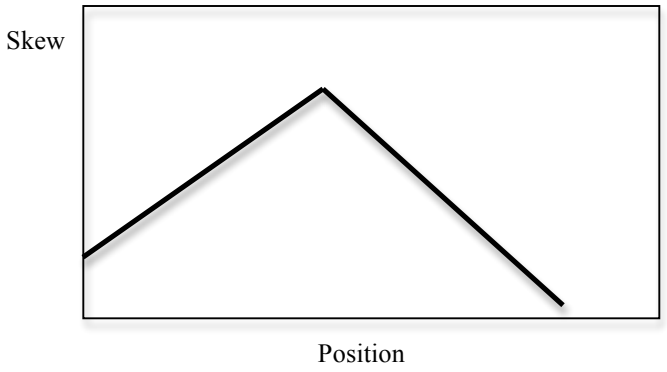
```
Input: A string P of length |P| and an alphabet A of size |A|;
Neighbors(P)
Q←P
Initialize S ← empty set
for j ← 1 to |P|
      for k ← 1 to |A|
              Q[j] ← A[k]
             add Q into S
output S
```

4. (15 pts) Devise an algorithm to generate the set of strings with the exact Hamming distance of *d* from a given string *P*. What is the running time of your algorithm in term of Big-O notation?

5. (20 pts) A k-mer *P* is defined as a *palindrome* is its reverse complement is identical to itself, e.g., ATCCGGAT. Many DNA-binding proteins are *dimers*, formed by two identical protein domains in opposite directions. As a result, the DNA sequences recognized by these proteins are palindromes. Devise an algorithm to output all palindromes of a given length k (e.g., k=8 in the above example) in an input DNA sequence *Text*. What is the running time of your algorithm in term of Big-O notation?

6. (10 pts) When you analyzed a newly sequenced bacterial genome, you obtained the following skew diagram. Can you tell where is the replication origin? Justify

your answer. (Hint: many bacterial genomes are circular.)



7. (20 pts) Given 10 instances of DNA binding sites as following (DNA source: H-NS, Histone like, nucleoid-associated DNA-binding protein),

CGCCTGAATA
CGAGAAAGTT
CGCCGGAATT
GGCATGAATA
TAAAGGAATC
TAATTTAATT
CAATTAAATT
GACATGAATC
TGGCTAATTT
CAACTGAATT

a) Building a profile (motif) matrix;
b) Compute the entropy H for model;
c) Given another sequence S_0, CAAATTATTT, compute $Pr(S_0|\text{profile})$ using the profile you generate.