**Bioinformatics*Algorithms*

HW #2

B. Mishra

February 27 2018 (due in 2 weeks, October 24))

My name is: .....

Q1. [+10] Consider the searching problem:

Input: A Genome $G \in \Sigma^*$ = A string (sequence of characters) over an alphabet of characters $\Sigma = \{a, t, c, g\}$; A reverse-complement restriction pattern k-cutters of length k, $p_k \in \Sigma^*$, such that Reverse(Complement(p_k)) $= p_k$.

Output: An index i such that $p_k = G[i..i + k - 1]$ or the special value `NIL` if v does not appear in G.

Write pseudocode for linear search, which scans through the sequence, looking for p_k. How will you optimize it, if you need to search for multiple p_k's?

Q2. [+10] In the same setting as the one described above: Let $|G| = n$ and $|p_k| = k$, what is the probability of a p_k occurring in G, assuming that G is a random string with all characters occurring equiprobably. Consider the distance between two consecutive occurrences of p_k in G: What is its average value? Variance?