Algorithms and Data Structures for Bioinformatics(BI-GY 7453)

Assignment - 3

Question 1 :

You are given two strings String1 and String2. In a single shift, you can rotate one line (String1) by 1 element such that its 1st element becomes the last and the second one becomes the first like “abcd” will change to “bcda” after the one-shift operation. You have to find the minimum shift operation required to get the common prefix of maximum length from String1 and String2. Do consider edge cases.

Test Case:
Input: String1 and String2 are input to the function String Match which takes two arguments : (String1, String2)

Output: Shift, Prefix

Example:
Input:
String1 = ’Bioinformatics’
String2 = ’This is Bioinformatics’

Output:

Shift = 8
Prefix = Bioinformatics

Question 2.

Align the following nucleotide sequences with the following general rules: Mismatch=0, Match=1, Gap=-1. Remember to pick the highest value calculated based on the two potential gaps (horizontal and vertical) and the match/mismatch (diagonal).
Global alignment: The first and last bases of the two sequences should be aligned. This is like the example done in class. Negative numbers are allowed. The traceback always begins with the very bottom-right corner.

		A	G	C	T	C	A	G
G								
C								
A								
G								
G								

Question 3 :

Write a script that implements the Rabin-Karp algorithm for string matching a Pattern to Text with a DNA

alphabet of A, G, C, T.

** Note in class we discussed how we can treat NT sequences as numbers. In slide 18 of lecture 7 it is mentioned that d is the total number of unique characters in the alphabet, in this case d is 4. This means ACGT can be 0,1,2,3. But what value of q can you use? Try a couple of them to see.

Text = "AGCATGCAGCAT" Pattern = "GCAG"

matches = rabin_karp(Text, Pattern, d, q)

print(matches)
This should output [2, 8], which are the starting indices of the matches in the Text.