Due Date: November 6, 2016 (23:55)

CENG 465
Introduction to Bioinformatics
Fall 2016-2017

Assignment #1

Programming Assignment on Dynamic Programming

Finding occurrences of a pattern P in a target string D using Dynamic Programming

Given a query string **P** and a target string **D**, your goal in this assignment is to write a program to find how many times **P** appears as a *sub-sequence* of **D**. Note that the term *sub-sequence* is not the same as the term *sub-string* and a *sub-sequence* may have other characters of **D** occuring in between the characters of **P**. For example, AT is a sub-sequence of ACGT. CT is also a sub-sequence of ACGT. However, TA is not a subsequence of ACGT. Formally, the problem can be stated as follows:

> Given two strings **P** and **D**, how many different sequences of increasing indices, *ind*, you can find, so that $\mathbf{D}[ind_1] + \mathbf{D}[ind_2] + \mathbf{D}[ind_3] + \ldots + \mathbf{D}[ind_{|\mathbf{P}|}] = \mathbf{P}$. Here, *ind* is an array of increasing integers, + is the character concatenation operation, **D**[*i*] is the i^{th} character of string **D**, and |**P**| is the length of the string **P**.

For example, if **P** is AT and **D** is AGTATCCTGTA, **P** occurs as a subsequence of **D** seven times, where the indices are [1,3], [1,5], [1,8], [1,10], [4,5], [4,8], and [4,10].

A dynamic programming solution for this problem has the following reccurrence equation, where F(*i*,*j*) shows the number of occurrences of the first *i* characters of **P** as a sub-sequence of the first *j* characters of **D**:

$$F(i,0) = 0 \quad 1 \leq i \leq |\mathbf{P}|$$
$$F(0,j) = 1 \quad 0 \leq j \leq |\mathbf{D}|$$

$$F(i,j) = \begin{cases} F(i,j-1) + F(i-1,j-1) & \text{if D[j] is equal to P[i]} \\ F(i,j-1) & \text{if D[j] is not equal to P[i]} \end{cases}$$

Since F can grow very quickly to very large numbers for certain **P** and **D,** in this assignment it is sufficient for you to report the last 5 digits of the count you compute. In other words, I will only be interested in F(|**P**|,|**D**|) % 100000.

You may write your code in any programming language of your choice.

Submission

Submit your source code only as a single file (for example, send only *.c, *.java, *.cpp, *.py) via ODTU-Class before the deadline. Late submission is -20 pts per day.