Lab Assignment

Due date: 3rd November 2021

Motifs, also known as Transcription Factor Binding Sites, are biological patterns of great interest. Motif Search algorithms identify the conserved motifs within a given set of DNA sequences. There are different versions of Motif Search Problem, such as Simple Motif Search, Planted Motif Search, Edited Motif Search, Quorum Motif Search. In this exercise, you are supposed to solve a simplified version of Edit-distance based Motif Search.

The problem is defined as follows:

Given to you are integers L and D and the alphabet set $\sum\{A, C, G, T\}$. Write a program to implement the following:

- Randomly generate 20 strings S_1, S_2 … S_{20} of length 600 each using alphabet set $\sum$.
- For each string i from 1 … 20
  For each substring M in string S_i, where |M| = L:
  If a neighbor of M occurs in each of the other 19 strings, then output M; where a string M' is considered as a Neighbor of M if Edit distance (M, M') <= D.

Edit Distance is defined as follows:

Given to you are two string X[1…p] and Y[1…q], and the ability to perform the following transformations Insertion, Deletion and Substitution. The cost of these transformation operations is as follows: cost of insertion and deletion is *indel* and cost of substitution is *sub*. Our aim to convert X to Y using the minimum number of transformation operations.

To solve this problem, you would apply the Dynamic programming approach. You need to create a matrix E[0…p][0….q], where E[i,j] represents the number of transformations required to convert X[1...i] to Y[1…j]. The value at E[p,q] will give you the Edit Distance.

You initialize the matrix E as follows:

E[i,0] = i for each $0 \leq i \leq p$ and

E[0,j]=j for each $0 \leq j \leq q$.

The other values in E are computed as follows:

$$E[i,j] = \min \begin{cases} E[i-1,j] + indel \\ E[i,j-1] + indel \\ E[i-1,j-1] + \begin{cases} 0, if\ X[i] = Y[j] \\ sub, if\ X[i] \neq Y[j] \end{cases} \end{cases}$$

Test for the following values: L=15, D=5 and submit the following:

1. Program file for the implementation
2. Input file titled “Data.txt”
3. Output File titled “ Out.txt”
4. Word file containing the Algorithm along with the analysis
5. Can your approach can be improved? Discuss any one significant improvement.

Note: L, D, indel and sub values should be input quantities.