HW: Longest common subsequence (LCS)

Description

In HW, you will implement a program to detect the longest common subsequence (LCS) using biological sequences. The human genome is 3.3M characters long, which embeds thousands of genes. By using LCS, we are able to detect specific sequence structures, such as miRNAs, which are known to target genes. An example is given below. The implementation of LCS must be based on dynamic programming (Wikipedia link). In the given biological sequence, a few miRNA like sequences are embedded. Thus, your program should be able to detect them and output the LCS of those sequences.

Example of LCS

> sequence
...TGACATCACCTCTTTCTTTCTCAACACACCCACAAAGGCACACACTGCTGCATAATTTTGCTTTTGTCTGA
GGAAGAAAAATTTAATAACGATACCAATTTTTATTTTTTAATTTTATGTATAAATTAGAAACTACATATGAGGA
GAATACCAGACGTTATTTTTTTGAACGACCACATACATAGCATACACATAATAAATTTAAAATGACATCACCT
CTTTCTTTCTCAACACACCCACAAAGGCACACACTGCTGCATAATTTTGCTTTTGTCTGAGGAAGAAAAATT
TAATAACGATACCAATTTTTATTTTTTAATTTTATGTATAAATTAGAAACTACATATGAGGAG...

The red sequence is the LCS (left), the blue sequence is the LCS (right) as shown in Figure 1. The black characters are the hairpin sequence. When folded, the structure will look like the sequence structure as below, which is similar to a miRNA gene.

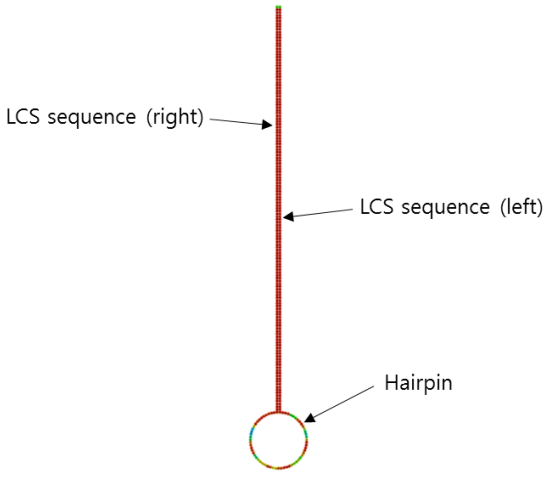

Figure 1. The secondary (folded) structure of the LCS sequence

Also, Insertion and deletion might be existing between two LCS.

> sequence
...TGACCAAATACAGGGCCCCTTAATATTCTTTAGCCGTTTTACCTTTGGAAAAGTTATATATTTTGATAATATG
TGATCTCTCTCGGCCTCACGCAAATGGGGAAGTAATCATTTCCTTTCTTAAATTCTC

ATTTGCTTGCAAATAGACAGTTTAAAACCACATTGTGACT (hairpin)

TGACCA~~AA~~TACAGGGGCCCCTTAATATTCTTTAGACCCGTTTTACCTTTG~~G~~AAAAGTTATATATTTTATGATAA
TATGTGATCTCTCCTCGGCCTCAC~~GCA~~AATGGGGAAGTAATCATTTCCTGGCTTCTTAAATTCTC
...

As you can see, red colored texts are insertions and deletions compared to first LCS.

What to implement and report

We recommend to implement the LCS algorithm in python language. The given biological sequence is 0.1M nucleodites (characters) long. Two LCS structure are separated by **<120 nt** hairpin sequence.

1) As a start, you need to implement dynamic programming with a certain window size. You should try a range of window size and select the best.
2) Run the dynamic programming on the given biological sequence. Report the number of LCS output from your program. How many LCS have you detected?
3) How many insertions or deletions are present?