Advanced DS&A - Assignment #2 (8 points)

[Find the LCS using Dynamic Programming approach] - write Python programs and use the Dynamic Programming skills to find out the longest common subsequence (LCS) of two DNA sequences.

Instructions:

1. The goal of this assignment is to use dynamic programming skills to solve real problems.
2. It is an individual assignment. Discussions with other students are encouraged, but you must implement the assignment by yourself. Also, you must write up a report to describe your ideas and detailed designs of your programs in your own words (copying is prohibited).
3. You can develop the code in a Jupyter notebook in Anaconda (Just a recommendation, not required).
4. You need to generate the report in a PDF file. The report should include one cover page indicating the assignment number, your student ID, and your name. The main content of your report should include the descriptions and the results. Please follow the order of the questions in the report.
5. **Deadline: December 14, 2023 (23:59)**
6. Compress these items “(1). Python source files (2). PDF report” in a ZIP file named as “studentid_assignment2.zip”. Turn in the ZIP file on TronClass (the upload directory will be announced)

Problem Statement

Given two DNA sequences (sequences of A, C, T, G characters), the longest common subsequence (LCS) problem is to find the longest subsequence (not necessarily contiguous) that exists in both of the input DNA sequences. For example, given sequences " ABCDEFG " and " XZACKDFWGH ", the subsequence " ACDFG " is the longest common subsequence.

There are two tasks in this assignment as follows.

Task 1 (40%)

Given two DNA sequences of sizes n1 and n2 respectively, implement the divide-and-conquer algorithm to find the **longest common subsequence (LCS)**. You can use any pair of the following sequences to test your result.

Task 2 (60%)

Given two DNA sequences of sizes n1 and n2 respectively, implement a **dynamic programming algorithm** to find the longest common subsequence. In addition, try to test long DNA sequences to compare the execution time with Task 1.

Sequence 1:
CGTGTTTGCTCTCACGAACTTGACCTGGAGATCAAGGAGATGTTTCTTGTCGAACTGGACAGC
GCTTCAACGGAACGGATCTACGTTACAGCCTGCATAA

Sequence 2:
TGAAAACGGAGTTGCCGACGACGAAAGCGACTTTGGGTTCTGTCTGTTGTCATTGGCGGAAA
ACTTCCGTTCAGGAGGCGGACACTGATTGACACGGTTT

Sequence 3:
TCCTTCGTCTGTGACTAACTGTGCCAAATCGTCTAGCAAACTGCTGATCCAGTTTAACTCACCA
AATTATAGCCGTACAGACCGAAATCTTAAGTCATATCACGCGACTAGCCTCTGCTTAATTTTTGT
GCTCAAGGGTTTTGGTCCGCCCGAGCGGTGCAGCCGATTAGGACCATGTAATACATTTGTTAC
AAGACTTCTTTTAAACACTTTCTTCCTGCCCAGTAGCGGATGATAATCGTTGTTGCCAGCCGGC
GTGGAAGGTAACAGCACCGGTGCGAGCCTAATGTGCCGTCTCCACGAACACAAGGCTGTCCG
ATCGTATAATAGGATTCCGCAATGGGGTTAGCAAGTGGCAGCCTAAACGATATCGGGGACTTG
CGATGTACATGCTTTGGTA

Sequence 4:
CAATACATACGTGATCCAGTTGTTATCCTGCATCGGAACATCAATTGTGCATCGGACCAGCATAT
TCATGTCATCTAGGAGGCGCGCGTAGGATAAATAATTCAATTAAGATGTCGTTTTGCTAGTATAC
GTCTAGGCGTCACCCGCCATCTGTGTGCAGGTGGGCCGACGAGACACTGTCCCTGATTTCTCC
GCTTCTAATAGCACACACGGGGCAATACCAGCACAAGCCAGTCTCGCAGCAACGCTCGTCAG
CAAACGAAAGAGCTTAAGGCTCGCCAATTCGCACTGTCAGGGTCGCTTGGGTGTTTTGCACT
AGCGTCAGGTACGCTAGTATGCGTTCTTCCTTCCAGGGGTATGTGGCTGCGTGGTCAAATGTG
CGGCATACGTATTTGCTCGA

Sequence 5:

AGCAGAAGGTTTGAGGAATAGGTTAAATTGAGTGGTTTAATAACGGTATGTCTGGGATTAAAG
TGTAGTATAGTGTGATTATCGGAGACGGTTTTAAGACACGAGTTCCCAAAATCAAGCGGGGTC
ATTACAACGGTTATTCCTGGTAGTTTAGGTGTACAATGTCCTGAAGAATATTTAAGAAAAAAGC
ACCCCTCATCGCCTAGAATTACCTACTACGGTCGACCATACCTTCGATTATCGCGGCCACTCTCG
CATTAGTCGGCAGAGGTGGTTGTGTTGCGATAGCCCAGTATAATATTCTAAGGCGTTACCCTGA
TGAATATCCAACGGAATTGCTATAGGCCTTGAACGCTACACGGACGATACGAAATTATGTATGG
ACCGGGTCATCAAAAGGTTATACCCTTGTAGTTAACATGTAGCCCGGCCCTATTAGTACAGTAG
TGCCTTGAATGGCATTCTCTTTATTAAGTTTTCTCTACAGCTAAACGATCAAGTGCACTTCCACA
GAGCGCGGTGGAGATTCATTCACTCGGCAGCTCTGTAATAGGGACTAAAAAAGTGATGATAAT
CATGAGTGCCGCGTTATGGTGGTGTCGGAACAGAGCGGTCTTACGGCCAGTCGTATGCCTTCT
CGAGTTCCGTCCAGTTAAGCGTGACAGTCCCAGTGTACCCACAAACCGTGATGGCTGTGCTTG
GAGTCAATCGCAAGTAGGATGGTCTCCAGACACCGGGGCACCAGTTTTCACGCCGAAAGCAT
AAACGACGAGCAGATATGAAAGTGTTAGAACTGGACGTGCCGTTTCTCTGCGAAGAACACCT
CGAGCTGTAGCGTTGTTGCGCTGCCTAGATGCAGTGTTGCTCATATCACATTTGCTTCAACGAC
TGCCGCCTTCGCTGTATCCCTAGACACTCAACAGTAAGCGCTTTTTGTAGGCAGGGGCACCCC
CTATCAGTGACTGCGCCAAAACATCTTCGGATCCCCTTGTCCAATCAA

Sequence 6:

ACTCATCGAATTCTTACATTTAAGACCCTAATATCACATCATTAGTGATTAATTGCCACTGCCAAA
ATTCTGTCCAGAAGCGTTTTAGTTCGCTCCACTAAAGTTGTTTAAAACGACTACCAAATCCGCA
TGTTAGGGGATTTCTTATTAATTCTTTTATCGTGAGGAACAGCGGATCTTAATGGATGGCCGCA
GGTGGTATGGAAGCTAATAGCGCGGGTGAGAGGGTAATCAGCCGTGTTCACCTACACAACGC
TAACGGGCGATTCTATAAGATTCCGCATTGCGTCTACTTATAAGATGTCTCAACGGTATCCGCAA
CTTGTGAAGTGCCTACTATCCTTAAACGCATATCTCGCCCAGTAGCTTCCCAATATGTGAGCATC
AATTGTTGTCCGGGCCGAGATAGTCATGTGCTCACGGAACTTACTGTATGAGTAGTGATTTGA
AAGAGTTGTCAGTTTGCTGGTTCAGGTAAAGGTTCCTCACGCTACCTCAAAGTAAGAGAGCG
GTCGTGACATTATCCGTGATTTTCTCACTACTATTAGTACTCACGACTCGATTCTGCCGCAGCCA
TGTTTCGCCAGAATGCCAGTCAGCATTAAGGAGAGCTCAGGGCAGGTCAACTCGCATAGTGA
GGGTTACATGTTCGTTGGGCTCTTCCGACACGAACCTCAGTTAGCCTACATCCTACCAGAGGT
CTGTGCCCCGGTGGTGAGAAGTGCGGATTTCGTATTTGCAGCTCGTCAGTACTTTCAGAATCA
TGGCCTGCACGGCAAAATGACGCTTATAATGGACTTCGACATGGCAATAACGCCTCGTTTCTAC
GTCAGGAGGAGAATAGTATAAACATAACTGCTGTCGGCAGAAGCGCCAAAGGAGTCTCTGAA
TTCTTATTCCCGAATAACATCCGTCTCCGTGCGGGAAAATCACCGACGGCGTTTTATAGAAGCC
TAGGGGAACAGATTGGTCTAATTAGCTTAAGAGAGTAAATTCTGGG