STEM Education

Task 1: DNA analysis

Student's Version

Time Allocation

A maximum of **60 minutes** in total are required for carrying out the performance task. It is suggested that you should spend 5-10 minutes to tidy up your written work and the work space.

Introduction

Bioinformatics is a growing interdisciplinary field providing techniques in understanding biological data and biological creatures. This technology can be applied to the forensics. During the criminal investigation, the DNA on the murder weapon and that of the suspects are compared, so that the murderer can be identified.

In this set of performance task, you are going to develop a good understanding of this technology, and apply your knowledge in basic biology, mathematics and engineering to design a solution through analysing the DNA sequence with a new algorithm called **Longest Common Subsequence (LCS)** algorithm.

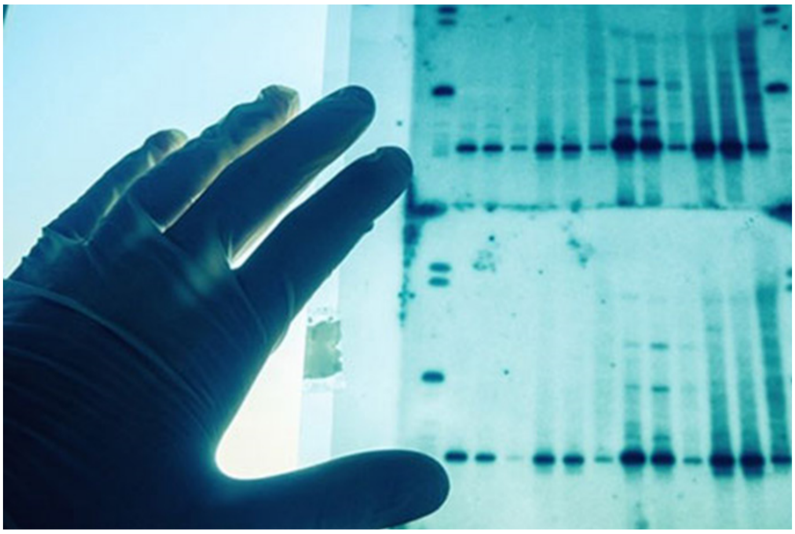
You should have read the pre-task reading about the LCS algorithm. You are allowed to refer to the reading about LCS algorithm while completing the task.

Please study the following case carefully.

CASE SCENARIO

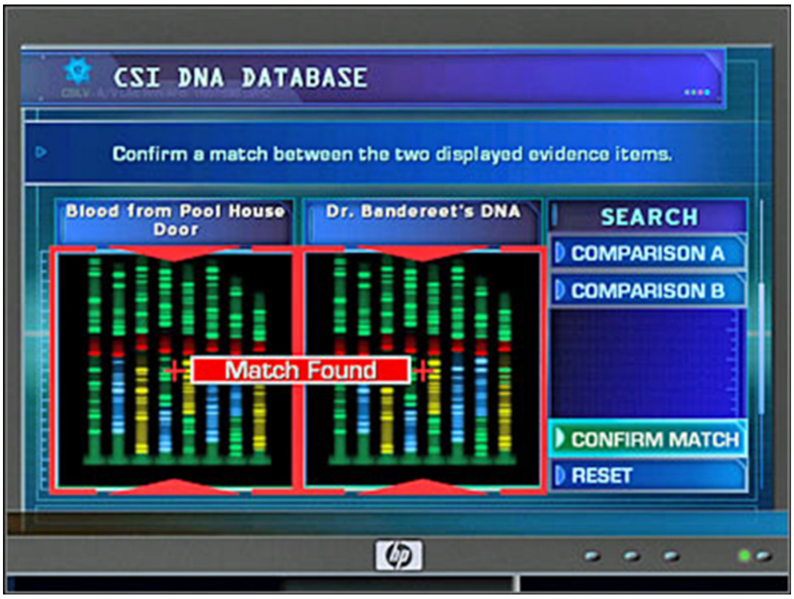
Source of image: https://www.tresham.ac.uk/course-display/full-time-course/?id=194&page=science

Recently, a homicide occurred in a research lab in which two research assistants, Dr. M and Dr. Y worked. They were good friends. After gathering information from some witnesses, the police suspected that Dr. M killed Dr. Y and then fled away. The police and forensic scientists had worked together and collected all the evidences including the DNA samples. Many evidences seem to confirm that Dr. M is the murderer, but the forensic scientists still have to carry out further experiments to compare the collected DNA samples with that of the suspect to confirm whether Dr. M is the right suspect of the homicide crime.



Source of image: https://www.allcriminaljusticeschools.com/forensics/dna-fingerprinting/

CASE SCENARIO (CONT.)



Confirm a match between the two displayed evidence items.

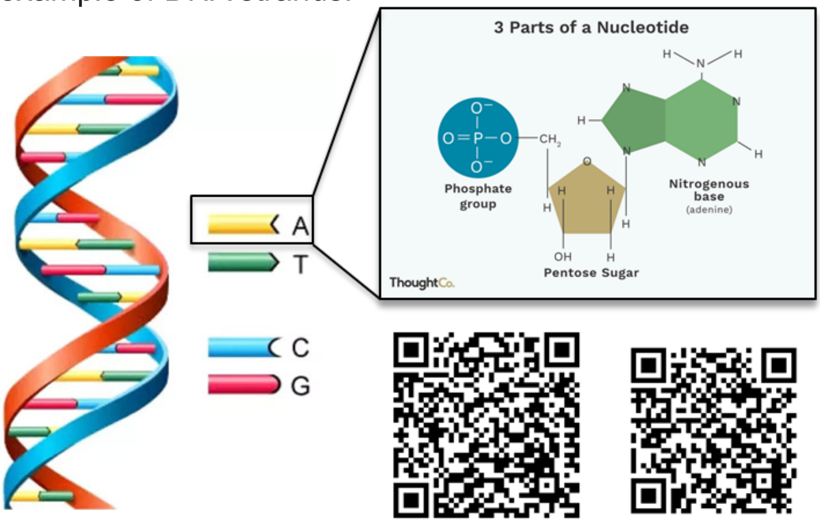
Source of image: https://guides.gamepressure.com/crimesceneinvestigationhardevidence/guide.asp?ID=3539

To solve this problem, you have been hired as a junior DNA forensic assistant. You are going to learn a new algorithm called Longest Common Subsequence (LCS) algorithm. With the help of computer programming, you think this simple method may be used to develop a new mobile application that can help the frontier police officers and the forensic scientists to easily compare two DNA samples and thus check the percentage of similarity match.

Before attempting this task you should have learned about the LCS algorithm. You are required to apply related knowledge to develop the mobile app.

Pre-task reading:
Understanding of LCS Algorithm

DNA (Deoxyribonucleic acid) is a macro-molecule composed of two nucleotide chains coiling around each other. All nucleotides are composed of three parts: the phosphate group, the pentose, and the nitrogenous base. The only difference between the nucleotides is the nitrogenous base including: cytosine [C], guanine [G], adenine [A] or thymine [T]. DNA sequence is basically represented as a succession of these four letters (C, G, A and T) to indicate the order of the nitrogenous bases in the DNA strands. A sample structure is given below as a visual example of DNA strands:



Source of image:
https://www.quora.com/Why-does-a-DNA-molecule-always-contain-equal-amounts-of-adenine-and-thymine-explain
https://www.thoughtco.com/what-are-the-parts-of-nucleotide-606385

The following are two different strands (S_1 and S_2) of DNA from two organisms:
S_1 = AATCCCCAGCTAG
S_2 = AAACGTACCTTAG

DNA analysis using Longest Common Subsequence (LCS) algorithm aims to compare the DNA of two (or more) different organisms for an understanding of their genetic similarity and differences.

In this task, you will be required to first understand the basics of LCS algorithm, and then answer the questions (Section 1 and 2) on paper.

Pre-task reading: Understanding of LCS Algorithm (cont.)

As mentioned previously, a DNA strand consists of a string of molecules and the sequence of the string of molecules can simply be expressed over the finite set {A, C, G, T}. Here is the definition of sequence and subsequence.

Section 1: Definition of Sequence and Subsequence

Each strand, denoted by $S = < s_1, s_2, \ldots, s_n >$ where n is a positive integer index, can be called a **sequence**. For example, *S* = <*AATCGCG*> is a sequence. Another sequence $Z = < z_1, z_2, \ldots, z_m >$, where m is also a positive integer index, is called a **subsequence** of *S* if there exists a strictly increasing sequence (i.e. matching and comparing from left to right) $< i_1, i_2, \ldots, i_j >$ of indices of S such that for all $j = 1, 2, \ldots, k$, we have $s_{i_j} = z_j$. For example, Z = <ATGCG> is a subsequence of S with corresponding index sequence <1,3,5,6,7> with a length of 5. Other subsequences are <AACCG>, <ACG> and <ACCG> but with different index sequences and lengths.

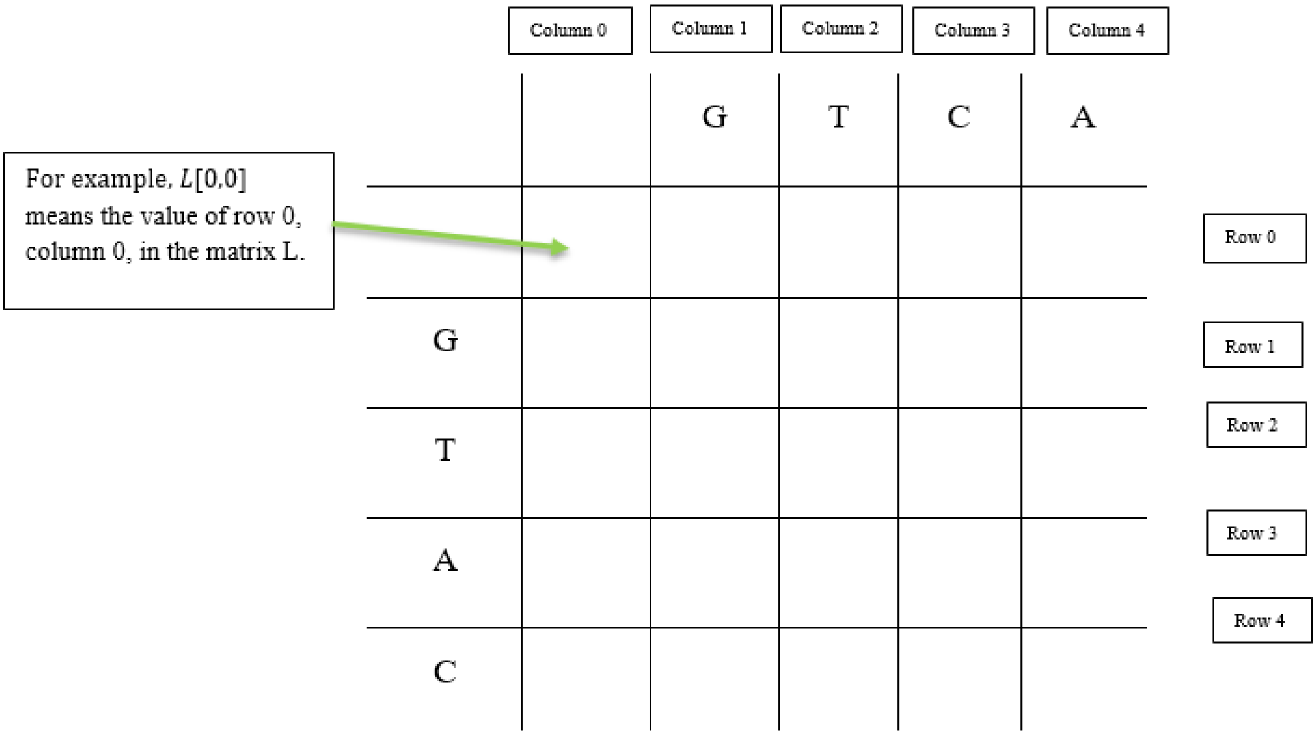
				Index sequence of S	Length
			<1234567>		
The sequence	S	=	<AATCGCG>		
Subsequence 1	Z_1	=	<A T GCG>	<13567>	5
Subsequence 2	Z_2	=	<AA C CG>	<12467>	5
Subsequence 3	Z_3	=	<A C G>	<147>	3
Subsequence 4	Z_4	=	<A C CG>	<1467>	4

Given two sequences X and Y, then Z is a **common subsequence** of X and Y if Z is a subsequence of X and Y. Furthermore, if Z has the maximum-length when comparing with all other common subsequences, then Z is the **longest common subsequence** (LCS) of X and Y. For example, if X = <AAAAACCCCCTTTTT> and Y = <AAAAAGCCTTTGGGGT>, then Z = <AAAAACCTTTT> is the LCS of X and Y. In finding the longest common subsequence, the one with the longest “length” should be the solution, but there could be more than one possible solutions. For example, if X = <ATTTCTG> and Y = <ATGATT>, then Z = <ATT> or <ATG> are the LCS of X and Y.

PRE-TASK READING: UNDERSTANDING OF LCS ALGORITHM (CONT.)

Section 2: Introducing the concept of LCS algorithm

Before we attempt to find the LCS, we can first understand how to find the length of LCS. Let's start with a simple problem. Suppose we have two DNA sequences, X = <GTAC> and Y = <GTCA>. We can structure it into a matrix as below where sequence X is written on the left-most column, and sequence Y is written on the top row:



For example, *L*[0,0] means the value of row 0, column 0, in the matrix L.

	Column 0	Column 1	Column 2	Column 3	Column 4	
		G	T	C	A	
						Row 0
G						Row 1
T						Row 2
A						Row 3
C						Row 4

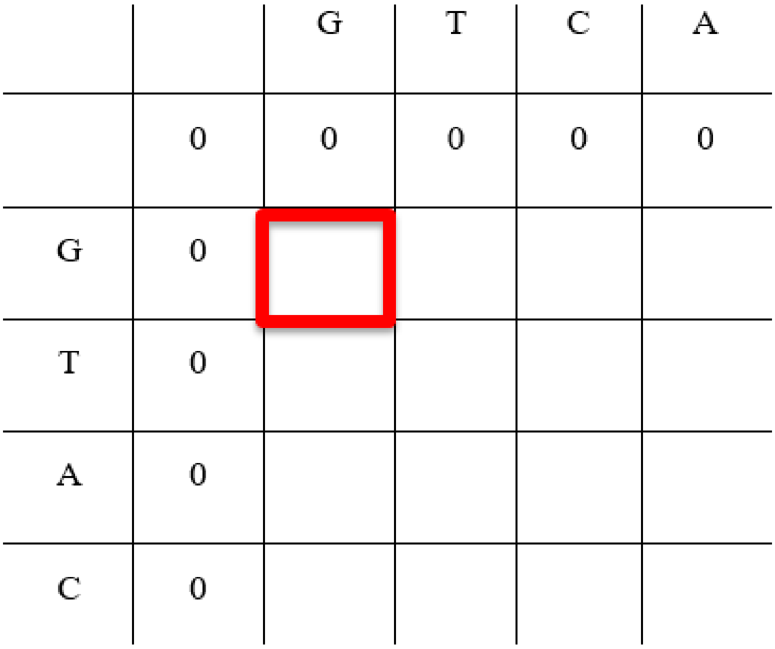
In order to find the length of LCS, let us try to fill up the matrix *L* with the value of *L[i,j]* in row *i* and column *j* of the matrix by the following rules:

1. *L*[0,j] = 0 and *L*[i,0] = 0
2. If *X*[i] = *Y*[j],then *L*[i,j] = 1 + *L*[i-1,j-1]. When i>0 and j>0
3. If not,then *L*[i,j] = max(*L*[i-1,j],*L*[i,j-1]).

PRE-TASK READING: UNDERSTANDING OF LCS ALGORITHM (CONT.)

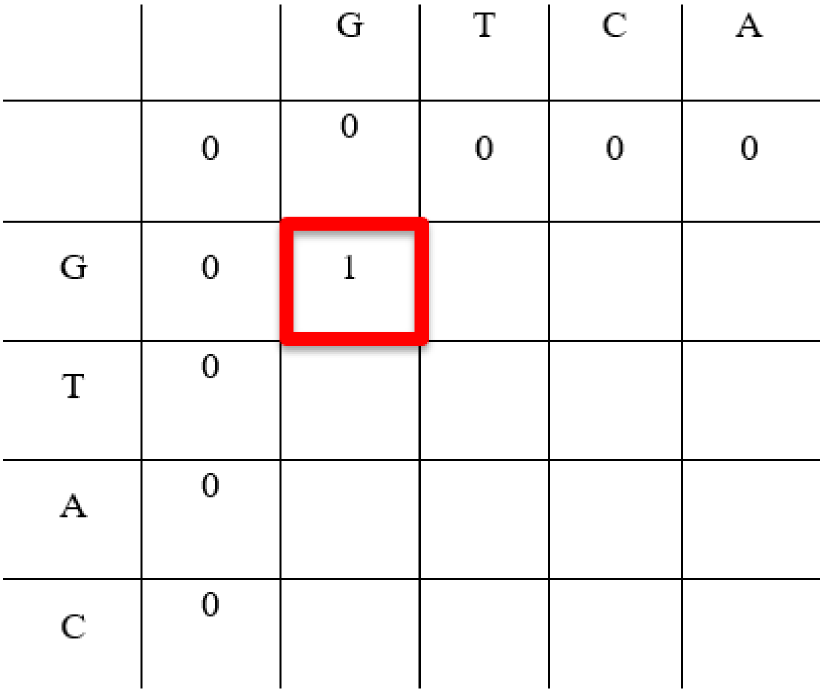
For these rules, we actually try to determine the length of the longest common subsequence (LCS) within the range by adding one more letters for comparison.

1. Fill up the zero in row 0 and column 0 using Rule 1:



		G	T	C	A
	0	0	0	0	0
G	0				
T	0				
A	0				
C	0				

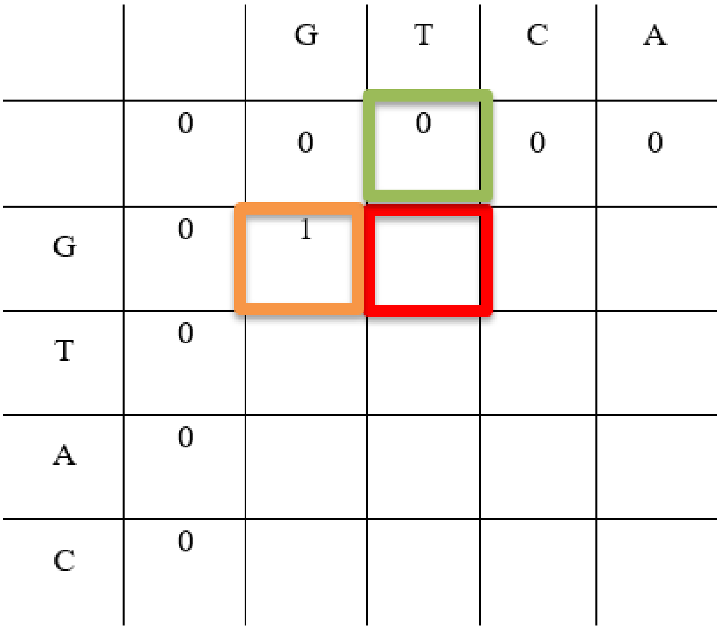
2. Put down *L*[1,1], as shown below:
Because *X*[1] = *G* and *Y*[1] = *G*, Rule 2 above applies. Thus, *L*[1,1]=1+0=1. Then we fill it into the matrix as follows:



		G	T	C	A
	0	0	0	0	0
G	0	1			
T	0				
A	0				
C	0				

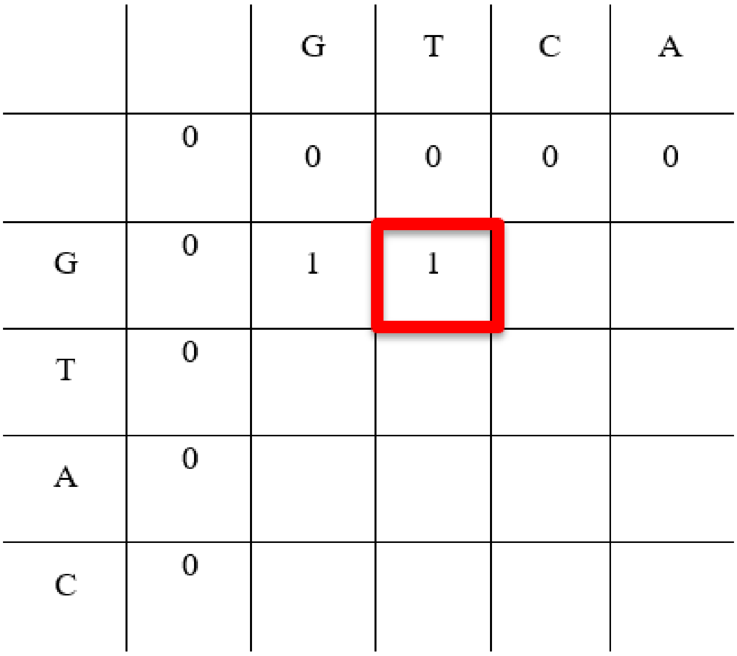
PRE-TASK READING:
UNDERSTANDING OF LCS ALGORITHM (CONT.)

After finishing this one, move on to the right. The next one is indicated below:



		G	T	C	A
	0	0	0	0	0
G	0	1			
T	0				
A	0				
C	0				

Because *X*[1] = *G* and *Y*[2] = *T*, hence *X*[1]≠*Y*[2]. Rule 3 above applies. Thus, *L*[1,2]=*max*(*L*[1-1,2],*L*[1,2-1])=*max*(*L*[0,2],*L*[1,1])=max(0,1)=1, because L[0,2] = 0 (in green box) and L[1,1] = 1 (in orange box), and 0 < 1. So we choose the larger number “1” as the answer. Then we fill it into the matrix as follows:



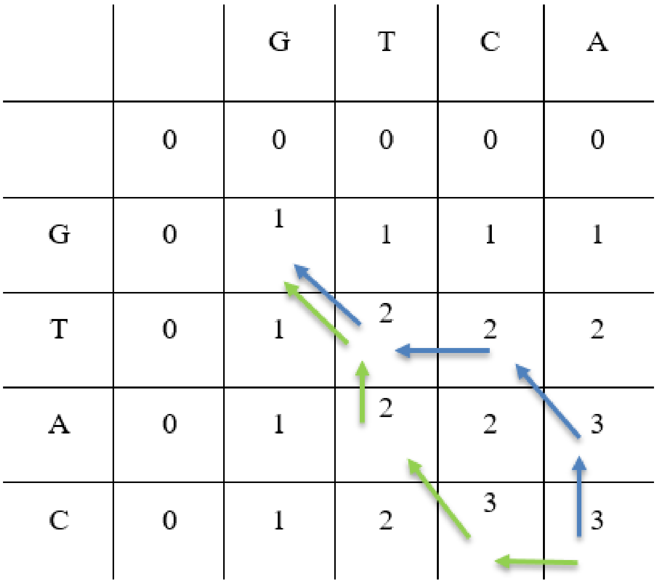
		G	T	C	A
	0	0	0	0	0
G	0	1	1		
T	0				
A	0				
C	0				

PRE-TASK READING: UNDERSTANDING OF LCS ALGORITHM (CONT.)

After going through all the steps, we have the following final matrix:

		G	T	C	A
	0	0	0	0	0
G	0	1	1	1	1
T	0	1	2	2	2
A	0	1	2	2	3
C	0	1	2	3	3

Finally, the final answer (the length of the LCS of X and Y) is 3.
How can we now find the sequence? Follow the path as indicated below, we will be able to find all letters from the sequence starting the lower-right corner (both blue and green arrows can give a possible solution):



		G	T	C	A
	0	0	0	0	0
G	0	1	1	1	1
T	0	1	2	2	2
A	0	1	2	2	3
C	0	1	2	3	3

Starting from the lower-right corner, follow the rules below:

1. If the cell directly above or directly on the left contains an equal value as the current cell, then move to the cell with an equal value. If both cells contains equal value, then move to either one of the cells;
2. If both of the cells (above and left) have values strictly less than the value in the current cell, then move diagonally up-left; Write down the current letter (which would be the same from the row and the column).

As illustrated above, by studying the blue arrows, the backward output sequence is ATG. After reversing it, we have GTA (or GTC if follows the green arrows) as the LCS.

QUESTIONS

1. Given M = <GTTCCCAGTGGCTAA> (Dr. M's DNA sequence), and Y = <TCCAGGCTATGCTAA> (the DNA sample obtained from the clothes found in Dr. Y's house):

   Help the forensic scientists to find the answer to identify one possible longest common subsequence Z using any method in which the expected length is 11.

2. To find the similarity percentage of two DNA sequences we can simply calculate it by

$$\frac{Length\ of\ Longest\ Common\ Subsequence}{Length\ of\ Original\ Sequence} \times 100\%$$

   For example, if the length of X is 5, and the length of the longest common subsequence with Y is 4, then the similarity will be:

$$\frac{Length\ of\ Longest\ Common\ Subsequence}{Length\ of\ Original\ DNA\ Sequence} \times 100\%$$

$$= \frac{4}{5} \times 100\% = 0.8 \times 100\%$$

$$= 80\%$$

   Compute the expected similarity percentage in Q1 (Rounding to the nearest 1 decimal place).

QUESTIONS (CONT.)

3. As a junior forensic assistant, you were asked to provide your method of finding the longest common sequence (LCS) for the DNA analysis. Describe your method and the process of finding the longest common subsequence of DNA sequence and explain how you could be certain with your answer.

4. Suppose there is a DNA sample with a sequence length of 4. What is the total number of all possible and distinct subsequences? (Note: The outcomes of two or more subsequences may look the same since some letters representing the nitrogenous base in the sequence may be the same. But they are regarded as different subsequences because those representing letters come from different positions in the sequence.)

5. List all the possible and distinct subsequences of the DNA sequence Z = <ATGC>.

6. Explain why the algorithm (or method) can be actually used to compute the LCS.

QUESTIONS (CONT.)

7. Using the LCS algorithm fill in the matrix provided and find the LCS and its length of the DNA sequence X and Y, where X = <GTTCCCAGTGGCTAA> and
   Y = <TCCAGGCTATGCTAA>.

Questions (cont.)

8. Reflect on the impact of this LCS algorithm toward the bioinformatics and forensic science in criminal investigation. Discuss the challenges and issues in adopting this computational solution for scientific investigation.

9. Describe how this DNA analysis relates the knowledge of all the disciplinary content in Science, Technology, Engineering, and Mathematics (STEM). Can this problem be solved optimally without at least one of the disciplines? Is there any other important knowledge we may need beyond STEM education to solve this DNA analysis problem? (Reference video to gain more insight: https://youtu.be/slUaVeNvuTk and https://youtu.be/W-WtMvNdEcM).