

STEM 教育

活動(1): DNA分析

學生版

時間分配

進行這個實作評量需要的時間，最多為**60分鐘**。建議你應該預留5-10分鐘來整理答案和工作空間。

引言

生物信息學是一門正在發展的跨學科領域技術，該技術可讓我們瞭解生物及其相關的數據。這門學科可應用於法證科學。當進行犯罪調查時，我們需要比對謀殺案凶器上的DNA 及不同疑犯的DNA，從而找出真兇。

在這次的實作評量中，你將會更了解此技術的原理，並運用它在基礎生物學、數學和工程領域的知識，分析DNA序列及使用**最長共同子序列 (LCS)** 的演算法來設計解決問題方案。

你應該已經閱讀了有關 LCS 演算法的「活動前閱讀資料」。在評量活動期間，你仍可參考 LCS 演算法的資料以完成此活動。

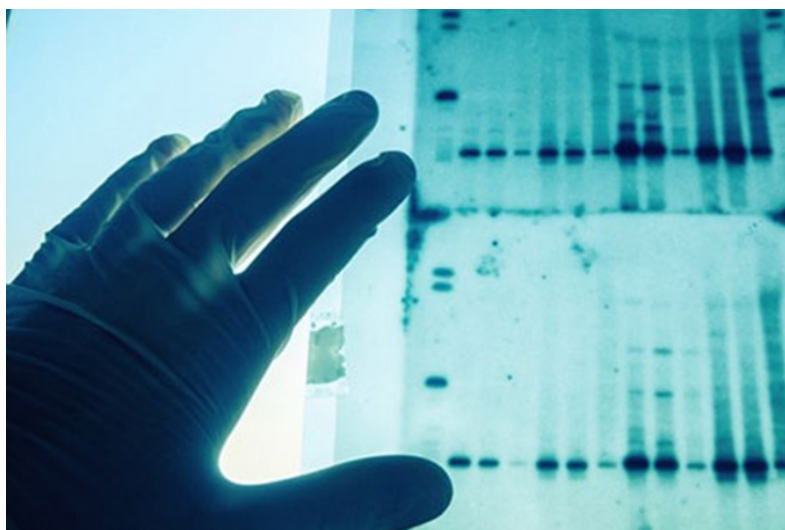
請小心細閱以下個案。

案例情景



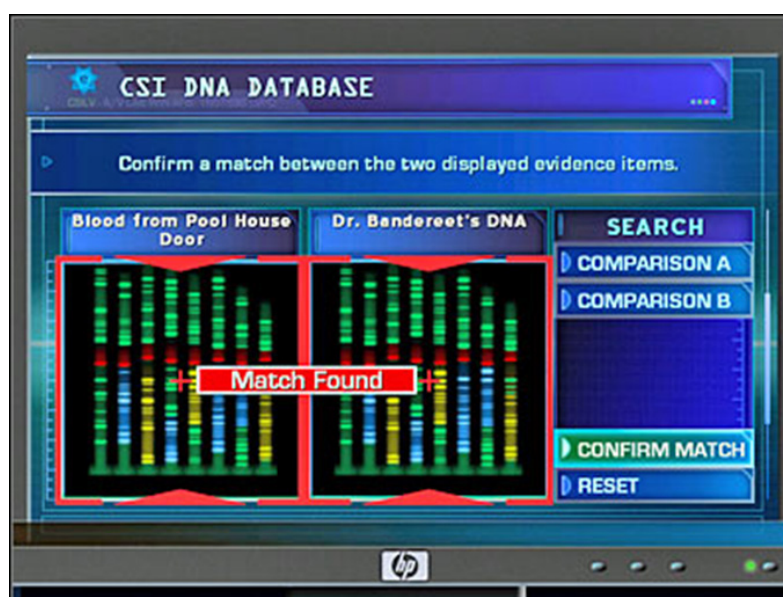
圖像來源: <https://www.tresham.ac.uk/course-display/full-time-course/?id=194&page=science>

最近，一個研究實驗室發生了一宗兇殺案，案中涉及兩名研究助理M博士和Y博士，他們彼此是好友。警方與相關的證人錄取口供後，懷疑M博士殺死了Y博士後逃離現場。警方和法證人員合作，收集了所有在現場留下的證據，當中包括一些DNA的樣本。雖然眾多證據似乎證實了真兇是M博士，但他們仍需對收集得來的DNA樣本與嫌疑犯的DNA進行分析比對，以確認M博士是殺人兇手。



圖像來源: <https://www.allcriminaljusticeschools.com/forensics/dna-fingerprinting/>

案例情景 (續)



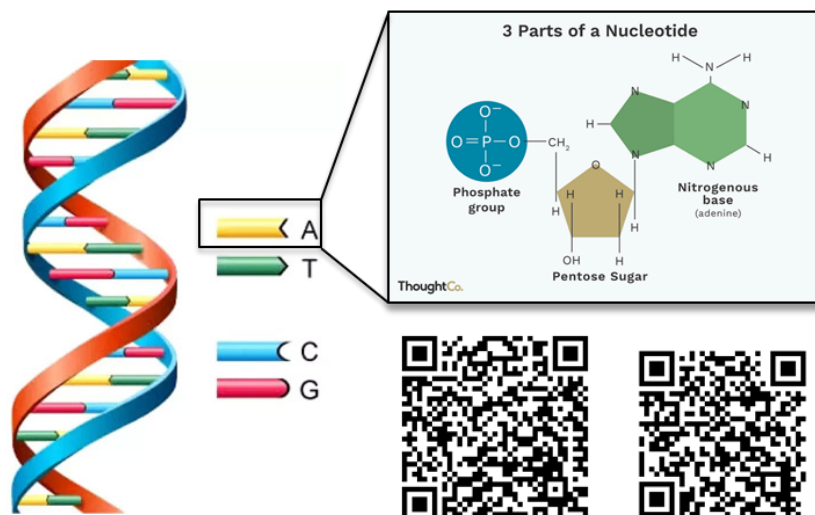
圖像來源: <https://guides.gamepressure.com/crimesceneinvestigationhardevidence/guide.asp?ID=3539>

為了解決這個問題，你被聘為初級DNA法證助理。你將學習一種稱為「最長共同子序列(LCS)」的演算法。你亦認為可利用電腦程式設計軟件開發一個流動應用程式，幫助前線警務人員和法證人員輕鬆地對比兩個DNA樣本相似匹配的百分比，及分析兩個DNA有多少的吻合程度。

在開始進行評量活動之前，你應該已對LCS演算法有一定的了解。你需要運用有關的知識來開發流動應用程式。

活動前閱讀資料： 理解LCS演算法

DNA(去氧核糖核酸)是由兩條核苷酸長鏈互相纏繞所組成的大分子。每個核苷酸均由三部份所組成，它們分別是戊糖、磷酸基及含氮鹼基。而不同類型的核苷酸唯一的差別是它們擁有不同的含氮鹼基，當中包括：胞嘧啶[C]、鳥嘌呤[G]、腺嘌呤[A]和胸腺嘧啶[T]。科學家們皆利用C、G、A、T四個字母代表相應的含氮鹼基，然後利用此四個字母排序，來表示DNA序列。下圖為DNA鏈的結構圖：



圖像來源:

<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>

下列是來自兩個不同生物的DNA鏈：

S1= AATCCCCAGCTAG

S2= AAACGTACCTTAG

利用最長共同子序列(LCS)演算法去進行DNA分析，目的是比較兩個(或更多)不同生物的DNA，以瞭解其遺傳相似性和差異。

在此評量活動中，你需要首先瞭解LCS演算法的基本知識，然後回答工作紙的問題(第一及第二部分)。

如前所述，DNA鏈是由一串分子組成，這些分子的排序會簡單地以「有限集」 $\{A、C、G、T\}$ 表示。下面是序列和子序列的定義。

活動前閱讀資料： 理解LCS演算法 (續)

第一部分：序列和子序列的定義

任何**序列**皆可表示為 $S = \langle s_1, s_2, \dots, s_n \rangle$ ，其中 n 是一個正整數索引。例如， $S = \langle \text{AATCGCG} \rangle$ 是一個序列。現有另一個序列 $Z = \langle z_1, z_2, \dots, z_m \rangle$ ，其中 m 是一個正整數索引，如果 S 存在有一個嚴格漸增的索引序列 (即順序從左至右匹配與比較) $\langle i_1, i_2, \dots, i_j \rangle$ ，使到對所有 $j = 1, 2, \dots, k$ 我們有 $s_{i_j} = z_j$ ，則 Z 便是 S 的子序列。例如， $Z = \langle \text{ATGCG} \rangle$ 是 S 的相應索引序列 $\langle 1, 3, 5, 6, 7 \rangle$ 的子序列，序列長度為 5。

S 還有其他子序列，例如 $\langle \text{AACCG} \rangle$ 、 $\langle \text{ACG} \rangle$ 和 $\langle \text{ACCG} \rangle$ ，相對而言，它們擁有不同的索引序列和序列長度。

序列	S	$\langle 1234567 \rangle$	S 的索引序列	序列長度
子序列 1	Z_1	$\langle \text{A T GCG} \rangle$	$\langle 13567 \rangle$	5
子序列 2	Z_2	$\langle \text{AA C CG} \rangle$	$\langle 12467 \rangle$	5
子序列 3	Z_3	$\langle \text{A C G} \rangle$	$\langle 147 \rangle$	3
子序列 4	Z_4	$\langle \text{A C CG} \rangle$	$\langle 1467 \rangle$	4

已知兩個序列 X 和 Y ，若 Z 分別是 X 和 Y 的子序列，則 Z 是 X 和 Y 的**共同子序列**。當 Z 與其他的共同子序列比較時， Z 的序列長度為最長， Z 會被稱為 X 和 Y 的**最長共同子序列(LCS)**。

例如，當 $X = \langle \text{AAAAACCCCCTTTTT} \rangle$ 和 $Y = \langle \text{AAAAAGCCTTTGGGGT} \rangle$ ，則

$Z = \langle \text{AAAAACCTTTT} \rangle$ 是 X 和 Y 的 LCS。在找尋最長的共同子序列時，是以“長度”最長的子序列為答案，但應注意 X 和 Y 可能有多於一個最長共同子序列(LCS)。例如，當 $X = \langle \text{ATTTCTG} \rangle$ 和 $Y = \langle \text{ATGATT} \rangle$ ，則 $Z = \langle \text{ATT} \rangle$ or $\langle \text{ATG} \rangle$ 都是 X 和 Y 的 LCS。

活動前閱讀資料： 理解LCS演算法 (續)

第二部分: 介紹LCS演算法的概念

在嘗試找尋 LCS 之前，我們需先瞭解如何找尋 LCS 的長度。讓我們從一個簡單的問題開始。假設我們有兩個 DNA 序列， $X = \langle GTAC \rangle$ 和 $Y = \langle GTCA \rangle$ 。我們可以將其建構為一個矩陣，如下圖所示，其中序列 X 寫在最左側的列上，序列 Y 寫入頂行：

例如, $L[0,0]$ 在矩陣 L 中表示行 0 列 0 的值。



	第 0 列	第 1 列	第 2 列	第 3 列	第 4 列
行 0		G	T	C	A
行 1	G				
行 2	T				
行 3	A				
行 4	C				

為了找出 LCS 的長度，讓我們嘗試用以下的規則，填寫矩陣L內每一個記錄 $L[i,j]$ 的值，而 $L[i,j]$ 是表示為在矩陣的第i行, j列中的值：

1. $L[0,j] = 0$ 和 $L[i,0] = 0$
2. 若 $X[i] = Y[j]$ ，則 $L[i,j] = 1 + L[i-1,j-1]$ 。(當 $i > 0$ 和 $j > 0$)
3. 若非，則 $L[i,j] = \max(L[i-1,j], L[i,j-1])$ 。

活動前閱讀資料： 理解LCS演算法 (續)

對於這些規則來說，我們實際上是嘗試在範圍內每次添加一個字母進行比較，以確定最長共同子序列的長度。

1. 使用規則一，將數值零填在零行及零列：

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

2. 填寫 $L[1,1]$ 如下：

由於 $X[1] = G$ 和 $Y[1] = G$ ，故此，規則2適用。因此 $L[1,1] = 1+0=1$ 。然後，我們將其填寫到矩陣中，如下所示：

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

活動前閱讀資料： 理解LCS演算法 (續)

一旦我們完成了這個，我們就向右移一格。如下圖所示：

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

因為 $X[1] = G$ 和 $Y[2] = T$ ，所以 $X[1] \neq Y[2]$ ，因此，規則3適用。即

$L[1,2] = \max(L[1-1,2], L[1,2-1]) = \max(L[0,2], L[1,1]) = \max(0, 1) = 1$ ，因為 $L[0,2] = 0$ (綠色框) 和 $L[1,1] = 1$ (橙色框)，和 $0 < 1$ 。所以我們選擇較大的數值「1」作為答案。然後，我們將較大值「1」填寫到矩陣中，如下圖所示：

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

活動前閱讀資料： 理解LCS演算法 (續)

依上述步驟，完成所有的空格後，我們最終獲得以下的矩陣：

		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

最後，答案 (X 和 Y 的 LCS 長度) 為 3。

我們現在如何能夠找到序列呢？按照下面所示的路徑，我們先從右下角開始，便能夠找到序列的所有字母 (藍色和綠色箭頭都能給出可行的答案)，方法如下：

		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

先從右下角開始，遵循以下規則：

1. 如果正上方一格或正左方一格的儲存格的值與目前儲存格的值相等，則可移往數值相同的一格。如兩個儲存格的數值均相同，則可任意向正上方或正左方移動一個儲存格；
2. 如果這正上方及正左方的儲存格的值均小於當前儲存格中的值，則沿左上方斜移一格；並記下當前行或列的字母 (此時，行和列的字母都是一樣的)。

如上所述，完成後，追蹤藍色箭頭方向，會得到輸出為 **ATG** 的反向序列，把它顛倒回順序方向後，得到的LCS是**GTA** (或追蹤綠箭頭方向而得到答案 **GTC**)。

問題

1. 已知 M = <GTTCCCAGTGGCTAA> (M博士的DNA序列) 和 Y = <TCCAGGCTATGCTAA> (從Y博士家中發現的衣物中獲得的DNA樣本)

試協助法醫學家，使用任何方法找出一個可能的最長共同子序列Z，預計序列的長度是11。

2. 為了找到兩個 DNA 序列的相似度百分比，我們可以簡單地通過以下方式計算

$$\frac{\text{最長共同子序列的長度}}{\text{原本序列的長度}} \times 100\%$$

例：如果 X 的原本序列長度為 5，而 Y 的最長共同子序列的長度為 4，則相似性將為：

$$\begin{aligned} &= \frac{4}{5} \times 100\% = 0.8 \times 100\% \\ &= 80\% \end{aligned}$$

試計算 問題 (1) 中的預期相似度百分比 (取小數點後1個位)。

問題 (續)

3. 你作為一名初級法醫助理，要為 DNA 分析提供一個找到最長共同子序列的方法。試描述你的方法和尋找 DNA 序列最長共同子序列的程序，並解釋你如何能夠確定你的答案是正確的。
4. 假設有一個 DNA 樣本的序列 Z 長度為 4。它總共有多少個不同的子序列？(注意: 雖然序列中的某些不同位置會有相同的含氮鹼基的代表字母，而導致兩個或以上的子序列看起來一樣，但由於字母是取自序列的不同位置，所以這些子序列會被認為是不相同的。)
5. 試列出 DNA 序列 Z = <ATGC> 所有不同的子序列。
6. 試解釋為甚麼所提供的演算法 (或方法) 可以用來計算 LCS (最長共同子序列)。

問題 (續)

7. 使用提供的 LCS (最長共同子序列) 演算法，填寫下列的矩陣，並找出 DNA X 和 DNA Y 序列的 LCS 及其長度。
(X = <GTTCCCAGTGGCTAA> 和 Y = <TCCAGGCTATGCTAA>)

問題 (續)

8. 反思該 LCS 演算法在刑事偵查中對生物信息學和法證科學的影響。討論採用這種計算解決方案而進行的科學探索所面臨的問題和挑戰。

9. 試描述進行基因分析時，如何需要將科學、科技、工程和數學 (STEM) 的學科知識聯繫起來，以解決基因分析遇到的問題。在解決基因分析問題時，如果缺少一個或以上和 STEM 有關的學科的幫助，會否遇到困難？另外，要解決 DNA 分析時所遇到的困難，除了 STEM 以外，還需要其他重要知識的幫助嗎？(可參考以下視頻，以獲得更多的洞見：
<https://youtu.be/sIUaVeNvuTk>和<https://youtu.be/W-WtMvNdEcM>)