考生編號_____ 分數_____

2023 年中華民國生物奧林匹亞國手選拔營實作試題

第 3 試場實作

生物資訊學實作

項目	數量
筆電(已安裝 MEGA-11 程式)	1 台
筆電桌面-存放**生物資訊學實作資料夾**(內含 mtCDNA SEQ.meg 檔案)	1 個
筆電桌面(存學生考號的資料夾)	1 個

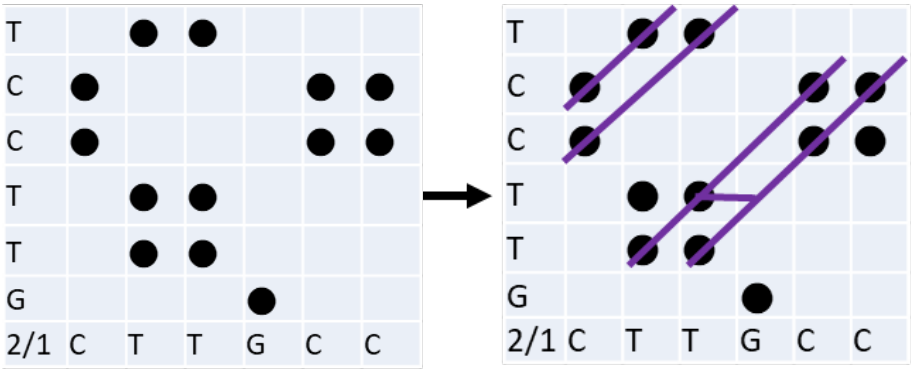
※ 請注意：

1. 請核對試卷上的考號是否和你的考號相同。
2. 本試卷(含封面、試題卷)共 6 頁，於交卷時全部繳回。
3. 作答時間共為 90 分鐘。
4. 請於本卷上作答，試題答案可寫在試卷背面，但請註明並標上題號。

第一題：

在生物資訊學的研究中，在生物信息學中，序列比對排列是一種將 DNA、RNA 或蛋白質序列排列以識別相似區域的方法，是最基本的工作項目之一，後續進行這些區域序列的功能、結構或演化關係分析。核苷酸或氨基酸鹼基的比對排列方法發展，點陣法(Dot-matrix methods)是其中一種快速且簡便的方法，點圖法可用於評估二個序列的相似性，利用點陣圖的方法，以 DNA 序列為例，通過在圖的 X 軸排列一個序列，在 Y 軸上排列另一個序列來比較兩個序列。當兩個序列的鹼基在圖中的相同位置匹配時，在相應位置繪製一個點。二個序列在兩個軸上的起點需要在同一點開始，X 軸上的序列對應到 Y 軸時若是相同鹼基則在此格中繪製點，如下圖例圖顯示，序列 1 序列 CTTGCC 列於 X 軸上，序列 2 序列 GTTCCT 列於 Y 軸上，X 軸和 Y 軸對應相同的鹼基位置繪製一個點，形成下列的結果。

一旦繪製點，以 45 度線將鄰近的點連接成線，相似序列的接近程度將決定對角線與顯示直接關係的連接線的接近程度，這種關係會受到二個序列間差異特徵的影響，例如移位，而在序列排列時加入空格方式顯示。



二序列排列的方法可整理成下列 5 種可能：

```
1 CTTGCC-   1 ---CTTGCC    1----CTTGCC    1 CTTGCC    1 CTTGCC-
    II II         I I           II          II I         II II
2 GTT-CCT   2 GTTCCT---    2 GTTCCT----    2 GTTCCT    2 GTT-CCT
```

而如何決定最佳的排列方法，則利用計算方法達成，計算式如下：

- $D = s + wg$

s 代表鹼基替換數目、g 代表空格的總數目、w 是空位長度加權分。因此下列四種排列方式的計算分別為：

```
1 CTTGCC-   1 ---CTTGCC    1----CTTGCC    1 CTTGCC    1 CTTGCC-
    II II         I I           II          II I         II II
2 GTT-CCT   2 GTTCCT---    2 GTTCCT----    2 GTTCCT    2 GTT-CCT
 D=1+1*2=2     D=1+3*6=19     D=0+4*8=32     D=3+0*0=3    D=1+1*2=3
```

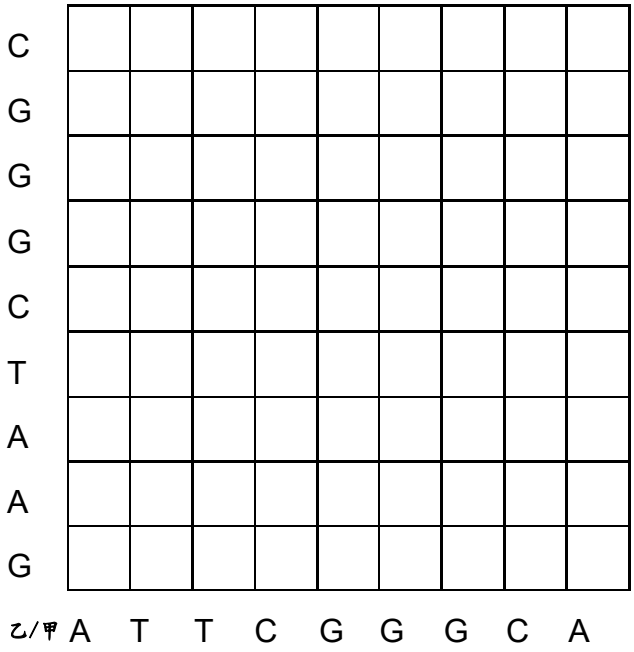
依據上列結果，此二序列最佳的排列情形則為：

```
1 CTTGCC-
    II II
2 GTT-CCT
```

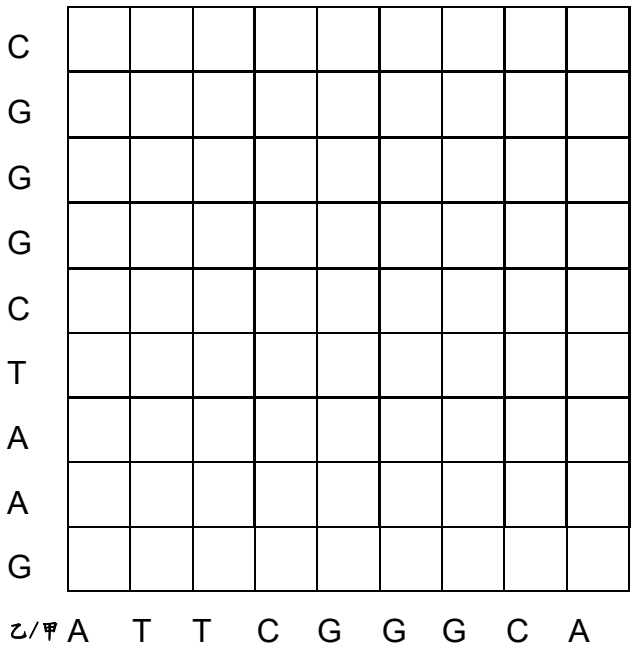
因為計算的 *D* 值為 2，為最小值

若有二序列如下：
序列甲：ATTCGGGCA
序列乙：GAATCGGGC
將此二序列進行比對排列，回答下列問題：

問題 1
利用點陣圖的方法，在下圖中繪製點：(5 分)



問題 2
利用問題 1 繪製的點陣圖結果，45 度線將鄰近的點連接成線，繪製出所有可能的序列排列方式：(5 分)



問題 3

利用問題 2 繪製出可能的序列排列方式，如上述題幹說明，將二序列可能排列方法整理寫出五種可能：(10 分)

問題 4

利用 $D = s + wg$ 計算方法，將問題 3 整理出的二序列可能排列方法以此計算式進行計算，s 代表鹼基替換數目、g 代表空格的總數目、w 是空位長度加權分，每一種可能的排列需要清楚寫出計算式與計算過程：(10 分)

問題 5

依據上列結果，寫下此二序列最佳的排列與原因。(4 分)

第二題：

利用 DNA 序列進行生物親緣關係分析已經是探討生物演化歷史的常用分析方法，在 mtCDNA SEQ.meg 檔案中，分別有人類(homo sapiens)、黑猩猩(chimpanzee)、倭黑猩猩(bonobo)、大猩猩(gorilla)、猩猩(orangutan)、蘇門答臘白掌長臂猿(sumatran)、長臂猿 (gibbon)等七種靈長類生物的某序列片段，利用此序列資料進行分析。在建構樹型關係圖的方法中，算術平均值不加權配對法(Unweighted Paired-Group Method Using Arithmetic Means, UPGMA)是一種常用的簡單方法，為研究不同樣本間的相似性，透過對樣本進行聚類分析建構樹型關係圖，UPGMA 作法首先將距離最小的兩個樣品聚在一起，來形成一個新的節點，其分支點位於兩個樣品間距離的 1/2 處，然後計算此分支點與其他樣品間的平均距離，再找出其中的最小兩個樣品進行聚在一起，反覆進行直到所有樣品都聚集在一起形成一完整的樹型關係圖。然而，要進行 UPGMA 建構時，第一步驟需要建構遺傳距離矩陣，因此，請依據下面題目設定，使用 MEGA-11 程式進行下列分析。

問題 6

此 7 物種的序列，總長為 3331 核苷酸，經過檢驗後，發現此序列前 120 個和後 131 個核苷酸有錯誤而需要移除，將序列去除前述有錯誤片段後，此片段的核苷酸鏈合長度佔原總長之百分比例為多少？Conserved 核苷酸位點百分比例為多少？變異(Variable)核苷酸位點百分比例為多少？Parsimony information 核苷酸位點百分比例為多少？Singleton 核苷酸位點百分比例為多少？(提示: 新的序列需另存至桌面考號的資料夾，檔名「考號-mtCDNA SEQ-1.MEG 及考號-mtCDNA SEQ-1.FAS」。)(各 2 分，共 10 分)

項目	答案
核苷酸鏈合長度佔原總長百分比例:	
Conserved(C) 核苷酸位點百分比例:	
變異(Variable)(V)核苷酸位點百分比例:	
Parsimony information(P) 核苷酸位點百分比例:	
Singleton(S) 核苷酸位點百分比例:	

問題 7

獲得此 7 物種的新的序列資料，使用 MEGA 11 來建構遺傳距離矩陣，設定使用 p-distance 取代模型，位點之間的速率設定為 Gamma Distributed，計算出兩兩物種間的遺傳距離，填入下表中。(各 1 分，共 21 分)

	homo	chimpanzee	bonobo	gorilla	orangutan	sumatran	gibbon
homo							
chimpanzee							
bonobo							
gorilla							
orangutan							
sumatran							
gibbon							

問題 8

獲得此 7 物種的遺傳距離矩陣，以算術平均值不加權配對法進行聚類分析來建構樹型關係圖，由上題資料找出哪二種物種的遺傳距離最近，填入下列表格？(各 2 分，共 4 分)

問題 9

依據算術平均值不加權配對法開始進行聚類分析與重新建構遺傳距離矩陣方法，承上題找出最近遺傳距離的二種物種，進行合併與重建新遺傳距離矩陣，將計算結果填入下方矩陣中，包含物種資訊和遺傳距離資訊(畫斜線格子 不需填入答案)。(各 1 分，共 21 分)

	1:	2:	3:	4:	5:	6:
1:						
2:						
3:						
4:						
5:						
6:						

問題 10

以此 7 物種的新的 DNA 序列資料，分別以 MEGA 11 程式的鄰位連接法(Neighbor-joining method, NJ)和最大似然度法(Maximum Likelihood method, ML)方式進行樹狀圖建構分析，鄰位連接法設定使用 p-distance 取代模型，位點之間的速率設定為 Gamma Distributed，最大似然度法則需要依據核苷酸取代模型分析獲得的 BIC (Bayesian Information Criterion)分數結果，選取出最符合的核苷酸取代模型(nucleotide substitution models)進行設定分析。以 Bootstrap 方法重覆分析，NJ 法進行 1000 次和 ML 法進行 100 次以獲得各分支的 Bootstrap 值，建構出 NJ 和 ML 樹狀圖，將含有 Bootstrap 值得 NJ 和 ML 樹狀圖手繪於下方，並比較與說明二圖的差異。(NJ 樹 3 分、ML 樹 3 分、比較說明 4 分)