Bioinformatics 生物資訊學

34th International Biology Olympiad

第 34 屆國際生物奧林匹亞競賽

3-10 July 2023, United Arab Emirates University

2023 年 7 月 3-10 日，阿聯酋大學

Practical Exam

實作考試

Total points: 98

總分：98

Duration: 90 minutes

時長：90 分鐘

General Instructions:

一般說明：

You have 90 minutes to complete **TWO tasks in this practical exam**.

您有 90 分鐘的時間完成本次實作考試中的兩項任務。

You can do the tasks in any order.

您可以按任意順序執行任務。

Task 1: Molecular phylogenetics (61 points)

任務 1：分子系統發育學（61 分）

Task 2: Genome editing (37 points)

任務 2：基因組編輯（37 分）

Important Information:

重要信息：

No answers on the exam paper will be graded.

試卷上的答案不會被評分。

You will use a dedicated browser application both to perform the bioinformatics analyses and to enter your responses. This application includes four different types of screens, namely question screen, "Notepad", "Input Sequences" and "Tools".

您將使用專用的瀏覽器應用程序來執行生物資訊學分析並輸入您的回覆。該應用程序包括四種不同類型的螢幕，即問題螢幕、“記事本”、“輸入序列” 和 “工具”。

You do not need to submit the exam: it will be submitted automatically, once the exam is over. There will be a timer showing the time remaining until the end of the exam.

您無需提交考試：考試結束後，考試將自動提交。會有一個計時器顯示考試結束前的剩餘時間。

You **MUST PRESS “Save”** after completing EACH question or else your answers will not be registered.

在完成每個問題後，您必須按“保存”，否則您的答案將不會被註冊。

You **MUST NOT** quit the Safe Exam Browser.

您**不得**退出安全考試瀏覽器。

Questions which are fully or at least partially answered and saved are highlighted in green in the side panel. 已全部或至少部分回答並保存的問題在側面板中以綠色突出顯示。

"Reset" button will erase answers provided for the currently selected question. It will never erase answers provided for other questions. Using this button, you can resubmit your answers as many times as you want.「Reset」按鈕將清除目前選定問題的答案，但不會清除其他問題的答案。使用這個按鈕，您可以隨時重新提交答案。

You can search specific text strings using "Ctrl" + "F", but you must first click on the text field (the window within the window) to execute the search against the sequence. 您可以使用「Ctrl」+「F」來搜索特定的文字串，但必須先點擊文字欄位（視窗內的視窗）來執行對序列的搜索。

You can open multiple instances of the same tool or different tools and have several windows side-by-side (for instance, to compare results of different analyses). 您可以打開多個相同或不同的工具實例，並將多個視窗並排擺放（例如，比較不同分析結果）。

You can switch between Windows using "Alt" + "Tab".

您可以使用「Alt」+「Tab」在視窗之間切換。

If you minimise the electronic copy of the questions, you can quickly return to it by clicking the icon on the bottom left of the screen. 如果您最小化了問題的電子副本，您可以通過點擊螢幕左下角的圖標快速返回。

To ensure proper formatting of the obtained analysis output you can maximize the window.

為了確保所獲得的分析輸出的正確格式，您可以最大化視窗。

You do not need to modify the obtained analysis output after you have pasted it into the answer box.

將分析輸出粘貼到答案框後，您不需要修改它。

If you face any technical issues with your computer, raise your card.

如果您的電腦遇到任何技術問題，請舉卡。

If you get an error message after pressing “Submit” for any tool, this is because of your input being incorrect, so such queries will not be dealt with by the assistants.

如果在按下「Submit」後出現錯誤消息，那是因為您的輸入有誤，因此助手們無法處理此類問題。

Use the following cards to ask for water/washroom/help.

使用以下卡片請求水/洗手間/幫助。

Drinking water 飲用水	Washroom 衛生間	Other queries 其他詢問
💧	🚻	?

No paper, materials or equipment should be taken out of the laboratory.

任何紙張、材料或設備不得帶出實驗室。

Good luck!

祝您好運！

TASK 1. MOLECULAR PHYLOGENY 任務一 分子親緣關係

One of the great milestones in molecular phylogenetics was the work by Carl Woese (1977). He compared the sequence of the small ribosomal subunit rRNA (16S/18S rRNA), between different species, by digesting it with T1 RNase followed by hybridization of conservative loops.

Carl Woese (1977) 的工作是分子系統發生學的偉大里程碑之一。他通過使用 T1 RNase 消化小核醣體亞基 rRNA (16S/18S rRNA)，然後進行保守環間的雜合，並比較了不同物種之間的小核醣體亞基 rRNA (16S/18S rRNA) 的序列。

In this task, you will also use 16S/18S rRNA sequences, of 10 different organisms, to build a phylogenetic tree. To do that, you will first need to make an alignment of all corresponding sequences, and then use that alignment as input for the tree-building algorithm.

在此任務中，您將使用 10 種不同生物的 16S/18S rRNA 序列來構建系統發育樹。為此，您首先需要對所有相應序列進行比對，然後使用該比對作為演化樹推測的數據來源。

The input data for this task can be found by clicking on the “Input Sequences” dropdown menu and choosing the input labeled as “rRNA.fasta”. This file contains the sequences of both the 16S/18S and the 23S/28S rRNA for each species. The labels have either “16S” or “23S” ending, respectively. Hence, you should prepare a "*fasta*" formatted input which contains only the sequences of 16S/18S rRNA. For this you can copy-paste the provided data into the Notepad page of the Application and edit it appropriately.

透過單擊“輸入序列”下拉選單並選擇標記為“rRNA.fasta”的數據輸入，可以找到此任務的輸入數據。該文件包含每個物種的 16S/18S 和 23S/28S rRNA 序列。標籤分別以“16S”或“23S”結尾。因此，您應該準備一個“fasta”格式的輸入，其中僅包含 16S/18S rRNA 的序列。為此，您可以將提供的數據複製粘貼到應用程式的記事本頁面中並進行適當的編輯。

Notes: a) each sequence should have a header starting with “>”; b) you can enter empty lines between sequences in the "*fasta*" format to keep them visually separated. For more about the "*fasta*" format see Appendix 1.

注意： a) 每個序列應該有一個以“">”開頭的標題；b) 您可以在“fasta”格式的序列之間輸入空行，以使它們在視覺上分開。有關“fasta”格式的更多資訊，請參閱附錄 1。

Now align the 16S/18S sequences of each species using the “7. Sequence Alignment” tool. It will return an alignment in the "*clustal*" format (for more about it see Appendix 1).

現在請你使用"7. Sequence Aligment" 工具來對齊每一個物種的 16S/18S 序列。這將使你的序列變成一個"clustal" 格式的檔案 (請見附錄 1 了解更多)。

Q1.1 True or False? 正確或錯誤？

Q1.1.1 The alignment provides evidence for insertions or deletions happening during the evolution of 16S/18S rRNA genes. 1.0pt
序列比對為 16S/18S rRNA 基因演化過程中所發生的插入或缺失事件提供了證據。

Q1.1.2 *C. paramecium* 16S/23S rRNA is the shortest sequence in the alignment. 1.0pt
草履蟲 16S/23S rRNA 是比對中最短的序列。

Q1.2 Copy your resulting alignment and paste it into the input field of question Q1.2 for the results validation. 5.0pt
複製生成的序列對齊結果並將其貼到問題 Q1.2 的輸入字段中以進行結果驗證。

Copy your resulting alignment and paste it into the input field of the “9. Tree Builder” tool. Press “Submit” for tree building.

把你做出來的對齊序列複製並貼到"9.TreeBuilder" 這個工具的輸入區。按下" 繳交" 來重建演化樹。

When the analysis is finished, you will get the resulting tree in the output field in *"newick"* format (for more about the *"newick"* format see Appendix 1), as well as an image.

分析完成後，您將在輸出字段中獲得“newick”格式的演化樹（有關“newick”格式的更多信息，請參閱附錄 1）以及圖像。

Note: the tree is unrooted!

注意：樹是無根的！

Q1.3	Copy the "*newick*" tree you have obtained from the output of the “9. Tree Builder” tool and paste it into the input field of question Q1.3. 複製你從"9.Tree Builder" 工具的輸出區所獲得的“newick”樹，並將其貼到問題 Q1.3 的輸入區中。	5.0pt

Based on a tree similar to the one you just obtained, Woese proposed that all cellular life is divided into three domains: Bacteria, Archaea and Eukaryotes.

基於與您剛剛獲得的樹相似的樹。沃斯提出，所有細胞生命都分為三個域：細菌域、古菌域和真核生物域。

Q1.4	For each of the species in Q1.4, select the domain of life it belongs to, using letters "**A**" for Archaea, "**B"** for Bacteria, "**E"** for Eukaryotes. 對於 Q1.4 中的每個物種，選擇其所屬的域，使用字母“A”代表古菌，“B”代表細菌，“E”代表真核生物。	6.0pt

Q1.5 What can you conclude about the *Z. mays* (corn, *Zea mays*) sequence you used based on its position in the tree? For each statement indicate if it is True or False.

關於您在樹中的 Z. mays（玉米、Zea mays）序列，您能得出什麼結論？對於每個陳述，請指出它是對還是錯。

Q1.5.1	The sequence could be the sequence of the nuclear 18S rRNA gene. 該序列可能是核 18S rRNA 基因的序列。	1.0pt

Q1.5.2	The sequence could be the sequence of the chloroplast 16S rRNA gene. 該序列可能是葉綠體 16S rRNA 基因的序列。	1.0pt

Q1.5.3	The sequence could be the sequence of the mitochondrial 16S rRNA gene. 該序列可能是粒線體 16S rRNA 基因的序列。	1.0pt

Q1.5.4	The position of *Z. mays* sequence must be incorrect / an artifact. 玉米序列的位置一定是不正確的/人為的誤差。	1.0pt

Q1.6 True or False?

正確或錯誤

Q1.6.1 The tree demonstrates that Archaea is evolutionarily closer to Eukaryotes than to Bacteria. 1.0pt
這個演化樹顯示古菌在演化上比較接近真核生物而非細菌。

Q1.6.2 The tree can be rooted (the position of the Last Universal Common Ancestor of all cellular life on Earth can be identified) by including a viral sequence into the analysis. 1.0pt
這個演化樹可以藉由將病毒序列包含進來而讓演化樹有根（也就是地球上所有細胞形式生命的最後共同祖先的位置就因此可被識別）

Q1.6.3 The Last Universal Common Ancestor of all cellular life on Earth likely had a homolog of the 16S/18S rRNA. 1.0pt
地球上所有細胞形式生命的最後共同祖先可能有一個 16S/18S rRNA 的同源序列。

Q1.6.4 The T1 enzyme used by Woese cleaves single-stranded RNA. 1.0pt
Woese 使用的 T1 酶切割單鏈 RNA。

Q1.6.5 Based on the tree, the species closest to *P. syntrophicum*, is *M. formicicum.* 1.0pt
根據這個演化樹，最接近 P. syntropicum 的物種是 M. formicum。

Q1.6.6 A taxon that includes *H. sapiens* and *S. cerevisiae* but excludes *C. paramecium* is monophyletic (contains **all** the descendants of a given ancestor). 1.0pt
包含智人和釀酒酵母但不包含草履蟲的分類群是一個單系群（包含該共同祖先的所有後代）。

Molecular phylogenetics can also compare species on a finer scale. In this task you will explore the relationships between 5 bacteria species from the *Streptococcus* genus.

分子系統發育學還可以更精細地比較物種。在本任務中，您將探索鏈球菌屬 5 種細菌之間的關係。

For this, you will use amino acid sequences of two proteins: DnaJ and Cas9.

為此，您將使用兩種蛋白質的氨基酸序列：DnaJ 和 Cas9。

- DnaJ is a molecular chaperone involved in protein folding.
- DnaJ 是參與蛋白質折疊的伴護蛋白。
- Cas9 is an endonuclease which is a part of the CRISPR-Cas9 system.
- Cas9 是一種核酸內切酶，是 CRISPR-Cas9 系統的一部分。

You have two "*fasta*" inputs, for all five *Streptococcus* species :

對於所有五種鏈球菌物種，您有兩個“fasta”輸入：

- Amino acid sequences of the DnaJ protein labeled “DnaJ-protein.fasta”
- 標記為“DnaJ- Protein.fasta”是 DnaJ 蛋白的氨基酸序列
- DNA sequences of the Cas9 locus labeled “Cas9-locus.fasta”.
- 標記為“Cas9-locus.fasta”是 Cas9 基因座的 DNA 序列。

For the Cas9 locus, you need to find the Cas9 open reading frames (ORF; the part of a sequence which can be translated) and the corresponding amino acid sequences using the “3. ORF Finder” Tool. To do this, copy the "*fasta*" sequence of a single Cas9 locus, together with the species name label from the "Input Sequences" menu, and paste it into “3. ORF Finder”. Specify the following parameters for an ORF search by entering specific numbers into the corresponding fields:

對於 Cas9 基因座，您需要使用“3. ORF Finder”工具找到 Cas9 開放性閱讀框（ORF；可轉譯的序列部分）和相應的氨基酸序列。為此，複製單個 Cas9 基因座的“fasta”序列以及“輸入序列”選單中的物種名稱標籤，並將其粘貼到“3. ORF Finder”中。請透過在相應欄位中輸入特定數字來指定以下執行 OR 搜索所需要的參數：

- **Parameter 1:** The minimum ORF length (in amino acid residues). Note that Cas9 proteins in all given species are longer than 1000 amino acid residues.
- 參數 1：最小 ORF 長度（以氨基酸殘基計）。請注意，所有 4 個給定物種中的 Cas9 蛋白都超過 1000 個氨基酸殘基。
- **Parameter 2:** The index of genetic code type which you would like to use to translate RNA to protein. The table below gives the available indices.
- 參數 2：您想要用於將 RNA 轉譯為蛋白質的遺傳密碼類型索引。下表列出了可用的索引。

Genetic code type 遺傳密碼類型	Genetic code index 遺傳密碼指標
The Standard Code 標準編碼	1
The Vertebrate Mitochondrial Code 脊椎動物粒線體密碼	2
The Yeast Mitochondrial Code 酵母菌粒線體密碼	3
The Invertebrate Mitochondrial Code 無脊椎動物粒線體密碼	5

The output of “3. ORF Finder” will be a table with 3 columns, showing ORF length, coding nucleotide sequence and encoded amino acid sequences for each Cas9 ORF.

"3. ORF Finder”輸出的檔案將是一個有 3 列的表格，顯示每個 Cas9 ORF 的 ORF 長度、編碼核苷酸序列和編碼氨基酸序列。

You will have to run “3. ORF Finder” five times; once for each species.

您必須執行“3. ORF Finder”五次；每個物種一次。

For each input species, you should copy the DNA and the protein sequences of the correct ORF and paste it into the "Notepad". Each sequence should have a "fasta" header identical to the header in the original input (“>>S.anginosus-Cas9” and so on; the corresponding DNA and protein sequences should have the same name).

對於每個輸入物種，您應該複製正確 ORF 的 DNA 和蛋白質序列並將其貼到“記事本”中。每個序列應具有與原始輸入中的起始序列相同的“fasta”起始序列（“>>S.anginosus-Cas9”等；相應的 DNA 與蛋白質序列應該有同樣的名稱）。

Next, rearrange the sequences in the Notepad so that you have grouped the five DNA sequences first, and then the five protein sequences. Keep both the amino acid and the coding DNA sequences in the "Notepad" –you will need them later.

接下來，在記事本中重新排列序列，以便首先將五個 DNA 序列分群，然後是五個蛋白質序列。將氨基酸和編碼 DNA 序列保留在“記事本”- 稍後您將需要它們。

Q1.7	What type of genetic code did you use? Choose the corresponding index number in Q1.7. 您使用什麼類型的遺傳密碼？在 Q1.7 中選擇相應的索引數字。	1.0pt

Q1.8	Paste the five coding DNA sequences into the input field of Q1.8 in the "*fasta*" format. The stop codon should NOT be included. 將五個編碼 DNA 序列貼到“fasta”中 Q1.8 的輸入欄位中格式。不包含終止密碼子。	5.0pt

Q1.9	When searching for ORFs, “3. ORF Finder” checks all possible reading frames (possible ways to divide the provided DNA sequence into codons) on both strands. How many reading frames in total does it check? Enter your answer as a number into the input field of Q1.9. 當搜索 ORF 時，“3. ORF Finder” 會檢查雙股上所有可能的閱讀框（將提供的 DNA 序列劃分為密碼子的可能方式）。它總共檢查了多少個閱讀框？將您的答案以數字輸入到 Q1.9 的輸入欄位中。	1.0pt

Follow the same workflow as for the rRNA sequences, to build trees based on the amino acid sequences of DnaJ (provided for you in the “Input Sequences” menu) and Cas9 (you have obtained these in the previous task).

按照與 rRNA 序列相同的工作流程，根據 DnaJ（在“輸入序列”功能表中為您提供）和 Cas9（您已在上一個任務中獲得這些）的氨基酸序列重建演化樹。

Q1.10	Paste the DnaJ protein alignment in "*clustal*" format into the input field of Q1.10. 將“clustal”格式的 DnaJ 蛋白比對粘貼到 Q1.10 的輸入欄位中。	1.0pt

Q1.11	Paste the Cas9 protein alignment in "*clustal*" format into the input field of Q1.11. 將“clustal”格式的 Cas9 蛋白比對粘貼到 Q1.11 的輸入欄位中。	1.0pt

Also save the Cas9 protein alignment in your Notepad as you will need it later on.

將 Cas9 蛋白比對保存在記事本中，因為稍後您將需要它。

Q1.12	Look at the two alignments. Which of the two is expected to have fewer errors (non-homologous amino acid positions identified as homologous)? Choose DnaJ or Cas9 in Q1.12. 看一下兩條序列對齊的結果。你預期兩者中哪一個的錯誤較少（非同源氨基酸位置被識別為同源）？在 Q1.12 中選擇 DnaJ 或 Cas9。	1.0pt

Q1.13	Paste the DnaJ tree in "*newick*" format into the input field of Q1.13. 將“newick”格式的 DnaJ 樹粘貼到 Q1.13 的輸入欄位中	2.5pt

Q1.14	Paste the Cas9 tree in "*newick*" format into the input field of Q1.14. 將“newick”格式的 Cas9 樹粘貼到 Q1.14 的輸入欄位中。	2.5pt

Q1.15 True or False?

正確或錯誤？

Q1.15.1	The Cas9 protein is more evolutionarily conserved, compared to DnaJ. 與 DnaJ 相比，Cas9 蛋白在演化上更加保守。	1.0pt
Q1.15.2	The Cas9 tree is more likely to show the genome-average phylogenetic relationships between these species, compared to the DnaJ tree. 與 DnaJ 樹相比，Cas9 樹更有可能接近以這些物種的全基因體所重建的演化關係。	1.0pt
Q1.15.3	The difference between the trees for the two genes could be a result of horizontal gene transfer. 兩個基因樹之間的差異可能是基因水平轉移的結果。	1.0pt
Q1.15.4	Both Cas9 and DnaJ trees include a branch with *S. anginosus* and *S. mitis* as closest neighbors. Cas9 和 DnaJ 樹都包含一個分支，其中最近的物種是 S. anginosus 和 S. mitis。	1.0pt
Q1.15.5	If we did the analysis using some other gene from the same five species, we could get a tree with a different topology. 如果我們使用來自相同五個物種的其他基因進行分析，我們可以獲得具有不同樹型的樹。	1.0pt

Using bioinformatics tools, we can also explore selective pressure on DNA and protein sequences. One method is the dN/dS test (also known as the Ka/Ks test). Here, negative selection is defined as amino acid substitutions being on average deleterious, while positive selection is a scenario of amino acid substitutions being on average beneficial.

使用生物資訊學工具，我們還可以探索 DNA 和蛋白質序列的天擇壓力。一種方法是 dN/dS 測試 (也稱為 Ka/Ks 測試)。在這裡，負向天擇被定義為氨基酸取代一般來說是有害的，而正向天擇是指氨基酸取代一般來說是有益的情況。

The test compares the observed rates of non-synonymous (N; leading to a change in amino acid) and synonymous (S; not leading to a change in amino acid) nucleotide changes in a protein-coding DNA sequence, calculating their ratio (dN/dS). The rate of each type of substitutions (dN or dS) is defined as the number of observed substitutions of this type normalized (divided) by the total number of possible substitutions of this type. The table below shows the number of possible non-synonymous (N) and synonymous (S) substitutions for some codons.

該測試比較蛋白質編碼 DNA 序列中觀察到的非同義（N；導致氨基酸變化）和同義（S；不導致氨基酸變化）核苷酸變化的比率，計算它們的比率 (dN /dS)。每種類型的取代率（dN 或 dS）定義為可被觀察到的該類型的取代數被標準化 (除以) 該類型的可能取代總數。下表顯示了某些密碼子可能的非同義 (N) 和同義 (S) 替換的數量。

Codon 密碼子	Number of possible non-synonymous (N) substitutions 可能的非同義取代數量	Number of possible synonymous (S) substitutions 可能出現的同義取代數量
ATG	9	0
AAA	8	1
AGA	7	2
ACA	6	3

The dN/dS test can be applied to either a pair of sequences, or to a tree (in the latter case, dN/dS ratio can be calculated for each branch).

dN/dS 測試可以應用於一對序列，也可以應用於一棵樹（在後一種情況下，可以計算每個分支的 dN/dS 比率）。

Q1.16 True or False?

正確或錯誤？

Q1.16.1 The maximum number of possible single nucleotide synonymous changes for a codon in the standard genetic code is 4.
標準遺傳密碼中密碼子可能發生的單一核苷酸之同義變化的數量不會超過 4 個。 1.0pt

Q1.16.2 The described normalization is needed because the relative probability of N and S substitutions depend on the original sequence.
上述的標準化流程是有必要的，因為 N 和 S 替換的相對機率取決於初始序列。 1.0pt

Q1.16.3 This test assumes that synonymous substitutions are selectively neutral.
該測試假設核酸同義取代在天擇上是中性的。 1.0pt

Q1.16.4 One can directly compare dN/dS values between genes with moderately different total mutation rates.
人們可以直接比較總突變率略有不同的基因之間的 dN/dS 值。 1.0pt

Q1.17 What is the expected dN/dS value for a sequence that evolves completely neutrally (even non-synonymous substitutions are neither deleterious nor beneficial)? Enter your answer as a whole number into the answer field of Q1.17.
完全中性演化的序列的預期 dN/dS 值是多少（即使非同義取代既無害也無益）？將您的答案的整數輸入到 Q 1.17 的答案欄位中。 1.0pt

Q1.18 For each DNA sequence described below, choose if the result of the dN/dS test is expected to be equal, smaller or bigger than X, where X is the dN/dS value for a sequence evolving completely neutrally.

對於下面描述的每個 DNA 序列，選擇 dN/dS 測試的結果是否預期等於、小於或大於 X，其中 X 是完全中性演化的序列的 dN/dS 值。

Q1.18.1	Coding sequence of histone H4. 組蛋白 H4 的編碼序列。	1.0pt

Q1.18.2	Coding sequence of a viral surface protein that can be recognized by the host’s immune defense. 可以被宿主免疫防禦所辨識的病毒表面蛋白質編碼基因序列。	1.0pt

Q1.18.3	Coding sequence of a pseudogene (a gene which no longer produces a functional product). 假基因（不再產生功能性產物的基因）的編碼序列。	1.0pt

Now you should apply the dN/dS test to explore evolution of the Cas9 gene in the *Streptococcus* genus. The first step is to align the coding sequences by codons using the “1. Codon Alignment”tool.

現在您應該應用 dN/dS 測試來探索鏈球菌屬中 Cas9 基因的演化。第一步是使用“1. 密碼子比對”工具按密碼子比對編碼序列。

Copy the Cas9 protein alignment for the 5 species from your "Notepad" and paste it into the first field labeled “Format: *clustal*”.

從“記事本”中複製 5 個物種的 Cas9 蛋白比對，並將其粘貼到標有“Format：clustal”的第一個欄位中。

Next, copy the Cas9 coding DNA sequences from your Notepad (all five together including the *"fasta"* headers) and paste them into the second field labeled “Format: *fasta*”, and press “Submit”.

接下來，從記事本複製 Cas9 編碼 DNA 序列（所有五個序列一起，包括“fasta”標題）並將其粘貼到標有“Format：fasta”的第二個欄位中，然後按“繳交”。

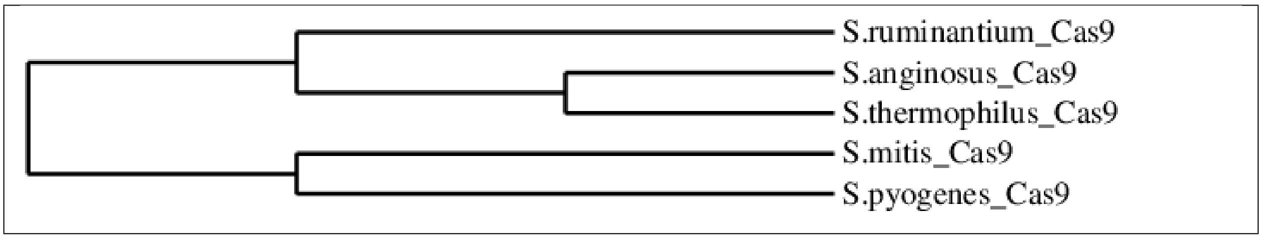
Q1.19	Paste the resulting Cas9 codon alignment data into the input field of Q1.19 將生成的 Cas9 密碼子比對結果貼到 Q1.19 的輸入欄位中	1.0pt

Now, the codon alignment can be used as an input for the dN/dS test. However, because running it for each branch of the tree would take too long, the results are already provided to you below.

現在，密碼子比對可用作 dN/dS 測試的輸入。但是，由於對樹的每個分支都執行一輪會花費太長時間，因此結果已在下面提供給您。

The cladogram below shows the clustering of the the five Cas9 sequences. Branch length conveys no information here. The table under the cladogram tabulates the dN/dS ratios for each branch of the cladogram.

下面的樹形圖顯示了 5 個 Cas9 序列的歸群結果。分支長度此處不代表任何資訊。樹形圖下方的表格列出了樹形圖中每個分支的 dN/dS 比率。



Branch leading to **樹枝連接至**	dN/dS ratio
S.ruminantium-Cas9	0.002
S.thermophilus-Cas9	0.028
S.anginosus-Cas9	0.011
S.pyogenes-Cas9	0.002
S.mitis-Cas9	0.022
(*S.pyogenes*-Cas9,*S.mitis*-Cas9)	0.112
(*S.thermophilus*-Cas9,*S.anginosu*s-Cas9)	0.009
((*S.thermophilus*-Cas9,*S.anginosus*-Cas9), *S.ruminantium*-Cas9)	79.086

Q1.20

Analyze the results of the dN/dS test of the Cas9 gene in the *Streptococcus* species.

分析鏈球菌屬 Cas9 基因的 dN/dS 測試結果。

True or False?

正確或錯誤

Q1.20.1	It can be concluded that Cas9 was under constant negative selection during its evolution in *Streptococcus*. 我們可以得出結論，Cas9 在鏈球菌中的進化過程中一直處於負向天擇之下。	1.0pt

Q1.20.2	During the evolution of Cas9, since the common ancestor of *S.thermophilus* and *S.anginosus* to modern-day *S.thermophilus* and *S.anginosus*, most of the non-synonymous substitutions were deleterious. 在 Cas9 的演化過程中，也就是起自 S. thermophilus 與 S. anginosus 的共祖至現生 S. thermophilus 與 S. anguinosus 這段過程，大多數的非同義取代是有害的。	1.0pt

Q1.20.3	The divergence in Cas9 amino acid sequence between ((*S.thermophilus, S.anginosus*), *S.ruminantium*) on the one hand, and (*S.mitis*, *S.pyogenes*) on the other, could be explained by positive selection. ((*S.thermophilus, S.anginosus*), *S.ruminantium*) 與 (*S.mitis*, *S.pyogenes*) 這兩個分支群在 Cas9 胺基酸序列上的差異可使用正向天擇來解釋。	1.0pt

Q1.20.4	The difference in Cas9 amino acid sequence between ((*S.thermophilus*, *S.anginosus*),*S.ruminantium*) on one hand, and (*S.mitis*, *S.pyogenes*) on the other, could be explained by a high error (non-homologous amino acid positions identified as homologous) rate in the alignment. ((*S.thermophilus*, *S.anginosus*),*S.ruminantium*) 與 (*S.mitis*, *S.pyogenes*) 這兩個分支群在 Cas9 胺基酸序列上的差異可以藉由序列比對中的高錯誤率 (非同源胺基酸被解讀為同源胺基酸) 來解釋。	1.0pt

END OF TASK 1

任務結束

TASK 2. GENOME EDITING 任務 2. 基因組編輯

Genome editing became a widespread technology thanks to the *Streptococcus pyogenes* Cas9 enzyme. In this approach, mammalian cells are transfected with a vector containing the sequence encoding Cas9 and a sequence encoding gRNA (guide RNA). The mechanism behind Cas9 application is described in Appendix 3.

基因組編輯成為一項廣泛應用的技術，這要歸功於鏈球菌 Cas9 酵素。在這種方法中，哺乳動物細胞被轉染以含有編碼 Cas9 酵素和編碼 gRNA（引導 RNA）序列的質體。Cas9 的應用機制請參見附錄 3。

Natural Cas9 uses two different RNAs: crRNA for DNA target sequence recognition and tracrRNA for proper enzyme assembly and crRNA binding. Both these functions can be combined in an artificial RNA, called gRNA.

自然界中的 Cas9 使用兩種不同的 RNA：crRNA 用於 DNA 目標序列識別，tracrRNA 用於正確的酵素組裝和與 crRNA 結合。這兩個功能可以結合在一起形成一種人工 RNA，稱為 gRNA。

The guide sequence consists of the target sequence (usually 20 nucleotides) and should be followed by the PAM sequence (protospacer adjacent motif; 5'-NGG-3', where N is any base). Therefore, only sequences with a PAM sequence at their 3' end can be targeted by the gRNA.

引導序列由目標序列（通常為 20 個核苷酸）組成，並且應該在其後緊接著 PAM 序列（原始間隔基本保守序列; 5'-NGG-3'，其中 N 代表任意鹼基）。因此，只有在其 3' 端具有 PAM 序列的序列才能被 gRNA 所做為目標。

In this task your ultimate goal is to design an experiment to create a mouse model of the human sickle-cell anemia disease, by editing the *Hbb* gene in mice. Specifically, you need to replace a part of the first exon of the mouse *Hbb* gene with the corresponding human sequence carrying the disease-causing mutation.

在這個任務中，您的最終目標是設計一個實驗，通過編輯小鼠的 Hbb 基因，創建一個人類鐮刀型貧血症的小鼠模型。具體來說，您需要將小鼠 Hbb 基因的第一外顯子的一部分，用攜帶致病突變的相應人類序列進行替換。

Initially, you need to design primers to amplify and clone the complete Cas9 coding sequence into a mammalian expression vector.

首先，您需要設計引子來擴增與選殖完整的 Cas9 編碼序列到哺乳動物表達質體中。

The primers should meet the following requirements:

引子應該滿足以下要求：

- Allow to amplify the complete *S. pyogenes* Cas9 coding sequence.
- 能夠擴增完整的 *S. pyogenes* Cas9 編碼序列。

- The part of the primer annealing to the Cas9 sequence should be 18-25 bases long.
- 引子與 Cas9 序列結合的部分應該長 18-25 個鹼基。
- The annealing part of both primers should have a melting temperature between 47-54°C.
- 兩引子的黏合部分的熔解溫度應在 47-54°C 之間。

- Have a G or a C on the 3’ end.
- 引子的 3' 端應帶有 G 或 C

- Have non-annealing 5’overhangs containing restriction sites for the restriction enzymes you plan to use for cloning. To ensure proper orientation, the forward and the reverse primer should be cut by different enzymes. **Note:** The multiple cloning site (MCS) of the expression sequence is described in Appendix 4.
- 請確保非黏合的 5' 突出端含有您計劃在選殖中使用的限制酶的限制位點。為了確保正確的定向性，正向引子和反向引子應該由不同的限制酶切割。註：表達序列的多重克隆位點（MCS）詳見附錄 4。
- Restriction sites included in the overhangs should be located right next to the annealing part of the primers, with no additional nucleotides in between.
- 突出端中包含的限制酶切點應該緊鄰引子的結合部分，中間不應有額外核苷酸。
- Contain 5 additional nucleotides on the 5’ end to improve activity of the restriction enzymes by ensuring proper binding to the DNA
- 引子 5' 端包含 5 個額外的核苷酸，以提高限制酶對 DNA 的結合能力。

- Allow cloning of the Cas9-coding sequence in the same frame as the FLAG-coding sequence present in the vector (i.e. a fusion protein of Cas9 and FLAG will be translated).
- 允許在質體中的 FLAG 編碼序列與 Cas9 編碼序列能在相同的框架中選殖（即 Cas9 和 FLAG 的融合蛋白將被轉錄）。

You should start with designing the annealing parts of the primers.

您應該從設計引子的黏合部分開始。

To do this for the forward primer you need to:

對於為正向引子進行設計，您需要以下步驟：

1. Open the“S.pyogenes-Cas9-cds”input which contains the sequence of the *S. pyogenes* Cas9 coding sequence and copy it. 打開包含 S. pyogenes Cas9 編碼序列的"S.pyogenes-Cas9-cds" 輸入，並將其複製。
2. Open the“6. Sequence Editor”tool and paste the sequence into the input field of the tool. 打開"6. 序列編輯器" 工具，並將序列粘貼到該工具的輸入框中。
3. Select a sub-sequence from the 5’ end of the Cas9 sequence. You can see the length of the selected subsequence in the “6. Sequence Editor”tool. Copy the selected sequence if it meets the requirements. 從 Cas9 序列的 5' 端選擇一個子序列。您可以在"6. 序列編輯器" 工具中看到所選子序列的長度。如果所選序列滿足要求，請將其複製。
4. Check its melting temperature using the “6. Tm Calculator”tool. 使用"6. Tm 計算器" 工具檢查其熔解溫度。

5. If all criteria for the annealing part are met, you can paste it into your Notepad. 如果黏合部分的所有標準都符合，您可以將其粘貼到記事本中。

The general procedure for the reverse primer is similar. However, using additional tools (see Appendix 2) might be needed. Note that both forward and reverse primer sequences should be written in 5' – 3' orientation.

反向引子的一般步驟類似。但是，可能需要使用其他工具（參見附錄 2）。請注意，正向和反向引子序列都應以 5' - 3' 的方向書寫。

Q.2.1

Enter the sequences of the annealing parts of the primers:

請輸入引子黏合部分的序列：

Q2.1.1 Forward Primer annealing part, in 5' to 3' orientation 4.0pt
正向引子黏合部分的序列（以 5' 到 3' 的方向）：

Q2.1.2 Reverse Primer annealing part, in 5' to 3' orientation 4.0pt
反向引子黏合部分的序列（以 5' 到 3' 的方向）：

The next step is designing the overhangs. For this, you first need to choose the restriction enzymes you want to use.

接下來的步驟是設計突出端。為此，您首先需要選擇要使用的限制酶。

Use Appendix 4 to find which enzymes can cut the ‘multiple cloning site’ of the vector. **Note**: the same sequence can be found in the "Input Sequences" menu labeled "MCS".

請使用附錄 4 查找可以切割質體中“多重選殖位點（multiple cloning site）”的酶。請注意：相同的序列可以在標有“MCS”的“輸入序列”選單中找到。

You should also investigate which enzymes may cut the Cas9-coding sequence. Copy the corresponding sequence and paste it into the input field of the “4. Restriction Mapper” tool. It will return the number of cut sites for commonly used restriction enzymes.

您還應該探討可能會切割 Cas9 編碼序列的酶。複製相應的序列，然後將其粘貼到“4. 限制酶圖譜”工具的輸入框中。該工具將輸出常用限制酶的切割位點數目。

Q.2.2

Based on the results you obtained choose the restriction enzymes that you will use for cloning.

根據您獲得的結果，選擇您將用於選殖的限制酶。

Q2.2.1 Forward Primer: 1.0pt
正向引子：

Q2.2.2 Reverse Primer: 1.0pt
反向引子：

Q2.3

Determine the reason why the following restriction enzyme **pairs** are inappropriate for cloning the Cas9-coding sequence into the mammalian expression vector.

確定以下限制酶組合於將 Cas9 編碼序列選殖到哺乳動物表達質體的不適合原因。

Q2.3.1 PmeI and XhoI 1.0pt
PmeI 和 XhoI

Q2.3.2 BamHI and HindIII 1.0pt
BamHI 和 HindIII

Reasons (you may pick more than one for each enzyme pair):

原因（您可以選擇每組酶組合的數個原因）：

A. At least one of the enzymes will cut *Cas9* gene inside the coding sequence.

A. 至少有一個酶會在編碼序列內切割 Cas9 基因。

B. At least one of the enzymes will cut the MCS twice.

B. 至少有一個酶會對多重選殖位點進行兩次切割。

C. The FLAG-coding sequence will be lost.

C. FLAG 編碼序列將丟失。

D. The FLAG-coding sequence will not be in-frame with the Cas9-coding sequence.

D. FLAG 編碼序列將無法與 Cas9 編碼序列保持在同一閱讀框架。

Now you have to design the primer overhangs to meet all the requirements listed previously.

現在您需要設計引子的突出部分，以滿足先前列出的所有要求。

Q2.4

Enter the primer overhang sequences in the fields below:

請在下方的欄位中輸入引子的突出序列：

Q2.4.1 Forward Primer overhang, in 5' to 3' orientation 4.0pt
正向引子的突出序列，以 5' 至 3' 方向為：

Q2.4.2 Reverse Primer overhang, in 5' to 3' orientation 5.0pt
反向引子的突出序列，以 5' 至 3' 方向為：

Q2.5

Indicate if the following statements regarding the cloning procedure and mammalian expression vector is **True or False**.

請指出以下有關選殖程序和哺乳動物表達質體的陳述是否正確 **(True)** 或是錯誤 **(False)**。

Q2.5.1 Cloning the Cas9 gene ($>$4000 bp) requires using Pfu polymerase (error rate 1.3x10^{-6}) instead of Taq polymerase (error rate 1.8x10^{-4}). 1.0pt
選殖 Cas9 基因（>4000 個鹼基對）所需使用 Pfu 聚合酶（錯誤率為 1.3x10-6）而不是 Taq 聚合酶（錯誤率為 1.8x10-4）。

Q2.5.2 Codon optimization (replacing naturally rare codons with more common codons encoding the same amino acids) may increase the expression level of Cas9. 1.0pt
密碼子優化（將自然稀有的密碼子替換為編碼相同胺基酸的常見密碼子）可能會增加 Cas9 的表達水平。

Q2.5.3 The vector should contain an antibiotic-resistance gene for the selection of bacterial colonies. 1.0pt
質體應該包含抗生素抗性基因，用於選擇細菌菌落。

Q2.5.4 Bacterial RNA polymerase can start transcription from the transcription start site adjacent to the multiple cloning site. 1.0pt
細菌 RNA 聚合酶可以從靠近多重選殖位點的轉錄起始位點開始轉錄。

In humans, sickle-cell anemia is caused by a single nucleotide mutation in one of the beta-globin genes, which results in the substitution of glutamate for valine at 6th amino acid position (E6V substitution; GAG codon changing to GTG).

在人類中，鐮刀型貧血症是由 β-球蛋白基因之一的單核苷酸突變引起的，這導致第 6 個氨基酸位置發生麩氨酸到纈氨酸的取代（E6V 取代；GAG 密碼子變為 GTG）。

In this task you need to design two nucleotide sequences:

在這個任務中，您需要設計兩個核苷酸序列：

1. A gRNA which will guide the Cas9 enzyme towards the proper mouse *Hbb* gene.

   一個 gRNA，它將引導 Cas9 酶定位到正確的小鼠 Hbb 基因。

2. A Homology Directed Repair (HDR) template (see Appendix 3) which will allow introducing the appropriate mutation. 一個同源定向修復（HDR）模板（見附錄 3），它將允許引入適當的突變。

The table below contains annotations for the GenBank file (accession number: X14061.1) (labeled “*M.musculus*-Hbb-complex” in the “Input Sequences” menu), containing the **mouse** genomic DNA sequence of the beta-globin complex. This sequence is 55856 bp, and includes several genes and pseudogenes (genes which do not produce functional products).

下表包含 GenBank 文件的注釋（存取號：X14061.1）（在“輸入序列”選單中標記為“M.musculus-Hbb-complex”的注釋），其中包含小鼠 β-球蛋白複合體的基因組 DNA 序列。該序列為 55856 bp，包含幾個基因和偽基因（不產生功能性產物的基因）。

Note: the coordinates of the exons are inclusive on both ends (a region defined as 1..2 spans 2 bp).

注意：外顯子的座標兩端都包含在內（定義為 1..2 的一個區域涵蓋 2 個核苷酸）。

The wild-type coding sequence of this **human** beta-globin is available in the “Input Sequences”menu labeled “*H.sapiens*-Hbb-cds”.

此種人類 β-球蛋白的野生型編碼序列可在“輸入序列”選單中標為“H.sapiens-Hbb-cds”

Type 類型	**Name 名稱**	**Expression 表達**	**Coordinates for exons 外顯子的位置**
Gene 基因	*Hbb-y*	**High expression level in early embryo**	12781..12872, 13200..13422, 14274..14402
Gene 基因	*Hbb-bh0*	Low expression level in early embryo	12781..12872, 13200..13422, 14274..14402
Gene 基因	*Hbb-bh1*	High expression level in late embryo	21102..21194, 21301..21522, 22331..22459
Pseudogene 偽基因	*Hbb-bh2*	None 無	23786..23882, 23963..24182, 24905..25030
Pseudogen 偽基因	*Hbb-bh3*	None 無	30435..30525, 30617..30707, 31409..31534
Gene 基因	*Hbb-b1*	High expression level in pups and adult	38339..38430, 38547..38769, 39424..39552
Gene 基因	*Hbb-b2*	Low expression level in pups and adult	53548..53652, 53757..53978, 54607..54735

Q2.6 Select the mouse gene to target with CRISPR-Cas9 which would probably create the best model of human sickle-cell disease. 1.0pt
選擇用 CRISPR-Cas9 進行小鼠基因的定向位基因編輯，可能會創建最佳的人類鐮刀型貧血症模型。

Q2.7

True or False?

正確或錯誤

You can perform additional analyses of these sequences using the available tools, if needed.

假如需要的話，您可以使用合宜的工具進行額外的序列分析。

Q2.7.1 The N-terminal methionine residue is present in the mature human beta-globin chain. 1.0pt
成熟的人類 β-球蛋白鏈中的 N 端有甲硫氨酸殘基。

Q2.7.2 The coding sequence of the target mouse beta-globin chain (chosen in Q2.6) has the same length as the coding sequence of the human beta-globin chain. 1.0pt
所選的目標小鼠 β-球蛋白鏈（在 Q2.6 中選擇）的編碼序列與人類 β-球蛋白鏈的編碼序列長度相同。

Q2.7.3 The sequences of the last 10 amino acid residues in the target mouse beta-globin chain and human beta-globin chain are identical. 1.0pt
目標小鼠 β-球蛋白鏈和人類 β-球蛋白鏈最後 10 個氨基酸殘基的序列是相同的。

Q2.7.4 Both mouse beta-globin pseudogenes result from reverse transcriptase activity. 1.0pt
兩個小鼠 β-球蛋白偽基因都是由反轉錄酶活性產生的。

Q2.7.5 The wild-type sequence of the target mouse beta-globin chain contains a glutamate in the 6th position (the target position) 1.0pt
目標小鼠 β-球蛋白鏈的野生型序列在第 6 個位置（目標位置）包含一個麩氨酸。

Q2.8 Choose the appropriate sequence for part of the gRNA hybridizing to the target sequence (excluding PAM; shown in dark blue in Appendix 3) from those provided in Q2.8. 1.0pt
從 Q2.8 中提供的選項中，選擇與目標序列（不包括 PAM，見附錄 3 中的深藍色部分）進行雜合的 gRNA 的適當序列。

Q.2.9 Choose the N nucleotide within the NGG sequence (PAM sequence) for the appropriate gRNA sequence. 1.0pt
從 NGG 序列（PAM 序列）中選擇 N 核苷酸，作為適當的 gRNA 序列。

Analyze the following HDR template sequence. Shaded fragments are flanking sequences which are homologous to the target sequence.

分析以下的 HDR 模板序列。陰影區域是目標序列區兩側的同源序列。

```
5′-CACAGCATCCAGGGAGAAAT-[160 nucleotides]-CAACCCCAGAAACAGACATC
            ATGGTGCATCTGACTCCTGTGGAGAAGTCTGCCGTTACTGCC
CTGTGGGGAAAGGTGAACTC-[160 nucleotides]-ACCCTTGGACCCAGCGGTAC-3′
```

This sequence is also available in the “Input Sequences” menu labeled “HDR-template”. There, each of the three parts of the sequence are separated onto different lines.

這個序列也可以在標記為"HDR-template" 的" 輸入序列" 選單中找到。在其中，該序列的三個部分，分別放在不同的行上。

Q.2.10

True or False?

正確或錯誤

Q2.10.1	The central sequence (non-shaded) in the HDR template should be divisible by 3. 在 HDR 模板中心序列（非陰影區域）應該是可被 3 整除的數。	1.0pt
Q2.10.2	Using this HDR template will change the total number of amino acid residues in the target mouse beta-globin chain. 使用這個 HDR 模板將改變目標小鼠 β-球蛋白鏈中的總氨基酸殘基數量。	1.0pt
Q2.10.3	Non-Homologous-End-Joining repair may induce frame shifting in the Hbb gene. 非同源末端結合修復可能會在引起 Hbb 基因中的框移。	1.0pt
Q2.10.4	Using this HDR template is likely to affect the amino acid sequence of the mouse embryonic beta-globin chain, encoded by the Hbb-bh1 gene. 使用這個 HDR 模板很可能會影響由 Hbb-bh1 基因編碼的小鼠胚胎 β-球蛋白鏈的氨基酸序列。	1.0pt

END OF TASK 2

任務 2. 結束

Appendix 1 Data format types 附錄 1 資料格式類型

FASTA

The *"fasta"* format is a format used to store DNA/RNA/protein sequences.

"fasta" 格式是用於存儲 DNA / RNA / 蛋白質序列的格式。

Each sequence has a label (header) that starts with a “>”sign. The actual sequence starts on the next line and goes on until the next label or the end of the file. An example is shown below.

每個序列都有一個以“>”符號開頭的標籤（標頭）。實際序列從下一行開始，直到下一個標籤或文件結尾。以下是一個示例。

>Examp1

AGTCGATCGACTAGCATCAGC

CACTACGTCAGCAT

>Examp2

AGTCGATGCACTAGCATCAGCCACTA

Note: Usually only four letters are used to represent both RNA and DNA sequences: A, C, G and T, as‘T’stands for both thymine and uracil. For amino acid sequences, single letter codes are used (see Appendix 5).

注意：通常只使用四個字母來表示 RNA 和 DNA 序列：A、C、G 和 T，其中'T' 代表胸腺嘧啶和脲嘧啶。對於氨基酸序列，使用單個字母的代碼（見附錄 5）。

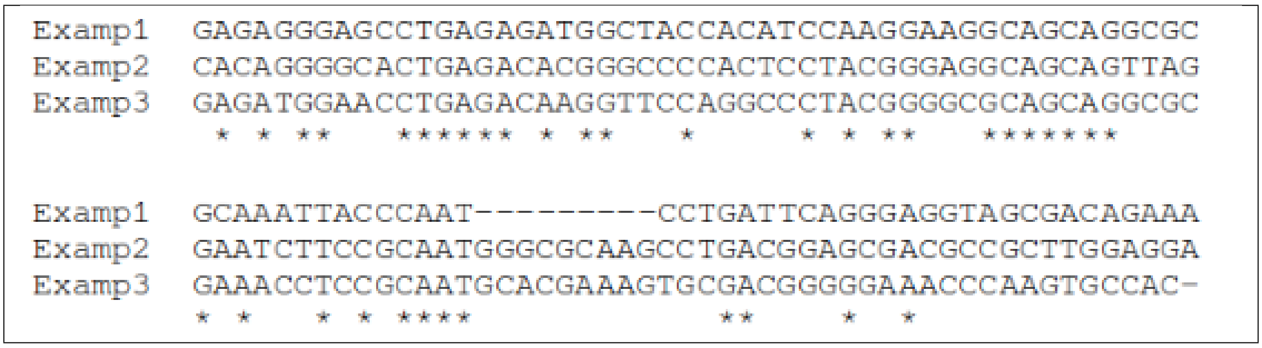
CLUSTAL

Below is an example of a sequence alignment using the *"clustal"* format

以下是使用“clustal”格式進行的序列對齊示例

CLUSTAL X (1.81) multiple sequence alignment

CLUSTAL X（1.81）多序列對齊



```
Examp1    GAGAGGGAGCCTGAGAGATGGCTACCACATCCAAGGAAGGCAGCAGGCGC
Examp2    CACAGGGGCACTGAGACACGGGCCCCACTCCTACGGGAGGCAGCAGTTAG
Examp3    GAGATGGAACCTGAGACAAGGTTCCAGGCCCTACGGGGCGCAGCAGGCGC
           * * **   ****** * **   *     * * **   *******

Examp1    GCAAATTACCCAAT---------CCTGATTCAGGGAGGTAGCGACAGAAA
Examp2    GAATCTTCCGCAATGGGCGCAAGCCTGACGGAGCGACGCCGCTTGGAGGA
Examp3    GAAACCTCCGCAATGCACGAAAGTGCGACGGGGGAAACCCAAGTGCCAC-
          * *   * * ****            **    *  *
```

The first line is a header. It specifies the alignment format. The header is followed by sequence blocks (in the example above there are two blocks). Each block has the sequence labels followed by the corresponding sequences. In this example there are three sequences labeled ‘Examp1’, ‘Examp2’and ‘Examp3’.

第一行是標頭。它指定了對齊格式。標頭後面是序列區塊（在上面的示例中有兩個區塊）。每個區塊都有序列標籤，後面跟著對應的序列。在這個示例中，有三個序列標記為'Examp1'、'Examp2' 和'Examp3'。

(A:0.3,(B:0.4,(C:0.3,D:0.2):0.2):0.1);

The dashes ‘-’ in the sequence indicate alignment gaps resulting from insertions or deletions during the evolution of the analyzed sequences. Beneath the last sequence there is a line with “*” symbols indicating positions conserved across all analyzed sequences.

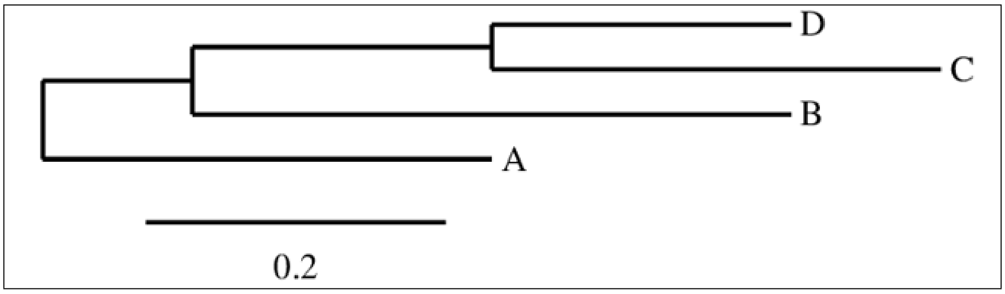
序列中的破折號'-' 表示在分析序列的演化過程中由於插入或刪除而產生的對齊間隙。在最後一個序列下面，有一行帶有"*" 符號標記的位置，表示在所有分析的序列中保持一切不變的位置。

NEWICK

Below is an example of a tree expressed in the *"newick"* format, followed by a graphical representation of the tree it describes:

以下是以"newick" 格式表示的樹的示例，後面是描述該樹的圖形表示：

(A:0.3,(B:0.4,(C:0.3,D:0.2):0.2):0.1);



In the "*newick*" format, characters to the left of the colon ‘:’ symbol are node or leaf labels; numbers to the right of the colon ‘:’ symbol indicate the length of the corresponding branch and brackets ‘()’ are used to group leaves and nodes into branches; the tree ends with a semi-colon symbol “;”.

在"newick" 格式中，冒號':' 符號左邊的字元是節點或葉子的標籤；冒號':' 符號右邊的數字表示相應分支的長度，括號'()' 用於將葉子和節點分組成分支；樹以分號符號";" 結束。

Appendix 2. Bioinformatics tools available in the application 附錄 2. 應用程式中提供的生物資訊學工具

Tool 工具	**Description 敘述**	**Input 輸入**	**Output 輸出**
1 Codon Alignment 密碼子對齊	Align two or more nucleotide sequences by codons 按照密碼子將兩個或更多的核苷酸序列進行對齊	1.A protein alignment ("*clustal*") 蛋白質對齊 2.Corresponding nucleotide codons sequences ("fasta") 相應的核苷酸密碼子序列	Nucleotide sequences with only aligned codons shown 只顯示已對齊密碼子的核苷酸序列
2 DNA to protein DNA 轉蛋白質	Translate DNA nucleotide sequence to protein amino acid sequence 將 DNA 核苷酸序列轉譯為蛋白質氨基酸序列	Nucleotide sequence (text) 核苷酸序列（text）	Encoded protein amino acid sequence 編碼的蛋白質氨基酸序列
3 ORF Finder 3 ORF 查找器	Find open reading frames in nucleotide sequence 在核苷酸序列中尋找開放閱讀框	1.Minimum ORF length (numeric) 最小 ORF 長度 2.Genetic code type (numeric) 遺傳密碼類型 3.Nucleotide sequence ("*fasta*") 核苷酸序列	ORF size (in amino acid residues), DNA coding sequence and encoded protein sequence ORF 大小（以氨基酸殘基表示）、DNA 編碼序列和編碼的蛋白質序列
4 Restriction mapper 限制酶圖譜定位器	Find restriction sites in a nucleotide sequence 在核苷酸序列中尋找限制酶切位點	Nucleotide sequence (text) 核苷酸序列（text）	Number of cut sites for the corresponding restriction enzymes 相應限制酶的切割位點數量
5 Reverse Complement 反向互補	Returns the reverse complement of a nucleotide sequence 返回核苷酸序列反向互補序列	Nucleotide sequence (text) 核苷酸序列（text）	Reverse complement nucleotide sequence 反向互補的核苷酸序列
6 Sequence Editor 序列編輯器	Returns the length, start and end position of a selected subsequence 返回所選子序列的長度、起始和結束位置	Nucleotide sequence (text) 核苷酸序列	Length, start and end position of a selected subsequence 所選子序列長度、起始和結束位置
7 Sequence Alignment 序列對齊	Performs alignment of two or more nucleotide / protein sequences 對兩個或多個核苷酸/蛋白質序列進行對齊	A set of at least two nucleotide / protein sequences ("*fasta*") 至少兩個核苷酸/蛋白質序列的集合	Aligned sequences ("*Clustal")* 對齊的序列
8 Tm Calculator 溶解溫度計算器	Calculates melting temperature for a nucleotide sequence 計算核苷酸序列的熔解溫度	Nucleotide sequence (text) 核苷酸序列	Melting temperature in degrees Celsius 熔點溫度（攝氏度）
9 Tree Builder 樹形構建器	Builds a phylogenetic tree based on a nucleotide / protein sequence alignment 根據核苷酸/蛋白質序列對齊構建親緣演化樹	Nucleotide/protein alignment ("*Clustal")* 核苷酸/蛋白質序列對齊（"clustal"）	Phylogenetic tree ("*newick*"+image*)* 親緣演化樹（"newick"+ 影像）

Appendix 3. Cas9: function and application 附錄 3. Cas9：功能與應用

To undertake genome editing of mammalian cells, scientists transform them with a vector to synthesise Cas9 enzyme and gRNA. Cas9 then cleaves DNA at the locus targeted by gRNA.

為了對哺乳動物細胞進行基因組編輯，科學家們用一個質體轉化它，以合成 Cas9 酶和 gRNA。然後，Cas9 酶在 gRNA 指定的位點上切割 DNA。

This can trigger the homology directed repair (HDR) pathway, which repairs the the double-strand break using complementary sequences (either sister chromosomes, or an artificially introduced repair template). The repair template can contain a desired mutation, or even an additional DNA fragment.

這可以觸發同源定向修復（HDR）途徑，該途徑使用互補序列（姊妹染色體或人工引入的修復模板）修復雙鏈斷裂。修復模板可以包含所需的突變，甚至可以是額外的 DNA 片段。

An alternative mechanism used by cells to repair double-strand breaks is Non-Homologous-End-Joining (NHEJ). NHEJ typically deletes or inserts several random nucleotides whilst fusing the cut DNA back together.

細胞用於修復雙鏈斷裂的另一種機制是非同源末端連接（NHEJ）。NHEJ 通常在將切割的 DNA 重新連接時刪除或插入數個隨機核苷酸。

An overview of the CRISPR-Cas9-mediated genome editing is shown in the figure below.

下圖顯示了 CRISPR-Cas9 介導的基因組編輯的概述。

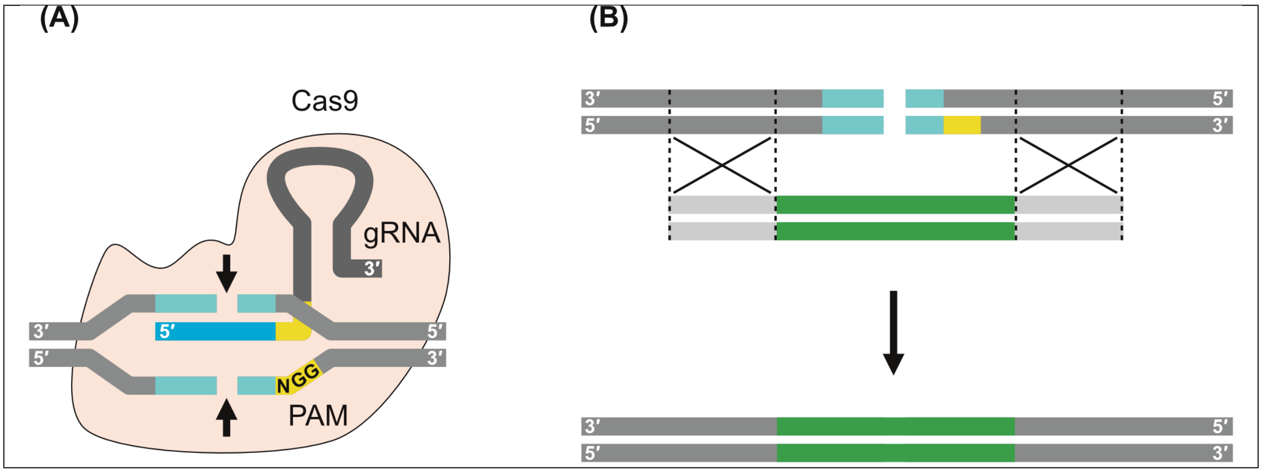


Figure **A.** shows how the Cas9 enzyme recognizes the cleavage site and introduces a double strand break into the genomic DNA. The target sequence is indicated in blue, while the PAM sequence (see main exam text) is shown in yellow. **B.** shows repair by HDR. Homologous recombination, which occurs between genomic DNA and flanking sequences in the HDR template, is indicated by crossed lines

圖 A 顯示了 Cas9 酶如何辨識切割位點並在基因組 DNA 中引入雙股斷裂。目標序列以藍色表示，而 PAM 序列（請參閱主要考試文本）以黃色顯示。圖 B 顯示了 HDR 修復。同源重組發生在基因組 DNA 和 HDR 模板中的鄰近序列之間，以交叉線表示。

Appendix 4. Multiple cloning sites of the vector 附錄 4. 轉錄載體的多重選殖位點

The figure below presents the sequence, with some features labelled, of the ‘multiple cloning site’ of the mammalian expression vector that you plan to use in your cloning.

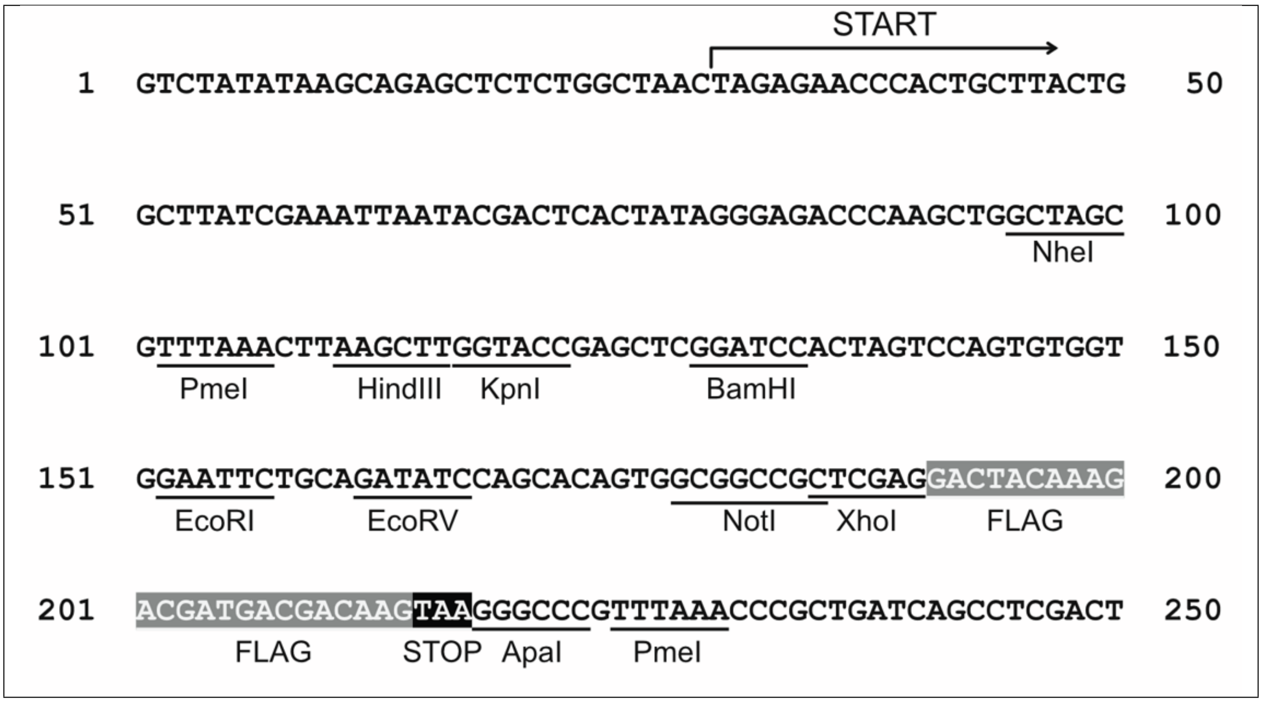
下圖顯示了你計劃在選殖中使用的哺乳動物表達載體的「多重選殖位點」的序列，其中一些特徵已標記出來。

First, the vector will be transformed into *Escherichia coli* for amplification. It is then isolated from bacterial cells, sequenced, and transfected into mammalian cells to produce recombinant Cas9-FLAG enzyme.

首先，該載體將被轉化為大腸桿菌進行增殖。然後，從細菌細胞中分離出載體，進行序列分析，並轉染入哺乳動物細胞中以產生重組的 Cas9-FLAG 酶。

FLAG is a commonly used artificial tag recognised by tool antibodies for immunodetection of recombinant proteins. In this case, western blotting may be performed on a crude lysate of transfected mammalian cells to confirm proper expression of the recombinant Cas9 enzyme.

FLAG 是一種常用的人工標記，可以被工具性抗體識別，用於免疫檢測重組蛋白。在這種情況下，可以對轉染的哺乳動物細胞的粗提取物進行西方墨點法檢視（western blotting），以確認重組 Cas9 酶的正確表達。



```
1    GTCTATATAAGCAGAGCTCTCTGGCTAACTAGAGAACCCACTGCTTACTG    50
51   GCTTATCGAAATTAATACGACTCACTATAGGGAGACCCAAGCTGGCTAGC    100
101  GTTTAAACTTAAGCTTGGTACCGAGCTCGGATCCACTAGTCCAGTGTGGT    150
151  GGAATTCTGCAGATATCCAGCACAGTGGCGGCCGCTCGAGGACTACAAAG    200
201  ACGATGACGACAAGTAAGGGCCCGTTTAAACCCGCTGATCAGCCTCGACT    250
```

Figure: Multiple cloning site (MCS) and promoter region of the vector.

圖：載體的多重選殖位點（MCS）和啟動子區域。

Legend: START –transcription start site

圖例：START –轉錄起始位點 FLAG –coding sequence for the FLAG epitope

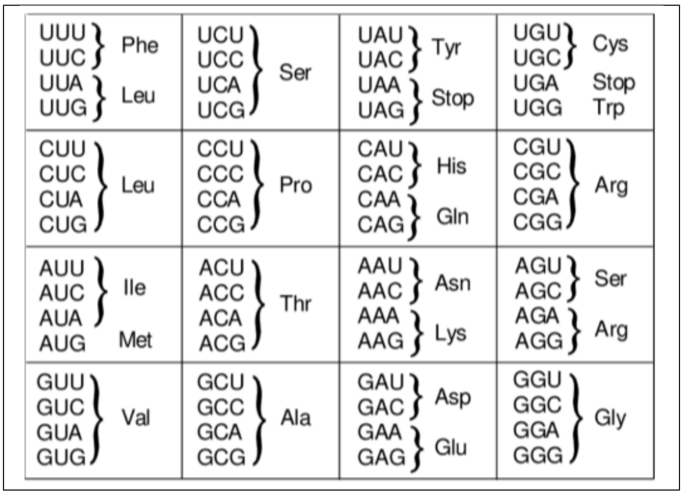
FLAG –FLAG 表位的編碼序列 STOP –stop codon for FLAG-coding sequence

STOP –FLAG 編碼序列的終止密碼子 Underlined sequence is recognized by the enzyme indicated under the sequence.

被底線劃出的序列是被指示在序列下方的酵素所識別的。Numbers either side of each line of sequence indicate the nucleotide position, within the vector sequence, at the start and end of that line, respectively

每行序列兩側的數字分別表示該行開始和結束的核苷酸位置，在載體序列內的位置。

Appendix 5. Standard Genetic Code table 附錄 5. 標準遺傳密碼表



UUU UUC Phe UUA UUG Leu	UCU UCC UCA Ser UCG	UAU UAC Tyr UAA UAG Stop	UGU UGC Cys UGA Stop UGG Trp
CUU CUC CUA Leu CUG	CCU CCC CCA Pro CCG	CAU CAC His CAA CAG Gln	CGU CGC CGA Arg CGG
AUU AUC AUA Ile AUG Met	ACU ACC ACA Thr ACG	AAU AAC Asn AAA AAG Lys	AGU AGC Ser AGA AGG Arg
GUU GUC GUA Val GUG	GCU GCC GCA Ala GCG	GAU GAC Asp GAA GAG Glu	GGU GGC GGA Gly GGG

Amino acids encoding with 3- and 1-letter codes

用 3 字母和 1 字母代碼編碼的氨基酸

Amino acid	3-letter code	1- letter code	Amino acid	3-letter code	1-letter code
Glycine 甘氨酸	Gly	G	Proline 脯氨酸	Pro	P
Alanine 丙氨酸	Ala	A	Valine 纈氨酸	Val	V
Leucine 白氨酸	Leu	L	Isoleucine 異白氨酸	Ile	I
Methionine 甲硫酸	Met	M	Cysteine 半胱氨酸	Cys	C
Phenylalanine 苯丙氨酸	Phe	F	Tyrosine 酪氨酸	Tyr	T
Tryptophan 色氨酸	Trp	W	Histidine 組氨酸	His	H
Lysine 離氨酸	Lys	K	Arginine 精氨酸	Arg	R
Glutamine 麩醯氨酸	Gln	Q	Asparagine 天門冬醯氨酸	Asn	N
Glutamate 麩氨酸	Glu	E	Aspartate 天門冬氨酸	Asp	D
Serine 絲氨酸	Ser	S	Threonine 蘇氨酸	Thr	T