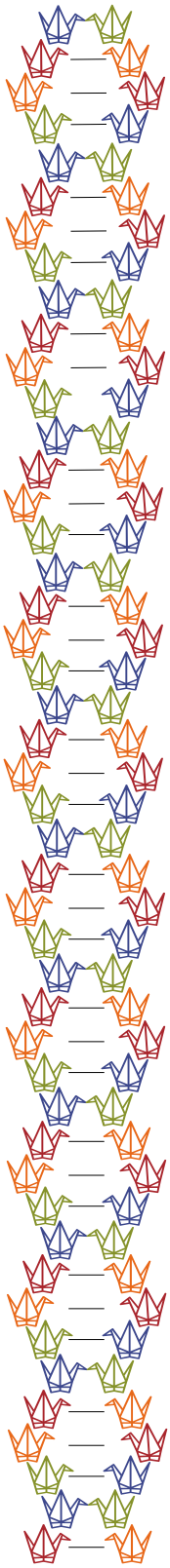

Practical Exam 2

Bioinformatics

signature

2020.8.11.

IBO Challenge 2020

A Substitute for The 31st IBO 2020 Nagasaki, JAPAN

Practical Exam 2, Bioinformatics

實作題 2，生物資訊學

Use the IBO Challenge 2020 Bioinformatics (IBOC) application to address the following questions. The IBOC application may be accessed using any Web Browser.
請使用 IBOC 的應用程式回答下列問題。你可使用任何網頁瀏覽器獲得 IBOC 應用程式。

Find the URLs of your country's server at:
這是貴國的伺服器
URL: **https://bit.ly/IBO2020file**

Or refer to the URL list at the end of this exam file/booklet.
或使用試卷末/手冊的網址

Then, enter the server using the following username and password
請輸入下列這組使用者名稱與密碼登入
Username: **ibo2020**
Password: **ibo2020binagasaki**

If you have a connection problem, try the following alternative servers;
如果你有連線問題，請使用如下替代伺服器：
Competitors in Asia, Oceania, or America: 18.181.30.53/ec2-user/ibo2020bi
亞洲、大洋洲、美洲的參賽者請使用：18.181.30.53/ec2-user/ibo2020bi
Competitors in Europe: 18.184.254.217:3838/ec2-user/ibo2020bi/
歐洲的參賽者請使用：18.184.254.217:3838/ec2-user/ibo2020bi/

IBOC applications should not be accessed except during exam hours. The above URL should not be shared with others even after taking the test.
除考試期間外，不可使用 IBOC 應用程式。即使在完成此測驗後，也不可與其他人分享上述 URL。

If you experience any problems using the IBOC application, reload the application using the reload button on your Web Brower.
如果你在使用 IBOC 應用程式時遭遇到任何問題，請你按 reload 按鍵重新載入網頁瀏覽器。

The IBOC application consists of 14 tabs. After starting the test, check first to see if you can open all tabs. Immediately after the application is loaded for the first time, it may take about 1 minute for a tab to be loaded.

IBOC 應用程序包含 14 個選項。 開始測驗後，請先檢查是否可以打開所有選項。 立刻開始查看該應用程序後，可能需要大約 1 分鐘的時間才能加載標籤頁。

This exam consists of three parts: Part 1: 13 marks, Part 2: 66 marks, and Part 3: 21 marks. The total score is **100 marks.** You will have **90 minutes** to answer all questions. These questions are designed to be solved in order, from beginning to end.
考試分為三部分：第一部分：13 分；第二部分：66 分；第三部分：21 分。 總分是 **100 分**。 您將有 **90 分鐘**的時間回答所有問題。 這些問題設計旨在從頭到尾依序回答。

Overview of the topics covered in the bioinformatics questions
生物資訊測驗題所包含的議題

Bioinformatics is a discipline that aims to elucidate how information is transmitted and interpreted within living organisms. Research scientists use bioinformatics tools, to convert observations of natural phenomena into digital format, and subsequently to visualize, analyze, and interpret this information using computers.
生物資訊學是一門旨在闡明遺傳訊息如何在生物體內傳遞和解釋的學門。科學家使用生物資訊學工具將對自然現象的觀察轉換為數位格式，然後使用電腦對這些資訊進行視覺化，分析和詮釋。

Digitization has benefited scientific progress in ways that are well beyond simply making the data easier to handle. First, the development of more powerful computers enabled the simultaneous analysis of amounts of data so large that the task would be impossible to perform manually. In addition, online sharing of the data has made it possible for researchers around the world to share their observations on various life phenomena more efficiently. In particular, research on genomic DNA sequences and on the three-dimensional structure of proteins benefited very early on from the advent of computing technology. Nowadays, the benefits of computers in the life sciences are widespread.
不僅僅是把資訊轉化得更容易處理，數位化已大大地加速了科學的進步。首先，功能更強大的電腦之發展使得科學家能夠同時分析龐大且無法手動操作的數據資料量。此外，在線上共享數據使世界各地的研究人員，可以更有效地共享他們對各種生命現象的觀察結果。特別是，基因組 DNA 序列和蛋白質三維結構的研究，從電腦分析計算技術的出現就非常早就受益。 如今，電腦在生命科學中應用的好處已得到廣泛應用。

Use the IBOC application to address the following questions. IBOC application may be accessed using any Web Browser.
請使用 IBOC 的應用程式來回答以下的問題。你可使用任何網頁瀏覽器來獲得 IBOC

Note: These application tools may return an error string written in red letters like the following: " Error: An error has occurred. check your logs or contact the app author for clarification". You will see this if the query data is formatted incorrectly.

注意：這些應用程序工具可能會出現如下以紅色字母表示的錯誤字符串：“錯誤：發生了錯誤。請檢查您的日誌或與應用程序作者聯繫以進行澄清”。 如果分析結果不存在或查詢數據格式錯誤，您將看到此訊息。

Part 1: Genome databases
第一部分：基因體資料庫

Read the following information about amino acid and nucleotide sequence databases, and answer the questions below.
閱讀以下有關氨基酸和核苷酸序列資料庫的訊息，並回答以下問題。

Genbank is known as one of the first DNA sequence databases and is maintained by NCBI (National Center for Biotechnology Information, USA). In Genbank, information for each gene is displayed in a format called the GenBank Format, as shown in ***Genes 1*** tab. Read the following section and answer the questions.
Genabank 是第一個由美國國家生物技術信息中心（NCBI）所維護營運的 DNA 序列資料庫。在 Genbank 中，每一個基因被呈現的格式被稱為 GenBank 格式，如同在 ***Genes 1*** 選項中所呈現的。請閱讀以下資訊並回答問題。

The ***Genes 1*** tab includes information about the human *HoxA5* gene as registered in the RefSeq database with the accession number NM_019102, and displayed in GenBank Format. In GenBank Format, each gene entry line begins with a **LOCUS** tag, and ends with two slashes (//). Most of the information is described as a combination of tags and data corresponding to the tags. A tag is a word shown at the beginning of a line, such as **LOCUS** or **DEFINITION**. For example, the data corresponding to the **LOCUS** tag is "NM_019102 1670 bp mRNA linear PRI 31-DEC-2019".
Genes 1 選項包含有關人 HoxA5 基因的資料訊息，該資料已在 RefSeq 資料庫中註冊，註冊號為 NM_019102，並以 GenBank 格式顯示。在 GenBank 格式中，每個基因輸入行均以 **LOCUS** 選項開頭，並以兩個斜杠（//）結尾。大多數訊息被描述為選項和對應於選項的數據的組合。選項顯示在行首的單詞，例如 **LOCUS** 或 **DEFINITION**。例如，對應於 **LOCUS** 選項的數據是“ NM_019102 1670bp mRNA 線性 PRI 31-DEC-2019”。

SOURCE tag shows the species name (Homo sapiens (human)). **FEATURES** tag corresponds to several sub-sections; **source**, **gene**, **exon**, and **CDS**. "/chromosome=7" in the source sub-section shows the chromosome number of which the gene of the entry is located. “/gene=HOXA5 in the gene sub-section shows the name of the gene of the entry. The **CDS** sub-section indicates that the region d for amino acids starts at the 75th base and ends at the 887th base. The "/translation=" part of the **CDS** sub-section contains the amino acid sequence translated from the coding sequence (CDS) of this gene. There are two **exon** sub-sections in the **FEATURES** tag, described as "exon 1..636" and "exon 637..1670". This means that the gene has two exons, and that the boundary between the two exons is located in between the 636th and 637th bases, for a final transcript length of 1670 bp. Finally, The **ORIGIN** tag contains the DNA sequence

of the genomic region that is transcribed as messenger (mRNA), written from the 5'- to the 3'-end.
SOURCE 選項顯示物種學名（智人（人類））。 **FEATURES** 選項對應於幾個次小節；來源，基因，外顯子和編碼區（**CDS**）。 在來源次小節" / chromosome = 7"顯示基因所在的染色體編號。 " / gene = HOXA5 在基因次小節中顯示了基因的名稱。 **CDS** 次小節指出，氨基酸的區域 d 在第 75 個鹼基處開始，在第 887 個鹼基處結束。 **CDS** 次小節的" / translation ="部分包含從該基因的編碼序列（CDS）翻譯而來的氨基酸序列。
FEATURES 標記中有兩個外顯子次小節，分別描述為"外顯子 1..636"和"外顯子 637..1670"。這意味著該基因具有兩個外顯子，並且兩個外顯子之間的邊界位於第 636 和 637 個鹼基之間，最終轉錄本長度為 1670 bp。最後，**ORIGIN** 選項包含轉錄為傳訊（mRNA）的基因組區域的 DNA 序列，從 5'-端寫入 3'-端。

Question 1. According to the amino acid sequence shown in the /translation= part of the above GenBank entry, the protein encoded by this gene starts with an M (methionine) at the first position, followed by an S (serine) at the second position, and an another S (serine) at the third position. Assuming the nucleotide sequence is exactly the same as the data shown in the above GenBank entry, select the most appropriate nucleotide sequences coding for the second and third position serines. [3 marks] **[No. 1]**

問題 1. 根據上述 GenBank 輸入資料，/ translation =顯示的氨基酸序列，此基因編碼的蛋白質在第一個位置以 M（甲硫氨酸）開始，在第二個位置以 S（絲氨酸）開頭。 位置，另一個 S（絲氨酸）位於第三位置。 假定核苷酸序列與上述 GenBank 輸入資料中顯示的數據完全相同，請選擇最合適的編碼第二個和第三個位置絲氨酸的核苷酸序列。[3 分] **[No. 1]**

	2nd position serine,	3rd position serine
1.	AAT,	TGT
2.	AAT,	AAT
3.	GAC,	GCA
4.	GAC,	GAC
5.	TCT,	TCT
6.	AGC,	AGC
7.	AGC,	TCT
8.	GCC,	CGG

In the GenBank entry, not only the amino acid sequence encoded by the target mRNA can be found in /translation=, this information is linked to another database (GenPept) as /protein_id="NP_061975.2".
在 GenBank 輸入訊息中，不僅可以在/ translation =中找到由目標 mRNA 編碼的氨基酸序列，且此訊息也可以連結/protein_id="NP_061975.2"的蛋白質形式鏈接到另一個數據資料庫（GenPept）。

NP_061975.2 is the accession ID of the GenPept database, also operated by NCBI. The GenPept data for NP_061975.2 is shown ***Proteins*** tab.
NP_061975.2 是 GenPept 數據資料庫的註冊 ID，亦由 NCBI 運營。 NP_061975.2 的 GenPept 數據顯示在 ***Proteins*** 蛋白質選項中。

Proteins consist of several regions with different characteristics functions (protein domains), which confer both structure and function. In the above entry, the protein encoded by the *HOXA5* gene has a total length of 270 amino acids, and the Region sub-section indicates that amino acids 176 to 181 constitute a protein domain called the "Antp-type hexapeptide". Another protein domain called "Homeobox domain" spans amino acids 199 to 251. Similarly, you can find an entry for the *HOXA6* gene in the ***Proteins*** tab.
蛋白質由具有不同特徵功能（蛋白質結構域）的幾個區域組成，這些區域賦予結構和功能。 在上面的輸入訊息中，由 HOXA5 基因編碼的蛋白質的全長為 270 個氨基酸，“區域”子區域表示第 176 至 181 位氨基酸構成了稱為“ Antp 型六胜肽”的蛋白質結構域。 另一個稱為“同源框結構域”的蛋白質結構域跨越 199 至 251 位氨基酸。您也可以在“***Proteins***”選項中找到 HOXA6 基因的輸入訊息。

One of the tools for discovering functional domains in amino acid sequences is hmmscan in Hmmer3 (Mistry *et al.*, 2013). It enables the discovery of common domains (eg. “Homeodomain”) in a given family of genes. In the ***HMMSCAN*** tab of the IBOC application, you can use hmmscan to examine the functional domains contained in the amino acid sequence of your protein of interest. This requires an existing protein domain database, and IBOC applications are designed to search the Pfam-A database. Let's check the position of the Homeobox domain of *HOXA5* examined in Question 1 using hmmscan.
用於發現氨基酸序列功能結構域的工具之一是 Hmmer3 中的 hmmscan 功能（Mistry et al.，2013）。 它使得能夠發現給定基因家族中的公共結構域（例如“同源結構域”）。 在 IBOC 應用程式的 ***HMMSCAN*** 選項中，您可以使用 hmmscan 檢查目標蛋白質的氨基酸序列中包含的功能結構域。 這需要現有的蛋白質結構域數據庫，並且 IBOC 應用程序旨在搜索 Pfam-A 數據庫。 讓我們使用 hmmscan 檢查問題 1 中檢查的 HOXA5 的 Homeobox 域的位置。

In the ***Protein Sequences 1*** tab, several amino acid sequences are shown in **FASTA format** which is one of the most commonly used formats to describe nucleotide and amino acid sequences. A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line (defline) is distinguished from the sequence data by a greater-than (">") symbol at the beginning.
在“***Protein Sequences 1***”選項中，以 **FASTA** 格式顯示了多個氨基酸序列，這是描述核苷酸和氨基酸序列的最常用格式之一。 FASTA 格式的序列以單行描述開頭，然後是序列數據行。 此描述線（defline）與序列數據的開頭是大於（>“）符號。

From the ***Protein sequences 1*** tab, copy the amino acid sequence of HOXA5 protein in FASTA format and paste it in the "Input sequence" box of the ***HMMSCAN*** tab. An E-value is calculated to estimate the probability of the result happening by chance, and the recommended threshold is 1e-5 (=0.00001). Clicking on "Exec hmmscan" will display the regions containing functional domains for HOXA5 (if present). If done correctly, the following should be displayed (Figure 1.):

從“***Protein sequence 1***”選項中，以 FASTA 格式複制 HOXA5 蛋白的氨基酸序列，並將其粘貼到 HMMSCAN 選項的“輸入序列”框中。 計算 E 值以估計偶然發生結果的機率分布，建議閾值為 1e-5（= 0.00001）。 單擊“ Exec hmmscan”將顯示包含 HOXA5 功能域的區域（如果存在）。 如果正確完成，應顯示以下內容（圖 1）：

target name:	target accession:	tlen	query name:	qlen:	E-value:	score:	#	of	from (hmm coord):	to (hmm coord):	from (ali coord):	to (ali coord):	description of target:
Homeodomain	PF00046.30	57	NP_061975.2homeoboxproteinHox-A5[Homosapiens]	270	4.5e-22	77.7	1	1	1	57	196	252	Homeodomain

Showing 1 to 1 of 1 entries Previous 1 Next

Figure 1. An example of Hmmscan search result. This figure shows the output of the hmmscan

圖 1. 此為 Hmmscan 搜索結果的示範案例。 該圖顯示了 hmmscan 的輸出。

The search results display the domain name and Accession ID in the "target name:" and "target accession:" columns, respectively. This information depends on the database provided, and in the case of the IBOC application, the domain name and Accession ID from the Pfam-A database are displayed. "tlen" is the total length of that particular domain in amino acids. Here, the homeodomain registered in Pfam-A is 57 amino acids.

搜索結果分別在“目標名稱：”和“目標註冊：”列中顯示結構域名稱和註冊號。 此訊息取決於所提供的數據庫，在 IBOC 應用程序的情況下，將顯示 Pfam-A 數據庫的結構域名稱和註冊 ID。 “ tlen”是該特定結構域在氨基酸中的總長度。 在此，在 Pfam-A 中註冊的同源結構域是 57 個氨基酸。

The name of the query amino acid sequence (the sequence you copied in, eg. NP_061975.2 homeobox protein Hox-A5 [Homo sapiens] in Figure 1.)" is displayed in the "query name:" column, and its length in the "qlen:" column. If the same domain is found more than once, the total number and serial number are displayed in the "of:" and "#:" columns, respectively. The "from (hmm coord):" column and the "to (hmm coord):" column indicate which part of the domain matched to the database. The "from (ali coord):" and "to (ali coord):" columns show the start and end positions of the domain of interest inside the query sequence. This time, the results show amino acids 1 to 57 of the PF00046 Homeodomain, which is exactly 57 amino acids long, meaning that the entire length of this domain was included in the query sequence.

查詢氨基酸序列的名稱（您在其中複製的序列，例如，圖 1 中的 NP_061975.2 同源異質盒蛋白 Hox-A5 [*Homo sapiens*]智人。）在“查詢名稱：”列中顯示，其長度在 如果多次發現同一結構域，則總數和序列號分別顯示在“ of：”和“ # ：”列中。“ from（hmm

coord）：”列“ to（hmm coord）：”列指示結構域的哪一部分與數據庫匹配。“ from（ali coord）：”和“ to（ali coord）：”列顯示結構域的開始和結束位置 這次，結果顯示 PF00046 Homeodomain 的 1 到 57 位氨基酸，恰好是 57 個氨基酸長，這意味著該結構域的整個長度都包含在查詢序列中。

Question 2. Find a Homeodomain (PF00046.30) in HOXA6 (NP_076919.1 homeobox protein Hox-A6 [Homo sapiens]). Use E-value: 1e-5 as threshold. [4 marks]
Hint: Use the sequence data from the ***Protein sequences 1*** tab and ***HMMSCAN*** tab.
問題 2. 在 HOXA6（NP_076919.1 同源異質蛋白 Hox-A6 [智人]）中找到一個同源結構域（PF00046.30）。 使用 E 值：1e-5 作為閾值。[4 marks] [4 分]

The homeodomain (PF00046.30) region in *HOXA6* gene starts (“from (ali coord):”) at **[No.2][No.3][No.4]**, and ends (“to (ali coord):”) at **[No.5][No.6][No.7]**.
HOXA6 基因的同源結構域（PF00046.30）區域從[no.2] [no.3] [no.4]開始（“從（ali 坐標）：”），結束於（“ to（ali 坐標）：”）。 ）在[5 號] [6 號] [7 號]。

eg) If you would like to answer "starts at **25**” and “ends at **143**”, the example of how to answer is as following,
例如）如果您想回答“開始於 25”和“結束於 143”，則如何回答的示例如下，

No. 2 : 0
No. 3 : 2
No. 4 : 5

No. 5 : 1
No. 6 : 4
No. 7 : 3

Next, let's find where the two exons of *HOXA5* are encoded in the DNA genome. It is convenient to use the NCBI-BLAST+ tool developed by NCBI to search for sequence similarity. To do this, you can choose between the following three programs. When **blastp** is selected, an amino acid sequence query is searched for in an amino acid sequence database. If **blastn** is selected, a nucleotide sequence query is searched against a nucleotide sequence database. Finally, when **tblastn** is selected, an amino acid query may be searched against a nucleotide sequence database.
接下來，讓我們找到 HOXA5 的兩個外顯子在 DNA 基因組中的編碼位置。 使用由 NCBI 開發的 NCBI-BLAST +工具方便地搜索序列相似性。 為此，您可以在以下三個程序之間進行選擇。 當選擇 blastp 時，在氨基酸序列數據庫中搜索氨基酸序列查詢。 如果選擇了 blastn，則針對核苷酸序列數據庫搜索核苷酸序列查詢。 最後，當選擇了 tblastn 時，可以針對核苷酸序列數據庫搜索氨基酸查詢。

The first three tabs of the IBOC application show the following nucleotide sequences.

The ***"Human Genome DNA 1"*** tab shows the nucleotide sequence of human chromosome 7 from positions 27,140,701 to 27,150,700.
The ***"Human Genome DNA 2"*** tab shows the nucleotide sequence of human chromosome 7 from positions 26,188,001 to 26,218,000.
The ***"B. burgdorferi B31 Genome DNA 1"*** tab is the sequence of the entire cyclic genomic DNA of a bacterium (*Borrelia burgdorferi* strain B31).
IBOC 應用程序的前三個選項顯示以下核苷酸序列。
“人類基因組 DNA 1”選項顯示了人類染色體 7 的核苷酸序列，位置從 27,140,701 到 27,150,700。
“人類基因組 DNA 2”選項顯示了人類 7 號染色體的核苷酸序列，位置為 26,188,001 至 26,218,000。
“ B. burgdorferi B31 Genome DNA 1”選項是細菌（*Borrelia burgdorferi* B31 菌株）的整個環狀基因組 DNA 的序列。

In the following questions we will refer to base positions within the analyzed sequence, not positions along the chromosome. For example, we will refer to the first base of the Human Genome DNA 1 sequence as 1, not 27,140,701. The genomic DNA of *B. burgdorferi* strain B31 is circular. Therefore, the first base was arbitrarily defined.
在以下問題中，我們將引用分析序列中的鹼基位置，而不是使用染色體的位置。 例如，我們將人類基因組 DNA 1 序列的第一個鹼基稱為 1，而不是 27,140,701。 *B. burgdorferi* 菌株 B31 的基因組 DNA 是環狀的。 因此，第一個基礎是任意定義的。

After copying the mRNA sequence of *HOXA5* (NM_019102.4) from the ***Predicted mRNA Sequences*** tab, paste it into the "Input sequence:" window of the ***BLAST*** tab. (Note: By convention, RNA sequences are shown as a complementary DNA sequence. This means a coding DNA strand and the transcribed RNA would be shown as identical sequences). Next, by selecting "Human Genome DNA 1" from the "Reference Sequence:" options, a similarity search is performed against “Human Genome DNA 1 ”. This step requires **blastn**. Clicking "Exec" gives the predicted locus of *HOXA5*. In the search results, "sseqid" displays the IDs of the hit sequence.
從“預測的 mRNA 序列”選項中複製 HOXA5 的 mRNA 序列（NM_019102.4）之後，將其粘貼到“ BLAST”選項的“輸入序列：”窗口中。 （注意：按照慣例，RNA 序列顯示為互補的 DNA 序列。這意味著編碼的 DNA 鍊和轉錄的 RNA 將顯示為相同的序列）。 接下來，通過從“參考序列：”選項中選擇“人類基因組 DNA 1”，針對“人類基因組 DNA 1”進行相似性搜索。 此步驟需要使用 **blastn**。 單擊“執行” ("Exec") 將給出 HOXA5 的預測基因座。 在搜索結果中，“ sseqid”顯示命中序列的 ID。

Question 3. Choose the location which is including the first exon of HOXA5 from the following options. Use “E-value:1e-5” as threshold for blastn. **Hint**: Use ***Predicted mRNA sequences*** tab, and ***BLAST*** tab. [2 marks] **[No.8]**

問題 3. 從以下選項中選擇包含 HOXA5 第一個外顯子的位置。 使用“ E-value：1e-5”作為閾值。 提示：使用“預測的 mRNA 序列”選項和“***BLAST***”選項。[2 分] [8 號]

1. 1 bp ~ 1,000 bp
2. 1001 bp ~ 2000 bp
3. 2001 bp ~ 3000 bp
4. 3001 bp ~ 4000 bp
5. 4001 bp ~ 5000 bp
6. 5001 bp ~ 6000 bp
7. 6001 bp ~ 7000 bp
8. 7001 bp ~ 8000 bp
9. 8001 bp ~ 9000 bp
0. 9001 bp ~ 10,000 bp

Question 4. Next, find proteins with a similar amino acid sequence to HOXA6. Using HOXA6 as a query, perform a blastp search on a human protein dataset (“human proteins” in the Reference sequence field). Use “E-value:1e-5” as threshold for **blastp**. Choose the entry with the highest sequence similarity from the following options, but not HOXA6 itself. Use “E-value:1e-5” as threshold for blastp. Hint: Use *Protein sequences 1* tab, and *BLAST* tab. [4 marks] **[No.9]**

問題 4.接下來，找到氨基酸序列與 HOXA6 類似的蛋白質。 使用 HOXA6 作為查詢基礎，對人類蛋白質數據資料庫（“參考序列”字段中的“人類蛋白質”）執行 blastp 搜索。 使使用 “ E-value：1e-5”作為 blastp 的閾值。

從以下選項中選擇序列相似性最高的輸入資料，而不是 HOXA6 本身。 使用“ E-value：1e-5”作為 blastp 的閾值。 提示：使用蛋白質序列 1 標籤和 BLAST 標籤。 [4 分] [9 號]

1. 1_06639
2. 1_05172
3. 1_07981
4. 1_09188
5. 1_12205
6. 1_12968
7. 1_15255
8. 1_15496
9. 1_17260
0. 1_18575

In order to obtain the protein sequence, you should go to the ***Entry Database*** tab within the IBOC application (you would need in questions 18 and 19). The amino acid sequence of the protein will be displayed upon entering the protein’s ID (eg. 1_05121) in the **Entry_ID** input field and choose an organism name from the **Sequence** selection.
Note: This feature may be useful for answering Question 18 and 19.

為了獲得蛋白質序列，您應該轉到 IBOC 應用程序中的 Entry Database 選項（在問題 18 和 19 中需要）。 在 Entry_ID 輸入字段中輸入蛋白質的 ID（例如 1_05121）並從“序列”選擇中選擇生物名稱後，將顯示蛋白質的氨基酸序列。

注意：此功能對於回答問題 18 和 19 可能很有用。

Part 2. Analysis of sequence features and motif discovery
第 2 部分。序列特徵分析和結構模體發現

The question below introduces methods to investigate the GC composition of nucleic acid sequences as a general property of genomic DNA. Read the following information about chromosomes and the GC composition of DNA, and answer the questions below.
以下問題介紹了研究核酸序列的 GC 組成作為基因組 DNA 的一般屬性的方法。 閱讀以下有關染色體和 DNA 的 GC 組成的訊息，並回答以下問題。

As their name suggests, chromosomes can be dyed using one of several staining methods. Among these methods, Giemsa staining results in a striped pattern of bands along the chromosome. The staining is usually performed during the **[No. 10]** of cell division, when the chromosomes are visible because of chromosome condensation. The coloration is darker in regions that have a **[No. 11]** AT content, and lighter and brighter in regions that have a **[No. 12]** GC content. Furthermore, regions with a **[No. 12]** GC content were thought to have more genes. Such a correlation between GC content and gene density was confirmed when the nearly complete sequence of human nuclear DNA (draft genome) was determined by the Human Genome Project in 2000. In addition, DNA tend to undergo denaturation (separation of the double helix into two single strands) at **[No. 13]** temperatures. GC pairs are denatured at a **[No. 14]** temperature than AT pairs. As described above, the GC content is a value that is both easy to obtain and very useful in various aspects of bioscience and biotechnology research.
顧名思義，可以使用幾種染色方法之一對染色體進行染色。在這些方法中，Giemsa 染色結果呈現出沿著染色體呈條紋狀的條帶。染色通常在細胞分裂的[No. 10]期間進行，當由於染色體濃縮而可見到染色體時。在 AT 含量為[No. 11]的區域中，顏色較深，在 GC 含量為[No. 12]的區域中，顏色較淺。此外，具有[No.12] GC 含量的區域被認為具有更多的基因。當 2000 年人類基因組計劃確定了人類核 DNA（基因組草稿）的近乎完整序列時，就證實了 GC 含量與基因密度之間的這種相關性。進一步來說，DNA 傾向於[No. 13]溫度時經歷變性（將雙螺旋分離為兩個單股）。 GC 對在溫度[No. 14]相較 AT 對易於變性。如上所述，GC 含量是一個易於獲得的值，在生物科學和生物技術研究的各個方面都非常有用。

Question 5. Pick the answer most appropriate for **[No. 10]**. [4 marks]
問題五 為[No. 10]挑選最佳的答案[4 分]

1. G0 phase
2. S phase
3. Metaphase

Question 6. Pick the answer most appropriate for **[No. 11], [No. 12], [No. 13],** and **[No. 14]**. [8 marks]

問題六. 為[No.11]、[No.12]、[No.13]與[No.14]挑選最佳的答案[8 marks]

1. higher
2. lower

From here, let us focus on the GC content of DNA sequences. The GC content (GC%) is represented by the following formula.

從這裡開始，讓我們關注 DNA 序列的 GC 含量。 GC 含量（GC%）由下式表示。

GC% = 100 × (([G] + [C]) / ([A] + [T] + [G] + [C]))

In this formula, [A] represents the number of adenosines (A), and similarly, [T], [G], and [C] each stand for the number of thymines (T), guanosines (G), and cytosines (C), respectively.

在此方程式中，[A]代表腺核苷（A）的數目，相似地，[T]、[G]和[C]分別代表胸腺嘧啶（T），鳥核苷（G）和胞嘧啶的數目（C）。

Question 7. For the DNA sequence shown in ***"Human Genome DNA 1"*** tab, calculate the GC content in the ten bases from the 1st to the 10th position included. [2 marks]

問題 7：對於“人類基因組 DNA 1”選項中顯示的 DNA 序列，請計算從第 1 位到第 10 位的 10 個鹼基中的 GC 含量。[2 分]

[No. 15][No. 16][No. 17] %

eg) If you would like to answer "32%", the example of how to answer is as following,

假設你想要回答”32%”，以下是回答方式的範例

No. 15 : 0
No. 16 : 3
No. 17 : 2

The usage of the ***“Count nucleotide”*** tab and the ***“Window search”*** tab is shown below.

“***Count nucleiotide”***選項與***”Window search”***選項的功能如下所示。

DNA sequences from the region of your choice can be extracted from "Human Genome DNA 1", "Human Genome DNA 2", and "*B. burgdorferi* B31 Genome DNA 1" by using the **Get Subsequence** function on the left side of the ***"Count Nucleotide" tab***.

Select **Human Genome DNA 1** from the **Target Sequence** selection list, then enter **1** and **10** in the "**Start**" and "**End**" fields, respectively. This will display the first 10 bases from the first to the tenth base of "Human Genome DNA 1". Note:This feature may be useful for answering Question 15 and 16.
使用“計數核苷酸”選項左側的“獲取子序列”功能，可以從“人類基因組 DNA 1” 提取你所選區域的 DNA 序列，“人類基因組 DNA 2”和“*B. burgdorferi* B31 基因組 DNA 1”中。

從“目標序列”選擇列表中選擇“人類基因組 DNA 1”，然後分別在“開始”和“結束”區塊中輸入 1 和 10。 這將顯示“人類基因組 DNA 1”的第一個至第十個鹼基的前 10 個鹼基。 此功能對於回答問題 16 可能有用。

Window search is a method used to examine the distribution of DNA features over the entire nucleotide sequence by shifting a window of fixed length by a fixed unit. The length of the window is called Window Size. The length of the window shift is called Step Size.
Window search 是一種通過將固定長度的 window 移動固定單位來檢查 DNA 特徵在整個核苷酸序列上的分佈的方法。 window 的長度稱為“Window Size”。 window 移位的長度稱為 Step Size。

Next, let's examine the sequence characteristics of "Human Genome DNA 1" and "Human Genome DNA 2" using the "***Window Search***" tab of the IBOC application.
接下來，讓我們使用 IBOC 應用程序的 ***Window Search***"選項卡檢查“人類基因組 DNA 1”和“人類基因組 DNA 2”的序列特徵。

Question 8. Let's check the GC content for the first 10,000 nucleotides of "Human Genome DNA 1" using window search.
問題 8：讓我們使用 window 搜索檢查“人類基因組 DNA 1”的前 10,000 個核苷酸的 GC 含量。

1. Open the ***"Window Search"*** tab and select "Human Genome DNA 1" from the **Reference Sequence** list.
   打開”***Window Search***”選項並從 **Reference Sequence list** 選擇”Human Genome DNA”
2. To check the total number of G's and C's combined, select "G+C" from the **Nuc.** selection list.
   為了要檢查 G's 和 C's 的總數，由 **Nuc.**的選項中選擇”G+C”。
3. Enter 1 in the **Start** field and enter 10,000 in the **End** field.
   在 **Start** 欄位中選項入 1 然後在 **End** 欄位選項入 10,000。
4. Set the “Window Size” to 100 bp, the “Step Size” to 100 bp, and the “Bin Size” to 10 %.
   把”Window Size”設定為 100 bp，把”Step Size”設定為 100 bp，然後把”Bin Size”設定為 10%。

5. With the above settings, the GC content will be calculated for every 100 bp window from the 1st base to the 10,000th base of "Human Genome DNA 1".
   依據以上的設定，GC 含量將會在每 100 bp、window 從"Human Genome DNA 1"的第 1 到第 10,000 鹼基的狀況下計算。
6. Finally, clicking on "**Show Chart 1**", displays the result as a histogram representing the frequency of GC content values in intervals of 10 % (corresponding to the histogram's **Bin Size**) in "**Histogram for the selected nucleotide(s)**".
   最後，按下"Show Chart 1"， 最後，單擊"顯示圖表 1"，在"**Histogram for the selected nucleotide(s)**".中以直方圖的形式顯示結果，該直方圖以 10%的間隔（對應於直方圖的 Bin Size)表示 GC 含量值的頻率。
7. Additionally, clicking on "**Show Chart 2**", displays the result as a plot representing the frequency of GC content values along the region above in "**Frequency for the selected nucleotide(s)**".
   此外，單擊**"Show Chart 2"**，將結果顯示為代表"**Frequency for the selected nucleotide(s)**"上方區域中 GC 含量值頻率的曲線圖。

Note: X-axis in Chart-2: nuceotide position (bp). Y-axis in Chart-2: frequency for the selected nucleotide(s).
注意：圖表 2 中的 X 軸：核苷酸位置（bp）。 圖表 2 中的 Y 軸：所選核苷酸的頻率。
Note: In Chart 2, Chart 3 and Chart4, the x-axis position may be displayed as "1e-4" instead of "10,000".
注意：在圖表 2，圖表 3 和圖表 4 中，x 軸位置可能顯示為" 1e-4"，而不是" 10,000"。
Note: If you have changed the parameters, click on "Show Chart 1(or 2, 3, 4)" again to reflect the change.
注意：如果更改了參數，請再次單擊"顯示圖表 1（或 2、3、4）"以反映更改。
Select the most appropriate item for [No. 18] in the following sentence. [4 marks]
在以下句子中為[No. 18]選擇最合適的項目。 [4 分]

Using the above settings, 100 bp windows with a GC content of **[No. 18]** % appear most frequently in "Human Genome DNA 1".
使用上述設置，"人類基因組 DNA 1"中出現[GC 含量為[No. 18]%]的 100 bp 窗口的頻率最高。

1. 10-20
2. 20-30
3. 30-40
4. 40-50
5. 50-60
6. 60-70
7. 70-80
8. 80-90

Question 9. As we saw in **Question 7. ~ 8.** the GC% of DNA varies from region to region. Find out where the 200 bp window with the highest GC content is located on "*B. burgdorferi B31* Genome DNA", and choose the corresponding location from the following options. Note that this genome DNA of this organism is circular and 910,724 bp long. Note that this **calculation may take some time** (about one minute per calculation) in IBOC applications. [5 marks] **[No. 19]**

問題 9：正如我們在問題 7～8 中看到的那樣。DNA 的 GC%隨區域的不同而不同。 找出具有最高 GC 含量的 200 bp window 在“ *B. burgdorferi* B31 Genome DNA”上的位置，然後從以下選項中選擇相應的位置。 注意，該生物的基因組 DNA 是環狀的和 910,724 bp 長。。注意，在 IBOC 應用程序中，此計算可能需要一些時間（每次計算大約需要一分鐘）。[5 分] [19 號]

1. 50,000 bp ~ 100,000 bp
2. 100,000 bp ~ 200,000 bp
3. 200,000 bp ~ 300,000 bp
4. 300,000 bp ~ 400,000 bp
5. 400,000 bp ~ 500,000 bp
6. 500,000 bp ~ 600,000 bp
7. 600,000 bp ~ 700,000 bp
8. 800,000 bp ~ 900,000 bp
9. The region including 900,000 bp ~ 910,724 bp and 1 bp ~ 50,000 bp

In the following section, we will investigate bacterial DNA replication using the methods for analysis such as **Window Search**, which we learned from the previous question.
在下一節中，我們將使用諸如 **Window Search** 之類的分析方法來研究細菌 DNA 的複製，這是我們從上一個問題中學到的。

Question 10. The replication of circular DNA starts at a single position and propagates in both directions (Figure 2).
問題 10：環狀 DNA 的複製從一個位置開始，並向兩個方向傳播（圖 2）。

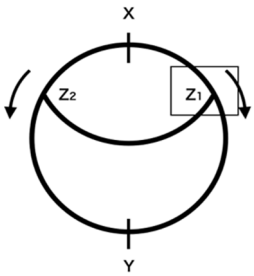

Figure 2. Schematic diagram of DNA replication in a bacterial circular genome. X is the start point of replication (Ori) and Y is the end point of replication. There are two replication forks, shown as Z1 and Z2.
圖 2.細菌環狀基因組中 DNA 複製的示意圖。 X 是複制的起點（Ori），Y 是複製的終點。 有兩個複製叉，分別顯示為 Z1 和 Z2。

If the whole genome sequence of the bacteria is available, to estimate the start and end points of replication, it is useful to examine the GC-skew. GC-skew is an indication of the bias in the balance between G's and C's on one strand of the DNA. It is expressed as the difference between the number of C's and the number of G's divided by the total number of G's and C's over a given sequence length.
如果可獲得細菌的整個基因組序列，以估計複製的起點和終點，則檢查 GC 偏斜將很有用。 GC 偏斜表明 DNA 的一條鏈上 G 和 C 之間的平衡存在偏差。 它表示為在給定的序列長度內，C 數與 G 數之差除以 G 和 C 的總數。

GC-skew = ([C]-[G]) / ([C]+[G])

Using the **Window Search**, look up the GC-skew of the *B. burgdorferi* B31 Genome DNA. In the "***Window Search***" tab, select "*B. burgdorferi* B31 Genome DNA" in the "Reference Sequence", fill the first base and the last base in the **Start** and **End** fields, respectively, and then choose ([C]-[G])/([C]+ [G]) for **Skew** field. Finally, press **Show Chart 3**, the GC-skew plot appears in the Chart 3 area. Note that **this calculation may take some time** (about one minute per calculation) in IBOC applications.
使用**"Window Search"**，查找 *B. burgdorferi* B31 基因組 DNA 的 GC 偏斜。 在"***Window Search***"選項中，在"參考序列"中選擇" *B. burgdorferi* B31 Genome DNA"，分別在"開始"和"結束"字段中填充第一個鹼基和最後一個鹼基，然後選擇（[[C]-[G]）/（[[C] + [G]）表示偏斜字段。 最後，按 Show Chart 3，GC-skew 圖出現在 Chart 3 區域中。請注意，在 IBOC 應用程序中，此計算可能需要一些時間（每次計算大約需要一分鐘）。

There are two switching points between high GC-skew and low GC-skew in the DNA. Choose the corresponding locations among the following options. [8 marks]
DNA 中的高 GC 偏斜和低 GC 偏斜之間有兩個切換點。 在以下選項中選擇相應的位置。[8 分]

1. 50,000 bp ~ 100,000 bp
2. 100,000 bp ~ 200,000 bp
3. 200,000 bp ~ 300,000 bp
4. 300,000 bp ~ 400,000 bp
5. 400,000 bp ~ 500,000 bp
6. 500,000 bp ~ 600,000 bp
7. 600,000 bp ~ 700,000 bp
8. 800,000 bp ~ 900,000 bp
9. The region including 900,000 bp ~ 910,724 bp and 1 bp ~ 50,000 bp

The two switching points are located in **[No. 20]** and in **[No. 21].**
兩個切換點位於[No. 20]和[No. 21]中。

Question 11. It is known that the replication start point, OriC, is in the vicinity of the region encoding the DnaA protein. Examine the region encoding the DnaA protein on the "*B. burgdorferi* B31 Genome DNA" and select the region that contains it.

Think about which tools you need to use to answer this question. You can use the **DnaA** protein sequence from the ***Protein Sequences 2*** tab.

[2 marks] **[No. 22]**

問題 11：已知複製起點 OriC 位於編碼 DnaA 蛋白的區域附近。 檢查*“ B. burgdorferi* B31 基因組 DNA”上編碼 DnaA 蛋白的區域，然後選擇包含該區域的區域。

考慮一下您需要使用哪些工具來回答這個問題。您可以從“蛋白質序列 2”選項中使用 DnaA 蛋白質序列。[2 分][22 號]

1. 50,000 bp ~ 100,000 bp
2. 100,000 bp ~ 200,000 bp
3. 200,000 bp ~ 300,000 bp
4. 300,000 bp ~ 400,000 bp
5. 400,000 bp ~ 500,000 bp
6. 500,000 bp ~ 600,000 bp
7. 600,000 bp ~ 700,000 bp
8. 800,000 bp ~ 900,000 bp
9. The region including 900,000 bp ~ 910,724 bp and 1 bp ~ 50,000 bp

Question 12. It is known that GC-skew values change significantly at both the replication start and end points of replication. Considering the answers to the previous questions, select the region that is most likely to contain the replication start point (OriC) for this organism from the following. [4 marks] **[No. 23]**

問題 12-已知在複製的複製起點和終點，GC 偏斜值都會發生明顯變化。 考慮到先前問題的答案，請從以下內容中選擇最有可能包含該生物體複製起點（OriC）的區域。[4 分] [23 號]

1. 50,000 bp ~ 100,000 bp
2. 100,000 bp ~ 200,000 bp
3. 200,000 bp ~ 300,000 bp
4. 300,000 bp ~ 400,000 bp
5. 400,000 bp ~ 500,000 bp
6. 500,000 bp ~ 600,000 bp
7. 600,000 bp ~ 700,000 bp
8. 800,000 bp ~ 900,000 bp
9. The region including 900,000 bp ~ 910,724 bp and 1 bp ~ 50,000 bp

From here, we will examine the DNA motifs involved in the regulation of gene expression in eukaryotes.

從這裡，我們將研究真核生物中涉及基因表現調控的 DNA 結構模體。

Gene expression can be regulated by the chemical addition of methyl groups to specific bases on DNA (Figure 3).

基因表現可以通過將添加甲基化到 DNA 上特定鹼基來調節（圖 3）。

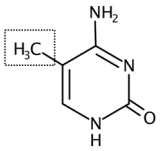

Figure 3. Structural formula of 5-methylcytosine, a methylated cytosine. The dotted area is the added methyl group.
圖 3. 5-甲基胞嘧啶（一種甲基化的胞嘧啶）的結構式。 虛線區域是添加的甲基。

In vertebrates, the cytosines of CpG dinucleotides are often subject to DNA methylation. A CpG dinucleotide, as the name indicates, consists of a consecutive C and G in the 5' to the 3' direction (Figure 4). 'p' between C and G, which stands for the phosphodiester bond so that CpG is distinguished from CG which is a nucleotide pair in a double-stranded DNA.
在脊椎動物中，CpG 二核苷酸的胞嘧啶經常經歷 DNA 甲基化。 顧名思義，CpG 二核苷酸由 5'到 3'方向上連續的 C 和 G 組成（圖 4）。C 和 G 之間的“ p”代表磷酸二酯鍵，因此 CpG 與 CG 有所區別，CG 是雙鏈 DNA 中的核苷酸對。

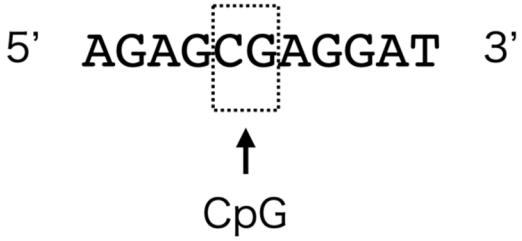

Figure 4. A CpG dinucleotide consists of a consecutive C and G in the 5' to the 3' direction. The dotted area is an example of a CpG dinucleotide.
圖 4. CpG 二核苷酸由 5'到 3'方向上連續的 C 和 G 組成。 虛線區域是 CpG 二核苷酸的例子。

A DNA region that is rich in CpG dinucleotides is known as a CpG island. CpG islands are regularly found in the genomic DNA of vertebrates, and may be located either around the exons of protein-coding genes, or a few hundred bp upstream of their transcription start sites. Not all CpG islands are methylated, but these methylation events are correlated with inactivation of the downstream gene.
富含 CpG 二核苷酸的 DNA 區域被稱為 CpG 島嶼。 CpG 島嶼經常出現在脊椎動物的基因組 DNA 中，可能位於蛋白質編碼基因的外顯子周圍，也可能位於其轉錄起始位點上游數百 bp。 並非所有的 CpG 島嶼都被甲基化，但是這些甲基化事件與下游基因的失活相關。

Several methods have been proposed to look for CpG islands. Among them, the **CpG-score** is expressed by the following formula, and can be calculated over a shifting window as demonstrated previously

已經提出了幾種方法來尋找 CpG 島嶼。 其中，CpG 分數由以下公式表示，可以如前所述在移動視窗上計算

CpG-score = ([CpG] / ([C] x [G])) x Window-Size

In this formula, [CpG] represents the number of CpG in the window, and similarly, [G], and [C] each stand for the number of guanosines (G), and cytosines (C) in the window, respectively.

在此公式中，[CpG]代表窗口中 CpG 的數目，類似地，[G]和[C]分別代表窗口中鳥苷酸（G）和胞嘧啶（C）的數目。

To obtain the CpG score, “Window-Size = 100” and “Step-Size = 1” are generally the standard parameters used. Subsequently, areas with a CpG score of 0.6 or higher are often considered as candidate CpG islands.

為了獲得 CpG 分數，通常使用“ Window-Size = 100”和“ Step-Size = 1”。 隨後，CpG 得分為 0.6 或更高的區域通常被視為候選 CpG 島嶼。

Question 13. As mentioned above, "Human Genome DNA 1" is the sequence from positions 27,140,701 to 27,150,700 on human chromosome 7, for a total of 10,000 bases.

問題 13：如上所述，“人類基因組 DNA 1”是人染色體 7 上從 27,140,701 到 27,150,700 位的序列，共有 10,000 個鹼基。

In the "***Window Search***" tab, select “Human Genome DNA 1” from the **Reference Sequence** field, and examine the distribution of the CpG-score. Enter the appropriate values for the **Start** and **End** fields. Use “Window-Size = 800” and “Step-Size = 1” as the parameters.

在"***Window Search***"選項中，從“**Reference Sequence**”字段中選擇“人類基因組 DNA 1”，並檢查 CpG 得分的分佈。 在“**Start**”和“**End**”字段中輸入適當的值。 使用“ Window-Size = 800”和“ Step-Size = 1”作為參數。

Select "**CpG-score**" from the **CpG-score** field shown in the lower right corner of the screen, and click **Show Chart 4**. This will display the distribution of CpG-scores for the target sequence region. Estimate the length of the longest candidate CpG island region found in this sequence, and choose the closest value from the following options. In this question, regions with a CpG score greater than 0.6 are considered to be candidate CpG islands. Note that this **calculation may take some time** (about one minute per calculation) in IBOC applications. [6 marks] **[No. 24]**

從視窗右下角顯示的"**CpG-score**"字段中選擇“ **CpG-score**”，然後單擊“**Show Chart 4**”。這將顯示目標序列區域的 CpG 分數分佈。 估計在此序列中找到的最長候選 CpG 島嶼區

域的長度，然後從以下選項中選擇最接近的值。 在此問題中，CpG 分數大於 0.6 的區域被視為候選 CpG 島嶼。 請注意，在 IBOC 應用程序中，此計算可能需要一些時間（每次計算大約需要一分鐘）。 [6 分] [24 號]

1. 200 bp
2. 800 bp
3. 1,400 bp
4. 2,000 bp
5. 2,600 bp
6. 3,200 bp
7. 3,800 bp
8. 4,400 bp
9. 5,000 bp

Question 14. Then, predict the transcription start site of human ***HoxA6*** gene in the Human genome DNA 1. Consider the most 5'-end-most position among the search results (also called "hits") of this full-length sequence as the transcription start site, and also consider the most 3'-end position among the hits of this full-length sequence as the end of transcript. Choose the closest answer from the following options. For the *HoxA6* sequence, use the sequence shown in the "***Predicted mRNA sequences***" tab.
問題 14：然後，預測人類 HoxA6 基因在人類基因組 DNA 1 中的轉錄起始位點。將全長序列的搜索結果（也稱為"hits"）中最 5'端的位置視為轉錄起始位點，並將此全長序列的匹配片段中最 3'端的位置視為轉錄物的末端。 從以下選項中選擇最接近的答案。 對於 HoxA6 序列，請使用"***Predicted mRNA sequences***"選項中顯示的序列。

The transcription start site of *HoxA6* is nearest to nucleotide position **[No. 25]**. [4 marks]
HoxA6 的轉錄起始位點最靠近核苷酸位置**[No.25]** [4 分]。

1. 1,000
2. 2,000
3. 3,000
4. 4,000
5. 5,000
6. 6,000
7. 7,000
8. 8,000
9. 9,000

Question 15. Referring to **Questions** 13 - 14 above, choose the **incorrect** statement from the following options [6 marks] **[No. 26]**
問題 15：參見上面的問題 13-14，從以下選項中選擇不正確的陳述。[6 分] **[No.26]**

1. The CpG island, which overlaps with the first exon of *HoxA5*, is longer than another CpG island which overlaps with the first exon of *HoxA6*.

與 *HoxA5* 的第一個外顯子重疊的 CpG 島嶼比另一個與 *HoxA6* 的第一個外顯子重疊的 CpG 島嶼長。

2. The intron of *HoxA6* is roughly consistent with the region containing the fewest CpG dinucleotides.
   HoxA6 的內含子與含有最少 CpG 二核苷酸的區域大致一致。

3. In this sequence, the region with the lowest CpG-score is roughly consistent with the region with the highest AT content.
   在此序列中，CpG 評分最低的區域與 AT 含量最高的區域大致一致。

4. Both *HoxA5* and *HoxA6* have a GC content of >60% in the first exon.
   HoxA5 和 *HoxA6* 在第一個外顯子中的 GC 含量均> 60％。

5. The intergenic region between *HoxA5* and *HoxA6* genes is approximately 1.7 kb in length.
   HoxA5 和 *HoxA6* 基因之間的基因間隔區域的長度約為 1.7 kb。

In the following question, we will examine the characteristics of functional sequence motifs found on genomic DNA.
在下面的問題中，我們將檢查在基因組 DNA 上發現的功能序列結構模體的特徵。

DNA contains a variety of functional sequence motifs. A famous example is the polyadenylation signal (AAUAAA sequence) encoded by the transcribed RNA in its 3' untranslated region (UTR). On mRNA precursors, a group of protein complexes recognizes this polyadenylation signal and truncates the 3’-end of RNA before adding a poly(A) tail. In other words, we can expect to find a polyadenylation signal in the last exon at the 3'-end of a coding sequence.
DNA 包含各種功能序列結構模體。 一個著名的例子是轉錄的 RNA 在其 3'非翻譯區（UTR）中編碼的多腺苷酸化信號（AAUAAA 序列）。 在 mRNA 前體上，一組蛋白質複合物識別出該多腺苷酸化信號，並在添加 poly（A）尾部之前截斷了 RNA 的 3'端。 換句話說，我們可以期望在編碼序列的 3'末端的最後一個外顯子中找到一個聚腺苷酸化信號。

If the location of the motif is known, an analysis method called “sequence logo” enables statistical evaluation of the consensus sequence. As an example, let's use the sequence logo program to visualize the polyadenylation signal motif. The 3'-end of genes are shown in FASTA format below. The sequence of the polyadenylation motif is shown in capitals. Ten bases either side are shown in lowercase. (Same sequences are shown on the bottom of the ***Sequence Logo*** tab).
如果知道了結構模體的位置，則可以使用一種稱為”***Sequence Logo***”的分析方法對共有序列進行統計評估。 例如，讓我們使用”***Sequence Logo***”程序可視化多腺苷酸化信號結構模體。 基因的 3'-末端以以下 FASTA 格式顯示。 多腺苷酸化結構模體的序列以大寫

字母表示。 小寫顯示了兩側的十個鹼基。 （相同的序列顯示在“序列徽標”選項的底部）。

>HoxA5_1638bp_1663bp
tggaacaaaaAATAAActttctattg
>CBX3_2027bp_2052bp
acagttgggaAATAAAagtttcatgt
>HNRNPA2B1_3591bp_3617bp
ggctgtccccAATAAAtgctgttcat
>Stk31_3238bp_3263bp
ggttgtgaaaAATAAAgatgtttggc
>Tra2a_1752bp_1777bp
agtagtctcaAATAAAagctaatttc

The figure below is a representation of the sequence alignment as a sequence logo (Figure 5).
下圖為序列比對結果的示意圖(Figure 5)

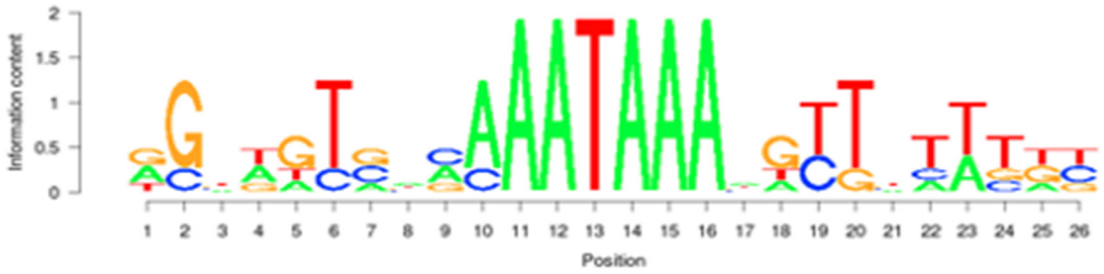

Figure 5. An example of a result of the sequence logo.

圖 5. An example of a result of the sequence logo

圖 5。 ”***Sequence Logo***”結果的示例

Using the ***Sequence Logo*** tab, make sure that you actually get the same diagram as above. (Copy the above sequences in the ***Sequence Logo*** tab and paste those into the "Input Sequence" window in the Sequence Logo tab, and then click on the “Exec Sequence Logo”).
使用”***Sequence Logo***”選項，確保您實際上獲得了與上述相同的圖表。 （將上述序列複製到”***Sequence Logo***”選項中，並將其粘貼到“序列徽標”選項的"Input Sequence"窗口中，然後單擊“Exec Sequence Logo”）。

Thus, the AATAAA sequence, a common polyadenylation-signal motif in all five sequences, is highlighted on the sequence log as you can see. In this example, only five gene sequences were used, but in the actual discovery of motifs, many more sequences need to be analyzed. Otherwise, you run the risk of highlighting sequences that just happen to be coincidental matches.
因此，如您所見，AATAAA 序列是所有五個序列中常見的多腺苷酸化信號結構模體，在序列日誌中特別顯示。 在這個例子中，僅使用了五個基因序列，但是在實際發現結構模

體中，需要分析更多的序列。 否則，您可能會特別顯示出恰好是巧合配對的序列。

Question 16. In eukaryotes, protein-coding genes consist of several exonic and intronic sequences on a genomic DNA. It is known that the boundary between exons and introns contains a distinctive sequence that allows the spliceosome (the enzyme that cuts out the intron) to recognize the intron sequences to be spliced. In addition to their commonality within eukaryotes, they also have species-specific and taxonomic specificity. The *HNRNPA2B1* gene has a typical exon-intron boundary sequence motif in the human genome. Determine the correct motif of exon-intron boundary and intron-exon boundary.You can use the information of the *HNRNPA2B1* gene in the GenBank format in the ***Genes 1*** tab, mRNA sequence in the ***Predicted mRNA sequences*** tab. Then choose the most appropriate answer from the following options. [13 marks]

問題 16：在真核生物中，具蛋白質編碼基因是由基因組 DNA 上的多個外顯子和內含子序列組成。 已知在外顯子和內含子之間的邊界線具有一個獨特的序列，該序列使剪接體（可切除內含子的酶）能夠識別需要剪接的內含子序列。它們除了在真核生物中具有的共通性之外，它們還具有物種專一性和分類學上的特殊性。 *HNRNPA2B1* 基因在人類基因組中具有典型的外顯子-內含子邊界線序列結構模體。確定外顯子-內含子邊界線和內含子-外顯子邊界線的正確結構模體。 您可以在“***Genes 1***”選項中使用 GenBank 格式的 *HNRNPA2B1* 基因訊息，在“***Predicted mRNA sequences***”選項中使用 mRNA 序列。然後從以下選項中選擇最合適的答案。[13 分]

Note that these represent the 4 bp motif (2 bp at the 3'-end of the exon and 2 bp at the 5'-end of the intron) that make up the main exon-to-intron boundary in human ([No. 27 – No. 30]), and the 3 bp motif (2 bp at the 3'-end of intron and 1 bp at the 5'-end of exon) that makes up the major intron-to-exon boundary in human ([No. 31 – No. 33]).
注意：這些代表形成人體最主要的外顯子至內含子邊界線的 4 bp 結構模體（外顯子 3'端 2 bp，內含子 5'端 2 bp） ([No. 27 – No. 30]) ，和 3 bp 的結構模體（內含子 3'末端 2 bp，外顯子 5'末端 1 bp）構成了人類主要的內含子與外顯子邊界線（ [第 31 號-第 33 號]）。

[Exon-side bases] - [Intron-side bases]

[[No.27], [No.28]] - [[No.29], [No.30]]

[Intron-side bases] - **[Exon-side base]**

[[No.31], [No.32]] - **[[No.33]]**

1. a
2. t
3. g
4. c

eg) If you would like to answer " **…exon… at** - gg …intron... ca - **t … exon …**", the answer should be as follows:
例如）如果您想回答"…外顯子……在-gg…內含子... ca-t…外顯子……"，答案應如下：

No. 27 : 1
No. 28 : 2
No. 29 : 3
No. 30 : 3
No. 31 : 4
No. 32 : 1
No. 33 : 2

Part 3: Task

第三部分：課題

Answer the following questions using the IBOC application.
使用 IBOC 應用程式來回答以下問題

Using the ***BLAST*** tab and the ***Entry Database*** tab, a sequence-similarity search can be performed against the protein data set of the following 7 organisms; 5 animals (human, mouse, fruit fly, octopus, and starlet sea anemone) and 2 unicellular organisms (choanoflagellate. and fission yeast). By examining some of the cadherin family genes in these organisms, answer the following questions.
使用 ***BLAST*** 選項和 ***Entry Database*** 選項，可以對下列 7 種生物的蛋白質數據資料集進行序列相似性搜索； 5 種動物（人類，小鼠，果蠅，章魚和星海葵）和 2 種單細胞生物（領鞭毛蟲綱和裂殖酵母屬）。 經由檢查這些生物中的一些鈣黏著蛋白家族(cadherin family)基因，回答以下問題。

Question 17. *E-Cadherin* and *N-Cadherin* are known to be representative of the cadherin family of genes. Information on these two genes are shown in the ***Genes*** tab, ***Proteins*** tab, ***Predicted mRNA sequences*** tab, and ***Protein sequences 1*** tab. Many of the cadherin family genes in the above animals have a tandemly duplicated (=repeatedly lined up) domain. Both E-Cadherin and N-Cadherin repeat this domain at least five times or more. The domain is given by the Accession ID of Pfam-A: (i)[PF *****]. Identify the domain, use the e-value of less than 1e-5 for the hmmscan threshold. (Note: Ignore the numbers after the dot in Pfam Accession ID.) [4 marks]
問題 17：已知 E-鈣黏著蛋白和 N-鈣黏著蛋白代表鈣黏著蛋白家族的基因。 這兩個基因的訊息顯示在"Genes"選項，"Proteins"選項，"Predicted mRNA sequences"選項和"Protein sequences 1"選項中。 上述動物中的許多鈣粘蛋白家族基因具有連續複製的（= 重複排列的）結構域。 E-鈣黏著蛋白和 N-鈣黏著蛋白二者都重複此結構域至少五次或更

多。 該結構域由 Pfam-A 的註冊 ID 可以獲得：（i）[PF *****]。 要鑑定此結構域，請為 hmmscan 閾值使用小於 1e-5 的 e 值。（注意：忽略 Pfam 註冊 ID 中點號後的數字。）[4 分]

(i) PF **[No. 34][No. 35][No. 36][No. 37][No. 38]**

eg) If you would like to answer “PF00001.12”, the answer should be as follows:
例如）如果您想回答“ PF00001.12”，答案應該如下：

No. 34 : 0
No. 35 : 0
No. 36 : 0
No. 37 : 0
No. 38 : 1

Question 18. A comparison of the domain structure of cadherin proteins in animals and non-animals shows that, (ii)[PF ***** . *] is added to the (iii) **[No.44]**-terminus next to the repeat of the (i) domain in animals. (Note: Ignore the numbers after the dot in Pfam Accession ID.)
問題 18：經由比較動物界生物和非動物界生物中鈣粘蛋白的結構域的結構可以獲得，（ii）[PF *****。 *在動物中（i）結構域重複的旁邊（iii）[No.44]末端添加。 （注意：忽略 Pfam 註冊 ID 中點號後的數字。）

(ii) PF **[No. 39][No. 40][No. 41][No. 42][No. 43]**

(iii) **[No.44]**

1. N
2. C

Question 19. From the set of protein entries for the choanoflagellate, find one gene that has multiple repeats of the above domain (i) in Question 17. To identify the domain (i), use the e-value of less than 1e-5 for the hmmscan threshold. [7 marks]
(The more repeats of domain (i) that are in the gene that you find, the higher score you will get.)
問題 19：在問題 17 中，從一組領鞭毛蟲的蛋白質輸入資料中找到一個具有上述結構域多個重複的基因（i）。要鑑定出結構域（i），請使用小於 1e-5 的 e 值 hmmscan 閾值。[7 分]
（在您發現的基因中，域（i）的重複次數越多，得分越高。）
註冊號。）

Hint: You may use following functions in ***HMMSCAN-tab*** as needed." ‘**Show [number] entries**’ function can be used to change the number of displayed lines. '**Search: [word]**' can be used to perform a string [word] search. The total number of rows in the

resulting table is displayed at the bottom left of the screen.
提示：您可以根據需要在 ***HMMSCAN*** 選項中使用下列功能。“'顯示[數字]個輸入訊息'功能可用於更改顯示的行數。'搜索：[單詞]'可用於執行字符串[單詞]]搜索，結果表中的總行數顯示在視窗的左下方。

(iv) **[No.45] _ [No. 46][No. 47][No. 48][No. 49][No. 50]**

// End of Practical Exam 2.
實作題 2 到此結束

URL list

#	Participants	ID	URL for Practical Exam 2
1	Iran	11	13.127.129.209/ec2-user/ibo2020bi
2	Hungary	12	3.122.239.215/ec2-user/ibo2020bi
3	Japan	13	54.250.182.124/ec2-user/ibo2020bi
4	Armenia	15	13.127.129.209/ec2-user/ibo2020bi
5	Russia	16	3.122.239.215/ec2-user/ibo2020bi
6	Kazakhstan	17	13.127.129.209/ec2-user/ibo2020bi
7	Philippines	18	52.79.248.78/ec2-user/ibo2020bi
8	Indonesia	19	54.255.220.196/ec2-user/ibo2020bi
9	South Korea	20	52.79.248.78/ec2-user/ibo2020bi
10	Nepal	21	13.235.100.64/ec2-user/ibo2020bi
11	Sri Lanka	23	13.235.100.64/ec2-user/ibo2020bi
12	Bangladesh	24	13.234.37.5/ec2-user/ibo2020bi
13	Pakistan	25	13.234.37.5/ec2-user/ibo2020bi
14	Thailand	26	52.79.146.192/ec2-user/ibo2020bi
15	Vietnam	27	54.255.220.196/ec2-user/ibo2020bi
16	Singapore	29	54.255.232.154/ec2-user/ibo2020bi
17	China	30	52.79.146.192/ec2-user/ibo2020bi
18	Chinese Taipei	31	52.79.248.78/ec2-user/ibo2020bi
19	Hong Kong, China	32	54.255.232.154/ec2-user/ibo2020bi
20	Syria	34	13.127.123.61/ec2-user/ibo2020bi
21	Saudi Arabia	36	13.127.123.61/ec2-user/ibo2020bi
22	Finland	44	35.156.199.58/ec2-user/ibo2020bi
23	Denmark	47	35.156.199.58/ec2-user/ibo2020bi
24	Iceland	48	18.184.71.168/ec2-user/ibo2020bi
25	Estonia	49	18.184.71.168/ec2-user/ibo2020bi
26	Latvia	50	3.124.195.33/ec2-user/ibo2020bi
27	Lithuania	51	3.124.195.33/ec2-user/ibo2020bi
28	Kyrgyzstan	53	15.236.134.99/ec2-user/ibo2020bi
29	Tajikistan	54	15.236.134.99/ec2-user/ibo2020bi
30	Uzbekistan	56	15.188.75.25/ec2-user/ibo2020bi
31	Azerbaijan	59	15.188.75.25/ec2-user/ibo2020bi
32	Georgia	60	3.124.195.33/ec2-user/ibo2020bi

33	Czech Republic	61	3.127.214.51/ec2-user/ibo2020bi
34	Poland	63	3.127.214.51/ec2-user/ibo2020bi
35	Bulgaria	64	35.180.21.57/ec2-user/ibo2020bi
36	Slovenia	65	35.180.21.57/ec2-user/ibo2020bi
37	North Macedonia	69	15.236.239.75/ec2-user/ibo2020bi
38	Turkey	72	15.236.239.75/ec2-user/ibo2020bi
39	Netherlands	73	15.236.131.4/ec2-user/ibo2020bi
40	Belgium	74	15.236.131.4/ec2-user/ibo2020bi
41	Germany	75	18.185.106.176/ec2-user/ibo2020bi
42	Switzerland	77	18.185.106.176/ec2-user/ibo2020bi
43	Luxembourg	78	3.126.249.168/ec2-user/ibo2020bi
44	United Kingdom	82	35.178.172.104/ec2-user/ibo2020bi
45	United States of America	83	3.128.202.209/ec2-user/ibo2020bi
45	Australia	84	3.104.47.102/ec2-user/ibo2020bi
46	Turkmenistan	89	3.126.249.168/ec2-user/ibo2020bi
47	Croatia	66	35.180.21.57/ec2-user/ibo2020bi
48	Canada	81	3.128.202.209/ec2-user/ibo2020bi
49	France	93	15.236.131.4/ec2-user/ibo2020bi
50	Afghanistan	95	15.236.239.75/ec2-user/ibo2020bi
51	Norway	45	35.178.172.104/ec2-user/ibo2020bi
52	El Salvador / Ibero-America	92	3.128.202.209/ec2-user/ibo2020bi