バイオインフォマティクス技術者認定試験

2024 年度　過去問と解説

2024 年度試験までに出題された過去問のうち数問をピックアップして解説します。分野によっては出題の傾向と今後の学習に関するコメントもありますのでぜひ今後の試験対策としてご活用ください。

【配列】

（1）

以下に示すデータは、フラットファイル形式で記された RefSeq のエントリーである。ここから読み取れる情報として、もっとも不適切なものを選択肢の中から 1 つ選べ。

```
LOCUS       NC_045512            29903 bp ss-RNA     linear   VRL 30-
MAR-2020
DEFINITION  Severe acute respiratory syndrome coronavirus 2 isolate
Wuhan-Hu-1,
            complete genome.
ACCESSION   NC_045512
VERSION     NC_045512.2
DBLINK      BioProject: PRJNA485481
KEYWORDS    RefSeq.
SOURCE       Severe acute respiratory syndrome coronavirus 2 (SARS-
CoV-2)
  ORGANISM  Severe acute respiratory syndrome coronavirus 2
            Viruses;    Riboviria;    Orthornavirae;    Pisuviricota;
Pisoniviricetes;
            Nidovirales;        Cornidovirineae;        Coronaviridae;
Orthocoronavirinae;
            Betacoronavirus; Sarbecovirus.

                                 == 中略 ==

FEATURES             Location/Qualifiers
     source          1..29903
                     /organism="Severe   acute   respiratory   syndrome
coronavirus
                     2"
                     /mol_type="genomic RNA"
                     /isolate="Wuhan-Hu-1"
                     /host="Homo sapiens"
                     /db_xref="taxon:2697049"
                     /country="China"
                     /collection_date="Dec-2019"
     5'UTR           1..265
     gene            266..21555
```

```
                     /gene="ORF1ab"
                     /locus_tag="GU280_gp01"
                     /db_xref="GeneID:43740578"
     CDS             join(266..13468,13468..21555)
                     /gene="ORF1ab"
                     /locus_tag="GU280_gp01"
                     /ribosomal_slippage
                     /note="pp1ab;    translated    by    -1    ribosomal
frameshift"
                     /codon_start=1
                     /product="ORF1ab polyprotein"
                     /protein_id="YP_009724389.1"
                     /db_xref="GeneID:43740578"

                                 == 後略 ==
```

1. 2019 年 12 月に中国でヒトより分離されたウイルス（Severe acute respiratory syndrome coronavirus 2）のコンプリートゲノム配列である。
2. 266 塩基から 21555 塩基にかけてタンパク質をコードする遺伝子があり、この間の塩基配列を抜き出してそのまま翻訳することでそのアミノ酸配列を得ることができる。
3. 分類学的には、コロナウイルス科（Coronaviridae）のオルソコロナウイルス亜科（Orthocoronavirinae）のベータコロナウイルス属（Betacoronavirus）に属する。
4. ゲノムサイズは約 30 kb で、直鎖状の一本鎖 RNA である。

[解説]

各選択肢について正否を検討する。

1. Source の欄に/collection_date="Dec-2019"、/country="China"、/host="Homo sapiens"とあり、また DEFINITION の欄に Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome.とあるため、この選択肢は正しい。

2. CDS の欄に、/ribosomal_slippage および/note="pp1ab; translated by -1 ribosomal frameshift"とあることから、この領域はリボソームフレームシフト(コドンの読み枠が途中でずれる現象)を伴って翻訳される。よって塩基配列をそのまま翻訳しても正しいアミノ酸配列が得られるわけではなく、誤った選択肢である。

3. ORGANISM の欄に、Coronaviridae; Orthocoronavirinae;　Betacoronavirus;とあるため、この選択肢は正しい。

4. LOCUS の欄に 29903 bp ss-RNA linear とあるため、この選択肢は正しい。

よって選択肢 2 が間違っている。

（2）
以下の文章は、アミノ酸の配列類似性を評価するスコアについて述べたものである。もっとも不適切なものを選択肢から一つ選べ。

1. PAM や BLOSUM などのアミノ酸類似性スコア行列において、負の値をとるアミノ酸対は、相同なアラインメント中に出現する頻度が、ランダムな場合に期待される頻度と比べて小さいことを意味している。
2. 局所アラインメントを行う場合、一般に非相同な配列対に対しては最適アラインメントのスコアは負の値となる。
3. アミノ酸類似性スコア行列 PAM100 においては、アミノ酸 100 残基当たり 100 回の置換が起きている程度の進化距離にある相同配列を基準としてスコアを定義している。
4. PAM や BLOSUM などのアミノ酸類似性スコア行列において、アミノ酸対 A、B について、A に対する B のスコアと B に対する A のスコアはつねに等しい。

[解説]
各選択肢について正否を検討する。
1. スコア行列においてはスコアの定義は対数オッズ $\log(q_{ab}/p_a p_b)$に基づいている。ここで、q_{ab} はアラインメント中でアミノ酸 a とアミノ酸 b が縦に揃う頻度であり、p_a 及び p_b はアラインメントにおけるアミノ酸 a, b の出現頻度を意味する。スコアが負の値であるとは、$q_{ab} < p_a p_b$ であるから、ランダムな場合に期待される頻度に比べても q_{ab} が小さいことを意味する。よって選択肢は正しい。
2. 局所アラインメント(ローカルアラインメント)では、アラインメントスコアが必ず 0 以上になるため、負の値になることはない。よってこの選択肢は正しくない。
3. PAM1 は 100 残基あたり 1 回の頻度で置換が起きる時間での置換確率行列であり、それを 100 乗したものが PAM100 である。よって選択肢は正しい。
4. 上記のスコアの計算方法から鑑みて、この行列は対称行列となる。よって選択肢はただしい。

よって選択肢 2 が間違っている。

（3）
以下は、異なるスコアづけによって同じ配列のグローバルアラインメントを行なった結果を表している。各スコア定義において、より高いスコアを持つアラインメントの組み合わせとして、もっとも適切なものを選択肢の中から一つ選べ。

スコア(A)：一致 +2、不一致 –1、ギャップ開始 –6、ギャップ延長 –1
スコア(B)：一致 +2、不一致 –2、ギャップ開始 –2、ギャップ延長 –2

アラインメント 1
```
ACAGTCGAGCAT
ACAG--G-GCAT
```

アラインメント 2
```
ACAGTCGAGCAT
ACA---GGGCAT
```

1. (A) アラインメント 1　(B) アラインメント 1
2. (A) アラインメント 1　(B) アラインメント 2
3. (A) アラインメント 2　(B) アラインメント 1
4. (A) アラインメント 2　(B) アラインメント 2

[解説]
実際にアラインメントスコアを計算する。
スコア A でアラインメント 1 を計算すると、一致箇所: 9, 不一致箇所: 0, ギャップ開始: 2 箇所, ギャップ延長: 1 箇所であり、そのスコアは 9*2 -6*2-1*1 = 5 となる。続いてスコア A でアラインメント 2 を計算すると、一致箇所: 8, 不一致箇所: 1, ギャップ開始: 1 箇所, ギャップ延長: 2 箇所であり、そのスコアは 8*2 – 1*1 - 6*1-1*2 = 7 となる。よってスコア(A)ではアラインメント 2 の方が高いスコアとなる。
次にスコア B でアラインメント 1 を計算すると、9*2 -2*2-2*1 = 12 となり、アラインメント 2 を計算すると、8*2 – 2*1 - 2*1-2*2 = 8 となる。よってスコア(B)ではアラインメント 1 の方が高いスコアとなる。

以上より、選択肢 3 が正解である。

（4）

RNA-seq データを用いて転写産物の発現量を測定する際は、転写産物にマッピングされたリードの数を、転写産物の長さや総リード数で補正する必要がある。TPM (transcripts per million)はそのような補正方法の一つであり、次の式で計算される。ただし、C_iは転写産物iにマッピングされたリード数であり、L_iは転写産物iの有効配列長とする。

$$T_i = \frac{C_i}{L_i}$$

$$TPM_i = T_i \frac{10^6}{\sum_i T_i}$$

さて、RNA-seq 実験を行い、次表のデータが得られたとする（簡単にするため、転写産物は表中の 3 つのみとする)。

転写産物	マッピングされたリード数	有効配列長
X	2,000	200
Y	3,200	400
Z	1,000	500

このデータを解析した結果を述べた次の文中において、空欄(a), (b)に当てはまる語句の組み合わせとしてもっとも適切なものはどれか。選択肢の中から一つ選べ。

TPM がもっとも大きかった転写産物は(a)であり、その値は(b)である。

1. (a) X (b) 5×10^5
2. (a) Y (b) 5×10^5
3. (a) X (b) 4×10^5
4. (a) Y (b) 4×10^5

[解説]

誘導に従って TPM を計算する。まず各転写産物の T_i を計算すると、$T_x = 10$, $T_y = 8$, $T_z = 2$ となる。この総和は 20 であることから、各転写産物の TPM_i を計算すると、
$TPM_x = 5 * 10^5$, $TPM_y = 4 * 10^5$, $TPM_x = 1 * 10^5$ となる。

以上より、選択肢 1 が正解である。

(5)

次世代シークエンサから得られる短い配列を参照ゲノム配列上にマッピングする検索手法として、接尾辞配列（suffix array）や、それを発展させたより高速な手法として、バローズ・ホイーラー変換（BWT: Burrows-Wheeler Transform）後の文字列が用いられている。BWT では、接尾辞の代わりに 1 文字ずつ巡回シフトした文字列を用い、それらを辞書順にソートする。こうして得られたソート済み文字列の最後の文字だけをつなげたものが BWT 後の文字列となる。

以下の図は、文字列「ATGCATGA＄」の BWT を計算する過程を表している。ただし「$」は文字列の最後を示す記号である。この文字列の BWT 後の文字列として適切なものを選択肢の中から一つ選べ。

A	T	G	C	A	T	G	A	$
T	G	C	A	T	G	A	$	A
G	C	A	T	G	A	$	A	T
C	A	T	G	A	$	A	T	G
A	T	G	A	$	A	T	G	C
T	G	A	$	A	T	G	C	A
G	A	$	A	T	G	C	A	T
A	$	A	T	G	C	A	T	G
$	A	T	G	C	A	T	G	A

$	A	T	G	C	A	T	G	A
A	$	A	T	G	C	A	T	G
A	T	G	A	$	A	T	G	C
A	T	G	C	A	T	G	A	$
C	A	T	G	A	$	A	T	G
G	A	$	A	T	G	C	A	T
G	C	A	T	G	A	$	A	T
T	G	A	$	A	T	G	C	A
T	G	C	A	T	G	A	$	A

1. AGC$GTTAA【正解】
2. TGCATGA$A
3. $AAACGGTT
4. $ATGCATGA

[解説]
左図が元の文字列の巡回シフトした文字列を並べたものであり、右図がそれをソートしたものである。その最後の文字列(一番右端の文字列)をつないだものが BWT 後の文字列であるため、AGC$GTTAA が BWT 後の文字列となる。よって選択肢 1 が正しい。

(6)
マルチプルアラインメントに関する説明として、もっとも不適切なものを選択肢の中から一つ選べ。

1　逐次改善法は、計算済みのマルチプルアラインメントをランダムに２つのグループに分割し、グループ間で再アラインメントを行うことを反復することにより、アラインメントを改善していく手法である。
2　逐次改善法を用いる事で、誤って挿入されたギャップの位置を修正する事ができる。
3　累進法（ツリーベース法）は、あらかじめ計算した近似的な系統樹（案内木）に従って、類似性の高い配列のグループから順にペアワイズアラインメントを組み上げていく手法である。
4　逐次改善法で作成したアラインメントを、累進法（ツリーベース法）を使ってさらに改善していくのが効果的である。

[解説]
マルチプルアラインメントでは、まず選択肢 3 のように累進法によってマルチプルアラインメントを組み上げる。しかし累進法では、一度誤って挿入されてしまったギャップを修正することができないと言う欠点がある。このため、選択肢 2 にあるように逐次改善法を用いてマルチプルアラインメントを修正する。逐次改善法とは選択肢 1 に示された手順である。

よって累進法を行ってから逐次改善法を行うというのが正しい手順であり、選択肢 4 は順序が逆となっており、不適切である。

(7)
モチーフの保存性を評価する指標として、モチーフの情報量（単位はビット）を考える。モチーフの各位置 i における情報量 $R(i)$は、エントロピー$H(i)$の減少量として定義され、位置 i における各文字 a の出現割合 $f(i, a)$を用いて以下のように表される。ただし、0log0=0 とする。

$$R(i) = 2 - H(i) = 2 - \left\{ - \sum_{a=\{\mathrm{A,T,G,C}\}} f(i,a) \log_2 f(i,a) \right\}$$

これについて述べた以下の文章中で、(a)〜(c)に入る語の組み合わせとして最も適切なものを選択肢の中から一つ選べ。

情報量 $R(i)$は、その位置においてすべての配列が同一の文字であるときが（　a　）で、（　b　）ビットとなる。以下の配列アラインメントに見られる 4 文字のモチーフにおいて、情報量の総和は（　c　）ビットである。

配列アラインメント

```
Seq1  A A C T
Seq2  A T C T
Seq3  A A T T
Seq4  A T T T
```

1. (a) 最大　(b) 2　(c) 6
2. (a) 最大　(b) 2　(c) 5
3. (a) 最小　(b) 0　(c) 5
4. (a) 最小　(b) 2　(c) 8

[解説]
式の定義より、情報量が最大であるとはエントロピーが最小の時である。よって全ての配列が同一の文字である時に最も情報量が最大となる。この時値を実際に代入して計算すると、その情報量は 2 となる。

次にアラインメントの情報量の総和を考える。第一列と第四列は全ての配列が同一の文字列であるためその情報量は 2 である。第二列と第三列は、2 つの文字が 2 個ずつ現れており、その情報量を実際に計算すると 1 となる。よってその総和は、2*2+1*2 = 6 となる。
以上より、選択肢 1 が正答となる。

(8)
あるタンパク質のアミノ酸配列について UniProt データベースに対してホモロジー検索を行い、それに基づいて機能予測を行った。その方法として、もっとも不適切なものを以下の選択肢の中から一つ選べ。
1　最上位でヒットした遺伝子と比べて類似性はやや低かったが、モデル生物のホモログが検出されていたため機能推定に用いた。
2　検出されたホモログの中には、オーソログとパラログが含まれていたため、その中からオーソログと思われるものを参照し、機能推定に用いた。
3　アノテーションに付与されたエビデンスコードに基づきその信頼度を判定した。
4　カイト・ドゥーリトル(Kyte-Doolittle)のスコア行列を用いて配列の類似度を算出した。

[解説]
各選択肢について正否を検討する。
1. モデル生物のホモログはアノテーション情報が豊富であり、非モデル生物はアノテーションが十分ではないことが多い。このため、スコアがやや低くてもモデル生物のアノテーションを参照することは有用である。
2. 種分化に伴うホモログであるオーソログは遺伝子機能が保存されている可能性が高いが、遺伝子重複に伴うホモログであるパラログは、遺伝子重複に伴い機能が分化している可能性が高い。このため、遺伝子機能推定にはオーソログを利用することが一般的である。
3. アノテーションには、実際に実験的な検証が行われているものから、計算機の予測にとどまるものまで含まれているため、そのエビデンスコードに基づいて信頼度を判定することは重要である。
4. カイト・ドゥーリトル(Kyte-Doolittle)のスコア行列とは、アミノ酸残基の疎水性を調べるための行列であり、配列の類似性検索に用いることはできない。

以上より、選択肢 4 が不適切な選択肢となる。

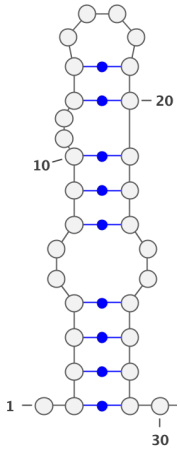
(9)
長さ 30 塩基からなる RNA の配列があり、各塩基に順に 1 から 30 までの番号がつけられている。その 2 次構造は、以下の 9 つの塩基対によって形成されているとする。

2–29	9–22
3–28	10–21
4–27	13–20
5–26	14–19
8–23	

このとき、この二次構造について述べた以下の記述のうち、もっとも不適切なものを選択肢の中から一つ選べ。

1　塩基 2 から 5 まではステムを形成している。
2　塩基 11 から 12 まではバルジループを形成している。
3　塩基 15 から 18 まではヘアピンループを形成している。
4　塩基 20 から 21 まではシュードノットを形成している。

[解説]
この二次構造を可視化すると次の通りとなる。



この二次構造にはシュードノット構造は存在しないため、4 は不適切である。一方、他の選択肢は正しい。

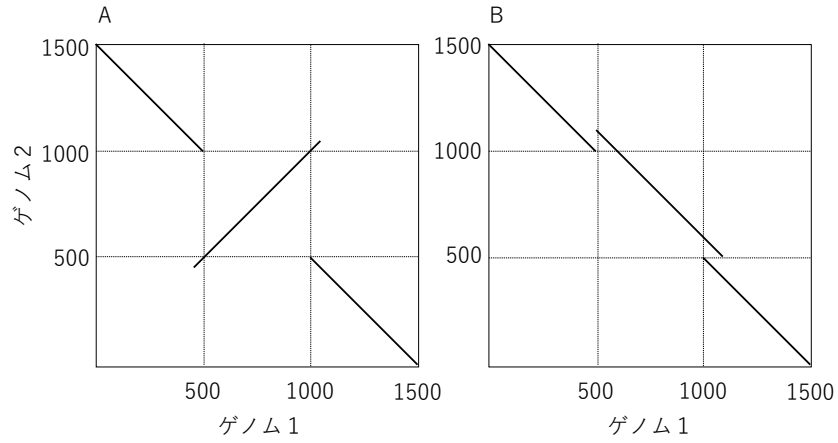
(10)
２本の近縁ゲノム DNA の配列断片を、BLAST を用いて比較したところ、以下の表に示す３つの相同領域が見つかった。ただし、表の各行の４つの数字は、それぞれアラインメントにおける配列１の開始位置、終了位置、配列２の開始位置、終了位置を表し、開始位置が終了位置より大きい場合は逆鎖とヒットしたことを意味する。

	ゲノム１	ゲノム２
領域１	1〜500	1500〜1001
領域２	461〜1040	461〜1040
領域３	1001〜1500	500〜1

これについて述べた以下の文章の(a)〜 (c)に入る語句の組み合わせとしてもっとも適切な

ものを選択肢の中から一つ選べ。

この結果を、ドットプロットを用いて表すと、下図の(　a　)の図が得られる。この中には大きな構造変化として(　b　)が見られ、それに付随する構造として、各ゲノム上に（　c　）がある。



1. (a) A　(b) 逆位　(c) 逆方向反復配列（inverted repeat）
2. (b) B　(b) 重複　(c) 逆方向反復配列（inverted repeat）
3. (a) A　(b) 逆位　(c) 直列反復配列（direct repeat）
4. (a) B　(b) 重複　(c) 直列反復配列（direct repeat）

[解説]
領域 1 及び 3 は逆鎖とヒットしており、領域 2 は順鎖とヒットしていることから、得られるドットプロットは A であることがわかる。またこのような事象を逆位と呼ぶ。

各ゲノムには複数箇所とアラインメント場所がある。たとえばゲノム 1 の 461~500 はゲノム 2 の 461~500 とアラインメントされる一方、ゲノム 2 の 1040-1001 ともアラインメントされる。これは反復配列であるが、向きが逆であるため、逆方向反復配列である。

これらの結果から、選択肢 1 が正しい答えとなる。

【構造】

1）

電子顕微鏡は試料に電子線を当てることで、試料の拡大像を取得する装置である。近年では検出器等の高性能化によって高精細な画像が得られるようになってきているものの、生体高分子の立体構造解析は容易ではない。この要因のひとつとして、三次元再構成の問題がある。電子顕微鏡で直接得られるデータはあくまで二次元像であり、立体構造を得るためにはこの二次元像から三次元像を復元する必要がある。高精細な三次元像を得るための様々な方法についての記述として、もっとも不適切なものを選択肢の中から一つ選べ。

1. 試料中の原子間の距離情報に基づいてエネルギー関数を定め、安定な三次元構造を計算によって探索する。
2. 試料を電子顕微鏡の装置内で回転させながら撮影し、コンピュータで三次元像を再構成する。
3. 解析対象の粒子を試料中に多数分散させて撮影することで、検出器に対して様々な方向を向いた粒子の像を多数得ることができる。画像から粒子ひとつひとつを検出して各像の方位を推定することで、三次元像を再構成する。
4. 類似した構造を持つと思われる試料の立体構造データが既に存在する場合は、その三次元像を参照することで、二次元像がどの方位から撮影されたものかを推定できる。多数の方位から得た二次元像についてこれを繰り返すことで、三次元像を再構成する。

解説：選択肢 2 は電子線トモグラフィー と呼ばれる方法の説明であり、単一粒子から複数の傾斜像を得る事で三次元構造を再構成する方法である。選択肢 3 は単粒子解析の説明であり、現在もっとも多用される三次元像再構成の方法の説明である。選択肢 4 は二次元の単粒子像を取捨選択し、三次元像の再構成を容易にする場合に用いられることのある手法である。従ってこれらの選択肢は、高精細な三次元像を得るための手法として適切である。一方、選択肢 1 は三次元再構成像が得られたのちに、分子モデルを構築・精密化する方法について述べているため、この 4 つの選択肢の中では最も不適切である。従って選択肢 1 が正解である。

2)

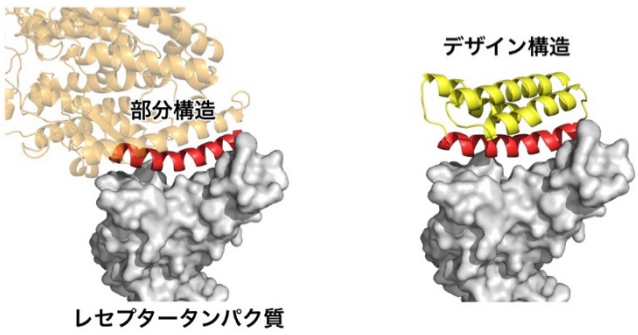
望みの立体構造や機能を持つタンパク質を人工的に設計（デザイン）しようとする試み、特に、コンピュータを用いた合理的デザインの試みは、1990 年代後期から活発化してきた。近年では機械学習の発展も相まって、さらに効率的な合理的デザインが可能となってきている。コンピュータによる合理的タンパク質デザインについて記述した以下の文章のうち、もっとも不適切なものを選択肢の中から一つ選べ。

1. 一般に、目標とする構造や機能を示すアミノ酸配列は単一ではなく、複数種類のアミノ酸配列があり得る。
2. デザインした立体構造の安定性の評価に用いられるエネルギー関数や最適化手法などの要素技術は、立体構造予測や分子動力学などで用いられるものと多くの共通点を持つ。
3. 膜タンパク質や複合体タンパク質をターゲットとする合理的デザインも行われている。
4. 分子量が大きいほど安定なフォールドを作りやすく、意図したフォールドを有するタンパク質のデザインが容易となる。

解説：一般に、あるフォールドや分子機能を持つタンパク質のアミノ酸配列は多数存在するので選択肢 1 は正しい。分子の構造安定性を評価する方法は分子動力学法などの *ab initio* 法が持ちられることが多いので選択肢 2 も正しい。単一サブユニットのタンパク質分子に比べると難易度は高くなるものの、膜タンパク質や複合体タンパク質のデザインも行われているので選択肢 3 も正しい。一方、方法にもよるが、タンパク質の安定性評価は残基間の組み合わせで評価する場合が多いので、通常は分子量が大きい(残基数の多い)タンパク質の方が安定性評価は困難になるので、選択 4 は不適切であり、これが正解である。

3)
あるレセプタータンパク質に結合するような小さな新規タンパク質をデザインすることを考える。以下の左図に示した既知の結合タンパク質には、レセプターと相互作用するαヘリックス様の部分構造(赤色)が存在している。これと類似した部分構造を持つタンパク質をデータベースから探索することでデザインの起点としたい(例えば右図)。この時、図に示したような α ヘリックス様の部分構造を有するタンパク質を検索する具体的な操作として、以下のような処理が考えられる。

a)検索を行う立体構造データベースとして(A)を用いる。
b)データベース内の全ての構造について手法(B)を用いて二次構造を決定し、必要な部分構造に相当する十分な長さの α ヘリックスを持たない対象を候補から除外する。
c)プログラム(C)を用いて構造アラインメントを行い、既知結合タンパク質と類似性の高い部分を含む立体構造を同定する。



上記の処理に現れる(Ａ）～（Ｃ)の組み合わせとしてもっとも適切なものを選択肢の中から一つ選べ。|

1. （Ａ）PDB　（Ｂ）STRIDE　（Ｃ）PSI-BLAST
2. （Ａ）PDB　（Ｂ）DSSP　（Ｃ）DALI
3. （Ａ）UniProt　（Ｂ）STRIDE　（Ｃ）PSI-BLAST
4. （Ａ）UniProt　（Ｂ）チョウ・ファスマン法　（Ｃ）TM-align

解説：選択肢に示されている語句を簡単に解説すると以下のようになる。
PDB：タンパク質などの生体分子の立体構造データベース
UniProt：タンパク質のアミノ酸配列のデータベース
DSSP：タンパク質立体構造上で二次構造を同定するアルゴリズム。主に水素結合パターンを評価する古典的な方法
STRIDE：タンパク質立体構造上で二次構造を同定するアルゴリズム。水素結合パターンと二面角を評価するより新しい方法
チョウ・ファスマン法：アミノ酸配列から二次構造を予測するアルゴリズム。アミノ酸毎の二次構造傾向を評価する古典的方法
PSI—BLAST: 塩基･アミノ酸配列のアライメントを計算するアルゴリズム BLAST のうち、位置特異的スコアマトリックス(position-specific scoring matrix, PSSM）を利用する方法
DALI：タンパク質立体構造の類似性を探索するアルゴリズム。クエリ構造に対してタンパク質立体構造のデータベースを検索する方法
TM-align: 2 つのタンパク質立体構造の最適構造重ね合わせを探索するアルゴリズム
　これを踏まえて問題文を解釈すると、(A)は立体構造データベース、(B)は二次構造同定、(C)は構造類似性探索の方法であると考えられるので、選択肢 2 が正解である。

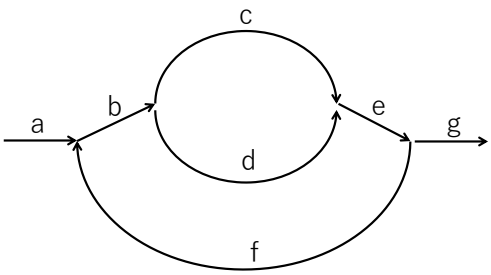
るスコアについて述べたものである。もっとも不適切なものを選択肢から一つ選べ。

5. PAM や BLOSUM などのアミノ酸類似性スコア行列において、負の値をとるアミノ酸対は、相同なアラインメント中に出現する頻度が、ランダムな場合に期待される頻度と比べて小さいことを意味している。
6. 局所アラインメントを行う場合、一般に非相同な配列対に対しては最適アラインメントのスコアは負の値となる。
7. アミノ酸類似性スコア行列 PAM100 においては、アミノ酸 100 残基当たり 100 回の置換が起きている程度の進化距離にある相同配列を基準としてスコアを定義している。
8. PAM や BLOSUM などのアミノ酸類似性スコア行列において、アミノ酸対 A、B について、A に対する B のスコアと B に対する A のスコアはつねに等しい。

【正解】2

【解説】局所アラインメントは、スコアが最大となるような部分配列間のアラインメントを求める。その際、スコアが負になるような部分アラインメントは捨てられることになる。このため、非相同な配列対の場合でも、アラインされる部分配列の長さが短く（従ってスコアが小さく）はなるが、スコアが負になることはない。

ある 1 本の DNA 配列から、十分長い k-mer を用いて de Bruijn グラフを作成すると、下図のような構造のグラフが得られた。このグラフから、繰り返し配列が存在するといえるのはどの部分か。もっとも適切なものを選択肢から一つ選べ



1. c, d
2. b, c, e, f
3. b, e
4. f

【正解】3

【解説】DNA 配列解析に用いられる de Bruijn グラフは、与えられた配列中に含まれるすべての k-mer に対して、配列上で隣接する（k-1 文字が重なる）k-mer どうしをつないで作ったグラフ（図では複数の k-mer が直鎖状につながった構造を一本の線で表してある）で、１本の配列に由来する de Bruijn グラフであれば、すべての辺を通るように一筆書きで辿ることで元の配列を再構成できる。その際、元の配列に（k 文字を超える長さの）繰り返し配列がなければ、同じ辺を 2 度通らないという狭義の一筆書きで辿ることができるが、繰り返し配列がある場合はその部分を複数回通る必要がある。問題のグラフを a から g まで一筆書きで辿ると、b と e の辺は必ず 2 回以上通る必要があるため、ここには繰り返し配列があることになる。なお、b, c, e, f のようなサイクルを繰り返し辿ることもできるが、繰り返さないこともできるため、このグラフだけから繰り返しがあるとはいえない。

【遺伝進化】

タンパク質をコードする遺伝子の進化の過程における塩基置換は、アミノ酸を変化させない同義置換と変化させる非同義置換に分類される。同義置換率(Ks)は同義サイトあたりの置換数、非同義置換率(Ka)は非同義サイトあたりの置換数である。この 2 つの置換率に関する以下の記述のうち、もっとも不適切なものを選択肢の中から一つ選べ。

1. Ka + Ks は同義・非同義置換を区別しない広義の塩基置換率に等しい。
2. Ka/Ks < 1 が純化淘汰を受ける大多数の遺伝子が示す値である。
3. Ka/Ks がほぼ 1 である場合、その遺伝子が偽遺伝子であることが疑われる
4. Ka/Ks > 1 である場合、その遺伝子は正の選択圧を受けている可能性があると解釈される。

解説

Ka，Ks はそれぞれ非同義置換数を非同義サイト数で割ったもの，同義置換数を同義サイト数で割った値である. 多重置換を補正するために, 根井・五條堀の方法や, Li-Pamilo-Bianchi の方法などがある．尤度を用いたコドン置換モデルの場合は ω というパラメータで表現されることが多い. 同義置換が中立であり，その遺伝子領域における突然変異率に等しいと仮定すると，非同義置換率はその遺伝子にはたらく自然選択の強さを知るための指標となる. 非同義サイト数, 同義サイト数は塩基配列によって異なり，多くの配列では非同義サイト数の方が同義サイトよりも多い．したがって，Ka と Ks を計算するときの分母はそれぞれ異なる. 分母が異なった割合同士をそのまま足し合わせることはできない. 問題はもっとも不適切なものを選ぶことを求めているので，選択肢 1 が明らかに誤っている.

その他の選択肢はすべて妥当であり，正しい解釈である．純化淘汰（負の自然選択）は，有害な変異が取り除かれる過程であり，正の自然選択は生存などに有利な変異が急速に集団に固定する仮定である. 遺伝子が機能をもっていない偽遺伝子の場合は，非同義置換は同義置換を同じ速さで起こると考えられるので，Ka/Ks の値は 1 に近くなる.

【オーミクス】

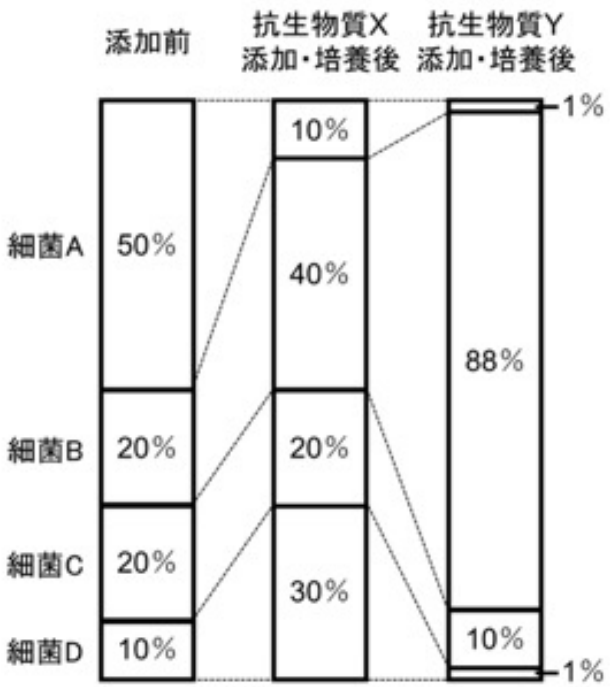
細胞に含まれる生体分子を一細胞単位で計測し解析する手法は、一細胞（シングルセル）解析と呼ばれる。この手法に関して述べた以下の文章のうち、もっとも不適切なものを選択肢の中から一つ選べ。

1. 一細胞単位でのトランスクリプトーム解析(scRNA-seq)に活用される技術であり、ゲノムは解析対象とはされていない。
2. 一細胞に含まれる計測対象の物質量は微量であるため、解析対象外の細胞に由来する物質の混入は、取得データの品質低下を招く要因となりうる。
3. 由来する生物種や組織によって細胞サイズは大きく異なるため、利用する計測手法で解析対象の細胞群を解析可能かどうかは、事前に検討する必要がある
4. 微量の生体分子を高感度で検出する技術や、多数の一細胞をハイスループットに処理する技術について、研究開発が進められている。

【正解 1】

オーミクス研究では、生物組織や微生物の細胞集団を一括して（バルクで）解析することが多い。一方で近年、1 細胞単位の解像度でデータを得る、いわゆるシングルセル解析技術が飛躍的に発展しており、関連するオーミクス解析が幅広く展開されている。例えば、1 細胞単位でゲノム配列を決定するシングルセルゲノム解析や、転写産物を解析するシングルセルトランスクリプトーム解析、代謝産物を解析するシングルセルメタボローム解析などが挙げられる。そのため選択肢 1 は誤りである。他の選択肢は正しい。

ある細菌叢の試料に対して 2 種類の異なる抗生物質を X, Y の順に添加して培養し、メタゲノム解析によって菌叢構造の変化を解析したところ、図のような結果になった。また、細菌数の測定をしたところ、各時点で試料の全細胞数は変化がほぼ無かった。これらの結果から考察できることとして、もっとも不適切なものを選択肢の中から一つ選べ。



1. 細菌 A は、抗生物質 X と抗生物質 Y により大半の細胞が死滅した。
2. 細菌 B は、抗生物質 X と抗生物質 Y の両方に耐性を持っている。
3. 細菌 C は、抗生物質 X に対して細菌 B とは異なる耐性機構を持っている。
4. 細菌 D は、抗生物質 X には耐性を持つが、抗生物質 Y には持たない。

【正解 3】

原核生物は抗生物質に対抗するために、様々な種類の耐性機構を持つ。例えば、抗生物質を細胞内で分解する機構、抗生物質を細胞膜外へ排出する機構、抗生物質が作用しないように DNA ポリメラーゼなどの在来酵素の構造を変化させる機構、バイオフィルム形成や細胞外膜の変化により抗生物質の細胞内浸透を抑制する機構、などが知られている。本設問の実験では、細菌 B は抗生物質 X や抗生物質 Y で処理した場合でも相対存在量が低下せず、むしろ増加していることがグラフから読み取れる。また、細菌 C は抗生物質 Y で処理した際は相対存在量が若干低下しているが、細菌 A や細菌 D ほどの極端な低下は見られない。このことから、細菌 B と細菌 C は共に、程度の差がはあれど、抗生物質 X・Y に対する耐性を持つ可能性が考えられる。ただし、この実験のみからその具体的な耐性機構を推測することは困難であり、その機構が両者で同じか異なるかは判断できない。そのため選択肢 3 が正解である。

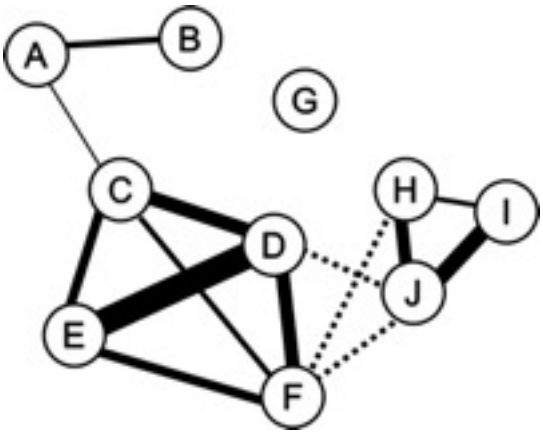
代謝のフラックス解析では、同位体標識した代謝物を用いて細胞内の物質変換の流れを追跡する。この解析について、もっとも不適切な説明を選択肢の中から一つ選べ。

1. 放射性同位体で標識すると精度は高くなるが、安全性や扱いの煩雑さから敬遠されている。
2. 代謝物を変数として量の増減を質量保存則に沿って記すと一次方程式になる。
3. ゲノムスケールのフラックス解析には、線形計画法がよく用いられる。
4. カーネル関数を用いて高次元に射影することで、重要成分を見出せる。

【正解 4】

カーネル関数を用いてデータを高次元に射影する手法はカーネル法と呼ばれ、サポートベクターマシン（SVM）や主成分分析など様々な手法と組み合わせて応用できる。これは代謝フラックス解析とは直接関係が無い手法であるため、選択肢 4 が誤りである。他の選択肢は正しい。

ある多数の観測データを用いて、各生物の存在量がペア間で相関するかどうかを調べる、いわゆる共起（co-occurrence）ネットワーク解析を実施したところ、以下のような結果を得られた。この結果の考察として最も不適切なものを選択肢の中から一つ選べ。なお、ノードは生物種、エッジは共起関係を表し、線の太さは共起の強さを、棒線は順相関を、点線は逆相関を示す。



1. A と B は、例えば生息環境を共有するような関係を持つことが予想される。
2. C,D,E,F は、例えば互いの代謝産物を与え合うような共同体的な関係を持つことが予想される。
3. G と H は、例えば両者が同じ餌を食べるといった競争的な関係を持つことが予想される。
4. D と J は、例えば一方がもう一方を駆逐するような排他的な関係を持つことが予想される。

【正解 3】

共起（co-occurrence）ネットワーク解析はノード間の相互作用を推定する手法の一つである。例えば、テキストデータから単語間の共起関係（同一文章で共に使われる頻度など）を可視化することでその関係性を調べたり、16S rRNA アンプリコン解析やメタゲノム解析などの微生物叢研究において微生物系統間の生態学的関係性を推定する際などに応用される。本文は、後者のような生態学的な文脈で実施した解析の解釈を問う設問である。例えば、相利共生（mutualism）や片利共生（commensalism）の関係を持つペア間では、同じ生息地（サンプル）で両者同時に観測されやすくなる、すなわち正の相関が見られることが期待される。逆に、片害共生（amensalism）や捕食―被捕食といった関係性を持つ場合、一方が増えれば他方が減る、すなわち逆相関が観測される可能性がある。本文の図では、ノード G と H にはエッジが引かれておらず、明確な共起関係は見出されていないため、ここから競争（competition）の関係性を類推することはできない。そのため選択肢 3 が正解である。