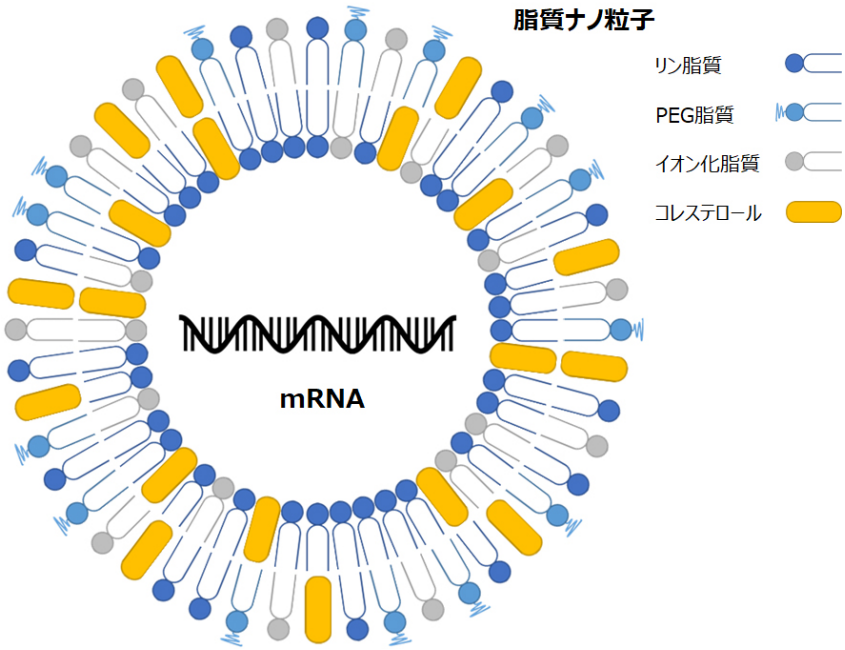
バイオインフォマティクス技術者認定試験

2023 年度　過去問と解説

2023 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。

【生命科学】

次の図は mRNA ワクチンの構造を示したものである。mRNA にはウイルスのスパイクタンパク質の遺伝子配列が組み込まれており、ワクチンによって体内で擬似的なウイルス感染が生じることで免疫応答を誘導すると考えられている。mRNA ワクチンの作用機序ついて述べた以下の記述について、もっとも不適切なものを選択肢の中から一つ選べ。



1. ワクチン接種により mRNA がヒト細胞に取り込まれると、細胞質で翻訳が行われ、リボソーム上で mRNA からウイルス抗原となるタンパク質が合成される。
2. mRNA ワクチンによるウイルス抗原の産生には、核内での転写の過程が必要ない。
3. mRNA ワクチンにより産生されたウイルス抗原は、細胞内でヌクレオチドに分解され、MHC クラス I 分子によって細胞表面に提示される。これを認識したキラーT 細胞が活性化し、ウイルスに対する細胞性免疫が誘導される。
4. 細胞外に出たウイルス抗原は抗原提示細胞に捕捉、貪食され、プロセシングを受けて MHC クラス II 分子によって細胞表面に提示される。これを認識したヘルパーT 細胞が活性化し、ウイルスに対する体液性免疫が誘導される。

【正解 3】

mRNA によって作り出されるのはタンパク質（ペプチド）なため、ウイルス抗原もペプチドである。ペプチドはアミノ酸から構成されているため、ヌクレオチドには分解されない。

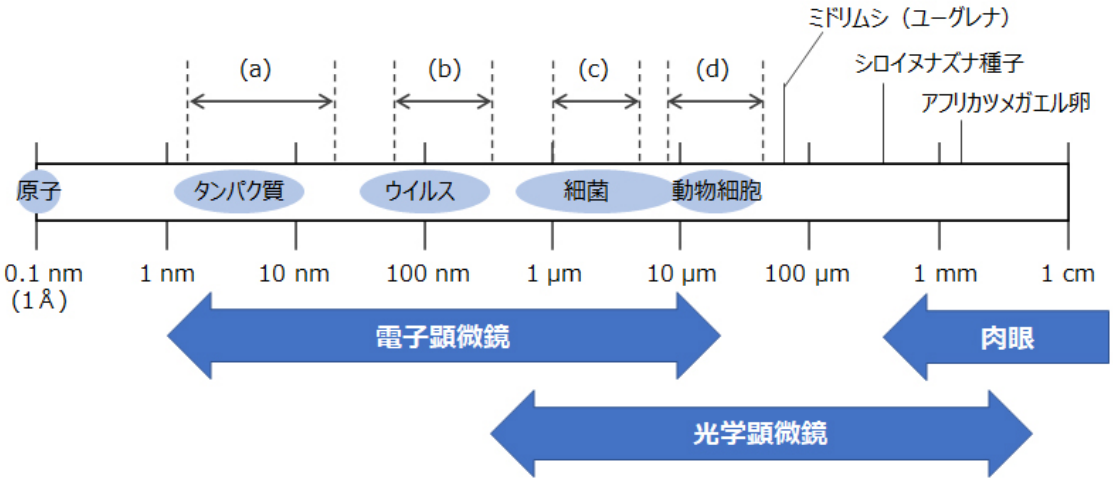
PCR(Polymerase Chain Reaction)法に関する以下の記述のうちで、もっとも不適切なものを選択肢の中から一つ選べ。

1. ごく微量の試料から DNA を増幅することが可能である。
2. 高熱環境に生息するバクテリアから単離された耐熱性の DNA ポリメラーゼがよく利用される。
3. 反応液の組成を変化させることで変性、アニーリング、伸長の 3 つのステップを繰り返し増幅を行う。
4. 増幅反応過程の DNA 量を逐次検出することで、サンプル中の標的 DNA 量を定量することが出来る。

【正解 3】

PCR は、反応液の組成ではなく温度を変化させることで、変性、アニーリング、伸張の核反応を行う。

次の図は、顕微鏡で観察できる対象の一般的な大きさを示したものである。(a)～(d) の範囲内に入るものの組み合わせとして、もっとも不適切なものを選択肢の中から一つ選べ。



1. (a) 抗体、GFP（green fluorescent protein）、ヘモグロビン
2. (b) SARS-CoV-2、マクロファージ、エクソソーム
3. (c) カビ胞子、ミトコンドリア、大腸菌
4. (d) 白血球、スギ花粉、ヒト ES 細胞

【正解】2

マクロファージは、白血球の一種であり、15～20μm の大きさであるため不適切である。

【配列】

（キーワード：次世代シークエンサーのマッピング）

以下は、次世代シークエンサを用いて得られたリシークエンスの配列データを用いて、リファレンス配列と照合して変異情報を抽出する際の処理の流れを表しており、A〜D にはデータフォーマット、x〜z には処理内容として、それぞれ以下の用語リストに挙げられた用語のいずれかが入るものとする。このとき、A, y, C に入る用語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。

[　A　]形式 →(　x　)→[　B　]形式 →(　y　)→[　C　]形式 →(　z　)→[　D　]形式

用語リスト

データフォーマット：　BAM, FASTQ, SAM, VCF

処理内容：　バイナリ化、バリアントコール、マッピング

（選択肢）

1. A. FASTQ,　y. バイナリ化,　C. BAM
2. A. FASTQ,　y. マッピング,　C. SAM
3. A. SAM,　y. バリアントコール,　C. VCF
4. A. SAM,　y. マッピング,　C. VCF

【正解】 1

【解説】よく使われるバイオインフォマティクスのツールにおいては、入出力ファイルに共通のフォーマットが使われることが多く、これにより同一の処理を行う異なるツールを適宜組み合わせて一連の処理を実施することが可能になっている。本問の穴埋めを完成させると以下のようになる。FASTQ 形式→マッピング→SAM 形式→バイナリ化→BAM 形式→バリアントコール→VCF 形式

次世代シークエンサから得られたショートリードをゲノムにマッピングする際の索引技術として、接尾辞配列が広く用いられている。対象とするゲノム配列の長さを N とした時、接尾辞配列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. 接尾辞配列とは文字列の接尾辞の開始位置を要素とする配列であり、その順序は接尾辞に関して辞書順に並び替えたものである。
2. 一度接尾辞配列を構築すれば、接尾辞配列を再構築することなく異なる k に対する k-mer の出現位置を検索することが可能である。
3. 接尾辞配列は $O(\log N)$ の時間計算量で構築することが出来る。
4. 接尾辞配列を利用することで、k-mer の文字列の検索を $O(k \log N)$ で行うことが可能である。

【正解】3

【解説】長さ N の文字列の接尾辞配列は、N 個ある接尾辞をソートして構築する。このため、単純な方法では $O(N^2 \log N)$ の時間計算量が必要となる（$O(N)$ の文字列比較を $O(N \log N)$ 回行う）が、より工夫された方法で $O(N)$ の時間で構築するアルゴリズムも存在する。しかし、文字列のごく一部のみを見てソートすることはできないため、$O(\log N)$ の時間計算量で構築することはできない。

原核生物と比べて真核生物のゲノム配列から遺伝子を予測するのは難しいため、いくつかの方法を組み合わせて予測を行うことが多い。以下の方法のうち、真核生物のゲノム配列から遺伝子を予測する方法としてはもっとも不適切なものを選択肢の中から一つ選べ。

1. RNA-Seq リードをゲノム配列に対してスプライス部位を考慮してマッピングする。
2. エキソンとイントロンを考慮した隠れマルコフモデルに基づいて遺伝子を予測する。
3. 近縁種のアミノ酸配列をゲノム配列に対してマッピングする。
4. ある 1 つのフレーム上に存在する開始コドンから 3 塩基単位でずらし、終止コドンが出現するまでを遺伝子として予測する。

【正解】4

【解説】選択肢 4 はオープンリーディングフレーム(ORF)と呼ばれ、原核生物のゲノムや、真核生物の転写配列上に存在する長い ORF は、そこがコード領域であることを示す有力な証拠となる。しかし、真核生物のゲノム上では一般にコード領域の途中にイントロンが挿入

されるため、この方法は有効ではなく、スプライシングを考慮した方法が必要になる。

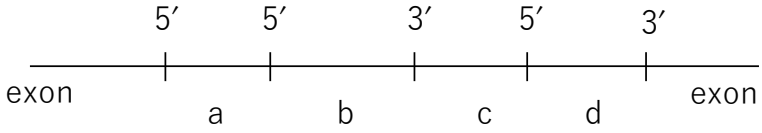
ゲノム配列を解読した後、リピート配列を検出することでゲノム配列の特徴を調べることができる。リピート配列の検出方法について、もっとも不適切なものを以下の選択肢から一つ選べ。

1. 既知のリピート配列データベース Repbase に対して RepeatMasker を用いてリピート配列を検出する。
2. N 末端にあるシグナルペプチドを検出する。
3. Tandem Repeat Finder を用いて縦列反復配列を検出する。
4. 分自身のゲノムへの相同性検索により多数ヒットする領域を新規のリピート配列として検出する。

【正解】2

【解説】リピート（反復）配列には、同じ配列が局所的に連続して出現する縦列反復配列と、ゲノム全体に散在して出現する散在性反復配列とがあり、それぞれ選択肢 3 および 4 のような異なるアルゴリズムで検出される。散在性反復配列はその類似性に基づいていくつかのファミリーに分類され、Repbase などのデータベースとして整理されており、それを用いて既知のリピート配列の検索を行うことができる。

簡略化した遺伝子予測問題を考えよう。ゲノム配列上にスプライス部位を予測したところ、イントロンの 5′ 末端（ドナー部位）および 3′ 末端（アクセプター部位）のモチーフを持つ部位が下図に示すように合計 5 つ見つかった。これらの部位で分けられる a〜d の領域について、コーディングエキソンおよびイントロンのポテンシャル（高い方ほど可能性が高い）をそれぞれ評価したところ、以下の表に示すような値になった。いま、この領域の両端はコーディングエキソンであるとし、またすべての境界で読み枠のずれが生じることはなく、単純に各領域のスコアの和で全体のポテンシャルが評価できるものとする。また遺伝子は図の左から右に転写されるもののみを対象とし、逆鎖側は考えないものとする。このとき予測される最適な遺伝子構造（領域 a〜d におけるコーディングエキソン(E)およびイントロン(I)の割り当てを順に並べたもの）はどれか。もっとも適切なものを選択肢から一つ選べ。



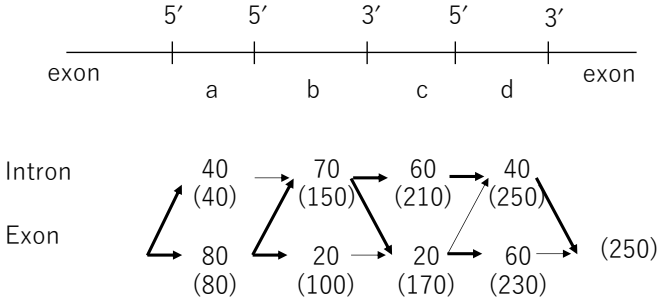
領域	a	b	c	d
イントロン (I)	40	70	60	40
エキソン (E)	80	20	20	60

1. (E)-E-I-I-I-(E)
2. (E)-E-I-E-I-(E)
3. (E)-E-I-E-E-(E)
4. (E)-E-I-I-E-(E)

【正解】 1

【解説】イントロンには 5′ 末端と 3′ 末端にそれぞれ保存配列があり、そのモチーフを用いてイントロン境界の候補を列挙できるが、イントロン領域を予測するには 5′ 末端で始まり 3′ 末端で終わるような正しい組合せを取る必要がある。選択肢の中では選択肢 4 のスコアが最も大きいが、b と c をイントロンとすると 5′ 末端で始まり 5′ 末端で終わることになるため、組合せが正しくない。残りの選択肢はいずれもイントロン末端の組合せが正しく、その中でスコアが最高の選択肢 1 が正解となる。

なお、選択肢 1 が最高値であることは以下のようにして確認できる。イントロンの開始(E から I への移行)は 5′ 末端のみで、イントロンの終了(I から E への移行)は 3′ 末端のみで可能であるという制約を考慮して下図のような状態遷移を考え、各段階でそこまでのスコアの和が最高値を取るパスを動的計画法で順次求める（太線で示されているのが各段階で選択されたパス、括弧内の数値がスコアの和)。これをトレースバックすることで選択肢 1 の結果が得られる。



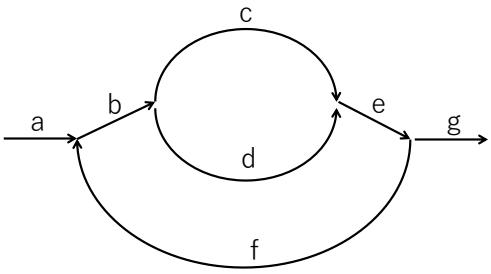
以下の文章は、アミノ酸の配列類似性を評価するスコアについて述べたものである。もっとも不適切なものを選択肢から一つ選べ。

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。

【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。

【構造生物学】

以下にいくつかの既出問題について解説します(以降は、これらの問題はこのままの形では出題されません)。

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。

【正解】4

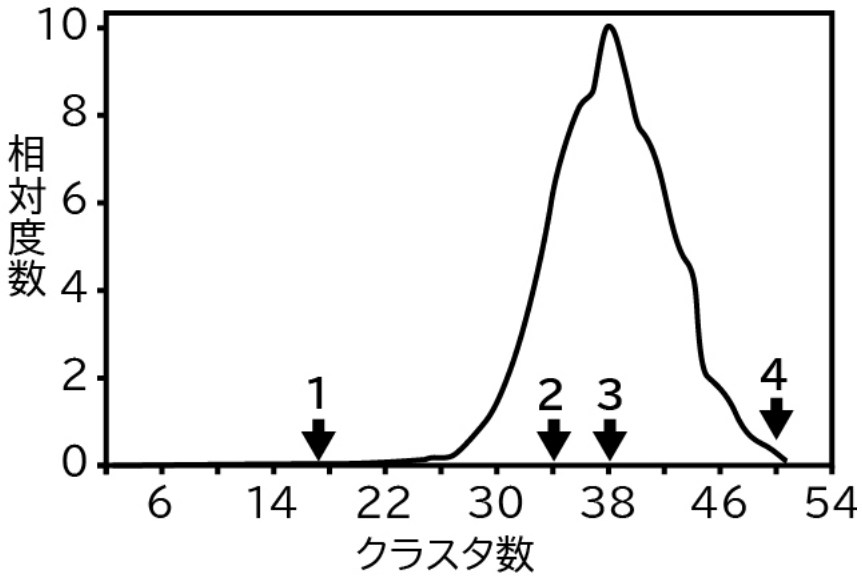
[解説]　人工的なタンパク質設計は主として、あるフォールドをとりうるアミノ酸配列の論理的な予測と、その配列に特定の機能(特異的なリガンドの結合や酵素活性など)を付与できるアミノ酸置換の予測からなる。経験的に、大きく異なるアミノ酸配列が類似したフォールドを取ることができる(タンパク質フォールド 1000 個説)ことから、解となるアミノ酸配列は単一ではない（選択肢 1 は正しい）。

手法としては、フォールド認識法などを応用して、フォールドの構造モデルに適合した(構造の安定化に寄与する)アミノ酸を選択する方法が良く利用される。従って、立体構造予測や分子動力学はこの手法の根幹となる(選択肢 2 は正しい)。またこの手法がアミノ酸配列(文字列)の生成にあたることから、近年ではトランスフォーマーなどの生成 AI の手法がよく利用される。デザインの対象は膜タンパク質やタンパク質間相互作用にも拡張できるので、選択肢 3 は正しい。

一方で上記のいずれの手法を用いた場合も、対象タンパク質の分子量が大きい場合には予測精度が低下する傾向がある。また一般に、分子量の大きいタンパク質は複数のドメイン(フォールド)が連結したものである場合が多く、全体構造の予測にはフォールド間の相互作

用の予測が必要になるので、より困難な課題となる。従って選択肢 4 が最も不適切であり、これが正解となる。

タンパク質分子表面で進化的に保存されたアミノ酸（表面残基）どうしは、立体構造上、近接する傾向にある。分子表面における保存残基のクラスタは、酵素の活性部位など、機能に重要な領域と一致することが多い。
図中に示された曲線は、pyruvate decarboxylase (1pvd)の表面残基の 15%（81 個）をランダムに選び、それらが空間的に近接している場合クラスタとみなすという操作を 5,000 回繰り返し、その都度形成されたクラスタの個数を計測することで示されたものである。次に、ある方法で進化的に保存された表面残基のみを 81 個選んだ場合、形成されたクラスタの個数は図中に示された 4 つの矢印の一つに対応した。矢印の位置としてもっとも適切なものを選択肢の中から一つ選べ。



J. Mol. Biol. (2002) 316, 139-154. Figure 2 を元に作図。

1. 矢印 1
2. 矢印 2
3. 矢印 3
4. 矢印 4

【正解】1

[解説] タンパク質表面上で進化的に保存された残基は、リガンド結合や酵素活性などの分

子間相互作用に必要なアミノ酸残基を多く含むと予測されるので、タンパク質の表面上で局在する傾向にある(クラスターまたはパッチと呼ばれる)。従って保存残基のクラスター数は、ランダムに選抜した残基の場合に比べて有意に小さくなると期待できる。図に示されたクラスタ数は、正規分布に近い分布を示し、矢印(選択肢)3 付近が「表面残基をランダムに 81 個選択した際に観察されるクラスター数の期待値」に相当すると予想できる。目分量ではあるが、矢印 2 は期待値 - 標準偏差程度、矢印 4 は期待値よりも大きいことから、選択肢 2〜4 は有意に小さいとは言えないので、選択肢 1 が最も適切(正解)である。

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。

【正解】1

[解説]　電子顕微鏡によるタンパク質の構造決定にはいくつか方法があり、選択肢 1 はトモグラフィー(単一粒子について複数の傾斜像を得ることで三次元構造を構成する)、選択肢 2 は単粒子解析(多数の粒子の二次元射映像から三次元構造を再構成する)について述べている。また選択肢 4 に述べる様に、類似した構造が既知である場合は、その構造から作成した二次元の射映像を、粒子像の選抜やそれぞれの粒子像の方位の推定に利用することができ

る。従ってこれらの選択肢は、いずれも適切である。
一方、選択肢 1 は原子間距離情報から三次元構造を構築する方法(核磁気共鳴法 NMR)について述べているので、もっとも不適切(正解)である。

【遺伝進化】

ヒト白血球抗原(Human Lymphocyte Antigen, HLA または Major Histocompatibility Complex, MHC) 遺伝子群のアレル(対立遺伝子)の組み合わせは骨髄移植などにおける組織適合性に重要であるので、識別のためにHLAアレルには系統的表記法が採用されている。HLA 表記法 DRB1*04:10:01:02 で表されるアレルと翻訳産物(タンパク質)が同一であるアレルを選択肢の中から一つ選べ。

選択肢

1. DRB1*04:05:01:01
2. DRB1*04:10:03:01
3. DRB1*09:10:01:02
4. DRB3*04:10:01:02

【正解】2

解説

HLA ハプロタイプのアレル表記は、遺伝子名につづくアスタリスク（*）のうしろから、コロン（:）で区切られた 4 つの区域によって表される。第 1 区域は HLA 抗原型（抗体応答性)、第 2 区域はアミノ酸配列、第 3 区域はコード領域の塩基配列、第 4 区域はコード領域以外の塩基配列による分類を示している。問題で問われている DRB1*04:10:01:02 と翻訳されたアミノ酸配列が一致するということは、遺伝子名から第 2 区画までが一致している必要がある。したがって正答は②となる。選択肢③、④については、第 2 区画は一致しているものの、上位の階層が異なっているため，アミノ酸配列は異なっていると考えられる。

【オーミクス】

次世代シークエンサのデータは、FASTQ という形式で扱う場合が多い。以下の FASTQ ファイルに関する記述として、もっとも不適切なものを選択肢の中から一つ選べ。

1. FASTQ は、FASTA 形式の拡張版でありテキストエディタで開くことができる。
2. FASTQ は、塩基配列に加えて各塩基のクオリティスコアを 0-9 の ASCII 文字で記録している。
3. FASTQ は、そのまま FastQC などのクオリティ評価ソフトウェアで分析できる。
4. FASTQ は、そのまま BWA などを用いたマッピング処理に使用できる。

正解：2

解説：FASTQ は、シークエンス配列データに加えて、Phred クオリティスコアを持つ、テキストベースのファイル形式である。Phred クオリティスコアは $Q = -10 * \log_{10}(p)$ で与えられる。ここで、p は誤った塩基がコールされる確率である。このクオリティスコアは、FASTQ では Q + 33 の数値に相当する ASCII 文字（すなわち ASCII コード上の「!」（ASCII コード 33）から「~」（ASCII コード 126）までのいずれかの文字）で表現される。

遺伝子ネットワーク解析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. ベイジアンネットワークはフィードバック制御をあらわすこともできる。
2. 遺伝子の相関ネットワークから因果関係を推定する研究もある。
3. ブーリアンネットワークはネットワークの時間変化をあらわすこともできる。
4. ベイジアンネットワーク上の確率変数は、連続値でなく離散値でもよい。

正解：1

解説：ベイジアンネットワークは、様々な事象の因果関係を条件付き確率のグラフとして表現したものである。ベイジアンネットワークは有向非巡回グラフで構成され、フィードバック制御のような巡回グラフ構造をあらわすことはできない。

ヒトゲノムの解析には、リファレンス配列とも呼ばれる標準ゲノムデータをよく使う。現在最もよく使われるバージョンは、ゲノム・リファレンス・コンソーシアム（略称 GRC）によるヒト(human)のビルド 38 であり、GRCh38 と名付けられている。標準ゲノムの記述として、もっとも不適切なものを選択肢の中から一つ選べ。

1. 標準ゲノムは比較対象にすぎず、その配列と異なるものが異常というわけではない。
2. 常染色体、性染色体の情報を含んでいるが、ミトコンドリア配列は含んでいない。
3. ヒトゲノムの多様性をできるだけ反映させるため、標準の染色体情報とは別に、多様なコンティグ配列も提供されている。
4. 日本国内では、日本人に特化した標準ゲノム作りが実施されている。

正解：2

解説：GRCh38 は常染色体、性染色体のほかに、ミトコンドリア配列も含まれた配列セットである。その他の文章は正しい。