

平成30年度

日本バイオインフォマティクス学会 (JSBi)

バイオインフォマティクス技術者認定試験

解説集

問 1 正解【2】

ヒトを含む哺乳類の場合、個体の性別は主に性染色体の組み合わせで決まり、オスは XY、メスは XX が一般的である。よってオスのマウスの場合 Y 染色体は父親（精子）由来となり、選択肢 2 が正解となる。

問 2 正解【1】

DNA の遺伝情報を mRNA に転写するのは、RNA ポリメラーゼである。よって選択肢 1 が正解となる。

問 3 正解【4】

ウイルスは非常に小さく（数十～数百 nm）、宿主の様々な生体防御機構をかいくぐって、外胚葉由来の組織（皮膚、脳、脊髄など）、中胚葉由来の組織（骨、筋肉、心臓、腎臓、脾臓、血管、血液など）および内胚葉由来の組織（消化管、肺、肝臓、膵臓など）の細胞に感染するものがある。よって選択肢 4 が正解となる。

問 4 正解【2】

減数分裂により細胞の染色体数は半分になる。よって選択肢 2 が正解である。

問 5 正解【1】

原核生物（真正細菌と古細菌）では、一般に 1 種類の RNA ポリメラーゼが rRNA、mRNA、tRNA の生合成（転写）を全て行う。真核生物では、一般に RNA ポリメラーゼ I が rRNA、RNA ポリメラーゼ II が mRNA、RNA ポリメラーゼ III が tRNA の転写を主に担当する。よって選択肢 1 が正解となる。

問 6 正解【2】

エピジェネティクスとは DNA 塩基配列の変化を伴わない、DNA やヒストンの化学修飾（メチル化やアセチル化など）による制御システムとその学問分野に関するものである。よって選択肢 2 が正解となる。

問 7 正解【2】

形質細胞に分化した 1 個の B 細胞からは 1 種類の免疫グロブリン抗体が産生される。よって選択肢 2 が正解となる。モノクローナル抗体はこのメカニズムを利用したものである。

問 8 正解【2】

細胞のシグナル伝達では細胞外のシグナル因子が細胞膜上の受容体に結合した後、受容体が活性化された分子が直接細胞内の状態を変化させる場合と、核に移行して遺伝子発現を

制御し、遺伝子発現の変化により間接的に細胞内の状態を変化させる場合があり、前者のほうが細胞は速く応答できる。よって選択肢 2 が正解となる。

問 9 正解【1】

RNA は一旦 cDNA に逆転写することで間接的に PCR 反応に使用する (RT-PCR 法)。近年では 1 つのチューブ内で逆転写と PCR 反応を連続的に行える技術も広く普及している。よって選択肢 1 が正解となる。

問 10 正解【2】

参照配列 (リファレンスゲノム) を使用せず、断片解読された配列のみでアセンブルする (つなぎ合わせる) 方法を *de novo* (デノボ) アセンブリという。すでに解読された生物種のゲノム配列に基づいて、異なる個体や近縁の種のゲノム配列を解析する方法はリシーケンセス等と呼ぶ。よって選択肢 2 が正解となる。

問 11 正解【4】

iPS 細胞は、初期化に必要な 4 つの遺伝子を既に分化した体細胞に入れて人工的に作り出す。この時、初期化細胞に患者自身の細胞を使用すれば、自身と同じ遺伝情報を持った多機能性幹細胞ができるため、iPS 細胞をもとに作った臓器や細胞を患者に移植しても、免疫系による拒絶は少ないと考えられる。よって選択肢 4 が正解となる。

問 12 正解【1】

ゲノム編集は文字通りゲノム配列を操作する技術で、変異の導入にはヌクレアーゼによる二本鎖 DNA の切断が必要である。標的部以外以外の意図しない場所が改変されるオフターゲット効果に注意しなければならない。選択肢 1 が正解となる。

問 13 正解【3】

cGMP は cAMP と同じく、細胞内での情報伝達に関与するセカンドメッセンジャー (二次情報伝達物質) で、ホスホジエステラーゼによって分解される。よって選択肢 3 が正解となる。

問 14 正解【3】

ヌクレオチドは核酸の構成単位で、糖に塩基とリン酸が結合した分子である。糖の 2 位の位置に OH があるものをリボヌクレオチド、ない (H に置き換わった) ものをデオキシリボヌクレオチドと呼び、DNA が複製される際には、デオキシリボヌクレオチドが基質となる。デオキシの場合 dATP (デオキシアデノシン三リン酸) のように頭に d を付けて違いを示す。よって選択肢 3 が正解となる。

問 15 正解【3】

普遍遺伝暗号には、翻訳終了を意味する終止コドンが 3 つ (UAA、UAG、UGA) ある。よって選択肢 3 が正解となる。なお、ヒトのミトコンドリアの遺伝暗号では、UGA は終止コドンではなく、トリプトファンを指定する (ヒトのミトコンドリアの遺伝暗号には、他にも普遍遺伝暗号と異なるコドンが幾つかある)。ミトコンドリアの遺伝暗号は生物種 (生物の系統) によってかなり多様であり、配列解析を行う際には注意が必要である。また、一部の原核生物および真核生物 (の核) では、普遍遺伝暗号とは異なる遺伝暗号が用いられており、これら生物種の配列解析を行う際にも注意が必要である。

問 16 正解【3】

トリプトファンの 1 文字表記は W である。なお、Y はチロシンの 1 文字表記である (チロシンの 3 文字表記は Tyr)。よって選択肢 3 が正解となる。

問 17 正解【4】

SNP は一塩基多型、CNV はコピー数多型、VNTR は反復配列多型を表す言葉である。GWAS はゲノムワイド関連解析のこと。よって選択肢 4 が正解となる。

問 18 正解【2】

指針では全ての研究者等の基本的な責務として、『提供者又は代諾者等のインフォームド・コンセントを受けて、ヒトゲノム・遺伝子解析研究を実施することを基本としなければならない』とされている。よって選択肢 2 が正解となる。

問 19 正解【3】

「個人情報の保護に関する法律についてのガイドライン・通則編」(平成 29 年 3 月 一部改正) には、特定の個人を識別できるゲノムデータとして、『互いに独立な 40 箇所以上の SNP から構成されるシーケンスデータ』と記載されている。よって選択肢 3 が正解となる。

問 20 正解【4】

ポリアデニル化とは、RNA の 3'末端に数十以上の A (アデニン) が連続で付加されること。付加された部分をポリ A 鎖と呼ぶ。よって選択肢 4 が正解となる。

問 21 正解【4】

例えば x を 0.5 にした場合には、選択肢 4 のみ計算結果が 1 になる。

問 22 正解【1】

XML はマークアップ言語とよばれる言語の一種で、プログラミング言語ではない。

問 23 正解【4】

含まれる  $a$  以上  $b$  以下の整数の数を  $occ$  とすると、列挙するには  $O(\log n + occ)$  時間が必要である。

問 24 正解【3】

ヒープと二分探索木の定義を確認する問題。選択枝 3 の二分木では、どの節点（葉をのぞく）に関しても「左の子の数字 < その節点の数字 < 右の子の数字」となっており、この二分木は二分探索木である。

問 25 正解【4】

1 は後順でたどった際のラベル列、2 は中間順でたどった際のラベル列、3 は幅優先探索した際のラベル列、4 は先行順でたどった際のラベル列である。

問 26 正解【4】

$S_1$  で終了する記号列は 4 のみである。

問 27 正解【3】

隣接行列はグラフの辺の本数が多いときは実用的であるが、辺の本数が少ないときは要素に 0 が多くなり、データに無駄ができてしまう。そのような場合、各頂点に接続する辺を連結リスト状に格納した隣接リストが有効であり、空間計算量は  $O(m + n)$  となる。なお、 $O$  記法の定義からは選択枝 4 も誤りではないが、より厳密な選択枝 3 がより適切であるため、選択枝 3 が正解。

問 28 正解【3】

RNA のシュードノットを含む二次構造は文脈自由文法で記述することができない。シュードノットを含まない二次構造は記述することができる。

問 29 正解【1】

このアルゴリズムは  $V$  と  $U$  の長さの和の線形時間で計算できる。 $O$  記法の定義からは選択枝 3 も誤りではないが、より厳密な選択枝 1 がより適切なため、選択枝 1 が正解。

問 30 正解【3】

UCF329B13、DRZ353B36、OPV214J66 の 3 件。

問 31 正解【2】

2 の組み合わせを選ぶと、すべて正答となる。

問 32 正解【2】

$n$  を試行回数、 $p$  を成功確率とすると、2 項分布は  $\binom{n}{x} p^x (1-p)^{n-x}$  と書ける。 $np \rightarrow \lambda$  となるように  $n \rightarrow \infty, p \rightarrow 0$  となる極限をとると、 $\binom{n}{x} p^x (1-p)^{n-x} \rightarrow e^{-\lambda} \lambda^x / x!$  となることが知られており、これはポアソン分布を表す。また、ポアソン分布は期待値、分散ともに  $\lambda$  となる。

問 33 正解【4】

標準正規分布の確率密度関数  $f(x)$  は、区間  $[0, +\infty)$  において狭義単調減少関数になる。また、あらゆる実数  $x$  において  $f(x) > 0$  を満たす。この 2 点から  $0 < f(y) < f(x)$  が導かれる。

問 34 正解【3】

正規分布のように、あらゆる実数  $x$  において  $f(x) > 0$  を満たすとき、累積密度関数は狭義単調増加関数になり、かつ、あらゆる実数  $x$  において  $F(x) > 0$  が成り立つ。よって  $0 < F(x) < F(y)$  となる。

問 35 正解【2】

確率の基本的な性質として、 $P(A|B) + P(\bar{A}|B) = 1$  が成り立つ。

問 36 正解【1】

ROC 曲線の AUC は曲線以下の部分の面積を表す。

問 37 正解【1】

最頻値は  $f(x)$  の最大となる点、中央値は  $F(x)=0.5$  になる点、平均値は 1 次モーメントである。この問題の例では、最頻値は 約 2.0、中央値は 10.91、平均値は 11.20 である。

問 38 正解【2】

中心極限定理によると、母平均  $m$ 、母分散  $v$  の母集団から  $n$  個の互いに独立な標本観測値から計算される標本平均の分布は、正規分布  $N(m, v/n)$  に漸近する。

問 39 正解【4】

$y_i = \frac{1}{1 + \exp(-z_i)}$  はシグモイド関数である。また、ReLU は  $y_i = \max(0, z_i)$  で定義され、勾配消

失問題に対処していることから、ディープニューラルネットワークに適している。

問 40 正解【2】

一般的な  $k$ -平均法のアルゴリズムにおいては、計算結果が収束するまでの計算時間の上限を与えることはできない。

問 41 正解【1】

一塩基置換はミスマッチを許した文字列照合によってマッピングできるので、スプリットアラインメントは必要ない。それ以外の選択枝のケースでは、リードを分割して、それぞれを参照ゲノム上の離れた位置と照合する処理が必要になる。

問 42 正解【1】

$k$ -mer 比較では完全一致が前提になるので、 $k$ を大きくするとミスマッチを含む配列を検出できない可能性が高くなり、検出感度が下がる（ $k$ を小さくすると検出感度が上がる）。一方、特定の  $k$ -mer はランダムな DNA 配列においても、各塩基の出現確率を  $1/4$  とすれば  $1/4^k$  の確率で出現する。このため、 $k$  が小さいと偽陽性のヒットを生じる可能性が大きくなる（ $k$ を大きくすると偽陽性の可能性が減る）。

問 43 正解【3】

図に示された最適アラインメントパスは、G のところにギャップが入っているので、3つのアラインメントのうち一番右のものに相当する。このアラインメントの XX のところに、各選択枝の配列の 2,3 文字目、GT、CC、CA、AC を入れて、ギャップの位置をずらしたときにスコアが下がる配列を探す。選択枝 3 の CA を入れたとき、A を一文字右にずらすと (G,A) のプリン塩基どうしが合い、(T,A) を合わせた時よりペナルティが小さくなって、スコアが大きくなる（下図）。

ずらす前	ずらした後
TCTGAC	TCTGAC
TCA-AC	TC-AAC

なお、プリン塩基間、またはピリミジン塩基間の変異はトランジション（転位）、プリン塩基とピリミジン塩基の間の変異はトランスバージョン（転換）といい、前者は後者より高頻度で起こることが知られている（→問 67）。

問 44 正解【4】

MAFFT は、DNA やタンパク質配列のマルチプルアラインメントを行う汎用的なソフトウェアである。

問 45 正解【2】

DNA の 2 本鎖においては、相補塩基対を形成しているため、 $A=T$ ,  $C=G$  の関係が厳密に成立する。ヒトゲノム中では、2 連続塩基 CG の出現頻度は、期待される頻度よりずっと低くなっていることはよく知られている (CG 配列の多くで C がメチル化されており、それが  $C \rightarrow T$  の変異を促すためと考えられている)。なお、ここで空欄 d は、前文の説明とは逆に、期待値が実際の出現頻度と比べてどうかを尋ねていることに注意。

問 46 正解【2】

相補鎖側(下)の配列を右から左に見ていくと、右から 8 文字目に ATG があり、そのままコドン 3 文字ずつ区切って読んでいくと、15 コドン目に TGA が見つかる。開始コドンはメチオニンをコードするが、終止コドンはアミノ酸をコードしないから、アミノ酸配列の長さは 14。

問 47 正解【3】

$k$  を十分大きくとったとき、多くの  $k$ -mer はゲノム中に高々 1 回出現することが期待されるため、 $k$ -mer の出現頻度分布から、シーケンサから得られたリード全体の長さがゲノムを何重にカバーしているかという「カバー度 (coverage)」が推定でき、これからゲノムサイズが推定できる。GC 含量や GC skew からはそうした推定はできない。フローサイトメトリー法では、ゲノムサイズ既知の生物の試料と蛍光強度を比較することにより、ゲノムサイズを推定できる。

問 48 正解【3】

選択肢 3 は、アノテーション可能な遺伝子を増やすという点で、偽陰性を避けるために有効な手段であるが、その結果誤った推定をするという偽陽性を増大させる危険性がある。このため、偽陽性を避ける他の工夫と組み合わせる必要がある。

問 49 正解【2】

局所的アラインメントにおいては、スコアが正であればアラインメントは長く伸びるが、スコアが負になるとアラインメントは打ち切られる。そこで、わずかなミスマッチでスコアが負になるようなスコア設定にすると、一致度が高く短い局所アラインメントが得られるようになる。そのためには、不一致のペナルティを、一致のスコアに対して大きくすればよい。

問 50 正解【1】

NW 法、SW 法 (正式名は問題文参照) は、それぞれ大域的、局所的アラインメントのアルゴリズムとして重要で、配列比較の様々な方法に使われているだけでなく、その高速化などが現在でも研究の対象となっている。KD 法は、各アミノ酸の水との親和性を数値化した

「疎水性指標」としても有名で、これを配列に沿ってプロットして得られる「疎水性プロット」は、膜貫通領域を予測する簡便な手法として現在でも使われることがある。CF法は、アミノ酸ごとに $\alpha$ ヘリックスや $\beta$ シートなどの二次構造要素を形成する傾向指数を、既知の立体構造を用いて統計的に算出し、これを用いて二次構造を予測する方法。KD法やCF法は、現在ではより洗練した方法に置き換えられて、その方法自体が使われることは少ないが、物理化学的指標を用いてアミノ酸配列の構造や機能を予測する方法の草分けといえ、そのアイディアは現在の様々な手法に通じている。

問 51 正解【4】

隠れマルコフモデルは、各矢印に付された遷移確率に従って状態を遷移し、各状態に付された塩基出現確率に従って塩基を出力する、という動作で配列を確率的に出力するモデル。問題となっている動作は、状態  $S \rightarrow 1$  の遷移(確率 1.0、以下同)、状態 1 で A を出力(0.3)、状態  $1 \rightarrow 2$  の遷移(0.1)、状態 2 で C を出力(0.2)、状態  $2 \rightarrow E$  の遷移(0.1)なので、これらの確率を掛け合わせればよい。なお、本問では経路が一通りしかないが、経路が複数存在する場合(たとえば AAC を出力するには、状態 1 で A、状態 2 で AC、または状態 1 で AA、状態 2 で C の 2 通りある)は各経路の確率の和をとる必要がある。

問 52 正解【3】

本問のモデルでは、状態  $1 \rightarrow 2$  の遷移がただ 1 回起こり、その確率は 0.1、それ以外は状態  $1 \rightarrow 1$ 、または状態  $2 \rightarrow 2$  の遷移であるが、その確率はいずれも 0.9 である。よって状態  $1 \rightarrow 2$  の遷移がいつ起こるとしても、遷移確率の積は一定であり、経路ごとの確率の違いを考えるには、遷移確率を無視して塩基出現確率だけを考えればよい。そこで、参考表を使って、各状態の塩基出現確率の積をとって比較する。選択肢の中で違っているのは 4 文字目から 6 文字目までなので、その間の確率の積を比較すればよい。選択肢 3 の場合は  $0.2 \times 0.4 \times 0.4 = 0.032$  となり、これが最大である。

問 53 正解【1】

過去にもしばしば出題されているコンタクトマップを読み解く問題である。ヘリックスは近傍のアミノ酸残基とのコンタクトが多いため、対角線上に太い領域として現れる。B、C、D の中で D のみが C 末端にヘリックスがあるため、構造 D とマップ E が対応する。あとは並行 $\beta$ シートが左下から右上へ、逆平行 $\beta$ シートが左上から右下へ向かう線として現れるため、配列上の何番目のストランド同士がどちら向きのシートを形成しているかを見れば、答えが分かる。

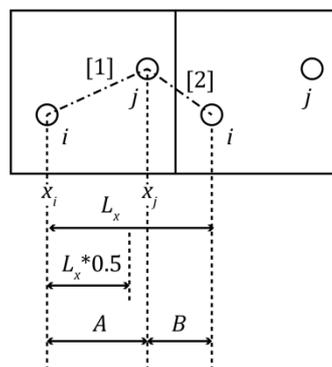
問 54 正解【2】

これは 5 つの要素の並べ方を数える問題と言い換えることができる。全ての可能な並べ方

は  $5!$  で計算でき、120 となる。ここで図 B、C、D 中において、下向きの要素が 3 つ、上向きの要素が 2 つある。上下を横切る構造を禁止するためには上向きと下向きの要素を交互に並べる必要があるため、必ず下向きで始まることになる。したがって並べ方の数は、 $3 \times 2 \times 2 \times 1 \times 1 = 12$  である。

問 55 正解【1】

分子シミュレーションでよく用いられる周期的境界条件において、隣り合うセルの中から最も近いペアを選ぶアルゴリズムを求める問題である。 $x$  軸だけを考えれば、原子  $j$  に隣り合う原子  $i$  は図中[1]か[2]で指すいずれかとなり、距離はそれぞれ下図中の距離 A および B となる。コード 1 行目で求められる  $r_x$  の値は同じセル内のペアの距離であり、すなわち距離 A である。一方 A と B の和はセルの長さ  $L_x$  となるため、 $A > L_x * 0.5$  のとき、 $A > B$  であり、求めるべき距離  $r_x$  は  $B = L_x - A$  となる。これが 2 行目の if 文の意味である。一方  $A \leq L_x * 0.5$  であれば、 $A \leq B$  であるため、求めるべき距離  $r_x$  は A となる。



問 56 正解【2】

アミノ酸配列と構造、機能の関連に関する問題である。 $\alpha$ ヘリックスは Ala や Leu などの小さな疎水性側鎖を好む傾向がある。DNA は負電荷を帯びているため、DNA 結合部位には多くの場合正電荷が多く含まれている。また天然変性領域は荷電性側鎖を好み、また翻訳後修飾部位を多くもつ傾向にある。なお、この配列は Ets-1 と呼ばれる転写因子の配列の一部である。

問 57 正解【1】

PDB フォーマットは 1 行 80 文字の固定長であり、要素はカラムごとに収められている。カラムに記載すべき内容がない時には空白文字が入られるため、ファイルを一見すると空白で区切られていると誤解しやすい。もちろん、要素の長さによっては要素間に空白が入らない場合もあるので、計算機での処理の際には注意が必要である。

問 58 正解【2】

この二つのタンパク質はどちらもグロビンフォールドに属し、 $\alpha$ ヘリックスのみを持ち $\beta$ シートを持たない。Aは同じ構造を持つサブユニットが向かい合って結合した二量体、Bは単量体であり、どちらもシングルドメインタンパク質である。そのため、Bをall- $\beta$ 型( $\beta$ シートのみを持つ構造)とした選択肢2がもっとも不適切となる。

問 59 正解【3】

ポリペプチド鎖ではペプチド結合は二重結合があり自由に回転できないため、 $C\alpha$ 原子-アミノ基 N 原子間の結合の回転角 $\phi$ (ファイ)角と $C\alpha$ 原子-カルボキシル基 C 原子間の結合の回転角 $\psi$ (プサイ)角の2つで、そのおおよその構造が決定される。R. A. ラマチャンドランによって提案されたラマチャンドランマップは、ペプチド主鎖二面角の分布図である。横軸に $\phi$ 角、縦軸に $\psi$ 角をとり、対応する点にそれぞれのアミノ酸残基をプロットする。Aは $\beta$ シート領域、Bは左巻きらせん領域、Cは $\alpha$ ヘリックス領域と呼ばれている。したがって、選択肢3が誤りである。アミノ酸の種類ごとにラマチャンドランマップを描くと、グリシンやプロリンでは、他のアミノ酸と異なる特徴的な分布となることが知られている。実験的に求めた立体構造だけでなく、予測により構築された立体構造の検証にも用いられている。1963年の提案当時より、明らかにされた立体構造の数と多様性は格段に増加している。結果として、現在入手可能な全タンパク質についてこのマップを作成すると、各領域は広がる。

問 60 正解【3】

タンパク質の分子機能を考える際、解析対象タンパク質の相同配列群を用いて算出されるアラインメントサイトの保存度と、解析対象もしくはその近縁タンパク質の立体構造の情報を組み合わせることで、より詳細な分子機能情報を得られる。選択肢1, 2, 4に記載された内容はいずれも正しい。酵素の基質認識に重要な役割を担うアミノ酸(基質認識部位)は、基質を認識する際、基質と接触するために分子表面に露出する必要がある。しかし、基質認識部位が常に分子表面に位置しているとは限らない。立体構造変化の結果として基質認識部位が分子表面に露出し、基質を認識できるようになることもある。したがって、選択肢3が誤りである。

問 61 正解【3】

これらのファイルフォーマットは、BowtieなどのプログラムでNGSリードをリファレンス配列にマッピングする際などに使用される。FASTQは配列記述フォーマットFASTAにベースコールのエラー確率をアスキーコードで示したものである。BAM(Binary Alignment Map)形式はNGSリードのマッピング結果をバイナリファイルに出力したもので、これのアスキーファイルをSAM(Sequence Alignment Map)形式と呼ぶ。BAM/SAM

ファイルなどに基づいて、リファレンス配列からの変異を出力したファイル形式が VCF (Variant Call Format) である。GTF (General Transfer Format) ファイルは遺伝子の注釈(アノテーション)の出力であり、特に ChIP-seq のピーク検出に関連しているわけではない。従って選択肢 3 が正解となる。

問 62 正解【3】

メンデル遺伝学の基礎知識を問う問題である。純系の丸・緑(RRyy)としわ・黄(rrYY)の F1 は RrYy、この表現型が丸・黄であるので、しわと緑は潜性(劣性)である。表現型分離比から形(丸かしわ)と色(緑か黄)は独立に遺伝していることがわかる。よって選択肢 3 が正解となる。

問 63 正解【4】

対立遺伝子頻度は、 $(A_1 \text{ 頻度}) = \{(A_1A_1 \text{ 個体数} = 10) \times 2 + (A_1A_2 \text{ 個体数} = 80)\} / \{(全個体数 = 100) \times 2\} = 0.5$ 、および  $(A_2 \text{ 頻度}) = 1.0 - (A_1 \text{ 頻度}) = 0.5$  と求められる。 $A_1A_2$  個体数が非常に大きいことからハーディー・ワインベルグ平衡が成立しないことは容易に推測できるが、 $A_1A_1$ 、 $A_1A_2$ 、 $A_2A_2$  個体数の期待値がそれぞれ  $100 \times 0.5^2 = 25$ 、 $100 \times 0.5^2 \times 2 = 50$ 、 $100 \times 0.5^2 = 25$  であることを踏まえて、 $\chi^2 \text{ 値} = (10 - 25)^2/25 + (80 - 50)^2/50 + (10 - 25)^2/25 = 36$  を求めても、有意水準を 1% とした場合に有意であることがわかる。よって選択肢 4 が正解となる。

問 64 正解【1】

UPGMA 法は、進化速度の一定性を仮定することで系統樹の根(ルート)の位置が特定される。一方、NJ 法は進化速度の一定性を仮定せず、外群を導入するなどしない場合は根(ルート)の位置は特定されない。これらから、選択肢 1 が正解であることがわかる。

問 65 正解【3】

系統樹 A には根(ルート)があることから、ニューウィック(Newick)表現は A, B, C をそれぞれ部分系統樹としたときに(A, B);のように表現される。無根系統樹の場合は(A, B, C);になる。従って選択肢 3 が正解となる。

問 66 正解【2】

選択肢 1 の記述は、系統樹 A, B のトポロジーが異なる(ヒトとクラスタするのがウシかマウスの違いがある)ことから妥当である。選択肢 2 の記述は、UPGMA 法も NJ 法も進化距離に基づく距離行列法であるので誤っている。選択肢 3 の記述は、UPGMA 法と NJ 法の系統樹トポロジーが異なる場合の典型的な理由を述べている。選択肢 4 の記述は、選択肢 3 について、この問題の場合の具体的な理由を述べていて、系統樹 B のマウスに至る枝が長

いことから、十分に妥当性のある推定である。よって選択肢 2 が正解となる。

問 67 正解【1】

選択肢 1 の記述は、一般的にはトランスポージョン変異の頻度が低いことが知られているので誤っている。選択肢 2 は変異記述法として一般化してきた HGVS に関する記述であり、c は coding DNA、数字は塩基番号を示し、変異を(変異前または参照配列の塩基)>(変異後の塩基)で示すので正しい。選択肢 3 の記述は、置換数と進化距離が比例しない典型的な理由を述べているので正しい。選択肢 4 の記述は、通常は  $dN < dS$  が成立することから正しい。よって選択肢 1 が正解となる。

問 68 正解【3】

選択肢 1 の記述は連鎖不平衡の定義であり正しい。選択肢 2 の記述も実際に観測される現象であり正しい。選択肢 3 の記述は、ハプロタイプとディプロタイプの用語を逆に用いているので誤っている。選択肢 4 の記述は、連鎖解析による遺伝子位置の推定法として正しい。よって選択肢 3 が正解となる。

問 69 正解【2】

定常状態で B と C が平衡に達して  $v_5=0$  の時が上限になると考えられるので、流入分が全て D にある 2 が正解。

問 70 正解【2】

$v=V/4$  を代入して y について解くことにより、 $y=3K$  となるので、2 が不適切。4 が少しわかりにくいかもしれないが、 $y=0$  で v が最大となり、y の増加と共に v が減少するので、y の増加が転写を抑制していることを意味する。

問 71 正解【3】

スケールフリーネットワークは多くの現象をネットワーク表現した際に見られるネットワークの 1 種であるが、統計学的な帰無仮説としてはランダムネットワークを用いることが適切である。

問 72 正解【4】

時間に関する微分方程式で、この式は物理現象でもよく見られる指数関数的増大を表すモデル式となっている。比例常数 k に応じて増加が大きくなるので、倍になる時間(ダブリングタイム)は短くなる。

問 73 【全員正解】

発現量を遺伝子の長さやシーケンサで読んだ総リード数で補正して FPKM (遺伝子の長さ 1kbp あたり、試料ごとの総リード数 1M あたりに換算したリード数) に直すと

	A (10M)	B (30M)
X (1kbp)	300 → 30	900 → 30
Y (2kbp)	500 → 25	1800 → 30

となる。1 と 4 の両方が正解となってしまうため、全員正解とした。

問 74 正解【3】

JASPAR は転写因子結合部位に関するデータベースであり、タンパク質間相互作用のデータベースではない。

問 75 正解【2】

RamDA-Seq は 1 細胞 RNA 解析に関する手法であり 2 の説明は間違っているが、この方法を知らなくても、他の基本的な解析手法を理解していれば消去法で正解にたどり着けると思われる。

問 76 正解【1】

腸内、皮膚、口腔内、土壌などの様々な環境に生息する細菌のバラエティを解析するメタゲノム解析の基本的な解析の流れを説明した文章である。(b)の SILVA データになじみがないかもしれないが、STRING データベースでは無いこと、(a)の OTU と ORF の違い、(c)の GATK がゲノム解析の手法であることを理解していれば正解にたどり着けると思われる。

問 77 正解【4】

NMR 法は質量分析装置に比べてスループットは高いが感度は高くないため、必要なサンプル量は NMR 法の方が多い。

問 78 正解【1】

t 検定は母集団が正規分布に従うことを仮定した検定法であるが、クラスタ解析した結果のクラスタ毎の分布が正規分布しているとは限らない。また、クラスタ数が 3 つ以上になった場合は、ANOVA を用いて検討する必要がある。

問 79 正解【4】

PCR 法での発現量と Ct 値の関係は頻出問題であるが、Ct 値は増幅曲線が一定のシグナルを超えるサイクル数なので、Ct 値が 1 減る毎に、元の量が 2 倍量ずつ発現していたことを意味する。今回は Ct 値が 2 低いので、 $2 \times 2 = 4$  倍量発現していることになる。

問 80 正解【4】

バイサルファイトシーケンス法はメチル化解析のために用いられる手法であり、ヌクレオソームの配置を調べるのは DNase-seq や ATAC-seq などの方法が用いられる。

試験問題に記載されている会社名または製品名は、それぞれ各社の商標または登録商標です。

なお、試験問題では、®および™を明記していません。