平成29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【3】
電気泳動を用いた DNA 分子の分離において、ゲル内の DNA 分子の泳動距離は DNA 分子の長さに関係するが、比例関係にはなく、分子の短いものほど移動時間に応じて長く進む。選択肢 3 はこの点で間違っていて不適切なので、正解は選択肢 3。

問 2　　正解【2】
細胞内の DNA 量に応じて蛍光強度が強くなる、つまり DNA 量が多いほど X 軸の大きい値をとる。S 期には DNA が複製されるので、S 期の細胞は DNA 量が増加している時期であることがわかる。M 期は分裂が進み、完了するまでは相対的に DNA 量は最大である。これらを満たすのは選択肢の 2。

問 3　　正解【2】
コルヒチンはチューブリンの重合阻害剤なので紡錘体が正しく形成されず、細胞分裂が完了できない。このため、全ての細胞が M 期で止まるので、正解の選択肢は 2。

問 4　　正解【2】
メンデルの遺伝の法則を構成する三つの法則、独立の法則、優性の法則、分離の法則についての理解を問うている。優性の法則とは、ある形質について、優性遺伝子のみを持った純系の親と、劣勢遺伝子のみをもった純系の親を交配させてできた雑種第一代は、優性の形質のみを発現することである。（注：現在「優性」を「顕性」と言い換えるべきだという意見が提唱されているが、この点についてより分かりやすくするものであろう。）

問 5　　正解【2】
ハーディ・ワインベルグの法則が成り立つ仮想集団では、集団内のランダムな交配が想定されている。したがって選択肢 2 の文章はもっとも不適切であり、これが正解となる。

問 6　　正解【3】
真核生物ゲノムにおいて、DNA のメチル化とは、DNA の一部のシトシンの六員環の 5 位の炭素原子にメチル基が付与されることである。動物ではシトシンの次にグアニンがくる CG 配列の C がメチル化されるが、植物では全てのシトシンがメチル化されうることが知られる。

問 7　　正解【2】
動物のホルモンは、親水性（水溶性）と親油性（脂溶性）の二つに大きく分けられる。ステロイドホルモンは典型的な親油性ホルモンであり、細胞膜を通過して細胞内のレセプターを介してシグナル伝達を行う。これに対し、親水性ホルモンは細胞膜を通過しないため、

細胞膜上のレセプターを介してシグナル伝達を行う。

問 8　　正解【4】
脊椎動物の軸索は束ではなく、1 本のニューロンをミエリン鞘が覆っている。

問 9　　正解【1】
環状の DNA をもつ細菌では、複製起点が一つ存在することが知られている。

問 10　正解【4】
mRNA の前駆体は、核内においてエキソンの切り出しを行い、そうして作られた成熟 mRNA が核外に輸送される。

問 11　正解【2】
ポリペプチドとは、多数のアミノ酸がペプチド結合によって繋がった重合体のことである。DNA が「デオキシリボヌクレオチドがフォスフォジエステル結合を形成して作るポリマーである」ことと勘違いしないようにしたい。正解は選択肢 2。

問 12　正解【3】
生物が行うシグナル伝達は様々な形で行われる。整理して理解しておきたい。動物のホルモンは主に循環器系を通してホルモンが伝達する。この点で不適切な表現があるので、正解は選択肢 3。

問 13　正解【3】
これまでに使われてきたゲノム編集技術についての質問である。CRISPR/Cas9 における DNA 結合ドメインは 1 本鎖の RNA（ガイド RNA）で構成されている。したがって選択肢 3 が正解。なお選択肢 4 については、今後法律の変更が行われる可能性があるので注意。

問 14　正解【1】
筋肉は主に中胚葉、神経は主に外胚葉から形成される。したがって選択肢 1 が不適切であり、正解。

問 15　正解【3】
再生医学に関して、iPS 細胞と ES 細胞を正しく理解しているかどうかを問うている。発生初期の胚盤胞から内部細胞塊を取り出して、それを培養して作成するのは ES 細胞である。したがって正解は選択肢 3。

問 16　正解【3】
文章を丁寧に読めば正解にたどり着く。Sog と chordin がホモログで、それぞれ節足動物と脊椎動物で神経を誘導すると書いてある。脊椎動物は背側に中枢神経系があるので、B が正解。節足動物では背腹が逆転していると書いてあることから、C が正解であるとわかる。したがって正しい選択肢は 3。

問 17　正解【1】
選択肢にある細胞小器官のうち、核ゲノムの DNA とは別に独自のゲノム DNA を持つものは、1 のミトコンドリアと、2 の葉緑体である。質問では動物細胞について問うているので、答えは選択肢 1 のミトコンドリア。

問 18　正解【1】
有性生殖とは一倍体配偶子の接合子によって二倍体の細胞である接合子が生じることである。この点で不適切な表現があり、正解は選択肢 1。

問 19　正解【1】
ヒトクローン胚のヒト胎内への移植は、日本では一切認められていない。この点で不適切な表現があり、正解は選択肢 1。

問 20　正解【4】
ヒトゲノム・遺伝子解析研究の業務を外部機関に委託することは、一定の範囲で認められている。この点で不適切な表現があり、正解は選択肢 2。
　なお、問 19, 20 でとりあげた生命倫理に関する問題は、ヒトゲノム解析や再生医学などの生命科学技術の進展に伴って、それらがヒトの個人や社会へ与える影響の大きさから重要視されるようになってきており、今後バイオインフォマティクス技術者がこれらの倫理的問題に直面する可能性も大きくなると予想される。初学者向けの参考書はまだ少ないが、文部科学省が定めた指針は、文部科学省・ライフサイエンスの広場「生命倫理・安全に対する取り組み」(http://www.lifescience.mext.go.jp/bioethics/seimei_rinri.html)のページで確認できる。

問 21　正解【1】
4 ビットの 2 の補数表現において、1111 は –1 を意味する。

問 22　正解【1】
たとえば、11/8 は 1。11 % 8 は 3 である。B の「11 番目」の要素は A の「1 番目」の要素の「3 番目」のビットに格納されている。(「××番目」はすべて先頭を 0 番と考えてい

ることに注意)

問 23　正解【4】
たとえば 8 ビットの 2 進整数内の「0 番目」のビットを取り出すには、7 ビット(7–0)ずらして、AND 演算を用いて一番右のビットを取り出せばよい。

問 24　正解【3】
コンパイラ型言語は C, C++, FORTRAN, Pascal などが知られている。一方、インタープリタ型言語では、Perl, Python, PHP, Ruby などがよく用いられている。

問 25　正解【3】
公開鍵暗号方式は，公開鍵から秘密鍵を推定するのは困難であることを前提とした方式であり、公開鍵は公開してよい。

問 26　正解【3】
任意の入力に対して遷移可能な状態が唯一つに決まるため、このオートマトンは決定性有限状態オートマトンである。また、s_0 と s_1, s_2 は、受理状態であるかどうかという点で状態として異なること、s_1 と s_2 は入力 1 に対して受理状態 s_0 に達する場合と、達しない場合とで異なることから、より小さな状態で同じオートマトンを表現することが不可能である。

問 27　正解【1】
この擬似コードは左側から深さ優先探索で左側から探索を行い、行きがけに‘(’、帰りがけに‘)’を出力している。このような括弧列は木のデータ構造のコンパクトな表現方法（実際には 01 のビット列で計算機には格納される）としてよく用いられている。

問 28　正解【3】
この擬似コードも問 27 同様深さ優先探索であるが、スタックを用いているために探索が右側から行うようになっており、探索順が問 27 と異なる。このように、深さ優先探索と一口に言っても異なる探索順があり得ることに注意が必要である。なお、この擬似コードでスタックをキューに置き換えると、問 27 と同じ探索順の深さ優先探索となる。

問 29　正解【3】
バブルソートは $O(n^2)$のアルゴリズムである。

問 30　正解【1】

連鎖ハッシュ法などで衝突があった場合、$O(1)$で探索できないことがある。

問 31　正解【2】

ある関数 $f(x)$が $\Theta(g(x))$であるとは、その関数が $O(g(x))$ ($g(x)$に比例した漸近的上限が存在)かつ $\Omega(g(x))$ ($g(x)$に比例した漸近的下限が存在)であることを表す。言い換えると、十分大きい x に対して、$c_1\, g(x) \leq f(x) \leq c_2\, g(x)$ となるような正の定数 c_1, c_2 が存在する。c_2 の大きさに制限はないので選択肢 2 は誤りだが、十分大きい n に対しては $n < c_1\, n^2$ となるから選択肢 3 は正しい。

問 32　正解【2】

Burrows-Wheeler 変換は BW 変換とも略され、部分文字列検索や文字列圧縮などの応用がある。これを用いた部分文字列検索用のデータ構造は FM-index ともよばれ、BWA などのゲノムマッピングソフトでも用いられている。BW 変換を活用した文字列圧縮ツールとしては bzip2 などが知られている。左上の行列の行を辞書式順序でソートすると次図のようになる。

A	C	A	T	C
A	T	C	A	C
C	A	C	A	T
C	A	T	C	A
T	C	A	C	A

問 33　正解【1】

ヒトのゲノム配列などにはリピート配列がきわめて多数含まれることが知られている。そのような、同じ部分文字列を多く含む文字列においては、その文字列を BW 変換した文字列内において同じ文字が連続して現れたり、近い場所でかたまって現れたりすることが多い。たとえば、「ACATC」という塩基配列が多数含まれる DNA 配列においては、部分塩基配列「CATC」の直前に（他の塩基と比較して）「A」が多数現れるため、「ACATC」の先頭の「A」は、BW 変換すると、連続して、あるいはかたまって現れる（特に、もし、循環シフトしたどの文字列内においても、それらの文字列内に含まれるどの部分文字列「CATC」の直前の文字も「A」であった場合には、「ACATC」の先頭の「A」は BW 変換すると必ず連続する）。bzip2 などは、この現象をうまく活用して文字列の圧縮を行っている。

問 34　正解【4】

リレーショナル・データベースが満たすべき制約は、キー制約、実体整合性制約、参照整合性制約など、様々である。ドメイン制約のみを満たしていても、整合性を満たしている

とは言えない。

問 35　正解【2】
平均値は、(4+5+4+7+6) / 5 = 5.2 である。中央値は、整列データ（4, 4, 5, 6, 7）における中央の値である 5 である。最頻値は 4 が 2 回、それ以外の数字が 1 回ずつ出現しているので、4 となる。

問 36　正解【2】
箱ひげ図の箱の部分は、25%点、中央値、75%点を表す。

問 37　正解【2】
命題 b は $m \bmod n = 0$ のとき、また、そのときのみ正しい。
$\{0,1,\ldots,m-1\}$ 上の一様分布から $\{0,1,\ldots,n-1\}$ 上の一様分布に従う値を得るには、$\{0,1,\ldots,m-1\}$ から$m-(m \bmod n)$ 未満の値が得られるまで独立なサンプリングを繰り返し、得られた値を n で割った余りを取る、などの方法が考えられる。

問 38　正解【4】
ROC 曲線の AUC は曲線以下の部分の面積を表す。

問 39　正解【4】
leave-one-out 法は、n の値がサンプル数に等しい場合の n-fold 法。また、一般論として、判別器の良さを検討する際には、同一データを訓練データとテストデータに用いてはならない。

問 40　正解【4】
連続する k 文字の出現頻度を要素とする特徴ベクトルで定義されたカーネル関数はスペクトラムカーネルと呼ばれる。

問 41　正解【2】
あるリードと同程度の類似度を示す領域が参照配列上に複数箇所ある場合、そのリードがどの箇所に由来するかを決めることができない。そのような場合、各箇所に由来する確率が等しいとして均等に割り振るのが望ましい出力方法の一つであり、選択肢 1 は一見乱暴にも見えるが、これを実現する簡便で効果的なアプローチである。これに対し、選択肢 2 はつねに同じ箇所にマッピングされるので、その箇所だけリード数が不当に高く、その他の箇所ではまったくマッピングされないことになり、好ましくない。選択肢 3 は、一つのリードを複数箇所にマッピングするので、リード数を数える際は後処理で調整する必要が

あるが、トランスクリプトーム解析のように、参照配列上の位置によって張り付き方が大きく異なる場合は、後処理によってより正確なリード数を推定するのが望ましい。一方、ゲノム上に極めて多く出現する繰り返し配列がある場合、リード数の正確な推定は困難で、かつ処理効率が大きく落ちる可能性があるので、選択肢 4 のように捨ててしまった方がよいこともある。

問 42　正解【4】
RNA 二次構造を「対応する括弧」で表現する方法の問題。問題文に従って 5′末端から順に変換していけばよい。

問 43　正解【2】
次世代シークエンサ由来の配列データは、（一部例外もあるが）配列に加えて塩基ごとのクオリティ値も保持して解析に用いるのが普通であり、そのため FASTA フォーマットを拡張した FASTQ フォーマットが標準的に用いられている。ただし SRA では、それ以外の付加的な情報も含めて一つのファイルにコンパクトに格納できるよう、独自に開発されたバイナリ形式である SRA フォーマットで配列データが公開されており、それを専用のツール(SRA Toolkit)を使って FASTQ フォーマットなどに変換するようになっている。

問 44　正解【1】
GEO (Gene Expression Omnibus) は、NCBI で作成されている、マイクロアレイや次世代シークエンサによる遺伝子発現その他の機能ゲノミクス関連データを集積したデータベースである。

問 45　正解【4】
マルチプルアラインメントの最適解は、多次元の動的計画法（各配列を辺とする多次元超立方体の格子を埋めていく計算）によって計算できるが、これには配列数の指数関数オーダーの計算量が必要になる。よく用いられるツリーベース法（選択肢 1 の方法）では、はじめに総当たりの比較を行うために配列数の二乗に比例した計算時間が必要になるが、これは近似的な解法であり一般に最適解は得られない。一方、高速フーリエ変換（FFT）は、2 つのベクトルの相互相関関数の計算を高速化するのに用いることができ、マルチプルアラインメントプログラム MAFFT では、これをアラインメント計算の高速化に利用している。

問 46　正解【1】
プロモーター配列に基づく転写開始点予測の難しさについて述べている。予測の間違いには偽陽性（予測したものが正しくなかったこと）と偽陰性（正しいものを予測しなかったこと）があるが、ここでは配列モチーフ検索によって予測された配列が、必ずしもプロモ

ーターにおける転写制御配列でないことを言っているので、偽陽性について述べていることになる。

問 47　正解【2】

位置特異的スコア行列（PSSM）は、配列モチーフの表現方法の一つで、対象となる配列アラインメント中の各位置における文字の出現頻度に基づいて類似性スコアを定義する方法。スコアを出現頻度の対数によって定義した場合、スコアの和が配列の対数尤度に相当するため、確率モデルに基づく推定と見なすことができる（実際には、選択肢 1 にあるような対数尤度比による定義がよく用いられる）。ただし PSSM ではギャップは考慮しないか、一定のペナルティとすることが多い。一方、隠れマルコフモデル（HMM）の一種であるプロファイル HMM は、位置ごとの挿入や欠失の入りやすさまで考慮して配列の尤度を計算できるため、そのようなパターンを表現するモデルとして PSSM より優れている。なお、PSSM は、転写因子結合部位など DNA の配列モチーフの定義に用いられる際には、「位置重み行列(PWM)」や単に「重み行列」と呼ばれることも多いが、本質的に同じものである。

問 48　正解【3】

ゲノム A の各遺伝子（a1, a2, a3, a4）について、ゲノム B で最大スコアを示す遺伝子をとってペアを作ると a1–b1, a2–b5, a3–b2, a4–b2 となる。このうち、b2 から見て最大スコアとなるのは a4 なので、a3–b2 は双方向が成立せず除外される。それ以外の a1–b1, a2–b5, a4–b2 は、それぞれゲノム B から見ても最大スコアとなっているのでこの 3 対が正解。a1, a3, a4 は、A と B の共通祖先以前に遺伝子重複で生じた相同遺伝子（パラログ）で、進化の過程で a3 の系列のみが B から欠失したと推定される。

問 49　正解【4】

系列 1 は 0.5 付近で変動しており、系列 2 は 0 の前後で変動している。GC 含量（G と C の頻度の和）は 0〜1 の値をとるのに対して、GC skew（G と C の頻度の差を示す量）は正負の値をとるので、系列 1 が GC 含量、系列 2 が GC skew と分かる。細菌ゲノムの多くにおいては、DNA の複製方向によって G と C の頻度に偏りが生じており、結果として複製開始点と終結点の前後において GC skew の正負が逆転する。これから D 点が複製開始点の候補となる。

問 50　正解【2】

アラインメントとコンセンサスパターンを対応させる問題。アライメント上部にマークされた保存部位におけるアミノ酸の種類と、それらの間隔に注意して照合し、合わないものを消去していけばよい。最初の 2 つの保存された C の間隔から選択肢 1 と 3 が消え、2 番目の C と次に保存された「F または Y」のカラムの間隔から選択肢 4 が消える。このよう

にして選択肢 2 が正解とわかる。なお、選択肢 2 の正規表現における[FYW]は、余分なアミノ酸 W が加わってはいるが、「F または Y」のカラムにマッチすることに注意。実際、ここには芳香族アミノ酸が適合すると考えられるので、あえてパターンを拡張して検索することが効果的なこともある。

問 51　正解【2】

動的計画法によるアラインメントアルゴリズムは、1）上から「ギャップ」で、2）左から「ギャップ」で、3） 左上から「一致」または「不一致」で、の 3 つの経路の中で最大のスコアの経路を選択する。これを踏まえて考える。まず、上辺および左辺は「ギャップ」しか取りえないので、ここからギャップのスコアは–2 と分かる。2 行 2 列のセルは、ここに上または左のセルから「ギャップ」で到達する場合、スコアは–4 になるはずなので、ここには左上から「不一致」で到達しているはずであり、これから不一致のスコアが–1 と分かる。同様にして、その右隣の 2 行 3 列のセルも、ここには上または左から「ギャップ」で到達したのではなく、左上から「一致」で到達したはずであり、これから一致のスコアは 1 と分かる。

問 52　正解【3】

前問のスコアを使い、動的計画法のアルゴリズムで、左端（4 行 1 列）のセルから順に埋めていく。順に–6（上から「ギャップ」で）, –3（左上から「一致」で）, –4（左上から「不一致」で）となり、–4 が答えになる。これより右側のセルは関係しないので計算する必要はない。

問 53　正解【2】

例年よく出題される PDF のデータフォーマットに関する問題の類題であり、mmCIF ファイルを読み取る内容となっている。今年度は例年と異なりクライオ電子顕微鏡による単粒子解析データからの出題とした。ここ数年クライオ電子顕微鏡による構造解析結果が高インパクト誌の誌面を賑わせており、ノーベル化学賞の受賞も相まって重要な解析手法として認識されつつある。単粒子解析法のデータの特徴などをよく把握しておくことが、これらの最新の成果を正しく解釈する上で重要であろう。単粒子解析では 1 つのサンプルに対して顕微鏡画像を複数取得し、その中から様々な方向から見た同一の粒子を特定し、分類、重ね合わせを行っていくことで、二次元画像から立体構造を再構成していく。問題の PDB エントリーでは、1,172 枚の画像データを用いている。選択肢にある"RELION"は代表的な再構成ソフトウェアの一種としてよく知られている。クライオ電子顕微鏡での解析結果は、原子モデルは PDB へ登録され、その元となった電子密度マップは EMDataBank へ登録される。

問 54　正解【2】

分子シミュレーションにおけるポテンシャルエネルギーに関わる問題である。このような計算は分子シミュレーションに限らず粒子系の力学シミュレーション一般によく見られるが、系に働くポテンシャルの性質に応じて効率の良いアルゴリズムを設計することが肝要である。ポテンシャルの式に現れる粒子間距離 r の次数を見れば、減衰の速さの違いがよく分かる。周期的な原子の分布を仮定する条件は「周期的境界条件」と言うが、それに基づいてフーリエ変換を用いて相互作用を計算する方法として Ewald 法がよく知られており、分子シミュレーションにおける静電相互作用計算のデファクトスタンダードとなっている。

問 55　正解【1】

天然変性領域についての知識を問う問題である。下線(a)の選択肢にある lock-and-key は Emil Fischer によって提唱された分子認識の特異性をなぞらえた概念であり、下線(a)の説明とは異なる。下線(b)における「ハブ」とはネットワーク科学またはグラフ理論の分野における用語である。多数の相手と繋がるハブという概念は「ハブ空港」やインターネット接続機器の「ハブ」などのように一般的な言葉としても広まっている。空欄(c)について、一塩基多型はゲノム上の変異を言い、細胞内の状況に応じて変わるものではないため、説明にはそぐわない。ここで示した p53 の例は、p53 自体の医学的な重要性はもとより、生物物理学としても非常に面白い例と言えよう。このような例が存在することを記憶に留めていただければ幸いである。

問 56　正解【3】

並行 β シートが α ヘリックスで接続された β-α-β 構造は、TIM バレルやロスマンフォールドなどの α/β タンパク質で観察される超二次構造である。注意深く観察すると、(B)と(C)のそれぞれ N 末端付近に、β-α-β 構造があることがわかる。図に 4 つの構造の PDB ID が記されているので、分子ビューワや PDB ウェブサイト上で是非目視で確認してもらいたい。（日本語の情報としては PDBj の万見（Yorodumi）http://pdbj.org/yorodumi/ が充実している。万見ウェブサイト内で分子ビューワも利用可能である。）

問 57　正解【1】

各 Tanimoto 係数を定義に基づいて計算する。

$$T(\mathbf{q},\mathbf{a}) = \frac{3}{3+5-3} = \frac{3}{5},\ T(\mathbf{q},\mathbf{b}) = \frac{2}{3+4-2} = \frac{2}{5},\ T(\mathbf{q},\mathbf{c}) = \frac{2}{3+3-2} = \frac{1}{2},\ T(\mathbf{q},\mathbf{d}) = \frac{3}{3+6-3} = \frac{1}{2}$$

となり、低分子 Q と低分子 A の間の Tanimoto 係数が最も大きい。よって答えは選択肢 1 の低分子 A となる。

問 58　正解【4】
タンパク質の内部に埋もれたアミノ酸と露出したアミノ酸の特徴についての問題である。膜タンパク質の脂質二重層と接する領域において、外側に露出し脂質と接するアミノ酸は疎水的、内側に埋もれたアミノ酸は親水的な傾向にあるので 4 が誤選択肢である。バイオインフォマティクス分野における古典的な問題である膜タンパク質と水溶性タンパク質の判別、膜貫通領域の認識に際して、これらの特徴の違いが考慮されている。

問 59　正解【2】
RMSD(Root Mean Square Deviation)は、対応する原子間距離の二乗平均の平方根で算出される。

$$\sqrt{\frac{1.0^2 + 3.0^2 + 0 + 3.0^2 + 1.0^2}{5}} = 2.0\text{Å}$$

問 60　正解【1】
様々なドメインを分類整理するために階層的な分類体系が有用である。そのような階層的体系に基づく立体構造分類データベースである SCOP と CATH において、両者の階層は良く対応している。しかし、CATH における「アーキテクチャー」は二次構造の空間配置の類似性は考慮するが、アミノ酸配列上での出現順序は考慮しない点において、SCOP の「フォールド」とは異なり、対応する階層が存在しない。

問 61　正解【2】
非メンデル遺伝は分離・独立・優性の法則に従わない遺伝であり、典型的にはミトコンドリア遺伝や多因子遺伝がこれにあたる。量的形質に関与する遺伝子で非メンデル遺伝する場合は存在するが、量的形質遺伝子であれば非メンデル遺伝という断定は成り立たない。従って選択肢 2 が正解となる。

問 62　正解【4】
メタアナリシスは既報データを統合して行う解析であるが、統合するデータ全体に存在する(例えば成功したケースしか報告されていないなどの)バイアスが除去されるわけではない。よって選択肢 4 が正解となる。

問 63　正解【2】
通常、同義置換率は非同義置換率より大きい。よって選択肢 2 が正解となる。

問 64　正解【1】

UPGMA 法では、まず最近接 OTU(この場合は種)をクラスタリングするので、この例ではヒトとウシがクラスタする(よって選択肢 2 と 3 はこの段階で間違っている)。次の段階では、クラスタ{ヒト、ウシ}からマウスへの距離 =ヒトおよびウシからマウスへの平均距離 = (0.3+0.4)/2 = 0.35、ニワトリへの距離 = (0.7+0.7)/2 = 0.7、ワニへの距離 = (0.7+0.7)/2 = 0.7、残りの現存種間の最小距離 = ニワトリ−ワニ間 = 0.4 から、クラスタ{ヒト、ウシ}とマウスがクラスタされることがわかる(よって選択肢 4 は間違っている)。この段階で(当然この操作を最後まで繰り返しても)選択肢 1 が正解として残される。

問 65　正解【3】

NJ 法では、単一のノードからすべての OTU が分岐した星型系統樹から、系統樹の総枝長が最小になる中間系統樹を選択しつつ、OTU/ノードの組を順次引き出してゆく方法である。UPGMA 法の場合と異なり、枝長を手計算で求めることは困難である。ただし NJ 法では、リーフノードのクラスタ(種 A と種 B が祖先ノード C に接続した部分系統樹)については、A–B の距離を d_{AB}、枝 A–C の長さを l_{AC}のように表すと、$d_{AB} = l_{AC} + l_{BC}$の関係が成り立つ(ただし進化速度の一定性を仮定しないので、UPGMA 法のようにかならずしも $l_{AC} = l_{BC} = d_{AB}/2$ とならない事に注意)。もし、$d_{AB} < l_{AC} + l_{BC}$であれば、余分な枝長の半分$(l_{AC} + l_{BC} - d_{AB})/2$ をノード C から上流に至る枝に加え、枝 A–C および枝 B–C からそれぞれ減ずれば、その他の枝長を全く変更せずに総枝長を減ずることができるので、NJ 法の作成手順に違反する。例として、選択肢 1、2、および 4 のクラスタ{ヒト、ウシ}は明らかにこれに違反しているので、NJ 法による系統樹ではないことがわかる。従って選択肢 3 が正解となる。

問 66　正解【4】

選択肢 1 と 2 の記述は系統樹作成法の説明として正しい。問 64(UPGMA 法)と問 65(NJ 法)の正解系統樹の顕著な違いは、後者でマウスに至る枝がかなり長くなっていることであり、これは進化速度が高くなっていることを表す。進化速度のばらつきは、手法によって異なる系統樹が得られる主要な原因の一つであるので、選択肢 3 の記述は誤っていない。一方、ある遺伝子の水平伝播が疑われる場合は、ある種の遺伝子が相当かけ離れた種の遺伝子に近い(クラスタリングする)場合であり、この場合はマウスの遺伝子は他とくらべて枝が長くなっており、妥当な種(ヒト)とクラスタリングしているので、特に水平伝播が疑われる状況ではない。よって選択肢 4 が正解となる。

問 67　正解【3】

センチモルガンの定義は、100 回あたり 1 回の組換えの起こる遺伝子間の距離である。よって選択肢 3 が正解である。

問 68　正解【3】

子供は両親から半数染色体を受け継ぐので、反復配列の反復数変化・転座・転移・重複がなければ、父親および母親で認められたバンドのそれぞれ 1 本を組み合わせたゲルイメージが得られるはずである。選択肢 3 には母親由来のバンドが認められないので、これが正解となる。

問 69　正解【1】
KEGG は代謝パスウェイを主に扱ったパスウェイデータベースであり、図はその一例としてアミノ酸合成系の一部を示している。KEGG の代謝パスウェイでは化合物が○で示され、その化合物を代謝する酵素タンパク質が四角で代謝酵素の酵素番号（EC 番号）と共に示される。遺伝子番号は遺伝子に付く ID 番号であり、酵素番号とは異なる番号であるため、1 が不適切な説明となっている。

問 70　正解【2】
代謝化合物の同定は様々な手法で行われるが、次世代シークエンサは塩基配列を解析する機器であり、代謝物の解析を行う機器ではないため、2 は不適切な説明である。

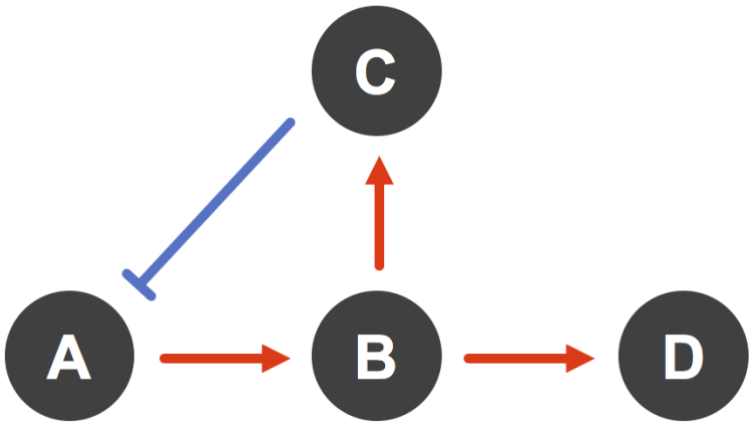
問 71　正解【3】
RNA-seq は次世代シークエンサを用いて発現している遺伝子のリード数をカウントするが、全体として読んだリードの中でのリード数であり、発現の絶対量が測定できている訳ではないため、3 は不適切な説明である。

問 72　正解【2】
Illumina 社は数多くのシークエンサを出しているが、スループットに優れた HiSeq と小規模な解析を迅速に行える MiSeq というシリーズがよく利用されている。1 ランあたりのリード数という観点では HiSeq の方が MiSeq よりも高い性能を有している。

問 73　正解【3】
ヒトゲノムのようによく解析されているモデル生物では、参照配列が存在し、その参照配列との差異としてゲノム解析を行うリシークエンスがよく行われる。これに対して、参照配列としての高品質なゲノムデータが利用できない生物では、長鎖シークエンサデータなどを用いて、参照配列に依存しない解析が行われることがあり、de novo 解析と呼ばれる。

問 74　　正解【4】

4 つの遺伝子破壊実験から予想される制御関係図は下記の図のようになり、遺伝子 A,B,C の間に正のフィードバックループは存在しないことが分かるので、4 が不適切な説明である。



問 75　　正解【3】

RNA-seq は発現している遺伝子の RNA を cDNA にして次世代シークエンサで解析する手法であり、遺伝子の発現量情報が得られる。メチル化状態は遺伝子の発現量の調整にも使われるが、特定の領域でのメチル化の有無を直接明らかにするためには、別途 Bisulfite sequencing などを行う必要がある。

問 76　　正解【1】

図は典型的なフィードフォワードループであり、遺伝子制御ネットワークに広く見られる構造である。この図のネットワークでは、抑制する方向の制御は含まれていないため、1 は不適切な説明である。

問 77　　正解【4】

細胞内ネットワークの推定としてベイジアンネットワークを用いることもあるが、遺伝子をノード、辺を相互作用として抽象化し、相互作用を条件付き事後確率として求めるモデルであり、選択肢 4 は不適切な説明となっている。

問 78　　正解【2】

過去に何度か出ている問題の改変であるが、固有値を正しく計算し、その値の正負から安定性に関する理解を問う問題である。固有値は 5,-3 であり、すべての固有値の実部が負で

は無いため、システムは不安定である。

問 79　正解【3】
網羅的な物質同定に液体クロマトグラフィーを使うことはあるが、X線による構造解析はスループットも低く 3 次元構造を決める手法であり、物質同定の目的で用いることはない。

問 80　正解【2】
同じ生物サンプルから得られるデータを、異なる手法で解析を行っても反復になっておらず、同じサンプルに対して繰り返し測定を行って始めて技術的反復となるため、選択肢 2 は不適切な表現である。

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。