平成 24 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】
たとえば真核生物でも、植物細胞は細胞壁を持っている。

問 2　　正解【2】
染色体 DNA は環状とは限らない。むしろ、真核生物では両端にテロメア構造をもつ線状構造が一般的である。

問 3　　正解【3】
カビやきのこは好気性従属栄養生物である。

問 4　　正解【1】
生体膜は両親媒性のリン脂質が二重層をつくっている。ラフトとは、その表面を筏のように、特定の成分が固まって漂う構造があるという仮説を指す。

問 5　　正解【3】
細胞内のタンパク質の輸送には、ミトコンドリアなどへの移行のように、サイトゾル中を直接移動するやり方と、小胞内に閉じ込められた形で移行する、いわゆる膜輸送系がある。

問 6　　正解【4】
ミトコンドリアも葉緑体も内膜と外膜の二重の膜で包まれている。葉緑体はこの他にチラコイド膜をもつ.

問 7　　正解【2】
グラフの A 区間の細胞が DNA 複製準備が行われてる G1 期の細胞、B 区間の細胞が DNA 複製が行われている S 期の細胞、C 区間の細胞が DNA 複製完了後に細胞分裂の準備を行っている G2 期および細胞分裂を行っている M 期の細胞に相当する。細胞周期を G1 期で阻害すると、この期の細胞が増え、それ以外の期の細胞は減っていく。

問 8　　正解【4】
メンデルの法則。胚乳の色では黄色が優性、種子の形では丸が優性の形質となり、それぞれが独立に振る舞う。雑種第一代には全部の対立遺伝子が含まれており、それらの間の組み合せと最終的な形質を数えればよい。

問 9　　正解【3】
チロシンはアミノ酸の一種で、タンパク質のリン酸化において重要な働きをするが、二次メッセンジャーではない。

問 10　　正解【1】
ウイルスは細胞のように分裂を繰り返して増殖するのではなく、宿主細胞に自己を一度に大量に複製させることによって増殖する。

問 11　　正解【2】
免疫グロブリンを産出するのは、T 細胞ではなく、B 細胞である。

問 12　　正解【1】
A と G はプリンと呼ばれる。

問 13　　正解【1】
鋳型 DNA は、文字通り合成のときの鋳型となるので、合成される DNA の相補配列を持つことになる。

問 14　　正解【2】
DNA 複製にかかわる基本的な酵素の役割を押さえておこう。

問 15　　正解【4】
キャップ構造はリボソームが翻訳を開始するために必要である。成熟 mRNA は核から細胞質に輸送された後、タンパク質合成に用いられる。

問 16　　正解【4】
rRNA はリボソームの構成成分となる RNA であり、コドンに対応するアミノ酸を運ぶのは、tRNA である。shRNA はヘアピン構造をもつ小さい RNA で、プロセシングされて、siRNA となる。

問 17　　正解【3】
HLA 型は完全に遺伝子の型として決まるもので、エピジェネティクスとは関わりがない。

問 18　　正解【3】
PCR を行うときには、まず高温状態にして、試料 DNA を 1 本鎖の状態に分離するので、二本鎖 DNA であっても問題なく増幅できる。

問 19　　正解【4】
サザンブロッティング法は、DNA 集団の中から、ある特定の配列をもつ DNA を検出するための方法なので、遺伝子の組織特異的発現を調べる目的には使えない。その目的には RNA の検出用のノーザンブロッティング法が用いられる。

問 20　　正解【1】
メタボローム解析では、細胞内に存在する代謝産物を網羅的に検出しようとするので、シークエンサではなく、質量分析計などが用いられる。

問 21　　正解【4】
分配則等を丁寧に追っていく事で分かる．

問 22　　正解【4】
無限大を表現するのは Inf (Infinite)． NaN (Not a number) は，0/0 の演算など非数を表すために使われる．

問 23　　正解【1】
HiSeq 2000 は，次世代シークエンサの種類であり，インタフェース名ではない．ちなみに，USB3.0 は USB2.0（現在多く使われている USB）の拡張規格，シリアル ATA はハードディスクの接続規格，IEEE 1394 は Mac では Firewire とも呼ばれる AV 機器や外部ハードディスクを接続する規格．

問 24　　正解【2】
ディスプレイの障害は，データの読み込み速度とは独立．

問 25　　正解【3】
外部ソートは，メモリに載らない巨大データに対しハードディスク等を利用してソートする手法の総称．実装としては，マージソートや基数ソートなどが利用される．

問 26　　正解【2】
スタックの状態を順に再現すると（右がスタックの上位，括弧内は POP で取り出されるデータ）
a => ab => abc => abcb => abc(b) => ab(c) => abb => ab(b) => a(b)

問 27　　正解【1】
クイックソートは最悪の計算量が O(n^2)．更に，任意の数列を O($n \log n$)でソートできる方法にはマージソートなども存在する．

問 28　　正解【2】
ヒープは，必ず親より子が大きい必要があるので，(a)は 2 以上 17 未満．(b)は 4 以上 15 未満．(c)は(a)以上である．

問 29　　正解【2】
二分操作は, 要素を節点もしくは葉に持つ二分木を作りつつ辿るので, 回数は木の高さの高々 O(log n)回．また，二分探索を実際に行うと，操作 0 回で見つかる値は 6．操作 1 回で見つかる値は，3 と 9 (それぞれ初めの操作で，6 を基準に左右に 3 要素ずつ分割した，中央の値)．残りは，操作 2 回目で見つかる．

問 30　　正解【4】
1〜3 は文字列の対応付けに利用される手法．4 は最小値・最大値を求める数値解析の手法．

問 31　　正解【3】
組が 3 以上は, 2402, 2403, 2405 の 3 名. そのうち, 評価点が 50 より大きいのは 2403, 2405. 出席日数が 150 日より多いのは，2403．これらの”OR”をとるので，2403 と 2405 の 2 名が答え．

問 32　　正解【1】
データをオブジェクトデータとして扱うのは，オブジェクトデータベース．

問 33　　正解【3】
確率密度は 1 を超えることがある（例えば，非常に分散が小さい場合）が，累積密度関数は 0 以上 1 未満の値のみを取る単調増加の関数である．

問 34　　正解【2】
尤度関数を最大にするのは，最尤推定．

問 35　　正解【1】
条件付き確率の定義より，P(A, B) = P(A) P(B | A)である．この確率の式が条件 C の基で起こるとすると，P(A, B | C) = P(A | C) P(B | A, C).

問 36　　正解【3】
帰無仮説が棄却されない場合は，対立仮説を採択できる有意な差が認められないだけで，帰無仮説が採択されるわけではない.

問 37　　正解【4】
箱のなかからボールを取り出す事象は，ポアソン分布や二項分布で表される．また，RNA-seq では，ポアソン分布に分散の項を追加したと解釈できる負の二項分布が用いられる事も多い.

問 38　　正解【4】
TELNET は平文で通信を行うため，盗聴される恐れが高く，SSH が用いられる．FTP も同様の問題があり，暗号化した SFTP が使われる場合も少なくない.

問 39　　正解【2】
特になし.

問 40　　正解【4】
SVM は凸最適化問題かつ二次計画問題として定式化されている．分離面は線形カーネルを用いた場合，平面，特にデータ，あるいは，写像空間が高次元の場合は，平面を一般化した超平面で表される.

問 41　　正解【4】
(A)と(B)のアラインメントで異なっている、中央 3 塩基のアラインメントスコアを考えると、(A)は $2m+2g$、(B)は $m+2u$ となり、(A)が(B)より大きくなる(m,u,g)の組み合わせは 4 のみ。

問 42　　正解【1】
非加重結合法では、距離が最小のクラスター対を併合し、併合されたクラスターと他のクラスターとの距離を平均値で置き換えるという操作を繰り返す。本問では、まず A と C が距離 0.2 で併合され、次に B と D、および(A,C)と E が距離 0.4 で併合され、最後に((A,C),E)と(B,D)が距離 0.6 で併合される。

問 43　　正解【4】
条件付き確率の公式 $P(A|B) = P(A \cap B)/P(B)$に当てはめて計算する。

$P(\mathrm{T}^1) = 0.05 + 0.06 + 0.07 + 0.07 = 0.25$（4 行目の合計値）、$P(\mathrm{T}^1 \cap \mathrm{G}^2) = 0.07$（4 行 3 列目の値）から、$P(\mathrm{G}^2|\mathrm{T}^1) = P(\mathrm{T}^1 \cap \mathrm{G}^2)/P(\mathrm{T}^1) = 0.07/0.25$ となる。

問 44　　正解【2】

「相同」は進化的に同一の起源に由来するという意味を持つ。一般にアラインメントをとる目的は、配列間で相同な部位を対応づけることにあるといえるが、実際にアラインメントされた配列間に常に相同性があるとはいえない。

問 45　　正解【4】

プロファイルの開始位置が指定されていることに注意。5 残基目以降、もっとも大きなスコアをとるアミノ酸が、順に V, D, F, ... となっていることに着目する。

問 46　　正解【3】

GC 含量の違いは、水平伝搬の存在を示す有力な手がかりの一つである。

問 47　　正解【3】

source の mol_type が genomic DNA となっているので、cDNA ではなく、ゲノム DNA から解読されたことがわかる。

問 48　　正解【2】

問題文中の式に当てはめて考えればよい。n が 8 倍になったとき、S を 3 大きくすれば 2^{-S} が 1/8 になるので、ちょうど打ち消して E が同じ値になる。

問 49　　正解【1】

GC skew 解析は DNA2 重鎖間での非対称性に着目した解析で、原核生物ゲノムの複製開始点の予測などに用いられる。なお、自分自身の配列に対して相同性検索を行うと、同じ位置を合わせる自明なアラインメントのほかに、異なる位置で類似した配列として繰り返し配列を検出することができる。

問 50　　正解【2】

ローカルアラインメント（スミス・ウォーターマン法）は、グローバルアラインメント（ニードルマン・ブンシュ法）と類似した動的計画法により求められるが、0 より小さいスコアになるとき 0 にリセットするという操作などが加わる。単に両端のギャップペナルティを 0 に設定するのはセミグローバルアラインメントと呼ばれ、部分配列どうしをアラインする際などに用いられる。

問 51　　正解【2】

アラインメント中のギャップ「.」には、欠失状態に対応するものと、他配列の挿入に対する穴埋めに用いられているものとがある。Seq4 の 2 文字目は、他配列がマッチ状態であることから欠失状態 D_2 に対応するが、5 文字目は Seq3 の挿入に対する穴埋めであり、対応する状態はない。あとは大文字をマッチ、小文字を挿入状態として対応するパスを探せばよい。なお、4 はパスではない（連続していない）ので不適。

問 52　　正解【3

結局、is_a の関係のみでつながるパスについては is_a、途中に part_of が入るパスについては part_of の関係になる。1, 2, 4 はこの規則を満たす（1 と 2 は同じノード間の異なる 2 つのパスに対応する）が、3 は矢印の向きが異なるためにパスをとることができず、推論規則が適用できない。

問 53　　正解【2】

解説：活性部位など機能に必須の残基や，ジスルフィド結合など立体構造の維持に不可欠な残基は，保存性が高く，同一のアミノ酸が縦にそろって並ぶ傾向がある．よって選択肢 1, 3 は適切である。また，タンパク質内部は密にパッキングする必要があるため，分子表面に比べ，残基の挿入や欠失は起りにくい.よって選択肢 4 は適切である．活性部位は解離基を持つ極性のアミノ酸（H,C,E,D,Y,K,R,S など）が担うことが多く，アラニン(A)などの疎水性のアミノ酸が活性部位となることは稀である。よって，選択肢 2 は不適切である．

問 54　　正解【4】

解説：構造決定法はヘッダーEXPDTA 行，エントリーID は HEADER 行，二次構造は HELIX 行や SHEET 行に記載されている．原子座標(x, y, z)は対応する原子の ATOM 行の 7~8 カラムの実数であり，MET 1 の Cα(CA)の z 座標値-14.049 の方が大きいので選択肢 4 が不適切である．

問 55　　正解【1】

解説：結合角(θ)はベクトル(化学結合 $\boldsymbol{a}$, $\boldsymbol{b}$)の内積が $\boldsymbol{a}\cdot\boldsymbol{b} = |\boldsymbol{a}||\boldsymbol{b}|\cos\theta$ である事を使って求めるのが一般的である。よって選択肢 1 が正解である。

問 56　　正解【3】

解説：RMSD は対応する原子間距離の二乗平均の平方根であるので，$\{(3.0^2 + 1.0^2 + 0.0^2 + 1.0^2 + 3.0^2)/5\}^{1/2} = 2.0$Å(選択肢 3)が正解である。

問 57　　正解【4】

解説：PDB は立体構造データベースである．β α β は二次構造がこの順番で出現する高頻度に観察される構造(超二次構造)を指す．CATH は SCOP と並ぶタンパク質構造分類のデータベースである．GEO(Gene Expression Omnibus)はアレイに基づく遺伝子発現情報のデータベースであるので，選択肢 4 は不適切である

問 58　　正解【3】

解説：分子力学法ではポテンシャルエネルギーを原子の中心点だけから簡便に計算する。原子軌道からエネルギーを計算する方法は，分子軌道法と呼ばれる．分子動力学法は，運動方程式を用いて分子構造の時間変化を計算する手法一般のことであり，使用するエネルギー関数には原理的には制限はない．よって，選択肢 3 が最も不適切である． 分子力学法のエネルギー，分子軌道法のエネルギーのどちらを用いても分子動力学法を行うことができるが，生体高分子の場合，計算コストの低い分子力学法のエネルギー(AMBER, CHARMM など)を用いることが多い．

問 59　　正解【1】

解説：蛋白質の核酸結合部位には正電荷を持つアミノ酸（アルギニンかリジン）が多く観察されるが，グルタミン酸とアスパラギン酸は通常負電荷を持つアミノ酸である．よって 選択肢 1 が最も不適切である．

問 60　　正解【3】

解説： アミノ酸配列の一致度が３０％以下であっても，立体構造が類似し，相同である（同じファミリー，スーパーファミリーに属する）と考えられている蛋白質ペアは数多く存在する．よって，選択肢 3 が最も不適切である．

問 61　　正解【2】

解説: SCOP はタンパク質立体構造を階層的に分類する．最上位階層はフォールド(類似した折り畳みパターンを持つグループ。相同でなくてもよい)であり，以降スーパーファミリー(相同である可能性が高い遠縁のグループ)，ファミリー(明らかに相同である近縁のグループ)と続く.よって，フォールドがスーパーファミリーに分類され，スーパーファミリーがファミリーに分類され，分類群の総数はこの順番で大きくなるので選択肢 2 が正しい．

問 62　　正解【1】

解説：リボン模型では，α ヘリックスはらせんを巻いた形で，β ストランドは矢印の形で表現される．配列上，最初のアミノ酸を N 端，最後のアミノ酸を C 端と呼ぶ．N 端から順番に二次構造をたどる事で，選択肢 1 が正解であることが分かる．

問 63　　正解【3】

解説 : リボン模型の見方は問 62 の解説の通りである. 免疫グロブリンフォールドは 2 層のβシートが特徴であり,逆平行のβシートが多い. TIM バレルフォールドはαヘリックスとβストランドが交互に現れ,全体として平行βシートの樽(バレル)状のフォールドとなる. ロスマンフォールドもαヘリックスとβストランドが交互に現れる平行βシートの構造であるが, 樽状ではなく開いた形状である. 以上を考え合わせると選択肢 3 が正解であることが分かる.

問 64　　正解【4】

解説:右図において, ヘリックス(長さ 1.5×40 =60Å)を斜辺, 膜の厚さ(30Å)を短辺とする直角三角形(斜辺:短辺=2:1. いわゆる 30-60-90° 直角三角形)が形成されることが分かるので, 問題の角度は 60°(選択肢 4)である.

問65　　正解【1】

パブリケーション・バイアスはメタアナリシスに混入する大きなバイアスのうちの１つである。

問66　　正解【3】

選択肢 1, 4 は r^2 の「平衡」と「もっとも極端な不平衡」の定義であり、選択肢 2 は「平衡」とは 2 座位の独立を意味すること・独立の場合の頻度は積であることから、これも定義であり、選択肢 1, 2, 4 はいずれも正しい。

問67　　正解【3】

「D'=1」は2つのSNPが作りうる4種類のハプロタイプのうち1つ以上のハプロタイプが存在しない (頻度が0) の状態を表す。選択肢3の場合は4種類のハプロタイプ頻度が 0 でないことが、数値の大小関係のみから確認できるから、3 は誤りである。

問 68　　正解【3】

現在の配列データは、親子関係ではなく類縁関係。祖先配列は直接知りようがなく、データとして導入できるのは、何らかの知識から注目している分類群よりも前に分岐した外群である。

問 69　　正解【2】

染色体を転座しただけでは、配列類似度をベストヒットで評価する際に大きな影響はない。

問 70　　正解【4】

遺伝子の同義置換速度は、世代時間の影響をうけるため、種により異なる。一方、同じ種のペアで、同義置換速度を計算すると、コードされているタンパク質によらずほぼ同じ値となる（同義置換はタンパク質レベルの選択圧かたは自由なので）

問 71　　正解【1】

進化距離から計算されるのは、近隣結合法。最大節約法と最尤法は、いずれもアラインメントの各サイトの状態（形質状態）によって計算されるが、再適性の尺度が異なる。

問 72　　正解【2】

系統樹のトポロジー探索については、「分子系統学への統計的アプローチ（Yang,Z-H 著、加藤、大安、藤　訳、共立出版）の 3.2 参照。

問 73　　正解【3】

ノードの次数分布がべき乗則に従うときに、そのネットワークはスケールフリー性を持つと言われ、さまざまな現実社会で見られるネットワークが持つ特徴として注目されている。スケールフリー性を持つネットワークでは、多数のエッジを持つ少数のノードが存在する事が知られており、ハブノードと呼ばれる。2 つ目の特徴として、ネットワークのノード間の距離に関する特徴として、スモールワールド性がある。スモールワールド性を持つネットワークでは、任意の２つのノード間の最短距離が、ノード数に対してログオーダーでしか増加しないことが知られており、巨大なネットワークに於いても少数のエッジをたどることで、異なるノードへとアクセスすることができる。3 つ目の特徴として、ノードの集まり具合を表す指標としてクラスター性がある。一般的にグラフにおいて、ノードが集まっている度合いを測る指標としてクラスター計数がある。ソーシャルネットワークなど現実社会に見られる多くのネットワークに於いては、ランダムグラフに比べて非常に高いクラスター係数を持つ事が知られている。

この設問においては、バラバシ-アルバートの提唱したネットワークの生成ルールに従って構成したネットワークが、これら３つの特徴のうち、スケールフリー性とスモールワールド性は持つが、クラスター計数は非常に低い事を説明する文章となっている。モデルに関する知識がなくとも、それぞれの性質がネットワークのどのような特徴を記述する指標かを理解していれば容易な問題である。

問 74　　正解【3】

一階の微分方程式を差分方程式に置き換え、初期値から逐次更新で数値積分を行うオイラー法に関する問題。1.0s の値を計算して、その値を用いて 2.0s の値を求める。

問 75　　正解【4】
INSDC は、国際的な塩基配列データベースで、International Nucleotide Sequence Database Collaboration の略称なので、不適切。

問 76　　正解【2】
ハイスループットな実験法では、日進月歩で改善が進んでいるが、相互作用しないのに相互作用が実験的に検出される偽陽性も、相互作用するのに実験的に検出出来ない偽陰性も多い。細胞内局在を考慮しないことも偽陽性の原因の一つであるが、問題文では、偽陰性となっている部分が不適切。

問 77　　正解【2】
技術的反復では、同じサンプルを複数に分けて測定する事であり、測定後のデータを異なる計算手法で解析し直すことではないので、２が不適切。

問 78　　正解【1】
サポートベクトルマシンは、訓練データセットをもとに学習する典型的な教師あり学習法である。

問 79　　正解【1】
エキソンアレイではエキソン領域（つまり既知の遺伝子領域）のプローブが必要であるし、cDNA アレイも既知の遺伝子領域のプローブが必要なので、未知の遺伝子発見には適さない。化合物マイクロアレイは、そもそも遺伝子検出の手法ではない。

問 80　　正解【4】
ダイナミックモデルにおける固有値は、システムに摂動を与えたとき、どれだけすばやく元の情報にもどれるかを示す尺度であり、システムのロバストネスを評価する指標の一つである。固有値から、そのシステムの挙動（安定/不安定）を判定することができる。