受験番号 Examinee number					

東京大学 大学院新領域創成科学研究科 情報生命科学専攻

Department of Computational Biology, Graduate School of Frontier Sciences
the University of Tokyo

平成 20（2008）年度

2008 School Year

大学院入学試験問題 修士・博士後期課程

Graduate School Entrance Examination Problem Booklet, Master's and Doctoral Course

専 門 科 目

Specialties

平成19 年 8 月 9 日（木）

Thursday, August 9, 2007

13：00～15：00

注意事項 **Instructions**

1. 試験開始の合図があるまで、この冊子を開いてはいけません。
   Do not open this problem booklet until the start of examination is announced.
2. 本冊子の総ページ数は 35 ページです。落丁、乱丁、印刷不鮮明な箇所などがあった場合には申し出ること。
   This problem booklet consists of 35 pages. If you find missing, misplaced, and/or unclearly printed pages, ask the examiner.
3. 解答には必ず黒色鉛筆（または黒色シャープペンシル）を使用しなさい。
   Use black pencils (or mechanical pencils) to answer the problems.
4. 問題は 12 題出題されます。問題 1～12 から選択した合計4問に解答しなさい。ただし、問題 1～12 は同配点です。
   There are twelve problems (Problem 1 to 12). Answer four problems out of the twelve problems. Note that Problem 1 to 12 are equally weighted.
5. 解答用紙は計４枚配られます。各問題に必ず１枚の解答用紙を使用しなさい。解答用紙に書ききれない場合は、裏面にわたってもよい。
   You are given four answer sheets. You must use a separate answer sheet for each problem. You may continue to write your answer on the back of the answer sheet if you cannot conclude it on the front.
6. 解答は日本語または英語で記入しなさい。
   Answers should be written in Japanese or English.
7. 解答用紙の指定された箇所に、受験番号と選択した問題番号を記入しなさい。問題冊子にも受験番号を記入しなさい。
   Fill the designated blanks at the top of each answer sheet with your examinee number and the problem number you are to answer. Fill the designated blanks at the top of this page with your examinee number.
8. 草稿用紙は本冊子から切り離さないこと。
   The blank pages are provided for making draft. Do not detach them from this problem booklet.
9. 解答に関係ない記号、符号などを記入した答案は無効とします。
   An answer sheet is regarded as invalid if you write marks and/or symbols unrelated to the answer on it.
10. 解答用紙・問題冊子は持ち帰ってはいけません。
    Do not take the answer sheets and the problem booklet out of the examination room.
11. 受験生の便宜のために問題の英訳が 19-33 ページに掲載されていますが、正式な問題文は日本語によるものです。
    For the convenience of applicants, problems are translated into English (pp.19-33). Note that Japanese version is the formal one.

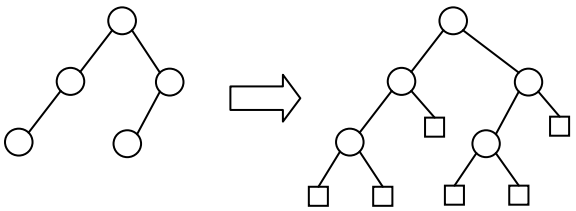
（このページは草稿用紙として使用してよい）
(Blank page for draft)

（このページは草稿用紙として使用してよい）
(Blank page for draft)

（このページは草稿用紙として使用してよい）
(Blank page for draft)

問題 1

二分木において、子のない節点に特別な節点を以下のように補完した構造を考えよう。○を「内部節点」、□を「外部節点」、節点から二分木の根までの距離を「深さ」と呼ぶ。また、内部節点の数を n とする。



（1）内部接点どうしを結ぶ辺の数が $n-1$ となることを証明せよ。

（2）外部節点数が $n+1$ であることを証明せよ。

（3）根からの深さが k で外部節点を 2 個持つ内部節点を一つ削除する操作を考える。外部節点の深さの総和の変化量と、内部節点の深さの総和の変化量を、それぞれ k を用いて表せ。

（4）外部節点の深さの総和が、内部節点の深さの総和+ $2n$ になることを証明せよ。

（5）n 個ある内部節点の深さの総和が最小、また最大となる構造はそれぞれどのようなものか、説明せよ。

問題2

ベクトル空間 V から同じベクトル空間への写像 T が

1. $T(\mathbf{x}+\mathbf{y}) = T(\mathbf{x})+T(\mathbf{y})$
2. $T(c\mathbf{x}) = cT(\mathbf{x})$

を満たす場合、T を V 上の線形写像と呼ぶ。

（1）V の基底を $\{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n\}$と書き、各基底の線形写像 T による変換先を $\{\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_n\}$ とする。これらの記法を用いて V 上のベクトル $\mathbf{x}$ の写像 $\mathbf{y}=T(\mathbf{x})$が行列演算で表現できることを示せ。

（2）写像 T の逆写像に対応する行列を逆行列とよび、T^{-1} で表現する。行列 A, B がともに逆行列を持つとき、その積 AB も逆行列を持ち、$(AB)^{-1} = B^{-1}A^{-1}$ であることを証明せよ。

（3）線形写像 T を二つの基底 $\{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n\}$, $\{\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_n\}$ を用いてそれぞれ A, B と行列表現する。基底 $\{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n\}$ から $\{\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_n\}$ への変換行列を P とすると $B = P^{-1}AP$ が成り立つことを証明せよ。

（4）*n* 次正方行列 A がスカラーλ に対して $A\mathbf{x} = \lambda\mathbf{x}$ を満たすとき、λ を A の固有値という。$A^2 = A$ を満たし、逆行列を持つ *n* 次正方行列 A の固有値を全て求めよ。

（5）行列 A の固有値を $\{\lambda_1, \lambda_2, ... \lambda_n\}$ とするとき、行列 A^m の固有値を求め、その導出を解説せよ。

問題3

下は 0 から M−1 までの正整数が入った長さ N の配列を分布数え上げ(カウンティング)ソートするアルゴリズムである。以下の問いに答えよ。

Program

```
for (int i=0; i< M; i++) C[i] = 0;
for (int i=0; i< N; i++) C[value[i]] ++;
for (int i=1; i< M; i++) C[i] += C[i−1];
for (int i=N−1; i >= 0; i−−) sorted[−−C[value[i]]] = value[i];
```

（１）配列 C の役割を記せ。
（２）プログラム３行目の処理目的を記せ。
（３）プログラム４行目において i を昇順でなく降順に変化させ sorted 配列を作成する目的は何か。
（４）このプログラム実行に要する計算量を M, N を用いて表せ。
（５）一般に n 個の要素のソーティングに要する計算量が少なくとも O(n log n)かかる理由を説明せよ。必要であれば、スターリングの公式 $n! \approx \sqrt{2\pi} e^{-n} n^{n+(1/2)}$ を用いてよい。

問題4

長さ n の配列に入った 1 から n までの数字をランダムに並び替える場合を考える。数字 i とその配列における位置とが全ての $i = 1,..., n$ について一致しない場合の数を $C(n)$ とする（ただし配列位置は 1 から始まるとする。）これから $C(n)$ を求めよう。

（１）数字 1 が配列位置 i (i≠1) にあり数字 i が配列位置 1 にある場合の数を、C を使って表せ。

（２）数字 1 が配列位置 i (i≠1) にあり数字 i が配列位置 1 にない場合の数を、C を使って表せ。

（３）上記の結果より C について成り立つ漸化式を求めよ。

（４）上記の漸化式を解け。

（５）$n \to \infty$ としたとき、配列位置とその中の数字が全て一致しない確率を求めよ。必要なら $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$ を利用してよい。

問題5

以下の(1)ー(4)の語句の組について、語句の意味と両者の関係を説明しなさい。(1)ー(4)それぞれについて２００字程度で記述すること。

(1) オーソログ　シンテニー
(2) 遺伝子発現プロファイル　マイクロアレイ
(3) ホモロジー検索　スコア行列
(4) QTL　DNA マーカー

問題6

２次元ドットプロット法は2つのDNA配列間で高度に保存された配列部分および構造的類似性や違い等を視覚的に同定する方法である。たとえば、ヒトとマウスのオルソログ遺伝子を含むゲノム領域（約5 kb）のドットプロットを図１に示す。ここでは一方の配列を20塩基のウィンドウ幅で10塩基ずつずらしながら、もう一方の配列の両鎖を全域スキャンし、配列類似度が90 %以上となるアラインメントを２次元にドットしている。すると図１に示したように保存領域 a, b, c が見かけ上多くのドットが連続した線として現れる。このことを参考に以下の問いに答えよ。なお、図中の矢印はすべて配列の5’から 3’ の方向を示す。また、実際の解析では多くの有意でないドットがバックグラウンドとして現れるが、この問題ではこれらのドットは表示しない。

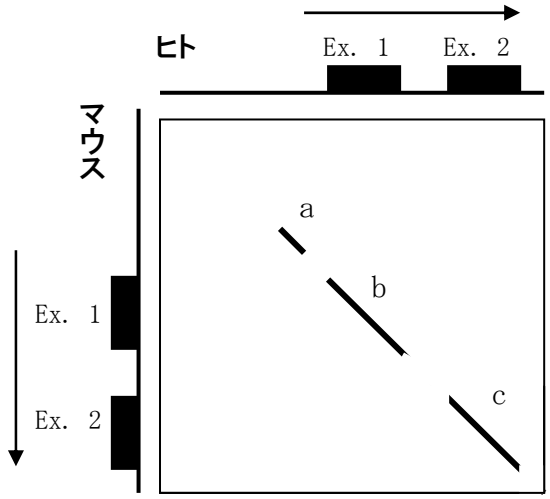


図１

（1）図１中のbとcは、ヒトとマウス間で保存された第１（Ex. 1）及び第２エキソン（Ex. 2）に由来する。bとcの間に存在する保存性が低い領域を何と言うか。また、第１エキソンの上流にある保存された領域aは何であると考えるのがもっとも適当か。

（２）配列２（約 500 kb）中には 90 %以上の配列類似度をもつ 2 つの同じ方向をもった重複領域（約 100 kb）が存在する（配列２上の矢印した部分)。配列２同士で配列類似度が 90 %以上の領域のみをドットプロットしたときの図はどのようになるか。解答用紙にフリーハンドで図２を書いて、その中にドット図を描け。ただし、線の長さ及び位置は必ずしも正確でなくても良い。

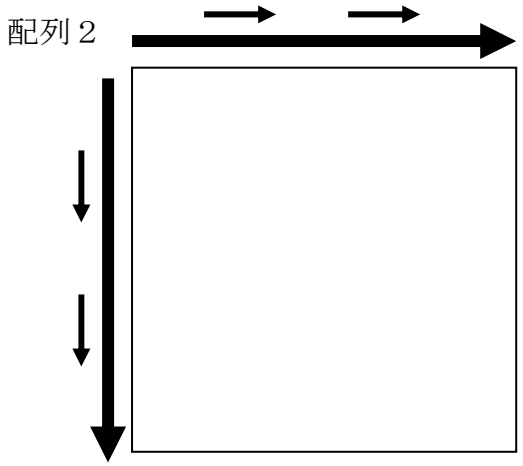


図２

（３）配列３同士をドットプロットしたとき図３のような結果を得た。配列３にはどんな種類の配列構造が存在するか、一般名で記せ。また、その配列の存在する位置と配列の向きを図２に示したような書き方で、解答用紙に図３を写して、配列３の上に書け。ただし、線の長さ及び位置は必ずしも正確でなくても良い。

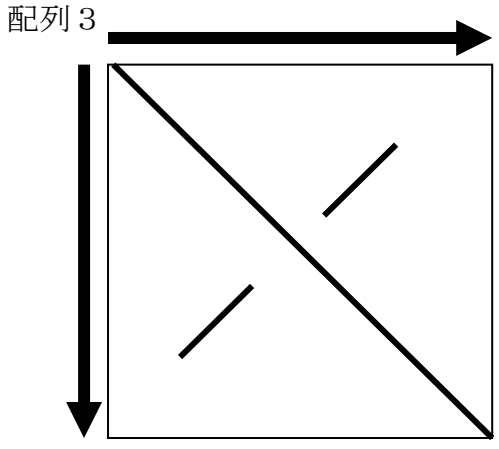


図３

（4） 細菌ゲノム４とゲノム５（ともに約 8 Mb）を比べると図 4 のドットプロットを得た。この結果から 2 つのゲノムについてその配列構造の関係を図式化し、考えられる互いの構造的違いを簡単に文章で説明せよ。ただし、線の長さ及び位置は必ずしも正確でなくても良い。

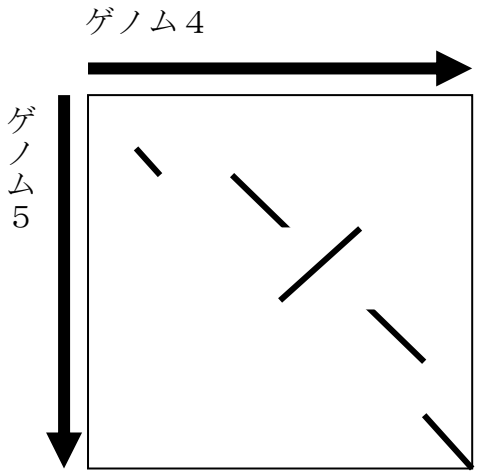


図４

問題7

外見上見分けがつかない 12 個の物質 $\{a, b, c, d, e, f, g, h, i, j, k, l\}$ のうち、1 つだけ重さが異なり、他の 11 個の重さは等しいと仮定する。以下の各々の場合について、天秤を 2 回だけ使って異なる重さの物質をみつけ、さらにそれが他の重さが等しい物質に比べ軽いか重いかを判定する方法を設計し、判定できる理由を述べよ。

（1） 重さが異なる物質が a, b, c の中にあることが分かっている場合。
（2） 2 つのグループ $\{a, b, c, d\}$ と $\{e, f, g, h\}$ の重さを天秤で比較したとき、等しいことが分かっている場合。
（3） グループ $\{a, b, c, d\}$ が $\{e, f, g, h\}$ より重いことが分かっている場合。

問題8

（1） 下記に 3 文字コードで示したアミノ酸残基の中から、A から E のそれぞれの条件を満たすアミノ酸残基ペアを重複のないように選べ。

アミノ酸残基

GLU PHE PRO VAL GLY CYS TYR MET ARG ILE

条件

A. 硫黄 1 原子を含む。

B. 側鎖 6 員環同士の相互作用によって蛋白質構造を安定化することができる。

C. 他のアミノ酸残基と比べて、このペアがとりやすい主鎖 2 面角の値の範囲は大きく異なる。

D. ペアの一方の側鎖のプロトン１つを１つのメチル基で置換することでもう一方が得られる。

E. 側鎖間で塩橋を形成することができる。

（2）同じリガンドを結合できる野生型の蛋白質とその変異型を考えよう。温度・圧力一定の条件のもとで、下記のように 4 種類のギブス自由エネルギー変化を定義する。

ΔG_1: リガンド非結合状態であった野生型蛋白質がリガンドを結合したときの自由エネルギー増加量。

ΔG_2: リガンド非結合状態であった変異型蛋白質がリガンドを結合したときの自由エネルギー増加量。

ΔG_3: リガンド非結合状態の野生型蛋白質と比較した、リガンド非結合状態の変異型蛋白質の自由エネルギー増加量。

ΔG_4: リガンド結合状態の野生型蛋白質と比較した、リガンド結合状態の変異型蛋白質の自由エネルギー増加量。

このとき以下の問いに答えよ。

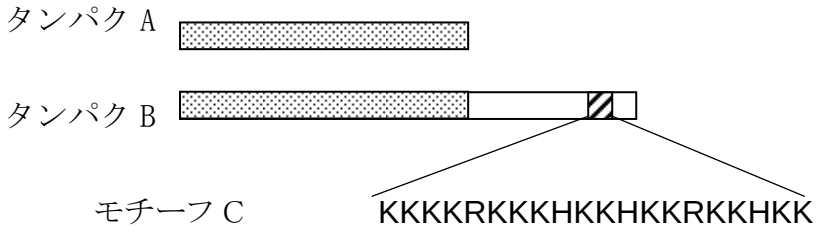
（２－１）野生型と変異型の蛋白質が安定にリガンドを結合するためには ΔG_1 と ΔG_2 の符号はそれぞれどうなればよいか答えよ。

（２－２）$\Delta G_1, \Delta G_2, \Delta G_3, \Delta G_4$ の満たすべき関係式を答えよ。

（２－３）野生型と変異型のどちらがリガンドをより安定に結合するかを判定するために最適な量αを定義し、その判定条件を説明せよ。

問題9

遺伝子 X は、ヒトの培養細胞 A と B で発現している。この遺伝子のエキソン 1 の全長をプローブとしたノーザンブロット解析を行ったところ、A 細胞では約 2.0 kb、B 細胞では約 3.3 kb の mRNA が検出された。さらに、それぞれの cDNA をクローニングして、全長のシークエンス解析を行うと、A 細胞の遺伝子 X の cDNA のタンパク質をコードする領域は 1,854 bp、B 細胞からの cDNA は 3,012 bp であり、最初の 1,356 bp は全く同一の配列であることがわかった。cDNA から予測される、A 細胞のアミノ酸配列（タンパク A）と B 細胞のアミノ酸配列（タンパク B）を比較したところ、タンパク B のカルボキシル末端領域には、特徴的なモチーフ配列（モチーフ C）がただ一つあることがわかった。



（1）　遺伝子 X は、染色体上の 1 つの遺伝子座から発現することがわかっている。しかし、A 細胞と B 細胞では異なる mRNA が発現している。このような機構を何というか。

（2）　図にはモチーフ C のアミノ酸配列をアミノ酸の 1 文字表記で示している。モチーフ C を構成しているアミノ酸に共通の特徴は何か。

（3）　細胞内における局在を調べるために、タンパク A とタンパク B の共通部分であるアミノ酸 1~35 番目を抗原とする抗体 a と、タンパク B のみに存在するアミノ酸 612~635 番目を抗原とする抗体 b の 2 種類の抗体を作製した。A 細胞および B 細胞をそれぞれ固定して免疫染色を行ったとき、抗体 a および b による A 細胞および B 細胞における予想される染色パターンを、根拠とともに示せ。

（4）　A 細胞と B 細胞の細胞全体の抽出物を用いて、抗体 a を使ったウェスタンブロット解析を行った。A および B 細胞で検出されると考えられるタンパク質の大きさをそれぞれ理由とともに示せ。アミノ酸 1 残基の平均分子量は 133 とする。

（5）　A 細胞からタンパク A を精製して、これを抗原としてモノクローナル抗体 c を作製した。抗体 c を用いてウェスタンブロット解析を行ったところ、A 細胞の抽出物では抗体に反応するバンドが検出されたが、B 細胞の抽出物ではバンドは検出されなかった。抗体 c は、どのようなエピトープ（抗原決定基）を認識していると考えられるか、２通り答えよ。

問題10

ヒトの細胞は 22 対の常染色体と 1 対の性染色体をもっている。細胞分裂期になると染色体は高度に凝縮し、特殊な染色方法で各染色体を区別できる。一方、間期の細胞では染色体の凝縮度は低く、(　A　) と呼ばれる領域では遺伝子の転写が起こっている。しかし間期の細胞でも、女性の細胞では2本のX染色体のうち1本は高度に凝縮した(　B　) になっており転写が不活性な状態になっている。
染色体はDNA合成期に複製され、細胞分裂期に 2 つの娘細胞に分配される。

（1）括弧（A）（B）に適当な語を入れなさい。

（2）ヒトのゲノムには約 30 億塩基対の DNA がある。1 塩基対の平均分子量を 635 として、細胞 1 個あたりに含まれる DNA の重さを、計算式を明記して求めなさい。アボガドロ定数は 6×10^{23} とする。

（3）染色体 DNA は高度に凝縮して直径数 μm の核内に納まっている。凝縮した DNA の最も基本的な単位であるヌクレオソームの構造を説明しなさい。

（4）X染色体の不活性化はメスの哺乳類細胞で一般に起こる現象である。ある一群のネコを多数観察した結果、メスでは体色が（白黒）（白茶）（白黒茶）の 3 通りがあったが、オスでは（白黒）（白茶）の二通りしかいなかった。（白黒茶）のネコがメスにしかいなかった理由を考察し、X染色体不活性化と関連させて説明しなさい。

（5）下線部について、DNA が染色体として機能するためには特徴的な DNA 塩基配列をもつ 3 つの領域が必要である。この 3 つの領域の名前をあげ、それぞれの機能を説明しなさい。

問題11

ショウジョウバエを化学物質Xで刺激すると、独特のY行動をとる。このY行動の変異体を単離するために、変異剤で処理したオスを正常のメスと交配させた。得られたF1集団から、Xを加えてもY行動を取らない表現型（L型）を示す変異体Aと、Xがないのに Y行動を取る表現型（H型）を示す変異体Bを単離した。このとき、以下の問いに答えよ。なおそれぞれの変異の原因遺伝子を*A*と*B*として、変異型アレルを*Am*と*Bm*、野生型アレルは+と表記する。またどちらの遺伝子も常染色体上にあり、互いに連鎖はないものとする。

（１）*Am*は、機能欠損型変異を持つと考えられるにも関わらず、優性を示す。なぜだろうか？考えられるメカニズムを簡潔に説明せよ。

（２）遺伝子*A*と*B*のどちらが遺伝学的により上流で機能するのかを調べる目的で、*A*のヘテロ接合体（*Am*/+）と*B*のヘテロ接合体（*Bm*/+）の交配を行なった。得られた集団における３種の表現型（野生型、L型、H型）の分離比が、*A*が*B*よりも上流で働いている場合と下流で働いている場合とで、それぞれどのようになると期待されるか、簡潔に説明せよ。

（３）F1集団中には、Y行動に関する遺伝子の劣性変異アレルを持つヘテロ接合体も含まれている筈である。それらを同定する為に、まず野生型表現型を示すF1個体Cと野生型の交配でF2集団を得た。次に、このF2集団をF1個体Cに戻し交配してF3集団を得た。F1個体CがY行動に関する遺伝子の劣性変異アレルを持っていた場合、F3集団の何%が変異型の表現型を示すと期待されるか、簡潔に説明せよ。

（４）問（３）の実験の結果、F1個体CはL型表現型を示す劣性変異アレル*c*とその野生型アレルのヘテロ接合体（*c*/+）であることが分かった。同様の交配実験によって、F1個体DもL型表現型を示す劣性変異アレル*d*とその野生型アレルのヘテロ接合体（*d*/+）であることが分かった。この２つの変異体の原因遺伝子*C*と*D*が同一遺伝子であるか否かを調べるには、どのような交配実験を行なえばよいかを、同一である場合とそうではない場合のそれぞれにおいて期待される結果とともに、簡潔に説明せよ。

（５）遺伝子 *B* と *C* のどちらが遺伝学的に上流で機能するのかを調べるためには、どのような交配実験を行なえばよいかを、*B* が *C* よりも上流で働いている場合と下流で働いている場合のそれぞれにおいて期待される表現型（野生型、L 型、H 型）の分離比とともに、簡潔に説明せよ。なお遺伝子 *C* も常染色体上にあって、*B* とは連鎖を示さないものとする。

問題12

真核生物のゲノム中の大きな部分は反復配列で占められており、その率はヒトでは全ゲノムの約 35 %、トウモロコシでは 50 %以上に及ぶ。反復配列の多くは転移性因子によって作られたと考えられている。あるタイプの転移性因子（「クラス 1」とする）はレトロウィルスのように DNA から特定の部分がいったん RNA に転写され、それが DNA に逆転写されてゲノムの別の場所にコピーされる。別のタイプの転移性因子（「クラス 2」とする）は特定の塩基配列にはさまれた DNA 部分が転移酵素によって切り出され、ゲノムの他の部分に移動して組み込まれる。

（1） 1 つのクラス 1 転移性因子が生殖細胞の中で 1 世代につき 1 回転移する場合、10 世代後にはゲノム中のコピー数はどの程度に増えると考えられるか。

（2） クラス 2 因子が毎回正確に元の場所から切り出され、新しい場所に移動するとしても、染色体中のクラス 2 因子のコピー数が増大することがある。どのような場合が考えられるか説明せよ。

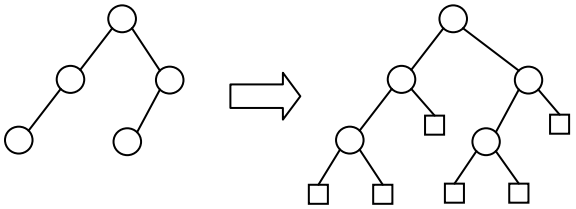
（3） 転移性因子が新しい位置に挿入された結果、あるタンパク質がその生物の細胞で産生されなくなった。考えられる理由を 2 つ述べよ。

（4） 転移性因子が新しい位置に挿入されても、表現型を示すような突然変異が観察されないことも多い。これはどのような理由によるか、考えられる理由を 1 つ述べよ。

（5） このような反復配列がゲノムの大きな領域を占めることによって、ゲノム配列の解析にどのような困難が生じるか、説明せよ。

Problem 1

Suppose an extended binary tree where nodes without children are augmented with special nodes as shown below. Let us call ○ as internal nodes, □ as external nodes, and the distance between a node and the tree root as depth. Let n be the number of internal nodes.



（１）Prove that the number of edges connecting internal nodes equals $n-1$.

（２）Prove that the number of external nodes equals $n+1$.

（３）Let us consider an operation to delete one internal node of depth k having two external nodes. Show the change in the sum of the depth of all external nodes, and the change in the sum of the depth of all internal nodes, respectively in terms of k.

（４）Prove that the sum of the depth of all external nodes equals the sum of the depth of all internal nodes + $2n$.

（５）Explain two tree structures of n internal nodes in which the sum of the depth of all internal nodes is the minimum, and the maximum, respectively.

Problem 2

A mapping T from a vector space V to itself is called linear if it satisfies

1. $T(\mathbf{x}+\mathbf{y}) = T(\mathbf{x})+T(\mathbf{y})$, and
2. $T(c\mathbf{x}) = cT(\mathbf{x})$.

（1） Let $\{ \mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n \}$ be a basis of V, and $\{ \mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_n \}$ be its projection by the linear mapping T. Using these notations, explain that $\mathbf{y}=T(\mathbf{x})$, i.e., the mapping of vector $\mathbf{x}$ in V by T, can be described by a matrix operation.

（2）Let T^{-1} be a matrix corresponding to the inverse mapping of T, and we call it an inverse of T. Let matrices A and B have their inverses. Prove that AB has its inverse and $(AB)^{-1} = B^{-1}A^{-1}$.

（3）Let A and B be the matrix representation of T using bases $\{ \mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n \}$ and $\{ \mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_n \}$, respectively, and let P be the transformation from $\{ \mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n \}$ to $\{ \mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_n \}$. Prove $B = P^{-1}AP$.

（4） When a square matrix of order *n* satisfies $A\mathbf{x} = \lambda\mathbf{x}$ for some scalar λ, we call λ an eigenvalue of A. Find all eigenvalues of matrix A that has an inverse and satisfies $A^2 = A$.

（5） Let eigenvalues of matrix A be $\{ \lambda_1, \lambda_2, ... \lambda_n \}$. Find all eigenvalues of matrix A^m and explain why.

Problem 3

The following algorithm is the counting sort of an array of length N containing positive integers from 0 to M−1. Answer the following questions.

Program

```
for (int i=0; i< M; i++) C[i] = 0;
for (int i=0; i< N; i++) C[value[i]] ++;
for (int i=1; i< M; i++) C[i] += C[i−1];
for (int i=N−1; i >= 0; i−−) sorted[−−C[value[i]]] = value[i];
```

(1) Explain the role of array C.
(2) Explain the purpose of the operation in the 3rd line of the program.
(3) Explain why the program generates the array 'sorted' in the decreasing order of i, instead of the increasing order.
(4) Show the computation time of running this program in terms of M and N.
(5) Explain why the computation time for sorting n elements requires at least O(n log n) time in general. You may use the Sterling's formula $n! \approx \sqrt{2\pi} e^{-n} n^{n+(1/2)}$, if necessary.

Problem 4

Suppose that we randomly permute numbers from 1 to n in an array of size n. Let C(n) be the number of combinations where the i-th position in the array does not contain i for all i = 1, ..., n. (We assume that the array index starts from 1.) Let us solve C(n).

（1）Suppose the number 1 is in the array position i ($i \neq 1$). Show the number of combinations where the number i is in the array position 1 using C.

（2）Suppose the number 1 is in the array position i (i≠1). Show the number of combinations where the number i is not in the array position 1 using C.

（3）From the results above, show the recurrence formula for C.

（4）Solve the recurrence formula of C.

（5）Find the probability where the i-th position in the array does not contain i for all i and $n \to \infty$. You may use $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$, if necessary.

Problem 5

Explain each of the following two terms and their relationship.　For each of (1)～(4), explain in about 100 words.

(1) “ortholog”　“synteny”
(2) “gene expression profile”　“microarray”
(3) “homology search”　“score matrix”
(4) “QTL”　“DNA marker”

Problem 6

Two-dimensional dot-plot is a method to visually identify the highly conserved regions and the structural similarities and differences between two DNA sequences. Figure 1 shows an example of the dot-plot analysis between human and mouse genomic sequences (about 5 kb in the length) containing their orthologous genes. The process of dot-plot analysis here includes scanning the sequence similarity for both strands of the entire region of one sequence by sliding another sequence every 10 bases with the 20 bases window size, and two-dimensionally displaying dots for the alignment-pairs that have more than 90% identity. As indicated by a, b and c in Figure 1, three conserved regions were identified as apparently seamless lines composed of many dots representing the alignment-pairs. Using these features of the dot-plot analysis, answer the following questions. Note that arrow lines in the figures indicate the 5' to 3' direction of DNA sequences. Also note that many less significant dots appear as backgrounds in the actual analysis but they were eliminated in this problem.

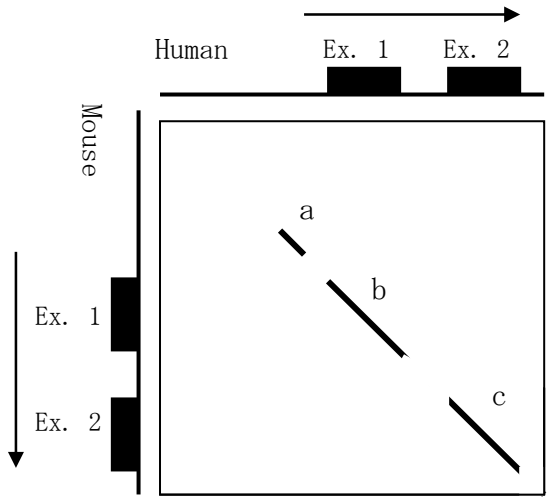


Fig. 1

1. In Figure 1, regions 'b' and 'c' represent the first (Ex. 1) and the second (Ex. 2) exons conserved between human and mouse, respectively. What is the name of the less conserved region between 'b' and 'c'? What is the conserved region 'a' located upstream of the first exon?

2. In Sequence 2 (about 500 kb in the length), two duplicate regions (about 100 kb for each) with more than 90% nucleotide identity are present in the same direction (shown by arrow lines). Draw freehand the result obtained from the self-dot-plot of Sequence 2 where only alignment-pairs with more than 90% identity are plotted. Note that you must draw Figure 2 in the answer sheet before you start plotting. It is not necessary to precisely copy the figure for the length and line positions.

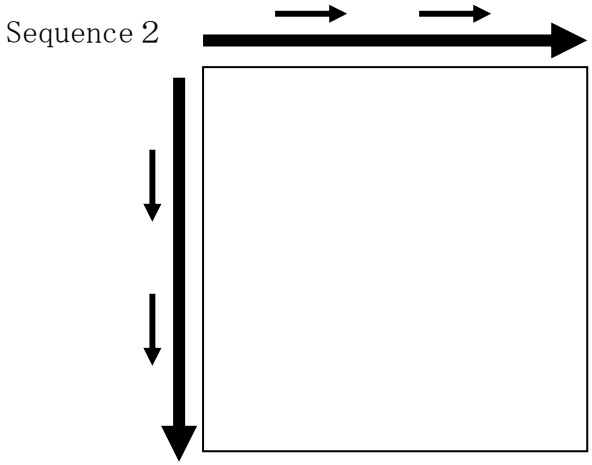


Fig. 2

3. The result of the self-dot-plot of Sequence 3 is shown in Figure 3. What type of sequence structure is present in Sequence 3? Answer its general name. Then copy Figure 3 in your answer sheet and draw arrows showing its position and direction for Sequence 3, as in Figure 2. It is not necessary to precisely copy the figure for the length and line positions.

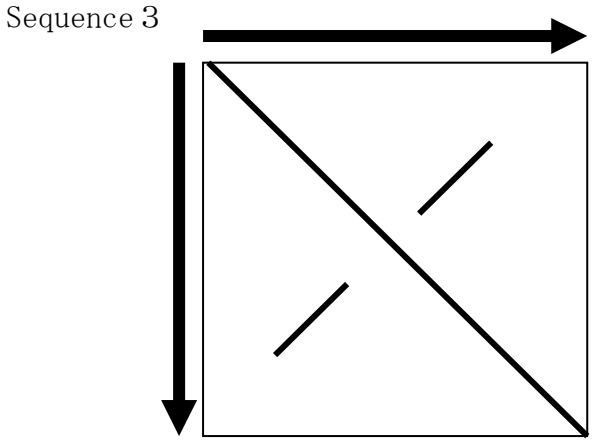


Fig. 3

4. When two bacterial genomes 4 and 5 of about 8 Mb were compared, the dot-plot in Figure 4 was obtained. From this result, draw schematically the structure of these bacterial genomes with clear indication of the differences between them, and explain them concisely. It is not necessary to precisely copy the figure for the length and line positions.

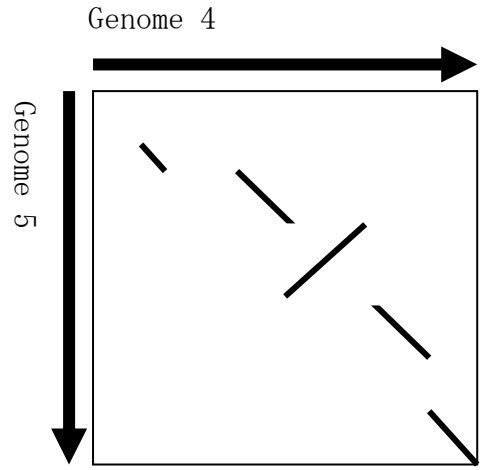


Fig. 4

Problem 7

Assume that among twelve substances { a, b, c, d, e, f, g, h, i, j, k, l } that are not distinguishable from their appearance, one of them is of different weight while the others have the same weight. For each of the following cases, design and explain a procedure that finds the substance of different weight and decides whether it is heavier or lighter than the others of the same weight by using a balance only twice.

(1) The case when one of a, b, and c is of different weight.
(2) The case when two groups of substances { a, b, c, d } and { e, f, g, h } are found to be equal in weight.
(3) The case when { a, b, c, d } is heavier than { e, f, g, h }.

Problem 8

(1) From the amino acid residues listed by three letter codes below, select a pair of amino acid residues which satisfy each condition from A to E without overlap.

Amino acid residues

GLU PHE PRO VAL GLY CYS TYR MET ARG ILE

Conditions

A. Include a sulfur atom.

B. Enable to stabilize protein structure by interaction between 6-membered rings.

C. Compared to other amino acids, this pair tends to take significantly different range of main-chain dihedral angles.

D. One of them is given by replacing a side-chain proton of another by a methyl group.

E. Enable to form salt bridge between the side chains.

(2) Let us consider a wild-type protein and its mutant that can bind to the same ligand. Under the isothermal-isobaric condition, four different Gibbs free energy changes are defined as follows.

ΔG_1: free energy gain when unliganded wild-type protein binds to the ligand.

ΔG_2: free energy gain when unliganded mutant protein binds to the ligand.

ΔG_3: free energy increase of unliganded mutant protein compared to unliganded wild-type protein.

ΔG_4: free energy increase of liganded mutant protein compared to liganded wild-type protein.

Answer the following questions.

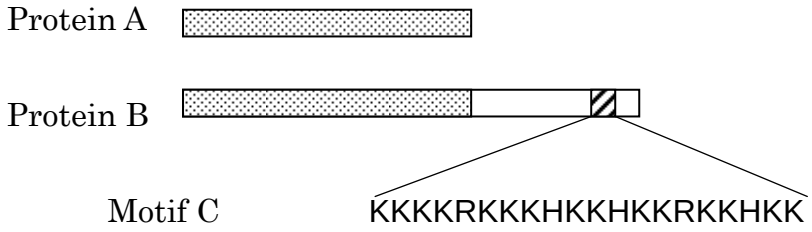
(2-1) For wild-type and mutant proteins to bind to the ligand stably, answer what the signs of ΔG_1 and ΔG_2 should be.

(2-2) Answer the relationship that ΔG_1, ΔG_2, ΔG_3 and ΔG_4 should satisfy.

(2-3) Define the quantity α that is the best suitable for judging whether wild-type or mutant protein binds to the ligand more stably, and explain the condition of the judgment.

Problem 9

Gene X is expressed in human cultured cells A and B. Northern blot analysis using the region coding for the exon 1 of gene X as a probe revealed that about 2.0 and 3.3 kb mRNA were expressed in A and B cells, respectively. By the molecular cloning and sequencing of each cDNA, it became clear that the sizes of protein coding region in A and B cells were 1,854 bp and 3,012 bp, respectively, and the first 1,356 bp was exactly identical between them. The predicted amino acid sequences from the cDNAs obtained from A cells (protein A) or B cells (protein B) revealed that only one distinct protein motif (motif C) exists in the carboxyl-terminal region of protein B.



(1) Gene X is known to be coded in a single gene locus. However, two different sizes of mRNAs were expressed in the cells A and B. What is this mechanism called?

(2) In the figure, the amino acid sequence of motif C is shown using the one letter code for amino acids. What is the common characteristic of amino acids composing the motif C?

(3) In order to investigate localization of protein A and B in the cells, two polyclonal antibodies a and b were prepared. Polyclonal antibody a was prepared using the common amino acid residues 1-35 of protein A and B as an antigen, and polyclonal antibody b was prepared using the amino acid residues 612-635 of protein B as an antigen. A and B cells were fixed and stained with each of these polyclonal antibodies. Explain the expected staining patterns of polyclonal antibodies a and b in the A and B cells, respectively.

(4) Using the polyclonal antibody a, western blot analysis was performed against the cell lysates from A and B cells. Explain the predicted sizes of each protein detected in A and B cell lysates. Let the average molecular weight of amino acids be 133 in the calculation.

(5) The monoclonal antibody c was prepared using protein A purified from A cells as an antigen. When western blot analysis was performed against the lysates from A and B cells, although a single band was detected against A cell lysate, no signal was detected for B cell lysate. Answer two candidates of the epitope recognized by the monoclonal antibody c.

Problem 10

Human cells contain 22 pairs of autosomal and a pair of sex chromosome. At mitosis, the chromosomes are highly condensed and can be distinguished by a particular staining method. The chromosomes are less condensed at interphase, and genes are transcribed at chromosomal regions called (A). In female cells at interphase, however, one of the two X chromosomes is highly condensed to be (B) where transcription is inactive.

Each chromosome is replicated during DNA synthesis phase and segregated into two daughter cells at mitosis.

(1) Fill the blank (A) and (B) with appropriate words.

(2) There are 3 billion base pairs of DNA in a human genome. Calculate the weight of DNA in a human cell, assuming that the average molecular weight of a DNA base pair is 635. Formula for calculation must be presented. The Avogadro constant is 6×10^{23}.

(3) The chromosomes are highly condensed and packaged into a nucleus with a diameter of several micrometers. Describe the structure of nucleosome that is the most fundamental unit of the condensed DNA.

(4) Inactivation of X chromosome generally occurs in mammalian female cells. When a pedigree of cats were observed, the patterns of body color were (White and Black), (White and Orange), and (White, Black and Orange) in female, whereas those were (White and Black) and (White and Orange) in male. Explain why the (White, Black and Orange) was observed only in female, in connection with the X chromosome inactivation.

(5) Regarding the underlined sentence, in order for a DNA molecule to be a functional chromosome, it must contain three kinds of DNA sequence elements. Describe the names of the three elements, and explain each of their function.

Problem 11

Fruitfly displays a unique behavior called Y when stimulated with a chemical X. To isolate mutants of this behavior, we crossed mutagen-treated males with untreated females. We isolated two mutants from the F1 population obtained by this crossing, namely mutant A that fails to display the Y behavior upon X stimulation (*i.e.* L phenotype) and mutant B that displays the Y behavior even in the absence of X (*i.e.* H phenotype). The genes responsible for mutant A and B are termed *A* and *B*, respectively. The mutant alleles for these genes are designated as *Am* and *Bm*, respectively, whereas the wild type alleles are designated as +. Note that both genes are autosomal and that no linkage was observed between the two genes. Answer the following questions.

(1) Although the *Am* seems to carry a loss-of-function mutation, it is a dominant allele. Why? Explain plausible mechanisms concisely.

(2) To learn which of the two genes *A* or *B* functions genetically more upstream, we crossed the heterozygous *A* mutant (*Am*/+) with the heterozygous *B* mutant (*Bm*/+). Explain the expected segregation ratio of the three phenotypes (*i.e.* wild, L, and H) in the population obtained by this crossing experiment in the following two cases, one in which *A* functions upstream of *B* and the other in which *A* functions downstream of *B*.

(3) The original F1 population likely includes heterozygotes for recessive alleles of genes involved in the behavior Y. To identify such individuals, an F1 individual named C that displays the wild-type phenotype was crossed with the wild-type fly to generate an F2 population. Subsequently, the F2 individuals were backcrossed with the F1 individual C to generate an F3 population. If the F1 individual C carries a recessive allele of a gene involved in the behavior Y, what fraction of the F3 population are expected to show a mutant phenotype? Explain the expected results.

(4) The crossing experiment in Question (3) revealed that the individual C is a heterozygote for a recessive allele *c* (*i.e. c*/+) and that the *c* allele is responsible for the L phenotype. A similar experiment revealed that an individual D is a heterozygote for a recessive allele *d* (*i.e. d*/+) and that the *d* allele is also responsible for the L phenotype. What kind of experiments should we do to learn whether or not the gene *C* is identical with the gene *D*? Explain the experiments and the expected results in the following two cases, one in which *C* is identical with *D* and the other in which *C* is not identical with *D*.

(5) What kind of crossing experiments should we do to learn which gene functions genetically more upstream, gene *B* or gene *C*? Explain the experiments with the expected segregation ratio

of the three phenotypes (*i.e.* wild, L, and H) in the following two cases, one in which *B* functions upstream of *C* and the other in which *B* functions downstream of *C*. Note that gene *C* is an autosomal gene showing no linkage with *B*.

Problem 12

A large portion of eukaryotic DNA – about 35% of the human genome and ≥ 50% of the maize (corn) genome – consists of repetitive sequences, most of which are presumed to be derived from transposable elements. One type of transposable element (called here as "class 1") functions like a retrovirus: certain portion of the DNA is first transcribed to RNA and then reverse-transcribed into the genomic DNA, resulting in a new copy in another locus of the genome. For another type (called here as "class 2"), a part of the DNA flanked by two specific sequences is excised by a transposase enzyme and inserted into another locus.

(1) Suppose that each class-1 transposable element can be transcribed only once in each generation. If there is one such element in the genome, how many copies will be observed after ten generations?
(2) The number of the class-2 transposable elements in the genome may increase even if it is always precisely excised and inserted to a new locus. Explain a conceivable reason.
(3) Suppose that a transposable element was inserted into a new locus, and the expression of a certain protein became no longer observed. Explain two conceivable reasons for this phenomenon.
(4) In many cases transposition of the element does not cause any observable mutations. Explain a conceivable reason for this.
(5) The fact that such repetitive sequences occupy a large portion of the genome causes difficulty in genome analysis. Explain what kind of difficulty may arise.

（このページは草稿用紙として使用してよい）
(Blank page for draft)

（このページは草稿用紙として使用してよい）
(Blank page for draft)