生物情報ソフトウエア論 **II**　（担当　森下）

試験日時　２０１４年９月４日　木曜日

１０：３０　～　１２：３０

途中退室して構いません。

問題１　スコア最大鎖問題と chaining 法

塩基置換率の高い（70〜90%）ゲノム配列の間に存在する古代から保存された微かな類似配列を検出するために、スコア最大鎖問題は定式化され、この問題を解くために chaining 法は設計された。一般に、微かに保存された類似配列は何らかの機能を担っていると考えられ、興味深い。以下の問に答えよ。

スコア最大鎖問題: 文字列 P の開始位置 sx から終了位置 ex-1 ($sx < ex$) まで連続した部分文字列を $P[sx, ex)$ と表記する(位置は自然数)。２つの文字列 P, Q で類似した部分文字列 $P[sx, ex), Q[sy, ey)$ の組を、類似度を示す実数スコア ($score$) とともに、

$$A = (P[sx, ex], Q[sy, ey], score)$$

と表現し、アラインメントと呼び、２次元平面上の開始点 (sx, sy) と終了点 (ex, ey) を結ぶ線で表現する。各要素 $sx, ex, sy, ey, score$ は $A.sx, A.ex, A.sy, A.ey, A.score$ と参照する。アラインメント A, B で $A.ex \leq B.sx$ かつ $A.ey \leq B.sy$ ならば $A < B$ と定義する。A_k で終了する昇順列 $A_1 < A_2 < \ldots < A_k$ ($k \geq 1$) を鎖 (chain)、鎖のスコアを $\Sigma_{i=1,\ldots,k} A_i.score$ ($A_k.chain\_score$ と表記)、鎖の長さを k と定義する。A_k で終了する鎖は複数ありうる。大きなスコアをもつ鎖は進化的に保存された領域を示唆する。スコア最大の鎖を選択する問題を、スコア最大鎖問題と呼ぶ。

(1) アラインメントの数を n とする。平衡木の検索、挿入、削除が $O(\log n)$ 時間で実行できることを仮定して、各アラインメントで終了するスコア最大鎖を $O(n \log n)$ で計算する chaining アルゴリズムを述べよ。

(2) なぜ最悪計算量 $O(n \log n)$ で実行できるか説明せよ。その際、以下の性質を合わせて証明せよ。

I. リスト Y のすべての組 ($C.ey$, C) は、終了点 y 座標 $C.ey$ と鎖のスコア $C.chain\_score$ 双方の値で同時に昇順にソートされている.

II. 処理済みアラインメント B で終了する鎖のスコア $B.chain\_score$ は、B で終了する鎖の中で最大.

(3) 動的計画法とは、部分問題の最適解から、より大きな問題の最適解を構成する一般的なアプローチを示す概念である。動的計画法は、アミノ酸や塩基配列の類似性を見つける際、しばしば使われるが、chaining も動的計画法の１つと考えられる理由を述べよ。一方、動的計画法に基づく Needleman-Wunsch, Smith-Waterman アルゴリズムと比較して、chaining アルゴリズムが有効に使える例題、およびあまり有効でない例題を、理由とともに述べよ。

問題２　人類遺伝学の基礎概念

パーソナルゲノムの 1 塩基変異を観測することは、低コストで実現可能になり(もうすぐ$1000/人)、疾患に関連する遺伝子上の変異を報告した論文が近年急速にふえている。変異を調べることで、ある疾患に罹患しているか否かを正確に判断できるようになり、診療の方針も明確になることが期待されている。このような意味で人類遺伝学は、新たな知識の集積が最も進んでいる生命科学分野の１つなのかもしれない。人類遺伝学に関する下記の基本的な知識について答えよ。

(1) 遺伝的距離とは何か、減数分裂、交叉、組換え、物理的距離のキーワードを使って述べよ。また遺伝的距離と組換え率の関係について数式化したホールデンのマップ関数を導け。

(2) アレル、遺伝子型、ホモ接合、ヘテロ接合について説明せよ。アレル A, B の頻度率を $p_{\mathrm{A}}, p_{\mathrm{B}}$、遺伝子型（unordered）$AA, AB, BB$ の頻度率をp_{AA}，p_{AB}，p_{BB} と記述するとき、以下のハーディー・ワインバーグ平衡がなぜ成立するか説明せよ。

$$p_{\mathrm{A}} = p_{\mathrm{AA}} + (1/2)p_{\mathrm{AB}}$$
$$p_{\mathrm{AA}} = {p_A}^2$$
$$p_{\mathrm{AB}} = 2p_{\mathrm{A}}p_{\mathrm{B}}$$

(3) ランダムな交配が繰り返されるとハーディー・ワインバーグ平衡が成立すると考えられる。成立しない場合、どのような理由が考えられるか？

(4) ハプロタイプと遺伝子型の違いを説明し、ハプロタイプ上に並んだアレルが連鎖不平衡にあるとはどのような状態を示しているか説明せよ。同一ハプロタイプ上のアレル A, B 間の連鎖不平衡の度合いを数値化した量として重宝されている D_{AB} を、数式を使って説明せよ。

(5) D_{AB} を正規化する必要性について説明し、正規化した D_{AB}' について述べ、D_{AB}' を使って定義される連鎖不平衡ブロックとは何か述べよ。

問題３　Genome Wide Association Study (GWAS)と個人ゲノム解析

全ゲノム相関解析（Genome-wide Association Study, 略して GWAS）は、ヒトゲノム全体からおおよそ 50 万個の位置を選択し、各位置の遺伝子型と対象疾患の相関関係を解析する手法である。中村祐輔博士（当時東大医科研、現在シカゴ大学）のグループが提案した方法論で、個人ゲノム解析の先駆けであり、ヒトゲノムの解読が進んでいた 2002 年に最初の論文が発表されている。2002 年ごろの技術では 50 万個の位置を観測することでさえ高コストであった（約 1 ドル／位置)。その後コストも下がり、2007 年には米国 23andMe 社が一般向け解析サービスを開始している。さらに 2009 年ごろからは次世代シーケンサーが普及しはじめ、各 1 塩基の遺伝子型を調べる時代が到来した。

(1) DNA 上のある位置におけるアレルを A, a とする。アレル A が優性遺伝形式の疾患の原因となるか否かを判定する際に使われるカイ二乗検定とは何か説明せよ。またこの優性遺伝形式が疾患に及ぼす効果の強さを測るために使われるオッズ比とは何か述べよ。

(2) 上の問題では予め 1 つの位置に焦点を絞ったが、ゲノム全体で疾患遺伝子の候補を探るために、DNA 上の n 個（たとえば $n = 500{,}000$）の位置で遺伝子型を観測しよう。どの遺伝子型が関心を持っている疾患と相関するかを調べるために、帰無仮説

「n 個のすべての位置の遺伝子型が疾患に相関しない」

を有意水準 α で検定し、この帰無仮説を棄却する遺伝子型が存在するのかどうかを探すことは自然である。もしこのとき

「各位置において遺伝子型が疾患に関連しない」

という帰無仮説を有意水準 α で検定してしまうと、どのような問題が起こるか述べよ。これを多重検定問題と呼ぶ。多重検定問題を回避するために Bonferroni 補正がしばしば使われる。この補正が、どのような工夫をするかを説明し、その妥当性を述べよ。

(3) DNA 上の n 個の位置を選択する際に、任意の 2 つの位置 A, B の連鎖不平衡の度合い D_{AB}' が 1 のときには、同時に選択すべきではない。その理由を説明せよ。

(4) GWAS は疾患関連遺伝子の探索に多大な影響を与えた研究である。たとえば 2007 年には Science 誌が Breakthrough of the Year に選んでいる（同年に報告された iPS もトップ 10 に選ばれている)。しかし成果は、必ずしも臨床検査として普及しているわけではないという批判もある。その理由を以下のキーワードを使って説明せよ。

オッズ比、エキソン、イントロン、遺伝子コード領域間、アレル頻度

(5) 次世代シーケンサーを使って遺伝子型を調べる手続きとその限界を、以下のキーワードを使って説明せよ。

エラー補正、標準ゲノム、ホモ接合、劣性遺伝、DNA 断片、アラインメント、ヘテロ接合、優性遺伝、構造多型、直列型繰り返し配列

問題４　連鎖解析

1987 年、メンデル性遺伝病の責任遺伝子領域を探索する連鎖解析アルゴリズムを、Eric Lander 博士と Phil Green 博士らは提唱した。実は、Lander 博士は 1980 年にオックスフォード大学、Green 博士は 1972 年にカルフォルニア大学バークレー校で数学の博士号を取得している。その後生物学へとアプローチし、この連鎖解析アルゴリズムを考案し、生命科学で最も利用されているアルゴリズムの１つとなった。

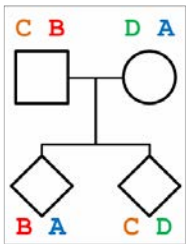
(1) メンデル性遺伝の状態を表現するために彼らが使った概念 IBD (Identical By Descent) とは何か、アレル、ordered/unordered genotype (遺伝子型)、3 世代等のキーワードを使って述べよ。また罹患同胞対解析 (affected sib-pair) とは何かを、IBD、常染色体劣性／優性遺伝形式等のキーワードを使って述べよ。

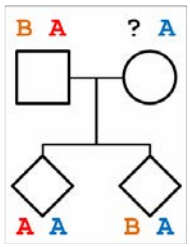
(2) unordered genotype, IBD および関連する確率を以下のように記述する。X_i $(i = 1,2,\dots,M)$ を観測する確率 $P(X_1 \cdots X_{j-1}|I_j)P(X_j|I_j)P(X_{j+1} \cdots X_M|I_j)$ を最大化する I_j を推定したい。Hidden Markov Model により推定する連鎖解析アルゴリズムを説明せよ。

i	DNA 上の位置
X_i	位置 i の Sib1/2 の unordered genotype
I_i	位置 i の IBD (I_i = 0, 1, 2) 観測困難なため予測
$P(I_i)$	位置 i の IBD が I_i となる確率
$P(X_i \,\&\, I_i)$	位置 i で X_i と I_i が起こる確率
$P(X_i \mid I_i)$	IBD が I_i のとき X_iを観測する条件付確率 $P(X_i \mid I_i) = P(X_i \,\&\, I_i)/P(I_i)$
$P(I_i \mid I_{i-1})$	I_{i-1} から I_i へ遷移する確率

(3) 位置 i と位置 i+1 の間の組換え率を θ とする。遷移確率 $P(I_i = 2 \mid I_{i-1} = 1)$ および $P(I_i = 0 \mid I_{i-1} = 2)$を求めよ。

(4) 連鎖解析アルゴリズムを構築する際に使われる IBD vector とは何か、右の２つの家系を例に述べよ。この２つの家系について、IBD vector を $\boldsymbol{I_i}$ 、unordered genotype X_i を **ABCD** および **AAAB** としたとき、確率 $P(X_i \,\&\, \boldsymbol{I_i})$ を各々求めよ。





問題５　区間推定

次世代シーケンサーの普及により、CpG サイトのメチル化状態、DNA が巻きついているヒストン８量体の修飾の状態を詳細に観測できるようになってきている。ゲノム全域に渡る調査を実施したのは ENCODE 計画であり、2010 年にショウジョウバエ、線虫、2012 年にヒトのデータを公開している。多様な修飾状態のどのような組合せが、プロモータ、エンハンサー、エキソン、イントロンのどの部位で顕著に出現するかが詳しく分析された。2013 年には、重要な発生関連遺伝子をコードするゲノム領域の多くは、低メチル化かつヒストン H3 の 27 番目のリジンがメチル化されることで初期胚における発現が抑えられており、細胞運命決定が進む過程で高メチル化へと変化し、発生関連遺伝子が転写されるようになるという現象が報告されている。

(1) バイサルファイト法により、全ゲノム中の CpG のシトシンメチル化状態を判定する方法を、データ分析の効率化を焦点にあてて説明せよ。また DNA がヒストン８量体に巻き付いている位置（ヌクレオソームコア）を Micrococcal Nuclease を使って網羅的に推定する方法を述べよ。

(2) ゲノムを区間に分割し、各区間の機能を推定することの意義を、以下の言葉を使って説明せよ。

DNA メチル化，ヒストン修飾，ヒストン H3，核スペックル，ポリコーム複合体

(3) DNA のメチル化状態やヒストン修飾状態は、同一の状態が 1 次元の区間としてブロック化して存在することが多く、このような区間を推定することが重要である。そこで実数値の列 $\{L_i | i = 1,2,\dots,n\}$ から、互いに交わらない m 個の区間の集合 $S =$ {区間 $I_j \mid j = 1,\dots,m\}$ で、重み $\sum_{i \in I_j \in S} L_i$ が最大となる S を求める問題について考える。m=1 の場合、$O(n)$ で計算するアルゴリズムを示し、なぜ $O(n)$ で計算できるか述べよ。

(4) m=2 の場合、重みが最大の 1 区間 $[c_1, c_2]$ に対して以下のどちらかの性質が成り立つことを証明せよ。

A) $[c_1, d_1]$ と $[d_2, c_2]$ が最適２区間となる $d_1 < d_2$ が存在。

B) $[c_1, c_2]$ との交わらない区間 K が存在して $[c_1, c_2]$ と K が最適２区間。

この性質を利用して、$O(n)$ 時間で最適２区間を計算するアルゴリズムを述べよ。