2H-2

塩基の配置情報を考慮に入れたDNA配列のグラフ表示とその応用

水田 智史[1]、山口 恭平[2]
[1]弘前大学大学院理工学研究科、[2]弘前大学理工学部電子情報工学科

1 イントロダクション

近年、データベースに登録されている生物学的配列の量は急速に増加しつつある。一般に配列比較にはアライメントが用いられるが、比較する配列の長さを L とするとアライメントの時間計算量は $O(L^2)$ であるため、それを軽減するための様々な改良が加えられる一方、アライメントによらない配列比較の手法も盛んに研究されるようになった。その代表的なものとして、配列中の一定の長さの語の出現頻度分布を比較することにより配列間の距離を評価する方法や、DNA 配列の場合は 4 種、アミノ酸配列の場合は 20 種の文字を何らかの方法で数値に置き換え、配列を 2 次元、もしくは 3 次元以上のグラフに表してその形状を比較することにより配列間の類似度を評価する方法などが挙げられる。本研究では、後者の範疇に属する新たな手法を提案する。

2 方法

2.1 DNA 配列の数値化

本研究では哺乳類 31 種の**ミトコンドリアゲノム**、および哺乳類 10 種と鳥類 1 種の **β-グロビン遺伝子**を対象として解析を行った。配列データはすべて GenBank[1] よりダウンロードして用いた。以下それぞれについて数値化の方法を述べる。

ミトコンドリアゲノム

4 種類の塩基 A、T、G、C のそれぞれにベクトル $\boldsymbol{V}(\mathrm{A}) = (1,1)$, $\boldsymbol{V}(\mathrm{T}) = (-1,1)$, $\boldsymbol{V}(\mathrm{G}) = (-1,-1)$, $\boldsymbol{V}(\mathrm{C}) = (1,-1)$ を割り当て、ゲノム配列の先頭から順番に各塩基に対応するベクトルを加えていきながら 2 次元のグラフにその軌跡を表示する。すなわち、ゲノム配列の i 番目の塩基に対応する点の座標 $\boldsymbol{R}_i = (x_i, y_i)$ を

$$\boldsymbol{R}_i = \sum_{k=1}^{i} w(s_k)\boldsymbol{V}(s_k) \tag{1}$$

によって求め、各点を線で結んでいく。ここで s_k は配列上の k 番目の塩基であり、$w(s_k)$ は後述する重みである。

β-グロビン遺伝子

β-グロビン遺伝子は 4 つのイントロン領域と 3 つのエクソン領域からなるが、本研究ではエクソン領域のみを対象とした。合計で約 450 塩基対と短いため、軌跡が横軸方向に対しては正の向きに単純に延びていくように $\boldsymbol{V}(\mathrm{A}) = (1,-1)$, $\boldsymbol{V}(\mathrm{T}) = (1,1)$, $\boldsymbol{V}(\mathrm{G}) = (1,-2)$, $\boldsymbol{V}(\mathrm{C}) = (1,2)$ を割り当て、式 (1) によって各塩基に対する座標を求める。

2.2 重みの導入

これまでの研究では、状況に応じて各塩基に割り当てるベクトルは様々なものが用いられているが、単純にそれらを加えることによってグラフ上の点の座標 $\boldsymbol{R}_i$ を求めている。しかし、DNA 配列おいては、個々の塩基の種類に加えて、それらが出現する順番にも重要な情報が含まれているはずである。そこで、その並びがもつ情報量を重みとして各ベクトルに掛け合わせることで、その情報を取り込むことにする。そのためにまず、塩基 s とその前に並ぶ合計 l 個の塩基の列に対し、次の条件付きの出現確率

$$P_l(s) = P(s|s_{(l-1)}s_{(l-2)}\dots s_2 s_1) \tag{2}$$

を求める。ここで s_j は s の j だけ前に位置する塩基である。例えば、$l = 1$ の場合は $P_l(s) = P(s)$ となり単純な塩基の組成比であり、$l = 3$ で考えると、$P(\mathrm{A}|\mathrm{TC})$

は塩基の並び “TC” の後に A が続く条件付き確率を表す。この条件付き確率から塩基 s に対する重み $w(s)$ を

$$w(s) = -\log P_l(s) \tag{3}$$

として定義する。

本研究では、ミトコンドリアゲノム、β-グロビン遺伝子のそれぞれについて、対象とするすべての配列を用いて $w(s)$ を求めた。

2.3 配列間距離の評価

式 (1) から求まる $\boldsymbol{R}_{L/4}$、$\boldsymbol{R}_{L/2}$、$\boldsymbol{R}_{3L/4}$、$\boldsymbol{R}_L$(L は配列長) の成分から構成される 8 次元のベクトル $\boldsymbol{F}$ を配列を表す特徴ベクトルとし、それを位置ベクトルとする点と点の間のユークリッド距離を配列間距離と定義する。すなわち、2 本の配列 A、B の特徴ベクトルをそれぞれ $\boldsymbol{F}^A = (r_1^A, r_2^A, \ldots, r_8^A)$、$\boldsymbol{F}^B = (r_1^B, r_2^B, \ldots, r_8^B)$ とすると、配列 A、B 間の距離 $D(A,B)$ は

$$D(A,B) = \sqrt{\sum_{i=1}^{8} (r_i^A - r_i^B)^2} \tag{4}$$

によって計算される。

3 結果

紙面の都合上、ミトコンドリアゲノムに対する一部の結果のみ示す。図 1 はヒト及びイヌのミトコンドリアゲノムをグラフ表示したものである。“No weight” は重みを付けない場合の結果で、“1mer”,..., “7mer” はそれぞれ $l = 1, \ldots, 7$ に対する結果である。グラフの形は重みを付けない場合と付けた場合では大きく異なるが、重みを付けた場合においては l への依存性はほとんど見られない。

図 2 は、$l = 1$ において式 (4) より求めた配列間距離に基づき、哺乳類 31 種に対して UPGMA(Unweighted Pair Group Method with Arithmetic mean) 法によって作成した系統樹である。重みを付けない場合は “Sheep” と “Cow”、“Tiger” と “Cat” が離れたグループに位置するなどの不具合が存在したが (図は割愛)、重みを付けることによってそれらが近いグループに位置するよう改善された。

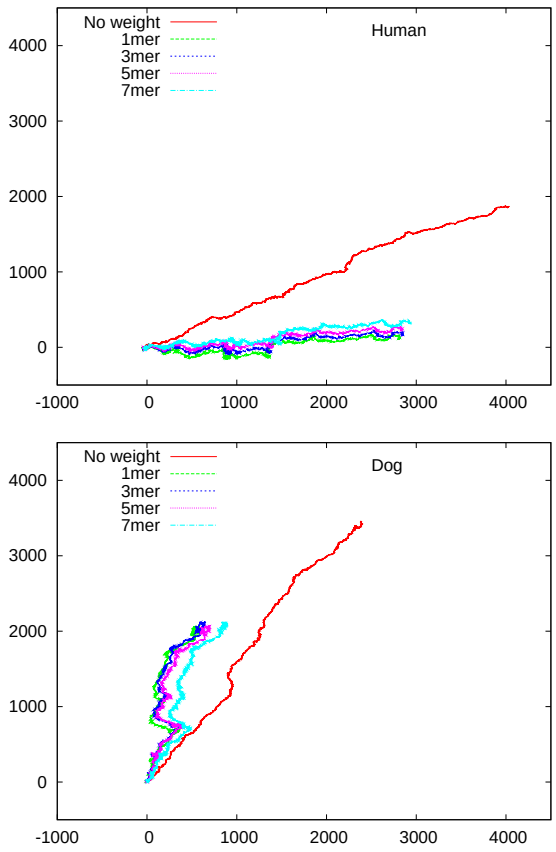

図 1: ミトコンドリアゲノム配列のグラフ表示

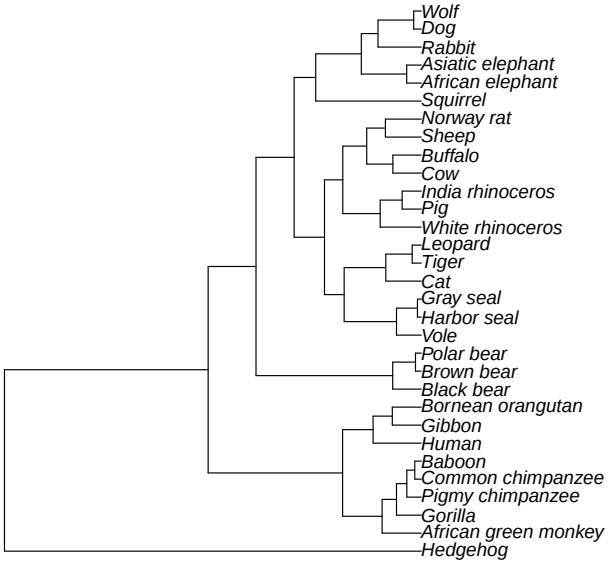

図 2: 哺乳類 31 種の系統樹 (UPGMA 法)

参考文献

[1] GenBank: http://www.ncbi.nlm.nih.gov/