2ZD-9

画像データ処理を用いたDNA配列間距離の評価

小堀裕聖† 水田智史†

†弘前大学理工学研究科

1 イントロダクション

配列間の類似性を調べる方法として一般的にアライメントが用いられている。アライメントの特徴として、比較する配列の長さがNとNの場合は計算量が$O(N^2)$となる。よって、ゲノム配列のような長い配列では膨大な計算量が求められるため、何らかの方法で計算量を減らす必要があり、アライメントを用いない様々な配列比較の手法が提案されている [1] [2] [3]。

本稿では、アライメントに依らない方法として、ゲノム配列の塩基（A,T,G,C）にベクトルを割り当てることによって数値化して、作成した二値画像の局所パターンから、配列間の類似性を定量的に評価する方法について検討する。

2 方法

2.1 使用するゲノム

Genbankから31種類の哺乳類のミトコンドリアゲノムをダウンロードして用いる。ミトコンドリアは、ほとんどの真核生物種が持っている、突然変異の頻度が大きい、ゲノムサイズが均一という性質を有しており、系統を探る目的に適していると考え使用する。

2.2 画像の作成

2.2.1 ベクトルの割り当て

配列の先頭を原点として塩基を取り出し、図1の2次元ベクトルを割り当てることによってゲノム配列を数値化することができる。

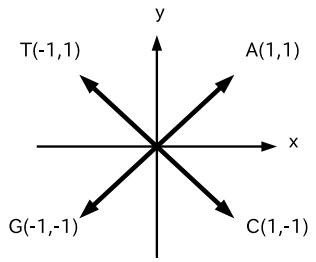

図 1: 割り当てるベクトル

2.2.2 重み付け

塩基にベクトルを割り当てるとき、例えば塩基Cを取り出す場合にAGCとTTCでは同じCでも持っている情報が等しくないのではないかと考え、配列の情報をより反映させるために重み付けをする。

重みには自己情報量$I(E)$を用いる。$I(E)$は事象Eが起こったときに受け取る情報の量で、事象Eが起こる確率を$P(E)$とすると、$I(E) = -\log P(E)$として定義される。また、確率$P(E)$には文字の並びから求めた条件付き確率を用いる。

2.2.3 条件付き確率

条件付き確率とは、ある事象が起こるという条件の下で別の事象が起こる確率である。例えば、AGが並ぶという条件の下でその後にCが続く場合の条件付き確率を$P(C|AG)$と表す。

本研究では3文字の塩基の並びを使用するため、64通り（AAA～TTT）の条件付き確率を31種類すべてのゲノムを用いて求める。

2.2.4 ゲノム配列の画像化

ゲノム配列間の距離を評価するため数値化した配列から図2のようなグラフ画像を作成する。画像形式には画像にする際の範囲と解像度を統一した二値画像を用いる。

図 2: アフリカゾウ (ミトコンドリア) の二値画像の一部

2.3 配列間距離の定義

2.3.1 特徴量の抽出

対象の画像全体から3 × 3の局所パターン（図3）の出現頻度をカウントして特徴ヒストグラム（図4）を作成する。3 × 3の二値画像なので512パターンがあるが、すべて白（ピン番号0）を除いた511パターンの出現頻度を特徴量として用いる。

Evaluation of the distance between DNA sequences using image data processing
Yusei Kobori†, Satoshi Mizuta†
†Graduate School of Science and Technology, Hirosaki University

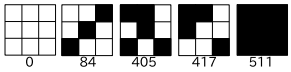

図 3: 局所領域の例と対応するピン番号

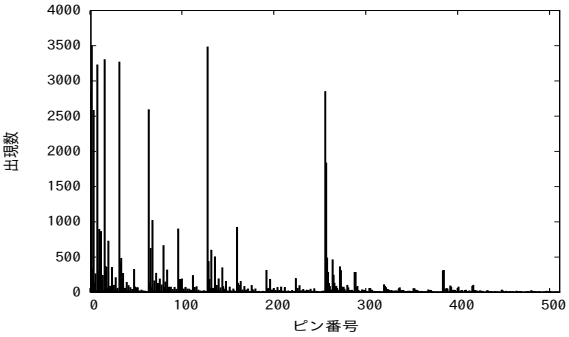

図 4: 特徴ヒストグラム（アフリカゾウ）

2.3.2 特徴ヒストグラム間距離

31 種類の各生物種から作成した特徴ヒストグラム間の類似度を求めるために、ヒストグラム交差法（HID）、マンハッタン距離（L_1D）、バタチャリヤ距離（BD）、ジェンセン・シャノン情報量（JSD）、ケンドールの順位相関係数（τ）の 5 種類の距離測度を用いて距離行列を作成する。

3 結果

3.1 系統樹の作成

図 5 は HID，L_1D によって求めた距離行列から非加重平均結合法（UPGMA）を用いて作成したの系統樹である。図 5 より近縁種同士が近く分類されていることが分かる。しかし、HID と L_1D，BD と JSD，τ の 3 種類の系統樹において齧歯目（ウサギ，リス，ハタネズミ，ラット）が分散した。

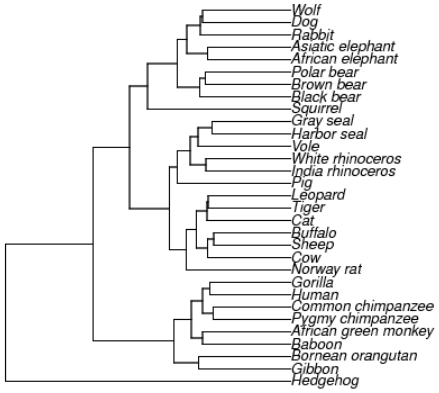

図 5: HID，L_1D に基づく UPGMA 系統樹

3.2 系統樹の比較

系統樹を定量的に評価するために系統樹間の Robinson-Foulds(RF) 距離を求めた。なお、今回は ClustalW によって作成した系統樹を基準とした。HID と L_1D、BD と JSD、τ で求めた RF 距離はそれぞれ 30, 40, 46 となり、HID 及び L_1D の系統樹が ClustalW で得られた系統樹に最も近いという結果が得られた。

3.3 距離測度間の相関関係

5 種類の距離測度間の相関関係（図 6）を見るために、それぞれの距離測度で計算した距離の間の相関係数を求めた。その結果を表 1 に示す。

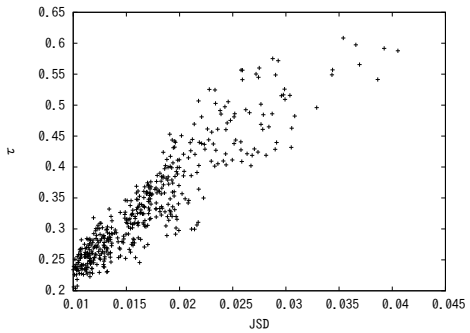

図 6: JSD と τ で求めた距離の散布図

表 1: 距離測度間の相関係数

	JSD	τ	L_1D	BD
τ	0.91797			
L_1D	0.98409	0.89977		
BD	0.99996	0.91717	0.98277	
HID	0.98987	0.90176	0.99898	0.98897

4 まとめと今後の課題

本研究では配列を数値化して画像にし、局所的なパターンの出現頻度分布によって配列間の距離を求めた。距離測度間の相関係数からヒストグラム交差法とマンハッタン距離、バタチャリヤ距離とジェンセン・シャノン情報量はほぼ同等な測度であることがわかった。今回は、HID と L_1D が最も良い結果となった。

問題点として、画像の解像度によって距離が変わってくるということが考えられるので、今後は、適切な解像度について検討する。

参考文献

[1] Vinga, H., and J. Almeida, “Alignment-free sequence comparison - a review,” *Bioinformatics* **19**, 513-523 (2003).

[2] Mantaci, S., et al., “Distance measures for biological sequences: Some recent approaches,” *Int. J. Approx. Reason.* **47**, 109-124 (2008).

[3] Huang, G., et al., “Alignment-free comparison of genome sequences by a new numerical characterization,” *J. Theor. Biol.* **281**, 107-112 (2011).