# 情報エントロピーを用いた生物進化の解析法

吉原郁夫[1)] 矢野佑貴[2)] 山森一人[3)] 相川勝[4)]

# Analysis of Organism's Evolutionary Relationships using Information Entropy

Ikuo YOSHIHARA Yuki YANO Kunihito YAMAMORI Masaru AIKAWA

Abstract

Organisms on the earth keep genetic informations to constitute a figure and the property of own in a macromolecule called DNA (deoxyribonucleicacid). For a long time, the base sequence of DNA had changed little by little, and the change laid the diversification of today's organisms. In this paper, We compared the similarity of the organisms by using information entropy to analyze it in a short time.

Key Words:
Information entropy, genome,

## 1 はじめに

地球上の生物は、自身の姿形や生物固有の性質を構成する遺伝情報を、DNA(デオキシリボ核酸)中に保持している。この DNA 中の塩基配列は、長い間受け継がれていくうちに突然変異などによる配列の置換・挿入・欠失により変化していき、この変化によって生物は進化していき、今日の生物の多様化が生じたと考えられる[1)]。従って、塩基配列の機能的、進化的性質を解析することは、生物進化の歴史を解明する上で非常に重要である。現在、生物の DNA の塩基配列のて解析を行うことにより、生物が進化してきた道筋や系統を理解しようとする研究分野は、分子系統学と呼ばれている[1)]。

こういった塩基配列を用いた生物の類似性解析の方法としては、配列の並びに着目し、複数の配列に対して CLUSTALW などのアラインメントツールを用いてアラインメントを行い、配列同士に関連性があるかを判断する方法がある[2)]。この塩基配列にパターンマッチングを適用する方法は広く利用されているが、解析の対象となる塩基配列が長いと計算量が膨大になり、非常に時間がかかるのが難点である[2)]。そこで本研究では、従来方法より時間をかけずに塩基配列の比較を可能とする方法として、各生物の持っている塩基配列データの一致を求めるのでなく、統計力学的な指標、たとえば DNA の塩基配列の不規則性を指標として生物種の類似性の比較できないか試みる。実験では、各生物の塩基配列を使用して各塩基配列中の各塩基の出現する確率を求め、それを基に情報エントロピーを計算し、各生物の特徴量とする。そしてその情報エントロピーからエントロピー進化率等を用いて遺伝行列を作成し、近隣結合法によって進化系統樹を作成する。そして、情報エントロピーによる生物種の類似性の比較とともに進化系統樹の類似性を数学的な値で示すため、Split Distance(スプリット距離) の比較法を用いることで、進化系統樹を比較する。

## 2 生物種の類似性判定方法

### 2.1 塩基配列

塩基配列とは、生物の細胞中に存在している DNA 中の塩基の並びのことを言う。生物の遺伝情報は親から子へと受け継がれていくが、その遺

[1)]情報システム工学科教授
[2)]情報システム工学科学生
[3)]情報システム工学科准教授
[4)]工学部教育研究支援技術センター技術職員

伝情報は細胞内のDNAに格納されている。DNAはリン酸・糖・塩基から構成された高分子であり、アミン (A) , シトシン (C), グアニン (G), チミン (T) の4つの塩基が糖とリン酸を介して鎖状につながっており、DNAの遺伝情報はこれらの4塩基の並びによって表現されている。塩基配列中の連続する3塩基をコドンと呼び、コドン中の塩基の並びによりアミノ酸が指定され、アミノ酸が順に結合してタンパク質を作る。アミノ酸を指定するコドンは塩基の並びにより全64種類存在し、64種類のコドンには同じアミノ酸を指定するコドンが存在しており、20種類のアミノ酸と終止コドンに翻訳される。

## 2.2 アラインメント

生物のDNA塩基配列の類似性を比較する際に、図1のようにできるだけ多くの塩基が一致するように配列間に空白を挿入して配列同士の対応づけを行う。これをアラインメントと呼び、配列間に挿入する空白をギャップと呼び、(-) で表す[2)]。

塩基配列にアラインメントを行うことによって、生物が進化してきた中で発生した塩基配列中の塩基の挿入、欠失、置換などの変化を評価することができ、遺伝子の機能、構造を知る手助けとなる。

図1のように2つの配列に対しアラインメントを行うことをペアワイズアラインメント、または単にアラインメントと呼ぶ。また、3本以上の配列に対しアラインメントを行うことをマルチプルアラインメントと呼ぶ。


図1 塩基配列のペアワイズアラインメント例

## 2.3 情報エントロピー

物理学でいうエントロピーとは、事象の複雑さ、曖昧さ、無秩序の程度を表す状態量である[3)]。この物理学のエントロピーの概念を情報理論に導入したものを、情報エントロピー、または平均情報量という。情報エントロピーとは、ある情報を得た事象の不確かさに対する尺度のひとつである[3)]。ある情報源 $X$ の事象が $x_1, ...., x_k$、それらの確率が $P(x_i)$ で与えられた時、その情報エントロピーは以下の式で定義される。

$$H(X) = -\sum_{i}^{k} P(x_i) log_2 P(x_i) \qquad (1)$$

情報エントロピーの値は、すべての $P(X_i)$ が等しい時に最大となり、その値は $H(X) = log_2 k$ である。

## 2.4 塩基配列からの標本の切り出し

本研究では配列の位置による情報エントロピーの変動を見るため、塩基長Lの塩基配列からN個の塩基を標本として切り出し、切り出した各標本から1塩基または3塩基の出現確率を計算、それを基に情報エントロピーを計算する。標本ごとの情報エントロピーの値を、標本の特徴量とする。

塩基配列から標本を切り出す方法は、まず対象の塩基配列の先頭から、あらかじめ決めていた切り出す標本サイズN個の配列を切り出す。これを図.2のように切り出す対象の塩基配列から1塩基ずつずらして標本の切り出しをL-N+1回行う。

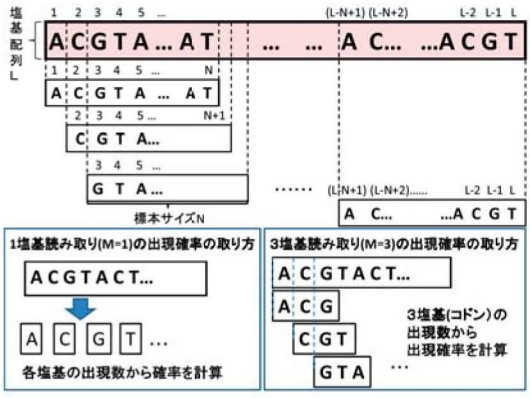


図2 塩基配列からの標本の切り出し、出現確率の計算例

## 2.5 エントロピー進化率

エントロピー進化率とは、2つの事象系X,Yの持つ情報量の相違を数学的に表わす尺度の一つで

あり、2つの確率事象系において導かれる[3)4)]。

例として、アラインメントされた二種類の塩基配列X,Yから標本として3塩基分を取り出し、3塩基に対応するアミノ酸に変換してアミノ酸配列を作成した場合のエントロピー進化率を計算する。このときアミノ酸配列では、アミノ酸20種とギャップの計21種が発生しうる。配列 $X, Y$ で各事象が発生する確率を $p_i, q_j (0 \leq i, j \leq 20)$ とし、$X$ と $Y$ の各事象が同時に起きるときの確率を $r_{i \cdot j} (0 \leq i, j \leq 20)$ とする。これらの事象系を基に、アミノ酸配列 $X$、$Y$ の相互情報量を求める。

$$I(X,Y) = \sum_{i,j} r(i,j) log_2 \frac{r(i,j)}{p(i)p(j)} \tag{2}$$

相互情報量 $I(X,Y)$ は、$X$ と $Y$ の間での相互依存の尺度を表す量である。この相互情報量と情報エントロピーを用いて $X$ から $Y$ のエントロピー比 $r(Y|X)$ を以下のように定義する。

$$r(Y|X) = \frac{I(X,Y)}{H(X)} \tag{3}$$

これは $X$ が持つ情報エントロピーと、$X$ から $Y$ へ伝達された情報エントロピーの比であり、$X$ に対する $Y$ の類似度を表す尺度である。これと同様のことを行って $Y$ から $X$ のエントロピー比 $r(X|Y)$ を計算し、二つのエントロピー比の平均の対称エントロピー比 $r(X,Y)$ を求める。この対称エントロピー比を基に $X$ と $Y$ の情報量の差異を表す量が、次式で求められる。

$$\rho(X,Y) = 1 - r(X,Y) \tag{4}$$

これがエントロピー進化率である。エントロピー進化率は $0 \leq \rho(X,Y) \leq 1$ をとり、塩基配列間の差異が大きくなるとこの値も大きくなる。エントロピー進化率を使った塩基配列の解析方法は、特定の箇所のアミノ酸あるいは塩基が何度も変化している可能性があるとき、特に役立つものである。

### 2.6 進化系統樹

生物の進化の解析において、生物種間の進化的関係を表現する方法としてよく用いられるのが、進化系統樹と呼ばれるものである。進化系統樹は木構造で表され、根を持つ有根系統樹と、根を持たない無根系統樹の二つに分けられる。

進化系統樹は塩基配列やアミノ酸配列を基に作成される。DNAデータによる進化系統樹の作成方法は「最節約法」や「距離行列法」などがあるが、本研究は距離行列法を用いる。

最節約法とは考えられるすべての進化系統樹を求め、その中から塩基配列やアミノ酸配列の変化が最小の進化系統樹を最適な系統樹として選択する方法である。

距離行列法とは、計算に用いるデータ中の生物種の各ペアについて、それらの値の差異を計算して距離行列を作成し、それを使って進化系統樹を構成する方法である。

### 2.7 近隣結合法

近隣結合法はNJ(Neighbor Joining)法とも呼ばれ、進化系統樹作成に用いる各要素の差異が、進化系統樹の枝の長さに反映するように進化系統樹を作るアルゴリズムである。これはすべての樹形を調べる検索方法と比べて計算量が少ないため効率が非常に良いことで知られ、最節約法などでは不可能なほどの大量のデータの計算を行うことができる。

まず近隣結合法に使用する生物種の塩基配列データを基にして、各生物ペアの配列データの値の差異を計算する。このときの各生物の塩基配列の差異を数値的に表わす尺度を遺伝的差異、あるいは遺伝距離と呼ぶ。そして、それらの遺伝距離の計算結果を基に距離行列を計算する。この距離行列を遺伝行列と呼ぶ。そして、遺伝行列中の最小の値を持つ要素を最も近隣の関係にある二つの生物の組として求め、その二つの生物間に節点を作成して二つの生物からの距離を計算し、その節点とほかの生物の遺伝行列を再度計算する。これをすべてが近隣関係になるまで繰り返し、最後に樹形を整えて進化系統樹の作成を終了する。

### 2.8 Split Distance

系統樹の構造を基に二つの系統樹の類似性を比較するために、本研究ではSplit Distanceに基づく比較法を用いる。Split Distanceにおける Splitとは、分子系統樹からある一つの枝を取り除いて、分割した結果得られる二つの部分木が持つ葉節点集合の組の数である[5)]。比較したい二つの系統樹から全てのSplitを求め、二つの木から求められたすべてのSplitのうち、両方に共通しないSplitの数を合計する。この値が、二つの系統樹の差異

の値 $SD$ であり、$SD$ が小さいほど、二つの系統樹は類似性が高く、$SD = 0$ の時に二つの系統樹は一致する。

## 3 実験内容

### 3.1 実験データ

今回の実験には、宮崎大学化学実験総合センターの RPG(Ribosomal Protein Gene Database:http://ribosome.miyazaki-med.ac.jp/) にて公開されているリボソーム蛋白質遺伝子から、実際には、以下に示した 12 種類のリボソームタンパク質遺伝子を持つ表.1 に示す 20 種類の生物の塩基配列データを用いる。

RPL3 RPL4 RPL5 RPL6
RPL7 RPL7A RPL8 RPL9
RPL10 RPL10A RPL17 RPL23

表 1 実験に使用する生物種

| 略称 | 生物名 | 略称 | 生物名 |
|---|---|---|---|
| Hs | ヒト | Fg | 赤カビ病菌 |
| Mm | マウス | Dd | キイロタマホコリカビ |
| Rn | ラット | Yl | アルカン資化酵母 |
| Fr | フグ | Sp | 分裂酵母 |
| Ci | ホヤ | Sc | 出芽酵母 |
| Dm | ショウジョウバエ | Cn | クリプトコッカス菌 |
| Ag | ガンビアハマダラカ | Um | トウモロコシ黒穂病菌 |
| Am | セイヨウミツバチ | Ro | クモノスカビ |
| Ce | センチュウ | At | シロイヌナズナ |
| Mg | イモチ病菌 | Cr | コナミドリムシ |

### 3.2 実験手順

ここでは、実験に用いる進化系統樹作成のための遺伝行列の作成手順を示す。

**Step1.** 読み込む塩基配列データ、標本サイズ、出現確率で扱う塩基の数を設定

**Step2.** 塩基配列データにマルチプルアラインメントを行う

**Step3.** 塩基配列データの各塩基を数字化

**Step4.** 標本サイズに基づき塩基配列の切り出し

**Step5.** 標本中の塩基の出現確率を計算

**Step6.** 各標本の情報エントロピーを計算

**Step7.** 情報エントロピーから遺伝行列を作成

### 3.3 実験結果

各生物種のリボソームタンパク質遺伝子の塩基配列から計算した情報エントロピーを基に、ユークリッド距離、差の絶対値、エントロピー進化率の三種類の方法から進化系統樹を作成し、生物種の類似性比較を試みた。切り出す標本のサイズを 64 と 128 とし、塩基の出現確率の取り方は 1 塩基または 3 塩基単位で行った。

図.3, 図.4 は、RPL10 遺伝子を基に計算した情報エントロピーのグラフである。標本サイズは 64 で切り出し、出現確率は 1 塩基単位で計算している。グラフの横軸は塩基配列上の塩基位置を示し、縦軸は情報エントロピーの値を示す。図.3, 図.4 を見てみると、ヒトとマウスの情報エントロピーの値は近く、類似性は高かった。一方、ヒトとセイヨウミツバチの場合は、セイヨウミツバチの情報エントロピーは変動が大きく、ヒトとマウスに比べて類似性が低かった。

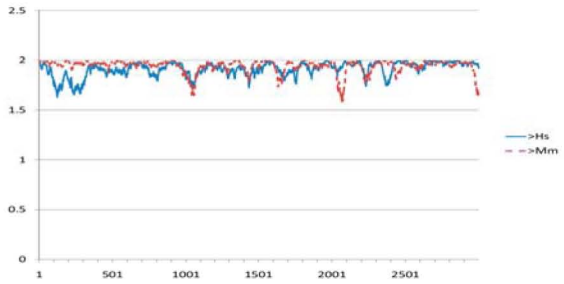


図 3 ヒトとマウスの情報エントロピー

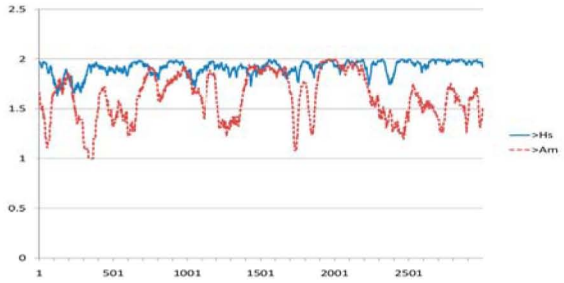


図 4 ヒトとセイヨウミツバチの情報エントロピー

これらの値を基に生物がどれだけ類似しているかを測るため、各情報エントロピーを基に相関係数を計算したのが表.2 である。

表 2　相関係数 1

| | Hs | Mm | Am |
|---|---|---|---|
| Hs | 1 | 0.166 | 0.072 |
| Mm | | 1 | -0.164 |
| Am | | | 1 |

表を見てみると、どの生物ペアも相関が強いとは言えず、生物の類似などは判断できない。そこで、より詳しく調べるために、各塩基配列にマルチプルアラインメントを行い、再度情報エントロピーを計算する。

図.5, 図.6 は、アラインメントを行った場合の情報エントロピーのグラフである。

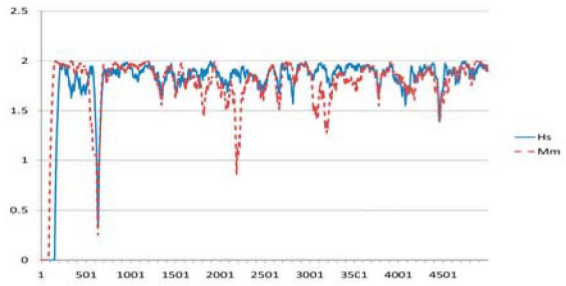


図 5　アラインメントありのヒトとマウスの情報エントロピー

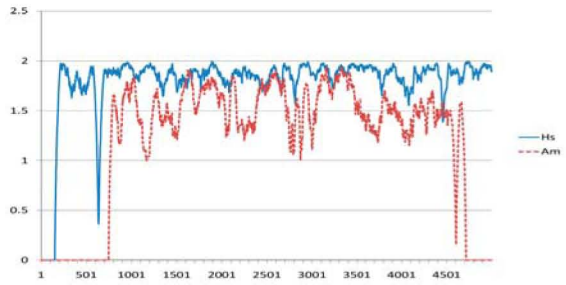


図 6　アラインメントありのヒトとセイヨウミツバチの情報エントロピー

アラインメントを行った塩基配列の情報エントロピーは、アラインメントを行うことによって入るギャップにより塩基の出現確率が変動し、アラインメントを行っていない場合に比べて値の変動が大きくなっており、また、アラインメントによって塩基配列の対応付けが行われたことにより元々配列が類似しているほど情報エントロピーの値が一致している部分が多くなっている。

また、これらの情報エントロピーの相関係数を計算すると、表.3 のようになる。

表 3　相関係数 2

| | Hs | Mm | Am |
|---|---|---|---|
| Hs | 1 | 0.717 | 0.399 |
| Mm | | 1 | 0.192 |
| Am | | | 1 |

この場合、各生物ペアの相関係数は全体的に相関が強くなっており、加えてヒトとマウスなどの進化の枝分かれが近いと考えられている生物の相関が、ヒトとセイヨウミツバチのように枝別れが遠いと考えられている生物の相関より強くなっている。図を見比べてみると、アラインメントを行っていない塩基配列を基にした情報エントロピーのグラフに比べて、アラインメントを行った場合のグラフでは、ヒトとマウスの情報エントロピーは一部の値の差異が大きくなっているものの、値が一致している部分、値が近い部分などが多くなり、全体的に類似性が見られる。また、ヒトとセイヨウミツバチの情報エントロピーの値は、アラインメントによる対応付けによって一部類似性が見られるようになったが、ヒトとマウスの情報エントロピーよりも情報エントロピーの差異が大きく、類似性は低い。以上のように、各生物の塩基配列を基に作成した情報エントロピーは、各生物ごとに違っており、グラフにして表すとヒトやマウスのように情報エントロピーの変動の小さい生物、セイヨウミツバチのように変動の大きい生物などそれぞれ違った特徴を示していた。

これら生物種間の類似性をより詳しく調べるために、ユークリッド距離、差の絶対値、エントロピー進化率を用いて進化系統樹の作成を行った. 図 7 はユークリッド距離を用いて遺伝行列を計算し、それを基に作成した進化系統樹である。

作成した各進化系統樹を CLUSTALW によって作成した進化系統樹とを比較してみると、ヒト、マウス、ラットなど一部に類似した箇所が見られるが、全体的にみると進化系統樹の形に類似性は見られなかった。

これら各進化系統樹がどの程度類似しているか、Split Distance を用いて計算した結果が表.4 である。表の系統樹作成方法についている数字は、情報エントロピー計算時に扱う塩基数と切り出す標本サイズである。

表を見ると、CLUSTALW と今回計算した各方

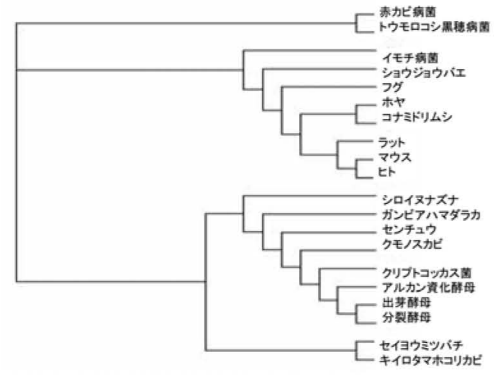


図 7　ユークリッド距離を用いて遺伝行列を計算した進化系統樹

表 4　Split Distance による各系統樹の類似度

| 系統樹の作成方法 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Clustalw | 0 | 43 | 37 | 36 | 35 | 32 | 37 | 35 |
| 進化率 3・64 | 43 | 0 | 46 | 45 | 45 | 43 | 46 | 44 |
| ユークリッド 3・64 | 37 | 46 | 0 | 22 | 18 | 25 | 10 | 22 |
| 差の絶対値 3・64 | 36 | 46 | 22 | 0 | 13 | 7 | 21 | 5 |
| ユークリッド 1・64 | 35 | 45 | 18 | 13 | 0 | 15 | 17 | 15 |
| 差の絶対値 1・64 | 32 | 43 | 25 | 7 | 15 | 0 | 24 | 6 |
| ユークリッド 3・128 | 37 | 46 | 10 | 21 | 17 | 24 | 0 | 21 |
| 差の絶対値 3・128 | 35 | 44 | 22 | 5 | 15 | 6 | 21 | 0 |

法はどれも類似性が高いとは言えず、差の絶対値を用いて作成した系統樹との類似度が比較的高いといった程度である。

以上の結果から、情報エントロピーを用いての生物種の分類は、グラフを用いての比較などでは生物種の類似性の分類などの指標の一つとして利用できると思われるが、進化系統樹を作成する各方法には、まだまだ研究の余地があると思われる。

## 4　おわりに

本研究では、DNA 塩基配列の不規則性に着目し、それをもとに生物種の類似性を測ることを目的とした。実験では各生物の塩基配列中の各塩基の出現確率をもとに情報エントロピーを求め、それを利用して進化系統樹を作成して生物種の類似性の比較を行った。

実験の結果、各生物の情報エントロピーの値は生物ごとに大きく違っており、生物種比較の指標の一つとして扱うことが出来るのではないかと思われる。一方で、情報エントロピーを基に作成した進化系統樹と CLUSTALW の類似性は低く、まだ研究の余地がある。

今後の課題として、情報エントロピーグラフを利用したの相関関数のさらなる計算、情報エントロピーの位置的比較など、計算した情報エントロピーを基にした比較実験の拡張があげられる。また、情報エントロピーを用いて系統樹を作成する場合の遺伝行列の作成に本研究ではエントロピー進化率、ユークリッド距離、差の絶対値を用いて進化距離行列の作成を行ったが、それらとは別の方法を用いた進化系統樹の作成に最適な遺伝行列の作成なども考えられる。。

## 参考文献

[1] 宮田隆, 五条堀孝, ゲノム情報を読む, 共立出版 (1997).

[2] 丸山修, 阿久津達也, バイオインフォマティクス - 配列データ解析と構造予測 - , 朝倉書店 (2007).

[3] 大矢雅則, 情報進化論, 岩波書店 (2005).

[4] 平野裕嗣, 山木隆史, 大矢雅則, “エントロピー進化率による HIV の解析”, 信学技報, No. IT98, pp. 61–66 (1998).

[5] 岸浪リサ, “類似構造検索機能を持つ分子系統樹データベースシステムの実現”, 情報処理学会論文誌, No. SIG 6, pp. 54–65 (1999).