画像圧縮技術を用いた生物種の類似性比較

吉原 郁夫[1] ・ 高島 弘明[2] ・ 山森 一人[3] ・ 菅原 研[4]

Similarity Comparisons of Seeds Using Image Compression Technology

Ikuo Yoshihara ,Hiroaki Takashima, Kunihito Yamamori, Ken Sugawara

Abstract

Discovering signs of the evolution concealed in the intron that exists in eukaryote's gene to clarify how the seed had evolved becomes a problem. Kind edge relation of the seed can be examined by comparing genomes of different seeds. It takes plenty of time to compare genomes in detail over the wide range. However, if the range for which it searches beforehand is understood, it is efficient. It aims to squeeze the range of the retrieval of the genome in this research.

This paper proposes a method of image compression for the genome sequence and an index comparing seeds. Seeds are compared with compression rates corresponding to different seeds. The experiments reveal the compressibility is related to similarity between seeds.

Keywords:
Genome, Seeds, Image Compression, DNA

1. はじめに

遺伝子は生物の設計図であり、生物が種の保存のために親から子に伝える自己の形質に関する情報が記されている。遺伝情報はDNAに塩基配列として刻まれており、アデニン(A)、チミン(T)、シトシン(C)、グアニン(G)の4つの塩基の組み合わせにより様々なタンパク質が作り出され、生物の細胞を構成し生命機能の維持を担っている。

生物には細胞に核を持つ真核生物と核を持たない原核生物があり、ヒトは真核生物に属する。真核生物の遺伝子の中にはタンパク質の合成に寄与するエクソンと直接寄与しないイントロンがあり、タンパク質の合成には直接寄与しないイントロン領域に進化の痕跡が隠されているのではないかと考えられている[1]。

共通の祖先をもつ生物のDNAには共通部分がある。2つの生物種が共通祖先から分かれてからの時間が短いほど、DNAの共通部分が多い[2]。現在の地球上の生物はすべて約40億年前に誕生した生命に起源し、それが進化して現存の多様な生物種が存在していると考えられている。その進化の道筋を一本の木になぞらえて、各種生物の系統としての位置関係を分かりやすく表現したものを系統樹と呼ぶ[1]。

本研究では、DNAの塩基配列の中の塩基を一つ一つ比較するのではなく複数の塩基を一まとまりとして比較を行う。従来、DNAを圧縮し比較する手段として塩基の並びをコンピュータ上で文字列として扱い辞書を用いて圧縮する手法がある[3]。そこでDNAの塩基配列を記したテキストデータをデータ圧縮技術を用いて圧縮し、その際の圧縮率を指標として比較を行う手法を試みる。また、イントロン、エクソン、イントロン+エクソンの3つの場合に分けて比較を行いそれぞれにどのような傾向が表れるのか調べる。

辞書を使った手法では塩基配列データの並びに着

1) 情報システム工学科教授
2) 情報システム工学科学部生
3) 情報システム工学科助教授
4) 東北学院大学教養学部助教授

目して圧縮しているが、提案手法では塩基配列データの規則性に着目して圧縮する。そこでアデニン(A)、チミン(T)、シトシン(C)、グアニン(G)4 つの塩基の並びの変化を白黒画像の階調変化に見立てて画像圧縮技術を用いて圧縮する。

2. 圧縮率による塩基配列の比較

画像処理において相続くフレームが類似している性質に基づき圧縮した際の特徴を以ってカット(シーンとシーンの変わり目)を検出する手法がある[4]。これから類似した塩基配列のデータを圧縮した際の圧縮率も似通っているのではないかと考え、データ圧縮を利用してスケッチの認識を行う手法[5]の考え方を参考に比較を行う。これは標本データを圧縮対象に付加して圧縮し、得られた圧縮率が小さいほど標本データに似ているとしてスケッチの認識を行っている。本研究では生物種間で共通する遺伝子を用い遺伝子から一定長塩基配列を切りだし圧縮し、得られた圧縮率をヒトと他の生物種で比較してヒトに近い生物、遠い生物でどのような違いが出るか確かめる。図 1 に提案手法の流れを示す

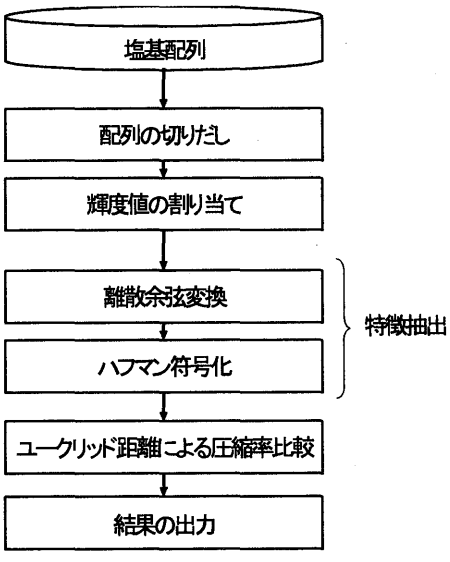

図 1　提案手法の流れ

2.1 塩基配列の切りだし

進化の痕跡を探るために遺伝子の塩基配列データからイントロン部分を切り出すが、エクソン部分、イントロン+エクソンとの違いがあるか調べるためにエクソン部分、イントロン+エクソンも個別に切り出す。遺伝子ごと生物種ごとに配列の長さが異なるため共通する遺伝子によって切り出す長さを変える。配列を切り出すときに切り捨てる部分が少なくなるように今回の実験では切り出す長さは 400～1200 塩基(以下単位略)とした。後述する輝度値の割り当てで 3 連続塩基に対して 1 つの輝度値を割り当てるが、3 つの組がどこから始まるか分からないため先頭位置を 1 つ、2つ、ずらして切りだし、元の配列から 3 つのデータを得る。切り出された配列の先端から 3 文字ずつずらしながら輝度値を割り当てるが、配列の終端では 3 つの組が作れない場合があるのでその部分のデータは切り捨てる。図 2 は配列切りだしの概念を示す。切りだす配列長の算出方法を以下の(A)～(C)に示す。

(A)塩基配列の長さがほぼ同じ場合

生物種ごとの塩基配列の長さがあまり変わらない場合は対応する遺伝子の組の中で一番短い塩基配列を持つ生物種の長さで切り取る。

(B)塩基配列の長さが不揃いな場合

塩基配列の長さが不揃いな場合は生物種の塩基配列の長さを平均した長さで切り取る。ただし、1200 を超える長さの塩基配列をもつ生物種は長さを 1200 として平均長を算出する。

(C)極端に短い配列がある場合

エクソンはイントロンに比べて短いので生物種によっては配列の長さが 400 を切る場合がある。この

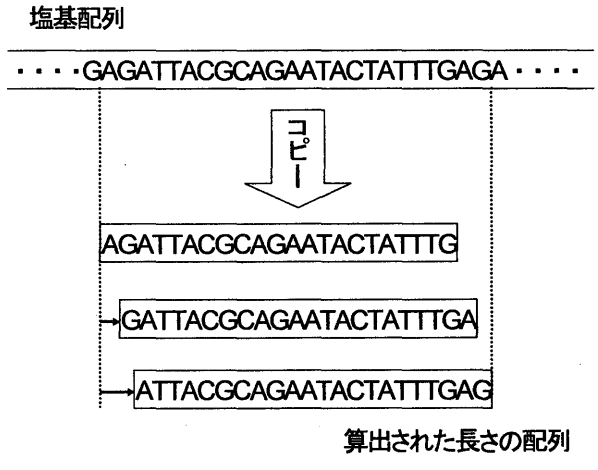

図 2　配列の切りだし

場合には圧縮した際の圧縮率が高くなるよう配列の

後ろに輝度の平均値を付けて補完し切り取る長さは全生物種で 400 とする。

2.2 輝度値の割り当て

一般的な白黒画像では輝度値は 0(黒)～255(白)の値をとること、また、塩基配列の 3 つの組で 1 つのアミノ酸を構成することから、輝度値の割り当ては 3 つの組の組み合わせ 64 パターンに対して 0～252 の間で 4 つおきに値を割り当てることとした。図 3 に塩基配列の文字列から数値への変換を示す。

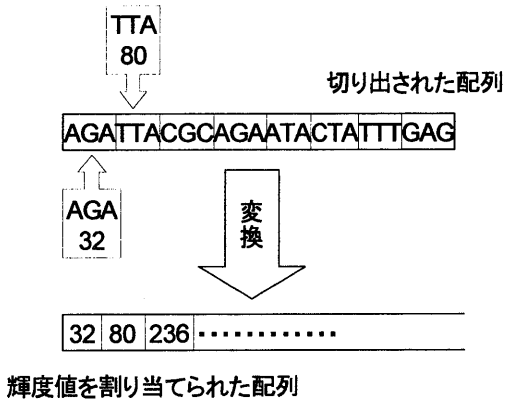

図 3　輝度値の割り当て

2.3 特徴抽出

圧縮方法として JPEG で使われている離散余弦変換とハフマン符号化を用いる[6][7]。これらを使うのはアルゴリズムが簡単で世の中に広く普及しているため実験方法の実行速度や精度に対する改善案が得やすいと考えたためである。

初めに切り出された配列を離散余弦変換で周波数成分に変換したあと低周波領域から高周波領域にかけて 5 等分する。次に分割された周波数領域をそれぞれハフマン符号化で圧縮し圧縮率を得る。1 つの遺伝子の塩基配列データから切り出される配列が複数ある場合は複数の配列から得られた圧縮率を平均してその遺伝子の圧縮率とする。図 4 に周波数領域による分割を示す。

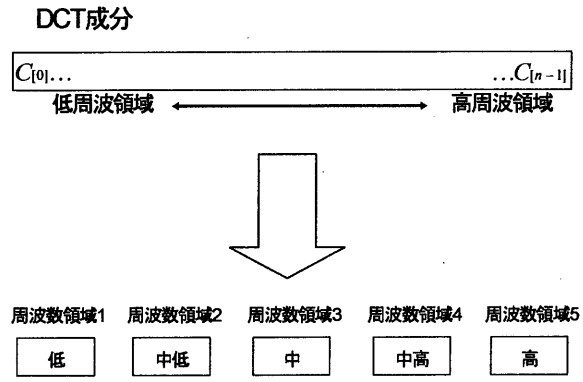

図 4　周波数領域による分割

2.3.1 離散余弦変換

信号の中にどの周波数成分がどれだけ含まれているのかを調べるために直交変換の一種で、離散フーリエ変換より符号化効率の良い離散余弦変換(Discrete Cosine Transform : DCT)を使用する。図 6 に周波数成分検出の例を示す。画像圧縮に使う場合は人間の目にはほとんど認識することができない高周波成分を切り捨てることでデータ量を大幅に小さくすることができる。しかし、本研究では、なるべく多くの特徴を拾いたいので高周波数成分を切り捨てることはしない。DCT の式(1)と基底(2)を以下に示す。

$$C[k]=\sum_{i=0}^{n-1} f[i]\phi_k[i] \quad (k=0,1,\cdots,n-1) \tag{1}$$

$$\phi_k[i]=\begin{cases}\dfrac{1}{\sqrt{N}} & (k=0)\\ \sqrt{\dfrac{2}{N}}\cos\dfrac{(2i+1)k\pi}{2N} & (k=1,2,\cdots,n-1)\end{cases} \tag{2}$$

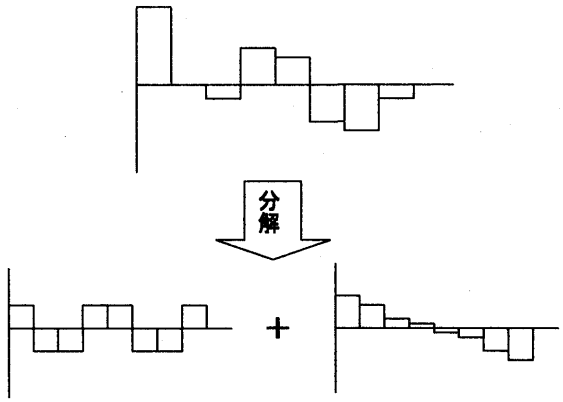

図6 周波数成分の検出

2.3.2 ハフマン符号化

ハフマン符号化はデータ圧縮の古典的な方法であり、出現頻度の高いデータには短い符号、低いデータには長い符号を割り当てることで全体のデータ量を減らすことのできる符号化方式。データの偏りが大きいほど圧縮が効く、データの偏りが少ない場合は逆にデータ量が増加する。例えば、AAAABBC という文字列があった場合、「A」「B」「C」、3 文字の出現率は順に「4」「2」「1」であり、出現率の低いものから束ねていくと図 7 のようになる。ここで、左に行けば 0、右に行けば 1 の符号を与えると A は 0、B は 10、C は 11 となり全体の大きさは 10 ビットになる。元のままでは 3 文字の識別に 2 ビット必要で、全体では 14 ビットになるため 4 ビットの削減となる。

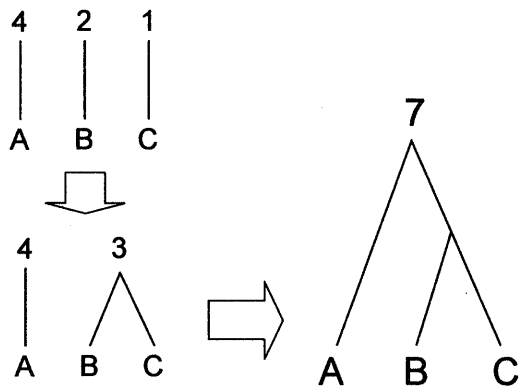

図7 ハフマン木の生成

2.4 評価方法

ヒトと他の生物種で共通する遺伝子の塩基配列データから得られた圧縮率を比較する。周波数領域により 5 つに分割した上で圧縮しているので、1 つの遺伝子につき 5 つの圧縮率が得られる。5 つの周波数領域において高周波領域のものほど圧縮率が高くなる傾向があるので周波数領域ごとに何らかの特徴が隠れていると考えられる、よって 5 つの圧縮率の値を 5 次元のベクトルとみなしてユークリッド距離を用いヒトとの距離を算出する、距離がヒトに近いほど、進化の過程で枝分かれしてからの時間が短いと考え ClustalW[8]で作成した系統樹と比較し、距離の遠近が系統樹上での位置に対応しているか確認する。

3. 塩基配列比較実験

3.1 実験データ

実験には、医学部より提供していただいたリボソームタンパク質を構成する遺伝子を用い、その中から真核生物であるヒト(Hs 動物)、マウス(Mm 動物)、ショウジョウバエ(Dm 動物)、出芽酵母(Sc 菌類)、分裂酵母(Sp 菌類)、シロイヌナズナ(At 植物)の 6 種類の生物種の遺伝子のデータを使用する。1 生物種あたりの遺伝子の種類は 76 種類である。系統樹上でこれらは根からヒトに至る経路で植物、菌類、ショウジョウバエ、マウスの順に分岐する。異なる

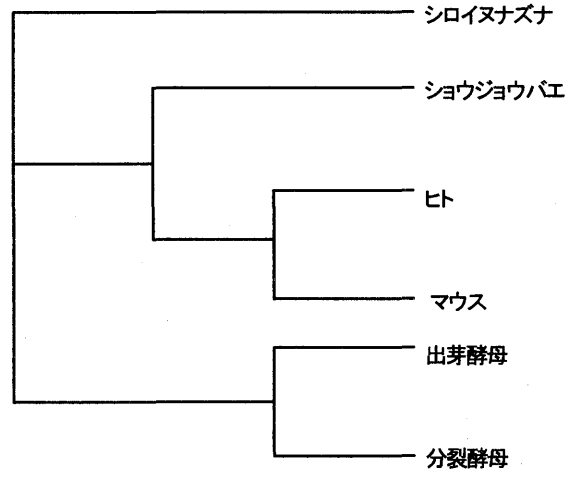

図8 ClustalW による系統樹

生物種間で互いに同じ名前を持つ遺伝子を比較する。図8に実験に使用した生物種の系統樹を示す。

3.2 圧縮率ベクトル比較

生物種間の比較はヒトを基準とし、遺伝子の塩基配列データを圧縮して得られた圧縮率ベクトルをヒトと他の生物種で比較する。表1に、生物種ごとのイントロン部分の平均圧縮率を示す。表2にエクソン部分、表3にイントロン＋エクソンの平均圧縮率を同様に示す。表1、表2、及び表3から、圧縮率の差に対して直接得点を与えて比較した場合、生物種ごとの違いが分かりにくいと思われるので、イントロン、エクソン、イントロン＋エクソンの3つの場合でそれぞれユークリッド距離がヒトに近い生物から順に、100、80、60、40、20、と等間隔に20点刻みで得点を与える。図9はイントロン、エクソン、及びそれらの組み合わせについて生物種ごとの得点の平均値を示したものであり、得点が高いほどヒトとの圧縮率の差が小さいことを示している。図9からイントロン、エクソン、イントロン+エクソンの3つの場合で概ね進化の系統樹上でヒトに近いものほど圧縮率の差が小さくヒトから離れるほど差が大きくなっていることが分かる。

表1　イントロン部分の平均圧縮率

生物種 / 周波数領域	Hs	Mm	Dm	Sc	Sp	At
低	67.6	67.4	67.6	67.5	67.1	66.8
中低	64.4	64.8	64.8	64.9	64.6	64.2
中	64.2	64.2	64.6	64.4	64.4	63.6
中高	63.9	64.2	64.3	64.3	64.1	63.0
高	63.5	63.8	63.8	63.8	63.6	63.6
平均	64.7	64.8	65.0	64.9	64.7	64.2

表2　エクソン部分の平均圧縮率

生物種 / 周波数領域	Hs	Mm	Dm	Sc	Sp	At
低	58.6	58.7	58.5	58.5	58.7	58.3
中低	57.8	57.7	57.4	57.0	57.4	57.5
中	57.8	57.9	57.2	57.3	57.3	56.9
中高	57.5	57.6	57.1	57.4	57.6	56.8
高	57.5	57.4	56.5	56.5	57.8	57.4
平均	57.9	57.9	57.5	57.3	57.7	57.4

表3　イントロン＋エクソン部分の平均圧縮率

生物種 / 周波数領域	Hs	Mm	Dm	Sc	Sp	At
低	69.6	68.8	69.9	69.7	69.5	68.9
中低	64.7	65.2	65.2	65.4	64.9	64.4
中	64.3	64.5	65.1	65.0	64.2	63.6
中高	64.1	64.1	64.4	64.5	64.4	63.4
高	63.3	63.0	64.0	64.1	63.7	62.7
平均	65.2	65.1	65.7	65.7	65.3	64.6

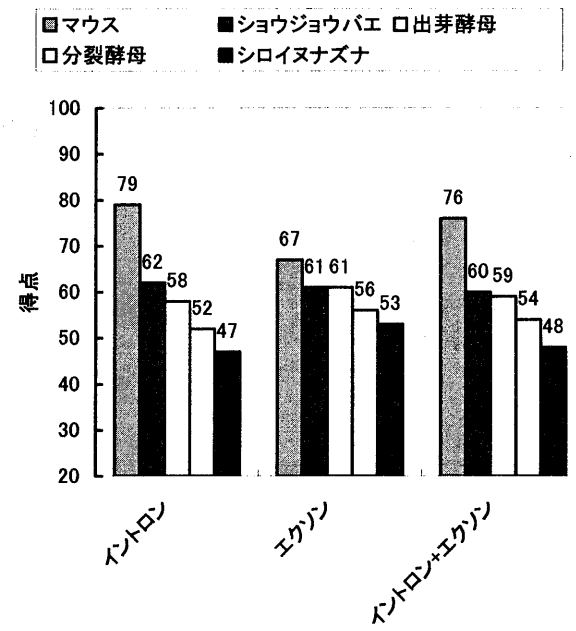

図9　ユークリッド距離による比較

4. おわりに

生物の進化を探るために画像圧縮技術を用い圧縮率を指標として比較する手法を試みた。まず、生物種の塩基配列データを画像とみなし、離散余弦変換により5つの周波数領域に分解する。次に各周波数領域でハフマン符号化により圧縮を行い、その圧縮率の差に基づき進化系統樹上の位置との比較をおこなった。

その実験結果からイントロン、エクソン、イントロン+エクソンの3つの場合で概ね進化の系統樹上でヒトに近いものほど圧縮率の差が小さくヒトから離れるほど差が大きくなっている。イントロンは生物種ごとの差がはっきりしているが、エクソンについて差はあるもののイントロンやイントロン+エクソンに比べ小さい。イントロン+エクソンで差が再び開いているのは高等生物になるほど遺伝子の中で多くの割合を占めるイントロンの影響が出ているためと考えられる。提案手法はおおまかに生物種の進化系統樹上の位置を推定することができたと思われる。

今後の課題としては、実験に使った遺伝子の数が少ないのでデータを増やして実験することである。

謝辞

フロンティア化学実験総合センターの剣持直哉助教授にはリボソーム遺伝子のデータを使用させていただいた。また、本研究の一部は(独)日本学術振興会科学研究補助金基盤研究 C(課題番号 17500146)による。

参考文献

[1]緒方宣邦, 野島博, “遺伝子工学キーワードブック改定第2版”, 羊土社, 2000

[2]岡崎康司, 坊農秀雅, “ゲノム情報はこう活せ!”, 羊土社, pp. 16-27, 2005

[3]S. chiba, K. Sugawara, “Estimation of Protein's Function by Evolutional Dictionary Method”, CEC2002, pp. 315-320, 2002

[4]有木康雄, “DCT特徴のクラスタリングに基づくニュース映像のカット検出と記事切りだし”, 信学論(D-Ⅱ), vol. J80-D-Ⅱ, no. 9, pp. 2421-2427, Sept. 1997

[5]近藤邦広, 加藤直樹, 渡辺俊典, “データ圧縮を利用したオンライン・スケッチ認 識手OSR”, 情処論, vol. 38, no. 12, pp. 24682478, Dec. 1997

[6]酒井幸市, “デジタル画像処理入門”, CQ 出版社, pp. 178-194, 2002

[7]奥村晴彦, “Javaによるアルゴリズム事典”, 技術評論社, pp. 408-412, 2003

[8] ftp://ftp.ebi.ac.uk/pub/software/