2V-3

データ圧縮率による生物種の比較法

高島　弘明† 吉原　郁夫† 山森　一人† 菅原　研†† 剣持　直哉†††
宮崎大学工学部† 東北学院大学教養学部†† 宮崎大学フロンティア科学実験総合センター†††

1. はじめに

真核生物の遺伝子の中にはタンパク質に翻訳されるエクソンと翻訳されないイントロンがあり、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。

ヒトと他の生物種のゲノムを比較することで、ヒトと共通している遺伝子を見出したり、種固有の遺伝子の有無、タンパク質に翻訳されない領域の意義を調べたり、既に機能や構造が分かっている遺伝子から未知の遺伝子の機能や構造を推測できる。

共通の祖先をもつ生物のゲノム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してからの時間が短ければ短いほど、ゲノムの共通部分が増えてくる[1]。

本研究では、ゲノム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纏まりとして簡易な比較を行う。そのためゲノムの塩基配列を画像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提案する[2]。また、イントロン、エクソン、イントロン+エクソンの 3 つ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。

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を検出する手法がある[3]。これから類似したゲノムを圧縮した際の圧縮率も似通っていると考え、データ圧縮を利用してスケッチの認識を行う手法[4]の考え方を参考に、圧縮方法として JPEG で使用されている離散コサイン変換(DCT)とハフマン符号化を用いる。本研究では生物種間で対応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に掛ける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、遠い生物をふるい分ける。提案手法の概念を（図１）に示す。

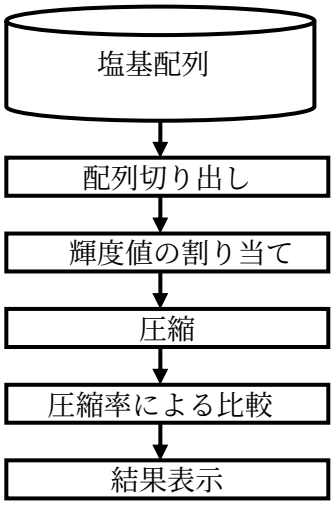

図 1 比較の手順

3. 実験

3.1 塩基配列の切りだし

実験には各生物種に共通して存在するリボソームタンパク質を用いる。実験データを(表 1)に示す。1 つの遺伝子につき、イントロン、エクソン、イントロン+エクソンの 3 つに対して 圧縮、比較を行うが、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の組によって 400～1200 の長さで切り出す。データの切り出し方を以下の(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 6 種類の生物種の塩基配列の長さを平均した長さで切り取

Comparison method of seed by data compression rate
†Hiroaki Takashima,Ikuo †Yoshihara and †Kunihito Yamamori
†Faculty of Engineering,,University of Miyazaki
††Ken Sugawara
††Faculty of Liberal Arts,Tohoku Gakuin University
†††Naoya kenmochi
†††Frontier Science Research Center, University of Miyazaki

表 1　実験データ

生物種（6 種類）	ヒト マウス ショウジョウバエ 出芽酵母 分裂酵母 シロイヌナズナ
遺伝子の種類	76 個

る。ただし、1200 を超える長さの配列をもつ生物種の配列の長さを 1200 として計算する。

(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一部の生物で配列の長さが 200 を切る場合がある。この場合には配列の後ろに輝度の平均値を付けて補完する。

3.2　輝度値の割り当て

輝度値の割り当ては塩基の三つの組に対して行い 0～255 の値を割り当てる。例えば、AAA なら 0、AAG なら 12、GCA なら 224 となるようにする。

3.3　圧縮

初めに切り出されたデータを離散コサイン変換に掛け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圧縮率ベクトルとみなす。1 つの遺伝子でデータが複数切り出された場合はデータから得られた圧縮率をそれぞれ平均してその遺伝子の圧縮率ベクトルとする。

3.4　比較結果

ヒトと他の 5 種類の生物種で共通する遺伝子から得られた圧縮率ベクトルを比較する。比較にはユークリッド距離を用い、距離がヒトに一番近い生物種に 100、次に近い生物種に 80 その次に近い生物種に 60 と 20 点刻みで得点を与える。76 個の遺伝子で比較した結果の平均値を(図 2)に示す。

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。その実験結果からイントロン、エクソン、イントロン+エクソン進化の系統樹の通りに並んでい

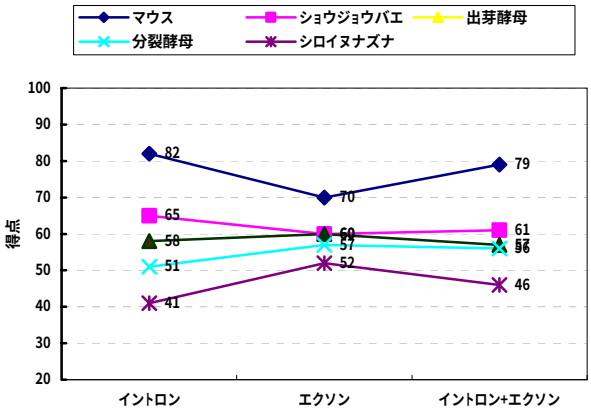

図 2 実験結果

る。イントロンの 3 つとも概ね進化の系統樹の通りに並んでいる。イントロンでは生物種ごとの差がはっきりしているが、エクソンについては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遺伝子の中で大多数を占めるイントロンの影響が出ていると考えられる。実験結果から提案手法は生物種の違いを調べられることが確認された。

今後の課題としては比較方法が用意したデータを相対的に比較したものであるので、生物種の類似性を距離に応じて判断できるようにすることである。

謝辞

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。

文　献

[1] 岡崎康司,坊農秀雅,”ゲノム情報はこう活かせ!”羊土社,2005

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

[3] 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

[4] 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・スケッチ認識手法 OSR,”情処論 , vol.38,no.12,pp.2468-2478,Dec.1997