論文の内容の要旨

応用生命工学専攻
平成 10 年度博士課程進学
氏　　名　相良純一
指導教官名　清水謙多郎

論文題目　統計的手法を用いた遺伝子配列解析と系統分類および分子進化への応用に関する研究

1　はじめに

本研究で用いられている主成分分析法 (Principal Component Analysis, PCA) と多次元尺度構成法 (Multidimensional Scaling analysis, MDS) は、多変量解析の一種として、よく用いられている手法である。PCA は、多くの情報をもつ個体間の相関関係を調べる手法であり、1 対 1 の相関関係だけでなく、全体あるいは部分の統計的相関を包括的に抽出することができる。この手法を配列解析に用いることにより、通常の相同性解析と比較し、複数の配列全体の統計的相関を抽出することができるという利点もある。MDS は多次元の情報を低次元 (二次元) 平面上に射影する手法であり、この手法を配列解析に適用することにより、配列の数だけ次元をもつ相関情報を、低次元の相関として可視化することができる。

本研究では PCA と MDS を再帰的に適用するなど手法の改良を行うとともに、塩基のコード化を行って、遺伝子の配列解析への適用を試みた。

2　手法

配列および配列中の塩基を数値的に取り扱うために、塩基をバイナリコードに変換し、遺伝子の塩基配列をこれらバイナリコードの列として表す。このバイナリコードの列として表された配列を各行とする行列がアライメント行列 $\boldsymbol{F}$ である。配列中の 4 種の塩基、アデニン (A)、シトシン (C)、グアニン (G)、チミン (T) は、それぞれ 4 ビットのバイナリコー

ド 1000、0100、0010、0001 に変換される。アライメント行列 $\boldsymbol{F}$ は、バイナリコードの列である塩基配列を縦に並べたベクトルであり、シーケンスプロファイルを形成する。

次に、配列間の相関をハミング距離で表す相関行列 $\boldsymbol{C} = \boldsymbol{F}\boldsymbol{F}^T$ を計算する。$\boldsymbol{C}$ の固有値 λ_p は、固有ベクトルを $\vec{u}_p$ とすると、$\boldsymbol{C}\vec{u}_p = \lambda_p \vec{u}_p$ ($\vec{u}_p$ は固有ベクトル) から求まり、これら λ_p と $\vec{u}$ の値を用いて主成分負荷量を計算することができる。

配列の各主成分負荷量 x_p^k は、$x_p^k = \sqrt{\lambda_p} u_p^k$ で求められ、これら x_p^k の値を 2 つの主成分ベクトルで定義される二次元平面上に射影することで、解析するすべての配列を二次元平面上にプロットすることができる。また、塩基の各主成分負荷量 y_p は、$y_p = \boldsymbol{F}^T \vec{u}_p$ で求められ、これら y_p の値を、2 つの主成分ベクトルで定義される二次元平面上に射影することにより、解析する配列に含まれるすべての塩基を二次元平面上にプロットすることができる。

上で求めた配列と塩基のそれぞれの主成分負荷量を 2 次、3 次と次々に求めていき、それらを二次元平面に射影することにより、配列と塩基の相関関係を同一の空間上で表すことができる。また、配列と塩基の相関関係をそれぞれがプロットされている方向で比較することができる。

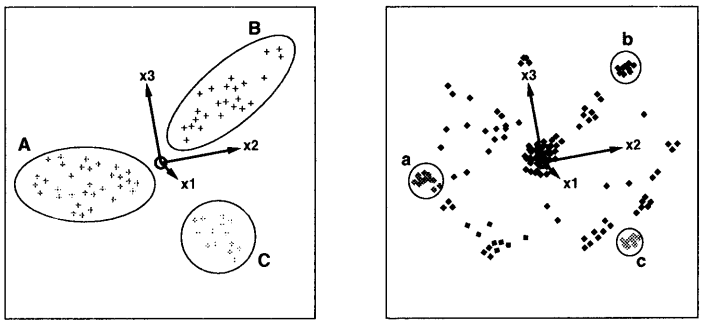

図 1: 塩基配列空間：x1 が第 1 主成分、x2 が第 2 主成分、x3 が第 3 主成分の方向を示す。x2 と x3 で形成される二次元平面に配列や塩基が射影される。(I) 配列をこの空間上に射影した図。各点は配列に対応している。(II) 配列中の塩基をこの空間上に射影した図。各点はアライメントされた配列中の 1 つの塩基に対応している。

3 大腸菌 tRNA の *identity* の予測

大腸菌の tRNA 遺伝子配列を、Class I と Class II の 2 つのグループに分類し、それぞれのグループに本手法を再帰的に適用した。

本手法により検出された特徴的な部位のうち、Class I ではおよそ 40% の部位が、Class II では 20 % の部位が従来の研究により明らかにされている *identity* と一致した。今回、新たに検出された特徴的な部位は T あるいは D ドメインといった tRNA の L 字型構造中のヒンジの部位に多くみられた。このことにより、tRNA の *identity* が同属のアミノアシル

シンセターゼとの結合に関与しているだけではなく、tRNA の微細な立体構造の違いの決定や tRNA のダイナミクスにも関与していることが示唆される。

4　大腸菌とメタン菌、酵母の tRNA 遺伝子配列の解析

大腸菌、メタン菌、酵母の tRNA 遺伝子配列に本手法を適用し、3 つの超生物界における tRNA 遺伝子配列の相関関係を解析した。解析の結果、大腸菌とメタン菌、酵母の細胞質の tRNA 遺伝子配列間の相関が高く、酵母のミトコンドリアの tRNA 遺伝子配列の相関が、それら 3 つの配列と相関が低いことがわかった。このことから、ミトコンドリアの tRNA 遺伝子が、他の 3 つの tRNA 遺伝子と早い時期に分岐したか別の祖先分子由来であると考えられる。

本研究により抽出された特徴的塩基は、生物種により多少の差はあるものの tRNA の二次構造における各ドメインのステム部位 (特に D ドメイン) に多く存在し、その多くは tRNA の L 字型構造を形成するために重要な高次水素結合を行う塩基であった。また立体構造の上でも、これらの特徴的塩基は tRNA のヒンジ部位の内側に集中し、しかも空間的に非常に狭い範囲に存在しており、このような部位に特徴的塩基が多く存在するということは、立体構造の微細な違いを進化の過程において維持していくことが重要であることを示している。

5　ミトコンドリア tRNA の分子進化に関する研究

ミトコンドリア tRNA 遺伝子配列に本手法を適用し、遺伝子配列の階層的な解析と特徴的塩基の抽出を行った。

データベースに登録されている全ての生物種のミトコンドリア tRNA の遺伝子配列を用いた解析では、それぞれの tRNA の遺伝子配列がコードしているアミノ酸ごとにまとまって二次元平面上にプロットされ、それぞれのアミノ酸の性質によって tRNA 遺伝子配列の配列パターンが似ていることが示された。また、単細胞・菌類の配列は原点付近に、動物の配列は原点から離れた位置に、植物の配列が両者の中間の位置に多く存在することから、tRNA 遺伝子配列が高等生物になるにしたがって他の生物種とは異なる特徴的な塩基を獲得していったと考えられ、これらの特徴的塩基がコードしているアミノ酸の性質の、また、生物種の分岐となっていると考えられる。

生物種ごとの tRNA 遺伝子配列の解析では、ミトコンドリア tRNA 遺伝子配列が、コードしているアミノ酸ごとにまとまっていることが示された。二次元平面上にプロットされた配列は動物、植物において高等な生物種になるにしたがって原点から離れたところに位置しており、このことは、進化にしたがって配列に変異が蓄積され、種間の遺伝子配列の違いが大きくなることを表している。また、動物、植物、単細胞・菌類ごとの解析結果を比較すると、(1) 射影された種間の配列の分離度は、植物や動物よりも単細胞・菌類のほうが小さい、(2) 単細胞・菌類の tRNA の配列パターンの相関が高く、動物や植物の tRNA の配列パターンの相関が低いことがわかた。このことから、単細胞から植物、節足動物、脊椎動

物の進化の過程において、tRNA 遺伝子配列が変異によって特徴的な遺伝子配列を徐々に獲得していったと考えられ、tRNA 遺伝子配列の祖先分子が単一あるいは数種であったことが示唆される。また、初期遺伝暗号におけるコドンの 20 種類のアミノ酸の翻訳が最小あるいは最小に近い数の tRNA 種で行われていたという tRNA の起源説とも合致している。

本研究により抽出された特徴的塩基は tRNA の二次構造における T あるいは D ドメインに多く存在した。これらの部位は、tRNA の立体構造における L 字型構造のヒンジの部分にあたり、tRNA の立体構造やダイナミクスを決めている。そのような部位に特徴的塩基が多く存在するということは、ミトコンドリア tRNA の分子進化において、L 字型構造の維持や変化が、種形成や種分化の大きな要因になっていると考えられる。

6 多変量解析を用いた未知配列の系統分類に関する研究

高 GC 含量グラム陽性菌の *gyrB* 配列に本手法を適用し、属の同定されていない配列の系統分類を行った。

本手法より得られた結果が、最尤法にもとづく分類とほぼ一致し、系統分類手法として本手法が有用であることが示された。また、複数の主成分負荷量を比較し、弱い配列パターンを抽出することにより、複数の分類の候補を挙げ、これにより分類の再検討が必要である配列や、新しい属と考えられる配列の検出に成功した。

本手法は最尤法と比べて計算のアルゴリズムが単純であるため計算コストも非常に小さく、簡便な系統分類予測の方法として適している。

7 まとめ

多変量解析の一種である PCA および MDS を再帰的に適用する方法を考案し、大腸菌 tRNA の *identity* の予測、大腸菌とメタン菌、酵母の tRNA 遺伝子配列の解析、ミトコンドリア tRNA の分子進化に関する研究、多変量解析を用いた未知配列の系統分類に関する研究に適用して、その有用性を示した。

今後は、本手法と隠れマルコフモデル (Hidden Markov Model) やニューラルネットワーク、相互情報量 (amount of Mutual Information) のような情報科学的手法とを組み合わせることで、より広範で柔軟な配列解析を行い、さらなる特徴的塩基の抽出や有意な塩基の絞り込みを行えることが期待できる。

参考文献

[1] Sagara, J.-I., Shimizu, S., Kawabata, T., Nakamura, S., Ikeguchi, M. and Shimizu, K., The use of sequence comparison to detect 'identities' in tRNA genes, *Nucleic Acids Res.*, 26:1974-1979, 1998.