4ZK-02

タンパク質配列のグラフ表示に適したアミノ酸指標の組み合わせの探索

土山 啓汰 †
弘前大学大学院理工学研究科 †

水田 智史 ‡
弘前大学大学院理工学研究科 ‡

1 研究の背景と目的

アライメントは配列比較に用いられる一般的な手法であるが、配列の入れ替えに対応せず計算量が大きいという問題点がある。そのため、本研究室ではアミノ酸にベクトルを割り当てグラフィカル表現を行う手法が提案されてきた [1]。これにより配列比較の直観的な評価を可能とし、また、定量的な評価も行う。

本研究の目的は、アミノ酸のベクトルの割り当て方に注目し、アライメントと同等の系統樹をより少ない計算量で作成することである。

2 方法

2.1 アミノ酸指標の除外

アミノ酸指標のデータベース AAindex [2] は合計で 566 の指標が採録されている。我々は、互いに独立しかつアミノ酸の特性を適切に表す指標を得るため、以下の 3 つの条件に合致する指標をあらかじめ除外した。

1. 指標内の異なるアミノ酸の間で重複した値をもつもの
2. 外れ値をもつもの
3. 指標間の相関係数の絶対値が 0.99 以上のもの

2.2 アミノ酸配列のグラフ表示

残った 111 種のアミノ酸指標のすべてのペアの相関係数を計算し、続いてアミノ酸指標の 2 個から 10 個組のすべての組み合わせに対して 得られる相関係数の絶対値の和が小さい順に 100 個選んだものを対象とする。指標の組は、組数が 2 個から 6 個のときは全探索、7 個から 10 個のときは焼きなまし法を用いて決定した。なお、アミノ酸指標は最小値 0、最大値 1 に正規化したもの、および平均 0、分散 1 に標準化したものを用いる。

選択した指標の組み合わせをもとにアミノ酸 20 種にベクトルを割り当てた後、配列の情報をより反映させるために重みを利用する。重みは

$$-\log_{10}\frac{\text{各アミノ酸の数}}{\text{アミノ酸の総数}}$$

により求めた。

得られたベクトルをもとに多次元座標群を作成する。図 1 はアミノ酸に与えられたベクトルの例で、図 2 はそれをもとに作成した配列 ACDEF のグラフである。

後に紹介する哺乳類 9 種類の ND5 配列を用いた 3 次元グラフは図 3 のようになる。3 個の指標の組み合わせの相関係数の絶対値の和が最小の指標の組を使用しており、各アミノ酸指標は標準化を行っている。

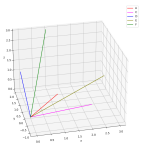

図 1 アミノ酸 A, C, D, E, D のベクトル

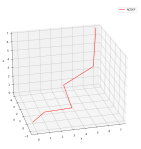

図 2 配列 ACDEF のグラフ

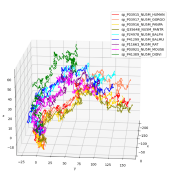

図 3 ND5 配列のグラフ

2.3 フィッティングにより配列間の距離の算出

求めた多次元座標群に最もよく適合できる直線を求める。この直線は座標群に対して主成分分析を行って得られた最大固有値に属する固有ベクトルであり、この方向ベクトルを特徴量とする。

配列のペアに対して求めた方向ベクトルのなす角を θ として、$\cos\theta$ を計算する。方向ベクトルを $\vec{a}, \vec{b}$ とする

Search for suitable combinations of amino acid indices for graphical representations of protein sequences

† Satoshi Mizuta, Graduate School of Science and Technology, Hirosaki University
‡ Keita Tsuchiyama, Graduate School of Science and Technology, Hirosaki University

と、$\cos\theta$ は以下の式で求められる。

$$\cos\theta = \frac{\vec{a}\cdot\vec{b}}{|\vec{a}||\vec{b}|}$$

$\cos\theta$ の値からアークコサインにより求めた θ を配列間の距離と定義した。

2.4 実験に用いたデータ

本研究で使用した生物種は参考文献 [3] で用いられている哺乳類の 9 種類 (Human, Gorilla, Pygmy chimpanzee (P.chi.), Common chimpanzee (C.chi.), Fin whale (F.wh.), Blue whale (B.wh.), Rat, Mouse,Opposum (Oposs.)) で、タンパク質は参考文献 [4] で用いられている、ミトコンドリア DNA にコードされている 13 種類のうち、配列長の長い順に 10 種類 (ATPase 6, COI, COII, COIII, Cyt-b, ND1, ND2, ND4, ND5, ND6) を使用した。

3 結果

3.1 系統樹の作成と比較

コサイン類似度で求めた距離から距離行列を作成する。系統樹の作成には UPGMA 法を用い、 Newick 形式で出力する。可視化は系統樹分析・可視化フレームワーク ETEToolkit を用いる。本研究で作成した系統樹と下図で示す Clustal Omega [5] で作成した系統樹との RF 距離で比較する。

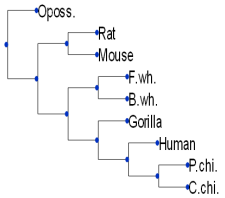

図 4 Clustal Omega で作成した系統樹

結果として RF 距離が 0 の系統樹は 81 個得られた。

また、 RF 距離が 0 の系統樹の作成に使われたアミノ酸指標の頻度上位 3 位は QIAN880138 (64 個)、NADH010107 (44 個)、RACS820102 (43 個) で、特定のアミノ酸指標が系統樹の作成に多く使われていることが判明した。

4 考察

4.1 RF 距離が 2 の系統樹

RF 距離が 2 の系統樹は 1216 個生成されたが、そのうち 1030 個は齧歯類が正確に分類できていないことが判明した。

4.2 計算量

配列長を N 、アミノ酸指標の個数を K とすると、Clustal Omega の計算量は $O(N\log N)$ であり、本研究の計算量は座標群の生成に $O(NK)$、分散共分散行列の計算に $O(N^2)$、固有値・固有ベクトル計算に $O(K^2)$、距離の計算に $O(K)$ である。ただし距離の計算以外は 1 回の前計算で済むため、配列比較の計算量は $O(K)$ である。

5 今後の課題

RF 距離が 0 の系統樹を生成した指標の組から最良の組を見つけることである。また、方向ベクトル以外の特徴量を利用すること、 RF 距離以外の基準を検討することが考えられる。

参考文献

[1] Yusei Kobori and Satoshi Mizuta, “Similarity estimation between DNA sequences based on local pattern histograms of binary images”, *Genomics, Proteomics & Bioinformatics* **14**(2), pp. 103–112 (2016).

[2] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa, “AAindex: amino acid index database, progress report 2008”, *Nucleic Acids Research* **36**(Database), pp. D202–D205 (2007).

[3] Agata Czerniecka, Dorota Bielińska-Wąż, Piotr Wąż, and Tim Clark, “20D-dynamic representation of protein sequences”, *Genomics* **107**(1), pp. 16–23 (2016).

[4] Chenglong Yu, Shiu-Yuen Cheng, Rong L He, and Stephen S-T Yau, “Protein map: an alignment-free sequence comparison method based on various properties of amino acids”, *Gene* **486**(1–2), pp. 110–8 (2011).

[5] Fábio Madeira, Matt Pearce, Adrian R N Tivey, Prasad Basutkar, Joon Lee, Ossama Edbali, Nandana Madhusoodanan, Anton Kolesnikov, and Rodrigo Lopez, “Search and sequence analysis tools services from EMBL-EBI in 2022”, *Nucleic acids research* **50**(W1), pp. W276–W279 (2022).