タンパク質配列グラフ表示画像の作成に適したアミノ酸指標の組み合わせの探索

土山啓汰 水田智史 *

これは非公式の情報処理学会東北支部研究会の発表資料テンプレートである。
uplatex,lualatex で動作することが確認されている。
なお、この abstract は再定義されているため、プリアンブルに記述すると動作しない。

1 イントロダクション

1.1 研究の背景と目的

本研究ではタンパク質のアミノ酸配列を扱う。アミノ酸はタンパク質の構成要素である。20 種類の異なるアミノ酸がタンパク質の合成に用いられる。各タンパク質の形状や他の特性はそこに含まれるアミノ酸の配列の仕方によって決定し、アミノ酸配列が似ているほど機能や性質が類似している可能性が高いと言える。

タンパク質のアミノ酸配列間の類似性を評価する方法として、一般的にアライメントが用いられている。しかし、配列長が N と N の場合は動的計画法により $O(N^2)$ の時間計算量となり、肥大な計算時間が必要となる。そのため、本研究室ではアライメントに依らないタンパク質アミノ酸配列比較の手法として、アミノ酸に何らかの 2 次元ベクトルを割り当てグラフィカル表現を行うことが提案されてきた。これにより、グラフが似ていれば類似性が高いといったような直観的な評価が可能となる。また、同時に定量的な評価も行う。

本研究では、この手法の精度を高めるために、アミノ酸のベクトルの割り当て方に注目し、遺伝子解析同等の系統樹を早く作成することを目的としている。

* 情報処理学会東北支部大学

2 方法

2.1 ベクトルの割り当て

タンパク質のアミノ酸配列を 3 次元座標群化するためにアミノ酸 20 種にベクトルを割り当てる。本研究では、AAindex に掲載されているアミノ酸指標 566 種類のうち、重複した値が使用されている場合を除外した 166 種類の指標を利用する。

具体的には、166 種類のうち 3 種類のアミノ酸指標を選択し、そのアミノ酸指標間の相関係数の絶対値の和を利用する。そこで使用したアミノ酸指標 3 種をそれぞれ x, y, z 座標に割り当てる。また、本研究では、絶対値の和を昇順に並べたリストを作成し、上位 1000 位までの指標を実験対象とした。

2.2 三次元座標群の作成

アミノ酸に割り当てたベクトルを元に三次元座標群を作成する。例として配列 ACDEF に対する座標群を表に示す。

表 1　3 次元座標群の例

	X	Y	Z
A	0	0	0
		↓	
C	0	0.83	1.083
		↓	
D	1.007	-0.177	2.942
		↓	
E	1.629	0.9	4.205
		↓	
F	2.205	-0.1	5.377
		↓	
	1.51	-1.305	6.79

2.3　重み

アミノ酸にベクトルを与えた後、配列の情報をより反映させるために重み利用する。重みには自己情報量を用いて

$$-\log 10 \frac{\text{各アミノ酸の数}}{\text{アミノ酸の総数}} \tag{1}$$

により求めた。この重みをそれぞれのベクトルに掛けたものを本研究では使用する。

表 2　重み

アミノ酸	重み	アミノ酸	重み	アミノ酸	重み
A	1.083	I	1.228	R	1.257
C	1.859	K	1.236	S	1.178
D	1.263	L	1.015	T	1.271
E	1.172	M	1.617	V	1.163
F	1.413	N	1.391	W	1.959
G	1.150	P	1.324	Y	1.535
H	1.643	Q	1.405		

2.4　距離行列の求め方

生物の配列のグラフ化行った後、それぞれのグラフに対して主成分分析を行う。本研究では、求めた主成分に対してコサイン類似度を適用した距離行列、スペクトル解析を行った後、コサイン類似度を適用した距離行列をそれぞれ求めた。

2.4.1　主成分分析

2.4.2　コサイン類似度

主成分分析で求めたそれぞれの第一主成分のなす角を θ として $cos\,\theta$ を計算する。方向ベクトルを $\vec{a}$、$\vec{b}$ とすると、$cos\,\theta$ は以下の式で求められる。

$$\cos = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|} \tag{2}$$

上記の式の値をアークコサインを用いて求めた θ の値を距離と定義した。

2.4.3　スペクトル解析

配列の位置を x 軸、主成分分析で求めた第二主成分を y 軸とした二次元グラフについて、離散フーリエ変換を用いて周波数成分を抽出する。

$$F(t) = \sum_{x=0}^{N-1} f(x) \exp(\frac{-i2\pi tx}{N}) \tag{3}$$

抽出した周波数成分のパワースペクトルを求める。

$$powerspectre = \sqrt{real^2 + imag^2} \tag{4}$$

パワースペクトルのコサイン類似度を求めた後、アークコサインを用いて求めた θ の値を距離と定義した。

3 結果

3.1 実験に用いたデータ

本研究で使用した生物種は以前の研究 [1] と同じ哺乳類の 9 種類で、以下の表の通りである。ミトゴンドリア DNA にコードされている NDAH デヒドロゲナーゼサブユニット 5（以下 ND5 タンパク質）を使用した。

表 3 使用した生物種

英名（和名）	accession no.	配列長
Human：ヒト	AP_000649	603
Gorilla：ゴリラ	NP_008222	603
Pygmy chimpanzee（P.chi.）：ボノボ	NP_008209	603
Common chimpanzee（C.chi.）：チンパンジー	NP_008196	603
Fin whale（F.wh.）：ナガスクジラ	NP_006899	606
Blue whale（B.wh.）：シロナガスクジラ	NP_007066	606
Rat：ドブネズミ	AP_004902	610
Mouse：ハツカネズミ	NP_904338	607
Opposum（Oposs.）：オポッサム	NP_007105	602

3.2 系統樹の作成

生物種に対して、Clustal Omega を用いてマルチプルアライメントで作成した系統樹と、コサイン類似度、及びスペクトル解析で作成した系統樹を robinson-foulds 距離で比較する。なお、系統樹の作成には UPGMA 法を用いる。

4 まとめと今後の課題