

標準化レーベンシュタイン距離法を用いた アミノ酸置換スコア行列である BLOSUM の改良と応用 Improvement and Application of BLOSUM which is Blocks of Amino Acid Substitution Matrix Using Standardized Levenshtein Distance Method

金城 和輝 小泉 範子 伊藤 利明
Kazuki Kaneshiro Noriko Koizumi Tosiaki Itou

1. 背景・目的

タンパク質は細胞の主要な成分であり、様々な生命現象の発現の役割を果たし、固有の立体構造と機能を持っている⁽¹⁾。この立体構造を解明することで医療などに役立っている。現在タンパク質の構造の分析手法の多くは、X 線結晶構造解析法と核磁気共鳴法で行われている。しかし、時間やコストがかかり、この問題を解決する手段として計算機上でタンパク質の構造を推定するプログラムが用いられている。タンパク質の構造を推定するプログラムを用いる際にアミノ酸置換スコア行列である BLOSUM が存在し、特に BLOSUM62 が組み込まれたものを用いられている。しかし BLOSUM62 には情報エントロピーの問題が存在し、アミノ酸のスコアの偏りがあると指摘されている。今回標準化レーベンシュタイン距離法を用いて BLOSUM62 の改良と応用を行う。

2. 方法

2.1. BLOSUM

BLOSUM62 の算出手順は、事前に用いるアミノ酸配列を同じ長さのアミノ酸配列に揃えたブロックを作成し、用いるアミノ酸配列の一致率が 62%を上回ったアミノ酸配列同士をクラスタリングする。ブロック内で同じ箇所同士で各アミノ酸が対応しているかを算出し、そのアミノ酸のペアの確率を P_{ab} とする。ペアになった各アミノ酸が出現する確率を P_a と P_b とする。アミノ酸のペアのスコアは (1)式へ代入し、情報エントロピーから算出されるスケール因子 λ を掛け算した値の整数値をスコアである S として用いる⁽²⁾。

$$S = \frac{1}{\lambda} \log_2 \frac{P_{ab}}{2P_a P_b} \quad (1)$$

2.2. 標準化レーベンシュタイン距離法

レーベンシュタイン距離法は二つの文字列がどれくらい違うかを表す距離法である。算出方法は一方の文字列を置換、挿入を行いもう一方の文字列に変形するのに必要な最小の編集回数で、二つの文字列の距離を求める⁽³⁾。しかし、この方法では文字列の長さによって距離が違ってしまいうため、二つの文字列の長いほうで編集回数を割る標準化レーベンシュタイン距離法がある。この方法を用いて対象のアミノ酸配列と事前に存在するアミノ酸配列のそれぞれ距離を求め、その中から平均値以下のもので再度ブロックを作成した。もう一度(1)式に代入して対象のアミノ酸配列に合わせた BLOSUM の改良である LBLOSUM0 を作成した。次に、BLOSUM を作成過程で算出される情報エントロピーとスコアから算出される情報エントロピーを BLOSUM62, BLOSUM0 と LBLOSUM0 それぞれ既存のプログラムで比較を行った⁽⁴⁾。

$$P_{ab} = 2P_a P_b \cdot 2^{\lambda S} \quad (2)$$

2.3. 人工設計タンパク質

人工設計タンパク質とは自然界に存在するタンパク質を再設計し、今までないタンパク質を設計されたものをいう⁽⁵⁾。今回、この人工設計タンパク質をタンパク質構造予測プログラムである I-TASSER を用いて BLOSUM62, BLOSUM0 と LBLOSUM0 のそれぞれでタンパク質の構造を推測した⁽⁶⁾。解析に用いた人工設計タンパク質を表 1 に示す。推測したタンパク質をそれぞれ NMR 法で解析したタンパク質の構造との一致率を検討した。比較の項目としてアミノ酸と原子のそれぞれの種類と順番を検討項目として、一致した数を I-TASSER によって推定されたタンパク質の総数で割った値を一致率として算出した。

表 1.人工設計タンパク質の名前とアミノ酸配列

人口設計タンパク質	アミノ酸の配列
1E65A	AECSVDIQGNDQMVFNTNAITVDKSCQFTVNLSPGNLPKNVMGHNWVLSTAADMQGVV TDGMASGLDKDYLPDDSRVIAHTKLIGSGEKDSVTFDVSKLKEGEQYMFFCTFPGHSA
1FSDA	QQYTAKIKGRTFRNEKELRDFIEKFKGR
1G6UA	SLAALKSELQALKKEGFSPEELAALESELQALEKKLAALKSKLQALKG
1J4MA	RGKWTYNGITYEGR
1JY4A	RGECKFTVPGR TALNTPAVQKWHFVLPGYKCEILA
1JY9A	TTTTTRYVEVPGKKILQTTTT
1L2YA	NLYIQWLKDGGPSSGRPPPS
1LQ7A	GSRVKALEEKVKALEEKVKALGGGGRIEELKKKWEELKKKIEELGGGGEVKKVEEEVKKL EEEIKKL
1MFTA	MDYLRELYKLEQQAMKLYREASEKARNPEKKSVLQKILEDEEKHIEWLETING

3. 結果

3.1 BLOSUM62, BLOSUM0 と LBLOSUM0 の平均値と標準偏差

BLOSUM62, BLOSUM0 と LBLOSUM0 の平均値と標準偏差の計算方法とし表 2 に BLOSUM62 の一部を示す. BLOSUM は 20 行 20 列から構成される行列であり, この全要素から平均値と標準偏差を求めた.

表 2.BLOSUM62 の一部

BLOSUM62	K	R	H	D
K	5	2	-1	-1
R	2	5	0	-2
H	-1	0	8	-1
D	-1	-2	-1	6

表 3 に BLOSUM62, BLOSUM0 と LBLOSUM0 の平均値と標準偏差を示す. BLOSUM62 は BLOSUM0 と LBLOSUM0 に比べてすべてのサンプルで標準偏差が低く, LBLOSUM0 はすべて標準偏差が高い値を示した. このことから, LBLOSUM0 は BLOSUM62 と BLOSUM0 よりスコアの差が大きいことが判明した. また, 解析に用いた PC のスペックは intel® Core™ i5-2320 CPU @ 3.00GHz メモリは 3.8GiB の 64bit で解析時間は BLOSUM62 で約 2 秒かかり, BLOSUM0 と LBLOSUM0 は共に約 1 秒ぐらいであり BLOSUM62 より早い解析処理が行われた.

表 3.BLOSUM62 と BLOSUM0 と LBLOSUM0 の平均値と標準偏差

人工設計タンパク質	BLOSUM62		BLOSUM0		LBLOSUM0	
	平均値	標準偏差	平均値	標準偏差	平均値	標準偏差
1E65A					-2.915	3.149
1FSDA					-2.570	2.982
1G6UA					-2.495	3.009
1J4MA	-1.138	2.261	-2.395	2.911	-2.528	3.082
1JY4A					-2.575	3.075
1JY9A					-2.755	3.180
1L2YA					-2.563	3.028
1LQ7A					-2.465	2.986

3.2 情報エントロピー

表 4 より, BLOSUM62 は BLOSUM0 と LBLOSUM0 と比べて情報エントロピーの差がすべて大きかった. しかし, BLOSUM62 は BLOSUM0 と LBLOSUM0 と比べて BLOSUM 作成時の情報エントロピーの値が低くなっていた. また, LBLOSUM0 は用いた人工設計タンパク質の内 3 つが BLOSUM0 を超えたが, ほかのサンプルでは低い値となった.

表 4.BLOSUM62 と BLOSUM0 と LBLOSUM0 のそれぞれの情報エントロピー

人工設計 タンパク質	BLOSUM 作成時の 情報エントロピー			スコアからの 情報エントロピー			情報エントロピーの 差の絶対値		
	BLOS	BLOS	LBLOS	BLOS	BLOS	LBLOS	BLOS	BLOS	LBLOS
	UM62	UM0	UM0	UM62	UM0	UM0	UM62	UM0	UM0
1E65A			1.819			1.836			0.017
1FSDA			1.640			1.656			0.016
1G6UA			1.597			1.537			0.059
1J4MA	0.698	1.517	1.597	0.608	1.552	1.585	0.090	0.035	0.013
1JY4A			1.633			1.627			0.005
1JY9A			1.659			1.705			0.046
1L2YA			1.671			1.617			0.055
1LQ7A			1.589			1.562			0.028

表 5 にスコアから求めた情報エントロピーで算出された BLOSUM62, BLOSUM0 と LBLOSUM0 の平均値と標準偏差を示す. 表 3 の各値を比べると BLOSUM62 の平均値と標準偏差の値に違いがあった. BLOSUM0 と LBLSOUM0 は変化がなかったことから, 情報エントロピーの差は BLOSUM62 のみ影響を与えることが判明した.

表 5. スコアの情報エントロピーから求めた.BLOSUM62 と BLOSUM0 と LBLOSUM0 の平均値と標準偏差

人工設計タンパク質	BLOSUM62		BLOSUM0		LBLOSUM0	
	平均値	標準偏差	平均値	標準偏差	平均値	標準偏差
1E65A					-2.915	3.149
1FSDA					-2.570	2.982
1G6UA					-2.495	3.009
1J4MA	-1.625	3.215	-2.395	2.911	-2.528	3.082
1JY4A					-2.575	3.075
1JY9A					-2.755	3.180
1L2YA					-2.563	3.028
1LQ7A					-2.465	2.986

3.3 人工設計タンパク質の一致率

表 6 に I-TASSER を用いて BLOSUM62, BLOSUM0 と LBLOSUM0 のそれぞれで推測されたタンパク質の構造の一致率を示す. 1E65A と 1G6UA で BLOSUM0 と LBLOSUM0 が BLOSUM62 より高い一致率を示した. また, 全サンプルにおいて BLOSUM62 が BLOSUM0 と LBLOSUM0 の一致率を上回ることにはなかった.

表 6.人工設計タンパク質の正答率一覧 (単位:%)

人工設計タンパク質	BLOSUM62	BLOSUM0	LBLOSUM0
1E65A	87.523	87.573	87.573
1FSDA	88.000	88.000	88.000
1G6UA	87.715	87.961	87.961
1J4MA	88.889	88.889	88.889
1JY4A	79.612	79.612	79.612
1JY9A	84.659	84.659	84.659
1L2YA	89.412	89.412	89.412
1LQ7A	85.427	85.427	85.427

4. 考察

BLOSUM62 は BLOSUM0 と LBLOSUM0 に比べて情報エントロピーが低かった. BLOSUM62 では BLOSUM の作成の過程でクラスタリングを行うため情報エントロピーが低くなったが, BLOSUM 算出の過程で丸め込み誤差によって情報エントロピーの差が大きくなったと考えられる. BLOSUM0 と LBLOSUM0 はクラスタリングの過程がないため丸め込みの誤差がなく情報エントロピーの差が小さくなったと示唆される. よって, スコアから求めた情報エントロピーで算出された BLOSUM62, BLOSUM0 と LBLOSUM0 の平均値と標準偏差に変化がなかったと思われる. LBLOSUM0 は距離の近いサンプルで行っているため情報エントロピーの差が小さくなった. しかし, サンプルによってはアミノ酸のペアの偏りが生じたため標準偏差が大きく, 平均値も低くなったと考えられる. 人工設計タンパク質は自然界に存在しないタンパク質の構造を有しているため, 既存の BLOSUM62 では相同性検出の際にアミノ酸のペアのスコアではサンプルが少ないため対応できなかった. しかし, BLOSUM0 と LBLOSUM0 のアミノ酸のペアのスコアは BLOSUM62 に比べて偏りがあるため高く, その結果多くのサンプルを選出でき, 自然界に存在しない構造にも対応できたのではないかと考える.

5. 展望

標準化レーベンシュタイン距離法を用いる際に文字列の変換と挿入を行っているが, 実際の生物界ではランダムにアミノ酸が変化するとは考えにくい. そのため今後は文字列の変換と挿入の頻度を算出し, 確率的な重みを考えた標準化レー

ベンシュタイン距離法の改良が必要であると考えます。また、アミノ酸の配列だけでなく遺伝子などの文字列にも応用を試みたい。

6. 参考文献

- (1) 林 典夫, 廣野 治子, et al., シンプル生化学 改定第 6 版, (2007), pp.39-40, 南江堂.
- (2) S Henikoff, J G Henikoff, Amino acid substitution matrices from protein blocks, *Proc Natl Acad Sci U S A*, Vol.15 (1992), pp.10915-10919.
- (3) Black, Paul E, et al., Levenshtein distance, *Dictionary of Algorithms and Data Structures [online]*, 2016
- (4) Sean R Eddy, Where did the BLOSUM62 alignment score matrix come from?, *Nature Biotechnology*, Vol.22 (2004), pp.1035-1036.
- (5) David S. Goodsell, Molecule of the Month, RCSB, 2005
- (6) Yang Zhang, I-TASSER server for protein 3D structure prediction, *BMC Bioinformatics*, Vol.9 (2008), p.40.