Vol.2015-MPS-103 No.67
Vol.2015-BIO-42 No.67
2015/6/25

保存度と近接度を用いたタンパク質機能部位予測

近藤　洋介[1,a)]　權　娟大[2]　宮崎　智[3,b)]

概要：タンパク質の重要な部位を予測するために多くの手法が開発されており，(i) 配列，(ii) 構造，(iii) 配列と構造に基づいた方法などがある．本研究では，(iii) の手法について共結晶構造を利用する手法を提案する．そこで，多重配列アライメントを文字型と座標型の二種類のデータ型で表し，それらの写像を考える．前者では (i) の手法で用いられるような保存度を，後者ではアミノ酸残基とイオンや分子の近接度を計算し，保存度と近接度によってどのアミノ酸残基が重要であるかを見積もる．翻訳伸長因子 Tu/1A タンパクに本手法を適用し，保存度と近接度の相関を調べた．このとき，保存度が小さければ近接度も小さく，保存度が大きければ近接度も大きいという傾向がみられたが，完全な線形の相関は得られなかった．このため，保存度と近接度によって測れるものは，互いに類似ではあるが，全く同様ではないと推測できる．本手法により，保存度と近接度に基づいて少し異なる観点からアミノ酸残基を評価することができる．

1. 序論

タンパク質が働くとき，イオンや分子が結合する特異的な部位が存在する可能性がある．結合部位を同定することは，そのタンパク質がどのように働き，どのようにイオンや分子と結合するかを調べるために重要である．そのような重要な部位を予測するために多くの手法が開発されている．その中の配列に基づいた手法では，タンパク質の重要な部位が変異に対して保守的になるという仮定に基づいて，保存度のようなものを用いるものもある [1], [2], [3], [4], [5]．現在では様々な保存度が提案されており，これらを立体構造上へ表示することで保存度と三次元空間上のアミノ酸残基の配置の間の関係が調べられている [6]．しかし，この関係は一つの代表構造上への表示により調べられることが多く，複数の立体構造からその関係を調べていない．したがって，様々なイオンや分子に結合するタンパク質や同じ祖先に由来するタンパク質群が考慮できない場合もある．そのため，本研究では，多重配列アライメントの写像について考える．

2. 写像の方法

2.1 データ型

$\boldsymbol{M}$ が多重配列アライメントを示し，$f_x : \boldsymbol{M} \to [\,0, \infty\,)$ があるとする．${}_i\boldsymbol{M} \in \boldsymbol{M}$ が文字型と座標型の二つのデータ型で表されるとする．

2.2 文字型

図 1 の問題はどのように ${}_1\boldsymbol{M} - {}_5\boldsymbol{M}$ をそれぞれ識別するかである．${}^t_i\boldsymbol{M} \in {}_i\boldsymbol{M}$ を時点 t の ${}_i\boldsymbol{M}$ とし，$t = 1, 2, \ldots, N+1$ とする．ただし，N は内部節の数である．${}^t_i\boldsymbol{M}$ は文字の多重集合を表し，時点 $t+1$ で分かれるとする．例えば，${}^t_4\boldsymbol{M}$ は

- ${}^1_4\boldsymbol{M} = \{R, R, L, R, R, R\}$
- ${}^2_4\boldsymbol{M} = \{R, R, L, R, R\}$
- ${}^3_4\boldsymbol{M} = \{R, R, L, R\}$
- ${}^4_4\boldsymbol{M} = \{R, R, L\}$
- ${}^5_4\boldsymbol{M} = \{R, R\}$

である．ここでは，${}^t_i\boldsymbol{M}$ に二種類以上の文字が含まれる場合に 1，それ以外を 0 とする．${}^1_i\boldsymbol{M}$ の数値のみのとき，${}_1\boldsymbol{M} - {}_5\boldsymbol{M}$ は識別できない．${}^1_i\boldsymbol{M}$ と ${}^2_i\boldsymbol{M}$ の数値を足すと，${}_1\boldsymbol{M}$ とそれ以外を識別できる．${}^1_i\boldsymbol{M}$，${}^2_i\boldsymbol{M}$，${}^3_i\boldsymbol{M}$ の数値を足すと，${}_1\boldsymbol{M}$ と ${}_2\boldsymbol{M}$ とそれ以外を識別できる．そこで，

$$f_1\left({}_i\boldsymbol{M}\right) := \sum_{t=1}^{N} \begin{cases} 0 & (\forall l \in {}^t_i\boldsymbol{M}, \exists \lambda \in \boldsymbol{A} \cup {}_i\boldsymbol{\Gamma}; l = \lambda) \\ 1 & (\text{その他}) \end{cases} \tag{1}$$

とする．ただし，$\boldsymbol{A}$ はアミノ酸シンボルの集合，${}_i\boldsymbol{\Gamma} \subset {}_i\boldsymbol{M}$ は ${}_i\boldsymbol{M}$ の中のギャップの集合である．$\boldsymbol{A}$ と ${}_i\boldsymbol{\Gamma}$ はそれぞれ $\boldsymbol{A} = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$ と ${}_i\boldsymbol{\Gamma} = \left\{{}^1_i\gamma, {}^2_i\gamma, \ldots, {}^G_i\gamma\right\}$ とする．ただし，G はギャップの数である．

[1] 東京理科大学　大学院薬学研究科　薬学専攻
[2] 東京大学　大学院農学生命科学研究科　応用生命化学専攻
[3] 東京理科大学　薬学部　生命創薬科学科
[a)] j3a12701@ed.tus.ac.jp
[b)] smiyazak@rs.noda.tus.ac.jp

Vol.2015-MPS-103 No.67
Vol.2015-BIO-42 No.67
2015/6/25

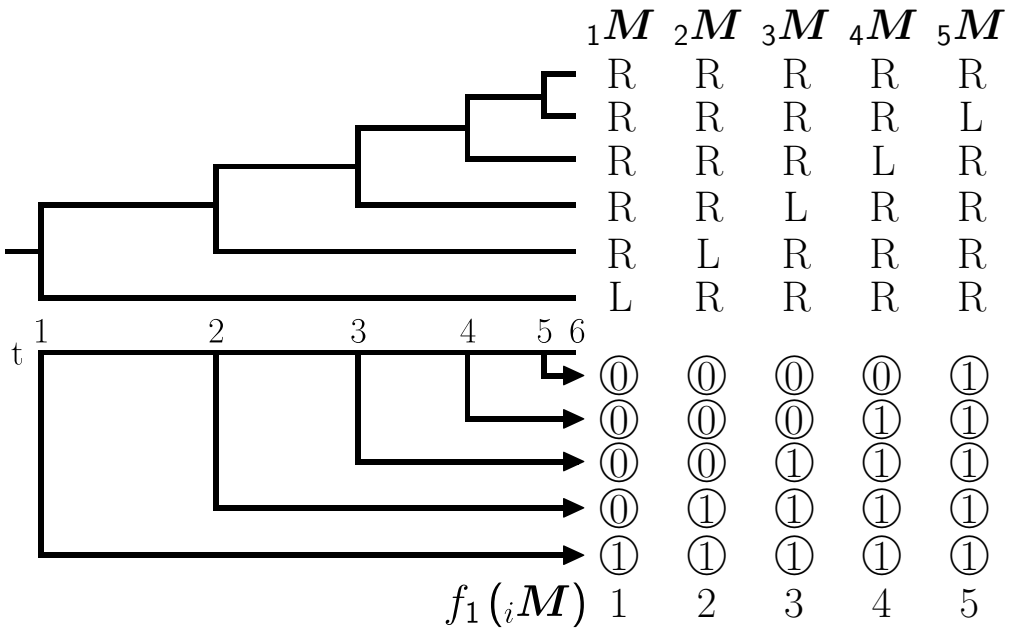

図 1　文字型の写像．$_i\boldsymbol{M} \in \boldsymbol{M}$ は全て 5 つの R と 1 つの L から成り，それぞれの文字には進化速度一定の仮定の基での有根系統樹の末端節が付いている．昇順の番号を時点 t として根から末端節に割り当てる．f_1 では，円の中の値を $_i\boldsymbol{M}$ に割り当てた後，それらの値の和を計算する．

座標型

$\mathbb{R}$ を実数の集合とし，$e : \mathbb{R}^3 \times \mathbb{R}^3 \to [\,0, \infty\,)$ があるとする．$\boldsymbol{R} \subset \mathbb{R}^3$ と $\boldsymbol{Q} \subset \mathbb{R}^3$ があり，

$$e\,(\boldsymbol{R}, \boldsymbol{Q}) := \min_{(\boldsymbol{r},\boldsymbol{q}) \in \boldsymbol{R} \times \boldsymbol{Q}} (\,\| \boldsymbol{r} - \boldsymbol{q} \|_2) \qquad (2)$$

とする．ただし，$\| \cdot \|_2$ はユークリッドノルムである．

あるタンパク質とイオンや分子が含まれる構造 k を考える．$_i^k\boldsymbol{R} \subset \mathbb{R}^3$ は構造 k 中のアミノ酸残基 i の原子座標を示し，$^k\boldsymbol{Q} \subset \mathbb{R}^3$ は構造 k 中のイオンや分子の原子座標を示す．K は構造数を示し，それらの配列はアライメントされているとする．$\left\{ {}_i^1\boldsymbol{R}, {}_i^2\boldsymbol{R}, \ldots, {}_i^{K-G}\boldsymbol{R} \right\} \subseteq {}_i\boldsymbol{M}$ は $_i\boldsymbol{M}$ の中の残基の集合，$\left\{ {}_i^1\gamma, {}_i^2\gamma, \ldots, {}_i^G\gamma \right\} = {}_i\Gamma \subset {}_i\boldsymbol{M}$ は $_i\boldsymbol{M}$ の中のギャップの集合とする．

$$f_2\,(_i\boldsymbol{M}) := \min_{{}_i^k\boldsymbol{R} \in {}_i\boldsymbol{M} \setminus {}_i\boldsymbol{\Gamma}} \left[e\left({}_i^k\boldsymbol{R}, {}^k\boldsymbol{Q} \right) \right] \qquad (3)$$

とする．

3. 材料と方法

3.1 データ収集

UniProtKB/Swiss-Prot release 2015_01 [8] から (1)‘Classic translation factor GTPase family. EF-Tu/EF-1A subfamily’ の注釈がある (2) 配列に ‘X’ を含まない (3) フラグメントでないの三条件を全て満たす 984 エントリーを抽出した．タンパク質構造データバンク (Protein Data Bank, PDB) [9] から (1) 上記の 984 エントリーの参照がある (2) X-線結晶構造解析で決定されたの二条件を両方とも満たす 68 エントリーを抽出した．免疫タンパクと結合している [10] またはキメラタンパクを形成している [11], [12], [13], [14] 14 エントリーを除外し，表 1 に示すように，103 鎖を含む 54 エントリーを残した．

3.2 f_1 と f_2 の計算

図 2 で $N = 984$ と $K = 103$ として，MAFFT 7 [15] を用いて配列をアライメントし，座標データを持つ残基を含む 477 の $_i\boldsymbol{M}$ を抽出した．

二配列間の距離を最尤法 [16] により計算した．このとき，アミノ酸置換モデルは Jones-Taylor-Thornton モデル [17]，平衡頻度は Dayhoff の方法 [18] を用いて計算した．二配列間の距離を全組み合わせについて計算した後，非加重結合法 [19] を用いて系統樹を作成し，$f_1\,(_i\boldsymbol{M})$ を計算した．

非対称単位に分けて，$f_2\,(_i\boldsymbol{M})$ を計算した．表 1 にそれぞれのエントリーの代表的なイオンや分子を示した．ただし，ナトリウムイオン，酢酸イオン，硫酸イオン，アンモニウムイオン，糖 (スクロース)，ジヒドロキシエーテル，グリオキシル酸，5-ブロモフラン-2-カルボン酸，β-メルカプトエタノール，水分子は，機能が明確ではないため除外した [7], [20], [21], [22], [23], [24], [25].

3.3 f_1 と f_2 の相関

$[\,0, \infty\,) \supset \boldsymbol{F} \ni f_1\,(_i\boldsymbol{M})$ は非負実数の部分集合であり，$f_1\,(_i\boldsymbol{M})$ の集合とする．$\boldsymbol{F} \ni v_1 < v_2 < \cdots < v_J$ と表す．

表 1　EF-Tu/EF-1A タンパクの PDB のエントリー

Subfamily	Organism	PDB ID	Resolution	Ions or molecules
EF-Tu	*Bos taurus*, mitochondrial	1D2E	1.94	GDP, Mg^{2+}
		1XB2	2.20	Elongation factor Ts mitochondrial
	Escherichia coli	1EFC	2.05	GDP, Mg^{2+}
		2HCJ	2.12	GDP, TAC, Mg^{2+}
		3U6B	2.12	GDP, Mg^{2+}
		2BVN	2.30	ENX, GNP, Mg^{2+}
		4G5G	2.30	Thiomuracin A derivative, GDP, Mg^{2+}
		1D8T	2.35	Thiocillin GE2270, GDP, Mg^{2+}
		3U6K	2.45	Thiocillin GE2270 analogue NVP-LDK733, GDP, Mg^{2+}
		1DG1	2.50	GDP, Mg^{2+}
		1EFU	2.50	Elongation factor Ts
		1EFM	2.70	GDP
		3U2Q	2.70	Thicillin GE2270 analogue NVP-LFF571, GDP, Mg^{2+}
		2HDN	2.80	GDP, TAC, Mg^{2+}
		1ETU	2.90	GDP, Mg^{2+}
		4Q7J	2.90	Elongation factor Ts, Qβ replicase
		1OB2	3.35	Phe-tRNA, GNP, KIR, Mg^{2+}
		2FX3	3.40	GDP, Mg^{2+}
	Pseudomonas putida KT2440	4J0Q	2.29	GDP, MES, MPD, Mg^{2+}
		4IW3	2.70	Putative uncharacterized protein, GDP, Mg^{2+}
	Thermus aquaticus	1EFT	2.50	GNP, Mg^{2+}
		1B23	2.60	Cys-tRNA, GNP, Mg^{2+}
		1TTT	2.70	Phe-tRNA, GNP, Mg^{2+}
		1TUI	2.70	GDP, Mg^{2+}
		1OB5	3.10	Phe-tRNA, ENX, GNP, Mg^{2+}
	Thermus thermophilus	2C78	1.40	GNP, PUL, Mg^{2+}
		2C77	1.60	Thiocillin GE2270, GNP, Mg^{2+}
		1EXM	1.70	GNP, Mg^{2+}
		4LBW	1.74	GNP, Mg^{2+}
		4H9G	1.93	GNP, Mg^{2+}
		1HA3	2.00	GDP, MAU, Mg^{2+}
		4LBV	2.03	GNP, Mg^{2+}
		4LBZ	2.22	GNP, Mg^{2+}
		4LC0	2.22	GNP, Mg^{2+}
		4LBY	2.69	GNP, Mg^{2+}
		1AIP	3.00	Elongation factor Ts
		4V5L	3.10	16S rRNA, 23S rRNA, Trp-tRNA, GCP, Mg^{2+}
		4V5P	3.10	16S rRNA, 23S rRNA, Trp-tRNA
		4V5Q	3.10	16S rRNA, 30S ribosomal protein S12, Trp-tRNA, GDP, KIR
		4V5R	3.10	16S rRNA, Trp-tRNA, GDP, KIR
		4V5S	3.10	16S rRNA, Trp-tRNA, GDP, KIR
		4V8Q	3.10	16S rRNA, 23S rRNA, Small protein B SMPB, tmRNA δ, GDP, KIR, Mg^{2+}
		4V5G	3.60	16S rRNA, 23S rRNA, 30S ribosomal protein S12, Thr-tRNA, GDP, KIR, Mg^{2+}
aEF1A	*Aeropyrum pernix*	3VMF	2.30	Peptide chain release factor subunit 1, GTP, Mg^{2+}
		3WXM	2.30	Protein pelota homologue, GTP, Mg^{2+}
	Sulfolobus solfataricus	1JNY	1.80	GDP
		1SKQ	1.80	GDP, Mg^{2+}
eEF1A	*Oryctolagus cuniculus*	4C0S	2.70	GDP, Mg^{2+}
	Saccharomyces cerevisiae	1F60	1.67	Elongation factor 1Bα
		2B7C	1.80	Elongation factor-1 β
		1G7C	2.05	Elongation factor 1-β, 5GP
		1IJE	2.40	Elongation factor 1-β, GDP
		2B7B	2.60	Elongation factor-1 β, GDP
		1IJF	3.00	Elongation factor 1-β, GDP

TAC; Tetracycline, ENX; Enacyloxin IIa, GNP; Phosphoaminophosphonic acid-guanylate ester, KIR; Kirromycin, MES; 2-(N-morpholino)-ethanesulfonic acid, MPD; (4S)-2-methyl-2,4-pentanediol, PUL; Pulvomycin, MAU; N-methyl kirromycin, GCP; Phosphomethylphosphonic acid guanylate ester, 5GP; Guanosine-5′-monophosphate.

t_j は閾値であり，

$$t_j \begin{cases} < v_1 & (j=0) \\ = \dfrac{v_j + v_{j+1}}{2} & (j=1,2,\ldots,J-1) \\ > v_J & (j=J) \end{cases} \tag{4}$$

を満たすとする．c_2 は $f_2\left({}_i\boldsymbol{M}\right)$ のカットオフとし，本研究では，$c_2 = 3$ Å とした．I_f は $f_2\left({}_i\boldsymbol{M}\right) > c_2$ を満たす ${}_i\boldsymbol{M}$ の数，I_t は $f_2\left({}_i\boldsymbol{M}\right) \leq c_2$ を満たす ${}_i\boldsymbol{M}$ の数である．$I_{fp}\left(t_j\right)$ は $f_2\left({}_i\boldsymbol{M}\right) > c_2$ かつ $f_1\left({}_i\boldsymbol{M}\right) \leq t_j$ を満たす ${}_i\boldsymbol{M}$ の数，$I_{tp}\left(t_j\right)$ は $f_2\left({}_i\boldsymbol{M}\right) \leq c_2$ かつ $f_1\left({}_i\boldsymbol{M}\right) \leq t_j$ を満たす ${}_i\boldsymbol{M}$ の数である．偽陽性率，感度，曲線下面積 (Area under the curve, AUC) をそれぞれ

$$p\left(t_j\right) = \frac{I_{fp}\left(t_j\right)}{I_f}, \tag{5}$$

$$q\left(t_j\right) = \frac{I_{tp}\left(t_j\right)}{I_t}, \tag{6}$$

$$AUC = \frac{1}{2}\sum_{j=0}^{J-1}\left[p\left(t_{j+1}\right) - p\left(t_j\right)\right]\cdot\left[q\left(t_{j+1}\right) + q\left(t_j\right)\right] \tag{7}$$

とする．

$\boldsymbol{F}_x \ni f_x\left({}_i\boldsymbol{M}\right)$ は $f_x\left({}_i\boldsymbol{M}\right)$ の多重集合とし，$\boldsymbol{F}_x \ni {}_i^1V_x \leq {}_i^2V_x \leq \cdots \leq {}_i^IV_x$ で表されるとする．ただし，I は ${}_i\boldsymbol{M}$ の数である．r が順位関数であり，

$$r\left({}_i^{j+k-1}V_x\right) = j - 1 + \frac{t_n+1}{2} \tag{8}$$

とする．ただし，$j = 1,2,\ldots,I$，$k = 1,2,\ldots,t_n$ であり，t_n は同順位の大きさである．ここで，スピアマンの ρ [26] は

$$\rho = \frac{T_1 + T_2 - \sum_{i=1}^{I}\left[r\left({}_i^lV_1\right) - r\left({}_i^mV_2\right)\right]^2}{2\sqrt{T_1T_2}} \tag{9}$$

である．ただし，$l = 1,2,\ldots,I$，$m = 1,2,\ldots,I$，

$$T_1 = \frac{\left(I^3 - I\right) - \sum_{n=1}^{N_1}\left(t_n^3 - t_n\right)}{12}, \tag{10}$$

$$T_2 = \frac{\left(I^3 - I\right) - \sum_{n=1}^{N_2}\left(t_n^3 - t_n\right)}{12} \tag{11}$$

である．ただし，N_1 は $\boldsymbol{F}_1$ の同順位の数，N_2 は $\boldsymbol{F}_2$ の同順位の数である．

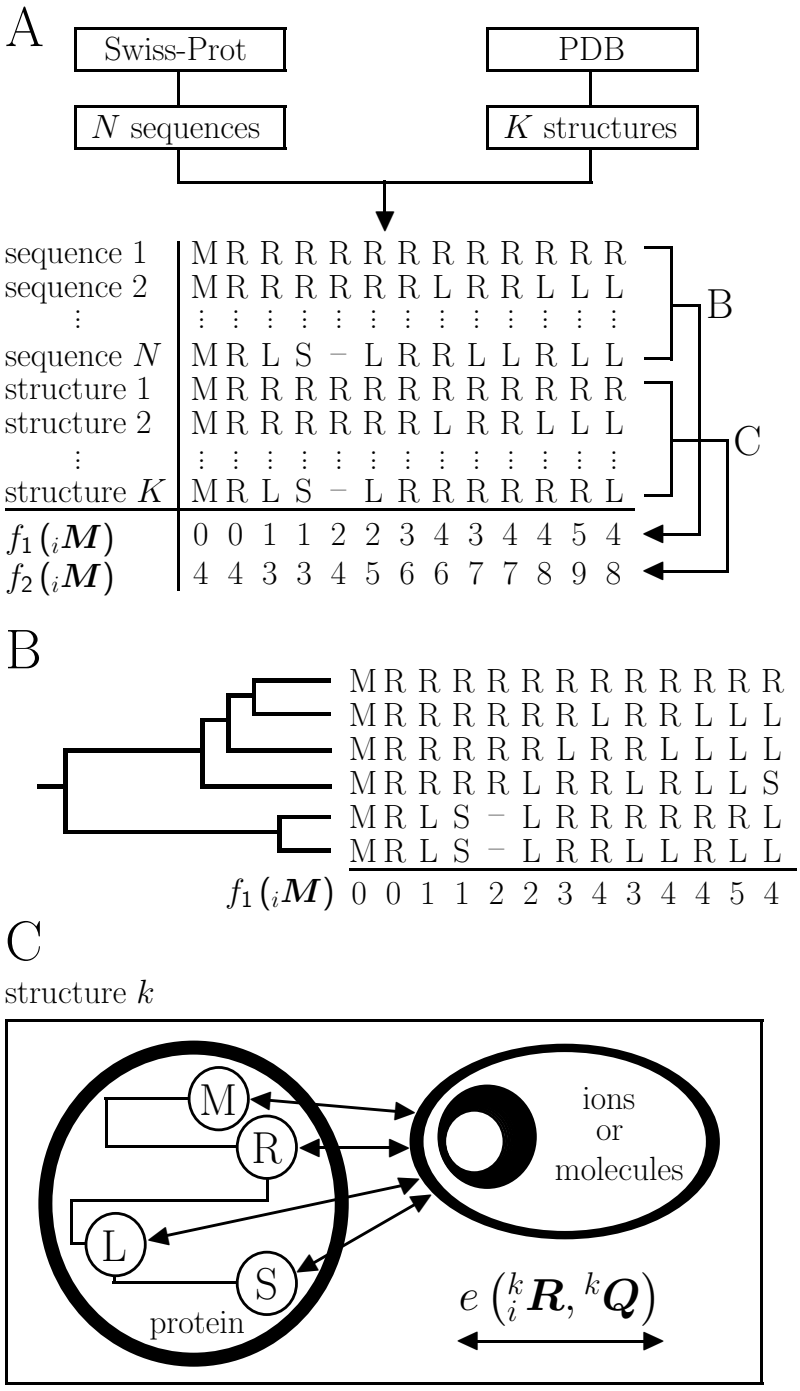

図 2 f_1 と f_2 **の計算．**(A) 配列と構造を Swiss-Prot と PDB からそれぞれ取得する．それらの全配列をアライメントした後，$f_1({}_i\boldsymbol{M})$ と $f_2({}_i\boldsymbol{M})$ を (B) と (C) によりそれぞれ計算する．(B) 配列から進化系統樹を作成した後，$f_1({}_i\boldsymbol{M})$ を計算する．(C) 構造 k の中で，${}^k_i\boldsymbol{R}$ と ${}^k\boldsymbol{Q}$ がそれぞれアミノ酸残基とイオンや分子の座標を示す．${}^k_i\boldsymbol{R}$ と ${}^k\boldsymbol{Q}$ の近接度を $e({}^k_i\boldsymbol{R}, {}^k\boldsymbol{Q})$ として K 構造に対して計算した後，$f_2({}_i\boldsymbol{M})$ を計算する．

3.4 可視化

$f_1({}_i\boldsymbol{M})$，$f_2({}_i\boldsymbol{M})$，AUC を Python のパッケージである matplotlib [27] を用いて可視化した．立体構造を VMD [28] を用いて可視化した．

4. 結果

図 3A では，図 3B の受信者動作特性 (receiver operating characteristic, ROC) 曲線 [29] を用いると $f_1({}_i\boldsymbol{M})$ と $f_2({}_i\boldsymbol{M})$ によって $\boldsymbol{M}$ を 4 つに分けれることを示している．このとき，$f_1({}_i\boldsymbol{M})$ と $f_2({}_i\boldsymbol{M})$ の AUC は 0.739 であり，スピアマンの ρ は 0.580 であった．図 3C では，左側は小さい $f_1({}_i\boldsymbol{M})$ と小さい $f_2({}_i\boldsymbol{M})$ が多いが，右側は大きい $f_1({}_i\boldsymbol{M})$ と大きい $f_2({}_i\boldsymbol{M})$ が多いことを示している．

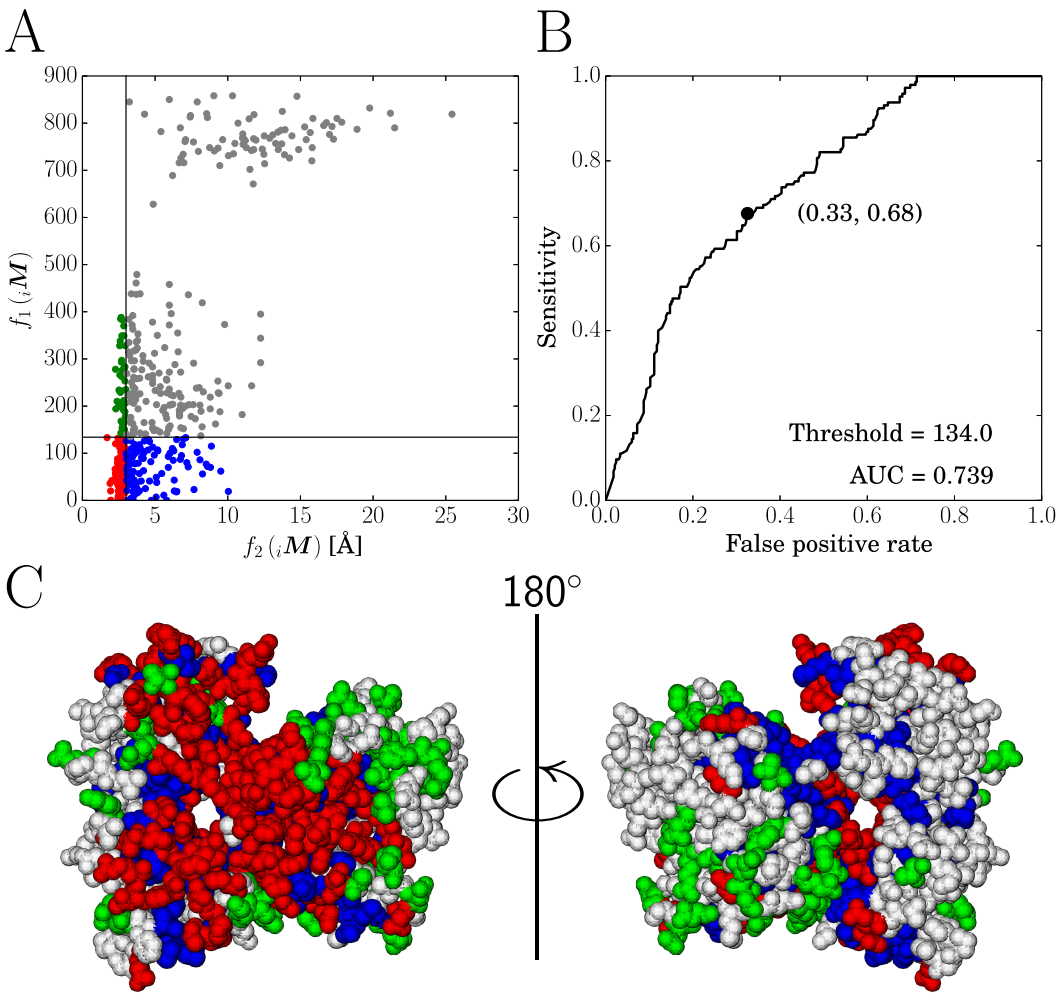

図 3 散布図，ROC 曲線，立体構造．(A) $f_1\,({}_iM)$ と $f_2\,({}_iM)$ の散布図．ただし，$f_1\,({}_{468}M) = 822.0$ と $f_2\,({}_{468}M) = 45.56$ は除いた．$f_2\,({}_iM)$ が 3 Å 以下であるかそれよりも大きいかで ${}_iM \in M$ を二つに分けた．(B) 前者を真，後者を偽として，ROC 曲線を $f_1\,({}_iM)$ から作成した．閾値は (感度 + 1 - 偽陽性率) が最大となるように決め，最終的に ${}_iM \in M$ を 4 つに分けた．(C) それらの分類を *Thermus thermophilus* EF-Tu [7] の立体構造上に表示した．

5. 考察

$f_1\,({}_iM)$，$f_2\,({}_iM)$，AUC，スピアマンの ρ の意味は以下のとおりである．$f_1\,({}_iM)$ が小さくなるのは文字が進化系統樹の根の近くでのみ分かれるときである．$f_1\,({}_iM)$ が大きくなるのは文字が進化系統樹の根の遠くで分かれるときである．$f_2\,({}_iM)$ が小さくなるのは少なくとも一つの ${}_iM$ の中のアミノ酸残基がイオンや分子の近くであるときである．$f_2\,({}_iM)$ が大きくなるのは全ての共結晶構造で ${}_iM$ の中のアミノ酸残基がイオンや分子の近くでないときである．AUC が 0.5 のとき，$f_1\,({}_iM)$ と c_2 のカットオフのもとでの $f_2\,({}_iM)$ の近くであるか近くでないかに相関がない可能性がある．AUC が 1 のとき，小さい $f_1\,({}_iM)$ と大きい $f_1\,({}_iM)$ に近くであることと近くでないことにそれぞれ相関がある．AUC が 0 のとき，大きい $f_1\,({}_iM)$ と小さい $f_1\,({}_iM)$ に近くであることと近くでないことにそれぞれ相関がある．スピアマンの ρ が 0 のとき，$f_1\,({}_iM)$ と $f_2\,({}_iM)$ に線形の相関がない可能性がある．AUC が 1 か-1 のとき，$f_1\,({}_iM)$ と $f_2\,({}_iM)$ にそれぞれ正か負の線形の相関がある．

EF-Tu/EF-1A タンパクはタンパク質の生合成に関わり [7]，本研究ではその機能に関わるイオンや分子を選択した．したがって，$f_2\,({}_iM)$ が小さいとき，${}_iM$ の中のあるアミノ酸残基がタンパク質の生合成に関わる領域の近くであることを示し，$f_2\,({}_iM)$ が大きいとき，${}_iM$ の中の全てのアミノ酸残基がタンパク質の生合成に関わる領域の近くではないことを示す．図 3A と 3C はその領域に近いか近くないかを示し，図 3B の $f_1\,({}_iM)$ の ROC 曲線では AUC が 0.739 となっている．これは，近くの領域は $f_1\,({}_iM)$ が小さく，近くない領域は $f_1\,({}_iM)$ が大きい傾向にあることを示している．さらに，スピアマンの ρ は 0.580 であ

り，$f_1({}_i\boldsymbol{M})$ が小さければ，$f_2({}_i\boldsymbol{M})$ が小さい傾向にあり，$f_2({}_i\boldsymbol{M})$ が大きければ，$f_2({}_i\boldsymbol{M})$ が大きい傾向にあることを示している．したがって，$f_1({}_i\boldsymbol{M})$ によってあるアミノ酸残基がイオンや分子の近くであるかないかを予測できる．ただし，完全な線形の相関は得られていないため，$f_1({}_i\boldsymbol{M})$ によって $f_2({}_i\boldsymbol{M})$ の全てを説明できるわけではない．このため，$f_1({}_i\boldsymbol{M})$ と $f_2({}_i\boldsymbol{M})$ によって測れるものは，互いに類似ではあるが，全く同様ではないと推測できる．また，$f_1({}_i\boldsymbol{M})$ か $f_2({}_i\boldsymbol{M})$ かによって測れるものの種類は異なり，それはイオンや分子と結合するための重要性あるいは立体構造を維持するための重要性であると考えることもできる．そのため，EF-Tu/EF-1A タンパクの重要な残基は近くの領域のみでなく近くない領域にも配置されている可能性もあると予測できる．

6. 結論

多重配列アライメントが文字型と座標型で表されるとしたときの写像の方法について述べた．本手法により，二種類の写像に基づいてアミノ酸残基を評価することができる．

参考文献

[1] Williamson, R. M.: Information theory analysis of the relationship between primary sequence structure and ligand recognition among a class of facilitated transporters, *J Theor Biol*, Vol. 174, No. 2, pp. 179–188 (1995).
[2] Lichtarge, O., Bourne, H. and Cohen, F.: An evolutionary trace method defines binding surfaces common to protein families, *J Mol Biol*, Vol. 257, No. 2, pp. 342–358 (1996).
[3] Landgraf, R., Fischer, D. and Eisenberg, D.: Analysis of heregulin symmetry by weighted evolutionary tracing, *Protein Eng*, Vol. 12, No. 11, pp. 943–951 (1999).
[4] Valdar, W.: Scoring residue conservation, *Proteins*, Vol. 48, No. 2, pp. 227–241 (2002).
[5] Mihalek, I., Res, I. and Lichtarge, O.: A family of evolution-entropy hybrid methods for ranking protein residues by importance, *J Mol Biol*, Vol. 336, No. 5, pp. 1265–1282 (2004).
[6] Ashkenazy, H., Erez, E., Martz, E., Pupko, T. and Ben-Tal, N.: ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids, *Nucleic Acids Res*, Vol. 38, No. 2, pp. W529–W533 (2010).
[7] Parmeggiani, A., Krab, I., Watanabe, T., Nielsen, R., Dahlberg, C., Nyborg, J. and Nissen, P.: Enacyloxin IIa pinpoints a binding pocket of elongation factor Tu for development of novel antibiotics, *J Biol Chem*, Vol. 281, No. 5, pp. 2893–2900 (2006).
[8] Consortium, T. U.: UniProt: a hub for protein information, *Nucleic Acids Res*, Vol. 43, No. D1, pp. D204–D212 (2015).
[9] Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. and Bourne, P.: The protein data bank, *Nucleic Acids Res*, Vol. 28, No. 1, pp. 235–242 (2000).
[10] Dai, S., Crawford, F., Marrack, P. and Kappler, J. W.: The structure of HLA-DR52c: comparison to other HLA-DRB3 alleles, *Proc Natl Acad Sci U S A*, Vol. 105, No. 33, pp. 11893–11897 (2008).
[11] Kidmose, R. T., Vasiliev, N. N., Chetverin, A. B., Andersen, G. R. and Knudsen, C. R.: Structure of the Q$\boldsymbol{\beta}$ replicase, an RNA-dependent RNA polymerase consisting of viral and host proteins, *Proc Natl Acad Sci U S A*, Vol. 107, No. 24, pp. 10884–10889 (2010).
[12] Takeshita, D. and Tomita, K.: Assembly of Q$\boldsymbol{\beta}$ viral RNA polymerase with host translational elongation factors EF-Tu and -Ts, *Proc Natl Acad Sci U S A*, Vol. 107, No. 36, pp. 15733–15738 (2010).
[13] Takeshita, D. and Tomita, K.: Molecular basis for RNA polymerization by Q$\boldsymbol{\beta}$ replicase, *Nat Struct Mol Biol*, Vol. 19, No. 2, pp. 229–237 (2012).
[14] Takeshita, D., Yamashita, S. and Tomita, K.: Mechanism for template-independent terminal adenylation activity of Q$\boldsymbol{\beta}$ replicase, *Structure*, Vol. 20, No. 10, pp. 1661–1669 (2012).
[15] Katoh, K. and Standley, D. M.: MAFFT multiple sequence alignment software version 7: improvements in performance and usability, *Mol Biol Evol*, Vol. 30, No. 4, pp. 772–780 (2013).
[16] Kishino, H., Miyata, T. and Hasegawa, M.: Maximum-likelihood inference of protein phylogeny and the origin of chloroplasts, *J Mol Evol*, Vol. 31, No. 2, pp. 151–160 (1990).
[17] Jones, D., Taylor, W. and Thornton, J.: The rapid generation of mutation data matrices from protein sequences, *Comput Appl Biosci*, Vol. 8, No. 3, pp. 275–282 (1992).
[18] Dayhoff, M. O. and Schwartz, R. M.: Chapter 22: A model of evolutionary change in proteins, *in Atlas of Protein Sequence and Structure* (1978).
[19] Sneath, P. H. A. and Sokal, R. R.: *Numerical taxonomy: the principles and practice of numerical classification*, W. H. Freeman, San Francisco (1973).
[20] Heffron, S. E., Mui, S., Aorora, A., Abel, K., Bergmann, E. and Jurnak, F.: Molecular complementarity between tetracycline and the GTPase active site of elongation factor Tu, *Acta Crystallogr D Biol Crystallogr*, Vol. 62, pp. 1392–1400 (2006).
[21] Nissen, P., Thirup, S., Kjeldgaard, M. and Nyborg, J.: The crystal structure of Cys-tRNA$^{\mathrm{Cys}}$-EF-Tu-GDPNP reveals general and specific features in the ternary complex and in tRNA, *Structure*, Vol. 7, No. 2, pp. 143–156 (1999).
[22] LaMarche, M. J., Leeds, J. A., Amaral, K., Brewer, J. T., Bushell, S. M., Dewhurst, J. M., Dzink-Fox, J., Gangl, E., Goldovitz, J., Jain, A., Mullin, S., Neckermann, G., Osborne, C., Palestrant, D., Patane, M. A., Rann, E. M., Sachdeva, M., Shao, J., Tiamfbok, S., Whitehead, L. and Yu, D.: Antibacterial optimization of 4-aminothiazolyl analogues of the natural product GE2270 A: identification of the cycloalkylcarboxylic acids, *J Med Chem*, Vol. 54, No. 23, pp. 8099–8109 (2011).
[23] Kobayashi, K., Saito, K., Ishitani, R., Ito, K. and Nureki, O.: Structural basis for translation termination by archaeal RF1 and GTP-bound EF1$\boldsymbol{\alpha}$ complex, *Nucleic Acids Res*, Vol. 40, No. 18, pp. 9319–9328 (2012).
[24] Groftehauge, M. K., Therkelsen, M. O., Taaning, R., Skrydstrup, T., Morth, J. P. and Nissen, P.: Identifying ligand-binding hot spots in proteins using brominated fragments, *Acta Crystallogr F Struct Biol Commun*, Vol. 69, No. 9, pp. 1060–1065 (2013).
[25] Vogeley, L., Palm, G., Mesters, J. and Hilgenfeld, R.:

Conformational change of elongation factor Tu (EF-Tu) induced by antibiotic binding - crystal structure of the complex between EF-Tu·GDP and aurodox, *J Biol Chem*, Vol. 276, No. 20, pp. 17149–17155 (2001).

[26] Spearman, C.: The Proof and Measurement of Association Between Two Things, *Am J Psychol*, Vol. 15, pp. 88–103 (1904).

[27] Hunter, J. D.: Matplotlib: a 2D graphics environment, *Computing Sci Eng*, Vol. 9, No. 3, pp. 90–95 (2007).

[28] Humphrey, W., Dalke, A. and Schulten, K.: VMD: visual molecular dynamics, *J Mol Graph Model*, Vol. 14, No. 1, pp. 33–38 (1996).

[29] Lasko, T., Bhagwat, J., Zou, K. and Ohno-Machado, L.: The use of receiver operating characteristic curves in biomedical informatics, *J Biomed Inform*, Vol. 38, No. 5, pp. 404–415 (2005).