情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2010-BIO-21 No.18
2010/6/19

計量学習による酵素活性部位予測

加藤 毅[†1] 諏訪和大[†2] 長野希美[†3]

活性部位の局所構造をもとに機能的に類似した酵素を探索する問題は構造生物学において重要な問題である．通常，類似部位を探索するために局所構造のテンプレートを用いるが，その予測性能はテンプレートに含める原子の選択に強く依存する．本論文では，計量学習によって原子の自動選択を行う算法を提案する．実験を通して提案法の有効性を示す．

Using Metric Learning Algorithms for Enzyme Active Site Prediction

Tsuyoshi Kato,[†1] Kazuhiro Suwa[†2] and Nozomi Nagano[†3]

It is an important problem to find functionally analogous enzymes based on the local structures of active sites. Conventional methods use templates of the local structures to search for analogous sites, but their performances depend on the selection of atoms to be included in the templates. We propose a new metric-learning-based algorithm that allows for the automatic selection of atoms. We show the usefulness of the proposed algorithm through experiments.

1. はじめに

新しくシーケンスされたゲノムの流入は，機能既知の遺伝子やタンパク質との大域的な配列比較や立体構造比較に基づく機能予測法の開発を加速させた[16]．酵素に関しては，多くの方法は EC 番号に基づくタンパク質の配列や立体構造をもとにした機能予測である[16]．

長年にわたって，EC 番号に基づく分類が使われてきた．EC 番号の分類は，基質および生成物の化学構造，また使われる補因子に基づく分類である[21]．しかし，EC 番号の分類は，タンパク質の配列や立体構造に関する情報を無視しているので，配列や立体構造と，機能の間に，関係性を見つけるのが困難な場合がある．相同な酵素は同じ祖先となる酵素から分岐進化されて同じ機能を持つ場合が多いが，異なるスーパーファミリーに属する非相同な酵素が収斂進化によって同じ酵素の機能を持つこともある．

酵素 trypsin および subtilisin は Ser-His-Asp 触媒残基を共有している．この 2 つの酵素は収斂進化から得られる**類似酵素**[22] の典型例である．Nagano は 131 スーパーファミリに対して 270 酵素の触媒機構を解析し，主に手動で酵素反応データベース EzCatDB[17] に登録した．この酵素反応の解析により，複数の**類似反応**が非相同酵素にみられることが明らかになった．EzCatDB では **RLCP 分類** という，酵素反応の階層的な分類を与える．これは EC 番号による分類に替わるものである．RLCP 分類では，反応の種類で酵素を分類している．基質の反応部位，触媒機構，および酵素の触媒部位が同一であると同一の反応と言える．RLCP 分類では，触媒機構が等しく，触媒部位が同じタイプでありさえすれば同じ反応クラスとなる[17]．EC 番号の分類では，触媒機構が等しく，触媒部位が同じタイプであっても，異なるクラスに分類をされることが多くあったが，RLCP 分類はこの問題点を解消したものになっている．

Gherardini ら[8] は類似酵素が活性部位を共有することは稀ではないとの報告をしている．これは，酵素機能の予測には，ドメインレベルや鎖レベルの大域的構造よりも酵素反応を反映している活性部位の局所構造に注目すべきであることを示唆している[16]．似たような活性部位を検出する局所構造比較法として主流な方法は**テンプレート法**である[2),5),7),11),14),15),18)–20)]．テンプレート法は，あらかじめターゲットとなるタンパク質立体構造の活性部位に含まれる原子を含めたテンプレートを作成しておき，機能未知のタンパク質立体構造に対してそのテンプレートと類似の局所構造を探索するものである．しかし，既存のテンプレート法は，次の問題点がある：(i) 予測精度は，テンプレートに含まれる原子数や原子の種類に依存する．どの原子をテンプレートに含めるべきか決定するのは非常に難しい場合がある．構造と機能の専門家が試行錯誤によってどの原子の組み合わせをテンプレートに含めるか決定しなければならない．(ii) 触媒部位にある原子は他の原子よりも触媒反応には重要である．これまでの報告[3] によると，触媒残基の側鎖は 92% の頻度で使われ，主鎖はたったの 8% である．

†1 東京大学大学院新領域創成科学研究科
Graduate School of Frontier Sciences, University of Tokyo
†2 ドットランプ
D. Trump
†3 産総研生命情報工学研究センター
AIST Computational Biology Research Center

Vol.2010-BIO-21 No.18
2010/6/19

さらに，電荷のある，もしくは，イオン化した残基は触媒に寄与しやすい[3]．しかし，これらの触媒に重要な原子は，機能予測には重要なのかは明らかではない．(iii) テンプレート法は，**部位マッチ**(同じ反応クラスとのマッチ) のほかに，非常に多くの**ミスマッチ**(異なる反応クラスとのマッチ，もしくは触媒部位以外の部位とのマッチ) をヒットしてしまう．ミスマッチの個数は減らすことは可能なのか？

本研究では，計量学習の考え方をテンプレート法に導入して触媒部位を効果的に検出できるテンプレートを生成し，探索精度を向上させる算法を提案する．酵素反応データベース EzCatDB にある触媒部位をもとにラフに作成したテンプレートを，計量学習によって原子を重みづけする．この重みづけは，探索精度を向上させるのみならず，どの原子が予測に重要かという情報を提供する．

記　　法

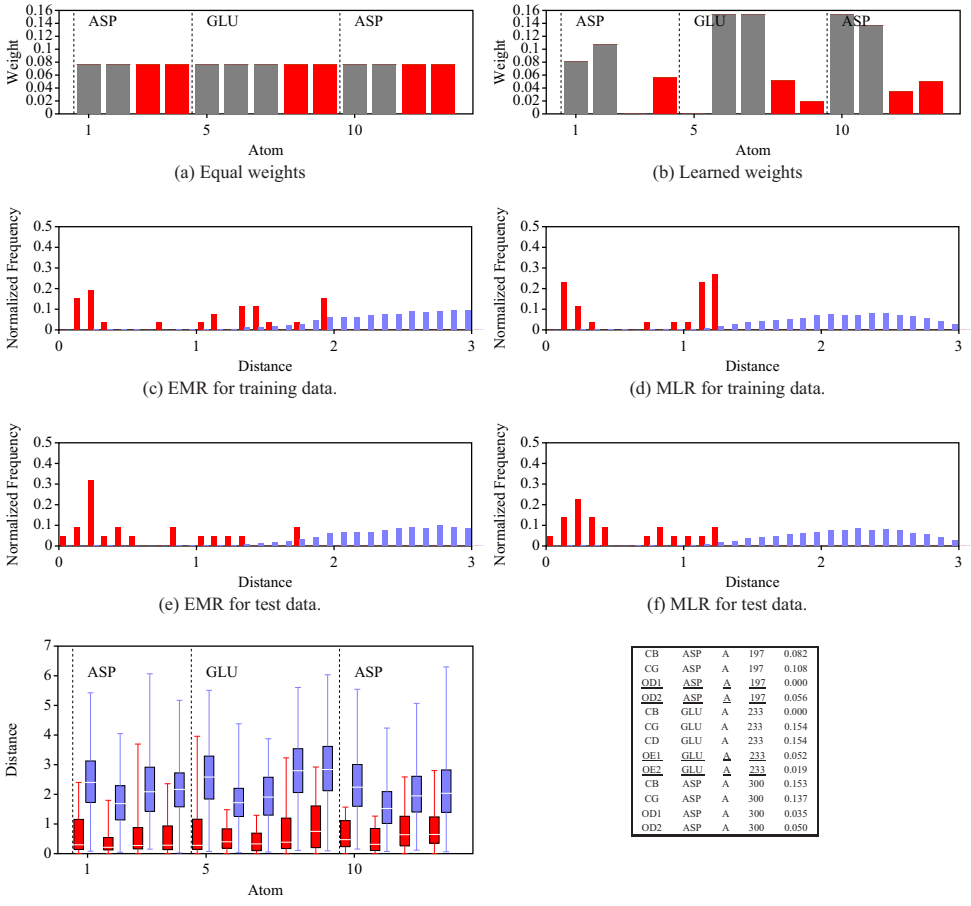
$\boldsymbol{\Delta}^n$ は n 次元空間における確率単体を表す：$\boldsymbol{\Delta}^n \equiv \left\{\boldsymbol{x} \in \mathbb{R}_+^n \mid \sum_{i=1}^n x_i = 1\right\}$．$\mathbb{O}^n$ は，$n \times n$ の回転行列を表す：$\mathbb{O}^n \equiv \{\boldsymbol{A} \in \mathbb{R}^{n\times n} \mid \boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}_n, \det(\boldsymbol{A}) = 1\}$．$\mathbb{N}_n$ は n 以下の自然数の集合である：$\mathbb{N}_n \equiv \{i \in \mathbb{N} \mid i \le n\}$．

2. 原　　理

2.1 問題設定

TESS[20] のようなテンプレート法は，機能未知のタンパク質立体構造からテンプレートと似た局所構造を探索する局所構造探索算法 (**LSS 算法**) である．まず，テンプレートはクエリとなる酵素の活性部位から原子を注意深く選択する．ここでは，選択された原子の集合を**クエリテンプレート**と呼ぶ．n 個の原子を含むクエリテンプレートを LSS 算法に入力すると，Protein Data Bank (PDB) のような立体構造データベースから似た形の部位を持つタンパク質とその部位を探索する．LSS 算法の出力は，表 2 に示すようなヒットした部位の集合である．ℓ はヒットした部位の個数である．従来の LSS 算法の使用方法は，ヒットした各部位に対して，**RMSD**(root mean square deviation) を計算し，その残差が閾値以下ならば部位マッチ，さもなければミスマッチと判別する．

本研究では，よりよい予測を得るために原子それぞれに重みづけすることを提案する．従来法はヒットとクエリがどのくらい似ているか測るために RMSD を用いる．RMSD について数式を用いて定義する．行列 $\boldsymbol{X}^{\text{query}}$ および $\boldsymbol{X}'$ でそれぞれクエリテンプレートとヒットを表すとする．クエリテンプレートに n 原子が含まれているとし，行列 $\boldsymbol{X}^{\text{query}}$ には各

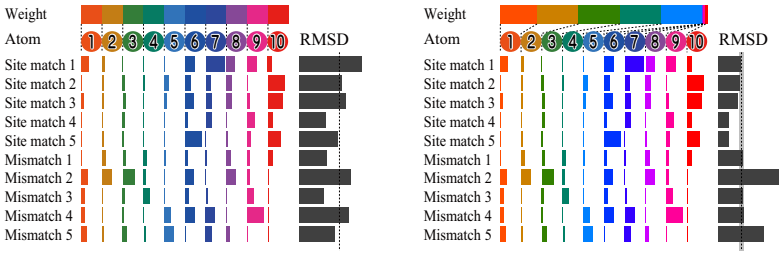


CB	ASP	A	197	0.082
CG	ASP	A	197	0.108
OD1	ASP	A	197	0.000
OD2	ASP	A	197	0.056
CB	GLU	A	233	0.000
CG	GLU	A	233	0.154
CD	GLU	A	233	0.154
OE1	GLU	A	233	0.052
OE2	GLU	A	233	0.019
CB	ASP	A	300	0.153
CG	ASP	A	300	0.137
OD1	ASP	A	300	0.035
OD2	ASP	A	300	0.050

図 1　クエリテンプレート 1jfh の結果．(a),(b) はテンプレートに含まれる原子の重みを示し，(a) は重みなし，(b) は計量学習によって得られた重みである．このテンプレートは 13 原子含んでいる．炭素は灰色，酸素は赤色で示している．(c),(d),(e),(f) は重みなし RMSD と重みつき RMSD の分布をプロットしている．赤が部位マッチ，青がミスマッチの分布である．(c) は，訓練用データに対する重みなし RMSD，(d) は，訓練用データに対する重みつき RMSD，(e) は，評価用データに対する重みなし RMSD，(f) は，評価用データに対する重みつき RMSD である．(g) は各原子に対する距離の分布である．赤が部位マッチ，青がミスマッチの分布である．(h) はテンプレートに含まれる原子のリストである．

Fig. 1　Results of query template 1jfh.

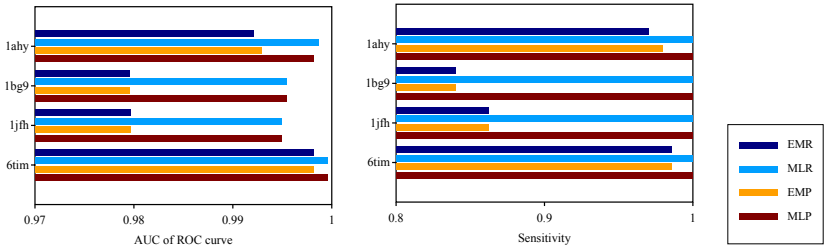
Vol.2010-BIO-21 No.18
2010/6/19



(a) 重みなし RMSD　　(b) 重みつき RMSD

図 2 計量学習の例．従来，部位マッチとミスマッチを分けるには重みなし RMSD を計算してきた．すなわち，各原子に対して距離を計算してその平均をとるのである．この例では，5 個の部位マッチと 5 個のミスマッチがある．重みなし RMSD は，(a) に示す閾値では，3 個のミスマッチと 2 個の部位マッチが誤って検出される．また，重みなし RMSD では，閾値をどのように動かしたとしても部位マッチとミスマッチとを分離することはできない．計量学習算法は部位マッチとミスマッチをできる限り分離するような重みつき RMSD を生成する計量を見つける．この例では，(b) に示すように，重みつき RMSD は部位マッチとミスマッチを完全に分離している．

Fig. 2　Example of metric learning.



(a)　　(b)

図 3 4 つのクエリテンプレート 1ahy, 1bg9, 1jfh, 6tim の結果．(a),(b) はそれぞれ AUC と感度をプロットしている．各テンプレートに対して，EMR, MLR, EMP, MLP を比較している．

Fig. 3　Results of four templates, 1ahy, 1bg9, 1jfh, and 6tim.

表 1 LSS 算法を使って生成されたデータセット．この表は LSS 算法を使って各クエリテンプレートに検出されたデータ数を示している．列 #mc/#in は，主鎖にある内部原子，内部原子の合計を与える．列 #mc/#out は，主鎖にある外部原子，外部原子の合計を与える．“#mtch” は部位マッチの個数，“#mis” はミスマッチの個数である．

Table 1　Datasets generated using the LSS algorithm.

Template	#mc/#in	#mc/#out	#mtch	#mis	Reaction type
1zio	0 / 26	4 / 15	11	22	adenylate kinase-type
1arg	4 / 21	0 / 0	58	7667	aminotransferase-type i1
1cq7	4 / 21	0 / 0	58	7518	aminotransferase-type i2
1ahy	0 / 21	4 / 4	46	2380	aminotransferase-type ese
1arg-2	0 / 21	4 / 4	43	2497	aminotransferase-type ese
1map	0 / 13	4 / 4	88	26944	aminotransferase-type sn
1ams	0 / 13	4 / 4	40	29104	aminotransferase-type sn
1ahg	0 / 21	4 / 4	35	2402	aminotransferase-type esi
1kcd	0 / 15	0 / 5	22	1360	polygalacturonase-type
2bvw	0 / 8	0 / 0	21	72738	lysozyme-type
1qk2	0 / 8	0 / 0	16	77283	lyzyme-type
1bg9	0 / 13	0 / 0	49	12967	α-amylase-type
1jfh	0 / 13	0 / 0	48	12977	α-amylase-type
1isw	0 / 18	0 / 2	27	863	xylanase A-type
1ka1	6 / 28	22 / 28	15	30	inositol-phosphatase-type
2oke	6 / 21	6 / 9	10	126	dUTP pyrophosphatase-type
1eo4	8 / 21	0 / 0	25	7162	restriction enzyme-type
1kfs	0 / 23	16 / 18	11	33	3’-5’ exonuclease-type
1rpa	0 / 31	0 / 6	14	8	acid phosphatase-type
1vcz	0 / 24	0 / 4	29	13	RNase-type
2dhc	4 / 36	4 / 6	13	15	dehalogenase-type 1
1g42	4 / 31	4 / 6	14	23	dehalogenase-type 2
1acb	4 / 16	4 / 4	504	4207	trypsin-type
1bls	4 / 18	4 / 10	10	143	cephalosporinase-type
2ace	6 / 18	6 / 7	43	528	cholinesterase-type
1af0	0 / 29	16 / 18	19	39	serralysin-type
3cpa	2 / 26	14 / 19	39	6	carboxypeptidase-type
1psa	0 / 8	0 / 0	333	62502	pepsin-type
6tim	0 / 16	0 / 0	63	4616	TIM-type i1
4tim	0 / 15	0 / 0	64	6309	TIM-type i2

原子の 3 次元座標が格納されている：

$$\boldsymbol{X}^{\text{query}} = [\boldsymbol{x}_1^{\text{query}}, \ldots, \boldsymbol{x}_n^{\text{query}}] \in \mathbb{R}^{3\times n}$$

ただし $\boldsymbol{x}_j^{\text{query}} \in \mathbb{R}^3$ は，クエリテンプレートにおける第 j 原子の 3 次元座標である．同様

Vol.2010-BIO-21 No.18
2010/6/19

表 2 計量学習算法の入力.
Table 2 Input of the metric learning algorithm.

	Atom1	Atom2	...	Atom$n$	Class
Site 1	$\boldsymbol{x}_{1,1}$	$\boldsymbol{x}_{1,2}$	$\cdots$	$\boldsymbol{x}_{1,n}$	y_1
Site 2	$\boldsymbol{x}_{2,1}$	$\boldsymbol{x}_{2,2}$	$\cdots$	$\boldsymbol{x}_{2,n}$	y_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Site ℓ	$\boldsymbol{x}_{\ell,1}$	$\boldsymbol{x}_{\ell,2}$	$\cdots$	$\boldsymbol{x}_{\ell,n}$	y_ℓ

に，$\boldsymbol{X}'$ は 3 次元座標の順序つき集合であり，

$$\boldsymbol{X}' = \left[\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_n\right] \in \mathbb{R}^{3\times n}$$

と定義される．ただし，$\boldsymbol{x}'_j$ はヒットにおける第 j 原子の 3 次元座標である．平均二乗残差はあらゆる剛体変換をかけた中で平均残差が最小の値で定義される[13]．数式で書くと

$$\min_{\boldsymbol{R}\in\mathbb{O}^3,\,\boldsymbol{v}\in\mathbb{R}^3} E_{\mathrm{unwei}}(\boldsymbol{X}^{\mathrm{query}}, \boldsymbol{X}'; \boldsymbol{R}, \boldsymbol{v})$$

のようになる．ただし，$E_{\mathrm{unwei}}(\boldsymbol{X}^{\mathrm{query}}, \boldsymbol{X}'; \boldsymbol{R}, \boldsymbol{v})$ はクエリの原子と剛体変換 (回転 $\boldsymbol{R} \in \mathbb{O}^3$，平行移動 $\boldsymbol{v} \in \mathbb{R}^3$) 後の対応する原子との距離の平均である：

$$E_{\mathrm{unwei}}(\boldsymbol{X}^{\mathrm{query}}, \boldsymbol{X}'; \boldsymbol{R}, \boldsymbol{v}) = \frac{1}{n}\sum_{j=1}^{n} \left\|\boldsymbol{x}_j^{\mathrm{query}} - \left(\boldsymbol{R}\boldsymbol{x}'_j + \boldsymbol{v}\right)\right\|^2.$$

RMSD は平均二乗残差に平方根をとったものであり，よく用いられている[13]．この関数は原子に重みづけせずに距離の平均をとっている．本研究の提案は重みつきの距離を使用することである．重みベクトルを $\boldsymbol{w} \in \boldsymbol{\Delta}^n$ であらわすとすると，**重みつき平均二乗残差**は次のようにあらわされる：

$$\min_{\boldsymbol{R}\in\mathbb{O}^3,\,\boldsymbol{v}\in\mathbb{R}^3} E(\boldsymbol{X}^{\mathrm{query}}, \boldsymbol{X}'; \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{w})$$

ただし

$$E(\boldsymbol{X}^{\mathrm{query}}, \boldsymbol{X}'; \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{w}) = \sum_{j=1}^{n} w_j \left\|\boldsymbol{x}_j^{\mathrm{query}} - \left(\boldsymbol{R}\boldsymbol{x}'_j + \boldsymbol{v}\right)\right\|^2.$$

である．重みつき平均二乗残差の重みを

$$\forall j \in \mathbb{N}_n : \qquad w_j = \frac{1}{n}.$$

とおくと，重みなし平均二乗残差になる．

原子の重みづけは，n 原子の座標集合の空間において，パラメトリックな計量[1] を調整しているとみることができる⋆1．その空間の次元数は $3n$ である．なぜなら，n 個の原子がそれぞれ 3 次元座標を持っているからである．重みづけをしない場合は，$3n$ 次元の空間でユークリッド計量を使っていることになる．予測精度は，重みベクトル $\boldsymbol{w}$ によって定められる計量に依存する．実際に，クエリテンプレートに含まれる原子のうち，予測に有効な原子もあれば，そうではない原子もある．クエリテンプレートのいくつかの原子は酵素反応の性質を保持するために位置が保存されている．これを幾何学的視点からみると，$3n$ 次元空間におけるいくつかの次元は予測に有効で，いくつかは予測に有効ではないことになる．これから述べる算法は，$3n$ 次元空間における計量を調整することにより，いくつかの原子を強調し，いくつかの原子を排除し，これによって予測性能を向上させる．

4 節において，LSS 算法によって得られるヒットを部位マッチかミスマッチか予測するには，学習によって計量を決める提案法のほうがよい性能を得られることを示す．計量を調整するにしても，ユークリッド計量を使うにしても，距離が閾値より小さければ部位マッチと予測し，さもなければミスマッチと予測する．計量の重みパラメータ $\boldsymbol{w}$ の値を決定するには，ヒットのうちすでに機能が判明しているものを使って計量学習を行う．計量学習の入力は表 2 に示すようなデータとなる．機能既知のヒット数は ℓ である．ベクトル $\boldsymbol{x}_{i,j} \in \mathbb{R}^3$ は第 i ヒットの第 j 原子の 3 次元座標を表す．変数 $y_i \in \{\pm 1\}$ は第 i ヒットのクラスラベルである：その値は部位マッチなら $+1$ であり，ミスマッチなら -1 である．

部位マッチ，およびミスマッチの添え字集合は，それぞれ $\mathcal{I}_+$，および $\mathcal{I}_-$ であらわす：

$$\mathcal{I}_+ \equiv \{i \in \mathbb{N}_\ell \,|\, y_i = +1\}, \qquad \mathcal{I}_- \equiv \{i \in \mathbb{N}_\ell \,|\, y_i = -1\}.$$

第 i ヒットは $3 \times n$ の行列

$$\boldsymbol{X}^{(i)} \equiv [\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in}] \in \mathbb{R}^{3\times n}$$

で表す．これは，表 2 の第 i 行に対応する．

次に，原子に重みづけを行う計量学習算法を示す．

2.2 計量学習算法

部位マッチとミスマッチがある閾値で完全に分離できるような距離が理想的である．その場合，すべての部位マッチの距離がある閾値 $\theta \in \mathbb{R}_+$ 未満で, i.e.

$$\forall i \in \mathcal{I}_+ : \qquad \min_{\boldsymbol{R}_i\in\mathbb{O}^3,\,\boldsymbol{v}_i\in\mathbb{R}^3} E(\boldsymbol{X}^{\mathrm{query}}, \boldsymbol{X}^{(i)}; \boldsymbol{R}_i, \boldsymbol{v}_i, \boldsymbol{w}) < \theta, \tag{1}$$

すべてのミスマッチの距離が θ より大きい, i.e.

$$\forall i \in \mathcal{I}_- : \qquad \min_{\boldsymbol{R}_i\in\mathbb{O}^3,\,\boldsymbol{v}_i\in\mathbb{R}^3} E(\boldsymbol{X}^{\mathrm{query}}, \boldsymbol{X}^{(i)}; \boldsymbol{R}_i, \boldsymbol{v}_i, \boldsymbol{w}) > \theta, \tag{2}$$

ことになる．図 2 は，重みなし RMSD と重みつき RMSD との差をあらわす例を示している．図では，5 個の部位マッチと 5 個のミスマッチを含んでいる．図 2(a) のように重みな

⋆1 正確に書くならば，非退化性は必ずしも成立しないので，擬距離空間となる．

Vol.2010-BIO-21 No.18
2010/6/19

し RMSD では部位マッチとミスマッチを分離できない場合でも，図 2(b) に示すように重みを調整することによって分離できることもある．

しかし，実際のデータの中には，どのように重みづけしても部位マッチとミスマッチを完全に分離できない場合もある．図 1 にクエリテンプレート 1jfh を使った場合の結果を示す．図 1(c) は部位マッチの重みなし RMSD の分布とミスマッチの重みなし RMSD の分布をプロットしている．図 1(d) は重みつき RMSD の分布をプロットしている．このデータセットでは，重みつき RMSD を使っても部位マッチとミスマッチを完全に分離することはできない．以上の理由による，(1) および (2) に与えた重みの条件は実際に使用する上では厳しすぎる．条件を緩和するために，各ヒットの不等式をある程度違反してもよいことにする．違反した量を定量化するために非負変数 ξ_i を導入し，不等式を次のように変更する：

$$\forall i \in \mathcal{I}_+ : \quad \min_{\boldsymbol{R}_i \in \mathbb{O}^3,\, \boldsymbol{v}_i \in \mathbb{R}^3} E(\boldsymbol{X}^{\mathrm{query}}, \boldsymbol{X}^{(i)}; \boldsymbol{R}_i, \boldsymbol{v}_i, \boldsymbol{w}) \le \theta + \xi_i,$$
$$\forall i \in \mathcal{I}_- : \quad \min_{\boldsymbol{R}_i \in \mathbb{O}^3,\, \boldsymbol{v}_i \in \mathbb{R}^3} E(\boldsymbol{X}^{\mathrm{query}}, \boldsymbol{X}^{(i)}; \boldsymbol{R}_i, \boldsymbol{v}_i, \boldsymbol{w}) \ge \theta - \xi_i.$$

これは次のようにまとめられる：

$$\forall i \in \mathbb{N}_\ell : \quad y_i \left(\min_{\boldsymbol{R}_i \in \mathbb{O}^3,\, \boldsymbol{v}_i \in \mathbb{R}^3} E(\boldsymbol{X}^{\mathrm{query}}, \boldsymbol{X}^{(i)}; \boldsymbol{R}_i, \boldsymbol{v}_i, \boldsymbol{w}) - \theta \right) \le \xi_i.$$

この条件のもとで，部位マッチに対する違反量の平均

$$\frac{1}{|\mathcal{I}_+|} \sum_{i \in \mathcal{I}_+} \xi_i$$

およびミスマッチに対する違反量の平均

$$\frac{1}{|\mathcal{I}_-|} \sum_{i \in \mathcal{I}_-} \xi_i.$$

の和で，**総和違反量**を測る．総和違反量を最小にするような計量を見つけるには，次のような最適化問題を解くことになる：

$$\begin{aligned} \min \quad & \frac{1}{|\mathcal{I}_+|} \sum_{i \in \mathcal{I}_+} \xi_i + \frac{1}{|\mathcal{I}_-|} \sum_{i \in \mathcal{I}_-} \xi_i \\ \mathrm{wrt} \quad & \theta \in \mathbb{R}_+, \qquad \boldsymbol{\xi} \in \mathbb{R}_+^\ell, \qquad \boldsymbol{w} \in \boldsymbol{\Delta}^n, \\ \text{subj to} \quad & \forall i \in \mathbb{N}_\ell : \\ & y_i \left(\min_{\boldsymbol{R}_i \in \mathbb{O}^3,\, \boldsymbol{v}_i \in \mathbb{R}^3} E(\boldsymbol{X}^{\mathrm{query}}, \boldsymbol{X}^{(i)}; \boldsymbol{R}_i, \boldsymbol{v}_i, \boldsymbol{w}) - \theta \right) \le \xi_i. \end{aligned} \tag{3}$$

この問題はかなり複雑な非線形最適化問題である．なぜなら，剛体変換の値 $(\boldsymbol{R}, \boldsymbol{v})$ と重みベクトル $\boldsymbol{w}$ が相互に依存しているからである．この制約を簡単化するために，剛体変換を固定してこの最適化問題を解くことにする．剛体変換は一時的に与えた重み $\boldsymbol{w}_{\mathrm{temp}}$ に対して最適化して決める：

$$(\boldsymbol{R}_i, \boldsymbol{v}_i) = \mathrm{argmin}_{\boldsymbol{R} \in \mathbb{O}^3,\, \boldsymbol{v} \in \mathbb{R}^3} E(\boldsymbol{X}^i, \boldsymbol{X}^{\mathrm{query}}; \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{w}_{\mathrm{temp}}). \tag{4}$$

本研究では，$\boldsymbol{w}_{\mathrm{temp}} = \mathbf{1}_n/n$ を選んだ．さらに，過学習を防ぐために，重みベクトルの ℓ_∞-ノルムの上限 $C \in \mathbb{R}$ を導入する．すなわち，制約

$$\|\boldsymbol{w}\|_\infty \le C$$

を最適化問題に加える．本研究の実験では，C の値は $2/n$ とおいた．この上限は正則化の効果がある (e.g. 文献[9])．これらをまとめると，計量を学習するための提案する算法は

$$\begin{aligned} \min \quad & \frac{1}{|\mathcal{I}_+|} \sum_{i \in \mathcal{I}_+} \xi_i + \frac{1}{|\mathcal{I}_-|} \sum_{i \in \mathcal{I}_-} \xi_i \\ \mathrm{wrt} \quad & \theta \in \mathbb{R}_+, \qquad \boldsymbol{\xi} \in \mathbb{R}_+^\ell, \qquad \boldsymbol{w} \in \boldsymbol{\Delta}^n, \\ \text{subj to} \quad & \forall i \in \mathbb{N}_\ell : \quad y_i \left(E(\boldsymbol{X}^{\mathrm{query}}, \boldsymbol{X}^{(i)}; \boldsymbol{R}_i, \boldsymbol{v}_i, \boldsymbol{w}) - \theta \right) \le \xi_i \\ & \|\boldsymbol{w}\|_\infty \le C. \end{aligned} \tag{5}$$

で与えられる．実際の計算には次の定理を利用する．

Theorem 2.1. 最適化問題 (5) は線形計画法[10] に帰着できる．

線形計画法は凸計画[4] の一種で，これを解くための効率的なソルバーを利用できる[6]．

図 1(d) は，テンプレート 1jfh へのヒットに対して，計量学習算法によって得られた重みつき RMSD の分布を示している．このデータに対しては部位マッチとミスマッチを完全に分離できる重みは存在しないものの計量学習算法はおおよそのヒットに対して部位マッチとミスマッチを分離する計量を獲得することに成功している．

3. 実験条件

計量学習算法の有効性を示すために，活性部位を探索する実験を PDB データセットに対して行った．48 のタンパク質立体構造を選んで活性部位のテンプレートを作成した．表 1 に 30 個のテンプレートに対する結果を示す．残りの 18 個は既知の部位マッチが 10 個未満しかなく，信用できる性能評価を行えなかった．クエリテンプレートは活性部位の原子が含まれるように作られる．本研究では，クエリテンプレートを作るために，まず，酵素の立体構造において酵素反応に寄与している残基を選んだ．酵素データベース EzCatDB[17] において，活性部位のそれぞれのアミノ酸残基は触媒残基，補因子結合残基，修飾残基，主鎖触媒残基の 4 種類に分類されている．触媒部位残基および修飾残基に関しては側鎖にある原子をテンプレートに含めた．補因子結合残基に関しては，すべての原子をテンプレートに含めた．主鎖触媒残基は主鎖の原子のみテンプレートに含めた．このように，テンプ

Vol.2010-BIO-21 No.18
2010/6/19

レート作成者は残基レベルの選択しかしないので，このように作製したテンプレートは作成者の能力や知識にあまり依存しない．この方法で作成したテンプレートを **Rough テンプレート**と呼ぶことにする．一方，従来は酵素反応にかかわる原子一つ一つを注意深く選択してテンプレートを作成していた．この従来の方法で作ったテンプレートを **Precise テンプレート**と呼ぶことにする．本報告では，Rough テンプレートは計量学習と組み合わせることにより，Precise テンプレートよりも精度よく予測できることを示す．本研究で用いた Precise テンプレートの原子の集合は Rough テンプレートの原子の集合の部分集合になっている．Precise テンプレートに含まれる原子を**内部原子**と呼び，Rough テンプレートにしか含まれない原子を**外部原子**と呼ぶ．

まず，LSS 算法のひとつである TESS[20] を PDB データセットに適用して，各テンプレートと類似の活性部位の候補を探した．現在，酵素データベース EzCatDB には 5,692 個の PDB 立体構造に対する機能が登録されている．ヒットしたすべての局所部位のなかで，その部位が EzCatDB にテンプレートと同じ反応クラスに属しているものを部位マッチとし，EzCatDB に未登録，もしくは，異なる部位に属しているものをミスマッチとした．これらを算法の性能評価に用いた．部位マッチとミスマッチの個数は表 1 に示す．

計量学習算法の性能を評価するために，データセット中の半分のタンパク質を無作為に選んで学習用に使い，残りを評価用とした． 2 種類の評価基準を採用した：一つは，**AUC**，もう一つは**感度**(sensitivity) である．AUC は，あらゆる閾値で陽性率と陰性率をプロットして得られる ROC カーブの下の面積である．感度は，特異度 (specificity) が 0.95 になるように閾値を定めた時の陽性率とした．AUC や感度を計算する際には，部位マッチを正例，ミスマッチを負例として扱っていることに注意．この手続きを 100 回繰り返し，平均の AUC と感度を調査した．

Rough テンプレートは Precise テンプレートよりも作成者の能力に依存しないので，Rough テンプレートで十分な予測精度が得られることがもっとも望ましい．計量学習をする場合，しない場合，Rough テンプレートを使う場合，Precise テンプレートを使う場合をそれぞれ区別するため，次の 4 つの用語を導入する．

Condition 3.1 (ユークリッド計量 Rough テンプレート (EMR))．*Rough* テンプレートを用いて，重みなし *RMSD* で予測を行う． □

Condition 3.2 (計量学習 Rough テンプレート (MLR))．*Rough* テンプレートを用いて，重みつき *RMSD* で予測を行う． □

Condition 3.3 (ユークリッド計量 Precise テンプレート (EMP))．*Precise* テンプレートを用いて，重みなし *RMSD* で予測を行う． □

Condition 3.4 (計量学習 Precise テンプレート (MLP))．*Precise* テンプレートを用いて，重みつき *RMSD* で予測を行う． □

表 3 ROC カーブの AUC.
Table 3 AUC of ROC curves.

	EMR	MLR	EMP	MLP
1zio	0.929 (0.040)	0.938 (0.048)	**0.942** (0.036)	0.935 (0.049)
1arg	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)
1cq7	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)
1ahy	0.992 (0.002)	**0.999** (0.000)	0.993 (0.002)	0.998 (0.001)
1arg_2	0.996 (0.001)	**1.000** (0.000)	0.997 (0.001)	**1.000** (0.000)
1map	0.938 (0.010)	**0.983** (0.002)	0.903 (0.014)	0.935 (0.007)
1ams	0.846 (0.028)	**0.902** (0.019)	0.763 (0.024)	0.824 (0.014)
1ahg	0.999 (0.001)	**1.000** (0.000)	0.999 (0.001)	**1.000** (0.000)
1kcd	0.828 (0.034)	**0.973** (0.012)	0.674 (0.066)	0.924 (0.028)
2bvw	0.635 (0.021)	**0.685** (0.023)	0.635 (0.021)	**0.685** (0.023)
1qk2	0.884 (0.060)	**0.912** (0.058)	0.884 (0.060)	**0.912** (0.058)
1bg9	0.980 (0.007)	**0.995** (0.001)	0.980 (0.007)	**0.995** (0.001)
1jfh	0.980 (0.007)	**0.995** (0.001)	0.980 (0.007)	**0.995** (0.001)
1isw	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)
1ka1	0.973 (0.033)	0.973 (0.039)	**0.975** (0.030)	0.974 (0.040)
2oke	**1.000** (0.000)	**1.000** (0.001)	0.999 (0.001)	0.997 (0.004)
1eo4	0.998 (0.002)	**1.000** (0.000)	0.998 (0.002)	**1.000** (0.000)
1kfs	0.985 (0.018)	**0.987** (0.022)	0.985 (0.018)	0.981 (0.023)
1rpa	0.843 (0.246)	0.893 (0.210)	0.843 (0.246)	**0.901** (0.217)
1vcz	0.732 (0.061)	0.974 (0.061)	**1.000** (0.000)	0.998 (0.008)
2dhc	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)
1g42	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)
1acb	0.997 (0.001)	**0.999** (0.001)	0.993 (0.002)	0.996 (0.002)
1bls	**1.000** (0.000)	**1.000** (0.001)	**1.000** (0.000)	0.997 (0.010)
2ace	**1.000** (0.000)	0.996 (0.009)	**1.000** (0.000)	0.999 (0.002)
1af0	**1.000** (0.000)	0.999 (0.006)	**1.000** (0.000)	0.999 (0.004)
3cpa	**0.992** (0.010)	0.982 (0.047)	**0.992** (0.010)	0.964 (0.083)
1psa	0.985 (0.003)	**0.997** (0.002)	0.985 (0.003)	**0.997** (0.002)
6tim	0.998 (0.002)	**1.000** (0.000)	0.998 (0.002)	**1.000** (0.000)
4tim	0.999 (0.000)	**1.000** (0.000)	0.999 (0.000)	**1.000** (0.000)

4. 実験結果

図 3 に 4 つのテンプレート 1ahy, 1bg9, 1jfh, 6tim における予測性能を示す．すべての場

Vol.2010-BIO-21 No.18
2010/6/19

表 4 特異値 0.95 としたときの感度.
Table 4 Sensitivities at specificity 0.95.

	EMR	MLR	EMP	MLP
1zio	0.666 (0.191)	0.668 (0.226)	**0.672** (0.196)	0.646 (0.222)
1arg	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)
1cq7	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)
1ahy	0.970 (0.028)	**1.000** (0.000)	0.980 (0.025)	**1.000** (0.000)
1arg_2	0.998 (0.009)	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)
1map	0.703 (0.041)	**0.883** (0.038)	0.740 (0.041)	0.672 (0.044)
1ams	0.574 (0.072)	**0.636** (0.075)	0.081 (0.036)	0.131 (0.052)
1ahg	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)
1kcd	0.593 (0.102)	**0.875** (0.062)	0.551 (0.113)	0.732 (0.116)
2bvw	**0.247** (0.110)	0.243 (0.113)	**0.247** (0.110)	0.243 (0.113)
1qk2	0.536 (0.196)	**0.666** (0.178)	0.536 (0.196)	**0.666** (0.178)
1bg9	0.840 (0.057)	**1.000** (0.000)	0.840 (0.057)	**1.000** (0.000)
1jfh	0.862 (0.045)	**1.000** (0.000)	0.862 (0.045)	**1.000** (0.000)
1isw	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)
1ka1	**0.944** (0.066)	0.918 (0.100)	**0.944** (0.066)	0.911 (0.103)
2oke	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)	0.988 (0.048)
1eo4	0.968 (0.041)	**1.000** (0.000)	0.968 (0.041)	**1.000** (0.000)
1kfs	0.752 (0.282)	**0.854** (0.229)	0.752 (0.282)	0.760 (0.259)
1rpa	0.707 (0.244)	0.803 (0.241)	0.707 (0.244)	**0.823** (0.245)
1vcz	0.465 (0.057)	0.910 (0.190)	**1.000** (0.000)	0.991 (0.046)
2dhc	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)
1g42	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)
1acb	0.990 (0.005)	**0.991** (0.004)	0.968 (0.009)	0.984 (0.006)
1bls	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)	0.990 (0.072)
2ace	**1.000** (0.000)	0.990 (0.019)	**1.000** (0.000)	0.995 (0.014)
1af0	**1.000** (0.000)	0.996 (0.044)	**1.000** (0.000)	0.991 (0.052)
3cpa	**0.978** (0.026)	0.948 (0.135)	**0.978** (0.026)	0.912 (0.192)
1psa	0.922 (0.019)	**0.989** (0.007)	0.922 (0.019)	**0.989** (0.007)
6tim	0.985 (0.016)	**1.000** (0.000)	0.985 (0.016)	**1.000** (0.000)
4tim	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)

合において，AUC と感度が計量学習によって向上している．6tim は計量学習を用いなくてもすでに高い予測精度を得ているが，さらに計量学習により予測精度が向上している．

表 3 に，本実験で用いたすべてのテンプレートにおける AUC を示す．赤い太字が最高 AUC，青い下線が最高 AUC の場合と統計的有意差がないことを示す．統計的有意差の検出には有意水準 1% として一標本 t-検定を用いた．この表は，計量学習の有効性を示すための豊富な証拠を提供している．30 テンプレート中，22 個のテンプレートで MLR が最高性能を得ている．MLR の AUC は 15 個のテンプレートで EMR よりも統計的に有意に性能がよく，7 個で同等の性能を得た．2ace だけが EMR のほうが有意に性能がよかった．9 個のテンプレート 1arg, 1cq7, 1isw, 2oke, 2dhc, 1g42, 1bls, 2ace, 1af0 において，計量学習を行わなくてもすでに部位マッチとミスマッチが完全に分離していた．このうちほとんどにおいて，計量学習を行ったとしても改悪されることはなかった．これらより，多くの場合によって計量学習によって性能がよくなり，計量学習を用いなくてもすでに十分高い予測性能が得られている場合においても計量学習によってかえって悪くなることはまれであることが示された．

閾値を特異度 0.95 としたときの感度も算出した．閾値を変えることにより，様々な特異度が得られる．AUC 値はあらゆる特異度で設定したときの平均値であり，しばしば予測性能の評価に使われている (e.g. 文献[12])．しかし，AUC には次のような欠点がある．表 1 で示したように，本実験で用いているデータセットは非常に多くのミスマッチをヒットするテンプレートが多い．この場合，特異度が低い閾値はおよそ無意味である．なぜなら，部位マッチを見つける際，RMSD が小さい順にヒットを精査していくとすると，順位が遅いヒットまで見ることはできないからである．閾値を特異度 0.95 としたときの感度も調査したのはこのような理由からである．表 4 にその感度を示す．感度の EMR と MLR の差は AUC のそれより顕著になった．テンプレート 1ka1 と 2ace を除いて，EMR の感度は MLR のそれより統計的に有意に高い感度を得ることはなかった．

図 1 は 1jfh の活性部位 (α-amylase) から作成したテンプレートの詳細な結果を与えている．このテンプレートは 3 残基に含まれる 13 原子からなる．図 1(c) は，訓練用データにおける重みなし RMSD の分布を示している．その分布を，24 の部位マッチと 6,486 のミスマッチのそれぞれの頻度の和が 1 になるように正規化して赤と青でプロットしている．ここで，unweighted RMSD では，部位マッチとミスマッチの分離が分かることが見てとれる．これら 24 の部位マッチと 6,486 のミスマッチに対して計量学習が算出した重みベクトルを図 1(b),(h) に示す．重みあり RMSD の分布は，図 1(d) に示すように，部位サイトとミスサイトの分離が改善している．評価用データ (計量学習には用いていないデータ) に対する重みなし RMSD と重みあり RMSD の分布を図 1(e),(f) に示す．このように評価用データに対しても部位サイトとミスサイトの分離がよい．これらは，提案する計量学習算法は過学習なしに汎化能力 (e.g. 文献[9]) を向上できることを示唆している．

図 1(g) はクエリテンプレートの各原子と各ヒットの対応する原子との距離の分布を Box plot であらわしている．2 原子 'OD1 ASP A 197' および 'CB GLU A 233', は特に分離

Vol.2010-BIO-21 No.18
2010/6/19

が悪い．この２つの原子に対する重みの値は 0 になっている．さらに，そのほかの酸素原子も重みは比較的小さい値になっている．これは，この原子の分布は部位マッチとミスマッチの分離の悪いからであろう．このように，計量学習算法はテンプレート原子の中から予測に重要な原子を自動的に選別することに成功している．

EMP や MLP の結果は，意外にも外部原子も予測には有効であることを示唆している．外部原子を含むクエリテンプレートは表 3,4 では，テンプレート名を青字の斜体文字で表示した．これらは Rough テンプレートと異なる予測を得る可能性がある．MLP で用いるテンプレートは，酵素反応に直接寄与する原子のみ含んでいて，外部原子は一つも含んでいない．しかし，外部原子に含まれる無関係な情報が計量学習を阻害されることはないにも関わらず，MLP は 2ace を除くと MLR より有意に優れた感度を得なかった．これは，外部原子は必ずしもあらかじめ除いておく必要はないことを示唆している．なぜなら，計量学習が不要な原子を自動的に除外するからである．

参 考 文 献

1) Amari, S. and Nagaoka, H.: *Methods of Information Geometry*, AMS and Oxford University Press (2000).
2) Barker, J.A. and Thornton, J.M.: An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis., *Bioinformatics*, Vol.19, No.13, pp.1644–9 (2003).
3) Bartlett, G.J., Porter, C.T., Borkakoti, N. and Thornton, J.M.: Analysis of catalytic residues in enzyme active sites., *J Mol Biol*, Vol.324, No.1, pp.105–21 (2002).
4) Boyd, S. and Vandenberghe, L.: *Convex Optimization*, Cambridge University Press (2004).
5) Chou, K.C. and Cai, Y.D.: A novel approach to predict active sites of enzyme molecules., *Proteins*, Vol.55, No.1, pp.77–82 (2004).
6) Dantzig, G.B.: *Linear Programming and Extensions*, Princeton University Press (2004).
7) Fetrow, J.S. and Skolnick, J.: Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases., *J Mol Biol*, Vol.281, No.5, pp. 949–68 (1998).
8) Gherardini, P.F., Wass, M.N., Helmer-Citterich, M. and Sternberg, M.J.: Convergent evolution of enzyme active sites is not a rare phenomenon., *J Mol Biol*, Vol.372, No.3, pp.817–45 (2007).
9) Hastie, T., Tibshirani, R. and Friedman, J.H.: *The Elements of Statistical Learning*, Springer (2003).
10) Hinrichs, C., Singh, V., Mukherjee, L., Xu, G., Chung, M.K. and Johnson, S.C.: Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset., *Neuroimage*, Vol.48, No.1, pp.138–49 (2009).
11) Ivanisenko, V.A., Pintus, S.S., Grigorovich, D.A. and Kolchanov, N.A.: PDBSiteScan: a program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins., *Nucleic Acids Res*, Vol.32, No.Web Server issue, pp.W549–54 (2004).
12) Kato, T., Tsuda, K. and Asai, K.: Selective integration of multiple biological data for supervised network inference, *Bioinformatics*, Vol.21, pp.2488–2495 (2005).
13) Kato, T., Tsuda, K., Tomii, K. and Asai, K.: A new variational framework for rigid-body alignment, *Structural, Syntactic, and Statistical Pattern Recognition*, Vol.3138, Springer Berlin / Heidelberg, pp.171–179 (2004).
14) Kleywegt, G.J.: Recognition of spatial motifs in protein structures., *J Mol Biol*, Vol.285, No.4, pp.1887–97 (1999).
15) Laskowski, R.A., Watson, J.D. and Thornton, J.M.: Protein function prediction using local 3D templates., *J Mol Biol*, Vol.351, No.3, pp.614–26 (2005).
16) Loewenstein, Y., Raimondo, D., Redfern, O.C., Watson, J., Frishman, D., Linial, M., Orengo, C., Thornton, J. and Tramontano, A.: Protein function annotation by homology-based inference, *Genome Biol.*, Vol.10, No.2, p.207 (2009).
17) Nagano, N.: EzCatDB: the Enzyme Catalytic-mechanism Database., *Nucleic Acids Res*, Vol.33, No.Database issue, pp.D407–12 (2005).
18) Stark, A. and Russell, R.B.: Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures., *Nucleic Acids Res*, Vol.31, No.13, pp.3341–4 (2003).
19) Torrance, J.W., Bartlett, G.J., Porter, C.T. and Thornton, J.M.: Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families., *J Mol Biol*, Vol.347, No.3, pp.565–81 (2005).
20) Wallace, A.C., Borkakoti, N. and Thornton, J.M.: TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites., *Protein Sci*, Vol.6, No.11, pp.2308–2323 (1997).
21) Webb, E.C.: *Enzyme Nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology*, Academic Press Inc., New York (1992).
22) Wright, C.S.: Comparison of the active site stereochemistry and substrate conformation in -chymotrypsin and subtilisin BPN', *J Mol Biol*, Vol.67, No.1, pp.151–63 (1972).