属性付き法線ベクトルを用いた蛋白質分子表面比較方式

兼田佳和[†] 庄治範匡[†]
大川剛直[†] 中村春木[††]

近年，蛋白質の機能の多くはその局所的な分子表面（活性部位）における形状や物性によって決定されることが明らかになってきた．このような背景から，機能未知な蛋白質の分子表面を入力とし，機能既知の活性部位をテンプレートとして比較することで，機能予測を行うシステムを開発している．分子表面データは数万個の頂点で構成されているため，単純に全頂点を使って位置合わせを行うと計算量が膨大になる．そこで本論文では，分子表面上の突起や窪みに着目して，その曲面の法線ベクトルに曲率や物性の情報を付加した属性付き法線ベクトルを導出し，効率的にマッチングする方法を提案する．また，相対位置関係と属性が類似しているベクトル組のみを抽出することで効率化を図る．本手法を抗体蛋白質6組に適用した結果，分子表面の比較に要する時間は平均約3分となった．また，抗体蛋白質103個を用いた機能予測に適用した結果，95.2%の機能予測精度が得られた．

A Method of Comparing Protein Molecular Surface Based on Vertical Vectors with Attributes

Yoshikazu Kaneta,[†] Norimasa Shoji,[†] Takenao Ohkawa[†] and Haruki Nakamura[††]

Recent researches have clarified that the function of a protein depends on its molecular surface. It suggests the possibility of the protein function identification based on the molecular surface comparison, in which a molecular surface of unknown protein is compared with many surfaces of known active sites as reference templates. This thesis presents an effective matching method by using vertical vectors with attributes of the curvature and the physical properties on projections and depressions. The vectors that should be matched are limited by extracting two vectors with the similar relative positions and the attributes of surface in order to reduce computational complexity. The proposed method was applied to 11 surface data. As a result, the mean calculation time was about three minutes. Furthermore, this method was applied to 103 surface data. The result of prediction showed 95.2% prediction accuracy.

1. はじめに

近年，分子認識や触媒作用といった蛋白質の多くの機能は，蛋白質の分子表面における形状，物性（静電ポテンシャル，疎水性，運動性，PHなど）が深く関わっていることが明らかになっている[1)]．また，蛋白質の化学反応は蛋白質全体ではなく，ある決まった機能を発現する原因となる部位（活性部位）によって引き起こされているという報告もある[2)]．このような背景から，分子表面形状に注目し，表面に出現する原子単位で蛋白質を比較する手法が提案されている[3)]．この手法は，分子表面に表れる原子の座標で比較を行うものである．しかし，蛋白質の機能は表面形状とともに，その物性も重要な関わりがあるため，これを含めた比較を行う必要がある．

そこで本論文では，形状と物性の情報を持つ蛋白質分子表面データを対象として，機能未知の蛋白質分子表面データと機能既知の蛋白質分子表面の活性部位データを比較する新しい手法を提案する．蛋白質分子表面の形状と物性を示す情報は，3次元空間上に配置された数万個の点の集合で構成されており，各点にはその部位の物性（静電ポテンシャル，疎水性）を示すデータが付加してある．この既知の分子表面データは，eF-siteデータベース☆で公開されている[4),5)]．

これらのデータを直接的にマッチングして位置合わせすると，膨大な計算時間が必要となる．また，酵素

† 大阪大学大学院工学研究科
Graduate School of Engineering, Osaka University
†† 大阪大学蛋白質研究所
Institute for Protein Research, Osaka University

☆ http://pi.protein.osaka-u.ac.jp/eF-site/

や抗体といった蛋白質の機能に関与する部位は，一般に大きな突起や窪み部分に存在することが報告されている[6),7)]．そこで，活性部位の突起や窪みの部分に着目し，その曲面に対する法線ベクトルを利用して，効率的にマッチングする手法を提案する．突起や窪み面の法線ベクトルは，その曲面の平均曲率，ガウス曲率[8)]を計算し，その値によって突起や窪みの位置を特定することにより求める．これに曲率と物性の情報を付加した属性付き法線ベクトルを定義し，マッチングに利用する．マッチング方法としては，入力蛋白質と活性部位の属性付き法線ベクトルから，各々2本のベクトルを選出し，移動，回転を行うことによって位置合わせする．ベクトルの選出にあたり，すべてのベクトルの組合せをマッチングするのは非効率であるため，2ベクトル間の相対的位置関係の類似しているものを選ぶために，2ベクトル間の距離，角度の4つの情報[9)]から，比較するベクトル組をベクトル間の距離や角度で切り分けたバケットに分割することで効率化を図る．また，厳密なマッチングを行うために，ベクトルによる位置合わせ後，反復改善法[10)]によって位置の微調整を行い，より最適な位置合わせを実現する．

2. 蛋白質の分子表面とその比較

2.1 蛋白質の分子表面

蛋白質は，1本のペプチド鎖が複雑に折り畳まった構造をなしており，蛋白質分子の表面に露出されるアミノ酸と内部に隠れるアミノ酸とが存在している．分子認識や結合といった蛋白質の機能の多くは，その分子表面での形状や物性が深く関わっており，主鎖の折り畳み方がまったく異なっていても局所的な表面物性が同一で，同じ酵素機能を持つ例が報告されている[1)]．こういった観点から，蛋白質の分子表面は，一般的には蛋白質の表面上で結合あるいはダイナミカルに動きうる分子が到達可能な露出表面を計算によって求めることができる．このようにして計算された分子表面の例を**図1**に示す．

分子表面データは，3次元空間上に配置した多数の三角形の集合による多面体で構成される[11)]．物性（静電ポテンシャル，疎水性）を示す情報は，これらの三角形の各頂点に付加される．分子表面データの例を**図2**に示す．このように，各データには，蛋白質のPDBコード☆，その分子表面を構成する三角形のポリゴンの個数，各頂点の座標 (X, Y, Z) と単位法線ベクトル (V_x, V_y, V_z) が記述され，さらに各座標における静電

☆ 4文字の英数字で表される蛋白質のID

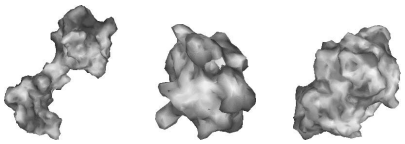


図1 蛋白質分子表面の例
Fig. 1 Examples of protein molecular surface.

＜データ例＞

```
1yee
24478
3
    8.30  15.60  -33.00   -0.09  -0.74  -0.66   -0.06  0.80
    8.36  14.09  -31.85    0.02  -0.15  -0.98   -0.08  0.80
    8.37  13.29  -31.82    0.09  -0.18  -0.97   -0.08  0.50
3
    9.90  17.44  -33.70    0.50  -0.00  -0.86   -0.05  0.80
    9.40  19.36  -33.11    0.37   0.57  -0.73   -0.02  0.80
    8.93  19.14  -33.48    0.06   0.69  -0.71   -0.02  0.80
                  ⋮
```

＜フォーマット形式＞

(蛋白質のPDBコード)
(蛋白質の全頂点数)
(1つのポリゴンを構成する頂点数)
(X) (Y) (Z) (V_x) (V_y) (V_z) (ポテンシャル値) (疎水性値)
⋮

図2 蛋白質分子表面データの例
Fig. 2 Example of protein molecular surface's data.

ポテンシャルや疎水性といった物性がそれぞれ0〜1の連続値として表される．

2.2 分子表面の比較

蛋白質の機能は局所的な表面特性によって決定されることが明らかとなり，機能を発現する部位（活性部位）がいくつか発見されている．このように，蛋白質の活性部位の分子表面と機能との間にはきわめて高い相関があり，機能の類似性を調べるためには，活性部位表面の比較が重要な役割を果たす．特に，機能や活性部位の位置が分かっていない蛋白質が与えられたときに，それらに類似する既知の活性部位を見つけることは，蛋白質の機能予測への発展も期待でき，とりわけ重要である．そこで，ここでは機能既知の活性部位の分子表面データ（蛋白質分子表面全体のうち機能発現に関わる活性な部位を切り出したデータ）をテンプレートと見なし，これに最も類似する部分表面を入力蛋白質表面内から特定し，その類似度を評価するという枠組みで分子表面の比較をとらえる．

3. 属性付き法線ベクトルを用いた分子表面比較手法

3.1 アプローチ

3次元データを比較する方法としては，これまでに3次元空間をシェル・セクタ状に分割し，ヒストグラムを調べて比較する方法[12)]や，3次元データをパラメー

タ化して比較する方法[13]などが提案されている．これらはいずれも大局的な観点から比較する方法であり，機能に関与する突起や窪み部分に重点を置きながら比較するのは困難であると考えられる．

また，3次元データを位置合わせするための最も原始的な方法として，3頂点を選び出して分子表面を重ね合わせる操作をすべての3頂点の組に対して網羅的に行う方法が考えられる．しかし，頂点数を n とすると，分子表面を重ね合わせる組合せは ${}_nC_3 \times {}_nC_3$ で $O(n^6)$ 通りあるため，分子表面データが数千〜数万個の頂点で構成されていることを考慮すると，莫大な組合せが存在するため，実用時間で計算を行うのは困難である．

ここで，分子表面を位置合わせするのに必ずしも頂点の組合せを考える必要はなく，分子表面上の法線ベクトルを利用することも可能である．すなわち，2つの分子表面から各々2つの法線ベクトルを抽出し，片方のベクトルの始点と方向を重ねた後，他方のベクトルを重ねれば全体の大まかな位置合わせができる．法線ベクトルの個数を n とすると，分子表面を重ね合わせる組合せは ${}_nC_2 \times {}_nC_2$ で $O(n^4)$ ですみ，頂点の重ね合わせより効率的な方法である．

法線ベクトルを利用したマッチングではいかに法線ベクトルを定義するかが重要となる．各頂点ごとに法線ベクトルを求めると，ベクトル数は数千から数万個となり，実用時間で処理できるほど少なくはならない．また，機能が突起・窪み部分に強く発現されることを考慮すると，突起・窪み部分の法線ベクトルのみを抽出するのが有効なアプローチと考えられる．

分子表面の突起や窪みの部分を計算機で導出する方法については，分子表面記述に関する研究など[14]，様々な分野で行われているが，曲面の曲がり具合を表す曲率[8]を用いる方法が一般的である．また，物性が類似している突起・窪み部分の法線ベクトルの組を使えば，より正確なマッチングが期待できる．

そこで，属性付き法線ベクトルを導入する．属性付き法線ベクトルは，凹凸の激しい高曲率な突起（窪み）部分の法線ベクトルで，かつその近傍における曲率と物性（静電ポテンシャルと疎水性）を付加したものと定義する．高曲率な部分に限定するのは，特に機能に関わる部位を重視するためと，ベクトルの組合せ数を減らすためである．また，属性を付加するのは，類似した表面形状を持つ法線ベクトルの組のみでマッチングすることで効率化を図るためである．属性付き法線ベクトルの概念図を**図3**に示す．

属性付き法線ベクトルのマッチングにおいて，2つのベクトル組を網羅的に調べるのは非効率である．そこで，入力蛋白質と活性部位で2つのベクトル組を選択したときに，それらのベクトル組が同様の関係，または同様の属性が含まれているものだけを選択することでマッチングの効率化を図る．

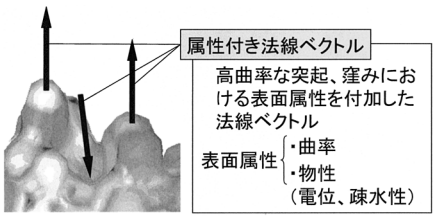


図3 属性付き法線ベクトル
Fig. 3 The vertical vector with attributes.

2つの属性付き法線ベクトルでマッチングした後は，活性部位の各頂点と最も近い点の座標と物性の誤差から非類似度を計算するが，2ベクトルによるマッチングでは大局的には最適なところに位置合わせできても，必ずしも厳密な最適位置にマッチングできるとは限らない．そこで，反復改善法[10]を利用し，非類似度の値が改善される方向に微調整して最適位置に近づける方法を採用する．

3.2 属性付き法線ベクトルを用いたマッチング方式の概要

属性付き法線ベクトルを用いたマッチング方式の概要を**図4**に示す．本方式は，属性付き法線ベクトル作成フェイズ，マッチングフェイズ，反復改善法による位置調整フェイズの3つのフェイズに大分される．属性付き法線ベクトル作成フェイズでは，入力蛋白質とすべての活性部位の属性付き法線ベクトルを作成する．マッチングフェイズでは，入力蛋白質と活性部位それぞれの属性付き法線ベクトルを使って，分子表面データの重ね合わせを行う．その概要を**図5**に示す．ベクトルを重ね合わせるには，入力蛋白質と活性部位それぞれからベクトルを1つ選択して重ね合わせた後，それを軸として残りのベクトルを近傍な位置まで移動させるように回転させることで実現できる．具体的には以下の手順で行う．

（1） 入力蛋白質と活性部位からそれぞれ属性付き法線ベクトルを2つずつ抽出する．各ベクトル組について，入力蛋白質のベクトルを $\vec{a}$, $\vec{b}$, 活性部位のベクトルを $\vec{p}$, $\vec{q}$ とする．

（2） 最初に重ね合わせるベクトル $\overrightarrow{\mathrm{v1}}$，後で重ね合わせるベクトル $\overrightarrow{\mathrm{v2}}$ をそれぞれ決定する．入力蛋白質のベクトルを $\overrightarrow{\mathrm{v1}_{\mathrm{orig}}}$, $\overrightarrow{\mathrm{v2}_{\mathrm{orig}}}$, 活性部位のベクトルを $\overrightarrow{\mathrm{v1}_{\mathrm{tmp}}}$, $\overrightarrow{\mathrm{v2}_{\mathrm{tmp}}}$ とすると，$(\overrightarrow{\mathrm{v1}_{\mathrm{orig}}}, \overrightarrow{\mathrm{v1}_{\mathrm{tmp}}}, \overrightarrow{\mathrm{v2}_{\mathrm{orig}}}, \overrightarrow{\mathrm{v2}_{\mathrm{tmp}}})$

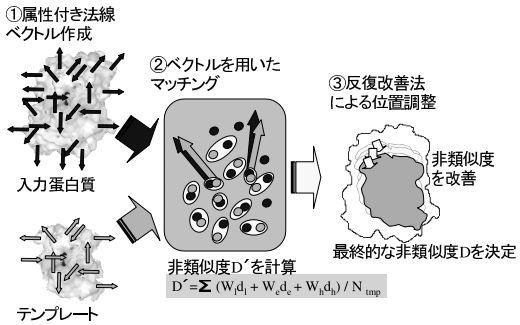


図 4 属性付き法線ベクトルを用いた分子表面比較方式の概要
Fig. 4 The metod of comparing protein surface by vertical vector with attributes.

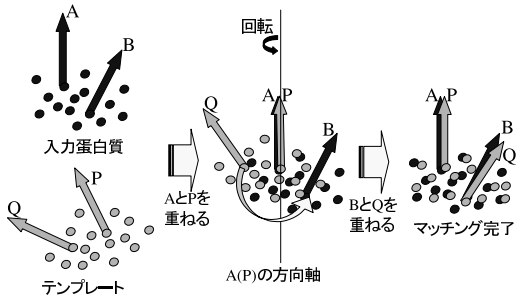


図 5 属性付き法線ベクトルのマッチングの概要
Fig. 5 The matching method by vertical vector with attributes.

は，$(\vec{a},\vec{p},\vec{b},\vec{q})$，$(\vec{a},\vec{q},\vec{b},\vec{p})$，$(\vec{b},\vec{p},\vec{a},\vec{q})$，$(\vec{b},\vec{q},\vec{a},\vec{p})$ の 4 通りが考えられる．それぞれについて (3)〜(4) を行う．

(3) 属性付き法線ベクトル $\overrightarrow{\mathrm{v1_{orig}}}$ と $\overrightarrow{\mathrm{v1_{tmp}}}$ を重ね合わせる．これは，ベクトルの始点を重ねるように活性部位を平行移動し，その後，ベクトルの向きが一致するように活性部位を回転させることで行う．

(4) ベクトル $\overrightarrow{\mathrm{v1_{orig}}}$ ($\overrightarrow{\mathrm{v1_{tmp}}}$) を軸として，$\overrightarrow{\mathrm{v1_{orig}}}$ と $\overrightarrow{\mathrm{v2_{orig}}}$ の作る面と，$\overrightarrow{\mathrm{v1_{orig}}}$ と $\overrightarrow{\mathrm{v2_{tmp}}}$ の作る面が等しくなるまで，活性部位を回転する．これにより最終的な位置合わせが完了する．

入力蛋白質と活性部位を重ね合わせたときの類似性の評価は，各頂点間の誤差を求めることにより行う．類似性を定量的に評価するために，二点 (i,j) 間の誤差 $d_{i,j}$ を以下のように定義する．

$$d_{i,j} = W_l \cdot d_{l_{i,j}} + W_e \cdot d_{e_{i,j}} + W_h \cdot d_{h_{i,j}}$$

ただし $d_{l_{i,j}}$，$d_{e_{i,j}}$，$d_{h_{i,j}}$ は，2 つの分子表面を重ね合わせたとき，最も近くにある点のペア (i,j) における座標，静電ポテンシャル，疎水性の差異である．また，W_l，W_e，W_h は，座標，静電ポテンシャル，疎水性に対する重みパラメータである．2 つの分子表面がどの程度似ていないかを示す非類似度 D は，活性部位に含まれる頂点の数 N_{tmp} で正規化した次の式で求める．

$$D = \frac{1}{N_{tmp}} \sum_{i,j} d_{i,j}$$

非類似度 D が大きければ，2 つの分子表面の形状と物性の差異が大きいことを示しており，小さければ，分子表面間の誤差が小さく，類似した表面形状と物性を持っていることを示している．

3.3 属性付き法線ベクトルの作成方法

3.3.1 平均曲率とガウス曲率

突起や窪みに対応した属性付き法線ベクトルを求めるには，分子表面の形状を判別するための方法が必要となる．これには，一般的に知られている平均曲率とガウス曲率[8]を利用することができる．平均曲率 H とガウス曲率 K は，

$$H = \frac{\kappa_1 + \kappa_2}{2}$$

$$K = \kappa_1 \cdot \kappa_2$$

と定義される．平均曲率 H は曲面が全体として凸か凹かを表しており，ガウス曲率 K は曲面の平面への展開のしにくさの程度を表している．平均曲率 H とガウス曲率 K が分かれば形状と曲率の大きさが判断できることから，これら 2 つの値を属性としてベクトルに付加することにする．

3.3.2 属性付き法線ベクトルの作成方法

属性付き法線ベクトル作成方法の概要を**図 6** に示す．曲率の値が大きくなる突起，窪み部分を見つけることは，分子表面の全頂点についてその周辺の形状を調べていくことで可能となる．また，その部分が突起であるのか，窪みであるのかを判別するには，平均曲率とガウス曲率を計算すればよい．これらの曲率は，注目点からそれに隣接する頂点の方向に沿って探索を行ったときに，探索経路を外接円で近似したときの円の半径を計算することで求めることができる．注目点に隣接する各頂点の方向について曲率を求めた後，その最大値と最小値から平均曲率とガウス曲率を近似的に求めることができる．ここで，曲率の小さいものを属性付き法線ベクトルから除外するために，閾値を設けて閾値以上の曲率を持つ頂点のみを採用する．さらに，突起や窪みの先端部分の頂点だけを残すために，探索経路の深さの差について閾値判定を行い，閾値以下の頂点のみ採用する．以下に具体的な作成手順を示す．

(1) 対象点 P に隣接する n 個の頂点を A_1〜A_n

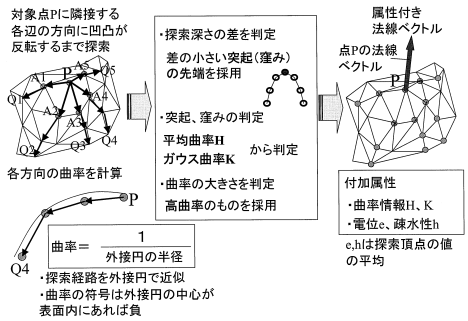


図 6 属性付き法線ベクトル作成方法の概要
Fig. 6 The method of creating the vertical vector with attributes.

とする．各頂点 A_k を探索開始点とし，$\overrightarrow{\mathrm{PA_k}}$ の方向へ頂点 A_k から先に隣接する頂点を探索していく．蛋白質を球に近似したときの中心点を O，探索頂点から面 OPA_k に垂直におろした線の交点を S としたとき，探索ごとに角 A_kPS を調べ，角の変化の増加と減少が転ずるところで探索を打ち切る．

（2） 探索開始点 A_1〜A_n から探索していったときの探索の終点を Q_1〜Q_n としたとき，各探索経路 A_kQ_k における探索深さ d_k を求める．探索深さ d_k は探索した頂点の数と定義する．探索深さ d_k の最大値 $\max(d_k)$ と最小値 $\min(d_k)$ の差が探索深さの差の閾値 T_d 以上の場合は，対象点 P を属性付き法線ベクトルの候補から除外する．

（3） 探索経路 PQ_1〜PQ_n における曲率 κ_1〜κ_n を計算する．曲率 κ_k は，探索経路 PQ_k を円で近似したときの曲率半径から求める．曲率の符号は近似した円の中心が分子表面内にあれば負，外にあれば正とする．曲率 κ_1〜κ_n の中で最大値を κ_{max}，最小値を κ_{min} とすると，平均曲率 H およびガウス曲率 K は

$$H = \frac{\kappa_{max} + \kappa_{min}}{2}$$

$$K = \kappa_{max} \cdot \kappa_{min}$$

で求まる．ここで，$H < 0$，$K > 0$（突起），$H > 0$，$K > 0$（窪み）以外の形になった場合は，属性付き法線ベクトルの頂点候補から除外する．

（4） 曲率 κ_k が曲率制限閾値 T_κ 以下の場合は，対象点 P を法線ベクトルの頂点候補から除外する．

（5） (2)〜(4) の判定を充たす頂点における法線ベクトルを属性付き法線ベクトルと決定する．ベクトルに付加する曲率属性は，近似的に求めた平均曲率 H およびガウス曲率 K とし，物性属性は探索経路 PQ_1〜PQ_n 上の頂点における静電ポテンシャル，疎水性の平均値（それぞれ e，h）とする．

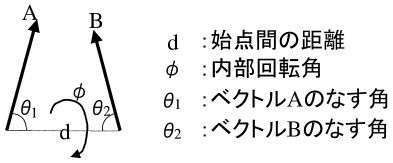


図 7 2 ベクトル間の相対的関係
Fig. 7 The relative relation of two vertical vectors.

3.4 属性付き法線ベクトルによるマッチングの効率化

マッチングする際，入力蛋白質と活性部位それぞれで抽出した属性付き法線ベクトルから 2 つずつ選択する必要があるが，その全ベクトルの組合せは，ベクトル数 n に対して $O(n^4)$ となり，これを網羅的に調べる方法では非効率である．本来，マッチングすべき理想のベクトル組は，2 ベクトル間の距離や角度といった相対的位置関係が同一で，その表面属性も同一なものであるから，このようなベクトル組だけを選択してマッチングする方が効率的である．まず，ベクトル組が持つ情報は次の 2 種類存在する．

- 2 ベクトル間の相対的関係
  属性付き法線ベクトルが持つ方向と始点座標の情報から，ベクトル間の 4 つの相対的関係，すなわち始点間の距離 d，内部回転角 ϕ，ベクトルのなす角 θ_1，θ_2 [9] を導出することができる（**図 7**）．
- ベクトルの持つ表面属性
  属性付き法線ベクトルは，平均曲率 H，ガウス曲率 K，静電ポテンシャル e，疎水性 h の属性を持っている．

これらを利用したマッチング数削減方法として，バケット分割による方法と閾値判定による方法を考える．

3.4.1 バケット分割によるマッチングの効率化

図 7 に示した 4 つの相対的関係は，ベクトルの組合せごとに決定される．入力蛋白質と活性部位で d，ϕ，θ_1，θ_2 が類似した値を持つベクトル組の組合せを抽出できれば，類似した関係を持つベクトル組どうしの効率的なマッチングが実現できる．ここで，バケット分割法を導入する．今回の対象の場合，d，ϕ，θ_1，θ_2 という 4 つの要素が索引となり，バケットに挿入するものがベクトル組となる．バケットテーブルを入力蛋白質，活性部位でそれぞれ作成し，同一バケットに含まれるベクトル組どうしでマッチングを行う．ベクトル数を n 個，バケット数を m 個とすると組合せ数は ${}_nC_2/m \times {}_nC_2/m$ で $O(\frac{n^4}{m^2})$ となり，計算量が大幅に削減できることが分かる．**図 8** にバケット分割法を用いた効率的マッチングの概要を示す．また，その具体的な手順を以下に示す．

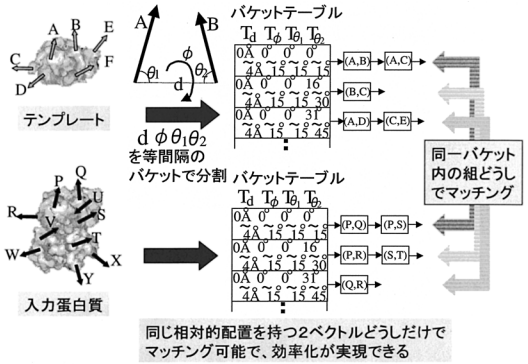


図 8 バケット分割法を用いた効率的マッチングの概要
Fig. 8 The effective matching by the method of bucket separation.

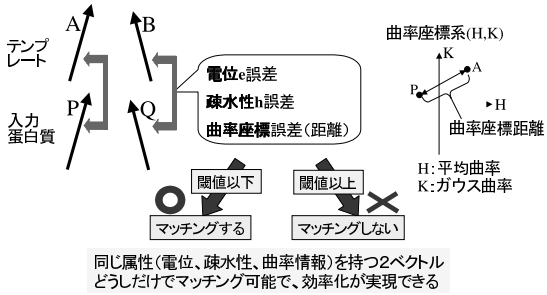


図 9 閾値判定法を用いた効率的マッチングの概要
Fig. 9 The effective matching by the method of threshold.

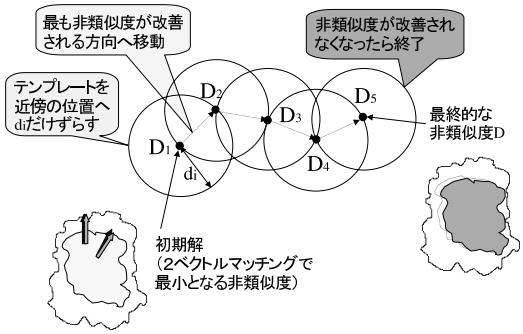


図 10 反復改善法を用いた位置調整法の概要
Fig. 10 The method of subtle adjusting by local search method.

(1) 入力蛋白質と活性部位のバケットテーブルを作成する．バケットテーブルは d，ϕ，θ_1，θ_2 の 4 つの要素からなる 4 次元テーブルであり，d は 0〜50 Å，ϕ，θ_1，θ_2 は 0 度〜180 度の範囲に固定する．各要素はあるバケット分割数 N_b で等間隔に分割する．

(2) 入力蛋白質および活性部位のバケットテーブルにすべての属性付き法線ベクトルの組を格納する．

(3) 入力蛋白質と活性部位において同一バケットにアクセスし，そのバケット内のベクトル組のみを使ってマッチングを行う．この操作をすべてのバケットについて行う．

3.4.2 閾値判定法によるマッチング数削減

属性付き法線ベクトルに付加した 4 つの表面属性が類似した値になるようなベクトル組を抽出するために，閾値判定法を導入する．判定には，曲率座標 (H, K)，静電ポテンシャル e，疎水性 h を使用する．曲率座標は，平均曲率 H を横軸，ガウス曲率 K を縦軸にとった座標系で，曲面の形状，曲率の大きさなどを表す．要素の差を求めるには，曲率座標については 2 つの座標間の距離を求め，静電ポテンシャルと疎水性については値の差をそのまま計算する．これらに対応した 3 つの閾値（曲率座標距離の閾値 T_{HK}，静電ポテンシャルの閾値 T_e，疎水性の閾値 T_h）を定め，閾値判定を行う．**図 9** に閾値判定法を用いた効率的マッチングの概要を示す．また，その具体的な手順を以下に示す．

(1) 入力蛋白質と活性部位の属性付き法線ベクトルの組をそれぞれ $(\vec{a}, \vec{b})$，$(\vec{p}, \vec{q})$ とし，$\vec{a}$ と $\vec{p}$，$\vec{b}$ と $\vec{q}$ を重ね合わせるものとする．また，$\vec{a}$，$\vec{p}$ に付加した表面属性をそれぞれ (H_a, K_a, e_a, h_a)，(H_p, K_p, e_p, h_p) とする．これらの値から，曲率座標距離 $d_{HK} = \sqrt{(H_a - H_p)^2 + (K_a - K_p)^2}$，静電ポテンシャル差 $d_e = e_a - e_p$，疎水性差 $d_h = h_a - h_p$ を求める．

(2) 閾値 T_{HK}，T_e，T_h を用いて $d_{HK} < T_{HK}$，$d_e < T_e$，$d_h < T_h$ を満たすかどうかを判定する．1 つでも満たさない場合，そのベクトル組によるマッチングは行わない．すべて満たす場合は (3) へ進む．

(3) $\vec{b}$ と $\vec{q}$ に対しても (1)，(2) と同様の判定を行い，成立すれば $\vec{a}$ と $\vec{p}$，$\vec{b}$ と $\vec{q}$ によるマッチングを行う．

3.5 反復改善法を用いた位置調整法

反復改善法による位置調整法の概要を**図 10** に示す．入力蛋白質と活性部位を非類似度が最小となるような 2 つの属性付き法線ベクトルでマッチングした位置が初期位置となり，その非類似度が初期解となる．近傍解の探索は，活性部位を微小距離 d_i だけずらした位置における非類似度を求めることで行う．この作業を非類似度が改善されなくなるまで繰り返す．改善されなくなった位置における非類似度を最終的な非類似度と決定する．この方法により，最適位置（非類似度が最小となる位置）により近づけることが可能となる．以下に具体的な位置調整手順を示す．

(1) 2 つの属性付き法線ベクトルで位置合わせし

て求めた非類似度の中で最小となる位置を初期位置 A_1 とし，その非類似度を D_{A_1} とする．

（2） A_1 における活性部位の各頂点座標 (x, y, z) について距離 d_i だけずらす処理を行う．あらゆる方向へずらすのは計算時間がかかってしまうので，x の値を $x-d_i$，x，$x+d_i$，y の値を $y-d_i$，y，$y+d_i$，z の値を $z-d_i$，z，$z+d_i$ とするような 26 通り（$(x-d_i, y-d_i, z-d_i), (x, y-d_i, z-d_i), \cdots, (x+d_i, y+d_i, z+d_i)$）の変更を行い，それぞれ移動した位置を B_1〜B_{26} とし，その位置における非類似度を D_{B_1}〜$D_{B_{26}}$ とする．

（3） 非類似度 D_{B_1}〜$D_{B_{26}}$ の中で最小のものを D_{min} とする．もし，$D_{min} < D_{A_1}$ ならば，（4）へ進む．もし，$D_{min} \geq D_{A_1}$ ならば，最終的な非類似度を $D = D_{A_1}$ と決定してマッチングを終了する．

（4） A_1 を D_{min} となる位置 B_k に移し，（2）に戻る．

4. 評価実験

本論文で提案した手法を eF-site データベースに登録されている抗体蛋白質 103 個に適用し，評価を行った．これらの蛋白質は，最も標準的な蛋白質立体構造データベースである PDB データベースに登録されているすべての蛋白質のうち，2.8 Å 以上の高い分解能で精度良く構造が決定されているもの[15]である．評価する項目は，属性付き法線ベクトルの導入によってどの程度計算時間が短縮されるか，そしてどの程度正確にマッチングできるかである．また，実際に蛋白質の機能予測問題に適用するとき，どの程度の機能予測精度が得られるかについても評価を行った．なお，実験で使用した計算機は COMPAQ 社の AlphaServer DS20（CPU：2 × Alpha21264 500 MHz，Memory：1.5 GB）である．

4.1 効率性と正確性の評価実験

この実験では，全ベクトル組をマッチングする方法，すなわちバケット分割と閾値判定によるベクトル組削減を行わずにマッチングする方法と提案手法とを比較し，計算時間がどの程度短縮されるかを検証する．また，最適な位置合わせと比べて，どの程度非類似度の値が変化するかを検証する．なお，最適な位置合わせは，蛋白質の機能に関する知見に基づき，グラフィックスディスプレイを用いて人手で慎重に位置合わせした後，網羅的に探索することで得たものである．実験の対象としたのは，抗体蛋白質 103 個の中からランダムに同じ機能を持つ蛋白質を 2 つずつ組にして取り出した 6 組（12 個：1a4k と 1a4j_cut，1yec と 1yeh_cut，1yee と 1yeh_cut，1yed と 1yeh_cut，1mlc と 1mlb_cut，2igf と 1igf_cut）である．ただし，‘1a4k’ は蛋白質のコードであり，‘1a4j_cut’ は蛋白質 ‘1a4j’ から切り出された活性部位であることを示している．具体的な実験の手順を以下に示す．

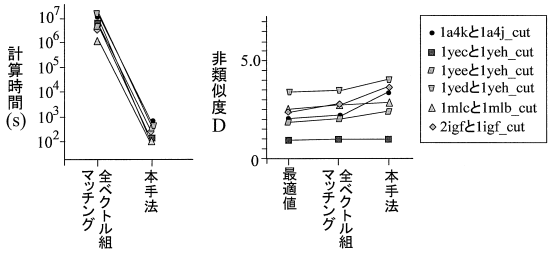


図 11 効率性と正確性の実験結果
Fig. 11 The evaluation of effectiveness and accuracy.

（1） 実験で用いる 6 つの機能グループを選択し，それぞれのグループから蛋白質を 1 つ抜き出す．

（2） （1）で抜き出したものとは別の蛋白質をそれぞれの機能グループから抜き出し，機能データをもとに活性部位である部分を切り出しておく．

（3） 提案手法と全ベクトル組マッチングの各々について，同じ機能グループに属する蛋白質と活性部位の組をマッチングして非類似度を算出し，それに要した時間を測定する．

実験結果を**図 11** のグラフに示す．図 11 の左側のグラフの縦軸は計算時間（秒）で，全ベクトル組マッチングと本手法を比較している．右側のグラフの縦軸は非類似度 D で，人手による最適値と全ベクトル組マッチング，本手法を比較している．

計算時間については，全ベクトル組マッチングでは平均は約 50 日となったが，本手法では平均は約 3 分となり，飛躍的に計算時間を短縮できたことが分かる．非類似度については，本手法でも全ベクトル組マッチングとほぼ同じ非類似度で，かつ最適値にも近い値が算出されていることが分かる．

4.2 機能予測精度の評価実験

本手法を抗体蛋白質 103 個を用いた機能予測に適用した．これらの蛋白質は，抗原との反応実験の結果から，同一機能を持つものをまとめていくと 82 のグループに分類される．82 のグループのうち，活性部位が複数登録されているものが 21 グループであることから，この 21 グループからの代表を 1 つずつ，合計 21 個の蛋白質を機能未知ものとして入力し，残り 82 個を活性部位としてデータベースに登録することにした．このため，21 個の入力蛋白質と同一の機能を持つ活性部位が必ずデータベースに存在する．機能

表 1 機能予測結果
Table 1 The result of function prediction.

蛋白質	第一候補	第二候補	第三候補
1a4j	**1a4k_cut**	1igi_cut	2jel_cut
1a6w	**1a6u_cut**	1ngp_cut	1yeh_cut
1ngq	**1ngp_cut**	1a6u_cut	1mfa_cut
1a7n	1vfa_cut	**1a7q_cut**	1dvf_cut
1kiq	**1vfa_cut**	1dvf_cut	1ucb_cut
2cgr	**1cgs_cut**	1mlb_cut	1ad9_cut
1dba	**1dbj_cut**	1ucb_cut	1lmk_cut
1ggc	**1ggi_cut**	1ucb_cut	1dvf_cut
1ifh	**1hil_cut**	1frg_cut	1ucb_cut
1hkl	**1gaf_cut**	1ucb_cut	1kel_cut
1igf	**2igf_cut**	1ucb_cut	1vfa_cut
1igj	**1igi_cut**	1plg_cut	1ucb_cut
1kem	**1kel_cut**	1flr_cut	1vfa_cut
2hrp	**1mf2_cut**	1igc_cut	1ucb_cut
1mlc	**1mlb_cut**	1ucb_cut	1dbj_cut
1mrc	**1mrd_cut**	1tet_cut	1ucb_cut
1cly	**1ucb_cut**	1hyy_cut	1dbj_cut
1cbv	**1nbv_cut**	1frg_cut	1plg_cut
1yec	**1yeh_cut**	1frg_cut	1plg_cut
1yed	**1yeh_cut**	1vfa_cut	1plg_cut
1yee	**1yeh_cut**	1dvf_cut	1frg_cut

予測実験の手順を以下に示す．

（1） 82 の機能グループから蛋白質を 1 つずつ選び出し，活性部位を切り出してデータベースに登録する．

（2） 機能未知な蛋白質を入力とし，提案手法を用いてどの活性部位と類似するかを調べる．

（3） 類似性の最も高い活性部位の持つ機能を予測結果とする．

本手法を適用した結果を**表 1** に示す．太字で示されている活性部位は，抗体蛋白質と同一機能を持つものであることを示している．表 1 のように，21 個中 20 個の蛋白質について機能予測が成功したが，蛋白質 ‘1a7n’ に関しては，同一機能を持つ活性部位 ‘1a7q_cut’ が第 2 候補に現れる結果となり，機能予測に失敗した．

失敗した蛋白質の機能を調べると，‘1a7q_cut’ と ‘1vfa_cut’ はどちらも ‘lysozyme’ というグループの抗原と結合することから，非常に類似した分子表面と機能を持っていることが分かった．このような活性部位を表面データのみから正しく機能予測するのは難しく，最終的にはいくつかの候補を提示し，人間の判断に委ねる必要があると思われる．また，‘1ucb_cut’ が第 2，第 3 候補に散見されるが，これは ‘1ucb_cut’ が糖と反応する蛋白質の活性部位であり，活性部位中心の静電ポテンシャルが中性に近いという平均的な特性を持つために，テストセット内の蛋白質のうち，多くのものと比較的高い類似度が得られたと思われる．

5. 結　　論

本論文では，蛋白質分子表面データの比較に基づく機能予測のためのマッチング方式として，突起，窪みに着目して正確かつ効率的なマッチングを実現する属性付き法線ベクトルを用いたマッチング手法を提案した．また，属性付き法線ベクトルの組合せ数をバケット分割法や閾値判定法を用いて削減することにより，効率化を図った．本手法を実際の蛋白質の分子表面データに適用した結果，全ベクトル組のマッチングでは約 50 日の計算時間を要するものが約 3 分で計算可能となり，かつ最適位置に近い位置にマッチングすることができた．また，抗体蛋白質 103 個を使った機能予測実験では，95.2%の予測精度が得られた．

一方で，1 対 1 の比較を行うのに平均約 3 分となったが，最悪 8 分かかるものもあった．これは，導出されるベクトル数が対象によって差が出るためである．そこで今後，ベクトル数が計算時間と位置合わせの正確性を考慮した適切な値となるようなベクトル作成方法を考案する必要があると考えている．

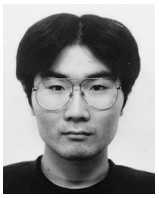
謝辞 日頃よりご指導いただく薦田憲久教授に深謝する．また本研究に関して議論いただいた木下賢吾博士に感謝する．本研究の一部は科学技術振興事業団および日本学術振興会科学研究費補助金からの助成による．

参 考 文 献

1) Goto, S., Nishioka, T. and Kanehisa, M.: LIGAND: Chemical Database for Enzyme Reactions, *Bioinformatics*, Vol.14, pp.591–599 (1998).
2) 中村春木：構造ゲノム科学 構造生物学によるゲノム情報解析へのアプローチ，蛋白質 核酸 酵素，Vol.44, No.3, pp.112–119 (1999).
3) Poirrette, A.R., Artymiuk, P.J., Rice, D.W. and Willett, P.: Comparison of protein surfaces using a genetic algorithm, *J. Computer-Aided Molecular Design*, Vol.11, pp.557–569 (1997).
4) 木下賢吾，中村春木：構造ゲノム科学からのアプローチ，実験医学別冊：プロテオーム解析法，羊土社 (2000).
5) Kinoshita, K., Furui, J. and Nakamura, H.: Identification of Protein Functions from a Molecular Surface Database, eF-site, *J. Struct. Funct. Genomics*, Vol.2, pp.9–22 (2001).
6) Laskowski, R.A., Luscombe, N.M., Swindells, M.B. and Thornton, J.M.: Protein clefts in molecular recognition and function, *Protein Sci.*, Vol.5, pp.2438–2452 (1996).

7) Peters, K.P., Fauck, J. and Frommel, C.: The automatic search for ligand-binding sites in proteins of known 3-dimensional structure using only geometric criteria, *J. Mol. Biol.*, Vol.256, pp.201–203 (1996).
8) 佐藤伊助：いろいろな曲線と曲面，裳華房 (1979).
9) Mizuguchi, K. and Go, N.: Comparison of Spatial Arrangements of Secondary Structural Elements in Proteins, *Protein Eng.*, Vol.8, pp.353–362 (1995).
10) 茨木俊秀：アルゴリズムとデータ構造，昭晃堂 (1989).
11) Connolly, M.L.: Solvent-accessible Surfaces of Proteins and Nucleic Acids, *Science*, Vol.221, pp.709–713 (1983).
12) Ankerst, M., Kastenmuller, G., Kriegel, H.P. and Seidl, T.: 3D Shape Histograms for Similarity Search and Classification in Spatial Databases, *Proc. 5th International Symposium on Large Spatial Databases* (1997).
13) Osada, R., Funkhouser, T., Chazelle, B. and Dobkin, D.: Matching 3D Models with Shape Distributions, *Proc. International Conference on Shape Modeling and Applications*, May 7-11, pp.154–166 (2001).
14) Lin, S.L., Nussinov, R., Fischer, D. and Wolfson, H.J.: Molecular Surface Representations by Sparse Critical Points, *PROTEINS: Structures, Function, and Genetics*, Vol.18, pp.94–101 (1994).
15) Shirai, H., Kidera, A. and Nakamura, H.: H3-rules: Identification of CDR-H3 structures in antibodies, *FEBS Lett.*, No.455, pp.188–197 (1999).

(平成 13 年 6 月 28 日受付)
(平成 13 年 10 月 16 日採録)



兼田 佳和

昭和 53 年生．平成 13 年大阪大学工学部電子情報エネルギー工学科卒業．同年同大学大学院工学研究科情報システム工学専攻博士前期課程入学．蛋白質の分子表面比較に関する研究に従事．

庄治 範匡

昭和 50 年生．平成 13 年大阪大学大学院工学研究科情報システム工学専攻博士前期課程修了．同年関西電力株式会社入社．

大川 剛直（正会員）

昭和 38 年生．昭和 63 大阪大学大学院工学研究科通信工学専攻博士前期課程修了．現在同研究科情報システム工学専攻助教授．工学博士．知識処理，バイオインフォマティクスの研究に従事．IEEE 等の会員．

中村 春木

昭和 27 年生．昭和 55 年東京大学大学院理学系研究科物理学専攻博士後期課程修了．現在大阪大学蛋白質研究所教授．理学博士．生物物理学，蛋白質工学の研究に従事．日本蛋白質科学会等の会員．