

分子表面データ間の距離を利用した効率的類似蛋白質検索方式

中 川 忠 輔[†] 兼 田 佳 和[†] 大 川 剛 直[†]

蛋白質の機能は局所的な表面部位（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知られている。このような背景から、機能未知な蛋白質の分子表面に対して既知の活性部位群から類似した分子表面が存在するかを調べることで、機能を予測することが可能である。現在、既知の活性部位は数千個存在するため、すべての活性部位と逐一類似性を評価するのは非効率的である。そこで本論文では、分子表面データ間の距離を定義することで、効率的に類似蛋白質を検索する手法を提案する。2つの蛋白質の分子表面上の頂点を1対1に対応付け、各頂点間距離の差の総和を求め、その総和が最小となる頂点の対応が求まったとき、その値を分子表面データ間の距離と定義する。ここで、頂点の対応付けの組合せを網羅的に調べると膨大な計算時間がかかることから、対応スコアを導入することで初期対応を求め、頂点ペアの交換により分子表面データ間の距離の改善を行う。各々の分子表面上の2頂点の相対位置関係が類似する頂点ペアに高い対応スコアを与える。この分子表面データ間の距離を利用して、類似する可能性のない検索対象を三角不等式を用いて除外することで検索の効率化を図る。提案した類似蛋白質検索方式を、実際の酵素蛋白質データに適用した。その結果、検索が効率的に行われ、従来手法よりも機能予測精度が向上したことから提案方式の有効性が確認された。

An Effective Similar Protein Retrieval Method Using the Distance between Molecular Surface Data

TADASUKE NAKAGAWA,[†] YOSHIKAZU KANETA[†]
and TAKENAO OHKAWA[†]

The shape and the property of a local part of protein molecular surface, named an active site, plays a crucial role in determining its function. Therefore, the function of protein can be predicted by comparing its molecular surface with that of many active sites whose functions are known. It's inefficient to compare the input surface with all of the stored active site data, because thousands of active site data have been stored in the active site database. This paper proposes an effective similar protein retrieval method using the distance space technique. The distance between two molecular surfaces is defined as the minimum value of the sum of all differences between the relative distance of two surfaces assuming that one-to-one correspondence of vertexes is found between two surfaces. The approximate distance is calculated by introducing the score for matching the vertexes. The distance space is constructed by evaluating the distance between all pairs of molecular surfaces in advance in order to perform efficient retrieval. The method was applied to a set of 310 enzyme proteins. The effectiveness of the method was confirmed by the fact that it can shorten the retrieval time to 8.6% of the total retrieval.

1. はじめに

蛋白質の機能の多くは、蛋白質分子表面上の形状や物性（静電ポテンシャル、疎水性等）に深く関与している¹⁾。また、これらの機能は、局所的な表面特性部位（活性部位）によって引き起こされている²⁾。そこで、既知活性部位の機能情報および分子表面データをデータベース化し、入力された機能未知の蛋白質表面に類似する活性部位をデータベースから検索すること

により、入力蛋白質の機能の分析や予測に役立てようという試みがなされている。このため、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し類似性を評価する様々な手法が提案されている^{3)~10)}。しかしながら、蛋白質の表面は非常に複雑な構造をしており、その比較には、多くの計算時間を必要とする。うえ、データベースには多数の活性部位が登録されているため、すべての活性部位と逐次的に類似性を評価するのは非効率であり、実用性が低い。

そこで、本研究では分子表面データ間の距離を定義することで、効率的に類似蛋白質を検索する手法を提案する。蛋白質間の類似性の評価のために、2つの蛋

[†] 大阪大学大学院情報科学研究科
Graduate School of Information Science and Technology,
Osaka University

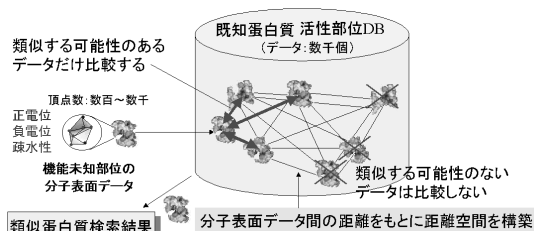


図1 類似蛋白質の検索

Fig. 1 Retrieval for similar protein.

白質の分子表面上の頂点を1対1に対応付け、各頂点間距離の差の総和を求め、その総和が最小となる頂点の対応が求まったとき、その値を分子表面データ間の距離と定義する。検索の際には、図1のようにデータベース内のすべての活性部位分子表面データ間の距離を事前に計算し、距離空間を構築しておくことで、類似する可能性のない検索対象を除外する。この方式により、高速な検索を実現し、実用的な検索システムを構築する。

2. 蛋白質分子表面データ

蛋白質の分子表面データは蛋白質機能部位表面形状データベース eF-site で公開されている⁸⁾。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⁾。これらの三角形の各頂点には、その点における物性を示す以下の情報が付加されている。

- 静電ポテンシャル

求める座標において、最も近くに存在する原子の種類とその原子が属しているアミノ酸の種類に基づき、連続体近似によるポアソン・ボルツマン方程式を解くことで、静電ポテンシャルを求めることができる。[-1.0, 1.0] の値で表現される。

- 疎水性

アミノ酸の種類により、疎水性の度合いが実験的に知られており、[0.0, 1.0] の値で表現される。疎水性の度合いは、各頂点の付近に存在する残基の性質によって決定される。

各蛋白質の分子表面データには、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、その座標における負電位、疎水性、正電位といった物性、および単位法線ベクトル (V_x, V_y, V_z) が記載されている。この分子表面データに基づく蛋白質表面の3次元画像の例を図2

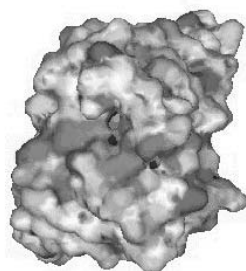


図2 蛋白質の分子表面

Fig. 2 Molecular surface of protein.

に示す。

3. 分子表面データ間の距離を利用した類似蛋白質検索方式

3.1 分子表面データ間の距離の定義

類似する2つの蛋白質の分子表面上の頂点は1対1に対応させることができ、一方の分子表面上の2頂点間の距離と、それに対応したもう一方の分子表面上の2頂点間の距離は類似した値となる。これを物性の面からも考慮することで、その2頂点間の距離の値の差から分子表面データ間の距離を定義する。すなわち、座標、静電ポテンシャル値、疎水性値からなる5次元ベクトルを考え、2ベクトル間の頂点間距離を利用することで、分子表面データ間の距離を定義する。なお、座標のみによる距離と静電ポテンシャル値、疎水性値ではスケールが異なっているため、重みパラメータを設定することで適切な座標系のスケーリングが行えるようにする。

3.1.1 定義

比較する2つの分子表面を S_1, S_2 とし、分子表面 $S_k (k=1, 2)$ 上のある頂点 p_i の x, y, z 座標、静電ポテンシャル値、疎水性値をそれぞれ $x_i(S_k), y_i(S_k), z_i(S_k), e_i(S_k), h_i(S_k) (i=1, 2, 3, \dots, N)$ とする。ただし、 N は分子表面の頂点の個数であり、2つの分子表面の頂点は同数で一意的に決まるものとする。また、頂点 p_i における x, y, z 座標、静電ポテンシャル値、疎水性値を要素とする5次元ベクトルを次のように表す。

$$\vec{v}_i = (x_i(S_k), y_i(S_k), z_i(S_k), e_i(S_k), h_i(S_k))$$

このとき、分子表面 S_k 上の2頂点 $p_i(S_k), p_j(S_k)$ の間の頂点間距離 $d_{ij}(S_k)$ を次式のように定義する。

$$d_{ij}(S_k) = \|\vec{v}_i(S_k) - \vec{v}_j(S_k)\|$$

$$= (W_l \cdot \{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2\} + W_e \cdot (e_i - e_j)^2 + W_h \cdot (h_i - h_j)^2)^{\frac{1}{2}}$$

ただし, 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. ここで, 頂点の対応行列 C を次のように定義する.

$$C_{ij} = \{0, 1\} \quad (i, j = 1, 2, \dots, N)$$

$$\sum_{j=1}^n C_{ij} = 1, \quad \forall i$$

$$\sum_{i=1}^n C_{ij} = 1, \quad \forall j$$

分子表面 S_1 上の頂点 $p_i(S_1)$ と分子表面 S_2 上の頂点 $p_j(S_2)$ が対応しているときに要素 C_{ij} の値は 1 となり, 対応していない場合には 0 となる.

ある頂点の対応 C における距離差 $|d_{i_1 i_2}(S_1) - d_{j_1 j_2}(S_2)|$ の総和 D_c は次式のように表すことができる.

$$D_c = \sum_{i_1=1}^n \sum_{j_1=1}^n \sum_{i_2=1}^n \sum_{j_2=1}^n C_{i_1 j_1} C_{i_2 j_2} |d_{i_1 i_2}(S_1) - d_{j_1 j_2}(S_2)| / N^2$$

ここで N^2 で割ることにより距離差を正規化し, 頂点数に依存しない値としている.

頂点の対応を変化させながら上式の計算を繰り返し, 距離差を最小にするような頂点の対応 C が求まったとき, その最小値を分子表面 S_1 , S_2 の間の距離 $D(S_1, S_2)$ と定義する. すなわち

$$D(S_1, S_2) = \min_C D_c$$

である.

3.1.2 ダミー頂点の付加

上記の $D(S_1, S_2)$ が距離として定義できるのは比較する頂点数が等しい場合に限定される. これは対応のとれない頂点が三角不等式を成り立たせない原因となりうるからである. そこで比較頂点数を事前に決定し, それに満たない頂点数の分子表面データにはダミー頂点を付加することで, 全体として比較頂点数を統一する.

ダミー頂点を付加するにあたって, その座標と電位, 疎水性といった物性を決定する必要がある. ダミー頂点は, 距離の公理を満たすための手段であり, 本来存在しないはずの頂点である. そのため, 分子表面を持つ本来の類似性を反映した距離を計算できなくなる可能性があることから, 距離の値にできるだけ影響を与

えないことが望ましい. そこで, ダミー頂点は分子表面データの全頂点の座標の平均値, すなわち分子表面の重心に存在することとし, 電位, 疎水性の値も全頂点の値の平均値として定義する.

3.2 分子表面データ間の距離の算出

以上より 2 つの分子表面上の頂点を対応付けすることにより分子表面データ間の距離が算出される. このとき, 比較頂点数を N 個とすれば, 各々の分子表面の N 個の頂点を対応付けるときの場合の数は $N!$ となり, すべての組合せについて D_c の値を計算するのは現実的でない. そこで距離の算出プロセスを頂点の初期対応決定フェイズと局所探索による頂点对応付けの改善フェイズに分け, 近似的に距離を算出することとする. 概要を図 3 に示す.

頂点の初期対応決定フェイズでは, まず一方の注目頂点がもう一方のどの頂点と対応しそうかを探索し, 対応しそうなものほど高いスコアで表しておき, そのスコアをもとに, 初期対応を決定する.

局所探索による改善フェイズではスコアの低い頂点の対応から順に別の頂点と 2 点交換することで分子表面データ間の距離がより小さくなる対応を求める.

3.2.1 頂点の初期対応の決定

分子表面上のある注目頂点と, それを除く他の頂点との間の相対位置関係が求まっていたとする. このとき, 2 つの分子表面が類似しているならば, 一方の注目頂点における $N - 1$ 個の相対位置関係と類似した相対位置関係を持つ頂点がもう一方の分子表面上に存在する. この頂点と注目頂点を対応させれば, 分子表面データ間の D_c の値が小さくなると考えられる. そこで, 以下に述べる方法により, 一方の注目頂点がもう一方のどの頂点と対応しそうかを探索し, 対応しそうなものほど高いスコアで表す. この対応スコアを用いて高いスコアを持つ頂点ペアを優先的に対応させる

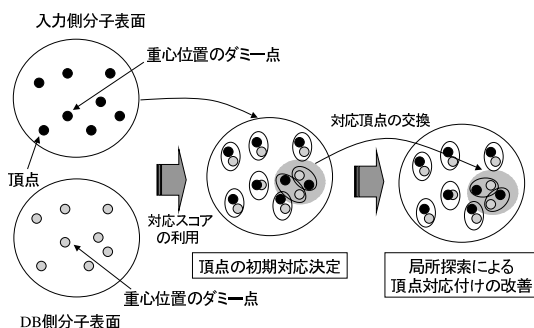


図3 分子表面データ間の距離の算出

Fig. 3 Calculation of the distance between molecular surface data.

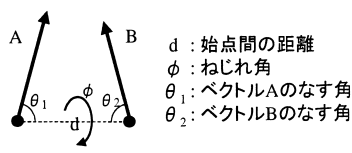


図4 2頂点間の相対的關係

Fig. 4 The relative relation between two vertexes.

ことで、頂点の初期対応を決定する。

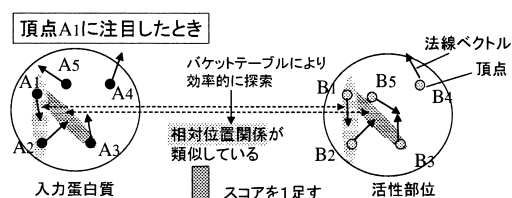
まず、分子表面上のある頂点とその他の頂点との間の相対位置関係を導出する。ここでは座標 (X, Y, Z) と単位法線ベクトル (V_x, V_y, V_z) を用いることで相対位置関係を求める。相対位置関係を表す要素としては図4に示すように、始点間の距離 d 、ねじれ角 ϕ 、法線ベクトルのなす角 θ_1, θ_2 の4つがある¹²⁾。これらの4つの値は2つのベクトルが分かれば一意に定まる値であり、類似した相対的位置関係を持つ頂点ペアを選出するのに利用できる。

具体的な手順を以下に示す。

- (1) 2つの分子表面それぞれに対し、分子表面上の2頂点の全組合せについて相対位置関係を示す4要素を求める。
- (2) 一方の分子表面の任意の2頂点を選択し、もう一方の分子表面上からその相対位置関係と類似する相対位置関係を持つ2頂点を探索し、存在すれば各頂点ペアに対応スコアに1を加える。
- (3) (2)の操作をすべての2頂点について行い、各頂点ペアの対応スコアを求める。

対応スコアの算出にあたり、2つの分子表面データからそれぞれ2頂点ずつ選択し、それらの相対位置関係をすべて調べると、全頂点の組合せは比較頂点数 N に対し $O(N^4)$ となる。そこで、相対位置関係が類似している頂点ペアどうしのみを選択することにより対応スコア算出の効率化を図る。

図4に示した $d, \phi, \theta_1, \theta_2$ は、各頂点組ごとに決定される。2つの分子表面上で $d, \phi, \theta_1, \theta_2$ が類似した値を持つ頂点組の組合せを高速に抽出できれば、類似した相対的位置関係を持つ頂点ペアどうしの効率的なスコアの算出が実現できる。ここで、バケット分割法を導入する。バケット分割法とは、索引となる複数の要素(上限と下限は固定)を一定の幅で複数のバケットに分割し、その中に膨大なデータを振り分けることで、近い値の要素を持つデータだけを高速に検索する方法である(1軸あたりのバケット分割数/近傍バケットへの探索幅)を m とすると、組合せ数は N^4/m^4 となり、効率を大幅に改善することができる。今回の対象の場合、 $d, \phi, \theta_1, \theta_2$ という4つの要素



	B1	B2	B3	...
A1	1+1			
A2		1		
A3			1	
...				

各頂点についても同様に類似相対位置関係をもつ頂点にスコアを与える

図5 対応スコアの算出

Fig. 5 Calculation of score.

が索引となり、バケットに挿入するデータが頂点組となる。十分細かい幅で分割されたバケットテーブルを2つの分子表面データについてそれぞれ作成し、同一バケットおよびその近傍にあるバケットに含まれる頂点組どうしについて4つの相対位置関係の値がある閾値以内であればその頂点ペアの対応スコアに1を加えることとする。図5にその様子を示す。こうすることで類似した相対位置関係を多く持つ頂点ほど高い対応スコアが与えられる。

頂点の対応スコアが求まれば、次に2つの分子表面の頂点の初期対応を決定する。対応スコアが高い頂点ペアほど対応している可能性が高いことから、ここでは貪欲法を用いて高い対応スコアを持つ頂点の対応を優先的に採用し、全頂点の対応を決定する。具体的な手順を以下に示す。

- (1) 各々の分子表面上のすべての頂点間の対応スコアを求め、対応スコア表に記述しておく。
- (2) 対応スコア表の中で最も高い値を持つ頂点どうしを対応頂点ペアとする。その対応頂点ペアを取り除き、対応スコア表の値を0とする。
- (3) (2)の操作を繰り返し、すべての対応頂点ペアを求める。

3.2.2 局所探索による頂点对応付けの改善

頂点の初期対応が決定されると、次に対応付けた頂点を交換することで D_c の値をより小さくする。すべての頂点の組合せを交換しながら分子表面データ間の距離を算出するのでは膨大な時間がかかるため局所的な探索を行う。

初期対応においては、対応スコアの高い順番に頂点对対応付けたため、一方のある頂点に対して、もう一方の頂点のうち対応スコアの最も高い頂点が必ずしも

選択されているわけではない．そこで対応スコアの低い頂点の対応から順番に別の頂点の対応との交換を試みる．このとき D_c の値が交換する前より小さくなればより適切な対応が得られたとして頂点ペアを交換する．具体的な手順を以下に示す．

- (1) 頂点ペアを対応スコアの低い順にソートし、番号を付ける．
- (2) 番号 1 の頂点ペアと番号 2 の頂点ペアを互いに交換する．
- (3) もし分子表面データ間の D_c の値が交換前より小さくなればその頂点ペアを採用し番号も付け直し、そうでなければもとの状態に戻す．
- (4) (2) において交換する頂点ペアを番号 (1,3), (1,4), ..., (2,1), (2,3), (2,4), ..., $(N-1, N)$ のように合計 $N(N-1)$ 回の頂点交換を試行する．このサイクルを反復し、最終的な距離を求める．ただし、 N は比較頂点数である．

3.3 距離を利用した類似蛋白質検索

バケット方式によって分子表面の 1 対 1 の比較を効率化することが可能となったが、データベース内の活性部位数が多くなると類似する活性部位を検索するのに時間がかかってしまう．そこで、分子表面間の距離 D が距離の公理を満たすことに着目し、入力に最も類似するデータベース内の活性部位を検索対象から除外することなく、検索対象を削減する方法を提案する．本検索方式の流れを図 6 に示す．

入力蛋白質が与えられると、まず活性部位をどのような順序で比較していくかを決定する．これは本検索方式が、できるだけ類似性の高い活性部位と早期に比較するほど検索対象の削減が可能となり、検索効率が向上するという特性を有するためである．本論文では、

比較する順位を決定するために、入力蛋白質とデータベース内の各活性部位との類似性を大雑把に求めることが可能な特徴領域抽出に基づく蛋白質表面比較方式¹³⁾を利用する．この方式は、類似蛋白質の分子表面データに、物性や形状の特徴がきわめて類似した部分（特徴領域）が多数存在することに着目し、抽出した特徴領域の比較により大まかな類似性評価を高速に実現するものである．この方式では必ずしも距離が最小となるものを第 1 位に選出することはできないが、比較的距離の近いものを上位に順序付けることが可能である．

次に、あらかじめ検索半径を無限大に設定したうえで、その順序に従って活性部位と比較し、そのつど検索半径と分割円の設定ならびに検索対象の除外を行う．検索半径とは、入力と、その時点ですでに比較されたそれに最も近い活性部位との分子表面データ間の距離である．図 7 に簡単のため距離空間を二次元上に描いた概念図を示す．検索半径を r 、新たに分子表面比較によって求めた活性部位との距離を d とすると、次の範囲にある活性部位は、三角不等式の定理によってそれまでに比較した活性部位よりも類似することはありえないため、検索対象から除外することが可能である．

- (i) 比較した活性部位から $(d+r)$ 以上の距離にある活性部位
- (ii) 比較した活性部位から $(d-r)$ 以内の距離にある活性部位

特に、入力と比較した活性部位が比較的類似している場合には (i) の範囲に存在する活性部位が多くなり、検索が効率化できる．また、入力と比較した活性部位があまり類似していない場合には、(ii) の範囲に存在する活性部位が多くなり、検索が効率化できる．

本検索方式の具体的な手順を以下に示す．

- (1) 特徴領域抽出に基づく蛋白質表面比較方式を

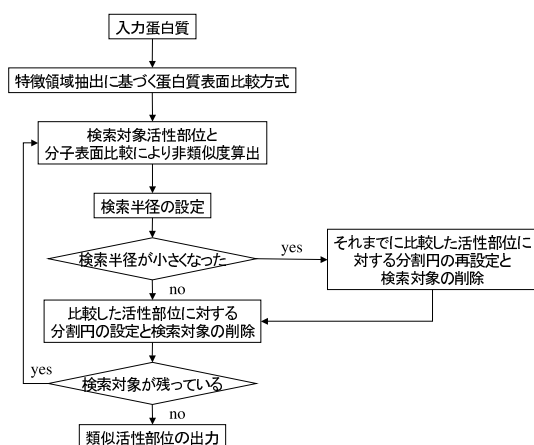


図 6 類似蛋白質検索の流れ
Fig. 6 The flow of retrieval.

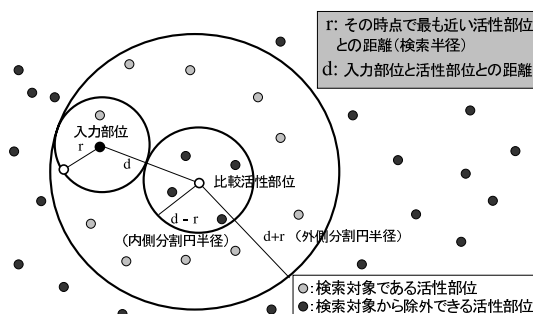


図 7 分割円の設定
Fig. 7 A setup of a division circle.

利用し, 入力活性部位 A_{in} とデータベース内の活性部位とを比較する. 類似性の高い順に並べたデータベース内の活性部位リストを $L = (A_1, A_2, A_3, \dots, A_n)$ とする. ただし, n はデータベース内の活性部位数である.

- (2) 検索半径 $r = \infty$, $A_x = A_1$ とする.
- (3) A_{in} と A_x とを比較し, その間の距離 d を求める. $d < r$ であれば, $r = d$, $A_{out} = A_x$ とする.
- (4) A_x からみて $(d + r)$ 以上の距離にある活性部位, および $(d - r)$ 以内の距離にある活性部位をリスト L から除外する.
- (5) リスト L が空であれば, 検索を終了して A_{out} を出力する. リスト L が空でなければ, L 内に存在する A_x の次に順位の高い活性部位を A_{next} とすると, $A_x = A_{next}$ としてステップ (3) へ戻る.

4. 評価実験

提案した類似蛋白質検索方式の有効性を検証するために, eF-site に登録されている酵素蛋白質の分子表面データを用いて評価実験を行う. なおここでの実験において用いた計算機は HIT 社 HPC-P4L (CPU: Intel Socket 423 Pentium 4 processor 1.7 GHz, Memory: 256 MB) である. また, 重みパラメータは距離・電位差・疎水性差のとりうる値の範囲が等しくなるように調整した値 $(W_l, W_e, W_h) = (1, 9, 18)$ とし, 1 軸あたりのバケット分割数を 80, 近傍バケットへの探索幅を 5 として実験を行った.

4.1 局所探索法における反復回数の評価

局所探索は合計 $N(N-1)$ 回 (N : 比較頂点数) の頂点交換の試行を 1 サイクルとし, このサイクルを反復することで最終的な距離の値を求める方法である. どれだけ反復するかを決定するため, 反復回数に応じて距離の値と比較が終了するまでの時間がどのように変化するかを調べる実験を行う. 実験対象は比較的良好な分子表面を持つ酵素蛋白質 5 組 (2lyo と 2lyz, 2csc と 2cts02, 1bpm と 1bpn, 1hel と 1hew, 1lmn と 1lmo) である. 局所探索をしない場合, 局所探索を 1 回する場合, 局所探索を 2 回反復する場合の 3 通りについて実験を行った結果を図 8 に示す.

図 8 より, 局所探索をしない場合に比べて局所探索をする場合にはより小さい距離の値を求めることができていくことが分かる. また, 局所探索が 1 回の場合と 2 回の場合ではほとんど距離の値が変化しないこと

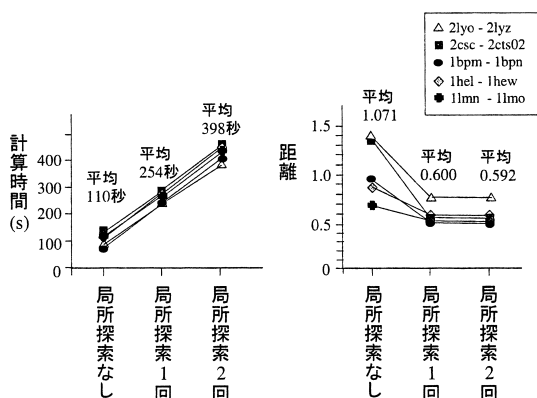


図 8 反復回数と距離・計算時間

Fig. 8 The repetition of the local search and the distance.

表 1 機能予測精度

Table 1 Functional prediction accuracy.

	正解 (個)	不正解 (個)	正解率 (%)
従来手法	250	30	89.3
本手法 (ダミー頂点無)	244	36	87.1
本手法 (ダミー頂点有)	267	13	95.4

が分かる. 一方, 局所探索の回数に比例して計算時間が多く必要となることを考慮すれば, 局所探索の繰返し回数は 1 回とするのが適当である. 本論文の以下の実験では, 反復回数をすべて 1 としている.

4.2 類似性判定結果の評価

データベース eF-site で公開されている酵素蛋白質のうち, 任意の 310 個を用いて機能予測実験を行う. ある蛋白質を入力としたとき, それ以外の 309 個を DB に格納し, 類似蛋白質の検索を行った. 入力には 310 個の蛋白質のうち同一機能を持つ蛋白質が自分以外に存在するもの 280 個を用いる. 同一機能を持つ蛋白質を検索結果として出力できた場合を正解とし, 従来手法^{9),10)}, 本手法でダミー頂点を用いない場合, 本手法でダミー頂点を用いる場合について, 正解の割合を比較する. 従来手法は分子表面を 3 次元空間上で重ね合わせて, 各頂点間の差を求めることで類似性を評価する手法であり, 算出される非類似度は距離の公理を満たさないが, 比較的精度良く類似性評価を行えることが確認されている. また, 本手法でダミー頂点を用いない場合, 頂点数が少ない側の活性部位に合わせて頂点の対応付けを行い, 対応付けされずに残った頂点数が多い活性部位の頂点については距離の計算に加味しないものとした. 実験結果を表 1 に示す.

表 1 より, 従来手法と本手法における機能予測の精度はそれほど変わらないことから, 本手法で定義した分子表面データ間の距離が妥当であることがいえる. ま

た、ダミー頂点を付加する場合の方がダミー頂点を付加しない場合に比べて精度が向上していることから、ダミー頂点が類似性評価に影響を与えない範囲で正しく定義されていることが分かる。

4.3 検索にかかる計算回数の評価

本論文では、入力が与えられたとき、データベースの中から最も類似している活性部位を検索することを目的としており、類似する可能性のない活性部位を検索対象から除外することで検索の効率化を図っている。ここでは酵素蛋白質 310 個の分子表面データを用いて、検索にかかる平均比較計算回数を、活性部位データをすべて検索に用いる逐次検索の場合と比べる。また検索結果が逐次比較の結果と一致しているかどうかについても検証する。なお、酵素蛋白質の分子表面データを使った場合、そのほとんどが 4~5 分の計算時間で比較可能であるため、平均比較計算回数によって検索速度を評価することが可能である。

表 2 に逐次比較および本手法について、同一機能蛋白質が自分以外にも存在する場合としない場合に分けて、310 個すべてを入力としたときの検索回数の総和と、蛋白質数で割った平均値を示す。表 2 より、同一機能蛋白質が自分以外にも存在する場合、逐次検索に比べて 11.6 倍速く検索可能であることが示された。また、存在しない場合にでも 5.6 倍速く検索可能である。このことから本方式が特に類似性の高い活性部位が活性部位データベース内に存在するときに検索効率が向上するという特性を示すことが分かる。

次に逐次検索で最も類似していると判定したものと、

表 2 平均比較計算回数

Table 2 The number of times of average comparison calculation.

	同機能蛋白質有	同機能蛋白質無
逐次比較	86520/280 309	9270/30 309
本手法 (ダミー頂点使用)	7448/280 26.6	1662/30 55.4

表 3 逐次検索と本手法による結果が異なる回数

Table 3 The number of times from which an output differs.

同機能蛋白質有	同機能蛋白質無
0/280	2/30

表 4 検索結果と分子表面データ間の距離

Table 4 Output and distance.

入力	逐次検索 (距離)	本手法 (距離)
2 pec	1 thl (1.181)	1 fps (1.185)
1 rmg	1 arx (0.924)	1 arp (0.968)

本手法によって効率的に検索した出力結果を比較する。表 3 に逐次検索で最も類似していると判定した結果を本手法によって出力できなかった数を示す。これより、同一機能蛋白質が自分以外にも存在する場合には完全に同じ結果が得られることが分かる。

表 3 において、結果が一致しなかった 2 例は、蛋白質 2 pec, 1 rmg を入力とした場合である。表 4 にそのときの分子表面データ間の距離を示す。表 4 より、お互いに比較的近い分子表面があったときに、必ずしも最小のものを選択できていないことが分かる。これは分子表面データ間の距離の値を近似的に求めているため、三角不等式が成り立たない場合があることに起因する。酵素蛋白質 310 個について任意の 3 個を取り出したときに、それらが三角不等式を満たすかどうかを網羅的に調べると、全組合せの 0.036% の場合において、三角不等式が成り立たなかった。このような問題を回避するためにはより厳密に分子表面データ間の距離を算出する必要がある。

4.4 データベース規模と計算時間の評価

類似蛋白質の検索を行う際には既知の蛋白質の活性部位をデータベースに登録しておく必要があるが、蛋白質の構造解明が進むにつれてその規模も徐々に大きくなっていくと考えられる。

そこでデータベースの規模を変化させたときに、平均比較計算回数がどのように変化するかを以下の 3 パターンに分けて調べた。ある 1 個の蛋白質を入力としたとき、それ以外の蛋白質の活性部位データを用いてデータベースを構築し、すべての蛋白質を入力としたときにその検索回数の平均を求める。

- A : 78 個 (同一機能を持つ蛋白質が自分以外に存在するもの 70 個)
- B : 155 個 (同一機能を持つ蛋白質が自分以外に存在するもの 140 個)
- C : 310 個 (同一機能を持つ蛋白質が自分以外に

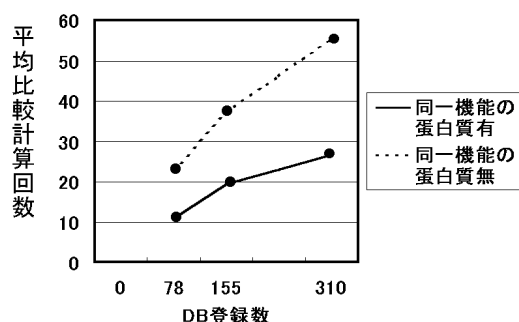


図 9 DB 規模に対する比較回数の変化

Fig. 9 DB scale—the number of comparison.

存在するもの 280 個)

各データベース規模に対する比較計算回数の変化を図 9 に示す。

図 9 より, 同一機能を持つ蛋白質が自分以外に存在するもの, および存在しないもの両方について, データベースの規模を拡大しても比較回数が線形に増加することがなく, 特に同一機能の蛋白質がある場合において比較回数の伸びが大きく鈍っていることが分かる。今後データベースの規模の拡大にともない各種機能の活性部位が網羅的に登録されることが期待でき, 入力蛋白質と同一機能を有する活性部位がデータベース内に存在する可能性が高くなると見込まれるため, 上記のような特性を持つ本手法の有用性が示唆される。

5. 結 論

本論文では, 類似蛋白質を検索するうえで, 活性部位データベースのすべてのデータと逐一類似性を評価するのは非効率であるという観点から, 分子表面データ間の距離を利用し, 効率的に類似蛋白質を検索する方式を提案した。2 つの蛋白質の分子表面上の頂点を 1 対 1 に対応付けることで, 分子表面データ間の距離を定義した。頂点の対応付けの組合せを網羅的に調べると膨大な計算時間がかかることから, 対応スコアとバケット法を導入することで初期対応を求め, 頂点ペアの交換により距離の改善を行った。分子表面データ間の距離を利用して, 類似する可能性のない検索対象を除外することで効率的な検索を可能にした。以上の 2 つの効率化手法(バケット法と検索対象の削減)による相乗効果でシステム全体としての効率化を実現した。また, 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, 分子表面比較回数を大幅に減らすことができ, また機能予測精度が十分高いことを確認し, 本手法の有効性を確認した。

謝辞 日頃よりご指導いただき薦田憲久教授に深謝する。また本研究に関してご議論いただいた中村春木教授に感謝する。本研究の一部は, 文部科学省科学研究費補助金および科学技術振興事業団バイオインフォマティクス推進事業からの助成による。

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, 中井謙太: バイオテクノロジーのためのコンピュータ入門, コロナ社 (1995).
- 3) Rosen, M., Lin, S.L., Wolfson, H. and Nussi-

nov, R.: Molecular shape comparisons in searches for active sites and functional similarity, *Protein Engineering*, Vol.11, No.4, pp.263-277 (1998).

- 4) Poirrette, A.R., Artymiuk, P.J., Rice, D.W. and Willett, P.: Comparison of protein surfaces using a genetic algorithm, *J. of Computer-Aided Molecular Design*, Vol.11, pp.557-569 (1997).
- 5) Ritchie, D.W. and Kemp, G.J.L.: Fast computation, rotation and comparison of low resolution spherical harmonic molecular surfaces, *J. Comp. Chem.*, Vol.20, pp.383-395 (1999).
- 6) Ankerst, M., Kastenmuller, G., Kriegel, H. and Seidl, T.: 3D shape histograms for similarity search and classification in spatial databases, *Proc. 6th Int. Symposium on Spatial Databases (SSD '99)*, Hong Kong, China, LNCS, pp.207-226, Springer Verlag, (1999).
- 7) Kriegel, H., Schmidt, T. and Seidl, T.: 3D Similarity search by shape approximation, *Proc. 5th Int. Symposium on Large Spatial Databases (SSD)*, Berlin, Germany, pp.11-28 (1997).
- 8) 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).
- 9) 兼田佳和, 庄治範匡, 大川剛直, 中村春木: 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, *情報処理学会論文誌*, Vol.43, No.1, pp.196-203 (2002).
- 10) Kaneta, Y., Shoji, N., Ohkawa, T. and Nakamura, H.: A method of comparing protein molecular surface based on normal vectors with attributes and its application to function identification, *Proc 4th Conference on Computational Biology and Genome Informatics*, Durham, NC, USA, pp.1213-1218 (2002).
- 11) Connolly, M.L.: Solvent-accessible surfaces of proteins and nucleic acids, *Science*, Vol.221, pp.709-713 (1983).
- 12) Mizuguchi, K. and Go, N.: Comparison of spatial arrangements of secondary structural elements in proteins, *Protein Eng.*, Vol.8, pp.353-362 (1995).
- 13) Nakagawa, T., Tanaka, T., Ohkawa, T. and Nakamura, H.: A filtering method for high-speed retrieval of similar active sites, *Proc. 4th Conference on Computational Biology and Genome Informatics*, Durham, NC, USA, pp.1207-1212 (2002).

(平成 14 年 4 月 1 日受付)

(平成 15 年 1 月 7 日採録)



中川 忠輔 (正会員)

昭和 52 年生。平成 12 年大阪大学工学部電子情報エネルギー工学科卒業。平成 14 年同大学大学院工学研究科情報システム工学専攻修士課程修了。同年(株)日立製作所に入社。



大川 剛直 (正会員)

昭和 38 年生。昭和 63 年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。現在同大学情報科学研究科マルチメディア工学専攻助教授。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。



兼田 佳和 (正会員)

昭和 53 年生。平成 13 年大阪大学工学部情報システム工学科卒業。同年同大学大学院工学研究科情報システム工学専攻修士課程。蛋白質の機能予測に関する研究に従事。
