蛋白質分子表面データを用いた特徴量の抽出と結合重要部位の推定

多森 凌太[1,a)]　深澤 葵[1,b)]　西出 亮[1,c)]　大川 剛直[1,d)]

概要：蛋白質は，主にポケットと呼ばれるくぼんだ部位で他の化合物と結合することが多く，その結合部位は，類似する化合物に結合する蛋白質同士で局所的に類似していることが知られている．この結合部位において，共通して存在する部分構造があれば，その構造が，その蛋白質における結合に重要な部位であると考えることが出来る．そこで，本論文では，ポケットにおける化合物と結合する際の重要な構造を推定するために，蛋白質の表面情報同士の距離に着目した新しい重要部位の推定手法を提案する．60 種の蛋白質に対して実験を行い，各蛋白質に対して重要な部位の抽出を行った結果，高い精度で重要部位が抽出でき，提案手法の有効性を示した．

Extraction of feature values using protein molecular surface data and prediction of remarkable structures

Abstract: Proteins often bind to other compounds at the concave site referred to as a pocket, and it is known that the binding site is locally similar with proteins binding to the same compound. In this paper, we propose a new method to predict the remarkable structures which has a common partial structure at this binding site, by focusing on the distance between the surfaces of proteins. Experiments were conducted with 60 kinds of proteins to extract remarkable structures for each protein. As a result, remarkable structures were extracted with high accuracy, and the effectiveness of the proposed method was confirmed.

1. はじめに

蛋白質の多くは，分子表面において低分子化合物 (リガンド) などと相互作用することで，様々な機能を発現している．その際，特にくぼんだ部位 (以下，ポケットと呼ぶ) で他の物質と結合する現象がしばしば見られ，これらの部位は，類似した物質に結合する蛋白質同士で局所的に類似していることが知られている [1], [2]．蛋白質の性質や相互作用するリガンドについて実験的な手法で解析を行うにはコストや時間がかかることから，計算機を用いた解析方法の開発が進められている．

本論文では，ポケットにおけるリガンドと結合する際の重要な構造 (以下，重要部位と呼ぶ) を推定する手法を提案する．類似リガンドに結合する類似した局所的な分子表面として，重要部位が抽出できると考え，提案手法では，ポケットの分子表面を特徴点の集合として表現し，あるポケットに類似する特徴点集合を多数の蛋白質から総当たり的に抽出する．そして，得られた複数の特徴点集合をもとに，バイクラスタリング処理により，多くの蛋白質に共通する局所的な特徴点集合を重要部位として発見する．特徴点集合の類似性を評価する際には，その特徴点がどのような周辺環境に置かれているかについて，特徴点からの三次元空間内での距離をもとに近傍領域を設定し，領域ごとに近傍のデータ点を参照して特徴量を作成する．提案手法では，特徴点周辺の形状と物性を反映した 3 種類の特徴量を導入し，また，ポケットについてより詳細な情報の比較を可能にする．

[1] 神戸大学システム情報学研究科
Kobe University Graduate School of System Informatics
[a)] r-tamori@cs25.scitec.kobe-u.ac.jp
[b)] aoi@cs25.scitec.kobe-u.ac.jp
[c)] nishide@port.kobe-u.ac.jp
[d)] ohkawa@kobe-u.ac.jp

2. 分子表面データを用いた特徴量の抽出と結合重要部位の推定

2.1 蛋白質のポケットの抽出

提案手法では，データベース eF-site[3] から取得できる

図 1　蛋白質ポケット.

efvet データを蛋白質の表面の性質を示した 3 次元画像点群データとして用いる．この efvet には，静電ポテンシャルと疎水性などの物性値が付与されており，蛋白質の構造を示す公共データベース Protein Data Bank(PDB)[4] の原子データを参照して計算された分子表面の性質を記述している．efvet では分子表面をポリゴンにより表現しているが，ポリゴンの各頂点（以下，ポリゴンデータ点あるいは単にデータ点）に対して，化学的な性質を付加して点群データを比較することは計算コストが大きすぎることから現実的ではない．また，同一リガンドに結合する部位は周辺部位を含めて完全一致するとは考えにくいことから，周辺の構造情報を比較し，類似性を評価しなければならない．

そこで，比較するデータの削減を目的にポケットの分子表面上のポリゴンデータ点のうち，近似球面の最大・最小曲率に基づいた抽出方法 [5] を用いて，特徴点を抽出する．このとき，ポケット内の特徴点のみを比較に用いるためにポケットの領域内に存在する特徴点のみを抽出する．ここでは，各蛋白質と結合するリガンドの位置情報を活用し，リガンドの結合部分から一定の距離内に存在するデータ点をポケット内の点として抽出する．蛋白質と結合するリガンドの結合した状態の位置情報は，PDB から入手できる．

ポケットが細長い場合を想定したモデルを図 1 に示す．左はポケットの中心から一定範囲内を抽出する手法，右はリガンドの位置情報を活用した手法で求めたポケットの範囲を色付き実線で示す．ポケットの中心から抽出した手法ではポケットの中心座標からの球で範囲を決定するため，このようなポケットが細長い場合や複雑な形状を持つ場合に赤色点線のような抽出できていない領域が存在する．一方，リガンドの位置情報を用いた手法では，結合している各原子から一定範囲を領域として切り出すので，適切にポケットを切り出すことができる．提案手法では，後者の手法を採用する．

2.2　特徴点に付加する特徴量

特徴点を比較 (マッチング) するために使用する特徴量については，次の 3 つの特徴量を導入する．

- 特徴点と近傍点のなす角度ヒストグラム
- 特徴点の接面と近傍点との高さ特徴ベクトル
- 近傍の性質を示す物性ヒストグラム

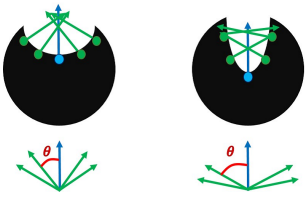

図 2　蛋白質モデルと法線ベクトル.

この 3 つの特徴量によって得られた特徴ベクトルを正規化し，マッチング処理の入力とする．以下で各特徴量について述べる．

2.2.1　特徴点と近傍点のなす角度ヒストグラム

法線ベクトルは，それぞれのデータ点において，蛋白質の外側垂直な方向を示す．この法線ベクトル同士のなす角は近傍点の表面と特徴点の表面がどれぐらい変化があるかを示している．2 つのベクトルがなす角が大きいほど近傍において表面の曲率が大きく，角が小さいほど変化がなく平面であることを示す．この特徴量を特徴点からの距離ごとに近傍点を分別してヒストグラムを作成することで，特徴点の周囲の形状を距離ごとに捉えることができる．このヒストグラムをここでは，「角度ヒストグラム」と呼ぶ．

この特徴量は，特徴点と近傍点の法線ベクトルを用いることで，求めることができる．特徴点の法線ベクトルを $\vec{n_a} = (n_{a1}, n_{a2}, n_{a3})$，近傍点の法線ベクトルを $\vec{n_b} = (n_{b1}, n_{b2}, n_{b3})$ とすると，2 つの法線ベクトルがなす角度 θ は，以下の式で求められる．

$$\theta = \arccos \frac{n_{a1}n_{b1} + n_{a2}n_{b2} + n_{a3}n_{b3}}{\sqrt{n_{a1}^2 + n_{a2}^2 + n_{a3}^2}\sqrt{n_{b1}^2 + n_{b2}^2 + n_{b3}^2}}$$

これで求められた θ は，必ず 0 から 180 の値をとる．提案手法では，近傍点全てと特徴点との全ての θ を一定の角度ずつ m 段階に分別し，さらに，2 点の距離から l 段階に分別する．計 $m \times l$ 段階のヒストグラムを作成する．この特徴量の概念をイメージするために，蛋白質表面を模した図形と法線ベクトルの例を図 2 として示す．ポケットの各表面の法線ベクトルだけを抜き出したものを図の下部に示す．青い点を特徴点として抽出しており，青い点と近傍の緑の点のそれぞれの法線ベクトルのなす角を求める．この 2 つの表面は，一見類似したポケットであるが，この類似度を用いることで，分別することができる．

一方，この類似度で似たヒストグラムを持つ蛋白質表面を模した図形の例を図 3 に示す．この 2 つの表面は，法線ベクトルのなす角度にのみに着目した, この特徴量のみを用いると，似た分布のヒストグラムを作成する．しかし，この 2 つの図は，明らかに区別されるべき形状であるが，この特徴量だけでは，分別することができない．そこで，この特徴量に加え，形状を示す特徴量として，特徴点の接面からの各データ点の距離を用いる．

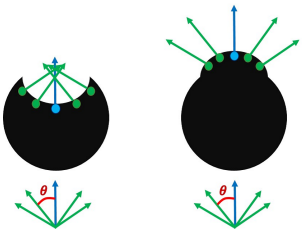

図 3　類似した角度ヒストグラムを持つ蛋白質モデル.

2.2.2　特徴点の接面と近傍点の距離

角度ヒストグラムにおいては，特徴点近傍の曲率の正負に関わらず，類似するヒストグラムが生成される可能性がある．

そこで，特徴点の接面からみたそれぞれのデータ点の高さを特徴量として導入する．特徴点の接面は，法線ベクトルから算出することができる．特徴点からみた近傍のデータ点の位置を，接面を基準として算出する．この値を特徴点とデータ点の距離ごとに分別して特徴ベクトルを作成することで，特徴点からみた距離ごとの標高の分布を捉えることが出来る．本手法では，この特徴量から得られる特徴ベクトルを「高さ特徴ベクトル」と呼ぶ．特徴点の法線ベクトル $\vec{n_a} = (n_{a1}, n_{a2}, n_{a3})$，特徴点の座標 $\vec{a} = (a_1, a_2, a_3)$，そして近傍点の座標 $\vec{b} = (b_1, b_2, b_3)$ としたとき，接面からみた近傍点の高さ h は以下の式で算出できる．

$$h = \frac{n_{a1}(b_1 - a_1) + n_{a2}(b_2 - a_2) + n_{a3}(b_3 - a_3)}{\sqrt{n_{a1}^2 + n_{a2}^2 + n_{a3}^2}}.$$

h は，大きい値をとるほど，接面を基準により膨らんだ位置にあることを示す．

提案手法では，2 点の距離から l 段階に分別し，その段階ごとに領域内の点の個数 k 個の点データの高さ h の平均値 $\bar{h} = \frac{1}{m}\Sigma_k h$ を特徴量とする．この特徴量は，l 次元の特徴ベクトルであり，ヒストグラムではないため，法線ベクトルのなす角を用いた特徴量と異なり，微小な値の違いを保持することが出来る．

この特徴量の概念を表現するために，蛋白質表面を模した図形と各近傍点の例と平均の高さ h のイメージを図 4 に示す．この特徴量は，特徴点から一定距離ごとの領域の，特徴点を基準とした近傍点の標高の平均が保持されている．この特徴量を用いることで，前節で導入した指標のみでは捉えられない，図 4 の 2 つの図形のポケットの表面を区別することが出来る．一方で，似た高さ特徴ベクトルを持つ表面を模した図を図 5 に示す．この 2 つの表面は，この特徴量では区別出来ないが，角度ヒストグラムを用いることで区別することが出来る．

2.2.3　物性を示す情報を用いた特徴量

物性を示す情報は，efvet から得られる情報であり，主に静電的あるいは疎水的な性質について表記されている．efvet に記述されている，静電ポテンシャルと疎水性の値を参照して，性質の特徴量を作成する，静電ポテンシャルの値は記述された値に応じて，c 段階に分類する．疎水性の値は Kyte らのアミノ酸疎水性指標 [6] に基づいて，d 種類に分類してヒストグラムを作成する．このときも，特徴点と近傍点の距離に応じて e 段階に分け，計 $c \times d \times e$ 次元のヒストグラムを作成する．本論文では，このヒストグラムを「物性ヒストグラム」と呼ぶ．この特徴量を用いることで，データ点距離ごとに性質を比較することができ，より詳細な比較が可能になる．

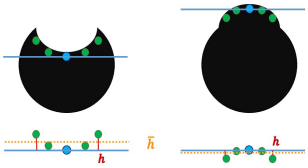

図 4　蛋白質モデルと特徴点の接面からの高さ.

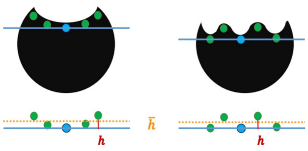

図 5　類似した高さ特徴ベクトルを持つ蛋白質モデル.

2.3　特徴点のマッチング処理

ポケットは特徴量が付与された特徴点の集合として表現される．ここでは，ポケット間の類似度を評価するため，特徴点を比較する．点同士の座標と特徴量を比較することで，特徴点の周辺が構造的かつ性質的に類似している点群同士を探索する．これにより，2 つのポケットを表現した特徴点群に対して，お互いのポケット間で類似しているとみなせる特徴点のペアの集合を高速に得ることができる．マッチングの手法については， Nishimura らの手法 [7] を汎用の CPU でシミュレートする．

提案手法では，特徴点の位置関係でマッチング候補に挙がった特徴点ペアにおいて，提案した特徴量を用いて，そのペアが類似しているかどうかを判定する．ここで，特徴点ペアが類似しているかどうかを判別するために特徴点間の類似度を定義する．特徴点ペアの特徴量ごと類似度をそれぞれ $s(1)$，$s(2)$，$s(3)$ としたとき，類似度 S を以下の式で定義する．

$$S = 1 - s(1) - s(2) - s(3).$$

角度ヒストグラムの類似度 $s(1)$，高さ特徴ベクトルの類似度 $s(2)$，物性ヒストグラムの類似度 $s(3)$ は，それぞれの重みを w_1，w_2，w_3 としたとき，それぞれ以下の式で定義する．

$$s(1) = \Sigma_{x=1}^{m*l} |P_1(x) - Q_1(x)|,$$

$$s(2) = \Sigma_{x=1}^{l} |P_2'(x) - Q_2'(x)|,$$

$s(3) = \Sigma_{x=1}^{c*d*e} |P_3(x) - Q_3(x)|$.

P は重要部位を推定したい蛋白質で，Q は比較対象の蛋白質である．$P_1(x)$，$P_2(x)$，$P_3(x)$，$Q_1(x)$，$Q_2(x)$，$Q_3(x)$ はそれぞれ P，Q における角度ヒストグラム，高さ特徴ベクトル，物性ヒストグラムの要素の値である．そして，全ての特徴点において，以下のように全ての特徴量の和が N になるように正規化されているものとする．

$$w_1 \Sigma_{x=1}^{m*l} P_1(x) + w_2 \Sigma_{x=1}^{l} P'_2(x) + w_3 \Sigma_{x=1}^{c*d*e} P_3(x) = N.$$

$$w_1 \Sigma_{x=1}^{m*l} Q_1(x) + w_2 \Sigma_{x=1}^{l} Q'_2(x) + w_3 \Sigma_{x=1}^{c*d*e} Q_3(x) = N.$$

また，$P'_2(x)$，$Q'_2(x)$ については，以下の式で示す．

$$P'_2(x) = 1 + \frac{P_2(x)}{dis(x)}.$$

$dis(x)$ は，距離のビン x における特徴点と近傍点群との平均距離である．これにより $P'_2(x)$，$Q'_2(x)$ は，必ず 0 から 2.0 の値を取る．

このとき，S がある閾値以上であれば，その特徴点のペアは，類似しているものとみなす．

2.4 バイクラスタリング処理を用いた重要部位の推定

類似したリガンドに結合する多くのポケットに対して，類似する特徴点集合が見られるとき，その特徴点は，そのリガンドに対する結合に際して重要な役割を果たす部位に対応していると考えることができる．そこで，ある蛋白質ポケットを構成する特徴点集合に対して，それと類似する特徴点集合を多数のポケットから抽出し，バイクラスタリング処理を行うことで，重要部位の抽出を実現する．重要部位を推定するためのそこで，重要部位を推定するためのバイクラスタリングの手法については，Nishimura らが提案した BISERS[7] を用いる．

3. 実験と結果及び考察

3.1 実験設定

評価実験には 60 種類の蛋白質を用いる．それぞれの蛋白質は，15 種類のリガンドのいずれかに結合することがすでに知られている蛋白質であり，1 つのリガンドに対して 4 つの結合する蛋白質が用意されたデータセットである．解析する蛋白質は，データセット内の他の全ての蛋白質と網羅的にマッチング処理を行う．計 60×59 回のマッチング処理により，60 種類の蛋白質と他の蛋白質を比較し，その結果を用いて，バイクラスタリング処理を行う．

提案手法の有効性を示すために，Nishimura らの手法 [7] を比較手法とし，同じ実験を行い，それぞれの重要部位について評価する．なお，提案手法のパラメータは，以下のよう設定する．

- ポケット抽出の際の切り出し距離 $dis = 7.5$Å.
- 角度ヒストグラムのパラメータ $m = 5, l = 10$.

The ligand binding site extracted from coordinates of「HETATM」atoms		
site_id	numbers of residues	detail
FMN_1bwk_A_401	29	FLAVIN MONONUCLEOTIDE binding site
chain	residue	ligand
A	PRO34-ARG38	FMN: FLAVIN MONONUCLEOTIDE
A	GLU71-GLY72	FMN: FLAVIN MONONUCLEOTIDE
A	GLN114	FMN: FLAVIN MONONUCLEOTIDE
A	TRP116	FMN: FLAVIN MONONUCLEOTIDE

図 6 リガンド結合部位 “HETATM” データ [7].

表 1 実験結果.

	平均 precision
比較手法 [7]	0.701
提案手法	0.705

- 高さ特徴ベクトルのパラメータ $l = 10$.
- 物性ヒストグラムのパラメータ $c = 4, d = 3, e = 5$.
- マッチング時の重みパラメータ $w_1 = 1, w_2 = 5, w_3 = 1$.
- マッチング時の類似度の閾値 $S = 0.60$.
- マッチング時の正規化の基準値 $N = 1.0$.

重要部位推定の評価として，PDB において結合に関与すると記録されている残基の情報を利用する．結合に関与する残基は，蛋白質や核酸の構成原子以外の原子，通称 “HETATM” が 5Å 以内に存在する残基であり, 図 6 のように記録されている．これを正解データとする．そして，重要部位を構成する各特徴点に対して，その点に空間的に最も近い残基が，上記の「結合に関与する残基」であれば，特徴点が正しく抽出できていると判断する．重要部位を構成する特徴点のうち，正しく抽出された特徴点の割合（precision）を算出し，手法の評価を行う．

3.2 実験結果と考察

実験結果から比較手法 [7] と提案手法を比較する．両手法の評価を行うために，precision の平均値を表 1 に示す．

この表に示すとおり，precision については，平均値は比較手法とあまり有意な変化は見られなかった．そこで，横軸に推定部位の点の個数，縦軸を precision とした散布図を図 7 に示す．オレンジの点，青の点はそれぞれ比較手法，提案手法の各蛋白質の実験結果を示している．

このように比較手法においては，抽出される特徴点の個数が圧倒的に少なく，多量の正解データのなかで，少量のデータのごく一部を正しく抽出できており，これが precision の値が高い理由である．これに比べ，提案手法は，比較手法を大きく上回る量の特徴点を重要部位として抽出している．これより，提案手法は重要部位の推定において，比較手法よりも良い結果を示していると考えることが出来る．

一方で，極端に precision が低い蛋白質も存在する．その原因として，リガンド側の結合部位が異なることが挙げら

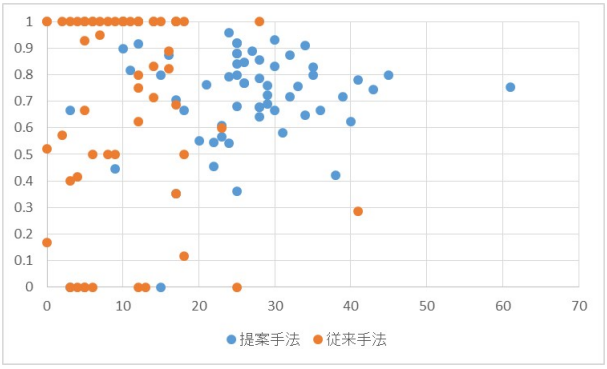

図 7 推定部位の点の個数と precision.

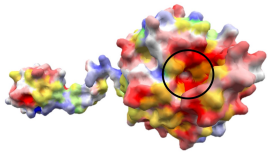

図 8 蛋白質 1xuz-A.

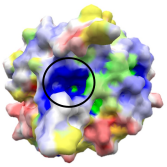

図 9 蛋白質 3fyo-D.

れる．具体例として，同じリガンドに結合する異なる 2 つの蛋白質を図 8，9 に示す．この 2 つは，図から見ても明らかであるように，表面の性質が異なっている．それぞれのポケットにおいて，同じリガンドに結合したとしても，リガンド側が同じ面で結合しているとは限らない．したがって，更なる精度の向上を図るには，同じリガンドに結合する蛋白質の中でもリガンド側の結合部位に応じて，区別して，バイクラスタリングを行う必要がある．また，同様にリガンド側の結合部位の解析も不可欠である．

4. おわりに

本論文では，蛋白質の表面情報を考慮した特徴量の抽出を行い特徴点を比較することで，蛋白質とリガンドとの結合に関する重要部位を推定する手法を提案した．

提案手法では，法線ベクトルと位置座標を用いた形状的な特徴量とアミノ酸の疎水性と分子表面データの静電ポテンシャルを用いた特徴量を距離ごとに分別して作成することで，特徴点からの距離に応じて変化する性質を分別して保持でき，より詳細な情報の比較を可能にした．

提案手法の有効性を示すために，重要部位の抽出実験を行った結果，precision の値を保持しながら，比較手法の約 2 倍の特徴点を重要部位として推定することが出来た．これにより，提案手法は，比較手法に比べて，より多くの重要部位を精度を保持した上で抽出できていることを示した．

提案手法における課題として，同一リガンドに結合する各蛋白質のポケットは，そのリガンドの結合状態や結合する部位が異なっていたとしても考慮されないことがある．これは蛋白質表面のみの情報では捉えることが出来ないことから今後は，リガンド側の構造・性質についての解析が必要となる．

参考文献

[1] 藤 博幸，“タンパク質の立体構造入門”，講談社 (2010).
[2] 藤 博幸，“はじめてのバイオインフォマティクス”，講談社 (2006).
[3] 木下 賢吾，中村 春木，“タンパク質分子表面形状と物性のデータベース eF-site による分子機能類似性検索”，生物物理，Vol.42，No.1，pp.20-23 (2002).
[4] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, “The Protein Data Bank”, Nucleic Acids Research, Vol.28, pp.235-242 (2000).
[5] H. T. Ho, D. Gibbins, “A Curvature-based Approach for Multi-scale Feature Extraction from 3D Meshes and Unstructured Point Clouds”, IET Computer Vision, Vol.3, No.4, pp.201-212 (2009).
[6] J.Kyte , R.F.Doolittle, “A simple method for displaying the hydropathic character of a protein”, Journal of Molecular Biology, Vol.157, Issue 1, pp.105-132 (1982).
[7] H. Nishimura, T. Ohkawa, “A New Biclustering Algorithm with Exclusive Random Selection of Columns for Predicting Recognition Spots on Protein Molecular Surfaces”, International Journal of Bioscience, Biochemistry and Bioinformatics , Vol. 8, No.1, pp.11-19(2018).