Vol.2016-MPS-108 No.34
Vol.2016-BIO-46 No.34
2016/7/6

# 蛋白質分子表面マッチングと項目集合からの排他的選択を考慮したバイクラスタリングを用いた重要特徴点抽出

西村 宏人[1,a)]　阪上 絢人[1]　大川 剛直[1,b)]

**概要**：蛋白質が他の化合物と結合する部位には，結合現象に大きく寄与する特徴的な局所部分が存在する．本稿では，蛋白質分子表面データから，このような重要部位を特定する手法 BISES (BIclustering based on Similarity and Exclusive Selection of column) を提案する．提案手法では，蛋白質分子表面を三次元の特徴点集合として表現し，重要特徴点を特定したい蛋白質（クエリ蛋白質）とそれ以外の多数の蛋白質（参照蛋白質）のマッチング結果から，特徴点の対応関係を示すバイナリ行列を生成する．得られたバイナリ行列に対して，列間類似度と項目集合からの排他的選択を考慮したバイクラスタリング手法を用いることで，重要特徴点の抽出を実現する．提案手法の有効性を複数のタンパク質からなるデータセットを用いて確認した．

# Extraction of hotspots from protein molecular surfaces by biclustering based on simrality between proteins and exclusive selection of column

Hiroto Nishimura[1,a)]　Kento Sakaue[1]　Takenao Ohkawa[1,b)]

## 1. はじめに

多くの蛋白質は他の化合物との結合により，機能することが知られている [1][2]．結合現象が観察される部分を結合部位と呼び，その形状や物性が結合現象に深く関係すると考えられている．そこで，本研究では，結合部位の中から結合現象に直接的に影響する重要部位（ホットスポット）があると考え，ホットスポットを予測する手法の開発を目的とする．

提案手法では，類似する化合物に結合する蛋白質は類似する構造を持つこと，また，ホットスポットと呼ばれる重要部位は，同様の結合現象において頻出して観察されることに着眼し，三次元点群データとして表現した分子表面を網羅的にマッチングし，その結果から類似する化合物に結合する蛋白質に共通して現れる部分構造を抽出することにより，ホットスポットの特定を実現する．すなわち，蛋白質の分子表面データを構成する頂点の中でも，より構造の特徴を表す頂点（特徴点）を抽出して三次元点群画像として表現する．そして，解析したい蛋白質（クエリ蛋白質）と過去に解析された多数の蛋白質（参照蛋白質）の間でパターンマッチングを行うことにより，クエリ蛋白質と参照蛋白質間で対応する特徴点のリストを得ることができる．このとき，ホットスポットの特徴である「多数の蛋白質間で類似して観測されるひとまとまりの点群領域」を抽出することは，クエリ蛋白質の特徴点を行に，参照蛋白質を列に配置して対応点の存在を表現したバイナリ表（クエリ特徴点対応表）からの極大バイクラスタ抽出問題と捉えることができる．現在，多数のバイクラスタリング手法が考案されている [3][4][5][6][7] が，クエリ特徴点対応表を対象としたバイクラスタリングでは，列が様々な蛋白質に対応しており，結合部位に関する既知の知見が利用可能であること，各参照蛋白質に対して，最も多くの特徴点対応が取れたマッチング結果だけでなく，次点以降の複数の結果を表す列が含まれていることを考慮する必要がある．しかしながら，これらの点を考慮したバイクラスタリング手法は見

[1] 神戸大学大学院システム情報学研究科
1-1 Rokkodai, Nada, Kobe, Hyogo 657–8501, Japan
[a)] n.hiroto@cs25.scitec.kobe-u.ac.jp
[b)] ohkawa@kobe-u.ac.jp

Vol.2016-MPS-108 No.34
Vol.2016-BIO-46 No.34
2016/7/6

当たらない．

そこで，列間類似度と，列のグループ化とグループ（項目集合）からの排他的選択を導入することにより，重要部位の抽出に適したバイクラスタリング手法 BISES (BIclustering based on Similarity and Exclusive Selection of column) を提案する．提案手法では，バイクラスタの評価値を列間類似度によって重み付けすることにより，類似する結合部位を持つ蛋白質に共通する重要部位をより効果的に抽出する．また，同一の参照蛋白質のマッチング結果を表す複数の列をグループとしてまとめ，バイクラスタに新たな列を追加する際には，グループからは単一の列のみを選択し，追加する．これにより，同一の参照蛋白質を包含するような意味の無いバイクラスタの拡張を回避する．

## 2. 蛋白質分子表面マッチング

### 2.1 分子表面データ

本研究で利用する分子表面データは蛋白質分子表面データベース eF-site[8] から取得可能な efvet という XML ファイルを元に作成する．efvet には，分子表面を構成する頂点群が座標情報や静電ポテンシャルなどの物性情報と共に記されている．

これらの頂点群の中から，より構造の特徴を表す点 (特徴点) を抽出する．この処理を行う理由として，分子表面データは巨大で，蓄積されたデータ量も莫大であることからマッチング処理のコストを削減する必要があること，また，蛋白質の結合の特性上，大まかな構造の類似性を見る必要があることが挙げられる．

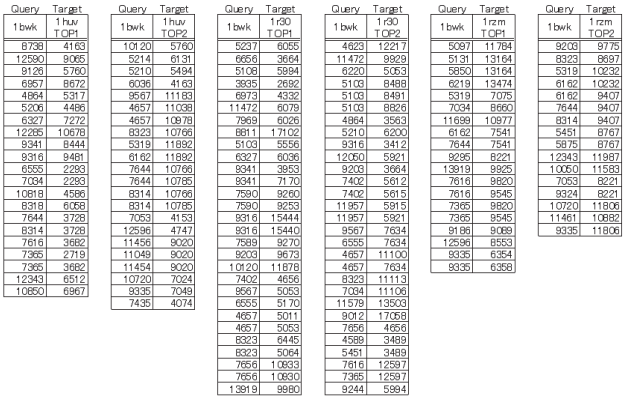
三次元画像からの特徴点の抽出手法として，SIFT アルゴリズム [9] や SURF アルゴリズム [10] があるが，蛋白質の構造が特徴的な凹凸を持つこと，また，スケールにより特徴が変化することを考慮するために，曲率に着目した曲面フィッティングによる特徴点抽出手法 [11] を利用する．この特徴点抽出によって得られた特徴点の情報を特徴点データと呼び，マッチングの入力データとして用いる．

### 2.2 特徴点マッチング

本研究では，蛋白質を構成する特徴点のマッチングに Set Operating Processor(SOP)[12] を利用する．SOP は文字列や二次元画像などの高速パターンマッチングに特化したメモリ型プロセッサである．入力として探し出したいパターンを与えると，SOP 内部で超並列に集合演算を行い，探索対象内で同様なパターンが現れる箇所が出力される．

具体的には，入力として蛋白質 $P$ の特徴点データ $\mathbf{p} = \{p_i | 1 \leq i \leq m\}$（クエリパターン）を SOP に与え，蛋白質 $Q$ の特徴点データ $\mathbf{q} = \{q_j | 1 \leq j \leq n\}$（参照パターン）に対してマッチングを行う．処理の詳細は省略するが，SOP により，クエリパターンの一部分 $\mathbf{p'} = \{p'_k | 1 \leq k \leq l\} \subseteq \mathbf{p}$ と参照パターンの一部分 $\mathbf{q'} = \{q'_k | 1 \leq k \leq l\} \subseteq \mathbf{q}$ との対応関係を瞬間的，かつ網羅的に発見することができる．ここで出力されるクエリ-参照パターン間で対応する特徴点のペアのリスト $\{(p'_k, q'_k) | 1 \leq k \leq l\}$ を特徴点対応関係と呼ぶ．また，クエリ-参照パターン間のマッチングは網羅的に行われるため，複数のマッチング結果をもとに，複数の特徴点対応関係を出力する．対応する特徴点数が多いほど，より適切なマッチングと考えられるため，全特徴点対応関係の中で対応するクエリの特徴点の数が最も多いものを TOP1 とし，以降多い順に TOP2，TOP3，... と表記する．

クエリパターンを蛋白質 1bwk，参照パターンを蛋白質 1huv, 1r30, 1rzm とした時の特徴点対応関係 TOP1, TOP2 の例を**図 1** に示す．左側はクエリパターンの対応する特徴点，右側は参照パターンの対応する特徴点を示す．数字は efvet データ内で用いられる頂点の ID であり，対応する特徴点 ID を示している．



| Query 1bwk | Target 1huv TOP1 |
|---|---|
| 8738 | 4163 |
| 12580 | 9065 |
| 9126 | 5760 |
| 6957 | 8672 |
| 4864 | 5317 |
| 5206 | 4486 |
| 6327 | 7272 |
| 12285 | 10678 |
| 9341 | 8444 |
| 9316 | 9481 |
| 6555 | 2293 |
| 7034 | 2293 |
| 10818 | 4586 |
| 8318 | 6058 |
| 7644 | 3728 |
| 8314 | 3728 |
| 7616 | 3682 |
| 7365 | 2719 |
| 7365 | 3682 |
| 12343 | 6512 |
| 10650 | 6967 |

| Query 1bwk | Target 1huv TOP2 |
|---|---|
| 10120 | 5760 |
| 5214 | 6131 |
| 5210 | 5494 |
| 6036 | 4163 |
| 9567 | 11183 |
| 4657 | 11038 |
| 4657 | 10978 |
| 8323 | 10766 |
| 5319 | 11892 |
| 6162 | 11892 |
| 7644 | 10766 |
| 7644 | 10765 |
| 8314 | 10766 |
| 8314 | 10765 |
| 7053 | 4153 |
| 12596 | 4747 |
| 11456 | 9020 |
| 11049 | 9020 |
| 11454 | 9020 |
| 10720 | 7024 |
| 9335 | 7049 |
| 7435 | 4074 |

| Query 1bwk | Target 1r30 TOP1 |
|---|---|
| 5237 | 6055 |
| 6656 | 3664 |
| 5108 | 5994 |
| 3935 | 2692 |
| 6973 | 4332 |
| 11472 | 6079 |
| 7969 | 6026 |
| 8811 | 17102 |
| 5103 | 5556 |
| 6327 | 6036 |
| 9341 | 3953 |
| 9341 | 7170 |
| 7590 | 9260 |
| 7590 | 9253 |
| 9316 | 15444 |
| 9316 | 15440 |
| 7589 | 9270 |
| 9203 | 9673 |
| 10120 | 11878 |
| 7402 | 4656 |
| 9567 | 5053 |
| 6555 | 5170 |
| 4657 | 5011 |
| 4657 | 5053 |
| 8323 | 6445 |
| 8323 | 5064 |
| 7656 | 10833 |
| 7656 | 10830 |
| 13919 | 9980 |

| Query 1bwk | Target 1r30 TOP2 |
|---|---|
| 4623 | 12217 |
| 11472 | 9929 |
| 6220 | 5053 |
| 5103 | 8498 |
| 5103 | 8491 |
| 5103 | 8826 |
| 4864 | 3563 |
| 5210 | 6200 |
| 9316 | 3412 |
| 12050 | 5921 |
| 9203 | 3664 |
| 7402 | 5612 |
| 7402 | 5615 |
| 11957 | 5915 |
| 11957 | 5921 |
| 9567 | 7634 |
| 6555 | 7634 |
| 4657 | 11100 |
| 4657 | 7634 |
| 8323 | 11113 |
| 7094 | 11106 |
| 11578 | 13503 |
| 9012 | 17058 |
| 7656 | 4656 |
| 4589 | 3489 |
| 5451 | 3489 |
| 7616 | 12597 |
| 7365 | 12597 |
| 9244 | 5984 |

| Query 1bwk | Target 1rzm TOP1 |
|---|---|
| 5097 | 11784 |
| 5131 | 13164 |
| 5850 | 13164 |
| 6219 | 13474 |
| 5319 | 7075 |
| 7034 | 8650 |
| 11699 | 10977 |
| 6162 | 7541 |
| 7644 | 7541 |
| 9295 | 8221 |
| 13919 | 9925 |
| 7616 | 9820 |
| 7616 | 9545 |
| 7365 | 9820 |
| 7365 | 9545 |
| 9186 | 9089 |
| 12596 | 8553 |
| 9335 | 6354 |
| 9335 | 6358 |

| Query 1bwk | Target 1rzm TOP2 |
|---|---|
| 9203 | 9775 |
| 8323 | 8697 |
| 5319 | 10232 |
| 6162 | 10232 |
| 6162 | 9407 |
| 7644 | 9407 |
| 8314 | 9407 |
| 5451 | 8767 |
| 5875 | 8767 |
| 12343 | 11987 |
| 10050 | 11583 |
| 7053 | 8221 |
| 9324 | 8221 |
| 10720 | 11806 |
| 11461 | 10882 |
| 9335 | 11806 |

**図 1** クエリ-参照パターン間のマッチング例

### 2.3 クエリ特徴点対応表

入力とする 1 つのクエリに対して，多数の蛋白質の 1 つ 1 つを参照蛋白質として入れ替えながら SOP によるパターンマッチングを行い，参照蛋白質毎に図 1 に示すような特徴点対応関係を出力する．

本研究の目的は，類似する結合物質を持つ蛋白質内で頻出する重要特徴点の抽出である．特に，類似する結合物質を持つ参照蛋白質との比較において，頻繁に対応が得られる特徴点は重要特徴点であると考えられる．そこで，クエリの各特徴点がそれぞれの参照蛋白質との比較において対応が得られたか否かを 1 または 0 で表したクエリ特徴点対応表を作成する．

図 1 の特徴点対応関係から，クエリ特徴点対応表を作成した例を**図 2** に示す．クエリ特徴点対応表の行ラベルはクエリの特徴点の ID を示し，列ラベルは参照蛋白質 ID とパターンマッチの順位を示す．TOP1 だけではなく，TOP2 以降も利用する理由としては，TOP1 が圧倒的に対応数が

多いわけではなく，TOP2 以降により適切な対応関係がある可能性があるからである．

| | | Target protein & the rank of pattern match | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1huv TOP1 | 1huv TOP2 | 1r30 TOP1 | 1r30 TOP2 | 1rzm TOP1 | 1rzm TOP2 |
| Query 1bwk feature points ID | 3935 | 0 | 0 | 1 | 0 | 0 | 0 |
| | 4589 | 0 | 0 | 0 | 1 | 0 | 0 |
| | 4623 | 0 | 0 | 0 | 1 | 0 | 0 |
| | 4657 | 0 | 1 | 1 | 1 | 0 | 0 |
| | 4864 | 1 | 0 | 0 | 1 | 0 | 0 |
| | 5097 | 0 | 0 | 0 | 0 | 1 | 0 |
| | 5103 | 0 | 0 | 1 | 1 | 0 | 0 |
| | 5108 | 0 | 0 | 1 | 0 | 0 | 0 |
| | 5131 | 0 | 0 | 0 | 0 | 1 | 0 |
| | 5206 | 1 | 0 | 0 | 0 | 0 | 0 |
| | 5210 | 0 | 1 | 0 | 1 | 0 | 0 |
| | 5214 | 0 | 1 | 0 | 0 | 0 | 0 |
| | 5237 | 0 | 0 | 1 | 0 | 0 | 0 |
| | 5319 | 0 | 1 | 0 | 0 | 1 | 1 |
| | 5451 | 0 | 0 | 0 | 1 | 0 | 1 |
| | 5850 | 0 | 0 | 0 | 0 | 1 | 0 |
| | 5875 | 0 | 0 | 0 | 0 | 0 | 1 |
| | 6036 | 0 | 1 | 0 | 0 | 0 | 0 |
| | 6162 | 0 | 1 | 0 | 0 | 1 | 1 |
| | 6219 | 0 | 0 | 0 | 0 | 1 | 0 |
| | 6220 | 0 | 0 | 0 | 1 | 0 | 0 |
| | 6327 | 1 | 0 | 1 | 0 | 0 | 0 |
| | 6555 | 1 | 0 | 1 | 1 | 0 | 0 |
| | 6656 | 0 | 0 | 1 | 0 | 0 | 0 |
| | 6957 | 1 | 0 | 0 | 0 | 0 | 0 |
| | 6973 | 0 | 0 | 1 | 0 | 0 | 0 |
| | 7034 | 1 | 0 | 0 | 1 | 1 | 0 |
| | 7053 | 0 | 1 | 0 | 0 | 0 | 1 |
| | 7365 | 1 | 0 | 0 | 1 | 1 | 0 |

⋮

**図 2** 図 1 から生成されたバイナリ表

## 3. 項目集合からの排他的選択を考慮したバイクラスタリング

重要特徴点の抽出は，クエリ特徴点対応表から複数のクエリの特徴点が，複数の参照蛋白質との比較において対応しているクラスタ，すなわちバイクラスタの発見により，実現される．クエリ特徴点対応表におけるバイクラスタリングでは，単純に極大バイクラスタを抽出するのではなく，列間類似度を反映し，排他的な列選択を行いながらバイクラスタの行や列を追加・削除 (以下，更新と呼ぶ) することが求められる．そこで，行や列の相関に基づく評価関数でバイクラスタを評価し，近傍探索を行い，段階的により良いバイクラスタへと更新するバイクラスタリング手法 PDNS(Pattern-Driven Neighborhood Search)[7] に着目する．そして，PDNS をもとに，

(a) 列間類似度の導入
(b) 列に対する項目集合の概念の導入と項目集合からの排他的な列選択

という 2 点の拡張を実施したアルゴリズム BISES (BIclustering based on Similarity and Exclusive Selection of column) を提案する．

### 3.1 列間類似度の導入

一般的なバイクラスタリングでは，全ての行，全ての列を対等に扱い，行間，列間の関係性は考慮しない．一方で，本研究で用いるクエリ特徴点対応表の列は参照蛋白質である．類似する化合物に結合する蛋白質の構造は類似する性質から，類似する化合物に結合する蛋白質同士の列間の関係性を考慮する必要がある．そこで，このような蛋白質間で優先的にバイクラスタを形成するため，列間類似度を導入する．

PDNS ではバイクラスタ $B(I', J')$（$I'$ は行の集合，$J'$ は列の集合）の評価値を計算する際に，全行相関値平均 $\frac{2\Sigma_{i\in I'}\Sigma_{j\in I', j\geq i+1}\ \rho_{ij}}{|I'|(|I'|-1)}$ と全列相関値平均 $\frac{2\Sigma_{k\in J'}\Sigma_{l\in J', l\geq k+1}\ \rho_{kl}}{|J'|(|J'|-1)}$ の内，値が大きい方を評価値とする ASR (Average Spearman's Rho) を導入している．ここで，$\rho_{ij}(i \neq j)$ はバイクラスタの $i$ 番目の行と $j$ 番目の行の相関を示すスピアマンの順位相関係数 [13] であり，$\rho_{kl}(k \neq l)$ はバイクラスタの $k$ 番目の列と $l$ 番目の列の相関を示すスピアマンの順位相関係数である．これをもとに，列間類似度が高いほど列相関値が高くなるように，列相関値を列間類似度により重み付けすることを考える．このとき，全行相関値平均と全列相関値平均を対等に扱うことができなくなるため，全行相関値平均と全列相関値平均を掛けた新しい評価値 ASR-CS（ASR-with Column Similarily）を導入する．ASR-CS では，列相関値が列間類似度により補正されるため，列間類似度が高い列同士を含む程，バイクラスタの評価値を高めることができる．$i$ 列目と $j$ 列目の列間類似度を $\mathrm{ColSim}(i,j)(0 \leq \mathrm{ColSim}(i,j) \leq 1)$ とする時，ASR-CS $(I', J')$ を次式で定義する．

$$\text{ASR-CS}\,(I', J') = \frac{2\Sigma_{i\in I'}\Sigma_{j\in I', j\geq i+1}\ \rho_{ij}}{|I'|\,(|I'|-1)} \frac{2\Sigma_{k\in J'}\Sigma_{l\in J', l\geq k+1}\ (\rho_{kl}\ \mathrm{ColSim}(k,l))}{|J'|\,(|J'|-1)}$$

### 3.2 項目集合からの排他的選択

クエリ特徴点対応表において，同じ参照蛋白質に対する異なるマッチング結果，TOP1 から TOP$t$ が列となっている．そこで，同じ蛋白質に対する結果を項目集合 (グループ) とし，排他的選択を行うことにより，同一蛋白質に対応する列が 1 つのバイクラスタに含まれることを回避する．

更新前のバイクラスタに列の更新を行うと，同一グループの複数の列を含む可能性がある．同一グループの列を 2 列以上含む時，同一グループの列を含む列集合と含まない列集合に分ける．同一グループの列を含む列集合において，グループから排他的に選択した列の全組合せに分岐し，各分岐先に同一グループの列を含まない列集合を追加する．各分岐先において行を更新し，更新されたバイクラスタを評価し，最も評価が高いバイクラスタを選択する．選択したバイクラスタに対して再び列の更新を行い，繰り返す．なお，列の更新後の同一グループの列を含む列集合において，行の更新後に選択されたバイクラスタに含まれない列は，同じ処理を省くために，その後の列の更新において追加しないようにする．

Vol.2016-MPS-108 No.34
Vol.2016-BIO-46 No.34
2016/7/6

それぞれ 2 列から成るグループ G1，G2，G3 で構成される行列に対して，排他的列選択を行いながらバイクラスタを更新する処理の例を図 **3** に示す．

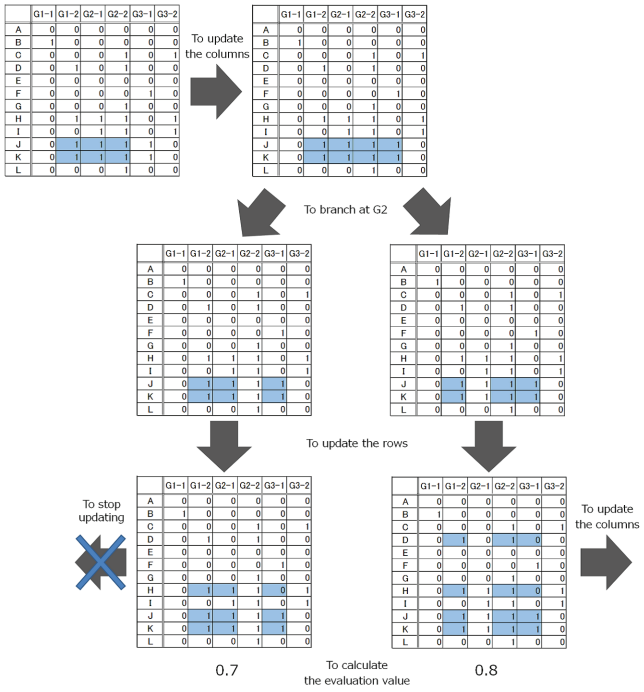



**図 3** バイクラスタの更新

以上の考えに基づくバイクラスタリング手法 BISES のアルゴリズムを Algorithm1 に示す．5 行目の $b \oplus mv_c(\alpha)$ はバイクラスタ $b$ において列を更新（1 を含む割合が $\alpha$ 以上である列のみを残し，さらに他の列において 1 を含む割合が $\alpha$ 以上である列を追加する）する処理を表し，また，10 行目の $b \oplus mv_r(\beta)$ は行を更新 (1 を含む割合が $\beta$ 以上である行のみを残し，さらに他の行において 1 を含む割合が $\beta$ 以上である行を追加する) する処理を意味している．20 行目は，得られた最良バイクラスタ $B^*$ に対し，構成する行と列を割合 $R$ だけランダムに変化させる処理であり，変化後のバイクラスタに対し，再度更新を繰り返す．

**Algorithm 1**

**Input:** $M$:data matrix, $B_0$:initial bicluster, $\alpha$, $\beta$:quality threshold, $Y$, $Z$:maximum number of iterations, $R$:a rate of perturb
**Output:** $B^*$:the best bicluster

1: $C \leftarrow \emptyset$
2: $b \leftarrow B_0$
3: **while** until we reach the maximum number of iterations $Z$ **do**
4: **while** until we reach the maximum number of iterations $Y$ **do**
5: $b' \leftarrow b \bigoplus mv_c(\alpha)$
6: Remove $c \in \{C\}$ from $b'$
7: Divide $b'$ into $b'_{group}$ having the same group's colmuns and $b'_{nongroup}$ not having the same group's columns
8: Branch to all combinations $D = \{d_1, d_2, \ldots, d_n\}$ of exclusively selected columns from the group for $b'_{group}$, and add $b'_{nongroup}$ to each $d_i \in \{D\}$
9: **for all** $d_i \in \{D\}$ **do**
10: $d' \leftarrow d \bigoplus mv_r(\beta)$
11: Compute evaluation value ASR-CS($d'$)
12: **end for**
13: $b$" $\leftarrow d_i$ having the maximum ASR-CS($d'$) in $D$
14: $C \leftarrow$ columns of $b'_{group} \notin b$"
15: **if** ASR-CS($b$") $>$ ASR-CS($b$) **then**
16: $B^* \leftarrow b$"
17: **end if**
18: $b \leftarrow b$"
19: **end while**
20: Generate a new bicluster $b$ by perturbing randomly $R$ of recorded best bicluster $B^*$
21: **end while**

## 4. 評価実験及び考察

### 4.1 評価実験

提案手法の有効性を示すために，5 種の化合物 (リガンド) に結合する蛋白質を 4 種ずつ，計 20 種の蛋白質を用いた評価実験を行った．**表 1** にデータセットを記す．

20 種のうち 1 つをクエリ蛋白質，残り 19 種を参照蛋白質としてマッチングを行い，20 種をそれぞれクエリ蛋白質として繰り返す．1 つの参照蛋白質に対して，TOP1 のみを用いてクエリ特徴点対応表を生成した場合と，TOP1〜3 を用いて表を生成した場合で，それぞれ評価実験を行う．TOP1 のみの場合はグループからの排他的選択が不要である．また，列間類似度の有無でも場合分けを行い，**表 2** に示す 4 つのバイクラスタリング手法 (a)〜(d) で実験を行い，抽出される重要特徴点の変化を検証する．

**表 1** 評価実験のデータセット

| Ligand | Protein |
|---|---|
| FMN | 1bwk(A), 1huv(A), 3txz(A), 3gfx(A) |
| GLC | 2osy(A), 2zid(A), 4gi6(A), 4hoz(A) |
| NAP | 1s1p(A), 3r7m(A), 4lau(A), 4gq0(A) |
| SAM | 1r30(A), 3cb8(A), 3t7v(A), 4njg(A) |
| PEP | 1rzm(A), 1xuz(A), 3fyo(D), 3tfc(A) |

**表 2** 評価実験に用いるバイクラスタリング手法

| | 列間類似度無 | 列間類似度有 |
|---|---|---|
| TOP1 | (a) | (b) |
| TOP1〜3 | (c) | (d) |

初期バイクラスタの最小サイズを $4 \times 4$ とする．アルゴリズムのパラメータは，$\alpha = 0.6$, $\beta = 0.6$, $Y = 10$, $Z = 10$, $R = 0.35$ とし，列間類似度として COMPLIG[14] を利用した以下の尺度を用いる．

化合物 $M$ の原子数を $\mathrm{Atom}_M$，結合数を $\mathrm{Bond}_M$，化合物 $N$ の原子数 $\mathrm{Atom}_N$，結合数を $\mathrm{Bond}_N$ とする．ただし，$\mathrm{Atom}_M + \mathrm{Bond}_M \geq \mathrm{Atom}_N + \mathrm{Bond}_N$ とする．化合物 $M$

Vol.2016-MPS-108 No.34
Vol.2016-BIO-46 No.34
2016/7/6

と $N$ を COMPLIG に入力した結果が，原子の対応数が $\mathrm{Atom}_{MN}$，結合の対応数が $\mathrm{Bond}_{MN}$ である時，列間類似度 $\mathrm{LigSim}_{M,N}$ が，次式により与えられる [15].

$$\mathrm{LigSim}_{M,N} = \frac{\mathrm{Atom}_{MN} + \mathrm{Bond}_{MN}}{\mathrm{Atom}_{M} + \mathrm{Bond}_{M}} \quad (0 \leq \mathrm{LigSim}_{M,N} \leq 1)$$

### 4.2 抽出した重要特徴点の評価

抽出された重要特徴点が，実際に結合に関与する重要部位であるのか検証する必要がある．結合に関与する残基の番号を記したデータが Protein Data Bank Japan（PDBj）[*1] に記載されている．

結合部位の位置はアミノ酸残基の番号として表されているため，得られた重要特徴点がどの残基に対応するかを調べる必要がある．本実験では，重要特徴点の座標を PDB ファイルに含まれる原子座標と比較し，三次元座標上で最も近い原子を対応する原子とする．そして，その原子を含む残基の番号を対応する残基とする．

抽出された重要特徴点に対応する残基が，“HETATM” 原子の座標から抽出されたリガンド結合部位に記載されている残基に含まれていれば適合しているとし，各蛋白質において手法 (a)〜(d) のそれぞれに対して適合率を算出する．その結果を**図 4** に示し，各手法における全蛋白質の適合率の平均 (平均適合率) を**表 3** に示す．

### 4.3 評価結果と考察

#### 4.3.1 評価結果

一部の例外を除き，提案アルゴリズム BISES を利用した手法 (d) の適合率が最大であり，全蛋白質の適合率を平均した平均適合率も，手法 (d) で最大である．また，列間類似度の有無に関わらず，一部の例外を除き，TOP3 まで用いる方が TOP1 のみを用いるよりも平均適合率が高くなっている．よって，クエリ特徴点対応表から重要特徴点を抽出する上で，提案手法の有効性が確認されている．

一方で，評価に用いた適合率の値は手法による差異の指標としての便宜的な数値であり，本当に重要部位であるホットスポットを抽出できたかどうかを明確に示すものではない．すなわち，実際の結合プロセスにおいては，最終的な結合箇所よりも，むしろ，そこから離れた箇所の認識が重要な役割を果たすことがしばしばある．このような箇所は重要部位と言えるが，結合部位ではないため，上記の適合率により，その抽出の可否を評価することができない．そこで，抽出された重要部位に対する個別的かつ定性的な評価を行う．

#### 4.3.2 FMN 結合に関する考察

FMN と結合する蛋白質の重要特徴点抽出結果を**図 5** に示す．これを見ると，いずれの蛋白質に関してもフラビン構造の結合部位より少し奥の窪んだ部分で重要特徴点が抽出されている．フラビン構造は電気的な性質が弱く，形状の特徴が結合に深く関与する．化合物は，このような構造を有する重要部位を認識し，そこに向かって結合しようとするが，エネルギーの関係上，その手前の部分で安定な結合状態に到達していると解釈でき，結合部位と重要部位が必ずしも一致しない例となっている．

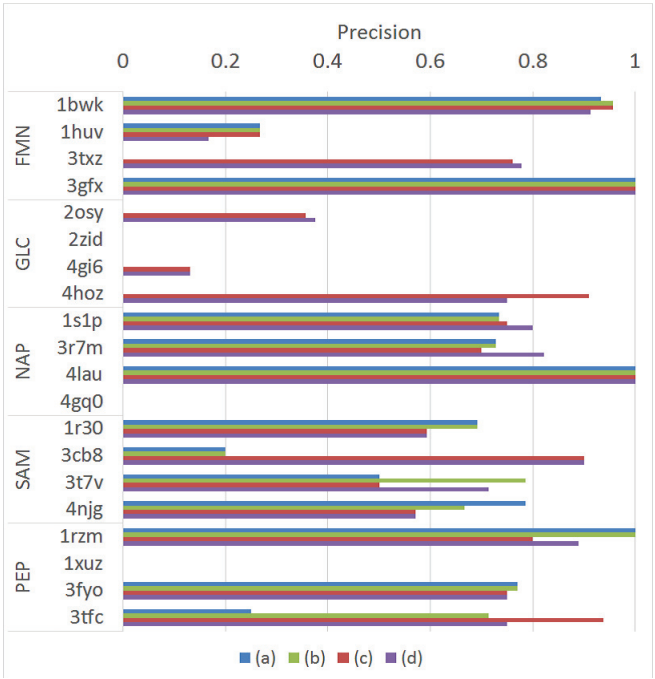



**図 4** 抽出した重要特徴点の適合率

**表 3** 抽出した重要特徴点の平均適合率

| | 列間類似度無 | 列間類似度有 |
|---|---|---|
| TOP1 | 0.443 | 0.476 |
| TOP1〜3 | 0.594 | 0.595 |

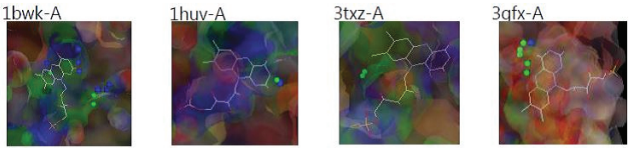



**図 5** FMN に結合する蛋白質の重要特徴点

#### 4.3.3 GLC 結合に関する考察

GLC と結合する蛋白質の重要特徴点抽出結果を**図 6** に示す．これを見ると，実際の結合部位よりも遠く離れた窪みの深い部分で重要特徴点が抽出されている．これも FMN と同様な理由が考えられるが，FMN よりも全体的に適合率が低いことからも，分子認識の目標とされる重要部位よりもさらに離れた箇所において結合している状況が想定される．

[*1] http://pdbj.org/

Vol.2016-MPS-108 No.34
Vol.2016-BIO-46 No.34
2016/7/6

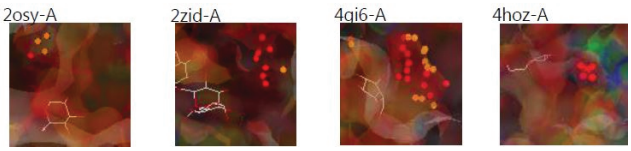



**図 6** GLC に結合する蛋白質の重要特徴点

## 5. まとめ

本研究では，蛋白質分子表面上の重要部位を特定するために，クエリ特徴点対応表から重要特徴点を抽出するバイクラスタリング手法 BISES を提案した．

提案手法の有効性を検証するために，20 種の蛋白質を用いた重要特徴点抽出の実験を行った．抽出された重要特徴点を，実際に結合に関与する部分のデータと対応させた結果，既存手法よりも有効であることが示せた．また，抽出された個々の蛋白質の重要特徴点について，結合化合物との関連性を含め，化学的な観点から考察した．この時，構造的特徴が結合に深く関与する蛋白質に関しては良好な結果を得たが，電気的性質に大きく左右される蛋白質などは良好ではなかった．

今後は，大規模なデータセットを用いて，前述のような蛋白質と同じ結合現象を行う蛋白質を多く含めることが必要であると考える．これに加えて，パラメータの変化による重要特徴点の変化，バイクラスタの評価方法を検討する．

本研究の一部は科学研究費・基盤研究 (B) 24300056 の補助による．

### 参考文献

[1] 藤博幸: 蛋白質の立体構造入門，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] A. Tanay, R. Sharan, R. Shamir: Discovering statistically significant biclusters in gene expression data, Bioinformatics, 18: pp. 5136-5144 (2002).
[4] X. Liu, L. Wang: Computing the maximum similarity biclusters of gene expression data, Bioinformatics, 23(1): pp. 50-56 (2007).
[5] A. Preli, S. Bleuler, P. Zimmermann, A. Wille, P. Buhlmann, W. Gruissem, L. Hennig, L. Thiele, E. Zitzler: A systematic comparison and evaluation of biclustering methods for gene expression data, Bioinformatics, 22(9): pp. 1122-1129 (2006).
[6] D.S. Rodriguez-Baena, A.J. Perez-Pulido, J.S. Aguilar-Ruiz: A biclustering algorithm for extracting bit-patterns from binary datasets, Bioinformatics, 27(19): pp. 2738-2745 (2011).
[7] W. Ayadi, M. Elloumi, J.K. Hao: Pattern-driven neighborhood search for biclustering of microarray data, BMC Bioinformatics 2012, 13: S11 (2012).
[8] 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).
[9] D.G. Lowe: Object recognition from local scale-invariant features, Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, Vol. 2, pp. 1150-1157(1999).
[10] H. Bay, T. Tuytelaars, L.V. Gool: SURF: speeded up robust features, Computer Vision-ECCV 2006, pp. 404-417(2006).
[11] 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, Issue 4, pp. 201-212(2009).
[12] 井上克己，レドゥクフン, 曽和将容, 範公可: 集合演算プロセッサー (SOP)-画像認識への応用-, 信学技報, vol.113, No.236, pp. 35-40(2013).
[13] E.L.Lehmann, H.J.M. D' Abrera: Nonparametrics: statistical methods based on ranks englewood cliffs, Prentice Hall, pp. 292-323 (1998).
[14] M. Saito, N. Takemura, T. Shirai: Classification of ligand molecules in PDB with fast heuristic graph match algorithm COMPLIG, Journal of Molecular Biology, vol. 424, Issue 5, pp. 379-390(2012).
[15] 白井剛: データベース解析によるタンパク質リガンドの多様性, 生化学, 第 85 巻, 第 8 号, pp. 671-678(2013).