Vol.2012-BIO-29 No.14
2012/6/28

3次元画像特徴量を用いた蛋白質分子表面比較

車谷 奈都実[1,a)]　大川 剛直[1,b)]

概要：蛋白質の機能と構造の関連を明らかにする上で，その立体構造を比較し，局所的に類似した部分を見つけることが重要である．本研究では蛋白質分子表面データを 3 次元画像へ変換し，そこから局所特徴点を検出して特徴量を算出することにより，蛋白質の局所構造間を比較する手法を提案する．提案手法を蛋白質の結合部位予測へ適用した結果，11 個中 6 個の結合部位の予測に成功することが示されており，その有効性を確認した．

キーワード：局所特徴量，蛋白質立体構造，SIFT，3 次元画像

Comparison of the protein molecular surface by using the local features in 3D images

Natsumi Kurumatani[1,a)]　Takenao Ohkawa[1,b)]

Abstract: To explain the relationships between functions and structures of proteins, it is important to identify locally similar sites on protein molecular surfaces by comparing protein 3D structures. In this paper, we propose a method of comparing protein structures, in which the molecular surfaces are regarded as 3D images and the similarity between them is calculated by detecting keypoints from the images and computing local features at each keypoint. We applied the proposed method to prediction of protein's binding sites, which shows the accurate prediction of binding sites in six out of eleven proteins.

Keywords: local features, protein structures, SIFT, 3D image

1. はじめに

蛋白質は，代謝や運動，免疫，遺伝と言った生命現象における重要な役割を果たしており，すべての生物が持つ重要な生体高分子の 1 つである．蛋白質の多様な機能は，立体構造を基盤としている．蛋白質には単体でその機能を発現するものもあるが，多くの蛋白質は他の蛋白質や生体高分子あるいは化合物と結合することにより機能を発現する．この結合は，蛋白質構造全体ではなく，蛋白質の局所部位においてしばしば観察され，このような局所部位のことを結合部位という．共通の機能を発現する蛋白質の結合部位は，立体構造上においても類似していることが多いため，蛋白質の局所構造の類似性を比較・評価することは，蛋白質の機能解析に非常に重要な役割を果たす．

蛋白質局所構造の比較には，アミノ酸配列を比較する方法や，立体構造を比較する方法など，様々な方法が存在する [1]．立体構造は蛋白質の機能に直接的に関連するとともに，蛋白質の進化において，構造はアミノ酸配列よりも保存されることから，配列の比較に比べて，より重要な情報を提供できる可能性がある．特に，分子表面上で窪んだ形状をしているポケットは，蛋白質が相互作用する際に直接的に機能する極めて重要な部位である．そこで，本研究では，内部の原子配置ではなく，分子表面ポケットを対象とした比較手法について検討する．

蛋白質の分子表面は，3 次元空間内に配置された多数の頂点の集合として表現でき，各頂点には，その部分における様々な物性値が付与される．これらの物性値を色情報と

[1] 神戸大学大学院 システム情報学研究科
Kobe University, Graduate School of System Informatics
[a)] kurumatani@cs25.scitec.kobe-u.ac.jp
[b)] ohkawa@kobe-u.ac.jp

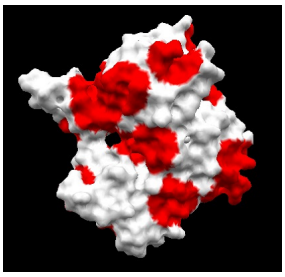

図 1 蛋白質ポケット
Fig. 1 Protein's pocket

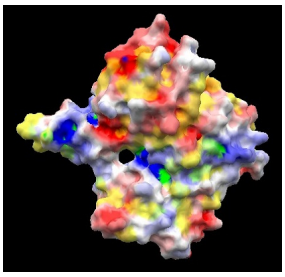

図 2 蛋白質分子表面の例
Fig. 2 Protein's molecular surface

して表すことにより，分子表面は，3 次元のカラー画像とみなすことが可能となる．このため，分子表面のポケットの比較には，3 次元画像の比較・認識手法の援用が期待できる．

これらのことを背景に，本稿では蛋白質の局所分子表面比較に，3 次元 SIFT アルゴリズム [2] と Point Feature Histogram[3][4] を用いた方法を提案する．3 次元 SIFT とは 3 次元画像から局所的な特徴点を抽出するアルゴリズムであり，検出された特徴点の持つ座標や色の情報を用いて Point Feature Histogram を算出し，それを特徴量として記述する．ポケットの比較においては，一方のポケットの 3 次元画像から抽出される複数の特徴点に最も類似する特徴点をもう一方のポケットから見つけ，対応する特徴点間の距離を計算し，これらを元に蛋白質のポケット間の類似度を算出する．このとき，比較する 2 つのポケットから抽出される特徴点の個数は一定ではなく，最近傍の対応付けは 1 対 1 にはならない．特徴点を対応させる基準となるポケットを特徴点数の数が多い方を選んで類似度を算出する方法と，2 つのポケットをともに基準として，両方から得た類似度を足し合わせることで新たな類似度として計算する方法の 2 通りの類似度算出方法を提案する．

2. 蛋白質の構造比較

蛋白質分子表面の凹凸構造の中でも凹んだ部分のことをポケットというが，これは蛋白質の結合部位の候補となるなど，蛋白質の機能に関連する重要な部位である．本研究では，蛋白質の分子表面のポケットを対象とした比較方法について検討する．

2.1 蛋白質のポケットの抽出

蛋白質のポケットは，分子表面の中から幾何学的条件などを考慮することで抽出する．蛋白質の立体構造データからポケットを自動抽出するサービスを提供するウェブサーバである LIGSITE[*1] を用いて蛋白質分子表面のポケットの部分を抽出した例を**図 1** に示す．図中の赤色の箇所がポケットである．

[*1] http://projects.biotec.tu-dresden.de/pocket/

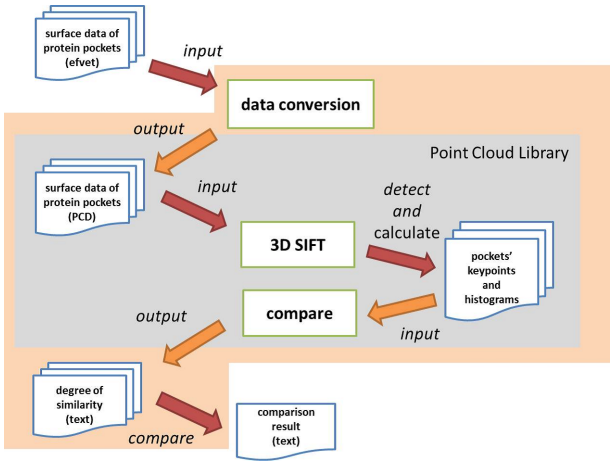

図 3 ポケット比較手法の全体の処理の流れ
Fig. 3 Entire processing

2.2 蛋白質の分子表面データ

蛋白質の分子表面に関するデータベースとして，eF-site[*2] がある．eF-site は，蛋白質の立体構造と機能の関係を統合的に考察することを目的とした蛋白質表面形状と物性に関するデータベースである．eF-site では，蛋白質の立体構造を原子の空間座標から計算した分子表面で表し，その表面上に蛋白質が作り出す静電ポテンシャルと，分子表面近傍の疎水性アミノ酸残基の有無を色付けして表現している [5]．**図 2** に蛋白質の分子表面の例を示す．eF-site の情報は，表面の形状と静電ポテンシャルの値を記述した efvet という名称の XML ファイルにより公開されている．本研究では，蛋白質の分子表面データを efvet から取得する．

3. 3 次元 SIFT による蛋白質ポケットの類似度比較

3.1 概要

分子表面の局所構造比較に 3 次元 SIFT と Point Feature Histogram[4] を使用する．3 次元 SIFT は 3 次元画像から SIFT 特徴点と呼ばれる局所的な特徴点を抽出するアルゴリズムである．この特徴点に対し，Point Feature Histogram と呼ばれる 250 個のビンで構成されたヒストグラムを算出し，その比較を行うことで，3 次元画像から類似している場所を発見する．3 次元 SIFT では 3 次元画像において，特徴的な性質を示す点を特徴点として抽出するため，ポケットの比較においては，ポケット内に含まれる特徴点のうち，多数の点が類似しているときに，類似したポケットと見なすことが可能である．蛋白質分子表面ポケットの比較手法の全体の処理の流れを**図 3** に示す．

[*2] http://ef-site.hgc.hp/eF-site

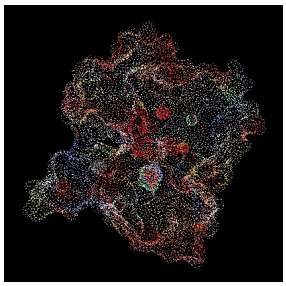

図 4 点群の 3 次元画像
Fig. 4 Point Cloud

3.2 3 次元 SIFT と特徴点の比較

3.2.1 3 次元 SIFT と Point Feature Histogram

3 次元 SIFT とは，SIFT アルゴリズム (Scale Invariant Feature Transform)[6] を 3 次元に拡張したものである．オリジナルの SIFT は 2 次元画像を対象に，2 次元画像の平滑化画像の差分から特徴点を検出し，その特徴点を中心とした領域から勾配方向を求めてオリエンテーションとし，それを元に 128 次元の特徴量を記述する．このアルゴリズムの利点は，画像のスケール変化，回転，明度変化に強い特徴点を取得する点である．本研究では，蛋白質のポケット表面が 3 次元画像と見なせることから，この SIFT アルゴリズムを 3 次元に拡張した手法を用いて特徴点を抽出し，特徴量の記述には Point Feature Histogram と呼ばれるヒストグラムを記述する手法を用いる．3 次元 SIFT の特徴量ではなく，Point Feature Histogram を特徴量として扱う理由は，分子表面に着目しているところにある．本研究で扱う蛋白質の 3 次元画像は**図 4** のような蛋白質分子表面の点群の 3 次元画像である．一方で，3 次元 SIFT では表面だけではなく，物体の中身も考慮して特徴量を計算する．Point Feature Histogram では物体の表面のみを考慮して特徴量を算出するため，計算量も少なくなり精度の高い結果が得られると考えられる [4].

3.2.2 特徴量の計算

本研究で対象とする蛋白質分子表面においては，疎水性や静電ポテンシャルなど，蛋白質の機能に関与する重要な物性値が色情報として付与されているため，特徴量の算出に，点の特徴と色情報を考慮したヒストグラム (RGB Point Feature Histogram) を用いる．1 つの特徴点を中心に k 個の近傍の点の 3 次元直交座標 (x,y,z) の情報を用いて 3 つの Feature を計算し 125 のビンで表されるヒストグラムを作成する．同様にして，1 つの特徴点を中心に k 個の近傍の点の RGB の色情報をもとに 3 つの Feature を算出し，125 個のビンで表されるヒストグラムを作成する．この 2 つのヒストグラムを合わせて 250 個のビンから構成されるヒストグラムを作成し，このヒストグラムを特徴量として扱う [3][4][7].

3.2.3 特徴点の比較

2 つの蛋白質ポケットの表面の類似性を評価するためには，各表面から得られた特徴点同士を比較する必要がある．そこで，2 つの点に対して特徴量 (ヒストグラム) が持つ 250 個のビンの値を 250 個の配列として扱い，配列の各要素の数値を比較し，数値の差の 2 乗和により距離を定義する．すなわち，特徴点 x と特徴点 y に対応する 250 次元のヒストグラムの値を $x = (x_1, x_2, x_3, \ldots, x_{250})$, $y = (y_1, y_2, y_3, \ldots, y_{250})$ と表現すると，距離 $d(x, y)$ は，

$$d(x, y) = \sum_{n=1}^{250} (x_n - y_n)^2 \quad (1)$$

と計算される．この距離を用いて特徴点に最も近い点を求める際に，Kd 木を利用し，高速化を図る [11].

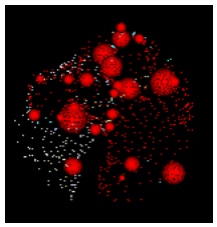

図 5 蛋白質ポケットの SIFT 特徴点
Fig. 5 SIFT keypoint of protein's pocket

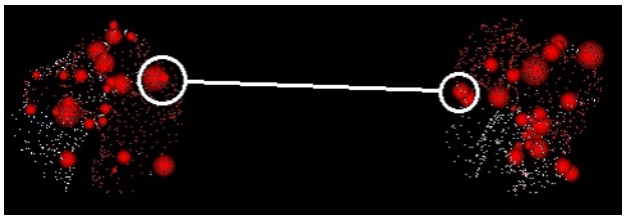

図 6 特徴点の比較
Fig. 6 Correspondence of keypoints

3.3 SIFT 特徴点と特徴量を利用した蛋白質ポケットの類似性比較

3.3.1 蛋白質ポケットの特徴点と特徴量の抽出

蛋白質ポケットの分子表面の 3 次元画像を処理することで，1 つのポケットの画像から，複数の特徴点と特徴量が検出される．**図 5** に，蛋白質ポケットから検出した特徴点の例を赤い球で示す.

蛋白質ポケットの画像から検出される特徴点の数は，画像の大きさや画像に描かれている内容によって異なり，蛋白質のポケットは，特徴点の集合により表現することが出来る.

また，2 つの類似するポケットのそれぞれから抽出された特徴点の中には類似した特徴点が存在することがわかる．**図 6** に例を示す.

3.3.2 蛋白質分子表面ポケットの比較

蛋白質の分子表面ポケット画像 P_A に対して検出された特徴点を $a_1, a_2, \ldots, a_n$，他のポケット P_B に対して検出さ

れた特徴点 $b_1, b_2, \ldots, b_m$ とすると，P_A と P_B を以下のように表すことができる．

$$P_A = (a_1, a_2, \ldots, a_n) \quad (2)$$

$$P_B = (b_1, b_2, \ldots, b_m) \quad (3)$$

ポケット間の類似性は，それぞれを構成する特徴点の類似性から評価する．具体的な手順を以下に示す．P_A の n 個の特徴点の内の任意の 1 点 a_k に対し，P_B の m 個の特徴点の内で最も距離が小さくなる点 $c(a_k, P_B)$ を Kd 木を用いて検索し，その距離 $d(a_k, c(a_k, P_B))$ を求める．この計算を n 個の特徴点全てに対してそれぞれ行う．

同様に，P_B の m 個の特徴点の内の任意の 1 点 b_i に対し，P_A の n 個の特徴点の内で最も距離が小さくなる点 $c(b_i, P_A)$ を検索し，その距離 $d(b_i, c(b_i, P_A))$ を求める．この計算を m 個の特徴点全てに対してそれぞれ行う．

以上のように求めた 2 つのポケット間で最も類似する特徴点同士の距離をもとに，ポケット間の類似度を計算する．このとき，対応する 2 つのポケットから抽出される特徴点の個数は，必ずしも同じではないため，最近傍の特徴点の対応付けは 1 対 1 にならない．そこで，本研究では特徴点対応付けの基準となるポケットを，特徴点数を利用して選択して類似度を算出する方法と，2 つのポケットを共に基準とし，両者から得られる類似度を総合的に用いる 2 通りの類似度算出方法を導入する．

(1) 特徴点数優先法

ポケット P_A と P_B の 2 つの画像を比較する際に，それぞれから抽出される特徴点の数である n と m の大きさを比較し，特徴点の多いポケットから抽出された特徴点を基準として，P_A と P_B 間のの配列から類似度 S_{AB} を算出する．以上のことを式に表すと，

$$S_{AB} = \begin{cases} \sum_{i=1}^{m} \frac{1}{d(b_i, c(b_i, P_A)) + \alpha} & (n < m) \\ \sum_{i=1}^{n} \frac{1}{d(a_i, c(a_i, P_B)) + \alpha} & (n \geq m) \end{cases} \quad (4)$$

となる．$c(a_k, P_B)$ は，P_B 内で点 a_k と最近傍の点を表し，$d(a_k, c(a_k, P_B))$ はその両者の間の距離である．この中の α は類似度に対する距離の影響の度合いを調整するとともに，距離が 0 になった時に逆数が無限大にならないための正の定数パラメータである．

(2) 類似度加算法

ポケット P_A と P_B の 2 つの画像を比較する際に，それぞれのポケットから抽出された特徴点を基準として算出された距離を元に両者の和を求めることにより，P_A と P_B の間の類似度 S_{AB} を求める．以上のことを式に表すと，

$$S_{AB} = \sum_{i=1}^{m} \frac{1}{d(b_i, c(b_i, P_A)) + \alpha} + \sum_{i=1}^{n} \frac{1}{d(a_i, c(a_i, P_B)) + \alpha} \quad (5)$$

となる．

4. 評価実験及び考察

4.1 評価実験

提案手法を，蛋白質の結合部位予測に適用することにより，その有効性を評価する．予測した結合部位が，実際の結合部位と合致していることを判断するため，実験では，結合部位が既知である蛋白質セットの中から，1 つの蛋白質の結合部位が未知であると仮定し，その結合部位を，残りの蛋白質結合部位を利用して予測する．評価実験に用いた 11 個の蛋白質を Table 1 に示す．なお，Protein 欄における記号は，蛋白質立体構造データベース PDB(Protein Data Bank)[*3]で用いられている PDB-ID であり，これによって，1 つの立体構造データが特定できる．また，表には各蛋白質の分子表面に存在するポケット数と，そのうちの結合部位に対応するものの個数についても示している．

本研究では，特徴点と特徴量の計算処理の実装に，Point Cloud Library(PCL)[*4]と呼ばれるライブラリを利用する．ここで扱われる 3 次元画像データは Point Cloud Data(PCD) と言われる，直交座標 XYZ 座標と RGB 色情報が記載された点群のデータであるため，efvet を PCD ファイルに変換してから実験を行う．

表 1 評価実験に用いた蛋白質
Table 1 Protein dataset for evaluation

Protein	Number of pockets	Number of binding sites
1yvn-A	30	2
1ek0-A	23	2
1csn-A	30	4
1f5n-A	30	1
1kv1-A	30	1
1y64-A	30	1
1ije-A	30	1
1k0k-A	13	2
1fy7-A	30	1
1ig8-A	30	2
1j4q-A	25	1

結合部位の予測は，以下の方法に従って実施する．すなわち，結合部位が未知と仮定した 1 つの蛋白質から，存在するすべてのポケット (以下，未知のポケットと表す) を抽出する．また，結合部位が既知の複数の蛋白質結合部位に相当するポケット (以下，結合部位ポケットと表す) を抽出する．1 つの結合部位ポケットと，複数の未知のポケットを比較し，ポケット間の類似度を算出する．この計算を全ての結合部位ポケットに対して行う．未知のポケットのうち，結合部位ポケットに対する類似度が最大となるポケッ

[*3] http://www.pdbj.org/
[*4] http://pointclouds.org/

トが実際の結合部位であるか否かを確認し，結合部位であった場合に正解とする．

4.2 結合部位予測結果と考察

特徴点数優先法 (keypoint prior) と類似度加算法 (similarity adding) の 2 通りの方法を用いて，結合部位の予測を行った．その結果を Table 2 に示す．比較対象として，MolLoc と物性値ヒストグラム (histogram) を用いた結合部位予測結果も記載する．MolLoc[8] は蛋白質の分子表面の構造比較サービスを提供しているウェブサーバである．MolLoc ではスピンイメージを用いて分子表面を比較する．物性値ヒストグラムによる比較 [9] では，提案手法のように特徴点を抽出することなく，ポケット全体を対象にその分子表面が持つ平均曲率・ガウス曲率・静電ポテンシャル・疎水性の 4 つのヒストグラムを用いて蛋白質分子表面の類似性を算出する．

表 2 結合部位予測結果
Table 2 Results

	keypoint prior	similarity adding	MolLoc[8]	Histogram[9]
1yvn	1	1	1	1
1ek0	1	-	1	14
1csn	4	2	1	5
1f5n	1	1	1	1
1kv1	2	1	1	3
1y64	1	1	1	1
1ije	-	-	1	-
1k0k	3	8	1	4
1fy7	1	2	1	9
1ig8	1	1	1	4
1j4q	6	7	-	-

表中の数字は実際の結合部位に相当するポケットの予測順位を示す．例えば，特徴点優先法の行の 1csn の列に着目する．この欄に記載されている値の 4 は，特徴点優先法を用いて結合部位に対する類似度の算出を未知のポケット全てに行い，類似度の降順に未知のポケットを並べたときに，実際に結合部位となる未知のポケットが登場する順位を表す．正解データが存在しなかった場合は，順位を表すことができないため，表に “-” と表した．

この結果より，特徴点数優先法では 11 個中 6 個の蛋白質において順位 1 位として結合部位予測に成功しており，類似度加算法では 11 個中 5 個の蛋白質において結合部位予測に成功していた．このことから，蛋白質の表面比較に局所特徴量の利用が有効であることがわかる．

さらに類似度加算法で結合部位の予測に成功している蛋白質は 1 つの蛋白質を除いて特徴点優先法でも成功していることから，蛋白質表面の比較においては特徴点優先法が類似度加算法よりも優れていると考えられる．

また，物性値ヒストグラムについては，11 個中 3 個の結合部位予測に成功している．提案手法で得られた類似度計算と比べると，劣った結果となった．これにより，蛋白質の局所構造比較には，局所部位全体から得られる特徴量よりも，SIFT のような，数個の特徴的な点を基準とした特徴量を抽出する手法が有効であることがわかる．

一方 MolLoc を用いて蛋白質を比較して結合部位を予測した場合, 11 個中 10 個の蛋白質において結合部位予測に成功していた．この結果を見ると MolLoc は提案手法に対して優位な結果を示しているといえる．この理由は，MolLoc の局所部位比較で用いられるスピンイメージでは点と点の比較を行う際に各点の位置関係を考慮しているためと考えられる．

これらの結果から，今後の課題として，Point Feature Histogram で点の特徴量を算出し比較する際に点の位置情報を考慮する，ポケットから抽出する特徴点を拡張あるいは選別する，そして類似度評価基準を変えることなどが挙げられる．

5. 結論

本研究では 3 次元局所特徴点を検出し，特徴量を算出することで蛋白質の分子表面ポケットを比較する手法を提案した．蛋白質ポケットの分子表面データを用いて 3 次元画像を作成し，3 次元画像に対して 3 次元 SIFT を適用することでポケット画像の特徴点を抽出し，Point Feature Histogram を用いて特徴量を算出し，これらを用いて，蛋白質ポケットの比較を行った．

ポケット同士の類似性を評価するために用いる類似度を求める手法を 2 通り提案した．ポケット画像から抽出される特徴点の数に重きをおいた特徴点数優先法と 2 つの画像の特徴点両方を考慮した類似度加算法の 2 つの類似度評価基準を導入した．これらの有効性を確認するために，提案手法を蛋白質の結合部位予測問題に適用する実験を実施した．

その結果，特徴点優先法では過半数，類似度優先法では約 5 割の精度で予測に成功した．一方で，蛋白質の局所部位から分子表面を作成して比較を行う既存手法に比べて，必ずしも良好な結果が得られていないことも判明し，今後は，点の特徴量を比較する際の位置情報の考慮，ポケットから抽出する特徴点の拡張や選別などについて検討する．

参考文献

[1] 藤 博幸：はじめてのバイオインフォマティクス, 講談社 (2006).
[2] 渡辺 弥壽夫, 中村 省吾：3 次元画像の SIFT 特徴量とその応用, 電子情報通信学会技術研究報告. PRMU, パターン認識・メディア理解, Vol. 109, No. 306, pp. 201–206 (2009).
[3] R. B. Rusu, N. Blodow, Z. Marton, A. Soos and M. Beetz：*Towards 3D Object Maps for Autonomous Household Robots*, Intelligent Robots and Systems, pp.

3191–3198 (2007).

[4] R. B. Rusu, Z. C. Marton, N. Blodow and M. Beetz：*Learning Informative Point Classes for the Acquisition of Object Model Maps*, Proceedings of the 10th International Conference on Control, Automation, Robotics and Vision (ICARCV), pp. 643–650 (2008).

[5] 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, pp. 20–23 (2002).

[6] D. G. Lowe：*Object recognition from local scale-invariant features*, International Conference on Computer Vision, Corfu, Greece, pp. 1150–1157 (1999).

[7] Point Cloud Library：*Point Feature Histograms (PFH) descriptors(online)*, 入手先 ⟨http://pointclouds.org/documentation/tutorials/pfh_estimation.php⟩ (2012.05.28).

[8] S. Angaran, M. E. Bock, C. Garutti, C. Guerra：*MolLoc: a web tool for the local structural alignment of molecular surfaces*, Nucleic Acids Research, Vol. 37, pp. 565–570 (2009).

[9] H. Monji, S. Koizumi, T. Ozaki and T. Ohkawa：*Interaction site prediction by structural similarity to neighboring clusters in protein-protein interaction networks*, BMC Bioinformatics, Vol. 12, Suppl. 1, 39 (2011).

[10] M. E. bock, C. Garutti, and C. Guerra：*Discovery of similar regions on protein surfaces*, Journal of computational biology, Vol. 14, No. 3, pp. 285–299 (2007).

[11] J. L. Bentley：*Multidimensional binary search trees used for associative searching.*, Commun. ACM 18, 9, pp. 509–517 (1975).