2006－BIO－6（8）
2006／9／15

座標基準点探索による蛋白質相互作用プロファイルの抽出

唐崎　太[†]　尾崎知伸[†]　大川剛直[†]

蛋白質の機能は，相互作用部位と呼ばれる局所的部位の立体構造と密接に関係し，相互作用部位の立体構造に類似した蛋白質はしばしば同様の相互作用を示す．蛋白質の相互作用部位の特定は蛋白質の機能解析に有効である．そこで，同一化合物と結合する相互作用部位に共通する特徴をプロファイルとして定義し，これをクエリとした類似した相互作用を示す可能性がある蛋白質を検索する手法が提案されている．プロファイルは，相互作用部位を同一空間上で重ね合わせることにより生成されるが，このとき，どのような相互作用部位立体構造の組み合わせを空間上に重ね合わせるかによって，結果として抽出されるプロファイルも変化する．つまり，プロファイルを抽出するには，立体構造が類似した蛋白質の相互作用部位の組み合わせを選択的に求め，重なり方が最大となるものを得ることが必要となる．

そこで，本研究では，複数の相互作用部位の組み合わせを探索することで，プロファイルの自動抽出を目指す，座標基準点探索を用いたプロファイル自動抽出方式を提案する．相互作用部位の組み合わせの発見に対しては，初期処理で得られる 2 つの相互作用部位を重ね合わせることから生成されるプロファイルを探索出発点として，クラスタリングに基づく方法で実現する．また，相互作用部位群の重ね合わせの発見に対しては，クラスタリングの処理で得られる相互作用部位群の重なり方を用いることで，初期の探索出発点に対する依存性を軽減し，相互作用部位群を対象とした座標基準点探索を実現する．

実験により，相互作用部位の組み合わせの探索が適切に行われ，また，設定した評価基準に基づいてより良いプロファイルを抽出でき，提案手法の有用性を確認した．

Extracting profiles of protein's interaction sites by searching coordinate reference points

Masaru Karasaki,[†] Tomonobu Ozaki[†] and Takenao Ohkawa[†]

The function of protein closely relates to the local structure that is called an interaction site. Proteins whose structure is similar to the interaction sites show the similar interactions. It is effective for the analysis on the protein's function to identify interaction sites. The common characteristic in interaction sites that bind to the same compound is defined as profiles. By using profiles as query, the protein which have the same interaction can be retrieved. Since profiles are generated with overlapping interaction sites, profiles to be extracted depend on the combination of interaction sites and on how they are overlapped.

Based on aggregative hierarchical clustering, we propose a method of extracting profiles by searching the combinations of interaction sites. We start the clustering at profiles which are extracted from two interaction sites and go on. When we overlap the interaction sites, we search the coordinate reference points from interaction sites by using previous reference points which are obtained in results of clustering.

Profiles that were extracted from the combination of some interaction sites show the validity of the proposed method.

1. まえがき

蛋白質の機能は，その表面局所部位で他の物質と相互作用することにより発現する[1)]．相互作用と立体構造には密接な関係があり，相互作用部位の構造が類似した蛋白質はしばしば同様の相互作用を示す．そこで，同一化合物と結合する相互作用部位に共通する特徴を抽出し，これを用いて類似した相互作用を示す可能性がある蛋白質を検索する手法が提案されている[2)~5)]．

この手法では，複数の相互作用部位間で共通する原子とその存在領域をプロファイルとして定義し，これをクエリとした検索が実現されている．プロファイルは，相互作用部位を同一空間上で重ね合わせることにより生成されるが，このとき，どのように空間上で構造を重ね合わせるかによって，結果として抽出される

† 神戸大学大学院 自然科学研究科
Kobe University Graduate School of Science and Technology

プロファイルも変化する．すなわち，相互作用部位の特徴を表すプロファイルを抽出するには，立体構造が類似した蛋白質の相互作用部位の組み合わせを選択的に求め，重なり方が最大となるよう，探索問題として定式化することが重要となる．

そこで，本研究では，複数の相互作用部位の組み合わせを探索することで，プロファイルの自動抽出を目指す，座標基準点探索を用いたプロファイル自動抽出方式を提案する．

相互作用部位の組み合わせの発見に対しては，クラスタリングの初期処理として得られる 2 つの相互作用部位から生成されるプロファイルを探索の出発点として，凝集型階層的クラスタリングに基づく方法で実現する．また，相互作用部位群の重ね合わせの発見に対しては，クラスタリングの処理で得られる相互作用部位群の重なり方を用いることで，相互作用部位群を対象とした座標基準点探索を実現する．

2. プロファイルに基づく類似相互作用蛋白質検索

2.1 概　　要

同一の化合物に結合する複数の相互作用部位は，原子の種類や数，位置が完全には同一でない．しかし，相互作用部位を構成する原子は，同種類の原子が同様の位置に存在することが多い．このことから，同一の化合物に結合する相互作用部位に共通して現れる原子を抽出し，この集合の特徴を**プロファイル**として定義し，クエリとして利用した類似蛋白質検索が提案されている[4),5)]．

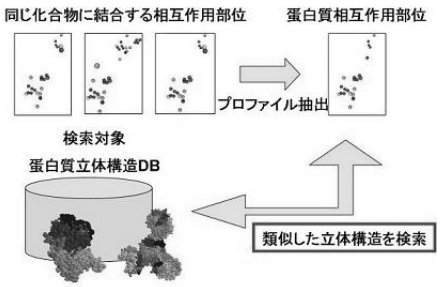

図 1　類似相互作用蛋白質検索方式の概要

プロファイルを用いた類似相互作用蛋白質検索方式の概要を**図 1** に示す．入力は同一の化合物に結合する複数の相互作用部位である．これらの相互作用部位群から検索のクエリとなるプロファイルを抽出する．検索の対象は，蛋白質の立体構造データベースである．プロファイルと，この対象となる蛋白質データベース中の蛋白質群と原子レベルでの構造比較を行い，検索を実現する．

2.2 プロファイルの定義

同じ化合物に結合する複数の相互作用部位において，種類や構造などの性質が類似した原子が共通して見られるとき，それらの原子を総括することで，一つの原子として扱う．これを**仮想原子**と呼び，プロファイルは，これらの仮想原子の集合として定義される．相互作用部位に関連する原子は，**表 1** に示す 10 種類のいずれかの性質を持つことが知られている[2)~8)]．

表 1　相互作用部位を構成する原子

残基	3 文字表記	原子	物性
リシン	LYS	N	塩基性
アルギニン	ARG	N	
ヒスチジン	HIS	N	
アスパラギン酸	ASP	O	酸性
グルタミン酸	GLU	O	
アスパラギン	ASN	N	非電荷極性
グルタミン	GLN	N	
セリン	SER	O	
トレオニン	THR	O	
チロシン	TYR	O	

さらに，相互作用部位の特徴として，抽出する仮想原子の空間座標配置があげられる．複数の相互作用部位において共通な原子の相対的な座標位置は，同一とは限らず，蛋白質によって変化している原子も存在する．この原子位置の変化の大きさをプロファイルを構成する仮想原子ごとに，その**存在可能エリア**として定義する．プロファイルを構成する各仮想原子が持つ存在可能エリアは，プロファイルと検索対象蛋白質との構造比較を行う際，原子の座標のずれに対する許容可能な誤差範囲として利用する．以上をプロファイルの属性と定義し，なおかつ，プロファイルを構成する各仮想原子の配置を，仮想原子間の相対距離行列によって表現する．

プロファイルを構成する仮想原子は以下の属性と相対距離行列を持つ．

- 仮想原子の種類：塩基性，酸性，非電荷極性や，属しているアミノ酸残基の種類
- 仮想原子の存在可能エリア：原子の中心位置と球の半径
- その他のプロファイル内の仮想原子との相対距離行列

プロファイルの一例を**図 2** に示す．図中の球体の大きさは仮想原子の存在可能エリアを表し，球体の中心

間の距離が仮想原子間の相対距離を表している．

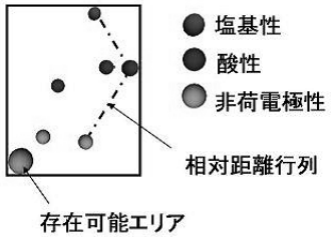

図 2　プロファイルの例

2.3　原子密度分布に基づくプロファイル抽出の概要

プロファイルは，**図 3** に示す手順で，同一の化合物と結合する複数の蛋白質相互作用部位立体構造情報をもとに抽出される．複数の相互作用部位を座標変換を行い，座標合わせを行うことで複数の相互作用部位を一つの空間上に重ね合わせる．この空間内の原子の密度分布を求めると，多数の相互作用部位に共通の原子が現れる領域ほど原子の密度が高くなる．この高密度の領域に着目し，プロファイルの仮想原子を抽出する[2),3)]．

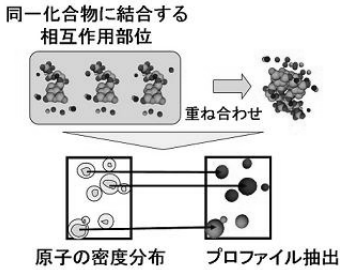

図 3　原子の密度分布に基づくプロファイル抽出方式の概要

原子の密度分布を求めるために，空間を複数の領域に分割して，それぞれの領域における密度を計算する．ここで，各蛋白質内の原子間最短距離が約 2Å であることに着目し，それぞれの領域内には，一つの蛋白質において原子がせいぜい一つしか存在しないように領域の大きさを 1.90Å と設定する．ただし，各領域は，領域内に含まれる原子が全て同一の種類，つまり酸性，塩基性，非電荷極性のいずれかであるように分割する．これを踏まえて，**原子密度**を式 (1) で定義する．

$$\text{原子密度} = \frac{\text{領域内に存在する原子数}}{\text{入力した相互作用部位数}} \qquad (1)$$

原子密度は，多数の相互作用部位に共通する原子が集まった領域ほど高くなる．求めた相互作用部位の原子密度分布において，密度が予め定めた閾値を超えた領域が，プロファイルを構成している重要な原子の集合であると考え，この領域をプロファイルの仮想原子として抽出する．このときの閾値を**原子存在確率**と呼ぶ．

プロファイルの仮想原子の座標はそれぞれの領域に含まれる原子の重心とする．ただし，各原子の質量は一定とする．また，仮想原子の種類は，各領域内に含まれる原子が全て同一の種類であるので，酸性，塩基性，非電荷極性のいずれかとする．更に，各領域内の原子の属するアミノ酸残基の種類がある一定以上の割合で一致する場合は，属するアミノ酸残基を仮想原子の種類として抽出する．

存在可能エリアは，領域内の原子の分布を元に，各領域内の原子が正規分布していると仮定して，重心を中心とした球内部に原子が存在する確率を一意に設定する．これにより，各領域の原子密度分布に基づいた存在可能エリアを決定する．

蛋白質の相互作用部位の原子の配置は三次元座標により表されているが，この座標は各相互作用部位によって異なる原点や座標軸を持つ．原子配置の比較を行うためには，まず複数の相互作用部位の座標を統一する必要がある． そこで，入力である相互作用部位立体構造データから 3 つの原子を選び，その 3 点を元にアフィン変換を用いることで空間を重ね合わせる．この座標変換の基準となる原子を**座標基準点**と呼ぶ．

3.　座標基準点探索を用いたプロファイル抽出

3.1　プロファイルの評価基準

理想的なプロファイルは，ある機能を持つ相互作用部位を表現するのに必要十分な仮想原子から成り立つものであると考えられる．しかし，必要十分な仮想原子は，事前に分からないので，何らかの別の基準でプロファイルの評価を考える必要がある．

ここで，プロファイルが持つべき性質を考える．少数の相互作用部位から作成されたプロファイルは，その相互作用部位に特化されすぎているので，プロファイル本来の役割を果たしているとは言えない．したがって，プロファイルは一定数以上の相互作用部位を表すものである必要がある．

一方で，少数の仮想原子から構成されるプロファイルは，過度に一般化されているので，様々な相互作用部位に合致してしまう．このため，プロファイルは一定数以上の仮想原子から構成される必要がある．ここで，プロファイルの元となった相互作用部位の数と，プロファイルを構成する仮想原子の数の間には，トレードオフの関係があることに注意が必要である．

これらの議論に基づき，本論文では，一定数以上の相互作用部位から作成され，一定数以上の仮想原子を持つプロファイルを対象に，その（相対的）評価を以下のように与える．

相互作用部位の集合 A から得られる仮想原子の集合を P_A とする．二つの相互作用部位の集合 A, B に対し，以下の条件が成り立つときに，A は B に比べて良いプロファイルであると定義する．ここで，$d(a_1, a_2)$ は原子 a_1, a_2 間のユークリッド距離である．

$$
\begin{aligned}
&(|P_A| > |P_B|) \vee \\
&(|P_A| = |P_B| \wedge \\
&\quad \max(\{d(a_1, a_2) | a_1, a_2 \in P_A\}) < \\
&\quad \max(\{d(b_1, b_2) | b_1, b_2 \in P_B\}))
\end{aligned}
$$

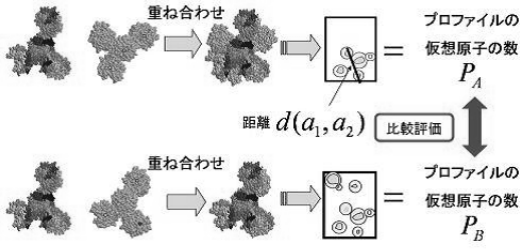

図 4　仮想原子の比較による評価

すなわち，より多くの密集した仮想原子によって構成されたプロファイルを良いプロファイルと定義する．

3.2　探索問題としての定式化

プロファイルは，複数の相互作用部位を重ね合わせることで抽出する．**図 5** のように，相互作用部位の組み合わせによって抽出されるプロファイルは変化する． また，どの座標基準点を使用して座標変換するかによって，空間上での相互作用部位の重なり方が変化し，抽出されるプロファイルも異なる．したがって，前述した基準で最良なプロファイルの抽出は，(1) 相互作用部位の組み合わせと (2) その重ね合わせの探索問題として定式化される．

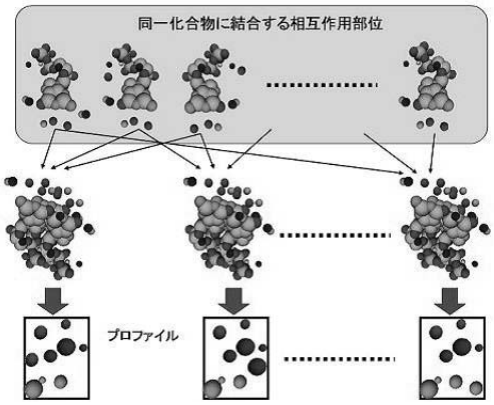

図 5　様々な組み合わせの相互作用部位によるプロファイル抽出

この問題に対し，本論文では，(1) 相互作用部位の組み合わせの発見に対しては，凝集型階層的クラスタリングに基づく方法を，また，(2) その重ね合わせの発見に対しては，相互作用部位群を対象とした座標基準点探索を提案する．また，(2) における相互作用部位群の重ね合わせの際に (1) 凝集型階層的クラスタリングの初期処理として得られる 2 つの相互作用部位から生成されるプロファイルを利用する．

以下では，まず，2 対 1 組の相互作用部位の重ね合わせについて説明し，次いで，凝集型階層的クラスタリングに基づく相互作用部位群の探索およびその重ね合わせの際の座標基準点探索について説明する．

3.3　2 対 1 組とした相互作用部位の座標基準点探索

ここでは，2 つの相互作用部位をどのように重ね合わせるかについて説明する．

いま，相互作用部位立体構造中の原子の総数を N とすると，$(N^3)^2$ 通りの座標変換が考えられ，その全てを調べるには多くの計算時間が必要となる． そこで，本論文では，相互作用部位ごとに座標基準点を探索することとする．更に，各相互作用部位中の座標基準点の探索においては，貪欲探索を用いる．これにより，探索する組み合わせ数が $(3N) * 2$ 通りとなり，計算量を大幅に減少させることが出来る．

以上の探索方式の手続きを示す． ただし，それぞれの初期座標基準点は，経験に基づいて，選択され与えられているものとする．

(1)　相互作用部位 A, B をそれぞれの初期座標基準点 $\langle A_a, A_b, A_c \rangle, \langle B_a, B_b, B_c \rangle$ により座標変換し，抽出されたプロファイル内の仮想原子の集合を P とする．

(2)　入力された各相互作用部位立体構造データ A, B に対して，以下 (3) ～ (5) の処理を繰り返す．

(3)　相互作用部位立体構造 S（A もしくは B）中の各原子 S_v に対して以下を行う．$\langle S_v, S_b, S_c \rangle$ の 3 点の座標基準点を用いて座標変換し，抽出されるプロファイル内の仮想原子の集合 P' を求める．$|P'| > |P|$ である場合，S_a を S_v で置き換える．また，P を P' に更新する．

(4)　(3) における $\langle S_v, S_b, S_c \rangle$ を $\langle S_a, S_v, S_c \rangle$ として相互作用部位立体構造 S 中の各原子 S_v に対して，(3) と同様の処理を行う．

(5) (3) における $\langle S_v, S_b, S_c\rangle$ を $\langle S_a, S_b, S_v\rangle$ として相互作用部位立体構造 S 中の各原子 S_v に対して，(3) と同様の処理を行う．

(6) P を出力する．

上記のアルゴリズムでは，3 つの座標基準点の組み合わせを探索する際に，2 点を固定した上で最適な座標基準点を求めることを繰り返している．すなわち，前の探索で得られた最適な座標基準点を次の探索の出発点としている，という意味で貪欲探索となっている．

3.4 凝集型階層的クラスタリングに基づく相互作用部位群の探索

先に述べたように，良いプロファイルの抽出には，適切な相互作用部位の組み合わせを選択することが重要である．しかし，M 個の相互作用部位の組み合わせ数は，2^M となるのでその組み合わせを全て考慮することは現実的ではない．そこで，凝集型階層的クラスタリングの考え方に基づき，良いプロファイルを生成する可能性が高いと期待される相互作用部位の組み合わせを選択的に求める．

以下に，凝集型階層的クラスタリングを用いた探索のアルゴリズムを示す．なお，α を最小仮想原子数，β を最小相互作用部位数とする．また，M は現時点におけるクラスタ数を表す．

(1) 入力された M 個の相互作用部位に対し，それぞれを単一の相互作用部位から構成される M 個のクラスタを生成する．

(2) 全てのクラスタの組み合わせに対して，任意の 2 つのクラスタの組 a, b の仮想原子 $P_{a\cup b}$ を求める．

(3) $M > 1$ かつ $|P_{a\cup b}| \geq \alpha$ を満たすクラスタの組 a, b が存在する限り以下 (4),(5) を繰り返す．

(4) 全ての組み合わせのうち，先述の基準に基づいて最も良いプロファイルを生成するクラスタの組 a, b を 1 つのクラスタに併合する．

(5) $M := M - 1$

(6) β 個以上の相互作用部位から構成されるクラスタからそれぞれプロファイルを抽出する．

上記の手順では，凝集型階層的クラスタリングの要領で，その時点で最も評価値の高いプロファイルが得られる相互作用部位群同士を併合している．このことは，探索の観点からは貪欲探索に相当する．一方，先述したとおり，抽出すべきプロファイルは，一定数以上の仮想原子を持ち，一定数以上の相互作用部位から構成される必要がある．(3) の条件 $|P_{a\cup b}| \geq \alpha$ は前者を，(6) の条件 β 以上は後者を表現している．

3.5 クラスタ間における相互作用部位の座標基準点探索

前節で示した，凝集型階層的クラスタリングに基づく相互作用部位群の探索では，(3) の処理において，相互作用部位群同士の重ね合わせが必要となる．このとき，各クラスタ内における相互作用部位立体構造データ数を M，相互作用部位立体構造中の原子の総数を N とすると，全ての座標基準点に対して座標変換を調べると，総数で $(N^3)^M$ 通りの組み合わせが存在する．この問題に対処するために，座標基準点探索の対象となる 2 つのクラスタ内にある相互作用部位をそれぞれまとめて座標変換を行うことにする．これは，クラスタ内の相互作用部位同士は，クラスタ生成の過程で，既に適切な重ね合わせ処理が行われていると考えられるためである．

一方， M 個の相互作用部位から構成されるクラスタ A と N 個の相互作用部位から構成されるクラスタ B を併合する際に，A に含まれる任意の相互作用部位 x と B に含まれる任意の相互作用部位 y に対する最適な座標基準点をその初期座標基準点とし，座標基準点探索を $M * N$ 回繰り返す．ここで x, y に対する座標基準点は，(2) の処理で得られていることに注意が必要である．これにより，初期座標基準点に対する依存性を軽減している．

4. 評 価 実 験

提案手法の有効性を確認するため，プロファイルの抽出実験を行った．入力対象として，化合物 Guanosine-5'-Diphosphate(GDP) に結合する蛋白質で，相互作用部位の原子の座標位置がほぼ同一である相互作用部位 10 個と化合物 2'-Monophosphoadenosine 5'-Diphosphoribose(NAP) に結合する相互作用部位 5 個を合わせた 15 個の立体構造データを用意した．これらのデータ 15 個を入力し，プロファイルを抽出する．このとき，最小相互作用部位数 β は 5，最小仮想原子数 α は試行の末，8 と設定した．なお，実験は全て CPU が Opteron 2.8GHz，メモリが 2GB である計算機を使用した．抽出されたプロファイル A, B をそれぞれ**図 6**，**図 7** に図示する．

図中の球体はそれぞれプロファイルの仮想原子であり，球体の濃淡は仮想原子の種類を表し，球体の大きさは各仮想原子の存在可能エリアを表している．

GDP に結合する相互作用部位 10 個のみを入力したときに得られるプロファイル，NAP に結合する相互作用部位 5 個のみを入力したときに得られるプロファイルはそれぞれプロファイル A，プロファイル B に

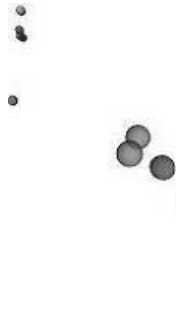

図 6 プロファイル A

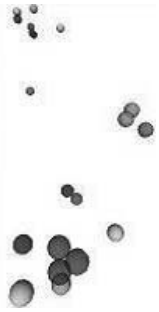

図 7 プロファイル B

類似する．すなわち，この結果は，提案手法において，相互作用部位の組み合わせの探索が適切に行われたことを示している．

一方で，相互作用部位 15 個を入力対象に，相互作用部位の組み合わせの探索を一切行わなずにプロファイルを抽出した場合と比較する．抽出されたプロファイルは，プロファイル A に類似する．これに対し，提案手法では，プロファイル A よりも仮想原子数が多いプロファイル B をも抽出している．このことは，提案手法が，設定した評価基準に基づいてより良いプロファイルを抽出したことを表している．

5. ま と め

本研究では，プロファイルの抽出を，相互作用部位の組み合わせおよびその重ね合わせ方の探索問題と定式化し，凝集型階層的クラスタリングに基づく探索アルゴリズムを提案した．また，予備実験を通じ，その有用性を確認した．

今後の課題としては，より多くのデータを用いた提案手法の定量的・定性的評価などが挙げられる．

参 考 文 献

1) 大島泰郎，西村善文，横山茂之，中村春木　編：構造プロテオミクス：蛋白質ネットワークの構造生物学，共立出版 (2002).
2) M.Matsumoto, Y.Nonomura, and T.Ohkawa: Automatic Profile Extraction Based on Frequency Distribution of Atoms for Retrieving Similar Interaction Protein, Proc. 5th International Conference on Intelligent Systems Design and Applications (ISDA 2005), pp.14–19 (2005).
3) 松本磨莉子，吉野公一，大川剛直：類似相互作用蛋白質検索のための原子頻度分布に基づくプロファイル抽出方式，平成 17 年電気関係学会関西支部連合大会，G12-13，p.G259 (2005).
4) Y.Nonomura, K.Yoshino, and T.Ohkawa : A Method for Retrieving Protein with a Local Structure Similar to Known Interaction Site Using Profile, Proc. 6th International Symposium on Computational Biology and Genome Informatics (CBGI 2005), pp.1307–1310 (2005).
5) 吉野公一，大川剛直：蛋白質－化合物複合体の相互作用部位プロファイルを用いた類似相互作用蛋白質検索方式，情報処理学会研究報告「バイオ情報学」，pp15–20(2005).
6) G.Kawamura, G.Nagakawa, and T.Ohkawa : Development of Protein-Compound Interaction Database on Grid Data Service Using the Three-dimensional Structure Data of Complex, in Abstracts of Pacific Symposium on Biocomputing 2004(PSB 2004), p.87(2004).
7) 野々村祐介，吉野公一，中江達哉，大川剛直：蛋白質－化合物複合体立体構造データに基づく類似相互作用蛋白質の検索方式，情報処理学会論文誌「数理モデル化と応用」，Vol.47，No.SIG 1(TOM 14)，pp.110–119(2006).
8) 野々村祐介，中江達哉，大川剛直：蛋白質－化合物複合体立体構造データに基づく類似相互作用蛋白質の検索方式，情報処理学会「数理モデル化と問題解決研究会」シンポジウム，pp.89–96 (2004).
9) T. Nakae, K. Yoshino, G. Kawamura, G. Nagakawa, and T. Ohkawa : Interaction-based Protein Retrieval by Complementary Use of PIntDB, the Protein-compound Interaction Database, and Known Sequence Motifs, Abstracts of Pacific Symposium on Biocomputing 2004, p.108(2004).