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

野々村 祐介[†,☆] 吉野公一[†]
中江達哉[†] 大川剛直[††]

蛋白質の機能解明に向けて，相互作用部位の立体構造が類似する部分構造を持つ蛋白質の検索が重要な役割を果たす．このとき，同様の相互作用が厳密に同一の原子配置によってのみ行われるわけではないため，入力となる相互作用部位立体構造データをそのまま検索のクエリに用いると類似構造の発見が困難になる．そこで本研究では，空間上で近接しており物性の類似する原子をグループ化することによって入力データを抽象化し，抽象化したデータを用いて類似構造の探索を行う方式を提案する．提案した方式を，化合物 GDP，NADP，FMN に結合する蛋白質 1i4d，1nvt，1um0 を入力として，同じ化合物に結合する複数の立体構造データに適用した結果，類似の相互作用部位を検出し，また，グループ化を行わない手法に比べ誤り部位の検出数を抑えることができ，その有効性が確認された．

Retrieval of Protein with Similar Interaction Based on Structural Data of Protein-compound Complex

Yusuke Nonomura,[†,☆] Koichi Yoshino,[†] Tatsuya Nakae[†] and Takenao Ohkawa[††]

Retrieval of proteins that have similar structures at their interaction site plays an important role in elucidating their functions. This paper proposes a method for retrieving proteins with similar interaction using abstract data by grouping atoms located nearby and having similar physical properties. The proposed method was applied to the retrieval problem where GDP-binding protein 1i4d (PDB ID), NADP-binding protein 1nvt, and FMN-binding protein 1um0 were given as a query. The effectiveness of the method was confirmed by both the facts that it can extract similar interaction sites correctly, and the noises in detection are significantly smaller than in the method that does not group atoms.

1. 序　　論

蛋白質は，20 種類のアミノ酸が分岐なく数十〜数千個鎖状に連なった巨大で複雑な高分子である．1 本の鎖を形成するアミノ酸残基配列は，複雑に折り畳まった形態となり，三次元立体構造を形成する．この鎖の長さと残基の並び方，およびその折り畳まり方によって，様々な機能を持つ蛋白質が形成される．蛋白質は，通常，表面のごく一部において，他の蛋白質や化合物と結合することにより機能を発現する．この部分のことを相互作用部位と呼ぶ[1]．

蛋白質の立体構造情報は，データベース PDB（Protein Data Bank）☆☆に公共データとして登録されている[2)〜4)]．PDB には，蛋白質単体の立体構造のみではなく，蛋白質と化合物が結合した状態の立体構造も登録されている．このようなデータを解析することで，相互作用に関与している原子を特定し，相互作用部位の立体構造を得ることができる[5]．

蛋白質の機能は立体構造と深い関係があることから，相互作用部位の立体構造情報を入力とした類似蛋白質の検索は，蛋白質の機能解明に有用である．これまでにも，蛋白質分子表面に注目した類似蛋白質の検索方式が多数提案されているが[6)〜9)]，これらの手法は原子

† 大阪大学大学院 情報科学研究科
Graduate School of Information Science and Technology, Osaka University
†† 神戸大学大学院自然科学研究科
Graduate School of Science and Technology, Kobe University
☆ 現在，日産自動車株式会社
Presently with Nissan Motor Co., Ltd.

☆☆ http://www.rcsb.org/pdb/

単位で記述された相互作用部位に特化した構造情報の入力を想定していない．そこで本研究では，蛋白質–化合物複合体データから抽出した相互作用部位の立体構造情報を入力とした類似蛋白質の検索方式を提案する．

相互作用部位での結合は複数の原子の組における弱い結合が集まって形成されるものであり，同様の結合が厳密に同じ原子配置によってのみ行われるわけではない．そのため，相互作用部位を構成する原子の座標データをそのまま検索のクエリとして用いると，類似構造の発見が困難になる．そこで，空間上で近接する原子を集めた原子グループ単位で相互作用部位の構造を表現することにより，入力データを抽象モデルとしてとらえ，類似構造の探索を実現する．

2 章では，類似相互作用蛋白質検索の課題を明らかにする．3 章では，同一化合物に結合する蛋白質の相互作用部位の比較をもとに，相互作用部位構造情報の抽象モデルを導入する．4 章では，提案する検索方式の概要を述べ，検索対象蛋白質の内部原子を特定し，構造探索の対象から除外する方法と，入力データに類似する部分構造を探索する手法について述べる．5 章では，提案した手法を PDB のデータに適用し，手法の有効性を示す．

2. 蛋白質–化合物複合体の構造と立体構造データ

2.1 蛋白質の相互作用部位

蛋白質が機能を発現するとき，局所的な部分表面である相互作用部位で，特定の分子を識別し，結合する．すなわち，**図 1** のように，蛋白質の相互作用部位と化合物の形状が合わさり，各部位におけるこれらの物性が相補的になることで多数の弱い化学結合が生じ，分子を結合させる[10]．

相互作用部位表面には複数のアミノ酸残基の側鎖が配置されるが，これらの残基は配列上でひと続きになっているとは限らず，**図 2** に示すように，配列上で離れた位置にある残基が，鎖の折り畳みの結果，三次元空間上で 1 カ所に集まり相互作用部位が形成されることもしばしばある．

2.2 蛋白質の立体構造データ

蛋白質の立体構造に関するデータ数は，近年になって急激に増加しており，代表的なデータベースである PDB には，2005 年 5 月現在で 30,000 個以上の蛋白質データが登録されている．PDB データの 7 割程度は，蛋白質単独の立体構造情報ではなく，蛋白質と化合物が結合した複合体の三次元情報である．これを，ここでは蛋白質–化合物複合体データと呼ぶ．

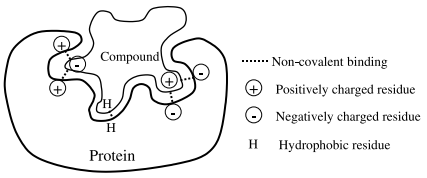


図 1 相互作用部位における蛋白質と化合物との結合
Fig. 1 Uniting of protein and compound in interaction site.

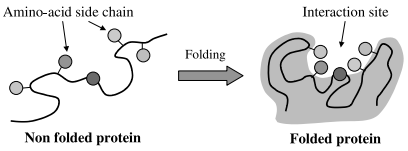


図 2 折り畳みによって形成される相互作用部位
Fig. 2 Interaction site formed by folding.

PDB に登録されている各データには，4 文字の英数字からなる PDB ID と呼ばれる ID コードがつけられている．本来，これらの ID コードは個々の複合体データを識別するために用いられるが，本論文では，便宜上複合体データ中の蛋白質そのものを指すときにも PDB ID を使用する．

2.3 相互作用部位構造データ

蛋白質–化合物複合体データをもとに，原子間の距離や官能基の性質を考慮することで相互作用部位を構成する原子を抽出することができる[5]．蛋白質–化合物の結合情報から抽出された，相互作用に関わっている原子ペアのデータの集合は，相互作用部位構造データと呼ばれ，PIntDB というデータベースに格納されている．PIntDB のデータ項目を**表 1** に，また，データの例を**図 3** 示す．PIntDB は，蛋白質名や化合物名から結合相手を検索する機能や蛋白質名と化合物名から二者の相互作用に関わっている部位を検索する機能等をユーザに提供する．

PIntDB の相互作用部位構造データには，原子間距離や官能基の特性を考慮することで，相互作用に直接関わっている原子の物性や配置に関する情報が含まれている．すなわち，蛋白質が機能を発現する際に最も重要な役割を果たす構造のデータに着目し，データの蓄積を行っている．これは，蛋白質全体のデータに着目している既存の立体構造データベースにはない特徴である．

2.4 類似相互作用蛋白質の検索

蛋白質の機能は，その相互作用部位において，形状や物性が相補的な化合物と特異的に結合することにより発現する．したがって，相互作用部位の形状や物性が類似している蛋白質は，類似の化合物に結合する，

表 1 相互作用部位を構成する原子対データの項目
Table 1 Item of atom pair data that composes interaction site.

	Data item	description
Protein	Protein name	Name that specifies protein
	Number of Residue-array	Number of residue in which position on peptide chain
	Residue	kind of residue
	Functional moiety	Position and kind of functional moiety
	Atom's name in PDB	Atom's name in PDB that takes part in interaction
	Atom's ID in PDB	Atom's ID in PDB
Conpound	Name of compound	Name of compound
	ID of compound	ID that identify compound
	Functional moiety	Position and kind of functional moiety
	Atom's name in PDB	Atom's name in PDB that takes part in interaction
	Atom's ID in PDB	Atom's ID in PDB
	Interatomic distance	Distance between diatomics that forms complex

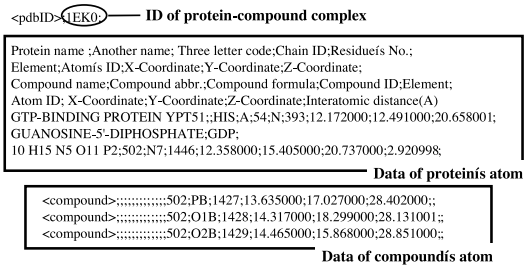


図 3 相互作用部位データの例
Fig. 3 Example of interaction site data.

すなわち類似の機能を持つ可能性があり，類似蛋白質の検索は，機能の解析に重要な役割を果たす．このような検索を実現するためには，蛋白質の相互作用部位の原子構造や物性を考慮する必要がある．PIntDB には相互作用部位のデータに特化したデータが集められており，原子単位での物性や配置に基づいた類似相互作用を有する構造の検索が期待できる．そこで本研究では，PIntDB の相互作用部位構造データを入力とした類似蛋白質の検索方式を提案する．

本検索方式の概要を**図 4** に示す．入力は PIntDB に格納されている相互作用部位構造データである．検索

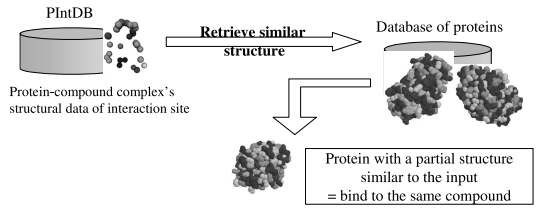


図 4 類似相互作用蛋白質の検索方式の概要
Fig. 4 Outline of a method for retrieving similar interaction protein.

対象は，必ずしも相互作用部位が特定されていない蛋白質の立体構造データであり，各原子の座標データが記述されている．この中から入力の構造に類似した部分構造を持つ蛋白質を見つけ出し，その蛋白質名および合致した部分構造情報を出力する．

相互作用は複数の原子の組における弱い結合が集まって起こるものであり，相互作用部位の構造が多少異なる蛋白質でも同一の化合物に結合する，すなわち類似する機能を持つことがある．類似相互作用蛋白質の検索では，そのような蛋白質も検索できる仕組みが必要である．

相互作用部位を構成する原子のうち，原子位置のずれが結合化合物の選別に比較的大きく影響するものと，あまり影響しないものがある．化合物のコンホメーションの変化に対応した類似相互作用蛋白質の検索を行うには，入力データの原子のタイプを見きわめる必要がある．そこで，あらかじめ入力データに前処理を施し，特に後者のタイプの原子に関して，曖昧性を許す記述に置き換えて抽象モデル化し，それをクエリとして検索する．

3. 相互作用部位構造情報の抽象化

3.1 アプローチ

相互作用部位には，類似の物性を持ち空間上で近接する複数の原子がしばしば見られる．これらの原子の集団は，同一の化合物に結合する別の蛋白質の相互作用部位で見ると，集団内の個々の原子の配置は異なっているものの，集団全体と化合物との位置関係はだいたい類似している．

そこで，個々の原子の厳密な位置は重要視せず，複数個の原子が集まって化合物との結合に機能的な役割を果たす原子の集団を，グループ化する．このとき，基本的には，距離が近く，物性が同じである原子をグループ化するが，最終的にはユーザの判断に委ねる．

どのグループにも属さない原子は，原子 1 個のグループと見なし，最終的に相互作用部位を複数のグループの集まりと考える．生成したグループに与えた

大きさや物性等の属性値およびグループ間の距離行列を記述したものを相互作用グループデータと定義し，類似構造探索のクエリとして用いる．

3.2 グループの属性とグループどうしの位置関係

生成されたグループに，以下の属性を割り当てる．

構成原子数 グループに含まれる蛋白質原子の数

大きさ グループ内の原子間距離の最大値

座標 グループ内原子の座標の平均値

物性 グループ内の原子がどの性質を持つかにより，次の 4 種類のいずれかをグループの物性とする．

- **acidic** グループ内の原子が酸性のみ，もしくは酸性と非荷電極性の場合
- **basic** グループ内の原子が塩基性のみ，もしくは塩基性と非荷電極性の場合
- **polar** グループ内の原子が非荷電極性のみの場合
- **acidic-basic** グループ内に酸性の原子と塩基性の原子をともに含んでいる場合

このような物性の決定方法の根拠については次節で述べる．

また，グループ G_i，G_j 間の距離 D_{ij} を以下のように定義する．

$$D_{ij} = \sqrt{(X_i - X_j)^2 + (Y_i - Y_j)^2 + (Z_i - Z_j)^2}$$

ただし，X_i，Y_i，Z_i はそれぞれグループ G_i の “座標” 属性における x 座標，y 座標，z 座標である．

3.3 グループの物性について

PIntDB には，蛋白質原子と化合物原子の距離だけでなく，それぞれの原子の化学的性質も考慮したうえで，相互作用原子が特定され，格納されている．その結果，相互作用部位を構成する蛋白質原子は，静電的特性を持つ**表 2** に示す 10 種類に限定される．

原則として，化合物の同一箇所に作用する蛋白質原子を 1 つのグループにする．もし，これらの原子の物性が塩基性・酸性・非荷電極性のいずれかに統一される傾向にあるならば，グループの物性は単純にグループ内の原子の物性をあてはめればよい．しかし，この傾向が見られないならば，グループの物性を決めるルールを定める必要がある．

そこで，以下の要件を満たす化合物 PLP（正式名称：Pyridoxal-5’-Phosphate）に着目し，PLP と結合している 27 種類の蛋白質の複合体データを集めた．

- 複合体の三次元構造データが比較的多数解析されている．
- 多くの有機化合物に共通する構成原子である「炭素」，「水素」，「窒素」，「リン」から構成されている．
- 相互作用部位を解析するうえで取り扱いやすい規模である．

これらの複合体データを用いて，同一の化合物原子に作用する蛋白質原子の物性を調べた．結果を**表 3** に示す．表中で，N1 や O1P，O2P 等は，化合物原子の識別 ID である．a，b，p はそれぞれ化合物原子と相互作用を行っている蛋白質原子の酸性・塩基性・非荷電極性を表す．たとえば，a と化合物原子 N1 の交差する要素が 11 であることから，化合物原子 N1 に

表 2 相互作用部位を構成する原子
Table 2 Atoms that composes interaction site.

Physical property	Residue	Atom
Basicity	Lysine	N
	Arginine	N
	Histidine	N
Acidity	Aspartic acid	O
	Gutamic acid	O
Polarity	Asparagine	N
	Glutamine	N
	Serine	O
	Threonine	O
	Tyrosine	O

表 3 化合物 PLP の各原子に結合する蛋白質原子の物性
（a：酸性，b：塩基性，p：非荷電極性）
Table 3 Polarities of protein atom that unites with each atom of compound PLP (a: acidic, b: basic, p: polar).

Physical property of atom	ID of compound atom						
	N1	O1P	O2P	O3P	O3	O4P	P
a	11	0	0	0	0	0	0
b	1	6	5	2	8	6	3
p	3	12	4	8	2	4	9
a-b	0	0	1	0	0	0	0
a-p	6	0	0	0	0	0	0
b-p	0	4	10	7	17	4	6

The element of m line n row shows the number of proteins that unite the protein atoms with the character of m line that unites with the compound atom of n row.

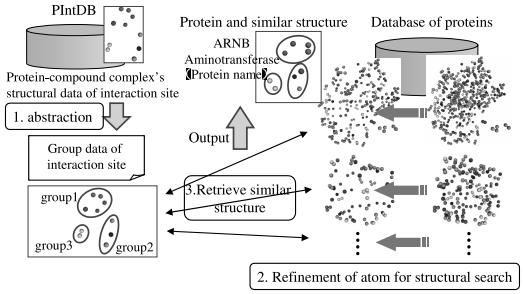


図 5 類似相互作用蛋白質検索方式
Fig. 5 A method for retrieving similar interaction protein.

作用する原子が酸性のみである蛋白質は 11 種類であることが分かる．また，a-b や a-p，b-p は 1 つの化合物原子に複数の物性の原子が結合していることを表し，たとえば a-p と化合物原子 N1 の交差する要素が 6 なので，N1 に作用する原子が酸性と非荷電極性ともに存在している蛋白質は 6 種類である．

表 3 から，同一の化合物原子に酸性原子と塩基性原子がともに結合している例はきわめて少ないが，酸性と非荷電極性，あるいは塩基性と非荷電極性がともに結合している例は多く見られることが分かる．そこで，非荷電極性原子を含むグループの物性は，同一グループに属する他の原子の物性に依存して決定する．

4. 類似相互作用蛋白質の検索方式

4.1 概　　要

類似相互作用蛋白質検索方式の概要を**図 5** に示す．クエリの相互作用部位構造データは，蛋白質原子と化合物原子の座標情報からなる．それらの座標情報をもとに，空間上で近接している蛋白質原子をグループ化し，グループの属性値やグループ間の相対的位置関係等を求め，相互作用グループデータを作成する．

一方，問合せ先である蛋白質立体構造データベース中の蛋白質は，化合物との相互作用部位が明らかにされていないため，構造比較対象となる蛋白質原子の数が膨大になる．構造比較対象の蛋白質原子の中には，蛋白質の表面に現れず，蛋白質内部に埋もれているものも多数存在し，それらは化合物との相互作用に直接関与しないので，構造探索の対象原子からは除外する．

残りの探索対象原子に対して，相互作用グループデータとの構造比較を行い，類似の構造を探索する．クエリとなるグループデータの中のすべての原子グループと属性・位置関係が一致する原子グループの集合を検索対象の中に発見することができた場合，一致した部分構造を出力する．

以下では，検索の前処理である蛋白質内部原子の除外手法と，検索処理のための構造比較手法の詳細について述べる．

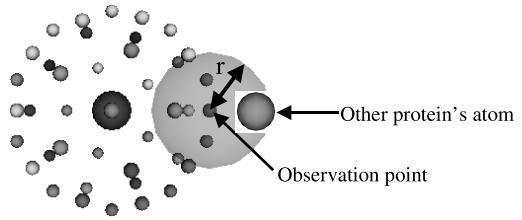


図 6 観測点を中心とした球の例
Fig. 6 Example of ball that centers on observation point.

4.2 構造探索対象原子の絞り込み

類似構造の探索では，探索の対象となる蛋白質原子は表 2 で示した 10 種類に限られるため，その数は蛋白質原子の総数の 10 分の 1 程度になる．しかし，蛋白質は数千個の原子から構成されているので，探索対象原子の数は数百のオーダとなる．その中には蛋白質内部に埋もれているものも含まれる．そこで，これらの内部原子を特定して探索の対象から除外することにより，検索の高速化，および不要な出力の減少による精度の向上を目指す．

4.2.1 蛋白質内部原子の特定

蛋白質の内部に位置する原子は，周辺を他の蛋白質原子に囲まれていることを利用して特定する．判定対象の原子を A とすると，A を中心とする半径 R [Å] の球面を想定し，その球面上に均等に 42 個の点を配置する．この点を周辺原子観測点と呼ぶ．この 42 個の点は，A を中心とした半径 R の球面に内接する Geodesic Dome（測地線ドーム）[11] と呼ばれる，正 20 面体から作られる 80 面対構造の頂点に一致し，これらの点で球体表面を近似的に表現することができる．

次に，周辺原子観測点を中心とした半径 r [Å] の球を想定し，それぞれの球内に他の蛋白質原子が存在するかを見る．その様子を**図 6** に示す．すべての観測点を中心とした球内に他の蛋白質原子が存在すれば，原子 A は蛋白質内部に埋もれていると判断する．

4.2.2 内部原子特定手法の評価実験

PDB の複合体データ 12 個に対して内部原子特定手法を適用し，検証実験を行う☆．なお，42 個の頂点座標の計算には，Geodesic Dome 設計ソフトウェアの “DOME 4.80”☆☆を用いた．実験対象の蛋白質は，化合物 PLP に結合する蛋白質と，PLP と構造がよく類似した化合物 4’-Deoxy-4’-Aminopyridoxal-

☆ $R = 3.2$ Å，$r = 2.7$ Å に設定した．
☆☆ http://www.applied-synergetics.com/ashp/html/domes.html

表 4 絞り込み実験の結果（各蛋白質構成原子に対する絞り込みとそれにともなう相互作用原子数の変化）
Table 4 Result of selection experiment of atoms.

	Number of atoms for search		Number of atoms in interaction site	
	Selection		Selection	
PDB ID	Before	After	Before	After
1a0g	376	306	8	8
1ahe	465	380	10	10
1aia	470	376	10	10
1aic	470	373	9	9
1amq	469	362	10	10
1bkg	458	373	8	8
1fg7	396	323	12	12
1kta	404	332	6	6
1mdo	373	299	8	8
2aat	467	409	6	6
2gsa	435	339	6	6
9aat	494	396	10	10

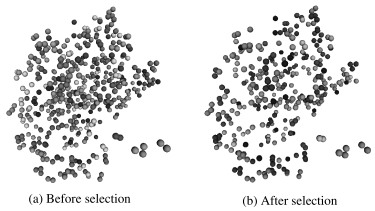


図 7 1aia における原子の絞り込み結果
Fig. 7 Result of selection of atoms in 1aia.

5'-Phosphate（略称：PMP）に結合する蛋白質で，相互作用原子が特定されているおり，結合の状態や蛋白質の大きさが同程度のものを用いた．本実験では，対象の原子数が 400〜500 前後の蛋白質で相互作用部位の構造が類似しているものを対象とした．そのため，PLP に結合する蛋白質のみを用いたのでは実験対象の数が不十分であるため，化合物の構造がよく類似している化合物 PMP に結合する蛋白質を補完的に用いることとした．**表 4** に，それぞれの複合体における，構造比較対象原子の数と，そのうち相互作用原子とされている原子の数を，絞り込み前と絞り込み後に分けて示す．

探索対象原子のうち 13〜23%を探索対象から削除でき，その中に相互作用原子は 1 つもないことから，内部に位置する原子のみを的確に除外することが確認された．一例として，1aia の構造比較対象原子の 3 次元配置を**図 7** に示す．

4.3 検索対象原子群の仮想グループを用いた類似構造の検索

クエリとして与えられた相互作用グループデータに類似する構造を，データベース中の 1 つの蛋白質の部分構造として特定する．このため，**図 8** のように探索対象蛋白質中の原子で仮想グループを順次生成しながら，相互作用グループデータのグループと属性ならびに位置関係が類似しているものを探索する．探索手続きは次の 3 ステップからなる．

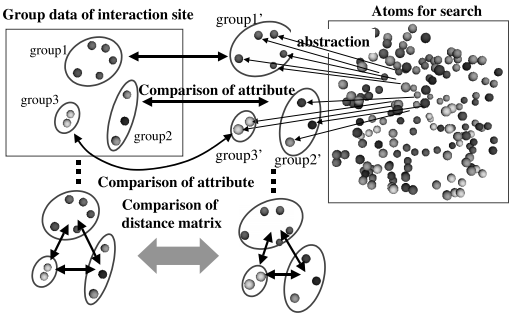


図 8 類似構造探索の概要
Fig. 8 Outline of retrieving similar structure.

ステップ 1 仮想グループとの一致性の判定
ステップ 2 仮想グループ間距離の判定
ステップ 3 類似構造の出力
それぞれの詳細について以下に述べる．

ステップ 1．仮想グループとの一致性の判定

はじめに，クエリ側の相互作用グループデータの中の 1 つの原子グループ G_{q1} に注目する．G_{q1} の原子数を N_{q1} とし，探索対象蛋白質に含まれる探索対象原子群のうち G_{q1} の物性と相反する物性を持つ原子を除いたものの中から，N_{q1} 個の蛋白質原子の組合せを網羅的に選び，それぞれを仮想的にグループと見なす．仮想グループの 1 つを G_{t1} とすると，G_{t1} の大きさと物性の属性が G_{q1} と一致するかどうかをを下記に従い，判定する．

(1) 大きさの一致性判定
誤差許容範囲定数 S_m 未満の誤差であれば一致すると判定する．

(2) 物性の一致性判定
G_{q1} と G_{t1} の物性が同じである場合は一致すると判断し，G_{q1} か G_{t1} のどちらか一方の物性が非荷電極性（polar）である場合にも，もう一方の物性にかかわらず一致すると判定する．

ステップ 2．仮想グループ間距離の判定

大きさ，物性の面で G_{q1} と一致するグループ G_{t1} が見つかれば，別の原子グループ G_{q2} に注目し，同様に仮想原子グループ G_{t2} を生成する．このときは，G_{t1}–G_{t2} 間の距離と G_{q1}–G_{q2} 間の距離との誤差が D_m 未満であれば一致しているとする．

ステップ 3．類似構造の出力

ステップ 1，2 を，クエリ側の原子グループすべてを注目し終えるまで続ける．G_{ti} は，それまでに見つけた仮想原子グループ $G_{t1} \ldots G_{ti-1}$ すべてとの距離関係が，G_{qi} と $G_{q1} \ldots G_{qi-1}$ のそれと一致するかどうかを判定する．また，ある注目原子グループに対して一致する G_{ti} が複数見つかった場合は，それぞれに対して次の注目原子グループに類似する G_{ti+1} を探索する．入力の相互作用グループデータの原子グループ数を n としたとき，G_{tn} が大きさ・物性・他のグループとの距離関係の面で G_{qn} と一致していると判断されたとき，$G_{t1} \ldots G_{tn}$ の原子グループのセットを，入力の相互作用部位構造に類似した部分構造として出力する．

4.4 検索対象原子の組合せ問題

ステップ 1 の仮想グループの生成において，探索対象原子の総数を N_{all} としたとき，その中から N_{q1} 個の原子を選ぶ組合せは ${}_{N_{all}}C_{N_{q1}}$ 通りであり，N_{q1} が増えるに従って組合せ数が膨大になる．しかしながら，それらの組合せの多くは，空間上で互いに遠く離れた複数の原子を含み，そのほとんどがグループの大きさの一致性を満たさない．そこで，以下に示すように探索対象原子が存在する空間を格子状に分割し，互いに遠く離れた格子の中の原子どうしは組合せを除外する．

グループ G_{q1} の大きさを S_{q1} としたとき，探索対象原子が存在する空間を，1 辺 $(S_{q1} + S_m)$ の格子状に分割する．そして，隣接する 8 個の小空間で構成される 1 辺 $2 \times (S_{q1} + S_m)$ の立方体の内部に存在する原子を対象にして，N_{q1} 個の原子の組合せを求める．次に，立方体を x 軸方向または y 軸方向または z 軸方向に格子 1 個分（立方体の 1 辺の半分）ずらして，同様に立方体内部の原子を対象に全組合せを求める．こうして探索対象原子が存在する空間内で立方体の位置を変えながら全領域を探索することで，$S_{q1} \pm S_m$ の大きさの条件を満たす蛋白質原子の組合せを不足なく求める．**図 9** に，格子状に分割された様子と，組合せを求める対象となる小領域の例を示す．

4.5 類似出力の統合

ステップ 3 で出力される部分構造 $G_{t1} \cdots G_{tn}$ の中には，ほぼ同一の箇所を表すものが複数個存在することがある．それらは，構成原子の大部分が同一である．**図 10** にその例を示す．これらは 1 つに統合し最終的な出力結果とすることで，冗長性を排除できる．

類似の出力部分構造を統合するにあたり，出力部分構造間の類似度を定義する．原子の識別 ID をもとに，Jaccard 係数法[12] に従い，同一原子の数および一方

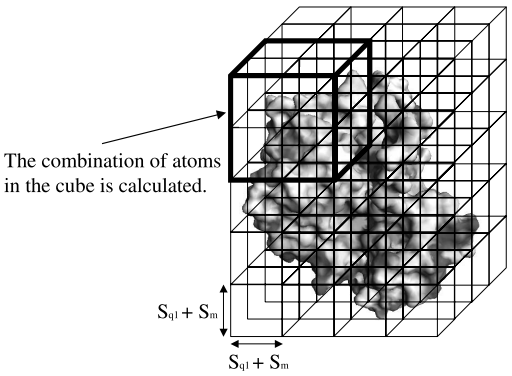


図 9 格子状分割の様子
Fig. 9 Lattice-shaped division.

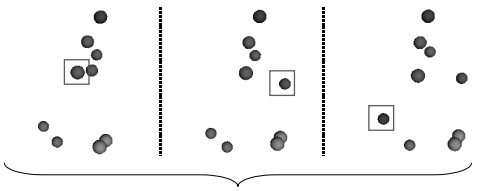


図 10 類似出力の例
Fig. 10 Examples of similar output.

の部分構造にはあるがもう一方の部分構造にはない原子の数から，部分構造間の類似度を Jaccard 係数として算出する．部分構造 i と j の Jaccard 係数 J_{ij} は，以下の式で定義される．

$$J_{ij} = \frac{N_{ij}}{N_{ij} + N_i + N_j} \quad (1)$$

N_{ij} は部分構造 i と j のどちらにも含まれている原子数であり，N_i，N_j はそれぞれ i のみ，j のみに含まれている原子数である．この方法は，部分構造内の原子数が異なっている場合でも類似度を計算することが可能である．さらに，非類似度 D_{ij} を以下の式で定義する．

$$D_{ij} = 1 - J_{ij}$$

D_{ij} をもとにクラスタリングを行って類似度の高い部分構造を統合する．新クラスタとその他のクラスタの非類似度計算方法には最長距離法を用いる．最長距離法を用いたクラスタリングでは各クラスタ内の要素数を比較的少なめに抑えることができ，巨大な部分構造が出力されるのを抑止する．

各クラスタ間の非類似度が T を下回るものがなくなるまでクラスタリングを行う．T は通常 0.667 程度に設定する．これは，部分構造内の原子数が等しいと仮定したときに，クラスタ内の部分構造は少なくとも半数以上の原子を共通に持つということを保証する値

表 5 評価実験に用いた 3 つのデータセット
Table 5 Three data sets used for evaluation experiment.

PDB ID	Compound	Complex's name	Usage
1i4d	GDP	Rac1-GDP complexed with Arfaptin (P21)	Input
1a2k	GDP	Gdpran-Ntf2 Complex	
1byu	GDP	Canine GDP-Ran	
1kk2	GDP	The Large γ Subunit Of Initiation Factor Eif2 From Pyrococcus Abyssi-G235D Mutant Complexed With GDP-Mg2+	Target
1rrg	GDP	Non-Myristoylated Rat ADP-Ribosylation Factor-1 Complexed With GDP	
1nvt	NADP	Crystal Structure Of Shikimate Dehydrogenase (Aroe Or Mj1084)	Input
1nyt	NADP	Shikimate Dehydrogenase Aroe Complexed With Nadp+	
1pqu	NADP	Crystal Structure Of The H277N Mutant Of Aspartate Semialdehyde Dehydrogenase	Target
1q7b	NADP	The Structure Of Betaketoacyl-[Acp] Reductase From E. Coli	
1ra9	NADP	Dihydrofolate Reductase Complexed With Nadp	
1um0	FMN	Crystal Structure Of Chorismate Synthase Complexed With Fmn	Input
1nni	FMN	Azobenzene Reductase From Bacillus Subtilis	
1nox	FMN	Nadh Oxidase From Thermus Thermophilus	
1v4b	FMN	The Crystal Structure Of Azor (Azo Reductase) From Escherichia Coli	Target
1x77	FMN	Crystal Struture Of A Nad(P)H-Dependent Fmn Reductase Complexed With Fmn	

である．

5. 評価実験

提案した類似相互作用蛋白質検索方式の有効性を検証するために，PDB に登録されている蛋白質–化合物複合体の立体構造データを用いて評価実験を行う．なお，実験に使用した計算機は Compaq 製 Evo Workstation W4000 SF（CPU: Intel Pentium 4 2.26 GHz 1CPU，Memory: 512 MB）である．

評価実験に用いたデータセットの詳細を**表 5** に示す．表にあるように 3 つの化合物（略称）GAP☆，NADP☆☆，FMN☆☆☆に結合する蛋白質を検索対象とした．1 つ目のデータセットは，化合物 GAP に結合する蛋白質（PDB ID）1a2k，1byu，1kk2，1rrg を検索対象とし，入力には蛋白質 1i4d を用いた．2 つ目のデータセットは，化合物 NADP に結合する蛋白質 1nyt，1pqu，1q7b，1ra9 を検索対象とし，入力には蛋白質 1nvt を用いた．3 つ目のデータセットは，化合物 FMN に結合する蛋白質 1nni，1nox，1v4b，1x77 を検索対象とし，入力には蛋白質 1um0 を用いた．

4 章で述べた実装法により，計算の効率化は行っているものの，計算量は化合物と結合する蛋白質原子の数に従って増加する．そこで本実験では，現実的な計算時間で検索が完了できる原子数（5〜8 個）で化合物と結合している蛋白質をデータセットとした．また，化合物自体の形が蛋白質との結合によって大きく変わるものは，提案方式で対応することが原理的に困難である．そこで蛋白質との結合により形がほとんど変化のない化合物を選択した．

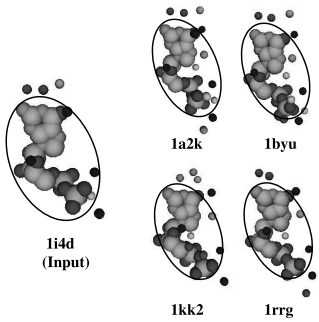
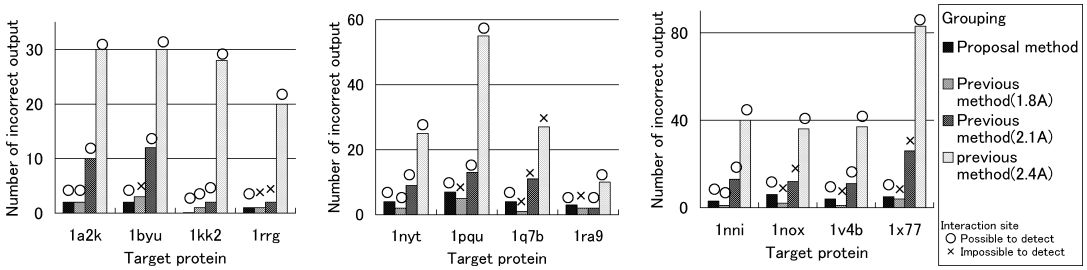


図 11 入力と検索対象の蛋白質の相互作用部位（○内は化合物原子，4 文字の英数字は蛋白質の PDB ID）
Fig. 11 Interaction sites of input protein and retrieval objects (ligand is in the circle, four letters are protein's PDB ID).

相互作用部位の原子の一例として，GDP に結合する 5 つの蛋白質の相互作用部位の様子を**図 11** に示す．それぞれの図の中心部において ◯ で囲まれた，一体となっている物体が化合物であり，その周辺に存在する複数の球状の物体が，蛋白質の相互作用部位原子である．

提案手法と比較する手法として，個々の原子の位置をそのまま照合する検索手法を利用する．具体的には，相互作用部位原子の原子間距離行列と原子の物性が一致する類似構造探索を行い，距離行列に誤差許容範囲を持たせることで，原子構造をある程度柔軟に一致させる．この際，誤差許容範囲 1.8 Å，2.1 Å，2.4 Å の

☆ Guanosine-5'-Diphosphate
☆☆ Nicotinamide-Adenine-Dinucleotide Phosphate
☆☆☆ Flavin Mononucleotide



(a). GDP に結合する蛋白質 (b). NADP に結合する蛋白質 (c). FMN に結合する蛋白質

図 12 3 つのデータセットで行った評価実験結果（異なる検索手法による不正解出力数と正解蛋白質検出の可否）
Fig. 12 Experimental results using three data sets.

表 6 計算時間（単位：秒）
Table 6 Calculating time (unit: Second).

PDB ID	proposed method	Permissible error range (Å) 1.8	2.1	2.4	compound
1a2k	3.2	4.6	9.8	21.5	GDP
1byu	5.3	5.8	12.7	30.4	
1kk2	10.5	15.5	30.2	69.0	
1rrg	2.7	3.9	7.2	15.4	
1nyt	79.4	94.1	147.4	239.9	NADP
1pqu	200.8	173.3	276.6	546.5	
1q7b	48.8	85.2	134.4	216.9	
1ra9	18.9	29.2	41.6	71.5	
1nni	169.4	73.3	134	259.5	FMN
1nox	440.5	104.7	169.1	346.3	
1v4b	170.8	82.7	151.5	282	
1x77	259.5	346.3	282	499	

3 つの値を用いて検索を行った．

提案手法では，グループの大きさに関する誤差許容範囲を 1.5 Å，グループ間距離行列の誤差許容範囲を 2.0 Å とした．検索対象蛋白質の既知の相互作用部位を正解とし，出力された部分構造と正解との類似度を式 (1) によって計算し，類似度が 0.45 以上のものが出力の中に 1 つ以上含まれていれば，正しく相互作用部位を検出したと見なす．3 つのデータセットに対し行った評価実験の結果を**図 12** (a)，(b)，(c) に，また要した処理時間を**表 6** に示す．表中の “error range” は，比較のための手法における誤差許容範囲を表しており，範囲設定ごとの計算時間が示されている．不正解出力数とは，正解との類似度が 0.45 未満の出力部分構造の数である．

比較に用いた手法では，いくつかの結果において誤差許容範囲を大きくしなければ正しく相互作用部位を検出できていない．相互作用部位を正しく検出するために誤差許容範囲を大きくすると，検索結果には多くの不正解が含まれている．提案手法を用いて検索を行った場合と同程度の不正解検出数で，正確に相互作用部位を検出している蛋白質は蛋白質 1a2k，1kk2，1nyt，1ra9，1nni の 5 つにとどまり，半分以上の蛋白質が正確に相互作用部位を検出することができなかった．一方，提案手法を用いた検索では，相互作用部位の個々の原子の位置に影響されることなく検索が行われ，不正解部位の検出を抑えながら，検索対象蛋白質 12 個すべての正確な相互作用部位の検出が行われている．さらに，提案手法では，12 個のうち 11 個の蛋白質で計算時間も短縮することができることが確認でき，個々の原子条件を用いた検索に比べ，計算時間や検索精度の点から，その有効性が示された．

6. 結　　論

本研究では，蛋白質–化合物複合体の相互作用部位立体構造データを入力として，蛋白質の立体構造データベースの中から類似の部分構造を持つ蛋白質を検索する方式を提案した．本手法を PDB の立体構造データに対して適用した結果，入力の相互作用部位に類似した構造を検索できることが確認できた．比較に用いた手法に比べ，正確に相互作用部位を検出し，不正解出力を抑えることができた．

今後の課題としては処理の効率化と蛋白質–化合物複合体以外の入力への対応があげられる．提案手法では注目すべき原子の削減等による効率化が図られているものの，任意の蛋白質を対象とした実用的な検索のためにはいっそう効率化をすすめる必要がある．今後は，相互作用部位原子のグループ化の条件に，結合への影響度合いを考慮したものを導入する等の，アルゴリズムの改良と計算の並列化等の工夫が必要になると考えられる．また，現時点では，蛋白質–化合物複合体のみを対象としているが，これを蛋白質–蛋白質複合体に対しても利用できるように拡張することで，三次元構造に基づく蛋白質相互作用解析等への応用が期待できる．さらに，将来的には相互作用部位の形状変

化への対応も課題としてあげられる．蛋白質が相互作用を起こす際，相互作用部位の形状が作用に合わせて少なからず変化するものがある．本論文における評価実験では，検索対象としてすでに複合体を形成している蛋白質を用いたが，実際には相互作用が発生する前の蛋白質が検索対象となるため，形状変化の大きな蛋白質に対しては，その変動傾向を考慮したフレキシブルな検索への対応が望まれる．

謝辞 本研究の一部は，文部科学省科学研究費補助金，および，科学技術振興機構バイオインフォマティクス推進事業（JST-BIRD）の助成による．

参 考 文 献

1) 伊藤隆司，谷口寿章：プロテオミクス タンパク質の系統的・網羅的解析，中山書店 (2000).
2) 中村春木，有坂文雄：タンパク質のかたちと物性，共立出版 (1997).
3) 金久 實：ゲノム情報への招待，共立出版 (1996).
4) 菅原秀明：あなたにも役立つバイオインフォマティクス，共立出版 (2002).
5) Kawamura, G., Nagakawa, G. and Ohkawa, T.: Development of Protein-Compound Interaction Database on Grid Data Service Using the Three-dimensional Structure Data of Complex, *Abstracts of Pacific Symposium on Biocomputing 2004* (*PSB2004*), p.87 (2004).
6) 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).
7) Ritchie, D.W.: Parametric Protein Shape Recognition, Ph.D. Thesis, University of Aberdeen (1998).
8) Ankerst, M., Kastenmuller, G., Kriegel, H.P. and Seidl, T.: 3D Shape Histograms for Similarity Search and Classification in Spatial Databases, *The 6th Int. Symposium on Spatial Databases* (*SSD*), Lecture Notes in Computer Science (LNCS), Vol.1651, pp.207–226, Springer Verlag (1990).
9) Kriegel, H.P., Schmidt, T. and Seidl, T.: 3D Similarity Search by Shape Approximation, *The 5th Int. Symposium on Large Spatial Databases* (*SSD*), pp.11–28 (1997).
10) Alberts, B., Bray, D., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P.（著），中村桂子，藤山秋佐夫，松原謙一（監訳）：Essential 細胞生物学，南江堂 (1999).
11) Kenner, H.: *Geodesic Math and How to Use It*, University of California Press (2003).
12) 西田英明，佐藤嗣二：実例クラスター分析，内田老鶴圃 (1992).

(平成 16 年 12 月 14 日受付)
(平成 17 年 5 月 21 日再受付)
(平成 17 年 6 月 10 日採録)

野々村祐介

昭和 56 年生．平成 15 年大阪大学大学院情報科学研究科マルチメディア工学専攻修士課程入学．平成 17 年同課程修了．同年日産自動車（株）入社．

吉野 公一

昭和 55 年生．平成 16 年大阪大学工学部電子情報エネルギー工学科卒業．蛋白質の構造比較に関する研究に従事．

中江 達哉

昭和 55 年生．平成 14 年大阪大学大学院情報科学研究科マルチメディア工学専攻修士課程入学．平成 16 年同課程修了．同年（株）日立製作所入社．証券の電子化に関する研究に従事．

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．大阪大学助手，講師，助教授を経て，平成 17 年神戸大学大学院自然科学研究科教授．工学博士．知的ソフトウェア，バイオインフォマティクスに関する研究に従事．IEEE 等の会員．