

タンパク質三次元モチーフ辞書構築支援ツールの開発

加藤 博明*, 宮田 博之, 内村 尚弘, 高橋 由雅, 阿部 英次

豊橋技術科学大学 知識情報工学系, 〒 441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1

*e-mail: hiro@cilab.tutkie.tut.ac.jp

(Received: March 30, 2004; Accepted for publication: May 18, 2004; Published on Web: August 10, 2004)

本研究ではアミノ酸配列レベルのモチーフ辞書 PROSITE に登録されている配列パターンに注目し, これに対応する三次元部分構造情報を網羅的に集積・整理するためのソフトウェアツールの開発を試みた. Protein Data Bank に登録された三次元構造既知のタンパク質構造情報を対象に配列モチーフ部位を検索し, その対応する三次元セグメントの情報を集積する. 次に, 共通の配列パターンを持つ一群の三次元セグメントに対し, 生成した類似度 (相違度) 行列に基づいて構造群をクラスタリングする. ここで, クラスタリング結果が変化する際の閾値間隔に注目し, それぞれのクラスタリング候補に対する優先度を定義するとともに, PROSITE に登録されている既知のモチーフ情報を利用したクラスタリング結果の絞込みについても併せて検討を行なった. これら一連の手順を自動化し, PDB の全エントリを対象とした三次元モチーフ辞書の構築を試みた. WWW ベースのインターフェースツールも併せて開発し, 辞書に登録されたモチーフの代表三次元パターンやその由来タンパク質の構造など, 容易に検索・参照できるようになった.

キーワード: タンパク質モチーフ, 三次元構造特徴, 構造類似性, 三次元モチーフ辞書, PROSITE

1 はじめに

タンパク質の三次元構造と機能との間には密接な関係があることはよく知られている. 特にモチーフと呼ばれるタンパク質構造中に特定の配置で存在する局所構造特徴は, 遺伝子配列の中でもよく保存されている部分であると考えられる [1, 2]. ヒトゲノム計画の進展, 並びにタンパク質構造決定技術の進歩に伴い立体構造のデータは急速に増加しており, その構造データベースはタンパク質の構造と機能との関係解明など分子生物学上の新たな知識獲得のための基盤としてその重要性はますます高まっている [3, 4]. しかし, タンパク質分子の構造の巨大さや複雑さ, さらには近年の急激なデータ数の増大から, 手作業によるモチーフの検索やその特徴解析はほとんど不可能となっている. そのため, これらのデータベースを有効に活用し, 三次元構造特徴の系統的な解析を行なうための方法論の確立, 並びに有効なコンピュータ援用技術の開発が切

望されている.

アミノ酸配列 (一次構造) レベルのモチーフ情報を文献等から広く収集して電子化したデータベースの一つに Bairoch による PROSITE がある [5]. PROSITE には酵素の活性部位やリガンド結合部位の他, タンパク質の細胞内の局在部位を決めるシグナル配列などが収められており, アミノ酸配列の正規表現パターン (PATTERN), 重み行列と配列アライメントのスコア (MATRIX), 自然言語による規則 (RULE), の 3 種類の方法によりモチーフが定義されている. そのうち最も一般的で数多く登録されているものが PATTERN によるモチーフである. 例えば, カルシウム結合に関連する EF-hand モチーフ [6] は, 「D-x-[DNS]-{ILVIFYW}-[DENSTG]-[DNQGHRK]-{GP}-[LIVMC]-[DENQSTAGC]-x(2)-[DE]-[LIVMFYW].」と定義されている. PATTERN では各アミノ酸残基は 1 文字コードで記述され, 正規表現 x はその位置で任意のアミノ酸と, [] はその中のどれかのアミノ酸

と、{ } はその中以外のいずれかのアミノ酸と対応することをそれぞれ意味する。また、要素の後の () 内の数字はその要素の繰り返しを表現し、例えば、x(2) は x-x, すなわち連続する 2 個の任意のアミノ酸とマッチすることを示す。この例では現れないが、 n 個から m 個まで ($n < m$) の任意のアミノ酸の並びは $x(n,m)$ で表現でき、ギャップ領域の指定などに利用されている。

これらの情報をもとに、例えばゲノムネットの DBGET システムでは Protein Data Bank (PDB [7]) をはじめ多くの分子生物学関連のデータベースを統合的に検索し、リンク情報を演繹的に利用し関連する情報を容易に取得することができる [8, 9]。ただし、これらの結果は基本的にそれぞれのデータベースにあらかじめ登録されているリファレンス情報に依存する。また、出力された大量の情報の中から、ある特定の配列モチーフに対応する典型的な三次元幾何パターンを手動で探し出すことは困難である。

筆者らは先に、PDB に登録された三次元構造既知のタンパク質構造情報を対象に、PROSITE の PATTERN をキーとして配列モチーフ部位を検索し、その対応する三次元部分構造 (セグメント) の集積を試みた。また、モチーフごとに、その対応セグメント群をその三次元構造情報をもとにクラスタリングし、それぞれの代表幾何パターンを決定するための方法を提案した

[10]。ただし、これら一連の作業、特にクラスタリング結果の評価とパターンの決定は対話的に trial and error で行なった。

本研究では、より合理的な構造クラスタリングの方法についての検討を行なうとともに、三次元モチーフ辞書構築のための一連の作業を自動化するソフトウェアツールの開発を試みた。また、構築した辞書を利用するための WWW インターフェースについても併せて実装した。

2 三次元モチーフ辞書構築支援ツール

2.1 概要

本研究ではタンパク質の三次元構造をその構成アミノ酸残基を単位として取り扱い、それぞれ炭素の座標で代表して表現する。具体的には、対象とする PDB ファイル群から、必要な構造情報 (ここでは各構成アミノ酸残基の種類とその炭素原子の三次元座標情報) を抽出し、鎖単位の結合表ファイルを生成する。この情報をもとにして、三次元モチーフ辞書を下記の手順で構築する (Figure 1)。

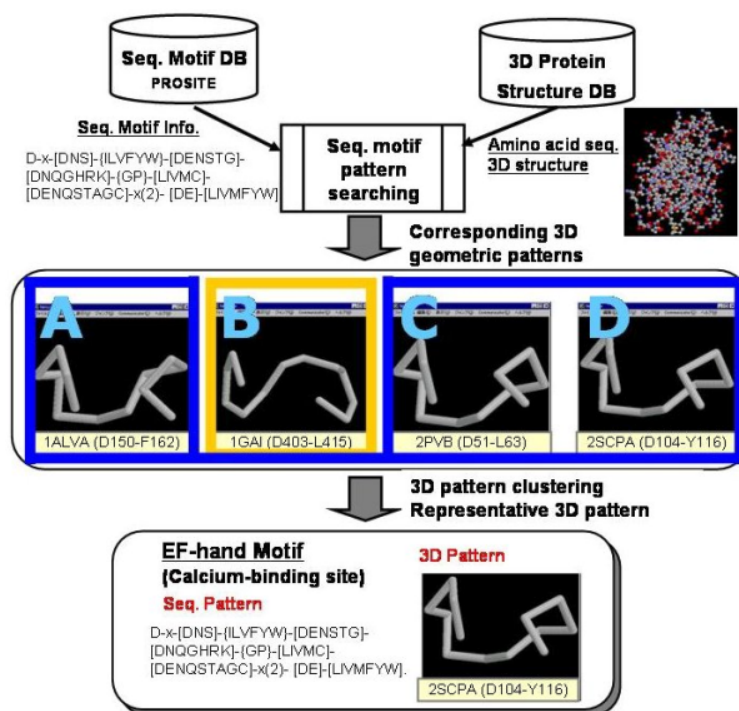


Figure 1. Basic concept of 3D motif dictionary of proteins in the present work.

- (1) PROSITE ファイルを読み込み，登録されている各配列モチーフについて，以下の処理を繰り返す行なう．
- (2) PATTERN 情報を基礎に，各タンパク質データに対して配列パターン検索を行ない，その対応三次元セグメントの情報を集積する．
- (3) これらの共通の配列パターンを持つ一群の三次元セグメントに対し，その全ての組み合わせのペア間で三次元構造比較（相違度の計算）を行ない，相違度行列を生成する．
- (4) 相違度行列に対し，閾値を変化させながら構造クラスタリングを行なう．この閾値間隔に注目して，クラスタリング結果をランキングする．
- (5) PROSITE に登録されているクロスリファレンス情報を参照し，クラスタリング結果の絞込みを行なう．
- (6) 採用されたクラスタリング結果にもとづき，その最もサイズの大きいクラスタの代表構造を，注目している配列モチーフの代表幾何パターンと定義する．
- (7) 対応三次元セグメントの情報（構成残基数，PDBコード，開始と終了の残基位置番号，対応アミノ酸残基とその三次元座標情報），クラスタリング結果，および代表幾何パターンなどの情報を辞書ファイルに登録する．

2.2 配列パターン検索

配列パターン検索では，一つのタンパク質構造中に注目している PATTERN に対応するモチーフ部位が複数箇所存在する場合にも，その全ての部位を抽出できるよう工夫した．ただし，その対応部位がオーバーラップするような領域では，その開始位置が N 末端側のセグメントで代表して表現する．また，正規表現 $x(2,4)$ など複数の可能な展開パターンがそれぞれオーバーラップしてヒットするような領域が存在する場合には，より長いセグメントで代表することとした．

2.3 三次元構造比較

二つのセグメント内の対応する 2 点（炭素）間のユークリッド距離の平均二乗誤差 (RMSD) を求め，こ

れをそれら二つの構造間の相違度として定義した．

$$RMSD(A, B) = \sqrt{\frac{\sum_{i=1}^n \sum_{j=1}^n (d_A(i, j) - d_B(i, j))^2}{n^2}} \quad (1)$$

ここで， n はセグメントのサイズ（アミノ酸残基数）， $d_A(i, j)$, $d_B(i, j)$ はそれぞれセグメント A, B の i 番目の残基と j 番目の残基の間の距離を表わす．

なお，PROSITE のパターン定義により，サイズ（構成アミノ酸残基数）が異なるセグメントが発生した場合は，より小さい方のセグメントを基準とし，配列順序を維持したまま可能な組み合わせを全て列挙する．そして，それぞれ上記の相違度の値を計算し，そのうちの最小値をこれら二つの構造の相違度と定義した．

2.4 三次元構造クラスタリング

共通の配列パターンを持つ一群のタンパク質三次元セグメントに対し，総当たりで三次元構造比較を行なうことによって相違度行列を生成する．ここである閾値を設定すれば，その閾値以内の構造同士は同じクラスタに属し，それより大きいもの同士は別々のクラスタに属するものとみなすことができる．クラスタリングは以下の手順により行なう．

- (1) データセット中の 1 番目のセグメントを代表構造とする．代表構造との相違度が設定した閾値以下となる全ての構造をこれと同じ候補クラスタに属するものとみなし，その構造数（クラスタサイズ）をカウントする．
- (2) 同様に 2 番目以降の三次元セグメントについてもそれぞれ代表構造として (1) を試行し，クラスタサイズが最大となるものを探索する．その候補クラスタを第 1 クラスタ，そのときの代表構造をクラスタの代表三次元パターンと定義する．
- (3) 第 1 クラスタに属する成分を除いた残りの部分行列に対して (1), (2) の処理を行ない，残りの要素がなくなるまでクラスタリングを繰り返す．

手順 (2) で，そのクラスタサイズが最大となる代表構造の候補が複数存在する場合には，(a) 同じクラスタとなる要素の相違度の和を求め，それが最小となるもの，(b) 差がない場合は，別のクラスタとなる要素の相違度の和が最大となるもの，(c) それも同じ場合はデータセット中でインデックス番号の若い方，を選択する．Figure 2 の場合，第 1 クラスタの代表構造は

“A”と決定される。

この場合のクラスタリング結果は設定する閾値に依存する。ここで、閾値は相違度行列の成分と比較されているため、クラスタリング結果が変化する閾値では離散値をとる。そのため、相違度行列内の重複分を除く全ての成分値を順次閾値としてクラスタリングを行なうことで、全ての可能なクラスタリング結果を列挙することができる (Figure 3)。

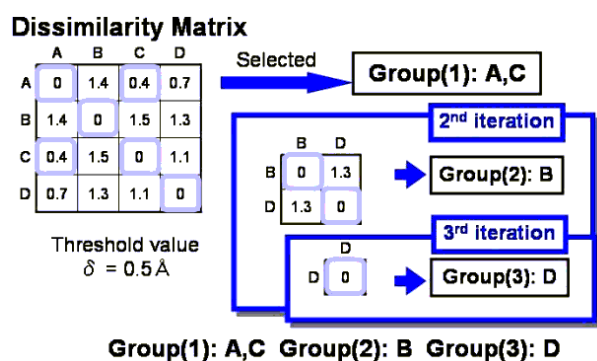


Figure 2. Clustering of 3D geometric patterns using the dissimilarity matrix.

Threshold (δ)	Clustering Result	Priority Value (P)
0.0	{A}, {B}, {C}, {D}	0.4 (= 0.4 - 0.0)
0.4	{A, C}, {B}, {D}	0.3 (= 0.7 - 0.4)
0.7	{A, C, D}, {B}	0.6 (= 1.3 - 0.7)
1.1	{A, C, D}, {B}	
1.3	{A, B, C, D}	0.2 (= 1.5 - 1.3)
1.4	{A, B, C, D}	
1.5	{A, B, C, D}	

Figure 3. Enumerated clustering results and their priority value.

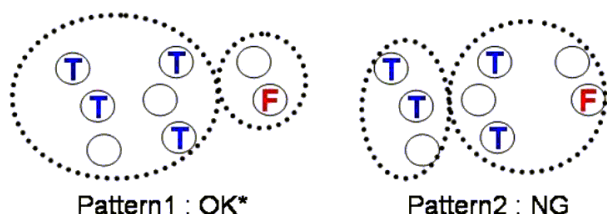


Figure 4. Refinement of the clustering patterns using the additional information of PROSITE.

列挙された各クラスタリング結果に対して、隣接するクラスタリング結果の閾値との差をそのクラスタリングの優先度と定義する。ただし、Figure 3 の 0.7 と 1.1 のように同一の結果を与える閾値の場合は、その値の小さいもので代表する。この優先度の順にソートして、指定された候補数だけクラスタリング結果を出力できるようにした。

2.5 クラスタリング結果の絞込み

PROSITE にはその該当モチーフパターンを含む既知の PDB または SWISS-PROT データベース [11] へのクロスリファレンス情報 (DR レコード) が定義されている [5]。そこには真のモチーフの構造として同定されているもの (真構造) と、配列パターンにはヒットするが他の理由によりモチーフとはみなされていないもの (偽構造) などの情報を含んでいる。本研究では必要に応じてこれらの情報を参照し、クラスタリング結果の絞込みができるように工夫した。具体的には、列挙されたクラスタリングのうち、真構造が全て一つのクラスタに、かつ偽構造がそれ以外のクラスタになるようなクラスタリングのみを残すようにした (Figure 4)。

このようにしてクラスタリングの候補が一つに絞り込めた場合はそのクラスタリング結果を、それ以外の場合には、そのうち優先度が最も高いクラスタリング結果を採用し、モチーフ辞書に登録するものとした。

2.6 開発環境

本ツールは、Microsoft Windows XP 上で、C++ 言語 (Visual Studio.NET 2002) を用いて開発を行なった。

3 三次元モチーフ辞書操作ツール

3.1 キーワード検索・参照機能

以上のようにして作成した三次元モチーフ辞書の情報は、モチーフの種類ごとに PROSITE の ID 番号 (例えば PS00018) に対応するディレクトリを構成している。各ディレクトリには、集積した対応三次元セグメントの一覧を記したファイル (インデックスファイル)、その構造情報ファイル群 (結合表形式)、及びクラスタリング結果の情報を記述したファイルが格納されている。

このような三次元モチーフ辞書の情報を以下の手順により参照することができる。

- (1) 登録されている PROSITE モチーフの検索
検索フォームにキーワードを入れ検索を開始する。
- (2) 検索結果の一覧
検索条件に該当するモチーフ名をリスト表示する。
ここから表示したいモチーフデータを選択する。
- (3) 登録モチーフ一覧
選択したモチーフに関する PROSITE 情報とそれに対応する代表幾何パターンをグラフィック表示する (Figure 5 左図)。
- (4) 注目モチーフの対応構造一覧
特定のモチーフを選択すると、それに対応する三次元セグメントの情報がリスト表示される。その際、構造クラスタリングの結果も併せて表示する。
- (5) 対応三次元セグメント群の表示
表示したリストの中から選択したセグメント群の三次元構造をグラフィック表示する。
- (6) 由来構造式の表示

セグメント (部分構造) のグラフィック表示のオプションとして、その由来タンパク質構造 (親構造) を表示することが可能である (Figure 5 右図)。また、三次元セグメントの結合表ファイルも表示できる。

キーワード検索機能では、モチーフに関するキーワードをクエリーとした検索だけでなく、タンパク質に関するキーワードや PDB コードをクエリーとして、辞書の登録内容を検索することができる。これにより、どのタンパク質がどのモチーフに対応しているのかを参照することが可能である。さらに、表示機能では、構造表示のウィンドウサイズの指定、画面 (ページ) 切り替え、モチーフ構造の向き (表示角度) の統一などの工夫をした。

なお、これらのインターフェースツールは Perl を用いて CGI として作成した。また、三次元構造情報の表示には、MDL 社の Chemscape Chime プラグイン [12] を使い、RasMol スクリプト [13] により部分構造や表示モデルの指定を行なった。



Figure 5. Snapshots of the WWW interface for the 3D motif dictionary system.

3.2 三次元部分構造検索機能

筆者らは先に、タンパク質三次元部分構造検索プログラム SS3D-P の開発を進めてきた [14]。本操作ツールでは、これと連携して、辞書に登録されている三次元幾何パターンをキーとしたデータベース検索を容易に行なえるようにした。

4 三次元モチーフ辞書の構築実験

PDB (Rel.102) の全エントリ (配列重複分を除く 25,980 鎖) と PROSITE (Rel.17.01) の PATTERN により定義された配列モチーフ (1,331 件) を用いて三次元モチーフ辞書の構築を試みた。その結果、少なくとも一つの部位がヒットした配列モチーフは 907 件あり、のべ 290,469 の三次元セグメントが抽出された。ヒットした部位の数とそのモチーフ数を Table 1 にまとめた。このうち対応部位の数が 3 以上 1,000 未満の配列モチーフについて、引き続き構造クラスタリングを行なった。この段階ではクラスタリング候補が多数存在するが、PROSITE の情報を用いて絞込みを行なうことでクラスタリングの候補数を大幅に減らすことができた (Table 2)。

Short-chain dehydrogenases/reductases (SDR) モチーフ (PS00061) [15] の例では、82 の対応三次元セグメントから、可能なクラスタリング候補が 87 種類列挙され、この絞込み処理の後、その数が 11 まで減少した。そのうち優先度が最も高かったクラスタリングは、閾値 3.05 (優先度 1.10) の条件によるものであり、

Table 1. Results of sequence motif pattern searching in the PDB entries.

Hit segments	0	1-2	3-99	100-999	1,000-
Motifs	424	167	705	26	9

Table 2. The number of clustering candidates and motifs after the refinement procedure.

Clustering candidates	0	1	2-4	5-9	10-
Motifs	4	324	212	117	72

結果として 6 個のクラスタが得られた。このときの各クラスタの代表構造を Figure 6 に示す。このうち、(1) は真構造を含むクラスタ、(6) は偽構造を含むクラスタである。これらの結果から、妥当な構造クラスタリングがなされていることが視覚的にも分かる。

5 おわりに

本研究では、PROSITE の各配列モチーフに対し、(1) 配列パターン検索、(2) 三次元構造比較、(3) クラスタリング候補の列挙と選択、(4) 代表三次元幾何パターンの決定、の一連の作業を自動的に行なうためのツールを開発するとともに、最新の PDB の全データセットを対象とした三次元モチーフ辞書の構築を行なった。また、構築した辞書を管理・参照するための WWW インターフェースも併せて開発した。これらのツールを用いて蓄積された三次元モチーフ辞書、特にモチーフの代表三次元パターンの情報は、アミノ酸配列とその立体構造、あるいはその機能との関係解明など、三次元分子構造特徴解析に基づく新たな知識発見のための有用な情報を提供するものとする。本研究で作成したモチーフ辞書と、そのインターフェースは下記の URL で公開している。

<http://prodb.cilab.tutkie.tut.ac.jp/services/>

なお、本研究は文部科学省科学研究費補助金・特定領域研究 B(2) (13131210)、及び財団法人堀情報科学振興財団の研究助成のもとに行われたものであることを明記して謝意を表す。

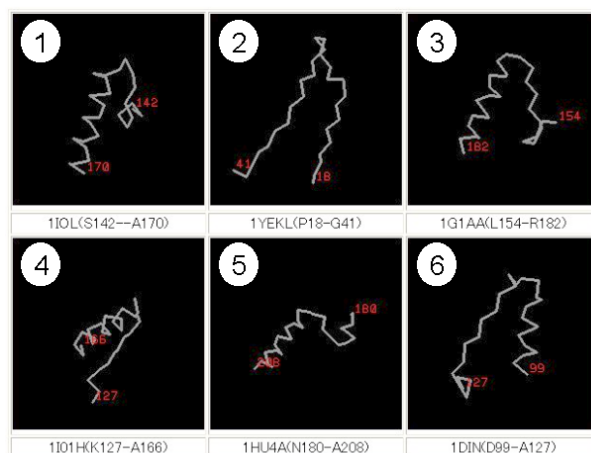


Figure 6. The representative geometric patterns for each cluster obtained from the SDR motif.

参考文献

- [1] Branden, C. and Tooze, J., *Introduction to Protein Structure*, Garland Publishing, New York (1991).
- [2] 中村春木, 中井謙太, バイオテクノロジーのためのコンピュータ入門, コロナ社 (1995).
- [3] Thornton, J. M. and Gardner, S. P., Protein Motifs and Database Searching, *Trends Biochem. Sci.*, **14**, 300-304 (1989).
- [4] 金久實, ヒューマンゲノム計画, 共立出版 (1997).
- [5] Bairoch, A., PROSITE: a Dictionary of Sites and Patterns in Proteins, *Nucleic Acids Res.*, **19**, 2241-2245 (1991).
<http://www.expasy.org/prosite/>
- [6] Kretsinger, R. H., Structure and Evolution of Calcium- Modulated Proteins, *CRC Crit. Rev. Biochem.*, **8**, 119-174 (1980).
- [7] Helen, M., *et al.*, The Protein Data Bank, *Nucleic Acids Res.*, **28**, 235-242 (2000).
<http://www.rcsb.org/pdb/>
- [8] 金久實, ゲノムネットのデータベース利用法, 共立出版 (2002).
- [9] Kanehisa, M., Linking Databases and Organisms: GenomeNet Resources in Japan, *Trends Biochem Sci.*, **22**, 442-444 (1997).
<http://www.genome.ad.jp/dbget/>
- [10] 加藤博明 他, タンパク質構造データマイニングのための三次元モチーフ辞書の作成, 人工知能学会論文誌, **17**, 608-613 (2002).
- [11] Boeckmann, B., *et al.*, The Swiss-Prot Protein Knowledgebase and its Supplement TrEMBL in 2003, *Nucleic Acids Res.*, **31**, 365-370 (2003).
<http://www.expasy.org/sprot/>
- [12] MDL Inc., <http://www.mdli.com/>
- [13] Sayle, R. A. and Milner-White, E. J., RASMOL: Biomolecular Graphics for All, *Trends Biochem. Sci.*, **20**, 374-376 (1995).
- [14] Kato, H. and Takahashi, Y., Three-Dimensional Structural Feature Search of Proteins, *Bull. Chem. Soc. Jpn.*, **70**, 1523-1529 (1997).
- [15] Joernvall H., *et al.*, Short-chain Dehydrogenases/Reductases (SDR), *Biochemistry*, **34**, 6003-6013 (1995).

Development of Software Tools to Construct a 3D Motif Dictionary of Proteins

Hiroaki KATO*, Hiroyuki MIYATA, Naohiro UCHIMURA,
Yoshimasa TAKAHASHI and Hidetsugu ABE

Dept. of Knowledge-based Information Engineering, Toyohashi University of Technology
1-1 Hibarigaoka, Tenpaku-cho, Toyohashi, Aichi 441-8580, Japan

**e-mail: hiro@cilab.tutkie.tut.ac.jp*

This paper describes a three-dimensional (3D) protein motif dictionary system that is closely related to the PROSITE sequence motifs. Because there were many different 3D motif patterns but having a particular PROSITE sequence pattern, we have investigated the approaches for quantitative comparison and clustering of such 3D structure segments. For a pair of 3D structure segments, the dissimilarity value was defined with the root mean squares of inter-residue distances. A conformational pattern clustering was employed for grouping the 3D patterns on the basis of the dissimilarity matrix. Some additional knowledge information described in PROSITE was also used to refine the clustering results. A 3D motif dictionary was constructed using all the data set of the Protein Data Bank. A graphical user interface for using the dictionary was also developed.

Keywords: Protein motif, 3D structural feature, Structural similarity, 3D motif dictionary, PROSITE