KP18

タンパク質三次元モチーフ辞書管理システムの開発

(豊橋技科大)　○内村尚弘、加藤博明*、高橋由雅、阿部英次

1. はじめに

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本要素としてその重要性はますます高まっている。筆者らは三次元分子構造特徴解析に基づく知識発見の視点から、Bairoch によるアミノ酸配列レベルのモチーフデータベース PROSITE [1]に対応する三次元モチーフ辞書の作成を進めてきた（図 1）[2]。本研究では、Protein Data Bank（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るとともに、これを利用したタンパク質三次元構造特徴解析を支援するための WWW ベースの管理システムの開発を行なった。

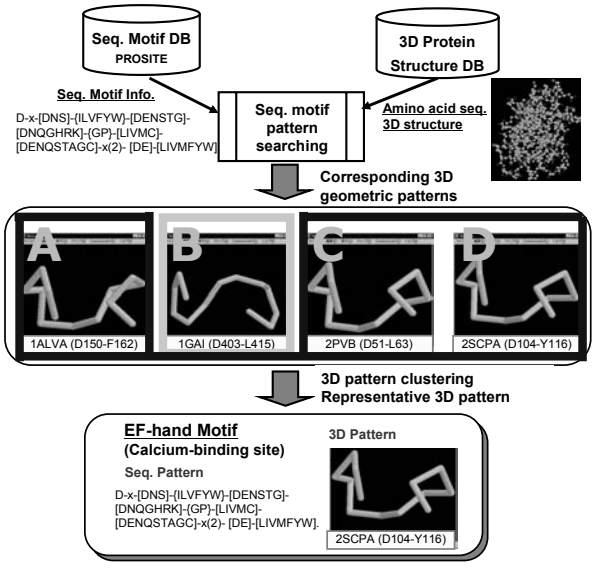

図 1　三次元モチーフ辞書構築の基本概念図

2. 三次元モチーフ辞書の概要

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（表 1）。この内、ヒット数が 3 件以上 1000 件未満(かつ、モチーフ部位の構成残基数が 100 未満)となったモチーフを対象に、先に提案した手法[2]で三次元構造クラスタリングと代表幾何パターンの決定を行なった。

表 1　構造ヒット数とモチーフ数

ヒット数	0	1,2	3～99	100～999	1000～
モチーフ数	424	167	705	26	9

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述したファイル（インデックスファイル）、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。なお、現在扱っているバージョンでは、1 モチーフあたり約 26 件の対応三次元セグメントが登録されている。

本研究では、構築したモチーフ辞書を対話的に利用・管理するためのシステムを開発した。
なお、本システムは全て Perl を用いて CGI として作成した。また、外部プログラムとして C++で作成した実行ファイルなども利用している。三次元構造情報の表示には、MDL 社の Chemscape Chime プラグイン[3]を用い、RasMol スクリプト形式により部分構造や表示モデルの指定を行った。

3. キーワード検索・参照機能

作成した管理システムでは、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（図 2 左）。

*hiro@cilab.tutkie.tut.ac.jp

(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がリスト表示される。その際、クラスタリングの結果も合わせて表示する。

(5) 対応三次元セグメント群の表示

表示したタンパク質のリスト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、その由来構造を表示することが可能である(図 2 右)。また、三次元セグメントの結合表ファイルの情報も表示できる。

図 2　登録モチーフ一覧(代表幾何パターン)とその由来構造式の表示画面例

なお、キーワード検索機能では、ユーザが指定可能な検索条件として、モチーフに関するキーワードをクエリとして検索できるだけでなく、タンパク質に関するキーワードをクエリとして、三次元モチーフ辞書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)などの工夫をした。また、各モチーフの代表幾何パターンの表示画面には、PROSITE エントリへのリンクを用意し、その詳細を参照できる。

4. 三次元部分構造検索機能

当研究室では、別途、三次元部分構造検索プログラムの開発を行なっている[4]。これと連携して本モチーフ辞書に登録された任意の三次元パターンをクエリとしたデータベース検索を可能とした。

5. データベース更新機能

(1) 再構築

PDB や PROSITE のデータは定期的に更新されている。それに合わせて、バッチ的にデータベースの再構築を行なうことができる。

(2) 代表幾何パターンの更新

構築した三次元モチーフ辞書のクラスタリングの結果を確認し、必要に応じて変更できる。具体的には、共通の配列パターンを持つ一群の三次元セグメントのグラフィック表示、及び、そのクラスタリング候補と代表幾何パターンの選択、の一連の手順を対話的に操作することで登録データの更新を行なうことができる(図 3)。

図 3　データベース更新機能の実行画面例

6. まとめ

本研究で作成した管理システムにより、三次元モチーフ辞書の内容を容易に参照・利用できるようになった。さらに、登録された情報を用いて、三次元モチーフ検索プログラムによる特徴解析が可能である。なお、作成したインターフェースは、http://prodb.cilab.tutkie.tut.ac.jp/services/より公開している。今後は、当研究室で開発している三次元共通構造特徴自動認識プログラムにより得られた共通構造と本モチーフ辞書の情報との比較・検討を行ないたい。

参考文献

[1] A.Bairoch, *Nucleic Acids Res.* ,**19**, 2241-2245 (1991).

[2] 宮田博之　他，第 30 回構造活性相関シンポジウム要旨集，107-108(2002).

[3] MDL Inc., http://www.mdli.com/

[4] H.Kato, and Y.Takahashi, *Bull. Chem. Soc. Jpn.*, **70**, 1523-1529 (1997).