情 報 学 基 礎 27－3
（1992. 9. 8）

演繹推論機能を用いた蛋白質立体構造解析システムPACADE

久原　哲[1]、佐藤賢二[2]、古市恵美子[3]、瀧口今日子[1]、高木利久[3]

[1]九州大学大学院農学研究科遺伝子資源工学専攻
[2]九州大学情報処理教育センター
[3]福岡女子短期大学
[4]東京大学医科学研究所ヒトゲノム解析センター

PACADEは演繹推論機能を応用した蛋白質高次構造解析用のデータベースシステムである。本システムは関係データベースと演繹推論システムDEEを結合したシステムであり、ルールの形で表現した構造モデルの検索を容易に行うことができる。また、データベースにはProtein Data Bankのデータの他、アミノ酸残基側鎖間の距離や二次構造間の角度などを格納している。本研究では演繹推論機能を用いて蛋白質の超二次構造の検索を行った結果について報告する。

A deductive database system PACADE for analyzing three dimensional and secondary structures of protein

Satoru.Kuhara[1], Kenji.Satou[2], Emiko.Furuichi[3], Kyoko.Takiguchi[1] and Takagi.Takagi[4]

[1]Graduate School of Genetic Resources Technology, Kyushu University, Fukuoka 812
[2]Educational Center for Information Processing, Kyushu University, Fukuoka 812
[3]Fukuoka Women's Junior College, Gojo, Dazaifu, Fukuoka 818-01
[4]Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo 108

We have developed a deductive database system PACADE for analyzing three dimensional and secondary structures of protein. The PACADE system consists of a relational database created from Protein Data Bank and a deductive engine DEE based on logic programming. It has the following features: (1) The system has an inference mechanism. This means by which users can easily write and check biological hypotheses using logical and declarative rules instead of procedural programs. (2) The relational database of the PACADE system stores data on both three dimensional and secondary structures of protein. The integration of this two level structure makes feasible an abstract representation of the protein structure. We describe herein the design, functions, and implementation of this PACADE system.

1．はじめに

蛋白質はその構造として1次、2次、3次これに加えて4次構造を持つことはよく知られている。1次構造（配列）については、１９６０年代から収集が開始されており、現在ではProtein Identification Resources(PIR)にまとめられ、急速に増加している。演者らはこの蛋白質の1次構造に遺伝子の1次構造を加えた配列解析システムGENASを九州大学大型計算機センターに構築している[1]。一方、蛋白質の立体構造データである結晶構造解析データはProtein Data Bank(PDB)に収集されている。このPDBにおいてもデータの集積速度は増加しており、１９９１年だけでも１５０件以上の登録がなされている。立体構造データの増加は蛋白質の構造機能解明の研究に拍車をかけるものと期待されている。この構造機能解明のためのデータベースとして現在までに、いくつかのデータベースシステムが構築されてきている。例えばBIPED[2]は関係データベース管理システムを用い、検索言語としてSQLを用いた蛋白質構造検索システムである。これに加えて、論理プログラミング言語であるPrologを質問言語として用いるシステムも構築されてきている。Morffew と Todd[3]はPrologを検索に応用し、Rawlingら[4]はPrologを構造のトポロジーの導出に用いている。最近はオブジェクト指向データベースも用いられてきておりGrayら[5]は質問言語としてPrologとDaplexを利用したオブジェクト指向データベースシステムを構築している。

しかしながら、これらの言語でも十分ではなく、SQLは再帰的検索が行えない、またPrologでは多量のデータが扱えない等の欠点を持っている。これらの短所を補う一つの方法として演繹データベースシステム[6,7]が研究されてきた。このシステムではPrologを用いたものとは異なり、全解探索が保証されているので論理的な部分を記述すればよく、記述がより簡単になる。またSQLと比べてもより詳細な質問の記述が可能である等の利点がある。

演者らはすでに演繹推論機能を持つ蛋白質構造データベースシステムPACADE[8]を構築している。このシステムはPDBのデータから1次、2次および3次構造のデータを関係データベースに格納し、演繹推論システムDEE[9]を検索に利用したシステムである。このシステムを用いて、構造に基づいた仮説をルールの形で書き、演繹推論機能を用い検証することが可能となっている。

本論文では、立体構造のトポロジカルモチーフとしてよく知られている超2次構造(Greek key, hairpin, meander, jelly roll等)[10] の検索を可能にするために、従来のPACADE に2次構造間の最短距離、角度を加え、超2次構造の検索を行った例を述べる。

2．システム設計

2.1 演繹データベースシステム

演繹データベースシステムは関係データベースと述語理論を組み合わせたシステムである。関係データベースは多量のデータを取り扱う部分を、述語理論の部分は質問の処理を受け持つことになる。

演繹データベースの推論機構はファクト、ルールおよび質問を基本としており、文法はPrologと同じもの用いている。例を挙げて説明すると

```
[F1] distance(6,14,4.631,'4ape').
[F2] distance(14,155,4.430,'4ape').
[F3] distance(74,83,4.024,'1sbc').
[F4] distance(83,86,4.865,'1sbc').
[R1] cluster(X,Y,P):-distance(X,Y,Dxy,P),'<'(Dxy,5).
[R2] cluster(X,Y,P):-cluster(X,Z,P),cluster(Z,Y,P).
[Q1] :-cluster(X,Y,'4ape').
```

F1, F2, F3およびF4はファクトである。"distance"は述語であり、引数は定数である。ファクトは関係データベースに格納されており、必要なときに推論機構によりデータベースから検索される。R1およびR2

はルールである。ルールはヘッドとボディーの部分からなる。特にR2は再帰的ルールとなっている。

演繹システムは関係データベースに格納されていない新しいファクトを生成できる。例えば、システムはR1のボディーの部分を満足する新ファクトF5, F6, F7およびF8を生成する。

[F5] cluster(6,14,'4ape').
[F6] cluster(14,155,'4ape').
[F7] cluster(74,83,'1sbc').
[F8] cluster(83,86,'1sbc').

また、R2を満足するF9とF10の新ファクトを生成する。

[F9] cluster(6,155,'4ape').
[F10] cluster(74,86,'1sbc').

推論機構は新ファクトが生成できなくなるまで繰り返し、最後に、質問を満足するファクトを利用者に解として返す。この例では、F5, F6 およびF9が解として返される。この例で、質問Q1については、ルールが再帰的であるのでSQLで置き換えることは不可能である。

このような点において、演繹データベースは従来のデータベースより有利であると考えられる。利用者は再起的な検索については再起的なルールを記述するだけでよく、解を得るためのコントロール部分を必要とはしない。この論理的な部分の記述だけで良いという点は試行錯誤的な検索を必要とされる蛋白質等の構造検索などに適していると考えられる。

2.2 システム構成

PACADEはSUNワークステーション上に構築されており、そのシステム構成を図1に示した。このシステムは3種類のファイル(Protein Data Bank file, Database, および Rule file) と3種類のモジュール(SYBASE, Extractor, およびDeductive Engine)から構成されている。

[Protein Data Bank file]はテキスト形式のファイルで種々の蛋白質構造データを格納している。データベースの整合性を保つためにDNA, RNAを除き, SEQRESレコード、2次構造のデータ、あるいはATOMレコードを持たないエントリーについても除外した。

[Database]データベースは関係データベースを用い、Extractor moduleによりProtein Data Baseから作成される。

[Rule file]ルールファイルはテキスト形式のファイルでDeductive Engineで使用するルールを格納している。

[Extractor]このモジュールはProetin Data Bankのデータを SYBASEの関係へ変換しデータベースへ格納する。データベースの整合性を保つため、insertion コードは除き、同じ理由により最初のalternation コードだけを採用した。

[SYBASE]SYBASEは関係データベース管理システムであり、Extractorによってデータベースが更新される場合には、関係を更新する。また、Deductive Engineからのコマンドによって、データベースを検索し結果をDeductive Engineに返す機能も有する。

[Deductive engine]Deductive Engine は2つのモジュールから構成されている。

(1) Rule Transformer

このモジュールは利用者が記述したルールをBottom-up Evaluatorがより効果的に処理できるようにルールの書き換えを行う。本システムではMagic set法およびNRSU法を採用している。[11, 12, 13、14]

(2) Bottom-up Evaluator

このモジュールはtrandformer からルールと質問を受取、semi-naive法を用いて評価を行い、解を利用者に返す。必要であればSYBASEを用いてデータベースを検索し、ファクトに変換する。

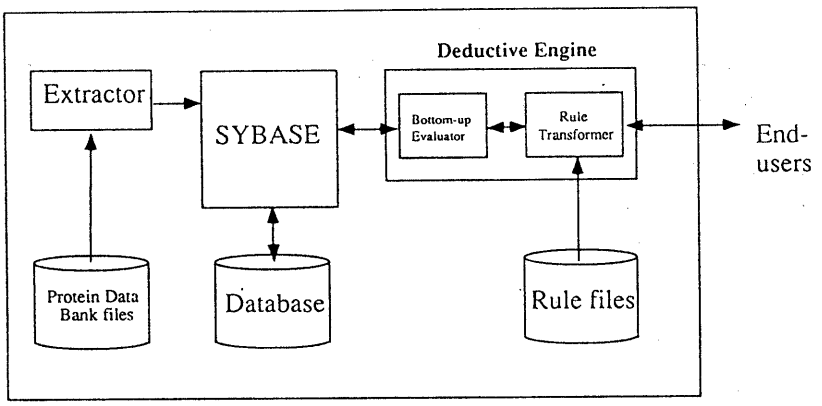

図１：PACADEのシステム構成。

２.３データ構造

PACADEにはすでに３次構造検索のため11の関係を作成していたが、これに加えて疎水性クラスター検索のためにhydrophobic_parameterを、超２次構造検索のために５個の新関係を追加した。

疎水性クラスター検索には、amino_acid, distanceおよびhydrophobic_parameterを使用した。利用した関係をファクト形式で例示すると次のようになる。

```
amino_acid(1, ser, 18, 23, '4ape').
distance(1, 2, 6.3, '4ape').
hydrophobic_parameter(gly, 0.10).
```

このファクトの引数の意味は次の通りである。

amino_acid：蛋白質4APEにおいて最初の残基はserであり、この残基は18番目から23番目の原子を有している。

Distance：蛋白質4APEにおいて最初の残基と2番目の残基間の距離が6.3 angstromである。

hydrophobic_parameter：gly残基の疎水性指標の値が0.10である。

メモリーの制約からdistanceに関するファクトは15 angstrom以内の距離を格納した。それ以上の距離の検索については短距離の組み合わせを検索することで代用している。

超２次構造の検索には、secondary_structure, min_distanceおよびangleの関係を使用した。ここで、ノードと呼ぶ単位を導入し、このノードの配列で蛋白質を表現した。ノードは２次構造を基にしたものでstrand, helixおよびrandom coilからなっている。使用した関係をファクトの形式で例示すると次のようになる。

```
secondary_structure(1, coil, 1, '3tln').
min_distance(2, 4, 4.401, '3tln').
angle(8, 10, 40.0, '3tln').
```

これらのファクトの意味としては

secondary_structure：蛋白質3TLNにおいて最初のノードはcoilで１残基を含む。

min_distance：蛋白質3TLNにおいて２番目と４番目のノードの距離が4.401angstromである。

angle：蛋白質3TLNにおいて8番目と10番目のノードの角度は40度である。

ここで、angleは２次構造間の角度の絶対値として表されている。２次構造間の角度はそれぞれの構造のＮ端側のＣαからＣ端側のＣαまでのベクトルの角度と定義した。

3．検索例

このシステムを利用した蛋白質の構造検索の例を示す。

3.1 疎水性 クラスターの検索

図２に今回使用したルールを示した。このルールの作成するにあったて採用した疎水性 クラスターの定義は以下の通りである。

(1)疎水性 指標[15]が2.0以上の残基を疎水性残基とする。

(2)２つの疎水性 残基が6angstrom以内にあるとき、その2残基間にhydrophobic connectが存在するとする。

(3)このhydrophobic connectが集合してhydrophobic clusterを形成る。

```
hydrophobic_connect(Xnum,Ynum,P):-
        hydrophobic_parameter(Xname,Xval),
        hydrophobic_parameter(Yname,Yval),'<'(2.0,Xval),'<'(2.0,Yval),
        amino_acids(Xnum,Xname,P),
        amino_acids(Ynum,Yname,P),
        distance(Xnum,Ynum,Dxy,P),'<'(Dxy,6.0).
```

図２：疎水性クラスター検索のルール

利用者が入力する質問は次の形式となる。

```
:- hydrophobic_connect(Xnum,Ynum,'4ape').
```

３種類の酸性プロテアーゼ(PDBのエントリーで、4APE, 3APPおよび2APR)にこのルールを適用した。その結果、３種類のプロテアーゼにおいては類似した形のクラスターが存在することが明かになった。酸性プロテアーゼにおいては、活性部位の領域に疎水性クラスターが存在すると考えられており、今回の結果はこの事実とよく一致していた。

このルールを５種類のセリンプロテアーゼに適用したが、共通のクラスターは検出できなかった。

3.2 超２次構造の検索

メンダーおよびロスマン構造を含む反復構造の検索

図３によく知られているβシート構造で構成されたヘアピン、メンダー、ロスマン構造を示した。これらの構造検索のためのルールを図４に示した。このルールでは、βシート構造だけではなくαヘリックス構造と混在した構造も検索可能なように作成した。ここでは、２つのノードの角度が120度から180度の範囲のものを逆平行と定義している。

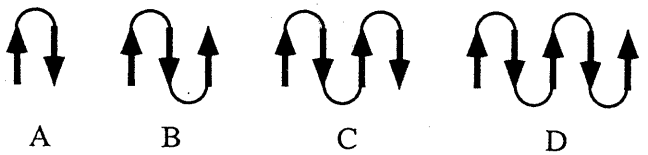

図３：超２次構造における反復構造。(A)はヘアピン構造；(B)はメンダー構造；(C)は４βシート構造のメンダー構造；(D)は５βシート構造のメンダー構造を表す。

```
meander_n(A,[B,C],3,P):-hairpin(A,B,P),hairpin(B,C,P).
meander_n(A,[B|L],Num1,P):-hairpin(A,B,P),neander_n(B,L,Num,P), Num1 = Num + 1.
hairpin(A,B,P):-not_coils(A,B,P),neighbour(A,B,P), double_anti_parallel(A,B,P).
not_coils(A,B,P):-not_coil(A,P),not_coil(B,P),B = A + 1.
not_coils(A,B,P):-not_coil(A,P),is_coil(A1,P),not_coil(B,P), A1 = A + 1,B = A1 + 1.
is_coil(X,P):-second_structure(X,coil,Xnum,P).
not_coil(X,P):-second_structure(X,strand,Xnum,P).
not_coil(X,P):-second_structure(X,helix,Xnum,P).
neighbour(X,Y,P):-min_distance(X,Y,Dxy,P),'<'(Dxy,20).
neighbour(X,Y,P):-min_distance(Y,X,Dxy,P),'<'(Dxy,20).
double_anti_parallel(X,Y,P):-anti_parallel(X,Y,P).
double_anti_parallel(X,Y,P):-anti_parallel(Y,X,P).
anti_parallel(X,Y,P):-angle(X,Y,Axy,P),'>'(Axy,120.0),'<'(Axy,180.0).
```

図４：反復構造検索のためのルール

バレル構造、特にグリークキーおよびジェリーロール構造の検索

図５に超２次構造の１種であるバレル構造と呼ばれるグリークキーおよびジェリーロール構造を示した。図６にこれらの構造を検索するためのルールを示した。このルールにおいてもβシートとαヘリックスの混在が可能な形となっている。このルールの作成にはRawlingsら[4]の表現法を参考にした。

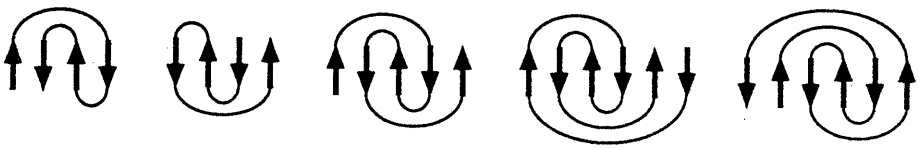

図５：超２次構造におけるバレル構造。(A)は４βシート構造のグリークキー構造；(B)は逆構造の４βシート構造のグリークキー構造；(C)は５βシート構造のグリークキー構造；(D)はジェリーロール構造；(E)は逆構造のジェリーロール構造。

```
greek_even_r(A,[],D,2,P) :- hairpin(A,D,P).
greek_even_r(A,L,D,Num1,P) :- not_coils(A,B,P), neighbour(A,D,P), double_anti_parallel(A,D,P),
        greek_odd(B,L1,D,Num,P), append([B],L1,L), Num1 = Num + 1.
greek_even_l(A,[],D,2,P) :- hairpin(A,D,P).
greek_even_l(A,L,D,Num1,P) :- not_coils(C,D,P), neighbour(A,D,P), double_anti_parallel(A,D,P),
        greek_odd(A,L1,C,Num,P), append(L1,[C],L),Num1 = Num + 1.
greek_odd(A,L,D,Num1,P) :- greek_even_r(A,L1,B,Num,P), greek_even_l(C,L2,D,Num,P), append(L1,[B],L),
        append([C],L2,L3), L = L3, Num1 = Num + 1.
```

図６：バレル構造検索のためのルール。

サーモライシンはαヘリックス構造から成るグリークキー構造を持っていることが知られている。図７に示したルールをサーモライシン(3TLN)に適用した。その結果、４αヘリックス構造から成るグリークキーが検索できた。同様にイムノグロブリンあるいはプレアルブミンにおいても、４βシート構造からなるグリークキーが検索できた。より複雑なジェリーロール構造もγクリスタリン(1GCR)から検索することができた。このγクリスタリン(1GCR)におけるジェリーロール構造を図７に示した。

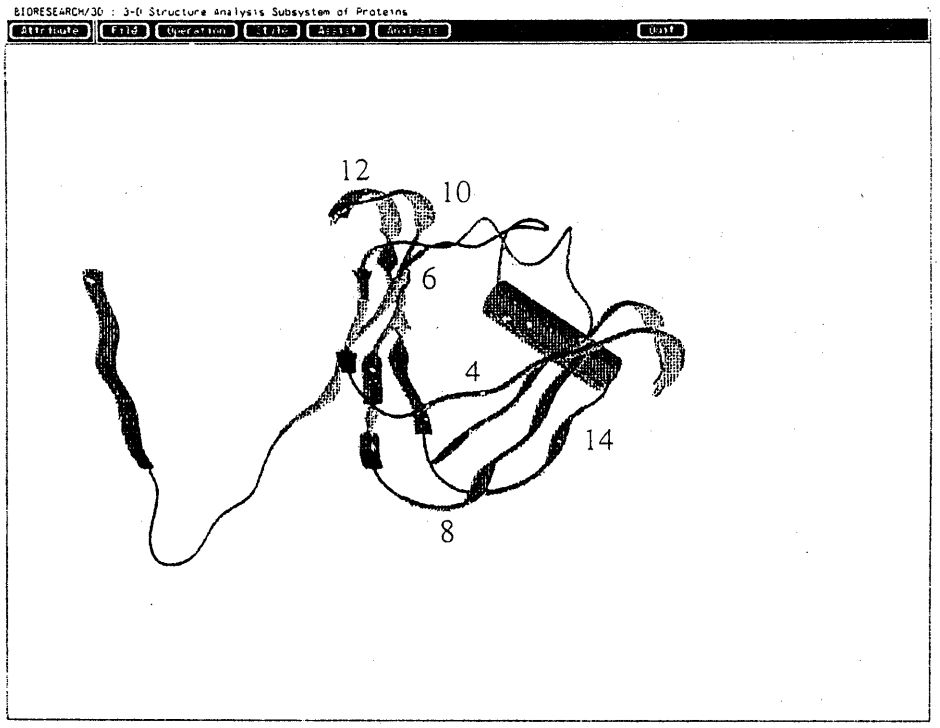

図７：γークリスタリン(1GCR)におけるジェリーロール構造。図中の番号はノードの番号を表している。

4．最後に

本論文では、演繹推論データベースシステムを用いた蛋白質構造検索の可能性を検討しきた。特に、構造に基づく仮説をルールという形で表現しデータベースの上で検証できる機能について３次構造、疎水性クラスターおよび超２次構造を用いて検討してきた。これらの例については図２、４および６で示したルールで検索が可能であり本システムの有用性が示された。

5．謝辞

PACADEの開発に当り多くの助言や協力をいただきました九州大学工学部の鈴木孝彦博士、五斗進氏に感謝いたします。本研究は文部省科学研究費重点領域研究「ゲノム解析に伴う大量知識情報処理の研究」の一貫として進められているものである。

参考文献

[1] Kuhara,S., Matsuo,F., Futamura,S., Fujita,A., Shinohara,T., Takagi,T. and Sakaki,Y. (1984) GENAS: a database system for nucleic acid sequence analysis. Nucleic Acids Res., 12, 89-99.

[2] Islam,S.A. and Sternberg,M.J.E. (1989) A relational database of protein structures designed for flexible enquiries about conformation. Prot. Engng., 2, 431-442.

[3] Morffew,A. and Todd,S. (1986) The use of Prolog as a protein querying language. Computers and Chemistry, 10, 9-14.

[4] Rawlings,C.J., Taylor,W.R., Nyakairu,J., Fox,J. and Sternberg,M.J.E. (1986) Using Prolog to represent and reason about protein structure. In Shapiro,E. (ed), Third International Conference on Logic Programming, Springer-Verlag, pp.536-543.

[5] Gray,P.M.D., Paton,N.W., Kemp,G.J.L. and Fothergill,J.E. (1990) An object-oriented database for protein structure analysis. Prot. Engng., 3, 235-243.

[6] Lloyd,J.W. (1987) Foundations of logic programming 2nd ed., Springer-Verlag, Berlin.

[7] Bancilhon,F. and Ramakrishnan,R.A. (1989) An amateur's introduction to recursive query processing strategies. In Mylopoulos and Brodie (eds), Readings in Artificial Intelligence and Databases, Morgan Kaufmann, pp.376-430.

[8] Kuhara,S., Satou,K., Furuichi,E., Takagi,T, Takehara,H. and Sakaki,Y. (1991) A deductive database system PACADE for the three dimensional structure of protein. In Proceedings of the 24th HICSS, Hawaii, pp.653-659.

[9] Takagi,T., Goto,S., Suzuki, T. and Ushijima, K. (1991) Applicability of a Deductive Database to CAD Systems. In Proceedings of Supplement 7th IEEE International Conference on Data Engineering, Kobe, Japan, pp.51-58.

[10] Richardson,J.S. (1981) The anatomy and taxonomy of protein structure. Advances in Protein Chemistry, Academic Press, pp.283-306.

[11] Beeri,C. and Ramakrishnan,R. (1987) On the power of Magic. In Proceedings of 6th ACM PODS, San Diego, California, pp.269-283.

[12] Bancilhon,F. and Ramakrishnan,R. (1988) Performance evaluation of data intensive logic programs. In Minker,J. (ed), Foundations of deductive databases and logic programming, Morgan Kaufmann, Los Altos, California, pp.519-543.

[13] Naughton,J.F., Ramakrishnan,R., Sagiv,Y. and Ullman,J.D. (1989) Efficient evaluation of right-, left-, multi-linear rules. In Proceedings ACM SIGMOD'89, pp.235-242.

[14] Kemp,D., Ramamohanaro,K. and Somogyi,Z. (1990) Right-, left-, and multi-linear rule transformations that maintain context information. In Proceedings of Far-East Workshop on Future Database System, Melbourne, pp.38-52.

[15] Zimmerman,J.M. (1968) The characterization of amino acid sequences in proteins by statistical methods. J. Theoret. Biol., 21, 170-201.