演繹データベースにおける類似検索の効率化について

3F－9

橋本真一＊　佐藤賢二＊＊　牛島和夫＊

＊九州大学工学部情報工学科　＊＊東京大学医科学研究所ヒトゲノム解析センター

1. はじめに

我々は、タンパク質の立体構造データを検索するための演繹データベースシステムPACADEを開発している。PACADEは通常の検索機能に加えて、類似した解を検索する類似検索機能を備えている。本稿ではマジックセット法を用いた類似検索の効率化について述べる。PACADEの類似検索機構では、与えられたルールを機構側で書換える。一方、マジックセット法は、質問処理の効率化のためにルールの書換えを必ず行う。よって、まずこれら二つのルール書換え手法を併用しても解が変わらない、すなわち質問等価であることを形式的に証明する。次に、どの程度効率が上がるかを、タンパク質中の類似部分構造を検索する例を用いて測定した結果を報告する。

2. PACADEと類似検索の概要

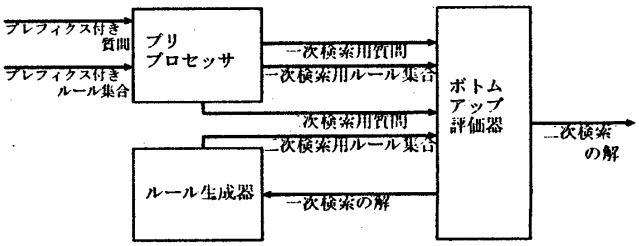
PACADEは、演繹推論機能を備えたタンパク質三次構造解析用のデータベースシステムであり、Protein Data Bankが提供するタンパク質の構造データを格納している。ユーザは、求める部分構造をルールとして記述することにより、データベース中に陽に記述されていない部分構造を検索できる。しかし、ユーザがルールで記述した部分構造を各種のタンパク質に対して検索する機能だけでは、タンパク質の機能と構造との相関関係を調べるためには不十分である。よって、機能と構造の関係をより明らかにするためには、ユーザが着目する機能を持つ特定のタンパク質の部分構造と良く似た部分構造を他のタンパク質に対して検索する機能が必要である。これを実現するために、我々はPACADE上で類似検索機構を開発した[1]。

PACADEの類似検索機構は、「多くの場合類似性の指標はルール中の変数の束縛値の中に含まれる」という観察に基づいて開発された。推論の途中で束縛される変数の値は解の特徴を表しているので、推論の過程でこれを記憶することによって解を特徴付け、類似した特徴を持つ解を探すことができる。ここで、PACADE上での類似検索を次のように定義する。

ユーザはシステムに質問とルール集合を与える。この時三つの付加的情報: 類似性の指標、許容誤差範囲、目標ドメインを、ある種の注釈（以下プレフィクス）として質問およびルールの中に記述する。類似性の指標は「何をもって似ていると認めるか; 類似の判断にどの変数が関係するか」許容誤差範囲は「どこまで似ていると認めるか」そして目標ドメインは「類似元がどこにあるか」を、それぞれシステムに示す。

まずシステムは質問中のプレフィクス付き定数によって限定される目標ドメイン中に解を探す。次にその定数を変数と置き換えることによって、先ほど見つかった解（類似元）と似た解を目標ドメインよりも広い探索空間で探す。

この類似検索システムは、ユーザの質問一回に対して二回検索を行う。タンパク質の類似構造検索を行う場合、ユーザの入力は、質問、ルール、類似性の指標、許容誤差範囲、目標ドメイン（着目するタンパク質名）、である。これらの入力を受けると、システムはまず着目するタンパク質に対して検索を行い、見つかった部分構造が推論の途中でどういう変数束縛を受けてきたかを記憶しておく（第一次検索）。次に、他のタンパク質に対して同じ部分構造を探す際に、記憶した束縛情報を用いて、第一次検索の結果見つかった部分構造と類似した部分構造を検索する（第二次検索）。最終的にシステムは、ユーザがルールで記述した部分構造のうち、着目するタンパク質中で見つかった部分構造と似た部分構造を他のタンパク質に対して検索した結果を解として出力する。下図はシステムの概要を示している。



例えば、第一次検索を行う前にユーザが与えた質問とルールは以下のように変換される。

変換前　　$:\!-anc\ (X, Y, @jpn)\ .$

変換後　　$:\!-anc\ (X, Y, jpn, R)\ .$

ここでancは述語、X,Y,Rは変数、jpnは定数である。@は質問中の定数に対するプレフィクスである。

変換前

$anc(X, Y, N) :\!- par(X, Y, \&+2-2\&Age, N).$

変換後

$anc(X, Y, N, [R0]) :\!- par(X, Y, Age, N),$
$make\_string\_rule([\text{"}anc(X, Y, N) :\!- par(X, Y, Age, N)\text{"},$
$\text{"}, Age > -2+\text{"}, Age, \text{"}, Age < 2+\text{"}, Age, \text{"}.\text{"}], R0).$

ここで&+2−2&はルール中の変数に対するプレフィクス

Studies on Optimization of Similarity Search in a Deductive Database.

Shin'ichi HASHIMOTO*, Kenji SATOU** and Kazuo USHIJIMA*

*Department of Computer Science and Communication Engineering,Kyushu University.

**Human Genome Center,Institute of Medical Science,Tokyo University.

である。組み込み述語 *make_string_rule* (以下 *msr*) は解の導出過程を記憶するために付け加えられる。

3. マジックセット法によるルール書換え

ある論理プログラムの論理的帰結を求める手段として、ボトムアップ評価とトップダウン評価という二つの方法がある。ボトムアップ評価は、質問の解になり得るものを全て計算し生成する。この方法は簡単に実装できるが、質問中の定数束縛を有効に使うことができないために、無駄な計算を行ってしまうことが多い。一方トップダウン評価は、質問に関連する部分だけを計算するので、ボトムアップ評価が行うような無駄な計算をしない。しかし、無限ループに入ってしまい評価が終わらない可能性がある。

この問題を解決するために開発されたのがマジックセット法である[2, 3]。マジックセット法ではまず論理プログラムを、トップダウン評価の動きを模倣するプログラムに書換える。この時与えられた質問中の定数束縛を用いる。そして書換えたプログラムをボトムアップ評価するという方法である。この方法はトップダウン評価を模倣するため、与えられた質問と関係する部分だけを計算するので効率的であり、しかもボトムアップ評価するので簡単に実装可能である。このように、マジックセット法はボトムアップ評価とトップダウン評価の利点を組合せ、欠点を補った方法である。演繹データベースでは、上で述べた論理プログラムのことをルール集合と呼ぶ。マジックセット法によるルール集合の書換えをマジックセット変換という。

4. 類似検索とマジックセット変換

4.1 第一次検索とマジックセット変換

第一次検索のための質問には、目標ドメインを表す定数が必ず存在する。また、マジックセット変換は質問に定数束縛がある場合に、前もってその束縛を有効に利用できるようにするためのものである。どちらも質問に定数束縛があるという前提の下で使用するため、これら二つを合わせて使うことができると我々は考えた。

プレフィクス付きルールを第一次検索用ルールに変換する変換規則により、各ルールボディの右端には必ず述語 *msr* を持つリテラルが置かれる。その第一引数に置かれる文字列定数は、変換前のルールの形を正確に表している (但しプレフィクスは除く)。よってルールボディの評価順序が左から右であるという仮定の下では、第一次検索用のルールを評価する各ステップにおいて、新しく導出される中間ファクトの最後の引数は、その中間ファクトの導出過程を正確に保持していることが保証される。

しかし第一次検索用ルールをマジックセット法で書き換えると、述語 *msr* の引数に文字列として埋め込んだルールの形と、その述語をボディに持つルールの形とが異なってしまう。故に、第一次検索のための変換とマジックセット変換が併用可能であるかどうかは自明ではない。

4.2 質問等価性の証明

命題 マジックセット変換を行わない場合に第一次検索用質問とルールの評価が停止し解を返すならば、マジックセット変換を行った場合にも第一次検索用質問とルールの評価も停止し同じ解を返す。

参考文献[2] 中にあるマジックセット法自体の正当性の証明を参考にして、上の命題の証明を行った。証明は誌面の都合で省略する。

5. 実験

実際にタンパク質の類似構造検索の一次検索において、第一次検索用に変換したルール集合をマジックセット変換した場合、そのマジックセット変換によってどれほど効率化がなされるかということについて実験を行った。ハードウェアは Heliostation 1030 (メモリ 64MB) で、推論にかかる時間だけを測定するために、検索に必要なファクトを全部メモリ上に事前にロードし、ディスクアクセスにかかる時間を排除した。

134 種類のタンパク質の構造データに対して、マジックセット変換を併用した第一次検索と、併用しない第一次検索を行った。以下の表では前者を magic、後者を no-magic として示している。マジックセット変換を併用した場合、約 400 倍の効率化がなされた。

	合計	平均 (合計／134)
ファクト数	132802	991
解の個数	2498	18.6
magic (秒)	145.28	1.08
no-magic (秒)	58372.7	435.6

6. おわりに

本稿では、類似構造検索における第一次検索用のルール変換とマジックセット変換とが併用可能であることを示した。また、タンパク質の立体構造データと、タンパク質中に含まれる類似部分構造を検索するルールを用いて、実際にマジックセット変換を行った場合と変換を行わない場合についてそれぞれ評価時間を測定し、比較した。その結果、マジックセット変換を行うことによりかなり効率が改善された。これは、タンパク質の立体構造解析という応用分野における本証明の有用性を端的に示している。今後の課題として、より制限条件を少なくした場合の証明、第二次検索の効率化、演繹データベースにおける他の類似検索への応用、などを挙げることができる。

参考文献

[1] Satou,K., Furuichi,E., Takagi,T., Kuhara,S. and Ushijima,K.: Similar Structure Search in a Deductive Database, *In Proceedings of International Symposium on Next Generation Database Systems and Their Applications*, pp.130-137, 1993.

[2] Ullman,J.: Principles of Database and Knowledgement-Base Systems volume I and II, COMPUTER SCIENCE PRESS, 1989.

[3] 森下真一: 知識と推論, 共立出版, 1994.