ICOT Technical Memorandum: TM-0739

TM-0739

DNA配列知識ベースシステム(KNOA)
論理型言語アプローチ—

小長谷明彦，横田　実(日本電気)

July, 1989

ICOT

Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan

(03) 456-3191～5
Telex ICOT J32964

Institute for New Generation Computer Technology

ＤＮＡ配列知識ベースシステム（ＫＮＯＡ）

DNA Sequence Knowledge Base System (KNOA)

―論理型言語アプローチ―

-Logic Programming Approach-

小長谷　明彦　　横田　実
Akihiko KONAGAYA　Minoru YOKOTA

日本電気株式会社Ｃ＆Ｃシステム研究所
NEC Corporation C&C Systems Research Laboratories

１．はじめに

近年、人の遺伝子解析プロジェクト[1]に代表されるようにＤＮＡ配列データの収集が急速に進められており、そのデータ量の多さと複雑さからＡＩ手法によるＤＮＡ配列データの解析に関心が寄せられている。ＤＮＡ解析は知識情報処理の観点からは２つの側面を持つ。一つは大量のＤＮＡ配列データの検索であり、もう一つは、蛋白質の構造予測に見られるような高度な推論である。

一方、第五世代計算機プロジェクトでは、１９８２年以来、論理型言語に基づく知識情報処理を目的とした推論マシンの研究開発が進められている[2]。論理型言語の特徴は推論と検索の両方の機能を備えている点にあり、この意味でＤＮＡ解析問題は、第五世代計算機に適した応用問題といえよう。

我々は、第五世代計算機プロジェクトの一環として論理型言語による知識表現、推論方式、知識ベース構築法に取り組んでおり、その応用例の一つとしてＤＮＡ配列知識ベースシステムＫＮＯＡを逐次型推論マシンＣＨＩ[3]上に試作した。本稿では、ＫＮＯＡの経験を基にＤＮＡ解析問題における論理型言語の有効性、推論マシンの有効性ならびに今後の課題について論じる。

本稿の構成は以下の通りである。はじめに、２節において、ＤＮＡ解析問題に関する簡単な説明を行う。次に、３節においてＤＮＡ解析問題における論理型言語と推論マシンの利点について述べる。そして、４節において、試作したＤＮＡ配列知識ベースシステムについて報告し、最後に５節において今後の課題について論じる。

２．ＤＮＡ解析問題

ＤＮＡとは４種の核酸塩基（ヌクレオチド）が２重螺旋を描いて結合した高分子であり、生命の根源である遺伝子を保持していることで知られている。ＤＮＡ内の遺伝子情報はＲＮＡにコピーされ、ＲＮＡから一本のアミノ酸鎖が合成され、さらにアミノ酸鎖が特異的な立体構造に折れ畳まれて蛋白質となる。

ＤＮＡ解析問題では、ＤＮＡがどのような遺伝子を持ち、どのような蛋白質を生成し、遺伝子が種の間でどのように変化しているかを扱う。これらを解析することにより、進化過程の解明、種の分類、遺伝病の治癒、医薬品の開発等様々な分野に役立てることができる。

ＤＮＡ解析問題の中で、計算機は様々な側面で利用されているが、我々が知識処理応用の観点から着目しているのは次に述べるホモロジー検索と立体構造予測である。前者はＤＮＡデータバンクの検索を、後者はＤＮＡ（アミノ酸）配列から蛋白質の立体構造の予測を行う。

（１）ホモロジー検索[4]

ホモロジー検索は、検索対象となるＤＮＡ配列あるいはアミノ酸の配列と似た構造を持つ配列をＤＮＡデータバンクから捜し出す操作である。現在、約２万種類のＤＮＡ配列がデータベースに登録され、未確認配列の発見、種の分類、進化過程の解明に利用されている。

ホモロジー検索が通常のデータベース検索ともっとも異なる点は、その目的が検索対象と一致する配列を見つけることではなく、類似性のある配列を検索する点である。この意味で、ホモロジー検索はあいまい検索の一種といえよう。一般的なホモロジー検索のアルゴリズムでは、ＤＮＡ配列あるいはアミノ酸配列を一次元の文字列に見立ててマ

ッチングを行い、一致する要素の個数あるいは割合で類似度を判定する。ただし、このマッチングにおいては要素の順番を入れ換えてはいけないが適当な空白（ギャップと呼ばれる）を挿入して文字列をずらすことが許されている。例えば、図１の例では、ギャップを挿入しなければ一致する要素の数は高々２つであるが、ギャップの挿入により半数以上の要素を揃えることが可能となる。このようなギャップの処理を可能とするために、ホモロジー検索では動的プログラミング法や最長文字列を優先的に一致させる方法が使用されている。

論理型言語の応用の観点からのホモロジー検索に対する興味は以下の２点である。一つは、ＤＮＡデータバンクのような大規模なデータを論理型言語でどのように取り扱うかである。ＫＮＯＡでは論理型言語の機能を活用するためにＤＮＡデータバンク全体をクローズデータベースで実現した。これの利点については３節で詳述する。

もう一つは論理型言語の機能がホモロジー検索においてどのように生かせるかである。ＫＮＯＡの経験によれば、文字列レベルでの比較においては特に「推論」等の高度な機能を必要としない。しかしながら、これは現在のホモロジー検索アルゴリズムが極めて「荒い近似」をしているからである。本来比較しなければならないのは、一致する要素の個数ではなく、各々のＤＮＡ配列から生成される蛋白質の形状のはずである。したがって、ＤＮＡ配列の類似性を立体構造レベルで比較するような高次ホモロジー検索においては論理型言語の推論機能が活用できると予想される。これを実現するために必要な技術が次に述べる立体構造予測である。

（１）挿入前

```
a w g k v g a h a g e y l a e a l
|           |
a l w g k v n h g e v g g e a l
```

（２）挿入後

```
a - w g k v g a h a g e y l a e a l
|   | | | |   | | | |     | | | |
a l w g k v n - h - g e v g g e a l
```

図１　ギャップの挿入処理

（２）立体構造予測[5]

立体構造予測とは、与えられたＤＮＡ配列またはアミノ酸の配列から対応する蛋白質の立体構造を予測することである。ＤＮＡ配列から蛋白立体構造の予測には、大きく分けてエネルギー計算による予測と経験的パターンによる予測の２つのアプローチがある。

エネルギー計算法とは分子間に働く様々な力（イオン結合力、水素結合力、バンデールワース力）についてその極小値を計算することにより、蛋白質の安定構造を求めようという方法である。これは、完全にシミュレーションすれば正確な構造が計算できるが、数万の分子量を持つ蛋白質においては計算量が膨大であり、計算機パワーのさらなる増強とアルゴリズムの改良が必要とされている。

我々が興味を持っているのは、むしろ、後者の経験的パターンによるアプローチである。現在、約２００種の蛋白質についてはその立体構造がわかっており、どういうアミノ酸の列がどのような構造をとるか調べることができる。経験的パターンによるアプローチでは、この対応関係を解析し、特徴をパターンとして抽出し、蛋白質の構造を予測するための推論規則とする。

アミノ酸配列から立体構造を直接求めるパターンを抽出するのは困難なので、立体構造の予測を１次構造（アミノ酸配列）、２次構造（部分構造）、３次構造（立体構造）と段階的に行うのが一般的である。ここで、２次構造とは、蛋白質を特徴的な部分構造に分解したもので、通常はαらせん（らせん構造を構成する部分）、βシート（面構造を構成する部分）、ターン（折れ曲がっている部分）の３種とそれ以外の部分に分類される（図２）。

立体構造予測問題において我々は２つの立場を取っている。一つは論理型言語の機能が実応用においてどの様に利用できるかを評価することである。もう一つは、どの様なＡＩ技法が構造予測問題に適用できるかを研究することである。立体構造予測問題は知識処理応用の立場からは非常に興味深い問題であり、知識表現、推論方式、学習方式等様々な観点からのＡＩ技法の適用を考えている。

図２　２次構造例（Lactate Dehydrogenase）

３．　論理型言語アプローチの有効性

論理型言語はその言語の持つ性質上、ＤＮＡ解析の有用なツールとなる可能性を秘めている。ここでは、ＤＮＡ解

析問題における論理型言語および推論マシンの有効性について述べる。

3．1　論理型言語の有効性

論理型言語の最大の利点はその記述力の高さおよび生産性の高さにある。これは、ラピドプロトタイプを可能とする。また、ＤＮＡ解析問題では、クローズデータベース機能がＤＮＡデータバンクの実装に活用できる。

（1）ラピドプロトタイピング

論理型言語のプログラミング言語としての特徴はプログラム作成期間の短さにある。その理由として、データ構造や制御のレベルが高いこと、段階的にシステムが構築できること、対話的にデバッグできること等があげられる。また、ＤＮＡ解析問題では、論理型言語の持つデータベース機能（クローズデータベース）や推論機能を直接利用できることが開発期間の短縮に貢献している。先端的なＤＮＡ検索システムの研究にはスクラップアンドビルドが不可欠であり、ラピドプロトタイピングが可能な論理型言語は研究を大きく促進しよう。

（2）クローズデータベース

論理型言語が他のプログラミング言語ともっとも異なる点はクローズデータベースという高度なデータベース機構を言語内に内蔵していることである。次にＤＮＡデータバンクの実装においてクローズデータベースのどのような点が有効に活用できるかを示す。

クローズデータベースの第１の利点は文字列やリストといった可変データ構造が扱えることである。ＤＮＡデータバンクでは配列データや遺伝子の機能情報のように不定長あるいは不定個のデータが多数存在する。このようなデータは通常のデータベースでは扱いが困難となるが可変データ構造により容易に実現できる。

第２の利点は述語を用いて高度な検索機能がプログラムできることである。ＤＮＡには配列に付随して様々な機能情報が付加されているが、ユニフィケーションおよびバックトラック機能を利用することにより、これらの付加情報を活用するプログラムを容易に記述することができる。

第３の利点はクローズデータベースの拡張性にある。論理型言語では、データベースの検索とプログラムの呼び出しの区別がないため、クローズデータベースの機能を容易に拡張することができる。これにより、クローズデータベースを使って、様々なＤＮＡデータバンクのモデルを実験することができる。

3．2　推論マシンの有効性

以上に述べたように、論理型言語はＤＮＡ配列解析の道具としての優れた機能を有しているが一般に実行速度、メモリ使用量の点で従来言語よりも不利とされている。我々は、推論マシンＣＨＩを利用することにより、性能面でも十分実用的に利用できることを確認した。

（1）メモリ容量

論理型言語アプローチの問題点の一つはクローズデータベースで本当にＤＮＡデータバンクを実現できるか否かである。例えばGenBank 58.0[6]では配列数は約２万、総塩基数は２６００万塩基であり、１配列の最大塩基数は１７万塩基にものぼる。ＫＮＯＡの実装では、遺伝子に関する機能情報を含めて約２０万個のクローズ、３６Ｍバイトのヒープ領域を必要とした。これは、３２０Ｍバイトの主記憶を持つＣＨＩでは全く問題にならない量である。

ただし、ＤＮＡデータバンクに登録されている塩基数は近年急激に増加しているため、将来的には５節で述べるようにクローズデータベースの拡張が必要である。

（2）処理性能

処理性能に関する最大の興味はホモロジー検索を論理型言語で実行した際にどれだけの時間がかかるかである。仮にGenBank全体に対して検索を行うとすると、対象となる塩基数は現在約２６００万塩基である。遺伝子格納領域にある塩基列をアミノ酸に変換してから検索したとしたとすると検査対象となるアミノ酸の数は約３００万個となる。

ＫＮＯＡではいくつかのレベルのホモロジー検索機能を提供しているが、もっとも、高速なホモロジー検索では全アミノ酸配列を対象としても約１０分以内で検索することができる。これは、汎用ワークステーション上の手続き型言語で実現されたホモロジー検索システムと同等以上の性能である。推論マシンがホモロジー検索においてこのような高性能を発揮できる理由として、ＣＨＩのプロセッサの高速性[3]もあるが、汎用ワークステーション上のシステムは主記憶容量の制限から十分な性能を発揮できないことも一因している。いずれにせよ、上記の結果は推論マシンがＤＮＡ解析問題の研究のツールとして性能的にも十分実用に耐えうることを示したといえよう。

4．　ＫＮＯＡ

ＤＮＡ配列知識ベースシステムＫＮＯＡは、ＤＮＡ配列

解析問題における論理型言語アプローチの有効性を実証するために推論マシンＣＨＩ上に試作した知識ベースシステムである。ＫＮＯＡは階層型ＤＮＡデータバンクを中心に、ホモロジー検索システムと２次構造予測システムが有機的に組み合わされた構成となっている。以下、その特徴について述べる。

（１）　階層型ＤＮＡデータバンク

階層型ＤＮＡデータバンクの最大の特徴は、カテゴリ単位にＤＮＡ配列情報（知識）を管理できる機構を設けたことである。各カテゴリは継承機能を持ち、データの実体をコピーせずに抽象的なカテゴリを定義することが可能である。例えば、図３に示すように「霊長類」と「霊長類以外のほ乳類」を継承した「動物」というカテゴリを定義した

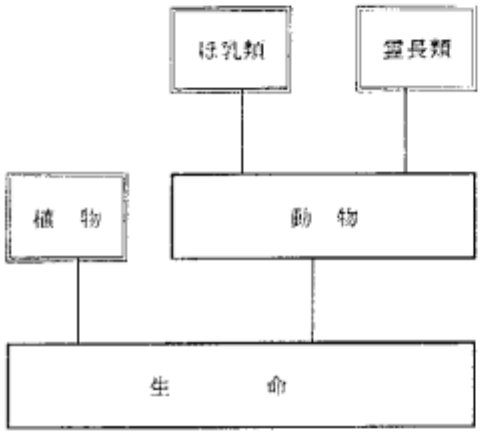


図３　階層型ＤＮＡデータバンク

り、さらに、「植物」を加えて「生命」というカテゴリを定義することができる。これにより冗長なコピーをせずに様々な検索空間を定義することができる。階層化の実現にはＣＨＩの多重名前空間機能[7]を利用した。

ＤＮＡ配列情報としては、塩基配列の他に、配列の登録日、出典、著者、配列の長さ、遺伝子の機能情報等が格納されている。

（２）　ホモロジー検索システム

ホモロジー検索に関しては、高速スクリーニング機構、アライメント位置表示システム、ドットマトリックスシステムの３つのシステムを実装した。高速スクリーニング機構は高速なホモロジー検索を実現するためのものであり、連続する２つのアミノ酸の出現パターンの頻度で類似度の計算を行なう。アミノ酸の種類は２０種なので、長さ２のアミノ酸列のパターンは４００種となる。出現頻度では、

図５　ドットマトリックス例
（縦軸　αヘモグロビン 横軸　βヘモグロビン）

アミノ酸配列の順序情報が失われてしまうので、類似性のないアミノ酸配列が含まれる可能性が高いが、アミノ酸配列の平均長は約３００であり、長さ２の出現頻度でも十分良い近似となる。

アライメント表示システムでは、最長一致文字列優先アルゴリズムを用いて配列のアライメントを行ない、挿入されたギャップの位置や対応するアミノ酸の位置の表示を行う（図４）。ドットマトリックスシステムではさらに全ての部分列の組合せに対する類似度を見ることができる。図５において、長い対角線は、その部分で強い類似性があることを示している。

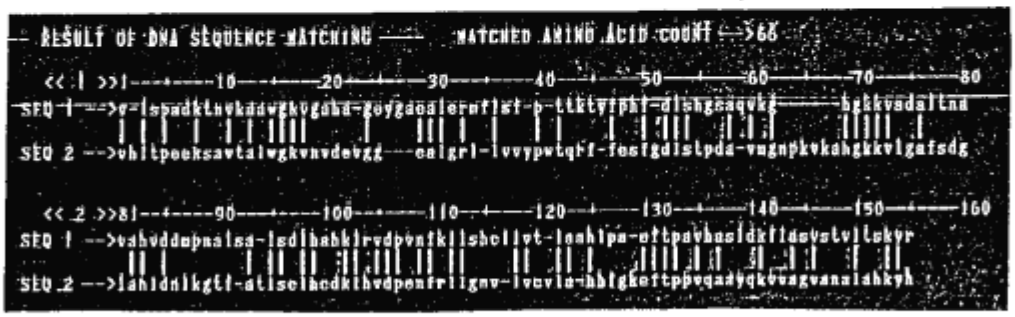


図４　アライメント例　（SEQ 1 αヘモグロビン，　SEQ 2 βヘモグロビン）

（3） 2次構造予測システム

立体構造予測システム実現の手がかりとして、アミノ酸配列からターンの位置を予測するエキスパートシステムを構築した。採用した推論規則は文献[8]による。推論規則はα／α型（αらせんの多い）の蛋白質の場合は約30個となっている。現在、立体構造の位置が判明している蛋白質について適用し、推論規則の妥当性を評価中である。予測結果の正当率は60～70%程度であり、推論規則に改善の余地が多いことを示している。

2次構造予測システムの表示画面の例を図6に示す。図中の星印が推論によりターンがあるとされた場所である。"h"はαらせんの位置、"t"がターンの位置、"－"がそれ以外の部分を表している。また、推論規則の妥当性を判定するために、立体構造が判明している蛋白質については主パス（アミノ酸配列の順にアミノ酸の中心位置を結んだ線）の立体透視図ならびにアミノ酸配列の立体表示モデルを表示し、推論結果との比較が容易にできるようにしている。

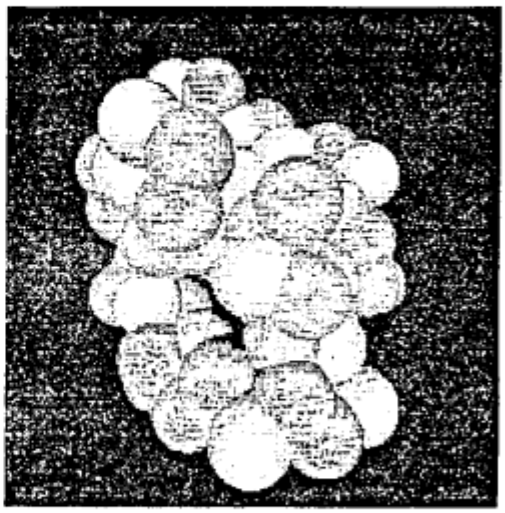
5．今後の課題

KNOAでは、DNA解析に必要な一部機能を実現しただけであり、実用利用にはさらなる改良拡充が必要である。逆に、論理型言語アプローチに基づきDNA解析問題を追求することにより論理型言語に必要な機能が浮かび上がってくる。このような課題としてはオブジェクト指向概念の導入、並列化、クローズデータベースの2次記憶化があげられる。

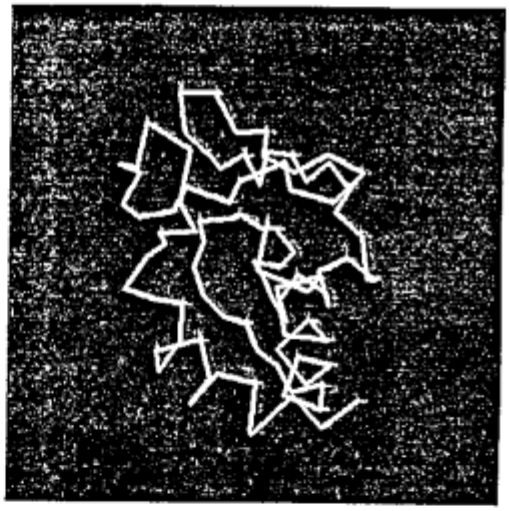
（1）オブジェクト指向概念

オブジェクト指向概念導入の目的は2つある。一つは、DNA配列データバンクの自然な反映であり、もう一つはイメージデータへの対応である。KNOAでは、論理プログラミングというパラダイムを追求するため、DNA配列の各情報はクローズの集合として表されている。しかしながら、DNA配列検索の目的は、配列の個々の情報を取り

アミノ酸配列立体表示



主パス（アミノ酸中心座標）



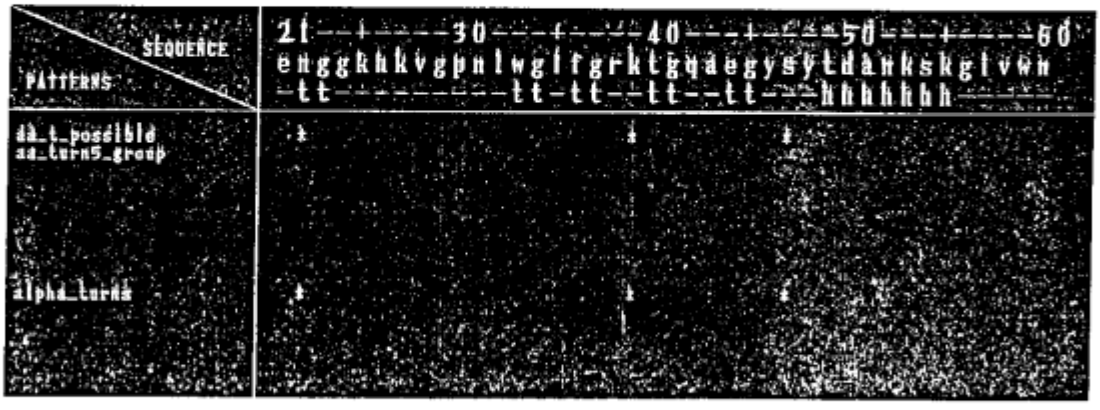


図6 2次構造予測例 (Cytochrome c' 31-78)

出すことよりも、配列そのものの検索にあると考えれば、ＤＮＡ配列という実体（オブジェクト）を導入し、配列情報をオブジェクトの属性と扱う方がより自然である。また、立体構造情報を扱うためには、イメージデータやグラフィックデータの扱いが不可欠である。これらの情報も操作性や処理の独立性を考慮するとオブジェクトとみなす方がより自然である。論理型言語にオブジェクト指向概念を導入するアプローチはこれまでにも試みられているが[9]、ＤＮＡデータバンクのような知識ベースの構築においてはデータベースの観点から論理プログラミングとオブジェクト指向概念を融合させる研究が必要であろう。

（２）並列化

ＤＮＡデータバンクは現在は２６００万塩基であるが、年間５００万－１０００万塩基の増加があり、さらに、近年その増加量はますます増加する方向にある。データ量の伸びが計算機の性能向上よりも速いとすると、ホモロジー検索の所要時間は次第に長くなり、最後には耐えきれないものになってしまう可能性がある。この問題を解決するためには、並列処理の導入が不可欠である。

ＤＮＡ配列検索は配列間の並列性や配列内での並列性と様々なレベルで並列性を引き出すことができるため、並列記号処理の研究材料としても着目できる。ただし、データ量が非常に多いため、実用レベルの性能を実現するためには並列度の向上だけでなくデータの供給方式やデータの分散化等についても十分検討する必要があろう。

（３）２次記憶クローズベース

ＫＮＯＡでは、クローズデータベースを利用することにより、高性能かつ高機能な検索を行うことができた。しかしながら、今後、ＤＮＡ配列データが増大し、主記憶に入りきれなくなったときにどのようにＤＮＡデータバンクを実現するかは大きな問題である。人の遺伝子解析プロジェクトでは３０～４０億塩基あるという人のＤＮＡを全て解明することを目的としている。ＤＮＡデータバンクがこのように増加した場合にはこれらのデータを全て主記憶上に展開するというアプローチは現実的でない。また、データの永続性や信頼性からも、データベースの本体は２次記憶上に保存されるほうが望ましい。

これを解決する手法の一つとして、クローズを内部表現のまま２次記憶中に格納できるようなクローズデータベースが考えられる。実現には解決すべき問題点が多いが、もし実現されればその応用範囲は広いといえよう。

６．まとめ

論理型言語によるＤＮＡ配列検索システムの実現についてその可能性と今後の課題を示した。論理型言語を用いれば、従来、個別に開発されてきたＤＮＡ配列情報のための関係データベース、ホモロジー検索システム、構造予測システム等を全て統一的に扱うことができ、性能面でも推論マシンでは、従来型のシステムと同程度以上の性能を達成することが可能である。実用化には、解決すべき点が多いが、先端的なＤＮＡ配列検索システムの研究には極めて有効なツールとなろう。

第五世代計算機プロジェクトでは、高度並列知識処理を実現するために並列オブジェクト指向言語を研究しており、今後はこのような観点からもＤＮＡ配列検索システムの研究を進めてゆく予定である。

謝辞

本研究の機会を与えてくださったＩＣＯＴ内田室長、Ｃ＆Ｃシステム研究所の石黒所長、大野部長に深謝いたします。また、ＣＨＩおよびＫＮＯＡの開発に携わった幅田主任、新部員をはじめとする多くの人々に感謝いたします。

参考文献

[1] Mapping our Genes, Johns Hopkins Univ. Press.

[2] Uchida,S. et al, "Research and development of the parallel inference system in the intermediate stage of the FGCS project", FGCS'88, 1988

[3] 幅田　他，"逐次型推論マシンＣＨＩ－Ⅱの性能評価"，情処計算機アーキテクチャ研究会，1989

[4] 金久，"タンパク質構造データベースと多変量解析"，ﾌﾟﾛﾃｨﾝｴﾝｼﾞﾆｱﾘﾝｸﾞ，化学増刊１１３，化学同人，1988

[5] 西川，"タンパク質の構造予測－現状と実現可能性"，ﾌﾟﾛﾃｨﾝｴﾝｼﾞﾆｱﾘﾝｸﾞ，化学増刊１１３，化学同人，1988

[6] Bilosfsky,H.S.,Burks,C.,"The GenBank(R)Genetic Sequence Data Bank", Nucl.Acids Res., vol.16, 1987

[7] Ito,H. et al,"Knowledge Base System in Logic Programming Paradigm", FGCS'88, pp.37-53, 1988

[8] Cohen,F.E. et al,"Turn Prediction in Protein Using a Pattern-Matching Approach", American Chemical Society, Vol.25, No.1, 1986

[9] 小長谷,"型付きユニフィケーションとクローズの対象指向解釈について", ｺﾝﾋﾟｭｰﾀｿﾌﾄｳｪｱ, vol.4, no.1, 1987