TR-725

MASCOT: 3次元ダイナミック プログラミングに基づいた 蛋白質のアライメントシステム

広沢　誠、星田　昌紀、石川　幹人、
戸谷　智之

January, 1992

ICOT

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

(03)3456-3191～5
Telex ICOT J32964

Institute for New Generation Computer Technology

ＭＡＳＣＯＴ：３次元ダイナミックプログラミングに基づいた蛋白質のアライメントシステム[1]

広沢 誠，星田 昌紀，石川 幹人，戸谷智之[2]
(財) 新世代コンピュータ技術開発機構 (ICOT)[3]

要約： 蛋白質の配列解析手法であるマルチプルアライメントは、未知の蛋白質の機能や構造を推定するのに有用な技術である。我々は、品質の良いマルチプルアライメントを作ることを目的とするシステムMASCOTを開発した。このシステムは、３次元ダイナミックプログラミングによるアラインメントを基本とした初期アライメントに、シミュレーテッドアニーリングによる洗練化を行う特徴を持つ。またアライメントの組合せ順序を規定するクラスター分析も工夫している。

1 はじめに

蛋白質配列の類似性を解析する典型的な手法であるマルチプルアライメントは、蛋白質の機能／構造予測や、生物種の進化系統樹の作成などに利用される重要な技術である。蛋白質配列は２０種類のアミノ酸を表す英文字の並びで記述されるが、マルチプルアライメントとは、複数の配列のうちで、性質の似ているアミノ酸がなるべく縦に揃うように、各配列を並べ合わせるものである。従来、このマルチプルアライメントは、おもに生物学者の経験に頼って行われていたが、近年、蛋白質配列が次から次へと決定され、計算機によるマルチプルアライメント手法の確立が急務とされている。

マルチプルアライメントの課題は、文字相互の類似性尺度が与えられれば、ダイナミックプログラミング（以下ＤＰと呼ぶ）の手法で、理論的には最適解が導けることが古くから知られている[1]。しかし、その計算量は膨大で、配列がｎ本あるｎ次元のＤＰは、配列の長さのｎ乗の計算量を必要とする（厳密には、さらにその何割かの計算が必要となる[2]）。そのため、200文字程の配列を十数本並べる標準的なアライメントの課題を計算機で解くには、通常、２次元のＤＰにより配列２本のアライメント結果を得て、これを複数組合せてマルチプルアライメントとしている[3]。当然ながら、こうした方法は計算量が少く、速やかに行えるという点では優れているが、類似性の低い蛋白質間のアライメントの品質は十分でないという問題がある。

これに対して我々は、多少時間がかかるが品質の良い、３次元ＤＰを基本とするマルチプルアライメントシステムの開発[4]、改良[5]を続けてきた。本システムMASCOT (Multiple Alignment System developed by ICOT) は、類似性の低い蛋白質配列を扱う時にも、品質の良いアライメントを導き出すことを目的としている。MASCOTの特徴は、（１）３次元ＤＰの結果を核とした初期アライメントの形成、（２）シミュレーテッドアニーリングを用いたアライメントの洗練化、（３）配列の３つ組を単位にした独自のクラスター分析である。

MASCOTの各モジュールは並列論理型言語ＫＬ１で記述されており、推論マシンPSI上で動作している。またMASCOTは、３次元ＤＰを使用することなどにより、従来法よ

[1] Multiple Alignment System for Protein Sequences based on 3-dimensional Dynamic Programming

[2] Makoto Hirosawa, Masaki Hoshida, Masato Ishikawa, Tomoyuki Toya

[3] Institute for New Generation Computer Technology (ICOT)

りも計算量が増大しているが、各モジュールの並列化可能な部分を並列推論マシン Multi-PSI で計算させることで、全体の計算時間の低減を図っている [6]。

本論文では、まず、第2章で、MASCOT の概要を紹介する。次に、第3、4、5章で、上に述べた MASCOT の特徴を、順に説明する。最後に、第6章で MASCOT の適用例について具体的に示す。

2 アライメントシステム MASCOT

我々が開発した MASCOT は、（1）“クラスター分割”、（2）“クラスター内アライメント”、（3）“クラスター間アライメント”、（4）“アライメントの評価”の4モジュールにより構成される（図1上）。この全体構成について概説しよう。

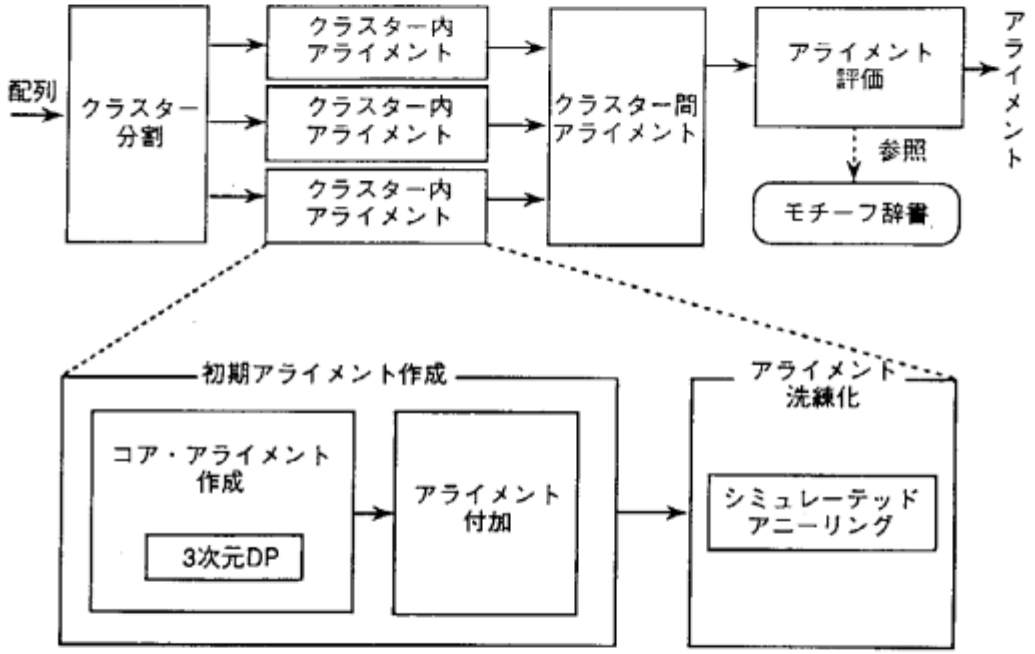

図 1: アライメントシステム MASCOT

マルチプルアライメントの問題は、配列のうちの類似性の高い部分を縦に揃えて並べ合わせる（アライメントする）のが課題であるが、通常、与えられた配列には、類似性の高いもの同士から、低いもの同士までいろいろとある。類似性の低いもの同士をむりやりアライメントすると、ノイズを拾ってしまい、アライメントの品質がよくないことが多い。そこでアライメントに際し、事前に類似性の高いもの同士をグループに分け、なるべく類似性の高いもの同士からアライメントすることが良く行われる [3]。

MASCOT においても、そうした処理を行っている。まず、“クラスター分割”において、配列の類似性に基づき、配列を複数のクラスターに分割する。次に、各クラスターごとのアライメントを“クラスター内アライメント”により求め、その結果の並べ合わせを、“クラスター間アライメント”において行い、全体的なアライメントを作成する。クラスター内におけるアライメントの組合せ順序と、クラスター間の組合せ順序はともに、“クラスター分割”にて行われるクラスター分析によって整合的に決定される（第5章参照）。

“クラスター内アライメント”（図1下）では、まず、初期のマルチプルアライメントを少ない配列のアライメントを組合せることで形成し、それを全配列にわたる類似性を評価しながら洗練化するという、2段階でマルチプルアライメントが行われる。

始めから全配列にわたる類似性を評価しながらアライメントするのは、計算量の問題で難しいため、従来も、少ない配列のアライメントを組合せたあとで、それを修正する方法がとられていた[3]。しかしながら、初期のアライメントの品質が十分でないと、その修正も十分に行えない。そこで、MASCOTでは、品質の良い初期アライメントを目指して、3次元DPを基本とした組合せ方法（第3章参照）を採用している。また、全配列にわたる類似性の評価にも、有力な確率的手法であるシミュレーテッドアニーリング（第4章参照）を導入し、洗練化能力の格段の向上を図っている。

“クラスター間アライメント”は、各クラスターのアライメントを結合して、全体のアライメントを作成する役割をしている。その際には、クラスター分析で決定された順序で、配列群間の2次元DPを繰返し行う方法をとっている。

“アライメント評価”は、“クラスター間アライメント”が済んだ時点で、複数のアライメントが生成されているときに、品質の良いアライメントを選択し、ユーザに提示するモジュールである。複数のアライメントが発生する場合とは、それまでの段階で評価が同得点であるアライメントが2つ以上得られた場合と、戦略的に評価が上位のアライメントを複数候補に残しておく場合がある。後者の戦略は、MASCOTでは生物学的にある程度定評のある評価尺度（Dayhoff Score [7]）を使用してはいるが、必ずしもそれが絶対ではないのが理由となっている。つまり、マルチプルアライメントの評価には、ほかの生物学的なノウハウも必要なのである。“アライメント評価”では、全配列にわたり同一の文字が縦に並んでいるカラムがどの程度あるか、それがモチーフであるかなどの、生物学的なノウハウも含めて総合的に判断する。モチーフとは、蛋白質の構造や機能の点で重要な意味をもつ部分にみられる、特徴的な配列パターンである。本システムは、モチーフ辞書（実験段階）を装備しており、既知のモチーフの情報がマルチプルアライメントに反映できるようになっている。

3　3次元DPによる初期アライメント作成

MASCOTでは、3次元DPの結果を核にしたコア・アライメントに、残りの配列を付加するかたちで初期アライメントが行われる。いちばん類似性の高い配列群がコアを形成するのが良い。どの配列がコアにあたるか、そのあと、どの順序で残りの配列を付加するかは、クラスター分析で決められる（第5章参照）。

少ない配列のアライメントを次々と組合せる方法をとると、コアにあたるアライメントがいかに正確かで、全体のマルチプルアライメントの性質が左右されがちである。そのため、可能な限り多くの配列を同時にDPして、最適のアライメントを求め、それをコア・アライメントにするのが望ましい。しかし、計算量の問題で、現実的に実現可能なDPは3次元DPまでである。そこでMASCOTでは、2つの3次元DPの結果を組合せて得られる配列4本のアライメントを、コア・アライメントとしている。

2つの3次元DPの結果を組合せる方法を、3次元DPマージ法[8]と呼んでいる。その手順について、小さな問題を例にとって図2で説明する。図では、3次元DPによって得られた3本のアライメント2組(Step0)を組合せて、4本のアライメント(Step3)を作る過程を示している。各アライメントにおいて、‘:’の右側の英文字が、それぞれのアミノ酸を表し、左側の見出しが配列の名前を表している。Step0を見ると、下の3本組のアライメント{HTLV-2, BLV, HIV}は、ところどころに‘-’で示されるギャップが挿入されて、類

似したアミノ酸を表す文字が縦に揃っている。一方、上の３本組のアライメント{HTLV-1, HTLV-2, BLV}は、ギャップが入らずに、そのまま類似文字が縦に揃っている。

```
Step0   HTLV-1  : LLQAIAHLGKPSYINT
        HTLV-2  : VLQAISLLGKPLHINT
        BLV     : LLEAIVHLGRPKKLNT
                    ^  ^^   ^^ ^   ^^
        HTLV-2  : VLQAI-SLLGK-PLH-INT
        BLV     : LLEAIVHL-GR-PKK-LNT
        HIV     : FL--L-KLAGRWPVKTIHT
                    ^    ^  ^ ^   ^

Step1   HTLV-1  : LLQAIAHL.GK.PSY.INT
        HTLV-2  : VLQAISLL.GK.PLH.INT
        BLV     : LLEAIVHL.GR.PKK.LNT
                  11111000011111111111
        HTLV-2  : VLQAI-SLLGK-PLH-INT
        BLV     : LLEAIVHL-GR-PKK-LNT
        HIV     : FL--L-KLAGRWPVKTIHT

Step2   HTLV-1  : LLQAIAHL.GK.PSY.INT
        HTLV-2  : VLQAISLL.GK.PLH.INT
        BLV     : LLEAIVHL.GR.PKK.LNT
        HIV     : FL--L-KLAGRWPVKTIHT
                    ^    ^  ^ ^   ^

Step3   HTLV-1  : LLQAIAHL-GK-PSY-INT
(Result)HTLV-2  : VLQAISLL-GK-PLH-INT
        BLV     : LLEAIVHL-GR-PKK-LNT
        HIV     : FL--L-KLAGRWPVKTIHT
                    ^    ^  ^ ^   ^
```

図２: コア・アライメントの作成

まず、最初の過程 Step1 では、{HTLV-1, HTLV-2, BLV}に含まれる HTLV-2 と BLV のアライメントと、{HTLV-2, BLV, HIV}に含まれる HTLV-2 と BLV のアライメントの一致部分を見つけ出す。図では、一致領域に'1'、矛盾領域に'0'を割り当てている。ここで、'.'はカラムを合わせるために挿入したギャップである。'0'が並んだ矛盾領域では、２組に共通な配列である HTLV-2 と BLV とのアライメントが、上の組で[SLL,VHL]である一方、下の組で[-SLL,VHL-]となって異なり、矛盾を示している。

次に、Step2 では、一致領域と矛盾領域に分けて、４本のアライメント{HTLV-1, HTLV-2, BLV, HIV}を作っていく。一致領域の４本のアライメントを作るのは簡単である。左側の一致領域では、上の組の[LLQAI,VLQAI,LLEAI]と下の組の[VLQAI,LLEAI,FL--L]を合わせて、[LLQAI,VLQAI,LLEAI,FL--L]という４本のアライメントを作る。右側の一致領域でも同様である。矛盾領域の上の組の[AHL,SLL,VHL]と下の組の[-SLL,VHL-,-KLA]では、以下の４通りの配列群間の２次元ＤＰを行い、得られた部分アライメントのうち、評価得点が最も良いものを矛盾領域の４本のアライメントとする。

1. 上の組の[AHL,SLL,VHL]と，下の組の[-KLA]のアライメント.
2. 上の組の[AHL,SLL]と，下の組の[VHL-,-KLA]のアライメント.
3. 上の組の[AHL,VHL]と，下の組の[-SLL,-KLA]のアライメント.
4. 上の組の[AHL]と，下の組の[-SLL,VHL-,-KLA]のアライメント.

この例では、１と２のアライメント結果は同一で、[AHL-,SLL-,VHL-,-KLA]となり、それが最良の評価を与える。そこで、それをその領域のアライメントとして、図の Step2 のような結果を得る。最後に、'.'を'-'に変えて最終的なアライメントとなる (Step3)。

さて、コア・アライメントが形成されたならば、残りの配列を１本ずつ、クラスター分析で決められた類似性の高い順に、次々と付加していく。このとき、すでに形成されたアライメントと、それに付加する配列とは、２次元ＤＰを用いてアライメントしている。すべての配列が付加されたら、初期のマルチプルアライメントの完成である。

まれに、４本のコア・アライメントをつくるときも、３次元ＤＰの結果を２つマージする方法でなく、３次元ＤＰのひとつ目の結果に、４番目の配列を２次元ＤＰで付加したほうが良いことがある。コア・アライメントの作成時には、つねにこの可能性をチェックし、評価得点の良い方を選んでいる。

4　シミュレーテッドアニーリングによる洗練化

MASCOTでは、3次元ＤＰを用いた良質な初期アライメントが得られる。この品質をさらに向上させるために、シミュレーテッドアニーリング（以下ＳＡと呼ぶ）という方法を用いて、初期アライメントの改善を図っている。

ＳＡとは、ある評価値が最良になるように試行錯誤的な探索を行う確率的手法の１つである [9]。ＳＡは、解空間の局所的最良値に探索者が捕われないように、温度という概念を導入している。温度が高い時には、評価値を下げる方向にも探索を行ない、温度が低い時には、ほとんど評価値を上げる方向のみに探索を行う。そして、温度を高温から徐々に下げていくこと（温度スケジューリング）により、高温において、大局的最良値の付近に移動でき、低温において、局所的最良値を探索することができる。ＳＡは組合せ最適化問題を解く、一般的で強力な手法である。

我々はマルチプルアライメントの課題を、組合せ最適化問題として定式化し、ＳＡでこの課題を解く方法を開発した [10]。ＳＡを用いれば、ランダムな初期状態からでも、満足のいくアライメント状態までに至らせることも、理論的には可能である。だが、通常規模のマルチプルアライメントの問題では、非現実的な時間を必要とするので、ある程度アライメントの済んでいる状態から始めるのが効果的である [11]。

ある程度アライメントの済んでいる状態からＳＡを始めるには、温度をあまり高温にせず、中低温から徐々に下げていかねばならない。そのため、解く問題の規模や性質に応じた、温度スケジューリングの工夫が必要である。MASCOTでは、各プロセッサに異なる温度を割り当てる温度並列ＳＡ [12] を採用して、この問題に対処した。温度並列ＳＡでは、温度ごとに割り当てられたプロセッサ間で、確率的に解交換を行うことで、いつでもその時点で最良の解が最低温度のプロセッサに得られる。それにより、事前に温度スケジューリングを考慮することなしに、与えられた時間において準最適な結果が手軽に得られる。

5　クラスター分析の方法

MASCOTは、3次元ＤＰを基本としてマルチプルアライメントを作成する。このため、どの3本の配列が互いに似ており、3次元ＤＰをした場合に品質の良いアライメントをもたらすかを見積もる必要がある。これまで配列を分類するときは、もっぱら、配列2本間の類似性を調べていたのに対し、MASCOTでは、配列3本間の類似性の指標をもとにして、クラスター分析している [13]。この配列3つ組の類似性指標によって、クラスターの核となる3つ組の選定から、アライメントの付加順序、クラスター間の組合せ順序までが、一貫して決定される。

クラスター分析にあたり、配列ｉ，配列ｊ，配列ｋの間の類似性に相当する指標 R(i,j,k) を、次のように算出する。まず、配列ｐと配列ｑの類似性 S(p,q) を、すべてのペアについて求める。現在、S(p,q) には、配列ｐと配列ｑとを2次元ＤＰしたときの評価得点を採用している。そして、S(i,j)、S(j,k)、S(k,i) を足し合わせたものを、類似性指標 R(i,j,k) とみなしている（ある種の重み付き足し算が妥当かも知れないが検討中である）。似た配列同士を一緒にアライメントしたほうが、一般に正確なアライメントが得られるので、この類似性指標は、その3つ組に対して3次元ＤＰを行ったときの、アライメントの信頼性と考えても良いといえる。

すべての3つ組について R(i,j,k) が求まったならば、それらを大きい順にソートした列を作る。この列ができれば、クラスター分割の開始である。この手順は基本的に、列の先頭か

ら抜き出した３つ組を調べ、その配列３本が、既存のクラスターに登録されているか否かをチェックする操作の繰返しである。以下の４つの場合に応じて、それぞれの処理を、列に３つ組がなくなるまで繰返すことにより、配列を複数のクラスターに分割できる。

(1) **配列３本のいずれもが既存のどのクラスターにも登録されてない場合**：新たにクラスターを作成し、その３つ組をそのクラスターのコアとして登録する。これを「コア登録」という。

(2) **配列３本のうち２本が既存のあるひとつのクラスターに登録されており、残りの１本はどのクラスターにも登録されてない場合**： その未登録の配列を，そのクラスターに付け加える。それを「付加登録」という。特別に、そのクラスターに配列が３本しかないときは、「付加登録」をせずに、新しい１本も登録済みの３本とともに「コア登録」する。

(3) **配列３本のうち２本が既存のあるひとつのクラスターに登録されており、残りの１本は別のクラスターに登録されている場合**： これらの２つのクラスターを結合することを登録する。これを「結合登録」という。

(4) **その他**： 何も行わずに、その３つ組を捨てる。

当然ながら、始めはクラスターがひとつもないので、始めの３つ組は (1) にあてはまり、「コア登録」される。その後の３つ組は (4) でない限り、おおよそ (1) あるいは (2) にあてはまり、いくつかのクラスターが「コア登録」や「付加登録」によって成長していく。この「コア登録」と「付加登録」の情報が、“クラスター内アライメント”へ渡され、「コア登録」の情報はコア・アライメントの作成に、「付加登録」の情報はアライメントの付加に利用される。

クラスターに未登録の配列がなくなったころに、(3) にあてはまる３つ組が現れ始める。この (3) によって得られる、クラスターの「結合登録」の情報が、“クラスター間アライメント”へ渡され、クラスターの結合順序を規定するのである。例外的に、未登録の配列がなくなる前に、(3) が現れる場合は、その未登録の配列を“クラスター間アライメント”において、付加アライメントするなどの処置をする。

6 統合システムの適用結果例

レトロウイルスの endonuclease という蛋白質の一部分の配列１６本を、MASCOT を用いてアライメントを行った。レトロウイルスは、成人白血病ウイルスや、エイズウイルス等を含むウイルス群であり最近非常に注目を浴びている。また、endonuclease という蛋白質は、これらのウイルスが増殖する際に必須の蛋白質である。クラスター分析の結果を図３、そしてこれに基づいたアライメントの結果の一部を図４に示す。図３の小さな丸はそれぞれ配列を表し、四角の枠はコアになる配列群を表す。クラスター内の丸枠は付加アライメントを、クラスター外の丸枠はクラスター間アライメントを行うことを示す。

まず、配列群は、“クラスター分割”により３つのクラスターに分割された。そして、各クラスターの配列群は、“クラスター内アライメント”により、それぞれアライメントされた。その後、“クラスター間アライメント”により、クラスター２のアライメントとクラスター３のアライメントが結合され、その結合結果に、クラスター１のアライメントが結合されて、最終的なアライメントが形成された。なお、クラスター１は、レトロウイルスの一群

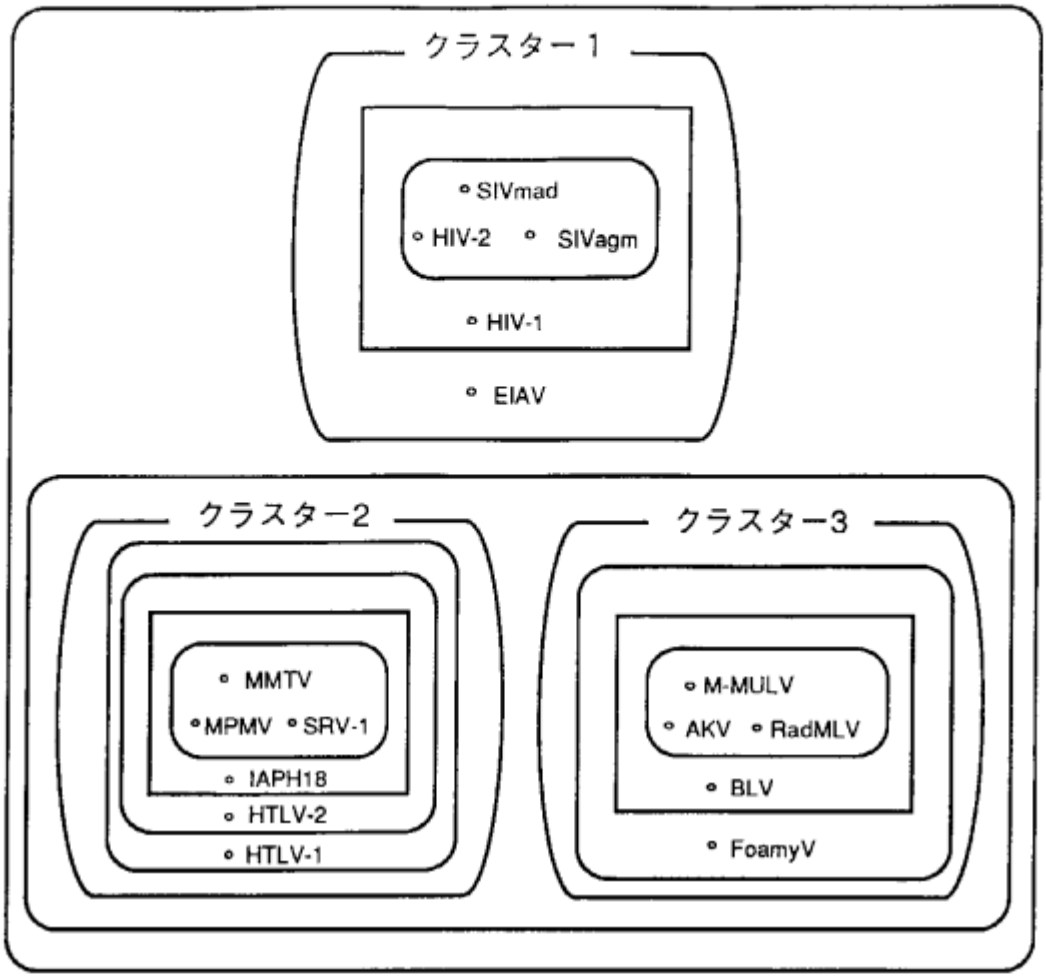

図 3: クラスター分析の結果

であるレンチウイルス（エイズウイルスが所属している）に属する種の配列より構成されている。そして、その他のほとんどの配列はレトロウイルスの一群であるオンコウイルス（成人白血病ウイルスが所属している）に属する種の配列である。これは、クラスター分析がうまくいったことを意味している。

このアライメント結果を見ると、各クラスターのコア・アライメントは、明確な配列パターンを持っていることが分かる。また、各クラスターの全体のアライメントにも、不明瞭であるが認識できる配列パターンが存在する。これは、品質の良いコア・アライメントを作ってから残りの配列を付加していくという戦略が有効であることを意味している。

また、１６本すべての配列のアライメントは、H..H.....C..C という Zinc Finger モチーフを捕えている。Zinc Finger は蛋白質がＤＮＡに結合する機能を持つ。このように、生物学的に重要な蛋白質配列の部位を特定できるアライメントを作成できることは、このシステムの適用結果を、蛋白質の機能／構造予測に利用できる可能性を示唆している。

```
<Cluster1>
HIV-2 : AHKGIGG--NQEVDH-L-VSQGI-RQVLF-L-----EK-IE-PAQEEHEKYHS---NVKELSHK-----
SIVmad: AHKGIGG--NQEIDH-L-VSQGI-RQVLF-L-----EK-IE-PAQEEHSKYHS---NIKELVFK-----
SIVagm: AHKGIGG--NEEIDK-L-VSKGI-RRVLF-L-----EK-IE-EAQEKHERYHN---NWKNLADT-----
HIV-1 : AHKGIGG--NEQVDK-L-VSAGI-RKILF-L-----DG-ID-KAQDEHEKYHS---NWRAMASD-----
EIAV  : ENK-I----SQRGDKGF-GSTGV-----FWV-----EN-IQ-EAQDEHENWHT---SPKILARN-----
        ahKgiGG  N   D  l VS Gi r  lF L        I   AQ  H  yH    n

<Cluster2>
SRV-1 : AHSGLPG-PIAHGNQ-K-A-D-LATK--T-VASNINTN-LE-SAQNAHTLHHL---NAQTLKLM-----
MPMV  : AHSGLPG-PIAQGNQ-R-A-D-LATK--I-VASNINTN-LE-SAQNAHTLHHL---NAQTLRLM-----
MMTV  : GHTGLPG-PLAQGNA-Y-A-DSL-TR--I-L-----TA-LE-SAQESHALHHQ---NAAALRFQ-----
IAPH18: AHSGLPG-PMSLGND-L-A-D-KATK--L-VATALSTH-AQ-AAKEFHKRFHV---TAETLRRR-----
HTLV-2: SHTNLPD-PISTFNE-Y-T-DSL-----I-LAP-L-VP-L--TPQGLHGLTHC---NQRAL-VS-----
HTLV-1: SHTNLPD-PISRLNA-L-T-DAL-----L-I-----TPVLQLSPAELHSFTHC---GQTALTLQ-----
         H gLPg P   gN    a D               t      a   H   H     a  L

<Cluster3>
AKV   : GYWVFQGKPVMP-DQ-F-VFE-------L-L-----DS-L-------HRLTHLGYQKMKAL-LDRGESP
RadMLV: GYWVFQGKPVMP-DQ-F-VFE-------L-L-----DS-L-------HRLTHLGYQKMKAL-LDRGESP
M-MULV: KYWVYQGKPVMP-DQ-F-TFE-------L-L-----DF-L-------HQLTHLSFSKMKAL-LERSHSP
BLV   : SHSS-ASHPIASLNN-Y-VDQ-------L-L------P-LE-TPEQWHKLTH--CNS-RAL-SRWPNPR
FoamyV: VSRP-EGVKIIP-PQ-S-DRQKI-----V-L-----QA---------HNLAHTG--REATL-LKIAN--
               p                     l L        l       H LtH       aL


<Cluster1>
HIV-2 : -F---GIPNLVARQ---IVNSCAQCQQ---------KG-EA-I-HGQVNAELGT-WQ---MDCTH-LEG
SIVmad: -F---GLPRLVAKQ---IVDTCDKCHQ---------KG-EA-I-HGQVNSDLGT-WQ---MDCTH-LEG
SIVagm: -Y---GLPQIVAKE---IVAMCPKCQI---------KG-EP-V-HGQVDASPGT-WQ---MDCTH-LEK
HIV-1 : -F---NLPPVVAKE---IVASCDKCQL---------KG-EA-M-HGQVDCSPGI-WQ---LDCTH-LEG
EIAV  : -Y---KIPLTVAKQ---ITQECPHCTK---------QG-SG-P-AGCVMRSPNH-WQ---ADCTH-LDN
              lP  VA     Iv  C  C           kG e    hGqV    g  WQ    DCTH Le

<Cluster2>
SRV-1 : -F---NIPREQARQ---IVRQCPICATYLPVPH-L--G-VN-P-RGLL--PNMI-WQ---MDVTH-YSE
MPMV  : -F---NIPREQARQ---IVKQCPICVTYLPVPH-L--G-VN-P-RGLF--PNMI-WQ---MDVTH-YSE
MMTV  : -F---HITREQARE---IVKLCPNCPDWGHAPQ-L--G-VN-P-RGLK--PRVL-WQ---MDVTH-VSE
IAPH18: -F---ALSRKEARE---IVTQCQNCCEFLPTPH-M--G-IN-P-RGIR--PLQM-WQ---MDVTH-IPS
HTLV-2: -F---GATPREAKS---LVQTCHTCQTINSQHH-MPRG-YI-R-RGLL--PNHI-WQ---GDVTH-Y-K
HTLV-1: -----GATTTEASN---ILRSCHACRGGNPQHQ-MPRG-HI-R-RGLL--PNHI-WQ---GDITH-FK-
         F      r  Ar    iV  C  C      p     G  n p RG    P    WG   mDvTH

<Cluster3>
AKV   : YY---MLNRD--KTLQYVADSCTVCAQVNASKAKIGAG-VR-V-RGHR--PGSH-WE---IDFTE-VKP
RadMLV: YY---MLNRD--KTLQYVADSCTVCAQVNASKAKIGAG-VR-V-RGHR--PGTH-WE---IDFTE-VKP
M-MULV: YY---MLNRD--RTLKNITETCKACAQVNASKSAVKQG-TR-V-RGHR--PGTH-WE---IDFTE-IKP
BLV   : -----ISAWD-PRS---PATLCETCQKLNPT----GGGKMRTIQRG-W--APNHIWQ---ADITH-YK-
FoamyV: LYWWPNMRKDVVKQL----GRCQQCLITNASNK--ASG-PI-L-RPDR--PQKP-FDKFFIDYIGPLPP
                D           C  C   N        G      Rg       h w     D t   k
```

レトロウイルスの endonuclease の配列の一部をアライメントした結果である。各配列は配列データベースである *PIR* より検索したものである。なお、各クラスターの配列の上の４本がコアを構成している配列である。各クラスター内のアライメントの下にコンセンサス（全配列に共通に存在するアミノ酸）を示してある。大文字はクラスター内のコンセンサスであり、小文字はコア内に限ったコンセンサスである。

図４：アライメント結果例

7　おわりに

3次元ＤＰを基本とした蛋白質配列のマルチプルアライメントシステムであるMASCOTについて紹介した。そして、このシステムの特徴である、（1）３次元ＤＰの結果を核とした初期アライメントの形成、（2）シミュレーテッドアニーリングを用いたアライメントの洗練化、（3）配列の３つ組を単位にした独自のクラスター分析について順に述べた。また、このシステムの適用例を示し、その結果、生物学的に重要な配列部位を特定できたことを述べた。

システムの適用例（図4）を見ると、コア・アライメントに明確なコンセンサスが存在するうえ、各クラスターのアライメントは、かなり明確なパターンを持っていることがわかる。このようなパターン情報を既存のモチーフ情報と比較しながら、付加アライメントやクラスター間アライメントを行うと、より品質の良い結果を得られそうである。現在、こうしたことを実現するための、知識処理的な手法も検討している。

文部省科学研究費補助金重点領域研究「ゲノム情報」の班員の方々からは、多くの助言をいただきました。ここに感謝の意を表します。また、本研究の機会を与えていただいた、ICOTの古川康一所次長、および内田俊一研究部長、新田克己室長にお礼申し上げます。

参考文献

[1] Needleman and Wunsch "A General Method Applicable to the Search for Similarities in the Amino Acid Sequences of Two Proteins", in J. Molecular Biology 48, 1970. pp443-453.

[2] 後藤修：核酸・蛋白質一次構造の計算機による解析，日本物理学会誌 Vol.38 No.6, 1983, pp.477-480.

[3] Geoffrey J.Barton "Protein Multiple Alignment and Flexible Pattern Matching" in *Methods in Enzymology 183*, Academic Press, 1990, pp.626-645.

[4] 石川、星田、広沢、戸谷、鬼塚、新田、金久：“並列推論マシンを用いたタンパク質の配列解析”，情報処理学会 情報学基礎研究会報告 23-2, 1991.

[5] 広沢、星田、石川、戸谷：“３次元ダイナミックプログラミングを活用した蛋白質のアライメントシステム”、第２回公開ワークショップ「ヒトゲノム計画と情報解析技術」論文集、1991.

[6] 戸谷、星田、石川、新田：並列３次元ダイナミックプログラミング法によるタンパクの配列解析，情報処理学会第５回プログラミング研究会報告、1991.

[7] Dayhoff, Hunt and Hurst-Calderone "Composition of Proteins" in *Atlas of Protein Sequence and Structure 5:3*, Nat. Biomed. Res. Found., Washington, D. C., 1978, pp.363-373.

[8] 石川、星田、広沢、戸谷、新田：“３次元ダイナミックプログラミングを用いたタンパク質の配列解析”情報処理学会第43回全国大会論文集，1991.

[9] S.Kirkpatrick, C.D.Gelatt and M.P.Vecci, "Optimization by Simulated Anealing", Science, Vol.220, No.4598, 1983.

[10] 戸谷、星田、石川、新田、金久：“並列シミュレーテッドアニーリングを用いたマルチプルアライメント”，情報処理学会第43回全国大会論文集，1991.

[11] 石川、戸谷、星田、新田、荻原、金久：“並列シミュレーテッドアニーリングを用いたマルチプルアライメント”、第２回公開ワークショップ「ヒトゲノム計画と情報解析技術」論文集，1991.

[12] 木村、瀧：時間一様な並列アニーリングアルゴリズム、電子情報通信学会 NC90-1, 1990.

[13] 広沢、星田、石川：“蛋白質配列間距離解析を用いた 蛋白質の相同性解析システム”情報処理学会第43回全国大会論文集，1991.