情 報 学 基 礎 23－2
（1991．9．10）

並列推論マシンを用いたタンパク質の配列解析

石川幹人，星田昌紀，広沢誠，戸谷智之，鬼塚健太郎，新田克己
（財）新世代コンピュータ技術開発機構

金久 實
京都大学化学研究所

新世代コンピュータ技術開発機構（ICOT）では、並列推論マシンを遺伝子情報処理の分野へ応用することを目的に、研究開発を行っている。今回、遺伝子情報処理の分野のなかでも、タンパク質の配列解析を行う３つの実験システムを構築したので、ここに報告する。３つの実験システムは、それぞれ３次元ダイナミックプログラミング、シミュレーテッドアニーリング、トーナメント法を要素技術として使用しており、どれも並列性を生かした実装を行っている。またそれらの利点を活用した、統合システムを構築したところ、そのシステムは生物学的な実用性の面から、多大な意義をもつ解析結果を提供することが判明した。

PROTEIN SEQUENCE ANALYSIS BY PARALLEL INFERENCE MACHINE

Masato Ishikawa, Masaki Hoshida, Makoto Hirosawa,
Tomoyuki Toya, Kentaro Onizuka, Katsumi Nitta
ICOT

Minoru Kanehisa
Kyoto University

We have studied application systems for parallel inference machine developed in ICOT. The systems analyze protein sequences and generate multiple sequence alignments. That analysis is very important in the field of genetic information processing. This paper reports four systems. The first three systems of them use different techniques in our own parallel ways: three-dimensional dynamic programming, simulated annealing and tournament method. The last system integrates the merits of the three techniques and makes it possible to generate as good alignments as biologists do.

1 はじめに

本論文は、ＩＣＯＴで開発した並列推論マシン上で動く、タンパク質の配列解析（マルチプルアライメント）システムについての報告である。本論文の構成は、まず本章で、問題領域の背景と課題を述べたあと、続く２、３、４章で、３種類のマルチプルアライメントシステムを紹介する。そして５章で、それらの３システムの比較をし、６章で、各システムの長所を生かした統合アライメントシステムを提案する。最後に７章で、まとめと今後の課題を述べる。

1.1 遺伝子情報処理

近年、DNA の核酸配列やタンパク質のアミノ酸配列の自動分析法が開発され、それらのデータを蓄えたデータベースが急速に膨らんでいる。たとえば現在 GenBank[Bilofsky 88] に蓄えられているデータ量は、およそ数千万塩基対であるが、アメリカを中心に進んでいるヒトゲノム計画は、三十億塩基対もある人間の全遺伝子を読みとろうとしている。現在の実験分析技術をもってすれば、今世紀中にもその読みとりが達成される可能性がある。しかし現状では、配列データを次々と蓄えてはいるが、それらの十分な利用がなされているわけではない。それは配列を読み取る実験技術の進歩に比べ、配列情報がいかなる意味を持つかを探る、遺伝子情報処理技術が未熟であるためといえる。そのため、今後の生物学の画期的な進歩には、遺伝子情報処理技術の確立が急務とされている。

情報処理学会においても、昨年の学会誌で、「遺伝子情報の解析とタンパク質の構造推定」なる特集が組まれ、注目されつつある遺伝子情報処理分野の最新の研究動向が紹介された。そのなかで、本論文に比較的近い内容の解説に [五條堀 90] があるので、あわせて参照されたい。また人工知能学会誌にも最近、解説が掲載されている [金久 91]。

遺伝子情報処理のうちで最も基本的な技術は、配列間の類似性を調べる解析処理といえよう。ここで問題となるのは計算量である。というのは、あるひとつの配列と、ある観点から類似とみなすべき配列をデータベースから見つけるには、数万から数百万の配列との類似性を評価しなければならない。さらに、数個の配列にわたって共通の特徴を抽出したいとなれば、組合せ的に計算量が増大する。そこで試みられるのが、並列処理による計算時間の低減である。すでにいくつかの並列処理を応用した配列解析の研究が報告されている [Iyengar 88]。

また最近では、並列処理による計算時間の低減だけでなく、生物学的な知見をヒューリスティクスとして盛り込む知識処理的な方法も注目されており、知識を盛り込み易い論理型言語による並列配列解析も始められている [Butler 90]。このように並列処理、とくに論理型言語を用いた並列知識処理には、今後の遺伝子情報処理の飛躍的な発展につながる重要技術であるといった期待が寄せられている。本論文で報告する並列推論マシン上のシステムもこの路線を目指したものである。

1.2 タンパク質と配列解析

よく知られているように、遺伝情報は細胞の内部にある DNA に格納されている。タンパク質は、この DNA の情報から翻訳生成されるアミノ酸配列が、空間的に折れ畳まって特異的な形状になったものである。タンパク質は生物の体を形成し、生命の代謝反応を司る重要な物質である。

タンパク質の構成要素であるアミノ酸には２０種類あり、それぞれ異なるアルファベットが割り当てられている。アミノ酸には、大きさ、水との親和性、酸性／塩基性、極性などの性質があり、どんな性質のアミノ酸がどんな順番に連なっているかで、タンパク質の構造や機能が決まってくる。タンパク質には大小さまざまなものがあり、短いのは数十個、長いのは数百個のアミノ酸が連なっている。

タンパク質の構造や機能は、実験を積み重ねて初めてわかるものとされている。事実、タンパク質の正確な構造は、まだ２百種類程度しか知られていない。一方、タンパク質のアミノ酸配列を調べる技術は、すでに確立されており、それを自動分析する機械も販売されている。現在では約１万種類のタンパク質について、その配列が決定されている。この数字は近年、ますます増大している。

このように多くのタンパク質の配列データが集まると、新たな可能性が見えてきた。類縁のタンパク質は類似したアミノ酸をもつ傾向が明らかになったのである。すると、未知のタンパク質であっても、それと類似の配列をもつタンパク質の構造や機能が既知であれば、それから未知の構造や機能を推測することが可能になる。そこで必要になるのは、配列間の類似性を解析する情報処理技術である。

1.3 マルチプルアライメント

もっとも基本的な配列解析のひとつは、複数の配列の類似する部分を縦に揃えて並べ合わせる操作で、マルチプルアライメント (Multiple Alignment) と呼ばれる。たとえば以下のようなタンパク質のアミノ酸配列が４本あったとす

る。

```
CCHU      GDVEKGKIFIMKCSQCHTVEKGGKHKTGPNLHGLFG
CCFS      ASFAEAPAGTTGAKIFKTKCAQCHTVKGHKQGNGLFG
CCZP      PYAPGDEKKGASLFKTAQCHTVEKGGANKVGPNLHGVFG
CCCRCF    PPKARAPLPPGDAARGEKLRAAQCHTANQGGANGVGYGLVG
```

ここで、左側の見出しが配列の名前で、右側の文字がひとつひとつのアミノ酸を表現する。アミノ酸は２０種類あり、２０種のアルファベットで識別される。最上段左からGDVEKは、それぞれグリシン、アスパラギン酸、バリン、グルタミン酸、リシンを意味している。これをアライメントすると、次のようになる。

```
CCHU      ----------GDVEKG-KIFIMKCSQCHTVEKGGKHKTGPNLHGLFG
CCFS      --ASFAEAPAG--TTGAKIFKTKCAQCHTV-KG--HKQG---NGLFG
CCZP      ------PYAPGDEKKGASLFKT--AQCHTVEKGGANKVGPNLHGVFG
CCCRCF    PPKARAPLPPGDAARGEKL---RAAQCHTANQGGANGVG---YGLVG
```

配列のところどころにギャップ“–”を入れることで、QCHTなどの共通文字が同じ列に並んでいるのがわかる。QCHTのように複数の文字が、複数の配列で共通になっている文字の組を配列モチーフと呼び、タンパク質のうちの重要な部分を指し示していると判断される。この背景には、タンパク質の配列のうち重要である部分には遺伝的変異が起きにくいという進化論的考え方[木村資生 86]がある。

一般のマルチプルアライメントでは、ひとつの列に同じ文字が揃うことは少なく、異なる文字でもそれらが表すアミノ酸の性質が似ていれば同じ列に置くことを許容して処理を行う。しかしアミノ酸の性質には親水性、疎水性、極性、酸性、塩基性、大きさなど多数あり、その類似性評価も多数の方法がある。現在最も広く使われている類似性評価尺度は、Dayhoffマトリックス[Dayhoff 78]である。この尺度は、当時知られていたアライメントをもれなく調べ、同じ列に該当アミノ酸対が並んでいることが偶然に対して何如に少ないかを数値化したものである。数値は確率の対数値になっているため、それらの足し算は複合事象の共起確率を算出したことに相当する。本論文の、３つのマルチプルアライメントシステムは、いずれもこの評価尺度を使用している。

マルチプルアライメントされた結果は次のように利用できる。第一に、先に述べたように、重要な配列の部分であるモチーフを見出して、データベースから新たな類似配列を検索する助けとできる。第二に、類似した構造を持つ配列をアライメントした結果から、共通の部分構造に対応するアミノ酸の性質の並びがわかり、その部分構造の形態を予測する助けとできる。第三に、マルチプルアライメントから進化系統樹を描くことができ、類似配列の各々がどのような遺伝的過程を経てきたかを推測することができる。このようにマルチプルアライメントは、遺伝子情報処理の基本技術と位置づけることができる。

Dayhoffマトリックスのような評価尺度が与えられていれば、それにおいて最適なマルチプルアライメントを求めるダイナミックプログラミング法が知られている[Needleman 70]。しかしそれは必要とする計算量が多く、配列が３本以上ではあまり実用的ではない。そのため通常は、２本ずつの比較を繰り返してマルチプルアライメントを行う方法が試みられている[Barton 90]。だが、それでは精度が十分でなく、難しいところは、生物学者の勘に頼っている。すなわち、計算機で出力されるマルチプルアライメントは、経験を積んだ生物学者が行うマルチプルアライメントのレベルには、いまだに達してはいないということができる。

我々は、生物学者レベルのマルチプルアライメントを行う計算機システムの構築を目指し、まず３つの実験システムを並列推論マシン上に作成した。各システムは課題解決に異なる手法を採用した。次章から、それらをひとつひとつ紹介する。

2　並列３次元ダイナミックプログラミング法

2.1　概要

本方式の特徴は、Dayhoffマトリックス等のある評価基準が与えられたとき、アミノ酸配列３本の最適なアライメントを得ることができること、及び計算量の増大の克服を図るために並列化を行ったことである。従来、２次元ダイナミックプログラミング法を用いたアミノ酸２本のアライメントは生物学者の間でよく用いられていた。本方式はそ

れを3次元（3本）に拡張し、並列化を行ったものである。本方式を用いると、アライメントされた3本の組を複数組み合わせて、配列4本以上のマルチプルアライメントを比較的精度良く行うことができる。

これまでにも、ダイナミックプログラミングの多次元化を試みた例が存在しないわけではないが [Murata 85] [Carrillo 88]、いずれも部分的に計算を間引いた近似計算であり、最適なアライメントを得るものではなかった。我々は、計算量の増大を並列処理によって補完し、近似計算することなしに、3次元ダイナミックプログラミングの高速処理を実現した。

2.2 ダイナミックプログラミングによるアライメント

ダイナミックプログラミング（以下DPと略す）は段階的に決定を行う特徴を持つ最適化問題を解くためのアルゴリズムのひとつである。いくつもの段階で決定を行う必要があり、その各段階における決定が直前の段階の決定にのみ依存するという形式に最適化問題を定式化できるとき、DPを用いると非常に効率良く最適解（コスト最小の経路）を求めることができる。もしこの種の問題を、最初にすべての場合を尽くしてそのうちの最小のものを選ぶという解法で解いた場合、指数オーダーの計算量がかかる。しかしDPを用いると多項式オーダーで解を得ることができる。

アミノ酸配列のアライメントは、このDPを用いて行うことができる。簡単のために配列2本のアライメントについてDPの概念的説明を、図1を用いて行う。たとえば、ADHE, AHIEという2つの配列をアライメントする場合、この2つの配列を図のような2次元のネットワークの辺に対応させる。斜め方向のアーク（矢）は、そのアークの位置に対応する2つのアミノ酸の類似度がコストとして割り振られる。この類似度には前述のDayhoffマトリクスを用いている。また縦および横方向のアークはギャップに対応し、ギャップを挿入するときのコストが割り振られる。

ギャップコストは経験上、ギャップの長さkに対して、a＋bkのような一次式を与えるのが適当であり、DPにも通常そうしたギャップコストが実装される。ギャップコストの効率のよい実装法や、それに伴う計算量の分析も考察されている [後藤修 83]。

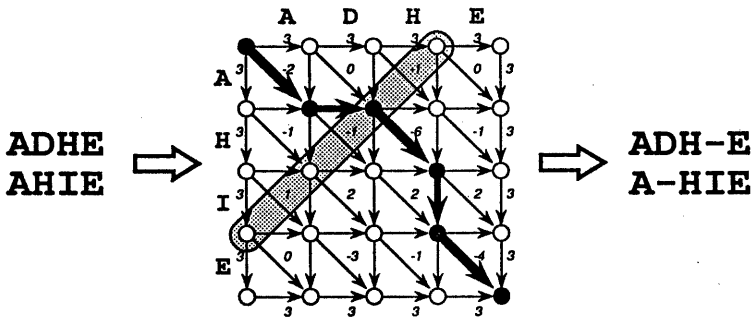

図1: 2次元DPマッチングによるアライメント

このように問題を定式化すると、最適なアライメントを求めることは、このネットワーク上のコスト最小の経路を求めることに対応する。図の例では太いアークで表された経路がコスト最小となる。この太いアークで表された経路を順に見ていくとAとAが対応し、Dに対応するもう片方のアミノ酸はなく（つまりギャップが対応し）、HにはHが対応し ... という具合に解釈することができる。結果として図の右側にあるアライメントが得られる。

コスト最小の経路は左上の端から右下の端に向かって（逆でも可能）各ノードに至る最短経路を段階的に決定していくことにより求めることができる。各段階は図1の右上がり斜め線上に存在するノード群に対応する。段階的に各ノードへ至る最短経路を求めていくと、いったん求まった部分的最短経路は、もはや変更されることがない。それゆえこの部分的最短経路を用いて次の段階の計算を行うことができる。具体的には、ある段階の各ノードの計算を行うためには、直前の段階の各ノード（2次元DPでは3つある）で求まった部分的最短経路のコストを参照して、今求めたいノードに至るコストをそれぞれ計算し、このうちの最小値を求めてそれをそのノードに至るコストとすればよい。直前のどのノードを選択したかという情報も記憶しておく。この操作を最後まで繰り返せば、ネットワーク全体の最短経路を求めることができる。

各段階で部分的最短経路を決定してしまうため、経路全体の組み合わせを考慮する必要はなく、ある段階から次の段階へ遷移するときの組み合わせだけを考慮すればよい。この性質があるために、全体の計算量が段階数の指数オーダーになることを回避できるわけである。これがDP手法の本質である。

2.3 並列3次元DPの実装

2次元DPによるアライメントを3次元に拡張すると、こんどは図2のような3次元のネットワーク内の最短経路問題として表現できる。3次元DPによるアライメントでは各段階が図の点線で囲まれた面の領域に存在するノード群に対応する。また2次元DPによるアライメントでは各ノードは直前の3つのノードと連結されていたが、3次元DPでは7つのノードと連結されている。この計算量の多い3次元DPを、並列推論マシン上で並列に動作させるプログラムを開発することにより、実行時間を削減させることができた。ここでは並列性を生かした実装の方法を簡単に説明する。詳細は[戸谷 91]を参照されたい。

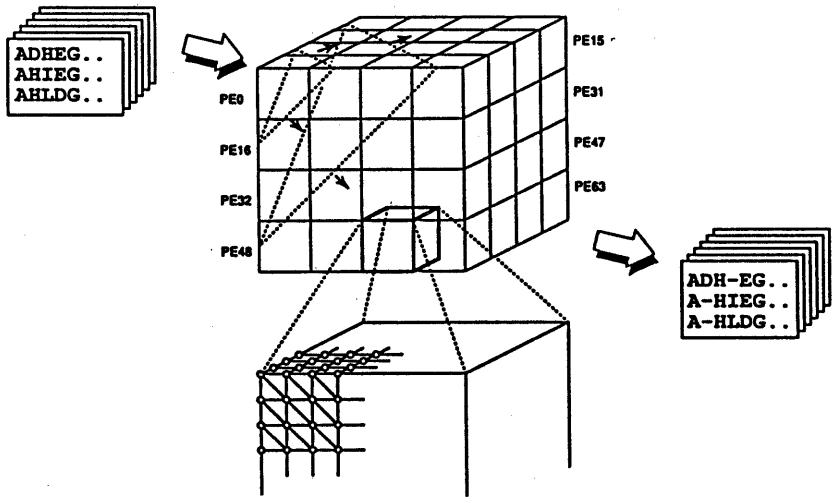

図2: 3次元DPによるアライメントとプロセッサへの割り当て

DPでは、ある段階の処理が終わらない限り次の段階の処理を行うことができない。つまりDPの各段階における処理は逐次的に伝播される。しかし各段階には複数個のノードが存在し、各ノードにおける処理は並列に実行可能である。また各段階に存在するノードの数が多くなればなるほど並列性が高い。3次元DPでは各段階に面状にノードが存在し、その数は非常に多いため並列性は高い。

さて、我々が使用する並列言語はKL1であり、その実行単位はプロセスである。したがって並列実行の主体である各ノードをKL1のプロセスで実現することにした。この各段階において計算を行うプロセスを段階が1つ進むごとに生成するのは効率が悪い。そこで3次元DPにおけるノードをKL1プロセスに、アークをKL1プロセスの通信路に、それぞれ対応させることを考えた。このように対応づけを行うと、3次元DPのネットワークをそのまま反映したKL1のプロセスネットワークを、あらかじめ生成してしまうことができる。

このようにプロセスネットワークを構築すると、各プロセスが隣接するプロセスとメッセージの授受を行うことによって全体の計算が進んでいくことになる。また、そのプロセスネットワークは、容易に3次元メッシュに分割でき、使用可能なプロセッサごとにそれを割り当て、並列実行を行うことが可能である。図2では縦、横、高さ、各方向にそれぞれ4分割されて、64個の要素プロセッサ（PE）にネットワークが割り当てられた例が示されている。メッシュに分割された各領域の処理は各PEが担当しており、ノードに対応するプロセスが多数割付けられている。また、このPEへのマッピングは、メッシュ構成でさえあれば、縦、横、高さ方向の切断数が可変であるように実装されている。

DPでは各段階の処理が波面状にメッシュの中を進んでいくことになる。すると、ある時点で波面を含まない部分のPEは稼働しなくなり、PEの利用効率が悪い。これを解消する方策としてアミノ酸配列3本の組を複数個、次々とネットワーク中に投入することを考えた。このようにするとパイプライン的な並列性が導入できるため、（最初と最後を除いて）PEが稼働しなくなることはほとんどなくなり、PE全体の稼働率を上昇させることができる。

整理すると、本並列処理方式には2種類の並列性が存在する。1つはDPの各段階におけるノード間の並列性で、各アライメントに対応する波面内の並列性ということができる。もう1つは複数のアライメントをパイプライン的に実行するときの並列性で波面間の並列性ということができる。

2.4 結果

アミノ酸５５個からなるタンパク３本のアライメント１組を、PE ６４台を用いて約４０秒で解決できる。パイプライン実行を行うと、２組以降は各々１０秒程度で求まる。これはプロセッサ１台での実行の、およそ３７倍の実行速度である。この３次元DPによるアライメント結果を組合せて、７本のマルチプルアライメントを形成したのが次のものである。

```
SMRV   :----G-FILATP-QTGEASKNVISH-VIHCLATIGKPHTIKTDNGPGYTGKNFQDF-CQKLQI-----
MMTV   :---YSHFTFATA-RTGEATKDVLQH-LAQSFAYMGIPQKIKTDNAPAYVSRSIQEF----LARW----
IAP    :----G-VMFATT-LTGEKASYVIQHCL-EAWSAWGKPR-IKTDNGPAYTSQKFRQF-CRQMDVT----
RSV    :------IV-VTQ-H-GRVTSVAVQHHWATAIAVLGRPKAIKTDNGSCFTSKSTREWLAR-WGIAH---
HTLV-1:---SG-AISATQKR-KETSSEAI-SSLLQAIAHLGKPSYINTDNGRAYISQDFLN-MCT--SLA----
HTLV-2:DTFSG-AVSVSCKK-KETSCETI-SAVLQAISLLGKPLHINTDNGPAFLSQEFQE-FC-----T----
BLV    :---H--A--S-A-KRGLTTQTTI-EGLLEAIVHLGRPKKLNTDQGANYTSKTFVR-FCQQFG-V-SLS
(score = -1093)
```

これは、(SMRV,MMTV,IAP),(MMTV,IAP,RSV),(IAP,RSV,HTLV-1),(RSV,HTLV-1, HTLV-2),(HTLV-1,HTLV-2,BLV)といった5組の3次元DPを行ったのち、それらの共通配列を利用して組合せていく。たとえば、(SMRV,MMTV,IAP)と(MMTV,IAP,RSV)の共通配列はMMTVとIAPであり、(SMRV,MMTV,IAP)のなかのMMTVとIAPのパターンと、(MMTV,IAP,RSV)のなかのMMTVとIAPのパターンとを比較して、4本のアライメント(SMRV,MMTV,IAP,RSV)を作る。

共通配列のパターンには、部分的に見ると、一致する部分と不一致な部分とが存在する。一致する部分は、他の1本の配列の影響にもかかわらず変化しない部分であるので、アライメントが強固な部分であり、信頼性も高いと考えられる。この一致部分に関しては、すぐさま、組合せが可能である。一方、不一致部分は、そのまま組合せることはできないので、なんらかの基準で優先づけをして、矛盾（不一致）の解消を行う。不一致部分は一致部分に比べて信頼性が低いと考えられる。以上のような手続きを繰り返して、5本以上のマルチプルアライメントが形成できる。

3 並列シミュレーテッドアニーリング法

3.1 概要

本システムは、汎用の組み合わせ最適化手法であるシミュレーテッドアニーリングをマルチプルアライメントに応用したものであり、並列化により事前の温度スケジューリングを不要にした、独自の方式を用いている。また本方式では、同時に可能性のある複数のアライメントを生成できるので、異なるアライメントを検討することにも有用である。本システムについて昨年、第一報を発表した[石川 90]。

3.2 シミュレーテッドアニーリングの基本アルゴリズム

シミュレーテッドアニーリングのアルゴリズムは、組み合わせ最適化問題でローカルミニマム（局所的にはコスト最小であるが、大局的にはそうでない点）に捕まらずに、グローバルミニマムを探索することを可能にするものである[Kirkpatrick 83]。

元来、アニーリングとは、物理系の焼きなまし過程を意味する。つまり、ある物質を高温から徐々に温度を下げることにより、非常に安定な物質が得られる過程を指している。シミュレーテッドアニーリングとは、この焼きなまし過程を模擬したアルゴリズムで、温度パラメータに依存して探索の範囲が決定される探索手法である。高温時においては、温度に依存させて、コストの悪化する方向の探索をも許し、徐々に温度を下げていくにつれて探索範囲を絞り込む。つまりコストの良くなるような方向にしか探索しなくなる。

具体的にアルゴリズムを説明すると、初期解X_0から順に次の様に解系列を生成していき、徐々に最適解に近い解を得ていく。まず、ある解X_nにランダムな微小変形を行うことで次の解の候補Y_nを作る。最小化を目的とする評価関数を$\mathbf{E}$とすると、評価値の変化は$\Delta E = \mathbf{E}(Y_n) - \mathbf{E}(X_n)$となる。$\Delta E \leq 0$ならば、無条件に$X_{n+1} = Y_n$とし、また、$\Delta E > 0$のような場合には、確率値として、$P = \exp(\frac{-\Delta E}{T_n})$を採用し、温度パラメータ$T_n$に依存させて、次の解を決定する。つまり、確率$P$で、$X_{n+1} = Y_n$とし、確率$(1-P)$で$X_{n+1} = X_n$とする。このオペレーションを多数回繰り返す。ここで温度パラメータ列$\{T_n\}$を適切に設定することにより、最適解を求めることができる。

この温度パラメータ列 $\{T_n\}$ を温度スケジュールと呼ぶ。限られた時間内に可能な限り最適解に近付くために、温度スケジュールを最適化することは非常に難しい問題である。そこで、次に述べる並列化 [木村 90][Kimura 91] により、温度スケジューリングを解消することを試みた。

3.3 シミュレーテッドアニーリングの並列化

各アニーリングプロセスごとに初期解を与え、それぞれのプロセスにおいては担当する温度パラメータで一定温度のアニーリングを行う。そして、温度を減少させることに対応させて、ある間隔置きに、隣接する温度を担当しているプロセス間で、解の交換を確率的に行う（図3）。その解交換を適当なタイミングで行うことにより、最適な温度スケジュールを見つけ出すことができる。その結果、適切な温度スケジュールを経た解が最終的に、低い温度プロセスに至る。

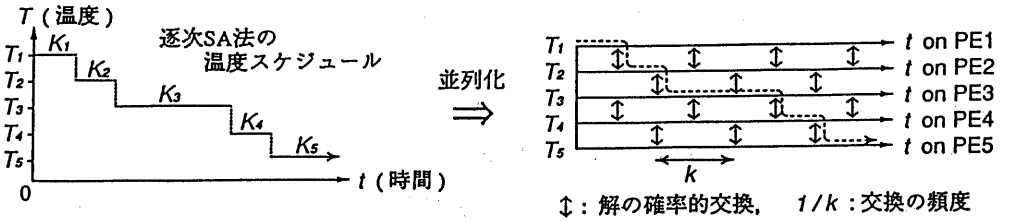

図 3: 並列シミュレーテッドアニーリング

温度パラメータ $T1$ において得られた解の評価値が $E1$、温度パラメータ $T2$ において得られた解の評価値が $E2$ の時 $\Delta E = E1 - E2$、$\Delta T = T1 - T2$ と置く。この時、プロセス間の解の交換確率は、

$$\mathbf{p}(T1, E1, T2, E2) = \begin{cases} 1 & \text{if } \Delta E \cdot \Delta T < 0 \\ \exp(\frac{-\Delta E \cdot \Delta T}{T1 \cdot T2}) & \text{otherwise} \end{cases}$$

として、定義される。これにより、各プロセス上における Boltzmann 分布に従う平衡状態を崩さずに解の交換を行うことが可能になり、十分に長い時間をかければ最適解が得られることが保証される。この関数により得られた確率値に従って、解を実際に交換するか、交換を見送るかの決定を下す。

3.4 マルチプルアライメントへの適用

アミノ酸配列のマルチプルアライメントをシミュレーテッドアニーリング法で実現した [金久實 89]。ここで、解の初期状態はアライメントしたい複数の配列であり、適当な数のギャップをあらかじめ配列の前後に付け加えた以下のものである。

```
SMRV   :-------GFILATPQTGEASKNVISHVIHCLATIGKPHTIKTDNGPGYTGKNFQDFCQKLQI-------
MMTV   :-------YSHFTFATARTGEATKDVLQHLAQSFAYMGIPQKIKTDNAPAYVSRSIQEFLARW-------
IAP    :-------GVMFATTLTGEKASYVIQHCLEAWSAWGKPRIKTDNGPAYTSQKFRQFCRQMDVT-------
RSV    :-------IVVTQHGRVTSVAVQHHWATAIAVLGRPKAIKTDNGSCFTSKSTREWLARWGIAH-------
HTLV-1:-------SGAISATQKRKETSSEAISSLLQAIAHLGKPSYINTDNGRAYISQDFLNMCTSLA-------
HTLV-2:-------DTFSGAVSVSCKKKETSCETISAVLQAISLLGKPLHINTDNGPAFLSQEFQEFCT-------
BLV    :-------HASAKRGLTTQTTIEGLLEAIVHLGRPKKLNTDQGANYTSKTFVRFCQQFGVSLS-------
(score = 877)
```

解に対する微小変形は次のように定義した。複数の配列のうちのある 1 本の配列に対して、ランダムにアミノ酸、ギャップをそれぞれ選択し、選択されたギャップを選択されたアミノ酸の反対側の隣に移動させる。ただし、両サイドのギャップは配列中にギャップが入り過ぎることを考慮して、ギャップがいくつあっても一つと見なして確率的に選択している。たとえば先の初期状態において、RSV の配列が選ばれ、中央部の A なるアミノ酸と、右端のギャップが選ばれた場合、次のように変形される。

```
SMRV  :-------GFILATPQTGEASKNVISHVIHCLATIGKPHTIKTDNGPGYTGKNFQDFCQKLQI-------
MMTV  :-------YSHFTFATARTGEATKDVLQHLAQSFAYMGIPQKIKTDNAPAYVSRSIQEFLARW-------
IAP   :-------GVMFATTLTGEKASYVIQHCLEAWSAWGKPRIKTDNGPAYTSQKFRQFCRQMDVT-------
RSV   :-------IVVTQHGRVTSVAVQHHWATAIAVLGRPK-AIKTDNGSCFTSKSTREWLARWGIAH------
HTLV-1:-------SGAISATQKRKETSSEAISSLLQAIAHLGKPSYINTDNGRAYISQDFLNMCTSLA-------
HTLV-2:-------DTFSGAVSVSCKKKETSCETISAVLQAISLLGKPLHINTDNGPAFLSQEFQEFCT-------
BLV   :-------HASAKRGLTTQTTIEGLLEAIVHLGRPKKLNTDQGANYTSKTFVRFCQQFGVSLS-------
(score = 866)
```

ランダムな選択については、合同法により生成される乱数列を使って選択を行っている。この微小変形操作により、ギャップが配列中に挿入された状態を生成し、残りの配列との比較により評価値を決定する。評価値 score の計算においては、各カラムにおけるアミノ酸の全組み合わせについて Dayhoff マトリックスの値を総和し、全カラムにわたって、それを合計したものを使用している。ギャップを含むペアのコストについては、ギャップの長さ k に対して、一次式 a + b k が設定でき、通常 a = 4, b = 1 にしている。

3.5 結果

以上のようなシステムを、並列言語 KL1 で記述し、並列推論マシンの実験機であるマルチ PSI 上に構築した。63の異なる温度プロセスを63台の要素プロセッサ（PE）に割り当てて、並列シミュレーテッドアニーリングを行った。先の例題を約1時間アニーリングしたところ以下の結果を得た。

```
SMRV  :-GFILATP--QTGEASK-NVISHVIHCLATIGKPHTIKTDNGPGYTGKNFQD-FC-Q-KLQI-
MMTV  :YSHFTFAT-ARTGEATK-DVLQHLAQSFAYMGIPQKIKTDNAPAYVSRSIQE-FL-A-RW---
IAP   :-G-VMFAT-TLTGEKAS-YVIQHCLEAWSAWGKPR-IKTDNGPAYTSQKFRQ-FC-R-QMDVT
RSV   :----IVVTQHGRVTSV--AVQHHWATAIAVLGRPKAIKTDNGSCFTSKSTREWLA-RWGIAH-
HTLV-1:-S-GAISA-TQKRKETSSEAISSLLQAIAHLGKPSYINTDNGPAYISQDFLN-MC-T-SLA--
HTLV-2:DTFSGAVSVSCKKKETSCETISAVLQAISLLGKPLHINTDNGPAFLSQEFQE-FC-T------
BLV   :---HASAK-RGLTTQTT---IEGLLEAIVHLGRPKKLNTDQGANYTSKTFVR-FCQQFGVSLS
(score = -1302)
```

4 トーナメント方式

4.1 概要

トーナメント方式によるマルチプルアライメントは、進化系統樹の考え方を用いてマルチプルアライメントを高速に求める方式である。つまり、よく似ている近いもの同士の配列は、アライメントしやすく、誤りも少ないことに注目して、近いものから順にトーナメント形式で、アライメントを決定していく手法である。

本方式では、現存する生物種（以下、生物と呼ぶ）が持つ特定のタンパク質に対応するアミノ酸配列が複数本ある場合に、これらの祖先として存在した生物の対応するアミノ酸配列（以下、祖先配列）を決定する。そして、この祖先配列を用いてマルチプルアライメントを求めるのである。また同時に、この過程において、生物の進化系統樹の例を構築することもできる。本方式は、トーナメント形式で共通配列を求める手法 [Smith 86] をマルチプルアライメントを求める手法に拡張したものである。他に進化系統樹を考慮してマルチプルアライメントを求める手法に [Hein 90] がある。本方式の詳細については [広沢誠 91] をも参照されたい。

トーナメント方式では配列間の類似度を求めるために $(n-1)^2$ 個（配列の本数を n 本とする）の2次元DPを行う必要があり、これは全体の計算量の大部分を占める。本方式では、これらのDPを並列に計算することにより高速にマルチプルアライメントを求めることができる。

4.2 トーナメントの決め方

本方式では、トーナメント形式で祖先配列を求めていく。つまり、祖先配列を求めていく順番がトーナメント状になっているのである（図4）。この順番は配列同士の類似度に基いて決まる。類似度としては2次元DPのコストを採用した。よく似た配列同士ほど、DPのコストが低いのである。

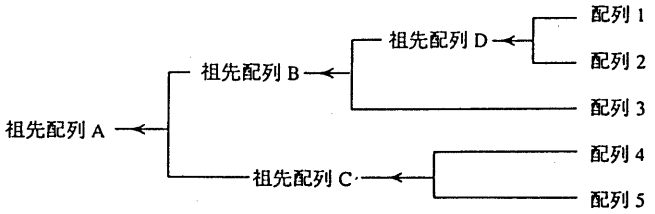

図 4: 祖先配列を求める

n個の配列の集合がある場合を考える。まず、配列のペアごとの2次元DPを全ての組合せで行う。このため、n（n－1）／2 回のDPを行う必要がある。そしてそのうちから、類似度が最大（コストが最低）である配列ペアを選び、それら二つの配列の祖先配列を求める。

次に、その二つの配列を配列集合から除き、代わりに、求められた祖先配列を配列集合に加える。そして、このn－1個の配列の、すべてのペアのDPコストを比較する。この時、重複する計算は前回の計算を利用できる。すなわち、実際に行うべきDPは、新しく加えた祖先配列と他のn－2個の間だけである。

その結果から、再び、類似度が最大である配列ペアを選び、それらの祖先配列を求める。以降、次々に祖先配列を求めて行く。こうして、結果としてn－1個の祖先配列を、トーナメント形式に求めることになる。祖先配列を求めるために必要なDPは、1個づつ減るので、i番目の祖先配列を求める時に必要なDPの数は n－i個である。

トーナメント方式では、DPを並列に行うことにより処理速度をあげることができる。膨大なDPの計算量に比較してその他の処理量を無視し、その並列効果を推定しよう。ここでは計算量を、1個のDPの計算量を単位として計ることにする。また、P個のプロセッサの内の1個を、管理プロセッサとし、その他の（P－1）個でDPを並列に行うとする。

まず、n個の配列のマルチプルアライメントの計算量は、最初に行うn個の配列間の $n(n-1)/2$ 個の DPと、その後に行う $(n-1)(n-2)/2$ 個の DPを合わせた $(n-1)^2$ 個のDPの計算量に対応する。そして、p個のプロセッサが全体として何サイクルのDPを行うかというと $\frac{(n-1)^2+\frac{n(p-1)}{2}}{p-1}$ 個である。したがって、並列効果の理論値は $\frac{p-1}{1+\frac{p-1}{2(n-1)}}$ となる。

この式から分かるように配列数が少ない場合には、並列効果は配列数が少ない場合は低いが、配列数が多くなるとプロセッサの数から1を引いた値に近づいていく。

4.3 祖先配列の求め方

最も類似しているものとして選ばれた二つの配列から祖先配列を求める方法について説明する。すでにそれらは2次元DPによってアライメントされているとする。

二つの配列の祖先配列を、この二つの生物（この生物を子孫、そして各配列を子孫配列と呼ぶ）の直接の祖先の生物が持っていた配列であると定義する。具体的には、祖先配列は、子孫配列をアライメントしたとき、対応する二つの文字に対して、それぞれ求めた祖先文字の並びとなる。たとえば、DIYA と DFA をアライメントして {DIYA, D-FA} が得られた場合、その祖先配列は、D と D の祖先文字、I と - の祖先文字、Y と F の祖先文字、A と A の祖先文字を求めることにより決められる。以下、子孫配列を構成する文字を子孫文字と呼ぶことにする。

祖先文字は、二つの子孫文字が祖先において、どのような文字であったかを示すものといえる。祖先文字としては、20個のアミノ酸を表す文字の他に、19個のアミノ酸のクラスを表す文字 $a \sim s$ を用いる（図5）。

$a \sim s$ は、子孫の二つの文字が異なる場合に、祖先文字がこの二つのどちらかであるという任意性があることを示すために用いる文字である。例えば、二つの文字が F と Y のどちらかであるという任意性があるということを a を用いて表す。類似しているアミノ酸は低い階層で1つのクラスにまとめられる。これとは反対に、類似していないアミノ酸は、高い階層において1つのアミノ酸にまとめられる。このクラス関係は、Dayhoff マトリクスから、類似性評価の高い順にアミノ酸をまとめて作ったものである。

次に祖先文字の決め方を述べよう。二つの子孫文字が同じ場合は祖先文字もこれと同じである。その他の特殊な場合には、以下に説明する Gap Handling Rule, Specificaation Rule, Generalization Rule といったヒューリスティクスを用いる。以下、一方の子孫文字を LetterA、他方の子孫文字を LetterB と表す。

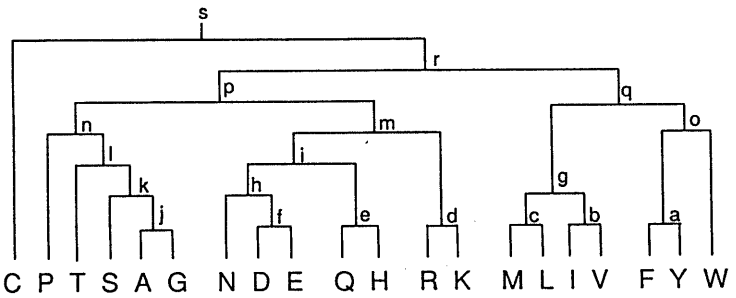

図 5: アミノ酸のクラスと文字の割当

1. Gap Handling Rule

   LetterA がギャップある場合、LetterB を祖先文字とするルールである。これは、祖先配列のこの場所にあった LetterB が生物 B の配列 B には継承されたが、生物 A の配列 A では 欠損が起こったことを意味している。

2. Specification Rule

   これは、図 5 において、LetterA が LetterB を包括している時に、LetterA の直接の子孫である LetterA1 または LettreA2 が LetterB に包括されているか、LetterB を包括している場合に、LetterB を祖先配列とするルールである。これは、取り得るアミノ酸の数が多い LetterA と、取り得るアミノ酸の数が絞られている LetterB がある時に、ある条件を満たすならば祖先文字として LetterB を選び、取り得るアミノ酸の数を絞ることを意味する。

3. Generalization Rule

   これは、Specification Rule を適用できない時に、図 5 において、LetterA と LetterB を包括する文字での中で、取り得るアミノ酸の数が最小の文字を祖先文字とするものである。

祖先配列を次々に求めていく方法を、DIYA, DFA, DIFT の祖先配列を求めることを例にして説明する。この中で、DIYA と DFA が最も類似しているとする。まず、DIYA と DFA のアライメントが必要であるが、これは類似度をＤＰにより求める時に計算されている。結果は {DIYA, D-FA} であるとする。D と D の祖先文字は D、I と - の祖先文字は Gap Handling Rule により I、Y と F の祖先文字は Genaralization Rule により a 、A と A の祖先文字は A であるので、祖先配列は、DIaA となる。その後に、この祖先配列と DIFT のアライメントを求める。結果は、{DIaA-, D-FAT} であるとする。なおアライメント時のクラス文字の評価は、クラスを構成するアミノ酸の平均値を使う。最後に、Specification Rule を用いて a と F から 祖先文字として F を求めれば、最終の祖先配列は DIFAT となる。

4.4 マルチプルアライメントの求め方

祖先配列が求まったならば、それからマルチプルアライメントを構成するのは比較的容易である。祖先配列を求めた手順と逆に、祖先配列を子孫配列のアライメントに置き換えていき、対応するカラムごとに整列させればよい。上記の例を用いて具体的に説明する。

DIYA, DFA, DIFT の祖先配列である、DIFAT を順次この配列の子孫配列に置き換えていく。まず、DIFAT を {DIaA-, D-FAT} に置き換える。そして、DIaA- に含まれる DIaA をこの子孫配列である {DIYA、D-FA} に置き換える。これにより、アライメントとして { {DIYA-, D-FA-}, D-FAT} 、つまり {DIYA-, D-FA-, D-FAT} が求まる。

前のふたつの章で扱った問題と同様な問題について、トーナメント法を適用して求めた解が次のマルチプルアライメントである。

```
SMRV   :-G-FILATRQTGEASKNVIS-HVI-HCL-A-TI-GKPHTIKTDNGPGYTGKNFQDFCQKL--QI--
MMTV   :YSHFTFATARTGEATK-DVLQHLAQSF--AY--MGIPQKIKTDNAPAYVSRSIQEFLARW------
IAP    :-G-VMFATTLTGE--K-A-S-YVIQHCLEAWSAWGKPR-IKTDNGPAYTSQKFRQFCRQM--DVT-
RSV    :---IV-VTQH-GRVTSVAVQHHWATAI--AV--LGRPKAIKTDNGSCFTSKSTREWLARWG--IAH
HTLV-1:------SGAISATQKRKETSSEAISSLLQAIAHLGKPSYINTDNGPAYISQDFLNMC----TSLA-
HTLV-2:-D--TFSGAVSVSCKKKETSCETISAVLQAISLLGKPLHINTDNGPAFLSQEFQEFC----T----
BLV    :-H------A-S-A-KRGLTTQTTIEGLLEAIVHLGRPKKLNTDQGANYTSKTFVRFCQQFGVSLS-
(score = -1031)
```

5 3つのアライメントシステムの比較

以上で説明してきた3つのシステムは、マルチプルアライメントの課題解決に、異なる手法を採用したため、それぞれが長所と短所を合わせ持つシステムとなった（図6）。

方　法	処理時間	解の信頼性	用　途
3次元DP	中（数分）	時間のわりには高い	ある程度良質の解を実用的な時間で出力
アニーリング	大（数時間）	時間をかければ非常に高い	アライメントの一部分を精度よく調整
トーナメント	小（数十秒）	ノイズに弱い	おおよその解を高速に発見

図6: アライメントシステムの比較

各システムの処理時間は、例として挙げている問題の規模であると、3次元ＤＰで数分、シミュレーテッドアニーリングで数時間、トーナメント法で数十秒といった具合である。トーナメント法が最も速いが、反面トーナメント法では、トーナメントの初期段階で近いもの同士のアライメントを誤ると、その誤差が後の段階のアライメント精度を大幅に下げてしまう問題点がある。そのため類似性の比較的高い配列群を、おおまかにアライメントするのに有効である。

シミュレーテッドアニーリングは、最も時間がかかっているが、さらに問題規模を上げると、妥当な解が得られるまで数日以上かかることになってしまう。そのためアニーリングは、実用的には、マルチプルアライメントの部分的な問題を対象にするのがよいと考えられる。3次元ＤＰは、他のふたつの方法の中間的性質を持ち、アニーリングほど時間はかからないうえに、トーナメントよりも信頼性の高い解を提供する。

これらのシステムは、生物学者が用途に応じて使い分けると有効利用できると考えられる。また各システムは、それぞれ違ったかたちで並列性を実現し、処理時間の逓減を図っており、並列処理方式の観点からも貴重な知見が得られている。

ウイルスの逆転写酵素の配列を、生物学者が人手で行ったマルチプルアライメントに以下のものがある[宮田　86]。

```
SMRV   :----GFILATPQTGE-ASKNVISHVIHCL-ATIGKPHTIKTDNGPGYTGKNFQDFCQKLQI---
MMTV   :--YSHFTFATARTGE-ATKDVLQHLAQSF-AYMGIPQKIKTDNAPAYVSRSIQEFLARW-----
IAP    :-----GVMFATTLTGEKASY-VIQHCLEAWSAW-GKPR-IKTDNGPAYTSQKFRQFCRQMDVT--
RSV    :------IVVTQH-GRVTSVAVQHHWATAI-AVLGRPKAIKTDNGSCFTSKSTREWLARWGIAH-
HTLV-1:---SGAISATQK-RKETSSEAISSLLQAI-AHLGKPSYINTDNGRAYISQDFLNMCTSLA----
HTLV-2:DTFSGAVSVSCK-KKETSCETISAVLQAI-SLLGKPLHINTDNGPAFLSQEFQEFCT-------
BLV    :--------HASAK-RGLTTQTTIEGLLEAI-VHLGRPKKLNTDQGANYTSKTFVRFCQQFGVSLS
(score = -1596)
```

これは、長いアライメントの一部分を抜き出したものであるが、2，3，4章で扱った問題と同一の部分である。各システムのアライメント結果を、この生物学者のアライメントと比べると、G,P,TDが縦に揃っている点など、どれも配列の特徴を共通して捕えているといえる。

しかし、Dayhoffの評価尺度で比べたところ、残念ながらどのアライメントシステムの結果も、高品質ではあるが、生物学者のそれと同等なscoreにまでは至っていなかった。そこで、さらに高度なアライメントシステムを模索した。

6 統合アライメントシステム

6.1 概要

生物学者のアライメントのレベルに相当する結果を出力することを目標に、3種のシステムの長所を統合したアライメントシステムを考案した。

トーナメント法は、配列間の距離を分析して近いものからアライメントする点はよいが、2次元DPを使用していたのでノイズに弱かった。それならば3次元DPを使用して、近いものからアライメントしたならばどうだろうか。一方、アニーリング法は、たくさんの配列にわたって類似部分を調整する能力があるが、初期状態からよい解に至るのにあまりにも時間がかかりすぎた。それならば初期状態として3次元DP法の組合せ結果を用いたらどうだろうか。このような発想から生まれたのが統合アライメントシステム（図7）である。

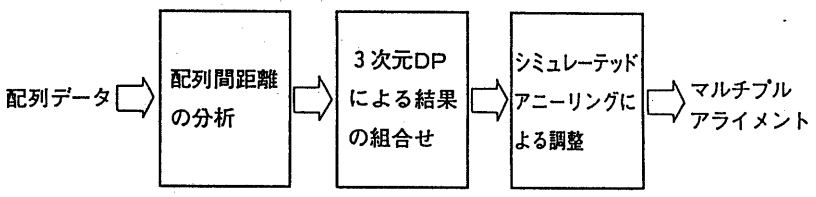

図 7: 統合アライメントシステム

6.2 統合システムの処理方式

統合アライメントシステムでは、まず配列3本のあらゆる組合せについて、その配列3本がどれだけ近接しているかを評価する。それには各組ごとに3次元DPを行い、アライメントに伴うコストから近接度を評価するのがひとつの方法である。より効果的なもうひとつの方法に、それぞれの組においてペアごとに3つの2次元DPを行って、それらのコストの和をとる方法がある。この方法の利点は、第1に処理量が小さいことである。処理量が小さいにもかかわらず、3次元DPを素直にやるのに比べ、結果はほとんど変わらないことがわかっている。第2の利点は、配列3本の分散を盛り込める点である。2次元DPのコストの和が同様であっても、配列が均等に近接している組の方が、あるペアのみが特別に他のペアより近接している組よりも、3次元DPの結果の信頼性が高いと推定できる。だから、配列3本の分散が小さければ、より近接度合が高いと評価できる。[広沢 91]も参照されたい。

配列3本の近接度合がすべての組合せで判明したならば、近接度合の高い3本組から、次の規則で3本組を順次サンプルしていく。始めに、最も近接度合の高い3本組をサンプルする。次に、その3本組と共通配列を2本持つ3本組のうち、最も近接度合の高い3本組をサンプルする。さらに、すでにサンプル済みの組に含まれる配列群と、2本の配列が共通配列になる3本組のうち、最も近接度合の高い3本組をサンプルする。そして、この操作を、サンプル済みの組にすべての配列が含まれるようになるまで繰り返す。こうして、3本組のサンプル列ができる。

3本組のサンプル列ができたならば、そこに含まれるすべての3本組について3次元DPを行い、3本のアライメントをサンプル個数求める。その結果を近接度合の高い3本組から順に、2章と同様に、共通配列2本のアライメントパターンを手がかりにして組み上げていく。共通配列2本のアライメントパターンが同一で矛盾がないときには、ただちに組み合わせることができる。しかし、それらのアライメントパターンが異なるときには、その部分について次の手順で矛盾解消をする。アライメント｛共通配列1，共通配列2，他の配列A｝とアライメント｛共通配列1，共通配列2，他の配列B｝が矛盾を起こしているとき、｛共通配列1，共通配列2，他の配列A｝と｛他の配列B｝との2次元DP、｛共通配列1，他の配列A｝と｛共通配列2，他の配列B｝との2次元DP、｛共通配列2，他の

配列A｝と｛共通配列1，他の配列B｝との2次元DP、｛他の配列A｝と｛共通配列1，共通配列2，他の配列B｝との2次元DPのうち、最もコストの低いアライメントを、該当部分のアライメントとする。このようにして、すべての矛盾箇所を解消すれば組上げが完成し、マルチプルアライメントが得られる。

3次元DPを用いただけであると、近い配列群の類似性しか考慮されていないので、場合によって、遠い配列群にわたる類似性が部分的に捕えられていない可能性がある。そこで、最後にシミュレーテッドアニーリングを用いて、複数配列にわたる類似性の評価に基づき、1時間程度のアライメントの微調整を行う。こうして得られたマルチプルアライメントが、統合システムの解析結果となる。

6.3 統合システムの処理結果

この統合システムの性能は比較的高く、先ほどのアライメント課題について、次のような良好な解を出力した。

```
SMRV   :----GFILATPQTGEASKNVI-SHVIHCLATIGKPHTIKTDNGPGYTGKNFQDFCQKLQI---
MMTV   :--YSHFTFATARTGEATKDVL-QHLAQSFAYMGIPQKIKTDNAPAYVSRSIQEFLARW-----
IAP    :----GVMFATTLTGEKASYVI-QHCLEAWSAWGKPR-IKTDNGPAYTSQKFRQFCRQMDVT--
RSV    :------IVVT-QHGRVTSVAVQHHWATAIAVLGRPKAIKTDNGSCFTSKSTREWLARWGIAH-
HTLV-1:---SGAISATQKRKETSSEAI-SSLLQAIAHLGKPSYINTDNGRAYISQDFLNMCT--SLA--
HTLV-2:DTFSGAVSVSCKKKETSCETI-SAVLQAISLLGKPLHINTDNGPAFLSQEFQEFCT-------
BLV    :-------HASAKRGLTTQTTI-EGLLEAIVHLGRPKKLNTDQGANYTSKTFVRFCQQFGVSLS
(score = -1755)
```

生物学者の解(score = -1596)よりも、Dayhoffの評価尺度において、上回った値(score = -1755)を得たため、生物学的に意義のあるシステムとなっていると判断できる。しかし、マルチプルアライメントには生物学的ノウハウがあり、必ずしも特定の評価尺度が絶対とはいえないので、これでもって統合システムが、生物学者よりアライメント能力に優るとはまったくいえない。むしろ、生物学者が検討するのに貴重な情報を提供するシステムといえよう。

さらに、より困難な課題にも挑戦した。その課題では下のようにジンクフィンガーという配列モチーフを抽出した。ヒスチジンHとシステインCがふたつずつ対になっている部分がジンクフィンガーと呼ばれ、通常DNAに結合する機能があることが知られている。ジンクフィンガーの周辺では配列の類似性が低く、そのモチーフは、マルチプルアライメントによる抽出が難しいとされている。

```
-------ILD--F-------HEKLLHPGIQKTTK-LF--GET--YYFPNSQLLIQNIINECSICNLAKTEHR---N-TDMPTKTT
-------LLD-FL-------HQ-LTHLSFSKM-KALLERSHSPYYMLNRDRTL-KNITETCKACAQVNASKS---A-VKQGTR--
LTDAL-LITP-VLQLSPAELHS-FTHCGQTAL-T-LQ--------GATTTEA--SNILRSCHACRGGNPQHQMPRGHI-------
VADSQATFQAYPLR-EAKDLHT-ALHIGPRAL-S-KA-------CNISMQQA--REVVQTCPHC---NSAPALEAG-VN------
-------ISD-PIH-EATQAHT-LHHLNAHTL-R-LL-------YKITREQA--RDIVKACKQCVVATPVPHL--G-VN------
-------ILT-ALE-SAQESHA-LHHQNAAAL-R-FQ-------FHITREQA--REIVKLCPNCPDWGSAPQL--G-VN------
```

統合アライメントシステムが、こうしたモチーフの抽出に成功したことは、同システムが、これまで知られていない重要なモチーフをも、新たに抽出できる能力を持ち合わせていることを意味する。すなわち、未知のタンパク質の構造や機能を予測するうえで、同システムが有力なツールとなることが期待できるのである。

7 おわりに

我々の開発した3つのアライメントシステムと、その統合アライメントシステムは、マルチプルアライメントの問題を高速に並列処理し、生物学的にも意義がある高品質の解を提供する。これらの事実は、遺伝子情報処理の分野に、並列推論マシンが有効に活用できることを実証しているといえよう。なお、各要素技術の詳細について、本年後期の情報処理学会全国大会でも発表の予定である(3S-6,7,8)。

今後の課題は、第1に問題規模をあげることである。現在の統合システムは、アミノ酸の個数にして百以下の比較的短いタンパク質しか対応できない。より高速な実装も試みる予定であるが、マルチPSIの上位機種であるPIMが完成すれば、実行速度が数十倍となり、アミノ酸2百個以上の長いマルチプルアライメントが可能になる見通しである。そうなれば、実用規模のほとんどのアライメント問題に対応できることが予想される。

第２の課題は、類似性の低い配列群のマルチプルアライメントに挑戦することである。類似性が低いとアライメントが難しく、生物学者もほとんど手掛けていない。この課題に挑戦すべく、配列間距離の高度な分析法を研究している。また、すでに知られているモチーフをアライメントの手助けにするため、モチーフを集めたモチーフ辞書を利用する、知識処理技術の検討も行っている。

謝辞

ＩＣＯＴ遺伝子情報処理ワーキンググループ（主査：金久實）に参加されている、委員の方々、オブザーバーの方々、そして講師として参加された方々には、多くの助言や批判をいただきました。ここに深くお礼申し上げます。

最後に、本研究の機会を与えていただいた、ＩＣＯＴ研究部長の内田俊一氏に感謝いたします。

参考文献

[Bilofsky 88] Bilofsky, H. S. and Burks, C. *"The GenBank Genetic Sequence Data Bank"* in *Nucleic Acids Research 16:5,* 1988, pp.1861–1863.

[五條堀 90] 五條堀、森山、内藤、河合：大量ＤＮＡデータを対象とした遺伝情報のコンピュータ解析，情報処理 *Vol.31 No.7,* 1990, pp.878–886.

[金久 91] 金久、新田、小長谷、田中：知識情報処理技術とヒトゲノム計画，人工知能学会誌 *Vol.6 No.5,* 1991.

[Iyengar 88] Iyengar, A. K. *"Parallel DNA Sequence Analysis"*, MIT/LCS/TR-428, 1988.

[Butler 90] Butler, Foster, Karonis, Olson, Overbeek, Pflunger, Price and Tuecke *"Aligning Genetic Sequences"* in *Strand: New Concepts in Parallel Programming,* Prentice-Hall, 1990.

[木村資生 86] 木村資生：分子進化の中立説，紀伊国屋書店，1986.

[Dayhoff 78] 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.

[Needleman 70] Needleman, S. B. and Wunsch, C. D. *"A General Method Applicable to the Search for Similarities in the Amino Acid Sequences of Two Proteins"*, in *Journal of Molecular Biology 48,* 1970, pp.443–453.

[Barton 90] Barton, J. G. *"Protein Multiple Sequence Alignment and Flexible Pattern Matching"* in *Methods in Enzymology Volume 183* Academic Press, 1990, pp.403–428.

[Murata 85] Mitsuo Murata *"Simultenious Comparison of Three Protein Sequences"* in *Proc. Natl. Acad. Sci. USA Vol. 82* 1985, pp.3073–3077.

[Carrillo 88] Himberto Carrillo and David Lipman *"The Multiple Sequence Alignment Problem in Biology"* in *J. Appl. Math. 48* 1988, pp.1073–1082.

[後藤修 83] 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.

[戸谷 91] 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.

[石川 90] 石川、戸谷、星田、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，＇知識情報処理技術とヒトゲノム計画＇講演要旨集,1990,A-4.

[Kirkpatrick 83] Kirkpatrick,S., Gelatt,C.D. and Vecci,M.P. *"Optimization by Simulated Annealing"* in *Science vol.220 no.4598* 1983.

[木村 90] 木村、瀧: 時間的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.

[Kimura 91] Kimura,K. and Taki,K *"Time-homogeneous Parallel Annealing Algorithm"* in *Proc. Comp. Appl. Math. 13 (IMACS'91)* 1990, pp.827–828.

[金久實 89] 金久實: 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法，分子生物学会年会，1989.

[Smith 86] Smith,R and Smith,T *"Automatic generation of primary sequence paterns form sets of related protein sequences"* in *Biochemistory Vol.87,* 1990.

[Hein 90] Jotun Hein *"Unified Approach to Alignment and Phylogenies"* in *Methods in Enzymology Volume 183* Academic Press, 1990, pp.626–645.

[広沢誠 91] 広沢誠：祖先配列を用いたマルチプルアライメント，ＩＣＯＴテクニカルメモ，*No.1069,* 1991.

[宮田 86] 宮田、藤、林田：コンピューターによる逆転写酵素遺伝子の探査，サイエンス，1986年２月号，pp.86–97.

[広沢 91] 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，1991.