並列シミュレーテッドアニーリングによるマルチプルアライメント[1]

石川幹人，戸谷智之，星田昌紀，新田克己[2]
(財) 新世代コンピュータ技術開発機構 (ICOT)

荻原 淳，金久 實[3]
京都大学化学研究所

1 はじめに

近年、DNA の核酸配列やタンパク質のアミノ酸配列の決定手法が確立され、配列データを蓄えたデータベースが急速に膨らんでいる。そのため、それらの大量の配列データを分類整理する必要性が生じているが、それに必要な配列解析技術は、まだ十分に確立されているとはいえない。

重要な配列解析技術のひとつに、複数の配列間の類似性を分析して配列の特徴を捉えるマルチプルアライメントがある。従来マルチプルアライメントは、熟練した者が人手で行っていたが、最近は計算機による自動化も試みられている。しかし、最適なマルチプルアライメントを求めるには、たいへん多くの計算を必要とするので、計算機による手法のほとんどが、ペアワイズのアライメントを組み合わせてマルチプルアライメントを求めている [Barton 90]。

我々は、シミュレーテッドアニーリングの手法を応用し、多数の配列にわたる類似性を同時に評価してマルチプルアライメントを求めるシステムを構築した。本システムについて昨年、第一報を発表した [石川 90]。本システムは、シミュレーテッドアニーリングを温度並列の方式で並列計算機上に実装したものであり、簡便に準最適なマルチプルアライメントを得られる特徴を持つ。相同性が低い配列群の処理には、とくにブロックオペレーションを用意している。

本論文の構成は、まず次章で、シミュレーテッドアニーリングの手法を紹介したのち、第3章で、我々が開発したマルチプルアライメントへの適用法について解説する。第4章では、我々が行ったシミュレーテッドアニーリングの並列実装法を説明し、続く第5章では、その実験結果を示す。また、相同性の低い配列群のアライメントのために考案したブロックオペレーションを、第6章で解説する。最後に第7章で、まとめと今後の課題を述べる。

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、組合せ最適化問題でローカルミニマム（局所的にのみエネルギー最小な点）につかまらずに、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摸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、エネルギーの悪化する（大きくなる）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。つまりエネルギーの良くなる（小さくなる）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エネルギー関数を E とすると、エネルギー値の変化は $\Delta E = \mathrm{E}(Y_n) - \mathrm{E}(X_n)$ となる。$\Delta E \leq 0$ ならば、良い変形であるので無条件に $X_{n+1} = Y_n$ とし、また、$\Delta E > 0$ のような場合には、エネルギーが悪化する変形である。そのため、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エネルギー値が収束するまで、多数回繰り返す。

ここで温度パラメータ列 $\{T_n\}$ は温度スケジュールと呼ばれ、それを適切に設定すれば、十分な時間ののち、最適解を求めることが理論的には可能である。しかし現実には、限られた時間内に準最適な解を求める場合が多い。そのための適切な温度スケジュールは、扱う問題ごとに異なり、それを設計するには少々骨がおれる。

3 マルチプルアライメントへの適用

マルチプルアライメントにシミュレーテッドアニーリングを適用するためには、マルチプルアライメントの問題を、組合せ最適化問題として定式化しなければならない。つまり、ある解の状態から近隣の状態へと移るオペレーションである

[1] Multiple Alignment by Parallel Simulated Annealing

[2] Masato Ishikawa, Tomoyuki Toya, Masaki Hoshida, Katsumi Nitta [Institute for New Generation Computer Technology (ICOT)]

[3] Atsushi Ogiwara, Minoru Kanehisa [Institute for Chemical Research, Kyoto University]

微小変形と、各状態における評価尺度にあたるエネルギー関数を定義する必要がある。我々の行った定式化について、以下に記す。なお我々は、タンパク質のアミノ酸配列のマルチプルアライメントを想定している。

3.1 微小変形の定義

マルチプルアライメントの結果は内側に不定数のギャップを含むので、微小変形において、ギャップをどのように扱うかが鍵となる。我々は、配列の頭部や尾部に、あらかじめ十分な数のギャップを付加する方法をとった[金久 89]。たとえば、アライメントの初期状態として、まったくギャップの入ってない状態を採用するならば、次のようなアライメント状態が初期状態となる。

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

そして、状態に対する微小変形は次のように定義する。複数の配列のうちのある1本の配列に対して、任意のギャップと任意のカラム位置をそれぞれランダムに選択し、選択されたギャップを選択されたカラム位置に移動させる。そして、間の部分の配列を、移動したギャップがあった方へ1カラム分移動する。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する。

たとえば先の初期状態において、RSV の配列が選ばれ、その配列の右端のギャップと、中央部の A なるアミノ酸のあるカラム位置が選ばれた場合、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--------
(energy = 866)
```

3.2 エネルギー関数の定義

マルチプルアライメントにシミュレーテッドアニーリングを適用した場合、取り扱う全配列にわたって同時に評価を行える利点がある。こうした評価の値を各状態に対して与えるのがエネルギー関数である。

現在エネルギー関数として、ある配列ペアにおいて、各カラムにおけるアミノ酸ペアについて Dayhoff マトリックス[Dayhoff 78] の値を総和し、それをすべての配列ペアについて、合計したものを使用している。Dayhoff マトリックスは、該当アミノ酸ペアが偶然に対して何如に少ないかを数値化したものであり、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。エネルギー関数は小さい程よいとする慣例から、マトリックスの数値の符号を反転して、類似しているアミノ酸ペアについて負の値となるようにしている。

Dayhoff マトリックスには、ギャップ対アミノ酸の評価の指定はないが、ギャップの長さ k に対して、一次式 $a+bk$ が設定できるのが、望ましい[後藤 83]。我々のシステムも $a+bk$ が設定できるようになっており、通常は、$a=4$, $b=1$ にしている。当然のことながら、ギャップ対ギャップのペアは無視される。

さらに、我々のエネルギー関数では、配列の頭部や尾部にあるギャップがアウトギャップとして特別扱いされており、アウトギャップ対アミノ酸の評価を特別に指定できる。その値を通常は 0 にしておくことによって、配列の水平位置がどんなに食い違っていても、ペナルティが与えられることがなくなる。そのため、類似部分の位置が配列によって、先頭部分にあったり、後尾部分にあったりする複雑なマルチプルアライメントが効果的に行える。

4 シミュレーテッドアニーリングの並列実装

シミュレーテッドアニーリング (SA) を並列に行うには、いくつかの方法がある。最も単純な方法は、通常の逐次的に動く SA を、利用可能な要素プロセッサ (PE) の数だけ独立に (異なる乱数で) 行い、そのうちの最もよい解を選ぶものである (単純並列 SA)。それに対して、我々が用いた方式[木村 90] は、各 PE ごとに異なる温度を割り当て、温度スケジュールがなかば自動的に行われる方式である (温度並列 SA)。温度並列 SA は単純並列 SA に比べ、同一時間で同等程度か、場合によっては同等以上の解を得られる[Kimura 91]。とくに、その時々における最良の解が刻々と得られる利点を持つ。この温度並列の方式を、以下に詳しく述べる。

各PEごとに初期解を与え、それぞれのPE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、ある間隔置きに、隣接する温度を担当しているPE間で、解の交換を確率的に行う。この解交換はより良い解はより低温のPEに移動するようになっており、それを十分な頻度で行うことにより、適当な温度スケジュールを経た解が最終的に、低い温度のPEに至る（図1）。

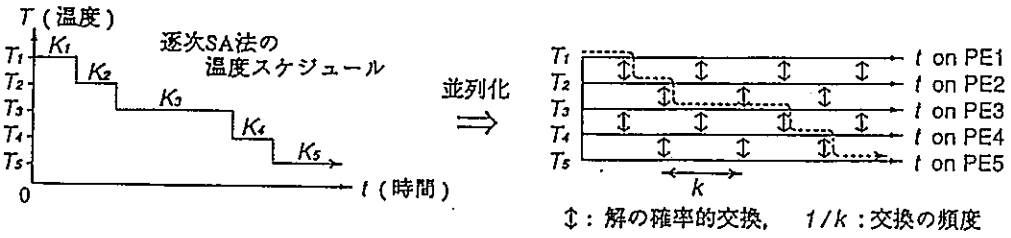

図 1: Temperature Parallel Simulated Annealing

この解の交換確率は、温度パラメータ $T1$ において得られた解のエネルギーが $E1$、温度パラメータ $T2$ において得られた解のエネルギーが $E2$ の時 $\Delta E = E1 - E2$、$\Delta T = T1 - T2$ とおけば、次式で定義される。

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

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せば、各温度に対する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。

5 実験と結果

以上で述べたようなシステムを、並列言語KL1で記述し、並列推論マシンMulti-PSIの上に構築した。高温から低温まで63の異なる温度を、63台の要素プロセッサ（PE）に割り当てて、温度並列SAを行った。先に示した初期状態から、約1時間アニーリングしたところ、次の結果を得た。類似部分をよく捉えている。

```
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---
(energy = -1302)
```

この問題に関して、温度並列SAによるエネルギー低下を、時間を追って、逐次SAや単純並列SAと比べたのが図2である。ふたつの並列SAが、逐次SAに比べ、かなり良いエネルギー値を示しているのがわかる。さらに温度並列SAは、単純並列SAに比べても、ほとんどつねに良いエネルギー値を示しているのがわかる。しかし、最終のエネルギー値には両者の差異はほとんどない。これは単純並列SAの温度スケジュールの設計が、割合に良かったためであろう。

この最終のエネルギー値の状態が多少不満であるとき、温度並列SAの長所がでてくる。このような場合、単純並列SAでは、再び始めから適当な温度スケジュールでやり直さねばならない。それに対し、温度並列SAは、つねに最低温度のPEをモニタしていれば、そこにより良い解が次々と現れるので、単に、もうしばらくの時間、待てばよい。

また、シミュレーテッドアニーリングで問題を解く場合に、初期状態として近似解を入れ、中低温からアニーリングすると効果的なことがある。このアニーリングには、その近似解に合わせた温度スケジュールが必要であるが、こうした場合にも、温度並列SAは温度スケジュールの吟味が不要で、気軽に使用できる。事実、他の方法で求めた近似解（energy = -1093）を初期状態にして、温度並列SAで約1時間アニーリングしたところ、良好な解（energy = -1755）が得られた[石川 91]。前に示したような、ギャップが全く入っていない初期状態からでは、3時間アニーリングしても、このエネルギー値には至らない。

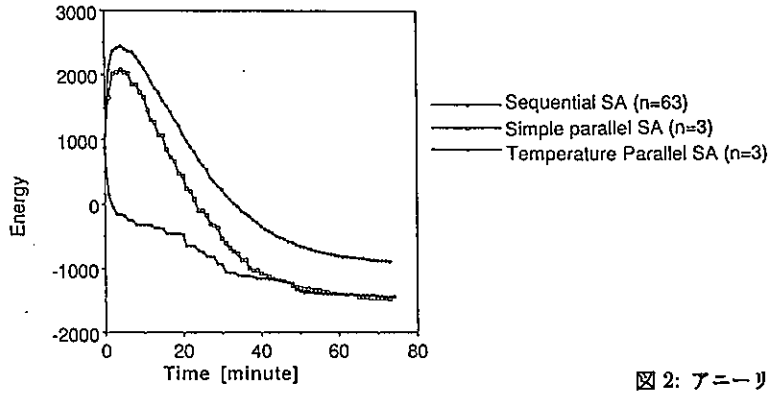

図 2: アニーリング経過の比較

6 ブロックオペレーションの導入

以上のシステムは、相同性の高い配列群のマルチプルアライメントは比較的うまく行えるのに対し、ギャップが固まりで入るような相同性の低い配列間のマルチプルアライメントは、苦手であった。それは第3章で定義したような微小変形では、ギャップの固まりが配列内部に形成されにくいためである。

その欠点を補うために、ギャップを長方形の固まりで動かすブロックオペレーション [荻原 90] を導入した。ブロックオペレーションは、あるギャップをランダムに選んだならば、そのギャップの横方向や縦方向にギャップの連なりを探し、矩形ブロックの単位でギャップ群を移動させる微小変形である。

従来の単独ギャップ移動モードの他に、横長ブロック移動モード、縦長ブロック移動モード、横優先矩形ブロック移動モード、縦優先矩形ブロック移動モードを設け、相同性の低い配列間のマルチプルアライメントに対処した。

7 おわりに

我々はマルチプルアライメントの課題を、シミュレーテッドアニーリングの手法で解決するシステムを、並列計算機上に構築した。本システムは、シミュレーテッドアニーリングを温度並列で実装しており、従来の方法で行ったマルチプルアライメントを精緻化するのに、とくに有効である。また、相同性の低い配列群のアライメント効率を上げるために、ブロックオペレーションを導入した。

今後の課題は第一に、マルチプルアライメントに熟練している研究者にインタビューして、アライメントをうまく行うための知識を抽出することである。その抽出された知識は、より効果的なオペレーションの導入や、より良いエネルギー関数の設定につながる。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。

参考文献

[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.

[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.

[金久 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.

[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.

[後藤 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.

[木村 90] 木村、霜: 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)* 1991, pp.827–828.

[石川 91] 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.

[荻原 90] 荻原、金久 : "パターン認識を取り入れたマルチプルアライメント法", 日本生物物理学会第 28 回年会, 1991.