並列シミュレーテッドアニーリングとタンパク質配列解析

1N−5

戸谷智之[1]，石川幹人[1]，星田昌紀[1]，荒木均[2]

1: (財) 新世代コンピュータ技術開発機構　2: 松下電器産業株式会社 情報通信東京研究所

1 はじめに

多数のタンパク質のアミノ酸配列の類似性を解析する有効な手法であるマルチプルアライメントは、組み合わせ最適化問題として定式化可能である [1]。シミュレーテッドアニーリング（SA）[2] は汎用の組み合わせ最適化問題の有効な解決手法のひとつである。また、順位検定を用いることにより、並列 SA において、実行時間の短縮を図ることができる [3]。我々は、昨年の温度並列 SA[4] に引き続き、この新たな並列 SA もマルチプルアライメントの問題に適用し、性能の考察を行った。

2 シミュレーテッドアニーリング

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。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、ある物質を高温の状態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。SA を用いて、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 $\{T_n\}$（クーリングスケジュールと呼ばれる。）は、任意の温度パラメータで一定回数微小変形を繰り返した後、温度を適切な比率で段階的に下げていくというものが単純ではあるが、かなり良い結果をもたらすとされている [2]。

3 シミュレーテッドアニーリングの並列化

SA は温度一定の微小変形を繰り返している間は、エネルギー平衡状態に向かう統計力学的な系としてとらえることができる。したがって、各温度で平衡状態に十分近付いた後に、温度を次の段階へと下げるようにしていくと、エネルギー最小の状態に収束することができる。このような SA の性質から、高温部では微小変形回数を多くする必要はないが、中、低温部で微小変形回数が少ないと局所的最小値につかまる可能性があり、この可能性は微小変形回数が少なくなればなるほど大きくなる。各温度で一定回数実行していたのでは、どうしても高温部で必要以上の微小変形を行うこととなり、全体の実行時間を無駄に伸ばしてしまいがちである。そこで、高温部では複数プロセッサにおける解のエネルギー分布から平衡状態か否かを検出することにより、変換回数を削減する手法を提案する。まず、各プロセッサには初期状態として、状態を一つづつ与え、同じ温度からアニーリングを開始する。各温度において一定回数の微小変形を繰り返した後、温度を適当な比率で段階的に下げていくというクーリングスケジュールの設定の仕方がもっとも簡単ではあるが、このシステムでは、各温度でアニーリングを開始する時点で、状態のエネルギーの上位と下位の 2 グループにグループ分けする（図 1(a)）。その温度での実行が進むにつれ、各プロセッサの状態は変化し、2 つのグループ間でエネルギー分布に差がなくなった時点（図 1(b)）で平衡状態に達したものと判断を下し、温度を次の段階に下げ、再び、エネルギー上下の 2 グループに分け直し、と同様な処理を繰り返す。エネルギー分布の違いの検出には順位検定 [3] を用い、閾値はパラメータとして与えることとした。

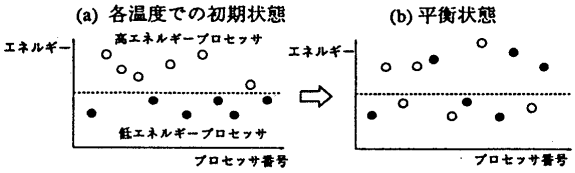

図 1: 平衡状態の検定

4 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 SA を適用するためには、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。

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。そこで、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。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

Parallel Simulated Annealing and its Application to Protein Sequence Matching

Tomoyuki TOYA[1], Masato ISHIKAWA[1], Masaki HOSHIDA[1], Hitoshi ARAKI[2]

1:Institute for New Generation Computer Technology (ICOT)

2:Matsushita Electric Industrial Co.,Ltd.

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。また、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に対応するため、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[5]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。

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[6]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マトリックスには、ギャップ対アミノ酸の評価値としての指定はないので、パラメータとして取り扱えるようにしている。ギャップを含むペアのコストについては、長さkのギャップに対して$a+bk$のような一次式が与えられる。aとbは可変であるが、上記のアミノ酸のDayhoff値とのかねあいから、通常は、$a=4$, $b=1$としている。以上のように定式化して、アライメントを行った。

5 システムの試作、及び評価

本システムは、並列論理型言語KL1で記述し、要素プロセッサ（PE）64台構成の並列推論マシンPIM[7]上で実行を行った。

本プログラムは、アニーリングプロセスとマネージャプロセスという2種類のプロセスから構成されている。アニーリングプロセスには初期状態、初期温度、減温比率、1温度における微小変形の上限回数などが与えられており、初期温度からアニーリングを実行するプロセスのことである。アニーリングプロセスはマネージャプロセスの置かれたPEを除く全てのPEに1つづつ置かれる。それぞれのアニーリングプロセスはマネージャプロセスとの間に通信路（ストリーム）を保持しており、指定したインターバルごとに各アニーリングプロセスからエネルギー値がマネージャプロセスに報告される。全てのPEのアニーリングプロセスから情報が届いた時点で、マネージャプロセスは温度を下げるか否かを順位検定ルーチンにより決定し、収束チェックをした後、各アニーリングプロセスに、決定された命令を与える。各アニーリングプロセスは、マネージャプロセスからの命令に従い、温度を維持、もしくは減温比率に従い温度を下げアニーリングを続行するか、収束条件に達した場合にはアニーリングを終了させるかする。

約60個のアミノ酸から構成される蛋白質の一部分の前後にあらかじめギャップを15個前後ずつ付け加えた、全長90の文字列、6本をサンプル問題として実験を行った。比較のために、逐次SA、単純並列SA、検定機能を付加した当SAのそれぞれに対して、Lu（：各温度での微小変形回数の上限）を変化させて、データを採取した（図2）。逐次SAの場合に乱数の違いにより生じた解の質のばらつきを防ぐ効果が両並列SAには現れている。並列SAの間では、従来からアライメントを行うのに一般的に利用されてきたTree-based Methodと比較し、同等以上の解をえることができるLuにおいては、大幅なアニーリング時間の短縮を実現することができている。

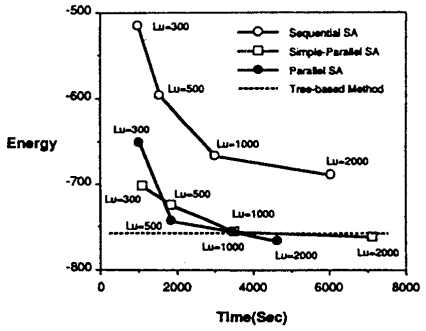

図2: 性能比較

6 おわりに

このシステムにより、乱数の影響を受け難い並列SAにおいて、高温における冗長な微小変形を削減することで、より短時間のうちに解を得ることが可能となっている。今後は、このシステムをより発展させることを検討中である。ひとつには、複数のアニーリングプロセスの中にはアニーリングを続けるうちにローカルミニマムに捕まるプロセスが現れてくるので、そのような解は捨てて、他の筋の良さそうな解をコピーしてアニーリングを続行することや、遺伝的アルゴリズム（Genetic Algorithm）風の色合いを加えることも検討中であり、更なるシステムの改善を計る予定である。

参考文献

[1] 金久：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.
[2] S.Kirkpatrick, C.D.Gelatt and M.P.Vecci : "Optimization by Simulated Annealing" Science, vol.220, no.4598, pp.671-681, 1983.
[3] 荒木、舘野、加藤、問藤：疎結合並列計算機上でのシミュレーテッドアニーリング情処研報, 91-AI-77-2, 1991, pp.7-14.
[4] 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タンパク質の配列解析 情報処理学会第43回全国大会 2-295 1991.
[5] 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 日本生物物理学会年会, 1990.
[6] Dayhoff, Hunt and Hurst-Calderone : "Composition of Protein" Atlas of Protein Sequence and Structure 5:3, Nat. Biomed. Res. Found., Washington, D.C., 1978, pp.363-373.
[7] Uchida : "Summary of the Parallel Inference Machine and its Basic Software" Proc. Int. Conf. on Fifth Generation Computer Systems, 1992, pp.33-49.