並列反復改善法によるタンパク質配列のアライメント

星田昌紀[1]，石川幹人[1]，広沢 誠[1]，戸谷智之[1]，十時 泰[2]

1: (財) 新世代コンピュータ技術開発機構　　2: (株) 情報数理研究所

1 概要

タンパク質配列のマルチプルアライメントの問題は、組合せ最適化問題と捉えることができ、実用的規模の問題では、大量の計算量を必要とする。生物学者が手作業でアライメントをする場合も、大変な労力が必要であり、高品質で高速な自動システムが望まれている。我々は、並列推論マシンを利用し、実用規模のアライメントを、現実的な時間内に実行するシステムを構築した。本システムは（ツリーベース）並列反復改善法を用いており、実用規模の問題においても、従来法より高品質なマルチプルアライメントを提供する。

2 はじめに

昨年、反復改善法 [1] という、新しい視点からのアライメント手法が提案された。この方法もやはり、要素技術に DP を用いているが、それを反復的に適用することにより、アライメントを徐々に改善していくというものである。我々は反復改善法の潜在的な能力に注目し、この方法をベースに新しい方式によるアライメントシステムを構築した（詳しくは [2] を参照されたい）。まず要素技術であるダイナミックプログラミング [3] について説明し、次に反復改善法についての解説を行う。さらに反復改善法の問題点について考察を行った後、我々の並列反復改善法について述べる。

ダイナミックプログラミング（DP）を適用するためには、図1のような2次元のネットワークの辺に、アライメントしたいタンパク質のアミノ酸配列を対応させる。斜め方向の矢には、その矢の位置に対応する2つのアミノ酸の類似度を割り振り、縦および横方向の矢には、ギャップを挿入するときのコストを割り振る。そうすると、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最良経路を求めることに対応する。各点に至る最良経路は、その後の経路に無関係に決定できるから、左上の端から右下の端に向かって、各点に至る最良経路を段階的に決定していけばよい。

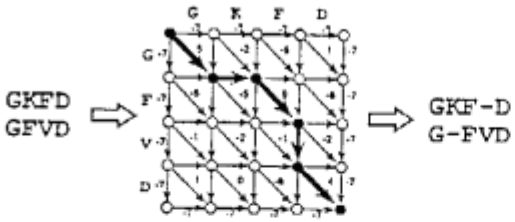

図1：2次元 DP による配列のアライメント

DP は、原理的には、そのまま多数の配列のアライメントに拡張できるが、n 本の配列を同時にアライメントする n 次元の DP は、概して、配列の n 乗の計算量と n 乗のメモリー量が必要であり、現実的に可能なのは3次元までである。そのためマルチプルアライメントは、通常、ペアワイズのアライメントを組み合わせて求めている。しかし、それでは精度が十分でない。

DP ではまた、アライメントが済んだ配列グループ間同士のアライメントに拡張することができる。2 のグループ間の2次元 DP を行うときは、それぞれの配列グループの中では、アミノ酸やギャップの縦方向の位置が変わらないように固定しておく。

3 反復改善法

反復改善法は、前述の配列グループ間の2次元 DP を反復的に適用することによりアライメントを徐々に改善する（図2）。まず、何らかの方法で初期状態となるアライメントを作成する。そして、これら N 本の配列を、ランダムに選んで2つのグループに分割する。この分割法は $2^{N-1}-1$ 通りある。分割された2つのグループ間に、2次元 DP を適用する。その結果は、それぞれ、必ずひとつ前の状態の得点より改善しているか、悪くても同じ得点である。改善が見られた場合は、改善された状態を、新しい初期状態として2次元 DP を適用する。この過程を1試行として、改善が行われる。この試行を繰り返すことにより、徐々にアライメントを改善していき、これを収束がみられるまで行なう。

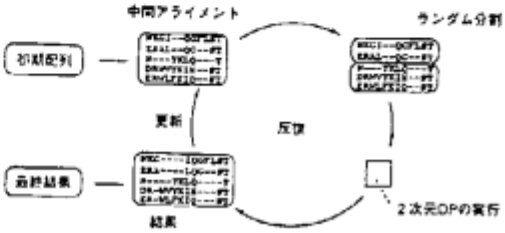

図2：反復改善法の手順

反復改善法は、強力な方法ではあるが、次の問題点も持っており、実用規模のマルチプルアライメントに適用するのが難しい。

- 問題点1：グループ間2次元 DP の再試行による速度低下
- 問題点2：配列の分割数の組合せ的爆発による速度低下
- 問題点3：アライメント結果の初期状態依存性

4 並列反復改善法

我々は、反復改善法に次の3点を拡張することにより反復改善法の問題点を克服し、実用規模の問題にも対処できるアライメントシステムを確立した。それぞれの拡張点について、順に説明する。

- 試行錯誤過程の並列化
- 配列グループを限定して分割する方法
- ツリーベース状に配列を組み合わせる方法

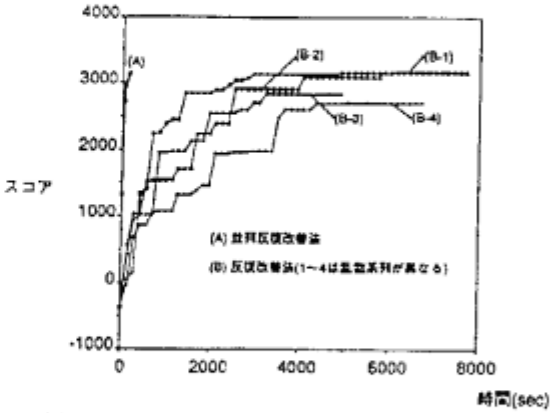

図3：並列反復改善法と反復改善法の比較（配列9本）

並列計算機を用いて、可能な分割を全て並列に実行することにより、再度試行を行う時間を節約し、反復改善法の全体の実行時間を短縮することができる [4]。このように並列化を行うことにより、

Protein Sequence Alignment by Parallel Iterative Method
Masaki HOSHIDA[1], Masato ISHIKAWA[1], Makoto HIROSAWA[1], Tomoyuki TOYA[1], Yasushi TOTOKI[2]
1:Institute for New Generation Computer Technology (ICOT)　2:Information and Mathematical Science Laboratory, Inc.

「問題点１」に対応できる。我々は、256台構成の並列計算機を用いて９本の配列の全ての分割方法255通りを同時に、グループ間の２次元DPをし、その結果で一番評価値のよいものを、次のサイクルの初期状態として採用するという方法を行った。この方法は探索問題における最急降下法に対応している。このような、並列化手法が拡張された反復改善法を「並列反復改善法」と呼ぶことにする。

８０文字のアミノ酸配列９本のアライメント問題を対象に、並列反復改善法を評価したところ、図３に示されるような高速化が達成された。グラフを見て分かるように、並列反復改善法(A)は非常に短時間の内に高得点に達しているのに対し、反復改善法(B)は、改善が行われない試行がかなりあるので、グラフが水平に近い部分が多い。同じ程度の得点に達する時間で比較すると、約60倍近い差が存在している。また、反復改善法は乱数を使った逐次改善法のため、その系列によっては、比較的悪いローカルミニマに陥るものも多くあることが分かる。(B-3) (B-4) はこのような例である。並列反復改善法では、探索空間内を一番良い方向を見極めながら探索するので、ある程度良い解へ安定して至る。

さて、我々の並列計算機で並列化を行っても、１０本以上の配列には一度には対応できないという問題は存在する。本当に実用化を考えた場合は、２０本くらいまで並列に実行可能なのが望ましい。我々は、「並列反復改善法」で実験を行った結果を分析するなかで、N本の配列がある場合、1本とN-1本、および、2本とN-2本という分割が主に改善に寄与し、N/2本とN/2本というような分割は、非常にまれにしか改善に寄与しないことを思い出した。このように、配列の本数に関して非常に偏った分割を行う手法を「限定分割法」と呼ぶことにする。限定分割法を使用すれば、プロセッサ256台構成の並列計算機で１０本以上の配列を扱うことが可能になる。この限定分割法を用いることにより、「問題点２」に対処できる。

限定分割法によって実用的な規模まで配列の本数を増やすことができたが、限定分割を行うことによる弊害が発生する場合も存在する。それは、複数の配列の中に、際だって類似性の高い配列グループが存在し、その配列グループが３本以上の配列からなる場合である。我々は、アライメントシステムMASCOTの開発設階において、類似性の高い配列同士はグループ分けすることと、各グループ毎の初期のアライメントを注意深く行うことが、アライメント全体の品質に大きく影響を与えるという知見を得ていた[5]。そこで、ツリーベースのアライメントの各段階において、並列反復改善を行ってアライメント状態を確実なものにしながら、配列を組み合わせていくという「ツリーベース並列反復改善法」の発想に至った。

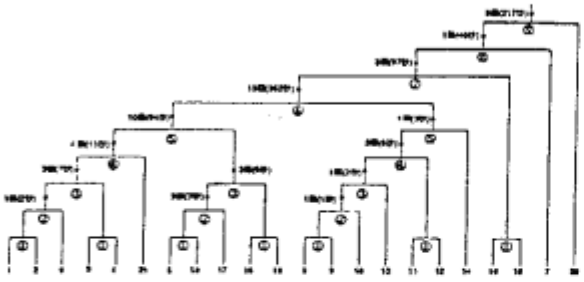

図４：ツリーベース並列反復改善法の手順

これまでの反復改善法、および並列反復改善法では、初期状態となるアライメントに関して考慮していない。ツリーベース並列反復改善法では、前もって配列群のうち、どの配列とどの配列が近い関係にあるかを調べてツリー（図４）を作り、そのツリーに従って、徐々に配列の本数を増やしながら反復改善を行っていく。つまり、反復改善法の「問題点３」に対処することができ、初期状態に依存しないアライメントを得ることできる。

ここで、並列反復改善法とツリーベース並列反復改善法との比較検討を行う。図５は、kinaseというグループのタンパク質（３０個）から、それぞれ類似性の比較的高い部分配列（８０文字分）を取り出した配列群（３０本）から、ランダムに選び出した２２本に対して、それぞれの実験を３０回行った結果を示している。並列反復改善法（PIA）と、ツリーベース並列反復改善法（TPIA）は、どちらも限定分割を用いており、１本限定と、1,2本限定の２種を、それぞれに対して試した。相対スコアとは、４つの異なる方法を行ったときの、スコアの平均値からの相対的な差を表している。図５の実験から次のことが分かった。

- TPIAは、PIAよりも総じて良い結果をもたらしている。
- PIAにおける1,2本限定（PE253台使用）のスコアと１本限定（PE22台使用）のスコアは、平均して大きな差異はない。
- TPIAにおける1,2本限定のスコアと１本限定のスコアは、安定して同じような値を与えており、ほとんどの場合、1,2本限定が若干良いスコアを与えている。
- TPIAは、PIAより３割ほど実行時間が短い。また、1,2本限定は、１本限定より、１割ほど実行時間が長い。

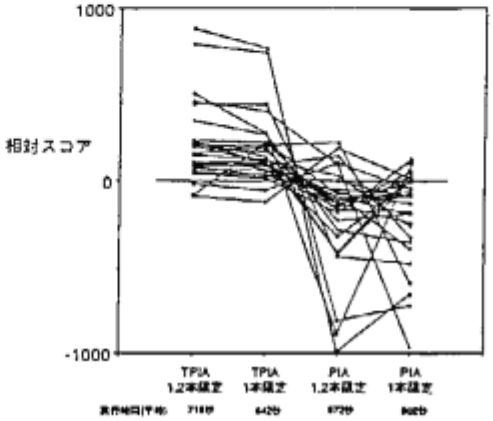

図５：PIAとTPIAの比較（限定分割あり、配列２２本）

図５には示していないが、典型的な従来法（ツリーベース法[6]）を用いて、同一の３０問を解く実験も行った。従来法は実行時間こそ平均408秒と、TPIAよりも４割ほど短いが、スコアは平均して、TPIAよりおよそ3000、PIAよりもおよそ2600悪く、３０問のなかでTPIA、PIAのいずれかより良いスコアを与えたものはひとつもなかった。

5　まとめ

我々は反復改善法に注目し、それをもとにして、実用規模の問題に高品質の解を与えるマルチプルアライメントシステムを構築した。そのために導入した要素技術は、並列化と、限定分割と、ツリーベース実行である。実験の結果、並列化と限定分割を導入したPIA（並列反復改善法）は、従来法より高品質のアライメントを与えることが判明した。さらにツリーベース実行も導入したTPIA（ツリーベース並列反復改善法）は、PIAよりも安定して高品質のアライメントを与えることもわかった。今回の実験は、比較的類似した配列部分を同じ長さに切って問題としたが、配列の長さが違う配列群をアライメントする場合には、類似した配列同士からアライメントしていくTPIAの方が、PIAより、ますます、きわだって良い結果を与えることが想像できる。今後はTPIAを、我々の標準システムとして評価改良を続けたい。

参考文献

[1] M.P. Berger and P.J. Munson : CABIOS,7, 1991, pp.479–484.

[2] 並田、石川、広沢、戸谷、十時：情報処理学会情報学基礎研究会 27-2, 1992.

[3] S.B. Needleman and C.D. Wunsch : J. Mol. Biol. 48, 1970, pp.443–453.

[4] Ishikawa, Hoshida, Hirosawa, Toya, Onizuka, Nitta : Proceedings of Fifth Generation Computer Systems '92, 1992, pp.294–299.

[5] 広沢、並田、石川、戸谷：情報学シンポジウム講演論文集, 日本学術会議, 1992, pp.77–85.

[6] J.G. Barton : Methods in Enzymology Volume 183, Academic Press, 1990, pp.403–428.