類似性を考慮した反復的マルチプルアライメント

7D−6

広沢 誠[1]、石川 幹人[1]、星田 昌紀[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

1 はじめに

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アライメントシステムが望まれている。

一昨年、反復改善法[Berger 91]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。この方法も従来法[Feng 87, Barton 90]と同様に、要素技術にダイナミックプログラミング[Needleman 70]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。しかしながら、反復改善法は、3、4本の配列のアライメントを作成する場合には、有効であるが多数の配列のアライメントを作る場合には計算量が大きくなるので実用的ではない。

我々は反復改善法を基に実用的なアライメントアルゴリズムを開発してきた。まず、我々は、並列反復改善法[星田 92]を開発し、さらに、これを基にツリーベース並列反復改善法[石川 92]を開発してきた。これらは、並列コンピュータ向きのアルゴリズムである。ツリーベース並列反復改善法は、それがベースとする並列反復改善法より品質の良いアライメントを作り出すが、計算量が大きいという問題点がある。我々は、ベースとする並列反復改善法を配列間の類似性を考慮し改良することにより、ツリーベース並列反復改善法の計算量を削減する方式を開発したので、発表する。

以下、まず、並列反復改善法、ツリーベース並列反復改善法の説明をしてから、新ツリーベース並列反復改善法の説明をする。最後に、以前のツリーベース並列反復改善法と新ツリーベース並列反復改善法の比較をする。

2 並列反復改善法

並列反復改善法[星田 92]は、配列グループ間に、ダイナミックプログラミング(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1)。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。

これらN本の配列を、何らかの基準によって2つのグループに分割し、分割された2つのグループ間に、2次元DPを適用し、アライメントを作ることができる。このアライメントの評価値は、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。

さて、この分割法は$2^{N-1}-1$通りあるので、可能な全ての分割方法に対してDPを行ない、一番良い評価値を持つアライメントを次のサイクルの初期状態するという過程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くことが可能である。評価値に収束がみられたら、その時点の状態を、最終アライメントとする。

しかしながら、$2^{N-1}-1$通りの分割方法を全て行なうと、配列の本数が多い時には、計算量が膨大になり現実的ではない。例えば、ICOTの256台の並列計算機で各分割に対するDPを並列に行なうと9本($2^{9-1}-1=255$)までの配列しか扱うことができないそこで、我々は、効果的な分割法をのみを採用することにした(限定分割法)。

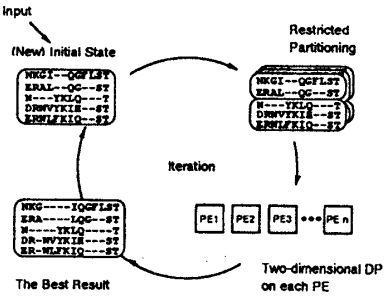

図1: 並列反復改善法

我々は、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は、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、あまり改善に寄与しないことを見い出した。例えば、1本とN-1本および2本とN-2本という分割法だけを行うとすると、22本の配列についてのグループ間DPが、一度に実行可能である(${}_{22}C_1+{}_{22}C_2=22+231=253$)。また、1本とN-1本というような分割法だけを行うとすると256本までの配列を扱える。以上の考察に基づき、我々は、前回は1本限定分割、1、2本限定分割を採用することになった。

3 ツリーベース並列反復改善法

ツリーベース並列反復改善法[石川 92]では、恣意的な初期状態を用いるのではなく、前もって配列群の中で、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方式である。このため、初期状態に依存しないアライメントが得られる。

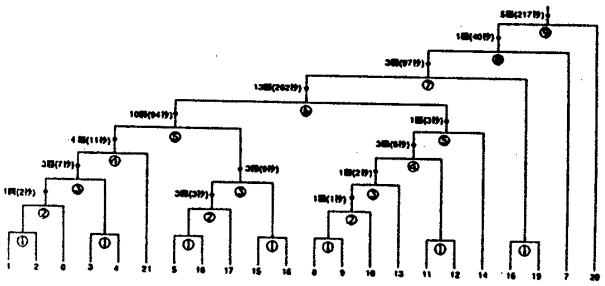

図2: ツリーベース並列反復改善法

この方法の手順を具体的に述べると、次のようになる。まず、配列の全てのペアについてDPを行い、そのスコアを求める。そして、そのスコアを配列間の類似性と考え、良く似た配列から順に結線したツリーを作成する。これが前処理であり、例えば図4に示すようなツリーが得られる。このツリー

Iterative Multiple alignment with similarity consideration
Makoto Hirosawa[1], Masato Ishikawa[1], Masaki Hoshida[2]
1:Institute for New Generation Computer Technology.　2: Matsushita Electric Industrial Co.,Ltd.

に従って、類似している配列から順に２次元DPを使い配列を組み合わせ、アライメントしていく。ただし、２次元DPを行った結果が配列３本以上のアライメントになるときは、そのあと収束するまで並列反復改善を行い、その時点のアライメントを確実なものとする。図２の黒い丸で示される点は、並列反復改善を行った個所を表している。それぞれ、1.2本限定で、何サイクル並列反復改善を行って、何秒かかったかを、「回、秒」で示している。

4 ツリーベース並列反復改善法の改良

我々は、1.2本限定のツリーベース並列反復改善法と１本限定のツリーベース並列反復改善法の比較検討を２５６個のプロセッサを持つ並列計算機を用いて行なった。実験は、kinaseというグループのタンパク質（３０種）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。プロセッサは分割数の数だけ用いた。つまり、1.2本限定では２５５個、１本限定では２２個のプロセッサを用いた。その結果、多くの場合において1.2本限定が１本限定より良い結果を出した。また、計算時間は双方とも同様であった（図３を参照）。

1.2本限定の問題点は分割数が配列の本数の２乗で増えるために、プロセッサの数が（分割数に比較して）少ない並列計算機、または、逐次計算機を用いて計算する場合には、計算時間が長くなることである。そこで、我々は分割数が１本限定と同程度であり、しかも、1.2本限定と同様に効果的な分割方法であるtree依存分割を考案した。

図２を流用してtree依存分割の説明をする。tree依存分割では、ツリーベース並列反復改善法において、アライメントに配列が加えられていくたびに系統樹を描き、類似性を考慮した分割方法を行なう。具体的には、系統樹のサブツリーを構成する配列と、そうでない配列に分割する。n本の配列には、$2n-3$本のサブツリーがあるので、$2n-3$個の分割法がある。図２におけるサブツリーは、例えば、{1},{2},{1,2},{0},{1,2,0},{3},{4},{3,4},{1,2,0,3,4},{21}{1,2,0,3,4,21}..... である。tree依存分割は、全ての１本限定分割を含んでる。また、配列の類似度が高い２つの配列に対する1.2本限定分割も含んでいる。

5 結果

図３はkinaseのデータに対するツリーベース並列反復改善法の結果を１本限定分割（２２個のプロセッサを使用）と、1.2本限定分割（２５５個のプロセッサを使用）とツリー依存分割（４１個のプロセッサを使用）の３方式について示した。相対スコアとは、３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。これから次のことが分かった。

- tree依存限定は１本限定に比較して、３つの場合を除いて良いスコアを出している。
- 1.2本限定は１本限定に比較して、４つの場合を除いて良いスコアを出している。
- tree依存限定と1.2本限定のスコアとスコアは、平均としては同様な値を与えている。1.2本限定のスコアの方のスコアが大きい場合の方が若干多い。ただし、１つの例において、1.2本限定のスコアが極端にわるい場合がある。
- 1.2本限定限定と１本限定の平均実行時間とはほとんど同じである。これに対して、tree依存限定の平均実行時間は他の２方式に比較して２割ほど長い。しかし、調査の結果サイクル数は他の２方式とほぼ同じであることがわかった。これは、1.2本限定分割を構成している不均等な分割のDPに較べて、tree依存限定に多少含まれている均等に近い分割（例えば、２２本を１０本と１２本の分割）においてのDPの時間が長いためであると思われる。２、３個均等に近い分割が含まれているためにtree依存限定のサイクルは長くなる（不均等分割に割り当てられたプロセッサは、自分の仕事が済んでも均等に近い分割に割り当てられたプロセッサの仕事が終了するまでアイドルとなる）。

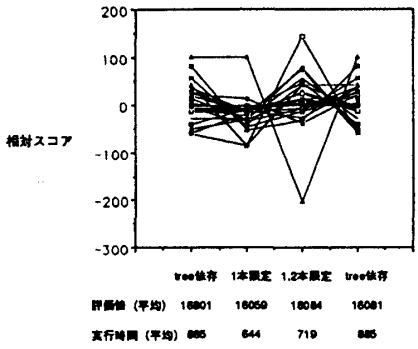

図３: ツリーベース反復改善法の比較

これらの結果を検討すると、まず、tree依存限定は１本限定に比較して優位であるということがわかる。また、tree依存限定は1.2本限定とほぼ同様のスコアを出すのにもかかわらず、必要とするプロセッサが少ないので、並列計算機を使う場合において、配列の数が多い時、または、使用できるプロセッサの数が少ない時には、有利である。特に、逐次計算機を用いる場合には、均等に近い分割のDPの終了を他のプロセッサが待つアイドル時間がなくなるために、tree依存限定は1.2本限定の速度比は、ほとんど分割数の比となる。したがって、分割比が大きくなる配列数が多い場合にはtree依存限定が優位である。

6 まとめ

我々は、以前発表したツリーベース並列反復改善法を、tree依存限定分割という方法をとることにより改良する方式を開発した。そして、この方式が以前の方式に対して優位であることを示した。

References

参考文献

[Berger 91] M.P. Berger and P.J. Munson: *CABIOS*, **7**, 1991, pp.479-484.

[星田 92] 星田、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.

[Needleman 70] S.B. Needleman and C.D. Wunsch: *J. Mol. Biol.* **48**, 1970, pp.443-453.

[Barton 90] J.G. Barton: *Methods in Enzymology Volume 183* Academic Press, 1990, pp.403-428.

[Feng 87] D. Feng and R.F. Doolittle: *J. Mol. Evol.*, **25**, 1987, pp.351-360.

[石川 92] 石川、星田、広沢、戸谷、十時: 第３回ゲノム情報ワークショップ講演論文集, 1992, pp.263-266.