遺伝的アルゴリズムを取り入れたタンパク質配列解析

2R−4

戸谷 智之、石川 幹人、星田昌紀*、小長谷明彦**

(財) 新世代コンピュータ技術開発機構, 松下電器産業 (株)*, NEC(株)**

1 はじめに

我々は、タンパク質配列の類似性を解析する手法のひとつであるマルチプルアライメントの問題の解決を目標に研究を行なってきた。昨年、並列反復改善法を用いたマルチプルアライメントシステムを開発し、従来以上に、高品質なアライメント結果を短時間のうちに獲得できることを可能にした。今回は、さらなる解の高品質化を目指し、並列反復改善法に遺伝的アルゴリズムをとりいれ、並列反復改善法の発展を試みた。それについて、報告を行なう。

2 並列反復改善法

並列反復改善法 [1] は、Berger-Munson の反復改善法のアイデアをもとにしている。これは、従来から行なわれてきた手法と比較して、高品質の解を得られる有効な方法であるが、N 本の配列群に対して、2^N-1 通りのグループ分割が考えれ、実用規模 (数十本) のアライメントを得るには莫大な時間を要してしまい、現実的とは言えなかった。我々は、それを並列化し、さらに独自のヒューリスティクスを導入することで、実用規模の問題を短時間のうちに解き、高品質のアライメントを獲得できるシステムに発展させた。(図 1)

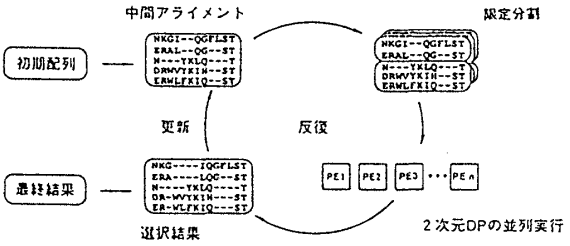

図 1: 並列反復改善法

(1) アライメントする配列群を 2 つのグループに分割する複数の組合せを生成する。
(2) それぞれの分割組合せで、分割された 2 つのグループ間でのグループ間 DP を行なうが、その処理を複数のプロセッサで並列に実行させる。
(3) 各グループ間 DP で得られた結果、アライメントの評価値として最も良いものを次のサイクルの配列群として、(1) の処理に戻る。

さらに、(1) の処理で、ヒューリスティクスを導入して、分割の場合の数を制限することで、計算時間の削減に成功した。それは、配列群を 2 つのグループに分割する際に、偏りをつけた分割の方が改善に効果があることが実験により確かめられたからであった。

このようにして、我々の並列反復改善法は、Berger-Munson の手法をより実用的な規模の問題で、質の高い解を短時間のうちに解決できるように発展させたのであった。ただ、実験を繰り返していくうちに、問題によって、比較的質の良くないアライメントが得られることがあった。解空間上で、局所解につかまってしまったと考えられるが、我々は、その解決に遺伝的アルゴリズムを取り入れることを考えた。

3 遺伝的アルゴリズムとそれを取り入れた並列反復改善法

遺伝的アルゴリズム (Genetic Algorithm 以下、GA) [2] は、生物の進化過程を発想のもとにした、組合せ問題の解決するための汎用の探索アルゴリズムである。 GA は、適応度の高い個体が生き残っていくように設計されたものであり、解空間上で複数の解 (個体) の集団を操作の対象としている。GA では、"mutation"、"crossover"、"selection" の 3 つのオペレーションが用いられる。図 2 のように、これらのオペレーションを組み合わせて行なわれるのが GA である。我々は、この GA の発想をもとに、並列反復改善法を拡張することを考えた。従来の反復改善法では、

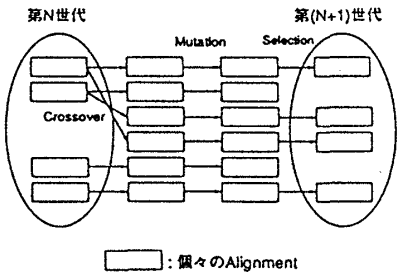

図 2: 遺伝的アルゴリズム

Parallel Iterative Aligner with Genetic Algorithm
Tomoyuki Toya, Masato Ishikawa, Masaki Hoshida*, Akihiko Konagaya**
ICOT, Matushita Electric Industrial Co.*, NEC Co.**

並列に実行を行なったグループ間DPの結果で、最も良い解だけを選ぶことを行なってきた。つまり、山登り法を行なってきたと言える。しかし、それでは局所解につかまることもあるので、それをGA的に発展させることで回避しようというものである。GAの適応度はアライメントの評価値に相当する。また、アライメントの解そのものを個体とした。我々は、GAにおける3つのオペレーションに相当するものを並列反復改善法では以下のように定義した。

mutation 個体の1部分を変化させるという意味で、解（配列群）の1本を抜きだし、その1本と抜かれた残りの配列群の間でDPによるアライメントを行なうこととした。GAの本来のmutationが持つ意味とは異なって、各個体を改善させるオペレーションとなっている。

crossover crossoverは、2つの個体間での解の混合と解釈できる。具体的には図3に示すようなオペレーションを行なう。選ばれた2つの個体（配列群）で、他の個体と“交換する配列”と“交換しない配列”をやはりランダムに決定する。そうして決定された“交換する配列”群を個体間で交換し、その後グループ間でDPを用いてアライメントする。

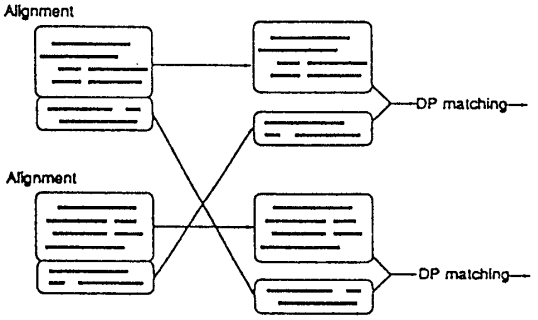

図 3: crossover

selection selectionは、今ある個体群から、指定した割合で、適応度の高い個体を残し、適応度の低いものは捨てる操作を行う。捨てられた分の個体数だけ、生き残った個体間でcrossoverを実行し、全体の個体数は一定を保つようにしている。

4 GA反復改善法の評価

我々は、並列論理型言語 KL1 を用いて、GA反復改善法のアライメントシステムを構築し、並列計算機PIM上で並列実行させた。

図4は、従来の並列反復改善法の結果とGA反復改善法の結果を比較するグラフである。問題は、いずれも80個のアミノ酸からなる22本の配列群であり、実用的な規模の問題である。

いずれも、256プロセッサをもつPIM上で実行した。

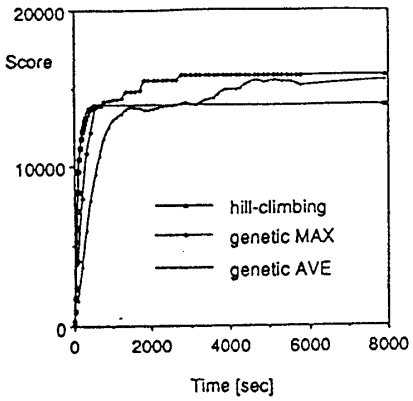

図 4: 実験結果

並列反復改善法では、ヒューリスティクスとして、分割する際に、“1本と残り”、もしくは“2本と残り”になる分割のみに制限する手法を用いている。この場合、22本の配列があるため、$253(_{22}C_1 +_{22} C_2)$の分割ができるため、マスタープロセッサ1つを加えた、254プロセッサでの並列実行ということになる。25サイクル約11分で改善が見られなくなった時点で実行は終了し、得られた解の評価値は13903であった。

GA反復改善法では、256プロセッサ使用可能なので、全体の個体数を255として実行した。各世代の時間は80秒とし次の世代に生き残る個体を上位90%とした。MAXとあるのは、各世代で最も高い評価値をプロットしたものである。AVEとあるのは、各世代での全個体の平均評価値をプロットしたものである。MAXを見ると、山登り法の場合とほぼ同じぐらいの時間で同レベルの解に達していることが分かるであろう。GAはさらに実行が続き、MAXの評価値は15775、AVEも15121にまで達した。

5 まとめ

今回は、GAの発想をもとに並列反復改善法の発展させたマルチプルアライメントのシステムを実装し、より質の高いアライメント結果が得られたことを報告した。今後は、個体数、生存率、世代時間などの、パラメータと性能についての相関について、実験を行なっていきたいと考えている。

参考文献

[1] 星田昌紀, 石川幹人, 広沢誠, 戸谷智之, 十時泰：“並列反復改善法によるタンパク質配列のアライメント”, 情報処理学会第27回情報学基礎研究会, 13-24, 1992.

[2] Goldgerg,D.E.: “Genetic Algorithms in Search, Optimization, and Machine Learning”, Addison-Wesley Publishing Company,Inc., 1989.