マルチプルアラインメントに対する遺伝的アルゴリズムの検討

河本　敬子† 筧田　祐一郎‡ 水野　陽介‡ 一野　天利† 谷澤　一雄† 堀部　和雄†
近畿大学　生物理工学部　知能システム工学科† 学生‡

1. はじめに

配列アラインメントは、異なる遺伝子やタンパク質をデータベースから類似の配列を検索するために用いられ、進化類縁関係や機能を推測するときの基礎となる。塩基配列またはアミノ酸がアラインできた場合、両者は類似した遺伝子やタンパク質であると一般にいえ、既知の構造あるいは機能を持つという推測が成り立つ。この情報は病気に対抗する新薬の設計に繋がる。

本研究では、マルチプルアラインメントに対する遺伝的アルゴリズムを用いた解法を検討することを目的としている。

2. マルチプルアラインメント

マルチプルアラインメントでは、3本もしくはそれ以上の文字列を各文字ごとに照合し、一致数が最大となるように整列させる。

図1は3本の配列時のマッチングとギャップの挿入例を示す。一致数をより大きくするために文字列間にギャップ（" -" ハイフンで表す）を挿入させる。文字列が一致したときやギャップを挿入したときに与える点数の合計をアラインメントスコアと呼び、最大となるアラインメントスコアが導かれるような最適なアラインメントを求める[1]。また、本研究では文字間に挿入するギャップはランダムに入れる。今回は以下のような前提と定義を用いた。

・対象は塩基配列とする。
・スコア設定は以下のようにする。
　マッチ時 +2点　ミスマッチ時 −2点
　ギャップ一致 0点　ギャップ挿入 −1点

```
ABCDEFG          ABCDEFG
ABEG      ⇒      AB--E-G
BCFGH            -BC--FGH
                  ★  ★
```

- ：ギャップ文字　★：マッチポイント

図1　アラインメントの例

3. 遺伝的アルゴリズム

遺伝的アルゴリズム（Genetic Algorithm：GA）とは生物進化（選択淘汰・突然変異）の原理に着想を得たアルゴリズムであり、最適化の一手法である。歴史的にみるとGAはHollandのAdaptation in Natural and Artificial Systems (1975) において導入された手法である[2]。一般的なGAの処理手順は以下のようになる。このような一連の処理の繰り返しによって、個体集団は全体として適応度の高い個体の集団へと収束する。

Step 1. 個体の集団である個体集団を初期化する。
Step 2. 個体集団の各個体を目的関数に従って評価し、適応度を求める。
Step 3. 各個体の適応度の低い個体を個体集団から取り除き、逆に適応度の高い個体をその高さに応じて増やす。これを淘汰と呼ぶ。但し、個体集団のサイズを変えない。
Step 4. 個体集団の各個体をランダムに2つずつペアにし、このペアに突然変異、交叉を施して新しい個体を作る。
Step 5. Step 2〜4が繰り返しの単位であり、世代と呼ぶ。現在の世代の処理が終わると次の世代の処理が終わると次の世代の処理に移るため、Step 2へ戻る。

本研究では、配列4つを1つの個体として表す。交叉方法は、1点交叉および2点交叉を用いた。図2は、2点交叉の例である。2個体（両親）に対して2箇所の交叉点をランダムに指定し、それらの箇所で両親の遺伝子を交叉させ、

Examination of Genetic Algorithm for Multiple Alignment Problem
Keiko Kohmoto†, Yuichiro Toita‡, Yosuke Mizuno‡, Takatoshi Ichino†, Kazuo Tanizawa†, Kazuo Horibe†
† Department of Intelligent Systems, School of Biology-Oriented Science and Technology, Kinki University
‡ Undergraduate Student, Kinki University

```
        親1                          子1
配列1 TA-|ACA|G-T              配列1 TA-|A-CA|G-T
配列2 -AT|CT-A|GC      ⇒       配列2 -AT|CT-A|GC
配列3 AT-|GT-A|G-A             配列3 AT-|G-TA|G-A
配列4 -GC|AAA-|G-C             配列4 -GC|A-AA|G-C

        親2                          子2
配列1 TA|A-CA|-GT              配列1 TA|ACA|-GT
配列2 AT|CT-A|GCG      ⇒       配列2 AT|CT-A|GCG
配列3 AT-|G-TA|--G             配列3 AT-|GT-A|--G
配列4 GC-|A-AA|-GC             配列4 GC-|AAA-|-GC
```

図2　2点交叉の例

子を生成する。交叉によって生成された子はエリート選択によって、適応度の高い個体を次世代に残し適応度の低い個体を淘汰させる。

4. 数値実験

本実験では、配列アラインメントに対する遺伝的アルゴリズムを用いた解法について検討するために、個体および個体の文字数、交叉法の違いによるスコアの増加傾向を調べた。試行回数は100、個体数は4、世代数は100とした。突然変異は行わず、選択方法はエリート選択とした。個体は、米国公的機関のNCBI[3]が運営するGen Bank上で公開されている遺伝子配列から引用した。遺伝子配列が似ているウチワサボテン属の植物（Opuntia）の4配列で構成される個体（文字数1000）、遺伝子配列が大きく異なると思われる菊、バラ、ユリ、チューリップの4配列で構成される個体（文字数660および2500）の3種類とした。スコアは配列1と2、1と3、1と4、2と3、2と4、3と4を比較し、それぞれの点数を合計したものを示している。

表1は文字数nおよび交叉方法の違いによる実験結果を示す。bestは100回の試行回数で得られた最良解、averageは解の平均値、best_gene_aveは最良解が得られたときの平均世代を示す。nの値に関わらず、収束するまでにはほぼ同じ世代数を必要とすることがいえる。

図3は文字数660の個体と文字数1000の個体におけるアラインメントスコアのaverageの違いを示す。交叉方法は1点交叉を用いた。図3から、文字数の多い個体の方がアラインメントスコアが高いことが分かる。これは、この個体の遺伝子配列が似ているためと考えられる。

図4は1点交叉と2点交叉におけるアラインメントスコアのaverageの違いを示す。文字数が600の個体を用いた。結果から、2点交叉の方がスコアが高いことが分かった。

表1　文字数・交叉方法別での実験結果

n	交叉法	best	average	best_gene_ave
660	1	-2913	-2958	2.95
	2	-2909	-2949	3.76
1000	1	-1990	-2655	2.40
	2	-2174	-2700	3.32
2500	1	-11437	-11452	2.62
	2	-11419	-11552	3.70

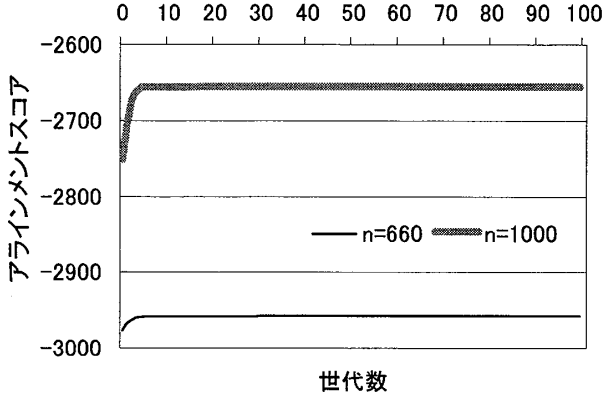

図3　文字数の違いによるスコア分布

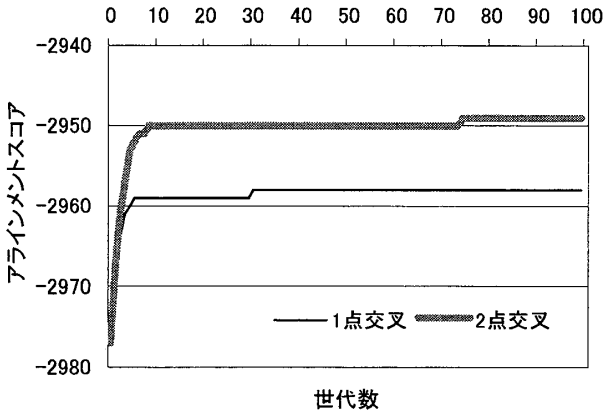

図4　交叉方法の違いによるスコア分布

5. 終わりに

本研究では、マルチプルアラインメントに対する遺伝的アルゴリズムを用いた解法について検討した。今後の課題としては、突然変異の検討、個体数を増やしてのアラインメント、様々な塩基配列やタンパク質のアラインメントを行う予定である。

参考文献

[1] 吉田孝廣，牧之内顕文，“類似配列検索のための配列アラインメントアルゴリズムの高速化”，電子情報通信学会，vol.104, no.176, pp.55-59, 2004.

[2] 坂和正敏，田中雅博，“遺伝的アルゴリズム”，朝倉書店，1995.

[3] NCBI: http://www.ncbi.nlm.nih.gov/