遺伝的アルゴリズムによるDNAのスプライス部位パターンの抽出

1E－9

謝 孟春　西野 順二　小高 知宏　小倉 久和

福井大学

1　はじめに

近年、遺伝的アルゴリズムが注目され、種々の応用研究が進められている。遺伝的アルゴリズムは生物進化における遺伝子の役割を模倣して、基本的な遺伝的操作 (交叉、突然変異、選択) を繰り返すことによって学習する。本研究では、GA の学習機能を利用して、スプライス部位のパターンの特徴を抽出することを試みた。まず、二つのコーディング方法を提案し、遺伝子の評価方法を設定した。それらに基づいて、遺伝的操作を設計した。異なる遺伝的操作がスプライシング部位の推定にどのような影響を考えるのかを検討した。また、遺伝子コーディング方法の違いによるスプライシング部位の推定能力の違いを比較した。

2　スプライス部位パターンの表現

2.1　データベースからのスプライス部位パターンの抽出

EMBL データベース配列中の GT(エクソン-イントロン接合部) あるいは AG(イントロン-エクソン接合部) を中心に、イントロン側 30 塩基、エクソン側 20 塩基を切り取りデータセットとする。エクソン側よりもイントロン側を大きくしたのは、エクソン側の配列には遺伝子のコーディングという制約があり、スプライシング情報はイントロン側に偏っていると考えられるからである。切り出された配列がスプライスに対応している場合は 1、対応していない場合は 0 のフラグを付加する。

GT あるいは AG パターンを抽出する時、スプライス部位に対応するかどうかによって、正例のデータと負例のデータに分ける。正例のデータはスプライシングの生じる配列であり、負例のデータはスプライシングの生じない配列である。すべての正例のデータと負例のデータをそれぞれほぼ半分に分けて、学習データセットと検査データセットを構成した。負例のデータは正例のデータ数の約 3 倍用意した。AG 部、GT 部それぞれ正例約 700、負例約 2200 である。

2.2　集団遺伝子のコーディング方法

本研究では、DNA の四つの塩基 A,G,C,T を含む配列を集団遺伝子とする。さらに、ドントケア記号を含めて、二つのコーディング方法を用意する。

コーディング方法 1 は、一種類のドントケア (D) を用いる方法である。コーディング方法 2 は、五種類のドントケア (D,E,F,H,Q) を用いる方法である。ドントケアとマッチするものの対応関係を次に示す。

コード	マッチするもの
D	A,C,G,T
E	T,C
F	A,G
H	T,A
Q	T,C,A

3　GA によるスプライス部位の抽出

3.1　遺伝的操作

各集団遺伝子に対して、GA で進化をさせるために、次のような遺伝的操作を適用する。

(1) 交叉

一点交叉を用いる。各遺伝子の適応度に比例して確率的に親遺伝子を選ぶ。同じ親遺伝子を選ぶことを許す。

(2) 突然変異

二つの突然変異方法を設定した。一つは逆位式で、もう一つは反転式である。

逆位突然変異はある確率で、ランダムに二つの遺伝子座を選び、その間の遺伝子すべてを反転して反対に並べかえることである。

反転突然変異はランダムに選んだ遺伝子座の値を対立遺伝子と入れ替えることである。

(3) 選択

各遺伝子の適応度によって、ルーレット戦略で子遺伝子を次世代に残す。適応度の高い遺伝子を次世

Extracting DNA Splice Sites By Genetic Algorithms
Mengchun Xie,Jyunji Nishino,Tomohiro Odaka,
Hisakazu Ogura
Fukui University

代に残すために、遺伝子集団には重複遺伝子を許す。また、集団中で最も適応度の高い遺伝子を一つそのまま次世代に残すというエリート保存戦略も用いる。

3.2 実験の結果

ここでは、コンピュータ・シミュレーションによって、GA でDNA のスプライス部位を抽出した結果を分析する。

シミュレーションの設定は次のようである。

- 遺伝子の長さ= 50
- 集団遺伝子のサイズ= 50
- 打ち切り世代数= 200
- 突然変異率= 0.1

(1) 異なる突然変異の結果

設計した二種類の突然変異方法でスプライス部位を抽出したときの、各世代のエリート個体の認識率の結果を図 1 と図 2 に示す。図 1 は逆位変異方法で、図 2 は反転変異方法である。閾値は 0.45 とした。図の横軸は世代数、縦軸は TP(実線)、TN (点線) である。TP はスプライス部位を正しく認識した割合を表し、TN は非スプライス部位を正しく認識した割合を表す。図の破線は進化におけるエリート個体の適応度の変化の様子を示している。

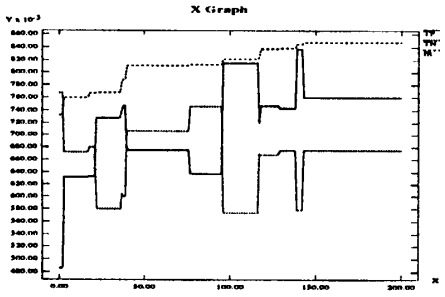

図 1: 逆位突然変異

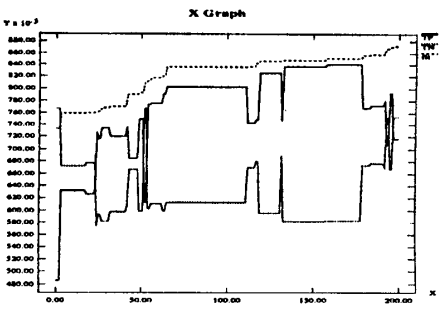

図 2: 反転突然変異

二つの突然変異の結果から見ると、エリート個体の適応度は進化とともに高くなるが、TP と TN の値は激しく上下している。特に、逆位変異より反転変異での収束はかなり遅い。それは、反転突然変異の場合には、集団遺伝子の全体の構成要素が変わるためと思われる。つまり、反転突然変異によって、親遺伝子に含まれない遺伝子の要素が現われ、近傍探索がうまくできる。

(2) コーディング方法 2 での結果

以上のシミュレーションはすべてコーディング方法 1 を用いた。ここでは、コーディング方法 2 を用いて、閾値を 0.6 とし、反転突然変異での TP、TN 及びエリート個体の適応度を図 3 に示す。

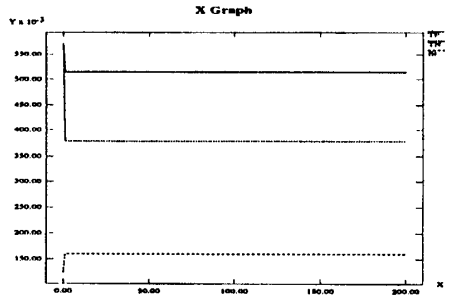

図 3: コーディング方法 2

コーディング方法 2 は遺伝子個体の成分が多いため、遺伝的操作を行う際に、評価関数に敏感ではなく、進化しないことがわかった。

4 おわりに

本研究では、一次元配列を遺伝子として、GA による DNA のスプライス部位パターンの抽出を試みた。適当な遺伝子コーディング方法と遺伝的操作によって、DNA のスプライス部位を抽出することができた。しかし、この方法では DNA のスプライス部位の認識率があまり高くならないことが結果より分かった。ドントケアを用いた一次元 DNA 配列の記述には限界があるため、より自由度の高い DNA 配列記述法が必要であると考えられる。現在、我々は、正規表現による DNA 配列記述を試みている。これによって、DNA の多次元的な構造を反映した表現が可能になると思われる。

参考文献

[1] 謝 孟春、小高 知宏、小倉 久和:遺伝的アルゴリズムを用いた DNA のスプライス部位の推定, 情報処理学会第 52 回全国大会講演論文集 (2), pp.57-58(1995.9)