3M－7

DNAスプライス部位のGAによる推定と遺伝子情報の分析に基づく遺伝的操作の改善

謝 孟春　西野 順二　小高 知宏　小倉 久和

福井大学

1　はじめに

DANの塩基配列およびタンパク質のアミノ酸配列には、さまざまな情報が潜んでいる。それらは、タンパク質の種類に関する情報であったり、進化に関する情報であったりする。われわれは、コンピュータを用いてDNAデータの解析を行うという遺伝子情報処理の手法として、遺伝的アルゴリズムの適用を試みた[1],[2]。しかし、単純なGAによるDNAスプライス部位の学習はあまり効率が良くなかった。認識率を高めるために、本研究では新たな配列の表現の方法を提案し、シミュレーションによって異なる表現方法の認識学習を比較した。また、GAにより得られたエリートを用いて、学習データについて分析した。

2　スプライス部位の表現と推定

2.1　データベースから実験データの作成

DNAスプライス部位を推定するための入力データとして、EMBLデータベース中の哺乳類のDNA配列を用いた。手法の有効性を検討するために、EMBLデータベース配列から学習データセットと検査データセットを作成した。前者はスプライス部位の学習モデルを生成するために使用し、後者は、生成された学習モデルの妥当性を判断するために使用する。学習データと検査データとの間に重複はない。

エクソンとイントロンの接合部位には一定の規則性、つまりGT-AG則がある。正例データを生成するとき、データベース配列中のGTあるいはAG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。データベースにおいてスプライシングの生じていないAGあるいはGT配列の部分を上の正例データと同様の形に、無所為に切り出し、負例データセットとする。もともとの意図は正の配列のならんかの規則を学習させるものであるから、あまり規則性のない負例の配列は学習しにくいと考えられるので、負例データ数は正例データ数の約3倍用意した。

Prediction of DNA Splice Sites by Genetic Algorithms and Improvement of Genetic Operations Using an Analysis of Gene Information
Mengchun Xie,Junji Nishino,Tomohiro Odaka,
Hisakazu Ogura
Fukui University

2.2　スプライス部位パターンの表現

本研究では、GAで進化させる遺伝子をGA遺伝子ということにする。次のようなGA遺伝子により、DNA配列のパターンを表現する。GA遺伝子は、A,T,C,Gといくつかのドントケア記号からなる記号列である。A,T,C,Gはそれぞれ DNA塩基配列における4種の塩基A,T,C,Gとマッチする。ドントケア記号は塩基A,T,C,Gのいくつかにマッチすることを意味する。ここでは、以前提案した二つの表現法[2]に加えて11種類のドントケア記号を使うGA遺伝子の表現法を検討する。この表現法における各記号の意味を表1に示す。

表1: 11種のドントケアを用いる表現法におけるドントケア記号の意味

コード	1	2	3	4	5	6
マッチする記号	A G C T	A G C	A G T	A C T	G C T	A G

コード	7	8	9	10	11	
マッチする記号	A C	A T	G C	G T	C T	

3　シミュレーションと分析

11種のドントケア記号の表現法に対して、GA学習システムを用いて、学習認識のシミュレーションを行った。以前提案した表現法と併せて、シミュレーションの結果を表2にまとめた。

結果から見ると、三つの表現法のいずれを用いても、80%前後の的中率でスプライス部位が推定できた。特に今回提案した11種のドントケア記号の認識率はかなり高くなったと思われる。11種のドントケア記号の表現法で得られたエリートGA遺伝子は

表 2: 異なる表現法での結果

	1 種の記号	5 種の記号	11 種の記号
S_0	0.6	0.8	0.8
TP	0.779	0.803	0.839
TN	0.810	0.694	0.832
TP'	0.817	0.788	0.874
TN'	0.804	0.722	0.823
f_{max}	1.620	1.510	1.697

87.4%(644/736) の正例と 82.3%(1944/2362) の負例を正しく認識した。

エリート GA 遺伝子を用いて正例の学習データと負例の学習データについて分析するために、例として、1 種のドントケア記号でのエリート GA 遺伝子を使った。このエリート GA 遺伝子とドントケア記号で解釈したスプライス部位配列の様子を図 1 に示す。

```
#C#TC###CT#TCCTTTTTTTTCCTTCCAGGT####ACG#G##C##A###

A A  AAA  A                    AAAA   A AA AA AAA
G G  GGG  G                    GGGG   G GG GG GGG
C C  CCC  C                    CCCC   C CC CC CCC
TCTTCTTTCTTTCCTTTTTTTTCCTTCCAGGTTTTTACGTGTTCTTATTT
```

図 1: 表現法 1 でのエリート (TP=81.7%,TN=80.4%)

まず、正例の学習データについて分析する。正例の学習データに対して、エリート GA 遺伝子を用いて正しく判別できたグループと正しく判別できなかったグループに分ける。二つのグループにおける各遺伝子座に現われた塩基の割合を図 2 に示している。実線はエリート GA 遺伝子を用いて正しく判別できたグループの各塩基の頻度で、破線はエリート GA 遺伝子を用いて判別を誤ったグループ各塩基の頻度である。点線はエリート GA 遺伝子がその遺伝子座に対して、その塩基にマッチするかどうかを表わす。マッチできれば 1 で、できなければ 0 である。

塩基 A,G の場合は、スプライス部位の左側には、正しく判別できたグループの A,G の割合が誤って判別したグループより低い傾向が見らるが、エクソン側には、逆の傾向が見らる。塩基 T の場合、イントロン側には、正しく判別できたグループの T の割合はかなり高いこともわかった。

同様に、負例の学習データについて分析すると、図 3 のような結果になる。塩基 A,G の場合は、正しく否定できたグループの A,G の割合がかなり高い。C,G の場合は、誤って否定したグループの C,G の割合が高いという傾向も見られる。

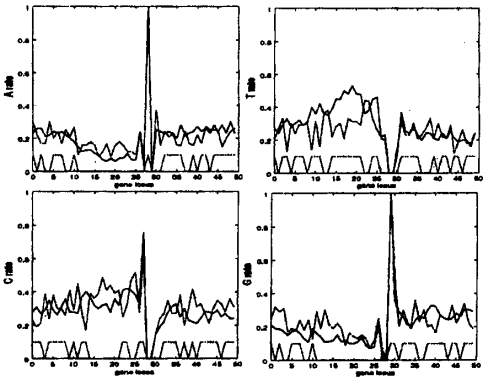

図 2: 正例データに対する塩基の頻度

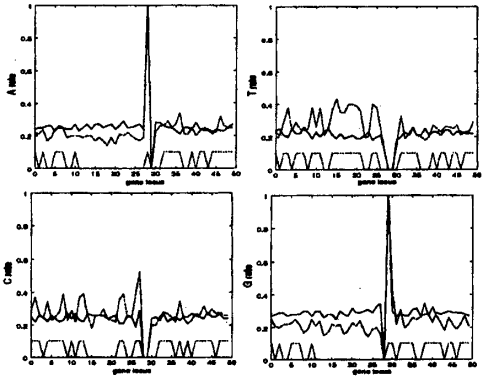

図 3: 負例データに対する塩基の頻度

4 おわりに

本研究では、GA を用いて DNA スプライス部位を推定する学習方法を試みた。きわめて簡単なパターン表現手法であるが、GA によってかなり高い認識率を得ることができ、スプライシング部位の周辺の構造を抽出できることも示された。DNA スプライシング配列の表現法は認識率にかなり影響を与えることがわかった。今後の課題として、より柔軟なパターン表現の可能な正規表現を利用して、スプライシング規則を表す方法を検討する予定である。

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 51 回全国大会講演論文集 (2), pp.57-58(1995.9)

[2] 謝 孟春、小高 知宏、小倉 久和: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, 情報処理学会第 52 回全国大会講演論文集 (2), pp.125-126(1996.3)