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

5K－5

謝 孟春　小高 知宏　小倉 久和

福井大学

1　はじめに

われわれ人間をはじめとして、地球上のほとんどの生物はDNAを持っている。DNA遺伝子はm-RNA前駆体への転写とスプライシングによる成熟m-RNAの生成、およびm-RNAからタンパク質への翻訳の2階段で発現する。DNAの塩基配列は、タンパク質のアミノ酸の配列順序に対応している。

遺伝子はイントロンによって分断されたエクソンとしてDNA中に存在する。イントロンはタンパク質に翻訳されないDNA領域で介在塩基配列と呼ばれ、エクソンが翻訳されるDNA領域で情報塩基配列と呼ばれる。m-RNA前駆体の配列から、イントロンを切り出す現象をスプライシングという。スプライシングは生物の進化や遺伝情報の発現にとって重要な意味を持つと考えられているが、その意義についてはまだ不明なことが多い。

DNAの情報は4種類の塩基(A:アデニン、T:チミン、C: シトシン、G:グアニン)の配列によって決定される。スプライシングによって切り出されるイントロンの両端の配列には、ほとんどの場合先頭がGT、後尾がAGのGT-AG則が見い出されている。しかし、この規則は必要条件であって十分条件ではないから、すべてのGT-AG対によってイントロンが切り出されるわけではない。遺伝子の塩基配列において、GTあるいはAGを含む部分配列があれば、スプライシングが生じるかどうかを判別するのが本研究の目的である。

本研究では、EMBLデータベースを対象として、遺伝的アルゴリズムを用いて、スプライス部位を推定する。まず遺伝子コーディングの方法及び各遺伝子の評価方法を検討する。それらに基づいて、遺伝的操作を設計する。GAで得られる結果からスプライス部位を推定する方法も検討する。

Evaluation of the Evolution of Building Block and the Role of Genetic Operators in GA
Mengchun Xie,Tomohiro Odaka,Hisakazu Ogura
Fukui University

2　GAによるスプライス部位の推定システム

本研究で作成したDNAのスプライス部位を同定するシステムの流れを図1に示す。

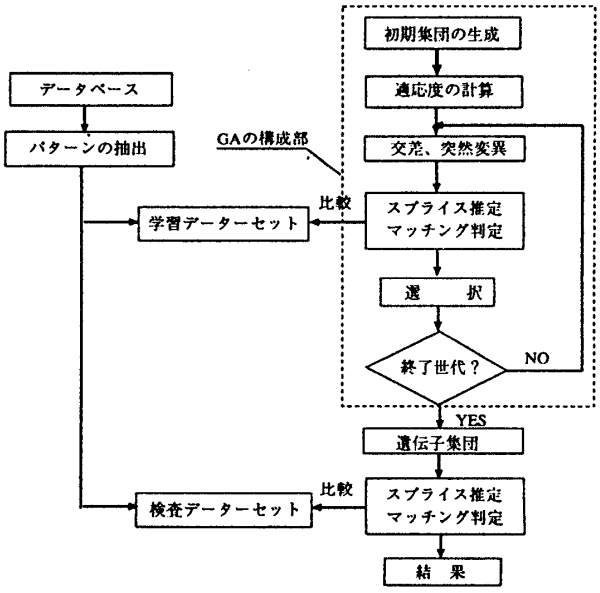

図 1: GAを用いたスプライスの推定

2.1　パターン抽出部

パターン抽出部は、データベースからGT,AGパターンを抽出し、学習データセットと検査データセットを作成する。EBMLデータベースからGTおよびAG部の前後、エクソン側20塩基、イントロン側30塩基を切り出す。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。その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。

本研究で用いたパターンデータセットは、次のような条件を満たしているものに限った。

エクソン、イントロンともに塩基配列が完全に解析されているもの。

エクソン側が20塩基以上あるもの。
イントロン側が30塩基以上あるもの。

2.2 GAの構成部

ランダムに初期遺伝子集団を発生し、基本3操作である交叉、突然変異、選択を繰り返すことにより評価関数の値が高い遺伝子集団を導き、進化させる。

2.3 スプライス推定

各パターンデータに対して、遺伝子個体のパターンとマッチさせることにより、マッチングスコアを求める。得られたマッチングスコアがあらかじめ定めた閾値より高いときはスプライス部位と判定し、低いときは非スプライス部位と判定する。この判定結果をスプライスのフラグと比較する。

学習データセットを用いて、GAにより弁別が十分にできるまで、繰り返し学習させる。塩基配列がGTの前後をとって学習させるとイントロンの先頭を推定できるようになり、AGの前後で学習させるとイントロンの後尾を推定できるようになる。先頭位置のスプライシングを学習した遺伝子集団と、後尾を学習した遺伝子集団がそれぞれ用意できる。

学習した遺伝子集団を用いて、検査データセットに対して弁別試験を行い、学習の効果を検討する。

3 遺伝的アルゴリズムの構成

3.1 DNA配列のコーディング方法

本研究では、GAにおける遺伝子個体のコーディングは、A,G,C,Tの4種類の記号と# (don't care)からなる塩基配列で表現する。#はA,G,C,Tのどれにもマッチする。遺伝子個体の長さは抽出したパターンの遺伝子の長さ(50塩基)である。

3.2 遺伝的操作の設計

DNA配列は、アミノ酸と対応している3個の連続塩基からなるコドンから構成されている。遺伝的交叉の方法としては、コドン配列を破壊しないコドン単位の交叉方法と、コドン配列を考慮しない方法とが考えられる。交叉戦略としては、ルーレット法により二つの親を選び、一点交叉、多点交叉、一様交叉などの方法で交叉させる。

突然変異は変更する位置をランダムに選びその位置の文字をランダムに自分自身と異る文字に書き換える方法と、部分配列の順序を逆転させる逆位の方法を用いる。

交叉と突然変異ののち選択を行い、ルーレット戦略あるいはランク戦略で次世代を構成する。世代交代ではエリート保存戦略を適用する。

3.3 マッチング・スコアの決定

マッチング・スコアSは、遺伝子長をM(=50)、マッチした遺伝子座数をiとして、次式で決める。

$$S = \frac{i}{M}$$

判別の閾値をS_0として、$S \geq S_0$のときスプライス部位と推定し、$S < S_0$のとき非スプライス部位と推定する。とりあえず$S_0 = 0.5$とするが、綜合的な優劣判断はROC解析により判断することができる。

3.4 評価方法の設定

遺伝子の評価は、学習データセットに対する正誤の回答数で行う。評価値は、正答に対してはそのマッチング・スコアを加え、誤答に対してはマッチング・スコアを減じる。評価値が負になった場合は、0とする。

$$\text{評価値} = \begin{cases} \sum(\text{正答のマッチング・スコア}) \\ \quad -\sum(\text{誤答のマッチング・スコア}) \\ 0 \quad (if \quad negative) \end{cases}$$

4 考察と課題

本研究では、遺伝的アルゴリズムを用いて、DNAのスプライス部位を推定する方法を検討した。この方法によって行ったコンピュータ・シミュレーション実験の結果については口頭で報告する。われわれは、以前、階層型ニューラルネットワークによって同様の試みを行った。その結果との比較も行う予定である。

ニューラルネットワークによる解析では、弁別のための配列規則がニューラルネットワークの構造として表現されているため、スプライシング規則が明示的には得られにくかった。GAでは記号表現による配列規則を用いてるので、スプライシング規則が明示的に得られると考えられる。

参考文献

[1] 小倉久和, 縣秀征, 古谷博史：階層型ニューラルネットワーク・シミュレーションによる遺伝子のスプライス部位学習, 福井大学情報センターニュース, Vol.7, No.1, pp.59-82(1993)