3ZH-1

特徴的塩基配列とステム構造に基づいた DNA からの snoRNA 遺伝子検出法

薛晨[1]　山森一人[2]　剣持直哉[3]　吉原郁夫[2]

宮崎大学工学研究科[1]　宮崎大学工学部[2]　宮崎大学フロンティア科学実験総合センター[3]

1. はじめに

タンパク質への翻訳情報を持たない非コード RNA と呼ばれる機能性 RNA の発見と役割の解析は、分子細胞生物学とバイオインフォマティクス双方において、現在最も重要な研究課題の１つである。基本的な代謝から個体発生、細胞分化までの様々な生命現象に関与する機能性 RNA が現在までに数多く見出されている[1]。機能性 RNA と疾患との関わりに関する研究成果も次々に報告されており[2]、創薬や再生医療分野などで大きな進展をもたらすことが期待されている。本稿では、機能性 RNA の１つである核小体低分子 RNA（snoRNA）をコンピュータによって自動検出する方法を提案する。

snoRNA は核小体内に存在し、リボソーム RNA 前駆体のスプライシングに関与する。snoRNA は図１に示したその構造から、Box C/D 型と Box H/ACA 型の２つに分類される[3]。２種類の snoRNA はそれぞれ独自の特徴的塩基配列を持つ。１つはボックスと呼ばれ、Box C/D 型 snoRNA は Box C（AUGAUGA）、Box D（CUGA）、Box C’（UGAUGA）と Box D’（CUGA）を持つ。Box H/ACA 型 snoRNA は Box H（ANANNA、N は任意塩基）と Box ACA（ACA）を持つ。もう一つは相補対（A-U と C-G）により形成されるステムと呼ばれる構造である。

提案手法では、特徴的塩基配列とステムの存在確率を評価することで snoRNA 遺伝子の自動検出を行う。

2. 提案手法

2.1. 提案手法の概要

snoRNA 遺伝子検出にあたり、まず対象とする塩基配列にボックスが含まれる確率を調べる。次に、ボックスが一定以上の確率で含まれる塩基配列のしかるべき位置に、相補対により形成されるステム構造が存在する確率を計算する。２つの確率の平均値がしきい値を超える場合、対象とする塩基配列に snoRNA 遺伝子が含まれると判断する。

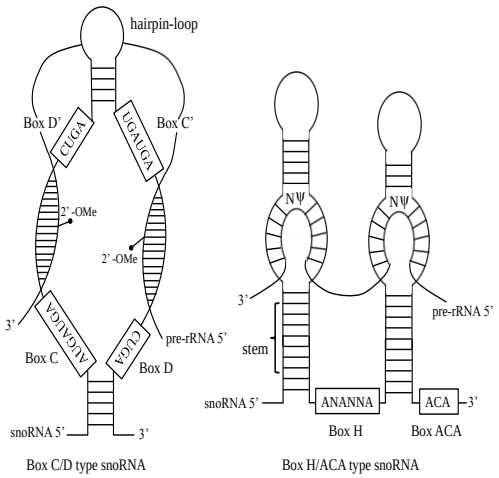

図 1. Box C/D 型 snoRNA と Box H/ACA 型 snoRNA の構造

2.2. ボックス存在確率の導出

ある塩基配列 X「ACCUGAU」の中に Box D が含まれるか否かを判定する場合について説明する。Box D は４塩基からなるので、X の先頭４塩基と Box D 配列をまず比較する。この場合、すべての塩基座において塩基が一致しないので、ボックスを構成する塩基配列と一致する割合を表す類似度は 0%となる。次に、切り出す領域を一塩基右にずらして「CCUG」と Box D を比較し、類似度 25%を得る。以下、X の末端となるまで同様の処理を行い、類似度の最大値を X における Box D の存在確率とする。この時、存在確率が 30%以下のボックスは存在しないものとして扱う。他のボックスについても同様の処理を行い、それぞれの存在確率を計算する。

2.3. ステム存在確率の導出

ある塩基配列 Y［･･･(Box D’)GUACGUAA(Box C’)･･･］におけるステム存在確率を求める場合について考える。ボックス間の配列

「GUACGUAA」の両端から内側に向かって、ステムを構成しうる相補対を探索する。８塩基からなるボックス間塩基配列がステムを構成すると仮定した場合、相補対数は４となる。実際には、(U-A)が１つ、(A-U)が１つ、(C-G)が１つの３つが含まれる。ステム存在確率は、構成しうる相補対数に対する、実際の相補対数として定義する。つまり、この例でのステム存在確率は 75%となる。

3. 実験と考察

3.1. 実験データ

データベース「snOPY 」[4]から取得した、長さ 120 の snoRNA 遺伝子全体を含む塩基配列 1000 個を正例として準備する。負例は正例と同じ生物種、同じ長さの、snoRNA 遺伝子から離れた位置にある塩基配列 1000 個とした。評価スコアとして、正例と負例それぞれを正しく判定できる確率である識別精度、正例を正しく判定できる確率である敏感度、負例を正しく判定できる確率である特異度を用いる。

3.2. 実験結果と考察

予備実験により、ボックス存在確率とステム存在確率の平均が 73%以上であれば snoRNA 遺伝子ありとして判定を行うこととした。正例 1000 個と負例 1000 個で実験を行ったときの識別精度、敏感度、特異度をそれぞれ表 1 に示す。

表 1. snoRNA 遺伝子の検出結果

型	識別精度	敏感度	特異度
Box C/D	0.856	0.877	0.835
Box H/ACA	0.547	0.407	0.687

表 1 の通り、Box C/D 型 snoRNA 遺伝子では約 86%という識別精度を得ることができた。しかし、Box H/ACA 型 snoRNA 遺伝子では約 55%という識別精度しか得られなかった。これは、Box H/ACA 型 snoRNA 遺伝子が持つ Box H 内に任意塩基を表す N が多く含まれており、その特徴を捉えにくいためと考えられる。

3.3. Box C/D 型 snoRNA 遺伝子検出実験

Box C/D 型 snoRNA の検出精度を評価するため、長さ 5000 の塩基データから正しく snoRNA 遺伝子の位置を特定できるかどうかの実験を行った。

データベース「snOPY」から snoRNA 遺伝子を含む 5000 塩基長のデータ 100 個を切り出し、あらかじめ snoRNA 遺伝子が何塩基目から何塩基目まで占めるか記録しておく。次に、提案手法で

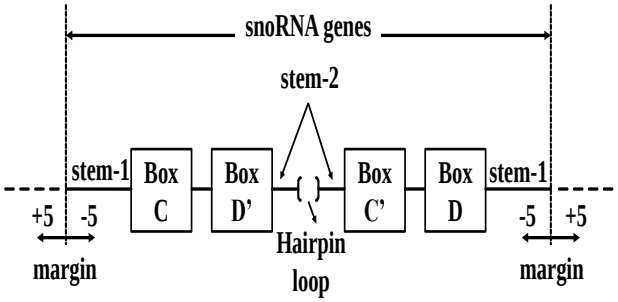

図 2. Box C/D 型 snoRNA 二次構造の展開図

snoRNA 遺伝子の有無を調べ、snoRNA 遺伝子が含まれていた場合はその位置を求める。これらの位置を比較し、検出された snoRNA 遺伝子の位置が正しいかどうか評価する。

図２に示すように、両端の stem-1 から±５塩基の検出誤差を許して実験を行った。その結果、87 個のデータで snoRNA 遺伝子を正しく検出することができた。このことより、提案手法の有効性が示された。

4. おわりに

本研究では、特徴的塩基配列とステム構造に着目した、DNA からの snoRNA 遺伝子検出法を提案した。実際の塩基配列を用いて識別精度を検証した結果、Box C/D 型 snoRNA について約 85.6%の識別精度を得ることができた。

今後の課題としては、Box H/ACA 型 snoRNA をより精度良く検出するため、snoRNA 遺伝子の修飾情報を加味した手法の開発などが挙げられる。

参考文献

[1] Eddy S (2001), Noncoding RNA genes and the modern RNA world. Nature Reviews Genetics.2, pp.919-929
[2] 剣持直哉 (2004), リボソームと疾患. 実験医学 増刊, 22, pp.200-205
[3] 剣持直哉 (2006), ゲノム情報に基づく snoRNA の解析：ゼブラフィッシュを用いたアプローチ. 機能性 Non-coding RNA, pp.83-98
[4] snoRNA Orthological Gene Database; http://snoopy.med.miyazaki-u.ac.jp/snorna_db.cgi

---

A method to detect intronic snoRNA genes using characteristic nucleotide sequences and stems structures
[1]Graduate School of Eng., Univ. of Miyazaki
[2]Faculty of Eng., Univ. of Miyazaki
[3]Frontier Science Research Center, Univ. of Miyazaki