Vol.2012-MPS-90 No.21
2012/9/20

構造モデルに基づく塩基配列からのboxC/D型snoRNA遺伝子検出法

山森一人[†1] 薛晨[†2] 岩切淳一[†3]
剣持直哉[†4] 吉原郁夫[†5]

boxC/D型snoRNAの一次構造，および二次構造を考慮した構造モデルに基づき，同遺伝子を高精度に検出する手法を提案する．提案手法では，boxC/D型snoRNAが持つ特徴的な塩基配列であるボックスと，ステム構造とよばれる相補塩基対で構成された二次構造の双方に着目し，ボックス存在位置等にモデルに基づく制約を課すことで高精度にsnoRNA遺伝子を検出する．実験により，提案手法は97.5%の精度でboxC/D型snoRNA遺伝子を検出できることを示した．

A Structure Model Based Method to Detect boxC/D type snoRNA Genes from DNA Sequences

Kunihito YAMAMORI,[†1] Xue CHEN,[†2]
Junichi IWAKIRI,[†3] Naoya KENMOCHI[†4]
and Ikuo YOSHIHARA[†5]

This paper proposes a method to detect boxC/D type snoRNA genes using structure model that consisting with characteristic base sequences and secondary structure. Our method utilizes both characteristic sequences called as boxes and secondary structure called as stems. We develop a structure model of boxC/D type snoRNA from actual sequences, and also determine a sequence include a boxC/D type snoRNA gene or not based on the matching ratio of the sequence and the model. Our method achieves 97.5% detection ratio.

1. はじめに

ゲノムはたんぱく質のアミノ酸配列をコードするコーディング領域と，それ以外のいわゆるノンコーディング領域に大別される．ゲノム配列解読当初，ノンコーディング領域はジャンクDNAとも呼ばれ，大部分は意味を持たないものと考えられていた．現在では，遺伝子発現調節のほか，たんぱく質の翻訳情報を持たないノンコーディングRNA遺伝子など，生体に必須な情報がこの領域に多く含まれることが明らかにされつつある．

ノンコーディングRNAに代表される機能性RNAの発見と役割の解析は，分子細胞生物学やバイオインフォマティクス双方において最も重要な研究課題の一つになっている[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．しかし，塩基配列の量は膨大で，医学的な検出方法では多くの手間と時間がかかる．そのため，コンピュータによるRNA遺伝子の自動検出法の開発が必要とされている．本報告では，機能性RNAの一つである核小体低分子RNA（snoRNA）[3)]を対象とし，構造モデルとの適合度に基づきsnoRNA遺伝子を自動検出する方法を提案する．比較実験により，サポートベクターマシン（SVM）とカーネル法を組み合わせた従来法に比べ，提案手法は大幅にsnoRNA遺伝子検出精度を向上させうることを示す．

2. snoRNAとその従来検出法

2.1 snoRNAの構造と機能

snoRNAはその構造上の特徴からboxC/D型とboxH/ACA型の二つに分類される．**図1**と**図2**にboxC/D型，boxH/ACA型の二次構造をそれぞれ示す．リボソームRNA（rRNA），またはその他のRNAに対し，boxC/D型snoRNAはメチル化，boxH/ACA型snoRNA

†1 宮崎大学工学教育研究部
Faculty of Engineering, University of Miyazaki
†2 ソフトバンク株式会社
SOFTBANK CORP.
†3 東京大学大学院新領域創成科学研究科情報生命科学専攻
Department of computational biology, Graduate School of Frontier Science, University of Tokyo
†4 宮崎大学フロンティア科学実験総合センター
Frontier Science Research Center, University of Miyazaki
†5 宮崎大学名誉教授
Emeritus Professor, University of Miyazaki

Vol.2012-MPS-90 No.21
2012/9/20

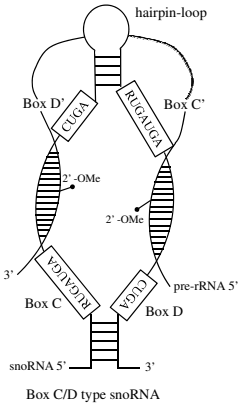

図 1 boxC/D 型 snoRNA の二次構造
Fig. 1 Secondary structure of boxC/D type snoRNA.

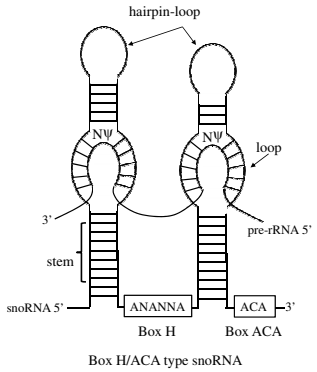

図 2 boxH/ACA 型 snoRNA の二次構造
Fig. 2 Secondary structure of boxH/ACA type snoRNA.

はシュードウリジン化などに関与している．本報告では，これら 2 種類の snoRNA 遺伝子のうち，boxC/D 型を対象とする．

boxC/D 型 snoRNA はおおむね 60〜100 塩基の長さで，両末端の塩基が相補塩基対（A-U，C-G）によるステム構造を持ち，これに隣接して特徴的な配列 boxC と boxD がある．また，ヘアピンループ直下にもステム構造を持ち，これに隣接してもう一セット特徴的な配列（boxC' と boxD'）を持つ場合もある．boxC の 3' 側と boxD の 5' 側に，標的となる rRNA と相補的な配列があり，rRNA のメチル化はこの領域で起こる．

snoRNA による rRNA 修飾の意義は今のところ不明であるが，修飾の異常，または snoRNA そのものが関わっている可能性のある疾患として，先天性角化不全症，B 細胞悪性リンパ腫，プラダー・ウィリー症候群などが報告されている[4]．

2.2 snoRNA 検出の従来法

塩基配列から snoRNA 遺伝子を検出する方法として，サポートベクターマシン（Support Vector Machine,SVM）[5] とカーネル関数[6] を組み合わせた手法がある．SVM は入力サンプルを 2 つのクラスに分類する手法であり，学習サンプルから最適な識別境界を決定する．snoRNA 遺伝子検出に用いられるカーネル関数として比較対象としたのは，文字列部分列カーネル（String Subsequence Kernel，SSK）[7] とステムカーネル（Stem Kernel，StK）[8] の 2 つである．

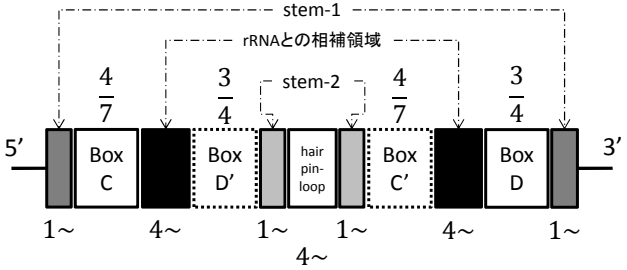

図 3 boxC/D 型 snoRNA 遺伝子の構造モデル
Fig. 3 Proposed structure model for boxC/D type snoRNA.

SSK は文字列に対するカーネルであり，文字配列の類似度を求めるものである．SSK では，文字列 x と y の類似度を，連続，または不連続の部分配列 k の出現頻度をカウントして評価する．StK は RNA 配列に特化したカーネルであり，RNA の二次構造を考慮した特徴空間をとり，任意の長さの連続，または不連続のステム構造候補の出現頻度を数えて特徴ベクトルとする．すなわち，ステム構造の間にギャップが挿入される場合も含めて，すべての可能なステム構造の候補をカウントする．

これら従来研究での snoRNA 遺伝子の検出精度は 65%程度[8] であり，原因として SSK はボックスと呼ばれる特徴的塩基配列だけを，StK はステムと呼ばれる二次構造だけを着目していることが考えられる．

3. 構造モデルに基づく boxC/D 型 snoRNA 検出法

3.1 構造モデルの設計

提案手法では塩基配列を構造モデルと比較し，モデルと一致する箇所に snoRNA 遺伝子が存在すると判定する．そこで，文献[2]，および snoRNA 遺伝子データベース “snOPY”[9] の塩基配列を調べ，**図 3** に示す構造モデルを作成した．図 3 の $\frac{n}{m}$ は m 塩基中 n 塩基以上が一致する必要があることを表し，$n\sim$ は当該箇所が n 塩基以上で構成されていることを表す．

まず，5' 端から説明する．5' 端には 3' 端とペアになる stem-1 が存在する．stem-1 は少なくとも 1 つの相補対からなるものとし，boxC は 5' 端から 1 塩基以上離れているものとする．続く boxC は 5' 端に最も近く，かつ boxC に最も多く塩基が一致する箇所に定める．このとき，boxC を構成する 7 塩基の半分以上，すなわち $\frac{4}{7}$ 以上が一致する必要があるとす

る．boxC に続き，標的 rRNA との相補的な塩基配列がある．この領域は 9∼20 塩基，すなわち片側で平均 4.5∼10 塩基からなる．そこで，標的 rRNA と相補的な配列は片側 4 塩基以上からなるものとした．次の boxD’ は，boxC 端から標的 rRNA との相補配列分の 4 塩基以上離れ，かつ boxD’ を構成する 4 塩基の半分以上，すなわち $\frac{3}{4}$ 以上の塩基が一致する箇所に定める．boxD’ 候補が複数ある場合については後に述べる．boxD’ の後には stem-2 がある．stem-1 と同じく，ここにも 1 つ以上の相補対があるものとする．stem-2 の次に，構造の折り返しとなるヘアピンループがある．「CentroidFold」[10] による予測結果から，この領域は 4 塩基以上からなるものとした．以下，1 塩基以上からなる stem-2，7 塩基中 4 塩基以上が一致している boxC’ 領域，標的 rRNA との相補的な 4 塩基以上の領域，4 塩基中 3 塩基以上が一致する boxD’ 領域，1 塩基以上の stem-1，と続く．

以上が，本研究で用いる boxC/D 型 snoRNA 遺伝子構造モデルである．

3.2 boxC/D 型 snoRNA の存在判定法

snoRNA 遺伝子の有無を判断する流れについて説明する．まず，対象とする塩基配列に各ボックスが含まれる平均確率 P_B を求める．次に，各ボックスが一定以上の確率で含まれる塩基配列において，相補塩基対により形成されるステム構造が存在する確率 P_S を計算する．式 (1) に示す通り，P_B と P_S の重み付き和がしきい値 P_{th} 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．

$$P_{th} = \alpha P_S + \beta P_B, \tag{1}$$

$$\alpha + \beta = 1.0. \tag{2}$$

ここで，α，β はボックス存在確率，ステム存在確率の寄与を調整するパラメータである．

続く 3.3 節と 3.4 節では，ボックス存在確率 P_B とステム存在確率 P_S の導出について詳しく説明する．

3.3 ボックス存在確率の導出

3.3.1 ボックス存在確率の定義

boxC/D 型 snoRNA 遺伝子におけるボックス存在確率の導出法について説明する．boxC/D 型 snoRNA は boxC/D を持つが，boxC’/D’ は持たない場合もある．そこで，各ボックスの存在確率をそれぞれ計算し，その平均をボックス存在確率 P_B とし，式 (3) で定義する．

$$\text{ボックス存在確率 } P_B = \frac{\sum \text{各ボックスの存在確率}}{\text{検出されたボックスの数}}. \tag{3}$$

以下，各ボックスの存在確率の計算方法について詳しく述べる．

3.3.2 boxC の探索

boxC の探索では，5’ 端から 3’ 端に向かって順に一塩基ずつ右にずらしつつ，boxC 存在確率 P_B^C を計算する．boxC 存在確率とは，ボックスを構成する塩基配列と対象配列が一致する割合であり，式 (4) で定義する．

$$P_B^C = \frac{1}{N}\sum_{i=1}^{N} d(C_i, T_i), \tag{4}$$

$$d(C_i, T_i) = \begin{cases} 1 & C_i = T_i, \\ 0 & C_i \neq T_i. \end{cases} \tag{5}$$

式 (4) において，N は比較する塩基数，C_i，T_i はそれぞれ対象ボックスと着目する配列の i 番目の塩基を表す．一塩基ずつずらしつつ計算した P_B^C を比較し，最も大きい P_B^C の位置に boxC が存在すると判断する．同じ P_B^C を持つ領域が複数あった場合，塩基配列の 5’ 側に最も近い領域を採用する．

boxC 領域は boxC/D 型 snoRNA 遺伝子に必ず含まれている．そこで，boxC を構成する “RUGAUGA”（R は A または G）の 7 塩基のうち，boxC 候補領域は少なくともこの半分以上の塩基が一致する必要があるとする．boxC 存在確率が $\frac{4}{7}$ 未満の場合は，当該配列は boxC/D 型 snoRNA 遺伝子を含まないと判断する．

3.3.3 boxD の探索

boxD の探索では，先に決定した boxC の末端から 3’ 端まで順に一塩基ずつずらしつつ，boxD の存在確率 P_B^D を計算する．P_B^D も式 (4) により boxC と同様に求める．なお，boxD は 4 塩基からなるので，式 (4) 中の N は 4 である．

一塩基ずつずらして計算した P_B^D を比較し，最も大きい P_B^D の領域を boxD 領域とする．このとき，同じ P_B^D を持つ領域が複数ある場合が考えられる．そこでまず，「rRNA と相補的な配列」の塩基数制限により boxC から 3’ 端側 4 塩基以内にある boxD 候補を除外する．それでもなお複数の boxD 候補が残った場合，boxC/D 両末端には stem-1 が存在することから，各 boxD 候補について計算した stem-1 存在確率が最大となる boxD 候補を boxD として採用する．stem-1 存在確率については 3.4.2 節で詳しく説明する．それでもなお複数の boxD 候補が残った場合，3’ 端に最も近い boxD 候補を採用する．これは，boxC/D 間の塩基数が多いほど，次に探索する boxC’，および boxD’ が存在する可能性が高くなり，中央部の stem-2 やヘアピンループを構成できる可能性が高くなるためである．

boxD は “CUGA” の 4 塩基からなる．boxC と同じく，対象配列の過半数，すなわち，boxD 存在確率が $\frac{3}{4}$ 未満の場合，対象配列は snoRNA 遺伝子を含まないものとして扱う．

3.3.4 boxC’ と boxD’ の探索

boxC と boxD の位置を決定後，boxC の 3’ 端から boxD の 5’ 端まで順に一塩基ずつ右にずらしつつ，boxC’/D’ の存在確率 $P_B^{C'}$，$P_B^{D'}$ を計算する．存在確率の計算には，boxC/D と同じく式（4）を用いる．求めた $P_B^{C'}$，$P_B^{D'}$ を比較し，最も $P_B^{C'}$，$P_B^{D'}$ が大きい個所に boxC’/D’ がそれぞれ存在すると判断する．同じ $P_B^{C'}$，$P_B^{D'}$ を持つ boxC’/D’ 候補が複数ある場合，以下の条件に当てはまる候補を除外する．

- 「rRNA と相補的な配列」の塩基数制限を用いて，boxC や boxD から 4 塩基以内にある候補を除外する．
- boxC と boxD の中間より boxC’ は 3’ 側，boxD’ は 5’ 側に存在するので，本来存在し得ない位置にある boxC’/D’ 候補を除外する．
- boxC’/D’ の間に最小 4 塩基からなるヘアピンループ領域を構成しえない候補を除外する．

上記の条件に当てはまる boxC’/D’ 候補領域を除外したのち，boxC’/D’ 候補間には stem-2 が存在することを考え，**図 4** のように残っている候補間で，3.4.3 節で説明する stem-2 存在確率を計算し，stem-2 存在確率が最大の boxC’/D’ 候補を boxC’/D’ と決定する．boxC’/D’ 間の stem-2 存在確率を評価したあとでもなお複数の boxC’/D’ 候補が存在する場合，boxC’ 候補と boxD’ 候補の間が一番離れている組み合わせを採用し，boxC’/D’ に決定する．これは，boxC’ と boxD’ が離れるほど，boxC’/D’ 間に stem-2 を構成できる塩基数が増えるためである．

boxC/D と異なり，$P_B^{C'}$ が $\frac{4}{7}$ 未満，$P_B^{D'}$ が $\frac{3}{4}$ 未満の場合，当該配列は snoRNA 遺伝子を含まないのではなく，boxC’ や boxD’ を含まないものとして扱う．

3.4 ステム存在確率の導出

3.4.1 ステム存在確率の定義

ステム存在確率の定義について説明する．boxC/D 型 snoRNA は boxC と boxD の両端にステム構造 stem-1 を持ち，boxC’ と boxD’ の間にもまた別のステム構造 stem-2 を持つ．boxC’ と boxD’ を含まない場合，stem-2 は boxC と boxD の間に構成する．したがって，前節までに確定した boxC/D や boxC’/D’ の間に存在するはずの 2 つのステム構造それぞれに対して stem 存在確率を計算し，その平均をとってステム存在確率 P_S を式 (6) で定義する．

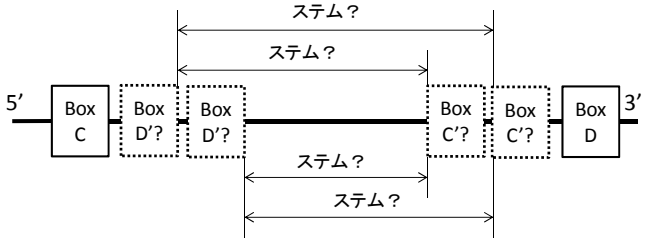

図 4 複数の boxC’/D’ 候補が存在する例
Fig. 4 An example of sequences including multiple boxC’/D’s.

$$\text{ステム存在確率 } P_S = \frac{\text{stem-1 存在確率} + \text{stem-2 存在確率}}{2}. \tag{6}$$

3.4.2 stem-1 存在確率の導出

boxC/D 端に存在する stem-1 の存在確率を求めるため，boxC 端から 5’ 側の 10 塩基，boxD 端から 3’ 側の 10 塩基に着目する．両端の塩基数が 10 未満の場合は，どちらか短い方に合わせる．着目する塩基数を 10 とした理由は，CentroidFold を用いた予備実験で stem-1 を構成する相補塩基対数の最大値が 7 と求められたため，余裕を含んで設定したものである．

stem-1 存在確率は，box 端に塩基の挿入や欠損が発生している場合を考慮して，塩基の挿入がない場合，boxC の 5’ 側に 1 塩基の挿入があった場合，boxD の 3’ 側に 1 塩基の挿入があった場合の 3 パターンで計算し，ステムを構成できる相補塩基対数が最も多い場合を採用する．例として，ある塩基配列 X「...AAAAGUCA（box C）...（boxD）GACUUUUU...」における stem-1 出現確率を求める場合について説明する．

boxC の 5’ 側端から 8 塩基，boxD の 3’ 側端 8 塩基すべてがステムを構成すると仮定した場合，相補塩基対数は 8 であり，これを分母とする．boxC 端と boxD 端に塩基挿入がない場合は (A-U) が 4 つ含まれることから，stem-1 存在確率は $\frac{4}{8} = 0.5$ となる．次に，boxC の 5’ 側に塩基挿入があった場合を考える．この場合 (G-C) が 1 つ，(A-U) が 1 つ，(C-G) が 1 つ，(U-A) が 4 つの計 7 となり，この場合の stem-1 存在確率は $\frac{7}{8} = 0.875$ となる．最後に，boxD の 3’ 側に塩基挿入があった場合を考える．この場合 (A-U) が 3 つが含まれるので，stem-1 存在確率は $\frac{3}{8} = 0.375$ となる．したがって，この例では boxC 端に塩基の挿入があった場合の 0.875 が stem-1 存在確率となる．

3.4.3 stem-2 存在確率の計算

snoRNA は boxC’ と boxD’ の間にステム構造 stem-2 を構成する．boxC’/D’ が含まれ

ない場合は，boxC と boxD の間に stem-2 を構成する．stem-2 存在確率を求めるとき，提案手法ではヘアピンループを考慮せずに行う．これは，ヘアピンループ部分の位置が不確定のためである．

stem-2 存在確率は，3.4.2 節で示した stem-1 存在確率の導出時と同じく，boxC' と boxD' の間に塩基挿入がない場合，boxC' 側に 1 塩基が挿入されている場合，boxD' 側に 1 塩基が挿入されている場合の三つのパターンで計算する．このうち，ステムを構成できる相補塩基対数が最も多い場合を採用し，stem-2 存在確率とする．

4. 実験と考察

4.1 実験に使用する塩基配列データ

評価実験を行うにあたり，boxC/D 型 snoRNA 遺伝子を含む正例データと，それを含まない負例データを作成した．これら評価実験で用いるデータは，データベース “snOPY”[9] から取得した．評価実験で用いる正例データは，snoRNA 遺伝子を含む長さ 120 塩基の配列で，1,100 個作成した．負例データは，正例データと同じ生物種の塩基配列から snoRNA 遺伝子をできる限り含まないように，snoRNA 遺伝子と離れた位置から長さ 120 塩基で 1,100 個作成した．作成したデータのうち，正例 100 個と負例 100 個は予備実験に使用する．作成した負例データは，塩基配列内に snoRNA 遺伝子が含まれていないか「CLUSTALW」[11] を用いて検証を行った．CLUSTALW は広く用いられている多重整列プログラムであり，これによりペアワイズアラインメントを行った後，二つの塩基配列の一致度を分析して重複がないことを確認した．

4.2 ボックス存在確率係数，ステム存在確率係数，およびしきい値の決定

予備実験として，3.2 節で述べた boxC/D 型 snoRNA 遺伝子の有無を判断するしきい値 P_{th}，および判定に係るボックス存在確率とステム存在確率の寄与を調整するパラメータ α，β を決定する．ここでは先に取り置いた正例データ 100 個と負例データ 100 個を用いて，α と β を 0.0 から 1.0 まで 0.1 刻みで変更し，各 α，β の組み合わせについて P_{th} を 0.1 から 0.9 まで 0.1 刻みで変更して最も高い検出精度（Accuracy）を示した α，β，P_{th} の組み合わせを決定する．なお，検出精度とはすべてのデータのうち，正例，負例を正しく判定できたデータ数の割合である．

各 α，β，P_{th} での検出精度の一部を**表 1** に示す．最も良く snoRNA 遺伝子を検出した (α, β, P_{th}) は，$(0.5, 0.5, 0.5)$，$(0.3, 0.7, 0.6)$，$(0.1, 0.9, 0.7)$ であり，表 1 中に太字で表す．評価実験では，この三つの組み合わせにおいて，それぞれ boxC/D 型 snoRNA 遺伝子の検出を行う．

表 1 α，β，P_{th} の組み合わせによる検出精度
Table 1 Accuracy for each combination of α, β and P_{th}.

		P_{th}								
α	β	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.5	0.5	0.920	0.920	0.920	0.945	**0.985**	0.920	0.785	0.530	0.500
0.3	0.7	0.920	0.920	0.920	0.920	0.945	**0.985**	0.955	0.725	0.500
0.1	0.9	0.920	0.920	0.920	0.920	0.920	0.925	**0.985**	0.920	0.670

表 2 boxC/D 型 snoRNA 遺伝子の検出精度
Table 2 Sensitivity, specificity and accuracy on boxC/D type snoRNA gene detection experiments.

α	β	P_{th}	感度	特異度	識別精度
0.5	0.5	0.5	0.951	0.990	0.971
0.3	0.7	0.6	0.955	0.994	0.975
0.1	0.9	0.7	0.953	0.999	0.976

4.3 boxC/D 型 snoRNA 遺伝子検出実験結果

予備実験で決定した (α, β, P_{th}) を用いて，boxC/D 型 snoRNA 遺伝子の検出実験を行う．正例データと負例データをそれぞれ 1,000 個を使用して実験を行い，感度（Sensitivity），特異度（Specificity），検出精度を比較する．感度とは正例を正しく判定できる確率，特異度とは負例を正しく判定できる確率である．

表 1 に示した 3 種類の (α, β, P_{th}) の組み合わせについて，snoRNA 遺伝子検出実験を行った結果を**表 2** に示す．

表 2 に示した通り，いずれの (α, β, P_{th}) の組み合わせにおいても，検出精度 0.97 以上を達成している．3 種類の (α, β, P_{th}) の組み合わせのうち最も感度が高い $(0.3, 0.7, 0.6)$ について，従来研究との比較を行った結果を**図 5** に示す．図 5 から，提案手法は感度，特異度，検出精度のいずれも従来研究より大幅に向上している．感度を比較すると，提案手法は SSK より 51.5%向上し，StK より 15.5%向上している．特異度について，提案手法は SSK より 22.2%向上，StK より 38.8%向上している．検出精度では，提案手法は SSK より 31%向上，Stk より 33.5%向上している．

4.4 長大配列からの snoRNA 遺伝子検出実験

本節では，提案手法により長い塩基配列上の snoRNA 遺伝子を正しく検出できるか検証する．データベース “snOPY” から，snoRNA 遺伝子を含む「host gene」からサンプル 100

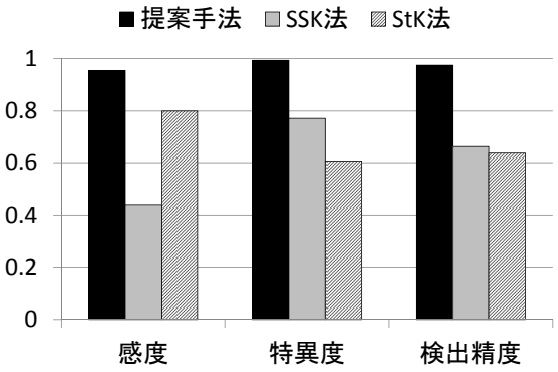

図 5 従来研究との比較
Fig. 5 Comparison proposed method with previous works.

個を取り出す．生物種はヒト，線虫，ネズミとシロイヌナズナの 4 種類である．長さは生物種ごとに異なっているが，少なくとも 2,000 塩基以上とした．

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ところ，100 サンプル中 94 サンプルで snoRNA 遺伝子を正しく検出することができた．正しく検出することができなかったサンプル 6 個のうち 4 つがヒトのデータであった．例として，“SNORD74L2” というヒト遺伝子について説明する．“SNORD74L2” では，4 塩基中 3 塩基以上が一致する boxD 候補が 2 つ見つかる．図 3 に示したモデル構造に基づくと，boxC の末端から boxD の先頭までは 26 塩基以上離れていなければならない．しかし，“SNORD74L2” で見つかる 2 つの boxD 候補はこの制限を満たしておらず，snoRNA 遺伝子として検出することができなかった．

5. おわりに

本報告では，boxC/D 型 snoRNA のボックスとステム両方に着目し，構造モデルとの一致確率に基づく塩基配列からの snoRNA 遺伝子検出法を提案した．構造モデルは，1 つ以上の相補塩基対からなるステム領域 2 個所，7 塩基中 4 塩基以上が一致している boxC/C’ 領域，4 塩基中 3 塩基以上が一致する boxD/D’ 領域，4 塩基以上の標的 rRNA との相補領域，4 塩基以上のヘアピンループ領域からなる．このモデル構造に基づき，各ボックスの存在可能位置などいくつかの制限を設定している．そのため，ボックス候補が多数存在しても，モデル構造によって存在しえない候補を除外することができる．

実際の塩基配列を用いて検出精度を検証した結果，boxC/D 型 snoRNA について 97.5%の検出精度を得ることができた．さらに，長さ 2,000 塩基以上の，ヒトや線虫，ネズミ，シロイヌナズナの塩基配列を使い，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ところ，100 サンプル中 94 サンプルで snoRNA 遺伝子を正しく検出することができた．

今後の課題として，構造モデルを改良しての精度向上や，任意位置での塩基の挿入・欠損を考慮したステム存在確率の計算などが挙げられる．

参 考 文 献

1) Eddy, S.R.: Non-coding RNA genes and the modern RNA world, *Nat. Rev. Genet.*, Vol.2, No.12, pp.919–929 (2001).
2) 河合剛太，清澤秀孔：機能性 RNA の分子生物学，クバプロ (2010).
3) Smith, C.M. and Steitz, J.A.: Sno Storm in the Nucleolus: New roles for Myriad small RNPs, *Cell*, Vol.89, pp.669–672 (1997).
4) 河合剛太，金井昭夫：機能性 Non-coding RNA，クバプロ (2010).
5) 澤田石翔太，三功浩嗣，土井晃一郎，山本章博：内包カーネルと配列分割法を用いた RNA 識別，情報処理学会研究報告（バイオ情報学）， Vol.2008, No.15, pp.71–78 (2008).
6) Schölkopf, B., Tsuda, K. and Vert, J.P.: *Kernel Methods in Computational Biology*, The MIT Press (2004).
7) Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N. and Watkins, C.: Text Classification using String Kernels, *J. Mach. Learn. Res.*, Vol.2, pp.419–444 (2002).
8) Sakakibara, Y., Asai, K. and Sato, K.: Stem Kernels for RNA Sequence Analyses, *Bioinformatics Research and Development* (Hochreiter, S. and Wagner, R., eds.), Lecture Notes in Computer Science, Vol. 4414, Springer Berlin/Heidelberg, pp.278–291 (2007).
9) RI Laboratory Frontier Science Research Center: snOPY, University of Miyazaki (online), available from ⟨http://snoopy.med.miyazaki-u.ac.jp/snorna_db.cgi⟩ (accessed 2012-7-20).
10) Sato, K., Hamada, M., Asai, K. and Mituyama, T.: CENTROIDFOLD: a web server for RNA secondary structure prediction, *Nucleic Acid Research*, Vol. 37, No.suppl 2, pp.W277–W280 (2009).
11) Thompson, J.D., Higgins, D.G. and Gibson, T.J.: CLASTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acid Research*, Vol.22, No.22, pp.4673–4680 (1994).