社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－BIO－12 (10)
2008／3／4

内包カーネルと配列分割法を用いたRNA識別

澤田石 翔太[1] 三功 浩嗣[2] 土井 晃一郎[2] 山本 章博[2]
[1]京都大学工学部
[2]京都大学大学院情報学研究科
〒606-8501 京都府京都市左京区吉田本町

タンパク質に翻訳されずに機能する機能性RNAの解析・発見に注目が集まり，機能性RNAの識別に対するカーネル関数がいくつか提案されている．本稿では，機能性RNA識別に対して内包カーネルと配列分割法を組み合わせた手法を提案する．内包カーネルとは文法における導出に基づいたカーネル関数であり，この考え方に基づいてRNAの二次構造を考慮したカーネル関数を構築している．配列分割法とは入力配列を分割してカーネル関数に適用する手法である．他のカーネル関数との比較実験を通して本手法の構造予測における有用性を示した．

Intentional kernel and sequence partition method for RNA classification

Shota Sawataishi[1] Hiroshi Sankoh[2] Koichiro Doi[2] Akihiro Yamamoto[2]
[1]Faculty of Engineering, Kyoto University
[2]Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan

Several kernel functions were proposed for the discrimination and detection of functional RNA sequences. In this paper, we propose a method combining the intentional kernel and the sequence partition method for RNA classification. The intentional kernel is based on derivations constructing structure, and we proposed an intentional kernel based on context free grammar for RNA sequences. The sequence partition method is to partition input sequences into substrings and apply the substrings to kernel functions. We compare our method with other kernel functions, and show efficiency of our method for RNA structure prediction.

1 はじめに

近年，ゲノム中でタンパク質に翻訳されずに機能する**機能性 RNA** の解析・発見に注目が集まっている [2]．機能性RNAは，遺伝子発現等の様々な制御を行うことで，複雑な生命活動を維持し，高等生物とそれ以外との違いに大きく寄与している可能性が高いと考えられている．

これまでに様々な機能性 RNA が発見されている．これらの機能性 RNA については，a, u, c, g で表される 4 種類の塩基の列で表現されたRNA配列が，機能や構造ごとにRNAファミリとして分類され，データベースに蓄積されている [3]．RNA配列の最大の特徴は，Watson-Crickの塩基の相補対 (a-u，c-g) により形成される二次構造である．通常は 1 本鎖のままで存在するRNA分子が，その鎖上の一部の相補対が結合することにより折りたたまれて安定した構造をな

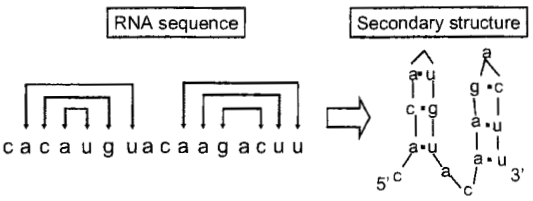

図 1: RNA 配列とその二次構造

す (図 1).

RNA 配列が与えられた際に，既存のどの RNA ファミリに属するかを識別する様々な手法が考案されており，その中には，近年注目されているサポートベクターマシン (Support Vector Machine, SVM) とカーネル関数を組み合わせた手法も，いくつか提案されている [5][7]．我々も先に，RNA 配列の二次構造を考慮したカーネル関数のクラス K_{RNA} を提案した [8][9]．K_{RNA} 内のカーネル関数は，内包カーネルの概念 [1] を基に設計し，[8][9] ではその代表的なカーネル関数 K^N_{RNA} の定義式及び計算法を与えた．

本稿では，内包カーネル関数 K^N_{RNA} と配列分割法を組み合わせた手法を提案する．配列分割法とは入力配列を分割して他のカーネル関数に適用する手法である．そして，文字列カーネル (String Subsequence Kernel, SSK) [6]，ステムカーネル [7] との比較実験を通して本手法の有用性を示す．

本稿は以下のように構成されている．2 章で準備としてカーネル関数と本稿で実験に使用するカーネル関数の紹介をする．3 章では内包カーネル K_RNA の紹介を行う．4 章では配列分割法についての提案を行う．5 章では計算機実験を行い，6 章では，本稿のまとめと，今後の課題を与える．

2 カーネル関数

線形判別関数を構成する学習手法を非線形の判別関数の構成に適用するには，訓練データ $\mathbf{x}$ を高次元の特徴空間 $\mathbf{R}^d$ に写像 ϕ によって射影し，その像に対して線形判別関数を構成するという手法が用いられる．SVM をはじめとするカーネル法では高次元の特徴空間上の 2 つのデータの内積を表すカーネル関数 $K(\mathbf{x}_i, \mathbf{x}_j) = \boldsymbol{\phi}(\mathbf{x}_i) \cdot \boldsymbol{\phi}(\mathbf{x}_j)$ だけを用いて判別関数を構成する．よって非線形の判別関数を求める際にも，高次元の特徴空間における訓練データの像の内積の値を知ればよい．この内積の値を与える関数がカーネル関数である．

以下，本稿で比較実験を行う先行研究におけるカーネル関数は次の二つである．文字列部分列カーネル (String Subsequence Kernel, SSK)[6] とは文字列に対するカーネルであり，長さ n の文字列の出現を特徴空間の座標としている．長さ n の文字列が連続または不連続に文字列中に現れるのをカウントしてギャップ部分の長さも含めた部分文字列の長さ m の文字列の出現に対して λ^m の重みづけをしている．この SSK は文字列 x, y に対して計算量 $O(n|x||y|)$ で計算できる．

ステムカーネル [7] とは RNA 配列に対するカーネルであり，RNA の二次構造を考慮した特徴空間をとっている．任意の長さの連続または不連続のステム構造の候補の出現頻度を数えて特徴ベクトルとしている．ギャップの長さに対するパラメータ λ, 相補対の距離に対するパラメータ α，ループの長さのパラメータ L などのパラメータが提案されている．他にもパラメータはあるが本稿中の実験では上記のパラメータのみを使用する．このステムカーネルは文字列 x, y に対して計算量 $O(|x|^2|y|^2)$ で計算できる．

3 RNA 識別のための内包カーネル K_{RNA}

本章では，まず内包カーネルの考え方について説明する．次に，K_{RNA} が用いている，RNA の二次構造を表現する文脈自由文法 (CFG) G_{RNA} を説明し，RNA の二次構造と G_{RNA} の導出クラスの対応関係を定義する．K_{RNA} は，G_{RNA} の導出クラスに対して内包カーネルを定義することで与えられる．

3.1 内包カーネルの考え方

内包カーネル (intentional kernel) は構造データを対象としたカーネル関数のクラスである．構

造データを対象とするカーネルの代表的なクラスである畳み込みカーネル (convolution kernel) とは対照的な設計方針に基づくものである．すなわち，**構造の輪郭**の方が，詳細な部分構造よりも類似性に強く反映されているという考え方に基づく．

内包カーネルは，もともと演繹推論とカーネルとを組み合わせるモデルとして提案された．そこでは，推論規則による式変形が重要な役割を果たす [1]．RNA 配列のような記号列に適用する場合には，推論規則は文法として表現する．より厳密には，対象とする構造データ全体の集合 D に半順序関係 $\succeq$ が定義されているとき，$x \in D$, $y \in D$ に対して，$E(x,y) = \{z \mid z \succeq x, z \succeq y\}$ を用いて表現されるカーネルを内包カーネルと定義する．内包とよぶ理由は，個々の半順序データが概念を表している場合，$a \succeq b$ を，“a から b が導出される” と解釈できるからである．典型的な内包カーネルとしては 2 つの一階述語論理の項を入力としたカーネル K_{TERM} があり，以下のように定義される

$$K_{TERM}(x,y) = \#(E(x,y)).$$

ここで，$\#$ は集合の要素数を表す．$K_{TERM}(x,y)$ の値は，一階述語論理の原子論理式に対する推論規則である，関数代入，定数代入，変数の単一化を，x, y の両方を導出する最汎な原子論理式に適用し，x と y の最小共通汎化に向かう全ての導出の個数である [1].

3.2 RNA 配列の二次構造を表現する文脈自由文法 G_{RNA}

K_{RNA} を設計するために必要な RNA の二次構造を表現するための文脈自由文法 (CFG) G_{RNA} と導出について [8] に沿って説明する．二次構造には分岐構造や疑似ノット構造，ヘアピン構造などがあるが，G_{RNA} はヘアピン構造のみを表現する．分岐構造は，ヘアピン構造の合成として表現することが可能である．G_{RNA} は，非終端記号 P, L, R, S, E を用いて構成される．P は塩基の相補対を表すために，L と R はそれぞれ左と右に一つの塩基を出力するという操作を表すために用いる．S は開始記号であり，E は空

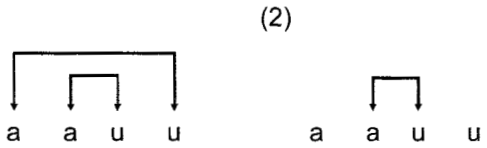

図 2: RNA 配列の二次構造

列を導出し，導出の終了を表す．G_{RNA} は以下の生成規則からなる．

$$\begin{aligned} S &\rightarrow P \mid L \mid R \mid E, \\ P &\rightarrow xPy \mid xLy \mid xRy \mid xEy, \\ L &\rightarrow xP \mid xL \mid xR \mid xE, \\ R &\rightarrow Px \mid Lx \mid Rx \mid Ex, \\ E &\rightarrow \epsilon. \end{aligned}$$

ここで，x は塩基を表し，$x \in$ {a, u, c, g} である．また，(x,y) は塩基の相補対を表し，$(x,y) \in$ {(a,u), (u,a),(c,g), (g,c)} である．

例 1 配列 aauu の導出の例として，

$$\begin{aligned} \sigma_1 :&\quad S \rightarrow P \rightarrow \mathrm{a}P\mathrm{u} \rightarrow \mathrm{aa}E\mathrm{uu} \rightarrow \mathrm{aauu} \\ \sigma_2 :&\quad S \rightarrow R \rightarrow L\mathrm{u} \rightarrow \mathrm{a}P\mathrm{u} \rightarrow \mathrm{aa}E\mathrm{uu} \rightarrow \mathrm{aauu} \end{aligned}$$

などがあり，それぞれ，図 2 の (1)，(2) の二次構造を表現する．

本稿では，RNA 配列の二次構造に対応する G_{RNA} の導出クラスを以下のように定義する．

定義 1 G_{RNA} の S から終端記号列への導出 σ に対して

終端記号 a, u, c, g と非終端記号 E を除去

という操作を行った結果得られる列を σ' と表す．このとき，G_{RNA} の任意の導出 σ_1，σ_2 に対して，

$$\sigma_1 \sim \sigma_2 \quad \Leftrightarrow \quad \sigma_1' = \sigma_2' = \sigma'$$

という同値関係 $\sim$ で得られる同値類 $[\sigma]$ を，**RNA 配列の二次構造に対応する G_{RNA} の導出クラス**と呼ぶ．以下では $[\sigma]$ を σ' と表す．

さらに，RNA 配列の二次構造に対応する G_{RNA} の導出クラスの同値類と，G_{RNA} の代表導出クラスを以下のように定義する．

定義 2 G_{RNA} の導出クラス σ' に対して以下の操作を行った結果得られる列を σ'' と表す.

- σ' の最後に現れる P，存在しなければ，S の後に続く L と R の繰り返しを除去
- σ' において L と R のみが現れる部分を，L を優先して並べかえる.

このとき，G_{RNA} の任意の導出クラス σ'_1，σ'_2 に対して，

$$\sigma'_1 \approx \sigma'_2 \quad \Leftrightarrow \quad \sigma''_1 = \sigma''_2 = \sigma''$$

という同値関係 $\approx$ で得られる同値類 $[\sigma']$ を，**RNA 配列の二次構造に対応する G_{RNA} の代表導出クラス**と呼ぶ．以下では $[\sigma']$ を σ'' と表す.

また，G_{RNA} の代表導出クラス σ''_1，σ''_2 に対して，半順序関係 $\succeq$ を以下のように定義する.

定義 3 G_{RNA} の代表導出クラス σ'_1，σ'_2 について，σ''_1 が σ''_2 の部分列であるとき，$\sigma''_1 \succeq \sigma''_2$ とする.

例 2 図 2 の (1)，(2) の二次構造に対応する G_{RNA} の代表導出クラスは，それぞれ

$$\sigma''_1 : \ S \to P \to P$$
$$\sigma''_2 : \ S \to L \to R \to P$$

と表現される．これにより，RNA の二次構造と G_{RNA} の代表導出クラスが一対一に対応する．また $\sigma''_3 : S \to P$ とすると，定義 3 により $\sigma''_3 \succeq \sigma''_1$ となる.

3.3 内包カーネル関数のクラス K_{RNA} の定義

内包カーネルの，最も外側に見える構造の輪郭から，より詳細な内部構造の輪郭へ注目を移していくという設計方針は，CFG の導出が，最も広い概念 (開始記号 S) から出発し，最終的に最も具体的な概念 (終端記号列) へ至るという性質に合致する．3.2 節において，RNA 配列の最大の特徴である二次構造と，G_{RNA} の導出クラスの対応関係が定義できたので，本節では，G_{RNA}

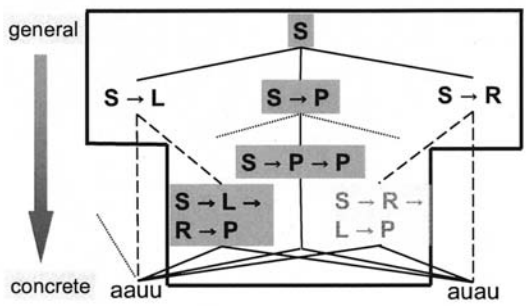

図 3: K_{RNA}(aauu, auau) の計算

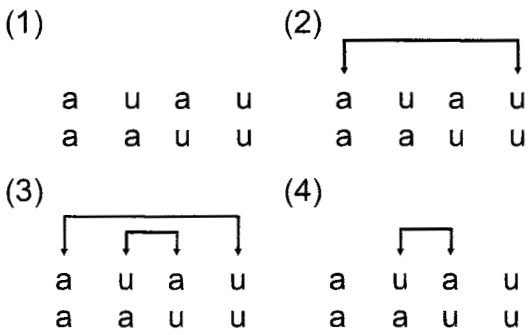

図 4: 配列 aauu と配列 auau の共通二次構造候補

の導出クラスに対する内包カーネル K_{RNA} を定義する．K_{RNA} は，2 つの入力配列 x，y の最も大まかな共通の輪郭，すなわち導出クラス S から注目し始め，G_{RNA} の生成規則を適用することで，共通な導出クラスを徐々に具体化する．最終的に，共通な代表導出クラスの個数を数え上げることで，カーネル値と定義する．共通な代表導出クラスの個数は，共通二次構造の候補数と等しいので，K_{RNA} は，2 つの入力配列 x，y の共通二次構造の候補数を数え上げることで，類似度と定めていることと等価である.

例 3 K_{RNA} の二つの入力配列 aauu と，配列 auau に対して，K_{RNA}(aauu, auau) = 4 となる．このとき，

$$\sigma''_1 : \ S,$$
$$\sigma''_2 : \ S \to P,$$
$$\sigma''_3 : \ S \to P \to P,$$
$$\sigma''_4 : \ S \to L \to R \to P,$$

の4つが共通な代表導出クラスである．図3にK_{RNA} (aauu, auau) を求める際の，導出クラス間の半順序関係を示す．$\sigma'_5 : S \to R \to L \to P$は定義2により，代表導出クラス$\sigma''_4 : S \to L \to R \to P$ の同値類であることに注意する．これら4つの共通な代表導出クラスに対応する共通二次構造の候補を図4に示す．

2つの入力配列 $x = x_0x_1 \cdots x_{|x|-1}$，$y = y_0y_1 \cdots y_{|y|-1}$ ($|x| \leq |y|$ とする) に対して，長さNの区切り窓を導入し，2つの長さNの部分配列s, tを抽出する．長さの等しい配列sとtが共通して取り得る二次構造候補の数を，$K_{common}(s,t)$とするとき，$K^N(x,y)$を以下のように定義する．

$$K^N_{RNA}(x,y) = \sum_{i=0}^{|x|-N} \sum_{j=0}^{|y|-N} K_{common}(x^{(i)}_N, y^{(j)}_N).$$

ここで，$x^{(i)}_N = x_ix_{i+1} \cdots x_{i+N-1}$ $(i = 0, \ldots, |x|-N)$, $y^{(j)}_N = y_jy_{j+1} \cdots y_{j+N-1}$ $(j = 0, \ldots, |y| - N)$ である．$|x| < N \leq |y|$ のときは，区切り窓の中央$N_{\lfloor N/2 \rfloor}$とxの中央$x_{\lfloor |x|/2 \rfloor}$を揃え，両端に空文字を挿入した状態で，$y$に対してスライドさせる．このとき，

$$K^N_{RNA}(x,y) = \sum_{i=\lfloor N/2 \rfloor - \lfloor |x|/2 \rfloor}^{|y|-N+\lfloor N/2 \rfloor - \lfloor |x|/2 \rfloor} K_{common}(x, y^{(i)}_{|x|})$$

とする．$N > |y|$ のときは，xの中央$x_{\lfloor |x|/2 \rfloor}$と$y$の中央$x_{\lfloor |y|/2 \rfloor}$を揃え，重なり部分に注目し，

$$K^N_{RNA}(x,y) = K_{common}(x, y^{(\lfloor |y|/2 \rfloor - \lfloor |x|/2 \rfloor)}_{|x|})$$

とする．$K^N(x,y)$は再帰式により表現し，動的計画法を適用することで，$O(N|x||y|)$で計算可能である [8, 9]．

前章で述べたカーネルが部分文字列を特徴にとっているのに対して，この$K^N_{RNA}(x,y)$は文字ではなく構造のみを特徴としている．ステムカーネルも相補対に関わる部分文字列を特徴としているのであって，構造だけを特徴としているわけではない．

4 配列分割法

配列分割法とは入力配列を分割してから他のカーネルに適用する手法である．以下，配列分割法に対する詳細な定義を与える．

定義 4 (配列の細分) $m \in \mathbf{N}$ が与えられたとき，文字列s ($|s| \geq m$)に対して文字列$s_1, s_2, \ldots, s_m$が$s = s_1s_2 \ldots s_m$を満たすときsの細分という．任意の文字列sに対してsの細分を与える写像を$\Delta'_m : s \mapsto (s_1, \ldots, s_m)$で表す．とくに，$s = x_0x_1 \ldots x_{n-1}$ $(x_i \in \Sigma)$に対して

$$s_i = x_{\lfloor n(i-1)/m \rfloor} \cdots x_{\lfloor ni/m \rfloor - 1}$$

となる細分をm等分という．

定義 5 (文字列の組み合わせ) $s_1, s_2, \ldots, s_m$を文字列とするときに，それらを組み合わせて得られる文字列

$$\begin{aligned} s'_1 &= s_{\gamma(1,1)}s_{\gamma(1,2)} \cdots s_{\gamma(1,m_1)} \\ s'_2 &= s_{\gamma(2,1)}s_{\gamma(2,2)} \cdots s_{\gamma(2,m_2)} \\ s'_3 &= s_{\gamma(3,1)}s_{\gamma(3,2)} \cdots s_{\gamma(3,m_3)} \\ &\vdots \\ s'_d &= s_{\gamma(d,1)}s_{\gamma(d,2)} \cdots s_{\gamma(d,m_d)} \end{aligned}$$

をs_1からs_mの**組み合わせ**という．ここで，$\gamma(i,j)$ $(1 \leq i \leq d, 1 \leq j \leq m_i)$は$1, \ldots, m$までの数($s$の添え字)をもれなく重複なく表す関数である．組み合わせを与える写像を$\Gamma : (s_1, \ldots, s_m) \mapsto (s'_1, \ldots, s'_d)$でという写像とみなすことにする．

定義 6 (文字列sの分割) sの**分割**とは，sのΔ'による細分$(s_1, s_2, \ldots, s_m)$に対して組み合わせΓを行って得られる文字列の組

$$\Gamma\Delta'(s) = (s'_1, s'_2, \ldots, s'_d)$$

である．d **等分割**とは，Δ'がm等分であって，組み合わせΓにおいて$s'_1, \ldots, s'_d$をつくる文字列の数$(m_1, m_2, \ldots, m_d)$がすべて等しいときをいう ($d = |\Gamma|$)．分割$\Delta = \Gamma\Delta'$とカーネル関数$K(x,y)$に対して，

$$K/\Delta(x,y) = \sum_{i=1}^{d} K(x'_i, y'_i)$$

と表記する．

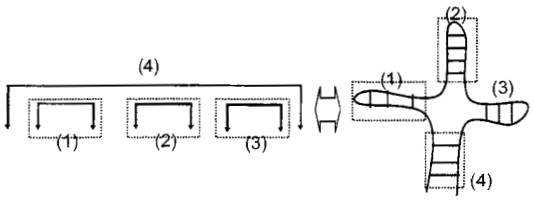

図 5: tRNA の二次構造

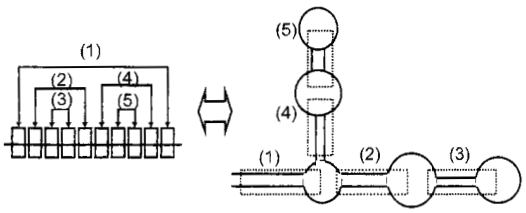

図 6: 5SrRNA の二次構造

命題 1 K がカーネル関数であるとき，$K/\Delta(x,y)$ はカーネル関数である．

命題 2 Δ を d 等分割とするとき，カーネル関数 $K/\Delta(x,y)$ の計算量は次のようになる．

- $O(n|x||y|/d)$ (K が SSK の場合)
- $O(N|x||y|/d)$ (K が K^N_{RNA} の場合)
- $O(|x|^2|y|^2/d^3)$ (K がステムカーネルの場合)

このように配列を分割することにより，計算量が小さくなり与えられた文字列が全体として似通っている場合には対応する部分同士がもとのカーネル関数で計算されることになり，識別率の向上も見込まれる．

ステムカーネルや K_{RNA} は分岐が考慮されていないことより，配列分割法においてRNAの二次構造に対応した分割を考えることができる．二次構造が既知であればその二次構造を考慮した分割を行えば性能の向上が見込まれる．tRNAの場合には図5のような二次構造をしているので，tRNAの配列 s を8等分に細分をして，$s'_1{=}s_2s_3$，$s'_2{=}s_4s_5$，$s'_3{=}s_6s_7$，$s'_4{=}s_1s_8$ と組み合わせる4等分割を行う．5S rRNAの場合は図6のような二次構造をしているので，tRNAの配列 s を10等分に細分をして，$s'_1{=}s_1s_{10}$，$s'_2{=}s_2s_5$，$s'_3{=}s_3s_4$，$s'_4{=}s_6s_9$，$s'_4{=}s_7s_8$ と組み合わせる5等分割を行うことが可能である．このような分割を[8]でも K^N_{RNA} と組み合わせて実験しているが，次章では配列分割の部分を独立させその有用性を示す．

5 計算機実験

K^N_{RNA} と配列分割法の組み合わせの性能評価のために，他のカーネルとの比較実験を行った．実験の手順は以下の通りである．カーネル関数は SVM^{light}[4] のオリジナルなカーネル関数として C 言語で実装し，CPU AMD Athron64X2 4400+，メモリ 2GB のマシン上で計算を行った．RNA データは Rfam[3] から tRNA ファミリ配列を100本抽出し，それを正例とする．負例として，正例サンプルの配列をランダムシャッフルした配列を同数用意する．これに対して 10-fold クロスバリデーションを行った．以下，実験結果を示していくが，各表では各実験に対して識別精度 (acc.) と適合率 (pre.)，再現率 (rec.)，計算時間 (sec.) を示している．また，表中の計算時間は 10-fold クロスバリデーションにおける 10 回の学習，テストの計算時間の平均を示している．

文字列カーネルを tRNA に対して適用した実験の結果を，表1に示す．結果を見ると，文字列カーネルをそのまま適用したときの識別精度は 92.0% だったのに比べ，配列分割法を適用した場合は識別精度があがり，計算時間は少なくなっている．しかし，学習中にカーネル関数を呼び出す回数の違いにより，分割を増やしていくのに反比例して計算時間が少なくなっているわけではなく，分割を増やしすぎると計算時間が増えていってしまう．

ステムカーネルを tRNA に対して適用した実験の結果を表2に示す．表中の d の列の t は前章で説明した tRNA の二次構造を考慮した分割を行ったことを表している．配列分割法と組み合わせることにより，識別精度が向上し計算時間も 7% に減少している．

表 1: 文字列カーネルによる tRNA の識別

d	n	acc.	pre.	rec.	sec.
1	9	92.0	97.6	86.0	216
2	7	94.0	99.0	89.0	31
3	6	94.5	97.0	92.0	29
4	7	92.0	97.8	86.0	31
5	5	95.0	98.8	91.0	32
6	6	96.0	100	92.0	39
7	5	96.5	99.0	94.0	44
8	4	96.0	96.9	95.0	47
9	5	96.5	100.0	93.0	54
10	4	96.0	99.0	93.0	58

表 2: ステムカーネルによる tRNA の識別

d	α	λ	L	acc.	pre.	rec.	sec.
1	0.1	0.3	0	74.5	75.8	74.0	45411
t	0.0	0.8	3	91.0	96.8	85.0	3218

K_{RNA} を tRNA に対して適用した実験の結果を表 3 に示す. 表 2 と同様に表中の d の列の t は前章で説明した tRNA の二次構造を考慮した分割を行ったことを表している. tRNA の二次構造を考慮した分割を行った場合には配列分割法と組み合わせることによってより高い識別精度を得ることができている. しかしながら, 等分割では精度は落ちてしまうがそれほど識別精度は落ちないのは文字列としての相同性があるためと考えられる.

また, 表 1, 表 2, 表 3 を比較すると K^N_{RNA} と tRNA の二次構造を考慮した分割を組み合わせた場合が最も識別率がよく, 計算時間も短かった.

5S rRNA に対しても Rfam から配列を 100 本をとり tRNA と同様に負例を 100 本正例をランダムシャッフルすることにより作成し、実験を行った. その結果を表 4 に示した. 表中で d のところに 5S と表記しているのは 5SrRNA 用の分割を行って実験を行ったという意味である. この表にみられるように, tRNA だけではなく他

表 3: $K^N_R NA$ による tRNA の識別

	N	acc.	pre.	rec.	sec.
1	23	90.5	94.5	86.0	216
t	16	97.0	98.0	96.0	9
8	8	86.5	88.2	85.0	30
4	16	86.5	95.3	77.0	31

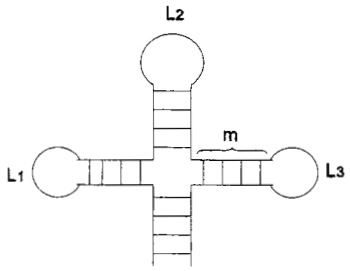

図 7: 擬似 tRNA

の二次構造をとる 5SrRNA に対しても配列分割法を適用することにより短い計算時間で高い精度を得ることができた.

また, 実データだけでなく文字列としては相同性が低いが二次構造は同じであるような配列の識別を行った. 実験結果を表 7 に示した. tRNA と似た構造を持つ擬似 tRNA の正例を以下のようにして作成した. 今まで正例として使用してきた 100 本の tRNA と同じ長さの分布の 100 本の配列とする. そして, 長さ n の擬似 tRNA を相補対をとる部分の長さをすべて $m = \lfloor n/8 \rfloor$, 各ループを $L_i = \lfloor (x-8m)i/3 \rfloor - \lfloor (x-8m)(i-1)/3 \rfloor$ としている (図 7 参照). 上記の条件のもとでランダムに配列を生成した. 負例は正例をランダムシャッフルして作成をした. K^N_{RNA} を使用した場合には高い識別率を示しているが, 等分割のように分割が正例の二次構造に対応していないときには実データを使用したときよりも顕著に識別率が落ちてしまう. これは実データと違い文字列の相同性を排除しているので, 擬似 tRNA を作成するのに利用した相補対以外の対になる塩基を利用して識別するのは難しいことを示している. また, 文字列カーネルによるこの擬似

表 4: K^N_{RNA} による 5SrRNA の識別

d	N	acc.	pre.	rec.	sec.
1	29	92.0	100.0	84.0	524
5S	14	96.0	98.1	94.0	24

表 5: KRNA による擬似 tRNA の識別

d	N	acc.	pre.	rec.	sec.
1	21	97.0	98.0	96.0	251
t	16	100.0	100.0	100.0	8
8	9	54.0	55.0	46.0	36
4	15	49.0	50.4	46.0	41

tRNA の識別率は 66.0%であったことを併記しておく.

6 まとめ

本稿では，RNA 識別に対して二次構造を考慮した文法に基づく内包カーネル関数 K_{RNA} と配列分割法を組み合わせた手法を提案し，その有用性を示した．RNA の二次構造が既知であるときには内包カーネルと配列分割法と組み合わせて適用した手法の有用性を示すことができた．しかし，構造が未知の場合，RNA の全長が与えられていない場合などにどう対処するかが今後の課題である．

謝辞

本研究の一部は日本学術振興会科学研究費補助金基盤研究 (B)19300046 の援助を受けている．

参考文献

[1] K. Doi, T. Yamashita, and A. Yamamoto. An efficient algorithm for computing kernel function defined with anti-unification. In *Proceedings of the 16th International Conference on Inductive Logic Programming (ILP2006), Revised Selected Papers (LNAI 4455)*, pages 139–153, 2007.

[2] S. R. Eddy. Non-coding RNA genes and the modern RNA world. *Nature Reviews Genetics*, 2(12):919–929, 2001.

[3] S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy, and A. Bateman. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research*, 33:D121–D124, 2005.

[4] T. Joachims. Making large-scale support vector machine learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods: Support Vector Machines*, pages 169–184. MIT Press, 1998.

[5] T. Kin, K. Tsuda, and K. Asai. Marginalized kernels for RNA sequence data analysis. *Genome Informatics*, 13:112–122, 2002.

[6] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. *The Journal of Machine Learning Research*, 2:419–444, 2002.

[7] Y. Sakakibara, K. Popendorf, N. Ogawa, K. Asai, and K. Sato. Stem kernels for RNA sequence analyses. *Journal of Bioinformatics and Computational Biology*, 5(5):1103–1122, 2007.

[8] H. Sankoh, K. Doi, and A. Yamamoto. An intentional kernel function for RNA classification. In *Proceedings of the 10th International Conference on Discovery Science (DS2007)*, pages 281–285, 2007.

[9] 三功 浩嗣, 土井 晃一郎, and 山本 章博. RNA 配列を識別するための内包カーネル関数の設計. 人工知能学会研究会資料 *SIG-FPAI A701-10*, pages 59–64, 2007.