2009－BIO－16 (10)
2009／3／6

ペア確率多重文脈自由文法による シュードノットつきRNA2次構造予測

田中 翔† 加藤 有己‡ 関 浩之†

†奈良先端科学技術大学院大学 情報科学研究科
‡京都大学 化学研究所 バイオインフォマティクスセンター

シュードノットを含む RNA の 2 次構造予測に対するアプローチとして，文脈自由文法 (CFG) より表現能力の高い形式文法 (MCFG, TAG 等) の構文解析アルゴリズムに基づく手法が提案されている．また，汎用性と精度の向上を目指し，複数の 1 次構造同士の比較解析に基づく 2 次構造予測法もいくつか提案されている．本稿では，比較解析ができるように MCFG を拡張したペア確率多重文脈自由文法 (Pair-SMCFG) を新たに定義し，これに基づく RNA の 2 次構造予測法を提案する．長さ 70 程度の RNA 配列に対して 2 次構造予測を行ったところ，RNA の特定のファミリーに対する文法の特化を全く行わないという条件下であっても，適合率 63.2%，再現率 62.0% という結果を得た．

Prediction of RNA Secondary Structure with Pseudoknots Using Pair Stochastic Multiple Context-Free Grammar

Sho Tanaka†, Yuki Kato‡ and Hiroyuki Seki†

†Graduate School of Information Science, Nara Institute of Science and Technology
‡Bioinformatics Center, Institute for Chemical Research, Kyoto University

Several methods for the prediction of RNA secondary structure including pseudoknots have been proposed based on parsing algorithms for formal grammars such as MCFG and TAG, of which generative power is greater than CFG. Also, comparative sequence analysis, which compares several RNAs and predicts their secondary structures, is a promissing approach. In this paper, we define pair-stochastic multiple context-free grammar (Pair-SMCFG) and propose a prediction method based on Pair-SMCFG. Pair-SMCFG is an extension of MCFG for comparative sequence analysis. Experimental results show that for RNA which have about 70 bases, the precision and recall of our algorithm are 63.2% and 62.0% respectively.

1 はじめに

RNA の 2 次構造予測問題に対する 1 つの有力なアプローチとして，形式文法における構文解析問題として定式化するという手法があり，特に文脈自由文法 (context-free grammar, CFG) を用いた手法が種々提案されてきた [4]．しかし CFG ではいくつかの塩基対が交差して現れるシュードノット (pseudoknot) 構造を表現できないため，CFG より生成能力の大きい文法に基づく予測法がいくつか提案されている．

また比較解析に基づくアプローチでは，2 本 (以上) の 1 次構造をお互いに比較しながら 2 次構造を予測していく．同時に 2 本の塩基配列情報を利用できるため，高い予測精度を期待できる．QRNA 法 [12] では，まず既存手法で 1 次構造の多重アラインメントを求め，その結果を利用して 2 本の RNA の 2 次構造を同時に予測する．しかし，QRNA 法では予測に Pair-SCFG と呼ばれる CFG に基づく文法を用いているため，シュードノットを含む 2 次構造予測は行えなかった．

本研究では，QRNA 法を拡張してシュードノットを含む RNA2 次構造予測のための比較解析法を提案する．以下の理由から，加藤らの SMCFG(確率多重文脈自由文法：stochastic multiple context-free grammar)[8] を用いた．(1) CFG の自然な拡張として簡潔に定義されており数学的性質も種々解明されている [7]．(2) 構文解析のための CYK 型アルゴリズムが提案されており [7][14]，RNA2 次構造予測に有用であることが実証されている [8][9]．

本研究ではまず，SMCFG に基づき，2 本の RNA の比較解析を行うための文法 Pair-SMCFG を定義した．Pair-SMCFG を用いて QRNA 法を拡張し，シュードノットを含む場合にも適用可能とした．提案法に基づく予測アルゴリズムを実装し実験を行ったところ，塩基数が約 70 の Corona_pk3 ファミリーに対して文法の特化を行わなくても適合率 63.2%，再現率 62.0% という精度を得た．

関連研究 まず 1 本の 1 次構造から 2 次構造を予測する手法として，[2, 8, 11, 13, 16] が挙げられる．n を入力 RNA の塩基数 (配列長) とすると時間計算量は単純シュードノット [1] の場合 $O(n^4)$, 再帰的な構造を考慮した場合 $O(n^5)$ である．ただしこれらの手法で高い精度を得るためには，文法を解析対象の RNA ファミリーに特化させる必要がある．次に，2 本の RNA を直接比較解析してそれらの 2 次構造を求める手法として，Dynalign[3] を拡張した [15] があるが実装は行われていない．この手法は精度は高いが時間計算量が単純シュードノットの場合でも $O(n^8)$ であり，実用性に乏しい．最後に，本研究は

表 1: Pair-SMCFG G_P.

グループ名	規則グループ	関数	遷移確率	出力確率
E	$W_v \to (\varepsilon, \varepsilon)$		1	1
S	$W_v \to J[W_y]$	$J[(x_1, x_2)] = x_1 x_2$	$t_v(y)$	1
D	$W_v \to SK[W_y]$	$SK[(x_1, x_2)] = (x_1, x_2)$	$t_v(y)$	1
$\mathrm{U_{1L}}$	$W_v \to UP_{1L}^{\binom{a_i}{b_i}}[W_y]$	$UP_{1L}^{\binom{a_i}{b_i}}[(x_1, x_2)] = (\binom{a_i}{b_i} x_1, x_2)$	$t_v(y)$	$P(a_i, b_i \mid t)$
$\mathrm{U_{1R}}$	$W_v \to UP_{1R}^{\binom{a_j}{b_j}}[W_y]$	$UP_{1R}^{\binom{a_j}{b_j}}[(x_1, x_2)] = (x_1 \binom{a_j}{b_j}, x_2)$	$t_v(y)$	$P(a_j, b_j \mid t)$
$\mathrm{U_{2L}}$	$W_v \to UP_{2L}^{\binom{a_k}{b_k}}[W_y]$	$UP_{2L}^{\binom{a_k}{b_k}}[(x_1, x_2)] = (x_1, \binom{a_k}{b_k} x_2)$	$t_v(y)$	$P(a_k, b_k \mid t)$
$\mathrm{U_{2R}}$	$W_v \to UP_{2R}^{\binom{a_l}{b_l}}[W_y]$	$UP_{2R}^{\binom{a_l}{b_l}}[(x_1, x_2)] = (x_1, x_2 \binom{a_l}{b_l})$	$t_v(y)$	$P(a_l, b_l \mid t)$
$\mathrm{P_L}$	$W_v \to BPL^{\binom{a_i a_j}{b_i b_j}}[W_y]$	$BPL^{\binom{a_i a_j}{b_i b_j}}[(x_1, x_2)] = (\binom{a_i}{b_i} x_1 \binom{a_j}{b_j}, x_2)$	$t_v(y)$	$P(a_i a_j b_i b_j \mid t)$
$\mathrm{P_R}$	$W_v \to BPR^{\binom{a_k a_l}{b_k b_l}}[W_y]$	$BPR^{\binom{a_k a_l}{b_k b_l}}[(x_1, x_2)] = (x_1, \binom{a_k}{b_k} x_2 \binom{a_l}{b_l})$	$t_v(y)$	$P(a_k a_l b_k b_l \mid t)$
$\mathrm{P_C}$	$W_v \to BPC^{\binom{a_i a_l}{b_i b_l}}[W_y]$	$BPC^{\binom{a_i a_l}{b_i b_l}}[(x_1, x_2)] = (\binom{a_i}{b_i} x_1, x_2 \binom{a_l}{b_l})$	$t_v(y)$	$P(a_i a_l b_i b_l \mid t)$

1 次構造の多重アラインメントを前提とした比較解析法である上述の QRNA 法 [12] を拡張したものである．1 本の解析の場合と計算量オーダは同一である一方，1 次構造の多重アラインメントで発生した誤りが最終結果に伝播するリスクがあるものの，精度は直接比較解析に匹敵する．

2 ペア確率多重文脈自由文法

2.1 SMCFG

確率多重文脈自由文法 (SMCFG) は，5 項組 $G = (N, T, F, P, S)$ からなる．N, T, F, P はそれぞれ非終端記号，終端記号，関数，生成規則の集合，$S \in N$ は開始記号である．各非終端記号 $A \in N$ には，正の整数 $\dim(A)$ が割り当てられる．これは A が終端記号列の $\dim(A)$ 項組を生成することを表す．ただし，$\dim(S)$=1 である．各 $f \in F$ は，$(T^*)^{d_1} \times \cdots \times (T^*)^{d_k}$ から $(T^*)^{d_0}$ への関数 (ただし各 $1 \le i \le k$ について d_i はある正の整数) である．f の関数値の各成分は，引数の成分と定系列との連接で表される．

P に含まれる生成規則は $A_0 \xrightarrow{p} f[A_1, \ldots, A_k]$ という形式をとる．ここで，$A_i \in N$ $(0 \le i \le k)$，$f : (T^*)^{\dim(A_1)} \times \cdots \times (T^*)^{\dim(A_k)} \to (T^*)^{\dim(A_0)} \in F$，$p$ は $0 < p \le 1$ の実数であり，この生成規則の適用確率と呼ばれる．左辺が同一の非終端記号である生成規則の確率の合計は 1 とする．規則 $A_0 \to f[\,]$ において，定数関数 f が $f[\,] = (\beta_1, \cdots, \beta_{\dim(A_0)})$ と定義されているとき，この規則を単純に $A_0 \to (\beta_1, \ldots, \beta_{\dim(A_0)})$ と表記する．

導出木を次のように定義する．

(D1) もし $A \xrightarrow{p} \overline{\alpha} \in P$ $(\overline{\alpha} \in (T^*)^{\dim(A)})$ ならば，根のラベルが A で $\overline{\alpha}$ を唯一の子として持つ木は，確率 p で $\overline{\alpha}$ を出力する導出木である．

(D2) もし $A \xrightarrow{p} f[A_1, \ldots, A_k] \in P$ で，根が $A_1, \ldots, A_k$ でラベル付けされた $t_1, \ldots, t_k$ が確率 $p_1, \ldots, p_k$ で $\overline{\alpha_1}, \ldots, \overline{\alpha_k}$ を出力する導出木であるならば，根のラベルが A で $t_1, \ldots, t_k$ を左から順に部分木としてもつ木は，確率 $p \cdot \prod_{i=1}^{k} p_i$ で $f[\overline{\alpha_1}, \ldots, \overline{\alpha_k}]$ を出力する導出木である．

根が $A \in N$ でラベル付けされ，$\overline{\alpha} \in (T^*)^{\dim(A)}$ を出力する導出木の確率の合計が q $(0 < q \le 1)$ であるとき，導出 $A \overset{*}{\Rightarrow} \overline{\alpha}$ の確率は q であるという．SMCFG によって生成される言語 G は，$L(G) = \{w \in T^* \mid$ ある $q > 0$ が存在して，導出 $S \overset{*}{\Rightarrow} w$ の確率は $q\}$ と定義される．

2.2 Pair-SMCFG

Pair-SMCFG は，終端記号集合が，ある記号集合 T の直積 $T \times T (= T^2)$ であるような SMCFG$(N, T \times T, F, P, S)$ である

あらかじめ 1 次構造のアラインメントが行われた 2 本の RNA の 2 次構造を表現する Pair-SMCFG $G_P = (N, T \times T, F, P, S)$ を表 1 に示す．$N = \{W_0, W_1, \cdots, W_m\}$ であり，W_v, W_y の添字は $0 \le v < y \le m$ を満たす整数である．開始記号 S を左辺にもつ規則は $S \to J[W_0]$ のみとする．また，$T = \{u, c, g, a, *\}$(4 種の塩基とギャップに対応) である．以降，$T \times T$ の要素を，$\binom{u}{u}, \binom{u}{c}, \binom{c}{*}$ 等と書く．さらに，$\binom{a_1}{b_1}\binom{a_2}{b_2}\cdots\binom{a_n}{b_n}$ を $\binom{a_1 a_2 \cdots a_n}{b_1 b_2 \cdots b_n}$ と略記する．

表 1 で，UP に添字のついた関数を UP と総称する．BPL，BPR，BPC を BP と総称する．関数 UP において，$\binom{a_i}{b_i} \in (T \times T) \backslash \{\binom{*}{*}\}$(つまり，ギャップ同士の対以外)，$BP$ において，$(\binom{a_i}{b_i}, \binom{a_j}{b_j}) \in \{($

$\binom{x_L}{y_L},\binom{x_R}{y_R}\ )|x_Lx_R, y_Ly_R \in \{ug, ua, cg, gu, gc, au\}\}$ とする．UP は対をなさない塩基単体を出力するものであり，BP は塩基対を出力するものである．例えば RNA 配列 $\binom{agauuu*}{ccuggaa}$ は，次のような手順で出力される．簡単のため，S 以外の非終端記号は A のみとする．

$A \overset{*}{\Rightarrow} BPC^{\binom{au}{ua}}[(\varepsilon,\varepsilon)] = (\ \binom{a}{u},\binom{u}{a}\ )$

$A \overset{*}{\Rightarrow} BPL^{\binom{gu}{cg}}[(\ \binom{a}{u},\binom{u}{a}\ )] = (\ \binom{gau}{cug},\binom{u}{a}\ )$

$A \overset{*}{\Rightarrow} BPL^{\binom{au}{cg}}[(\ \binom{gau}{cug},\binom{u}{a}\ )] = (\ \binom{agauu}{ccugg},\binom{u}{a}\ )$

$A \overset{*}{\Rightarrow} UP^{\binom{*}{a}}_{2R}[(\ \binom{agauu}{ccugg},\binom{u}{a}\ )] = (\ \binom{agauu}{ccugg},\binom{u*}{aa}\ )$

$S \overset{*}{\Rightarrow} J[(\ \binom{agauu}{ccugg},\binom{u*}{aa}\ )] = \binom{agauuu*}{ccuggaa}$

この導出の表す 2 次構造を図 1 に示す．

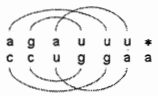

図 1: RNA 配列 $\binom{agauuu*}{ccuggaa}$ の 2 次構造

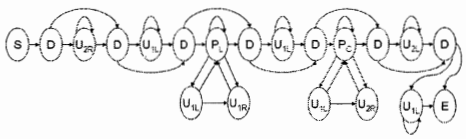

図 2: G_P の「遷移図」

次に，表1の規則の左辺，右辺の非終端記号 W_v, W_y は図 2 に基づいて定める．図 2 の頂点が左から非終端記号 $W_0, W_1, \cdots$ に対応する．頂点 W_v から W_y に有向辺があり，W_v のラベルが例えば U_{1L} であるとき，規則グループ $W_v \rightarrow UP^{\binom{a_i}{b_i}}_{1L}[W_y]$ が存在する．図 2 の自己ループは，左辺と右辺の非終端記号が同一である規則 (自己再帰規則) が存在することを意味する．このような規則はしばしば計算量の増大を招く．そこで自己再帰規則を連続する有限個の非再帰規則で近似する．これにより，1 つの非終端記号当たりの繰り返し回数は限定されるが，実行時間，必要なメモリ量共に劇的に減少する．

2.3 適用確率

ある規則 $W_v \rightarrow f[W_y]$ の適用確率は HMM との類比によって，W_v から W_y への遷移確率と f の定義式右辺に現れる終端記号 (の組) の出力確率の積で表すことにより，確率の算出を見通しよく行う．表1の出力確率 $P(x, y|t)$ は 2 本の RNA の位置の t において 1 本目，2 本目にそれぞれ塩基 $x, y \in \{u, c, g, a\}$ が現れる確率を表し，

$$P(x,y|t) = \frac{1}{3} \sum_{(x',x'',y',y'')\in\{U,C,G,A\}} \{ P^{COD}(xx'x'', yy'y''|t) + P^{COD}(x'xx'', y'yy''|t) + P^{COD}(x'x''x, y'y''y|t)\}$$

で計算される [12]．ここで，$P^{COD}(x_1x_2x_3, y_1y_2y_3|t)$ は，連続する 3 つの位置 t において 2 本の RNA の 1 本目，2 本目にそれぞれコドン $x_1x_2x_3$ および $y_1y_2y_3$ が現れる確率であり，次式で算出できる．

$$P^{COD}(x_1x_2x_3, y_1y_2y_3|t) \simeq \sum_{A_x,A_y} P(x_1x_2x_3|A_x) P(y_1y_2y_3|A_y)P(A_x, A_y|t)$$

ここで，$P(A_x, A_y|t)$ は，アミノ酸 A_x，A_y の同時確率であり，BLOSUM62[6] で定めた．また，$P(x_1x_2x_3|A_x)$ はコドン頻度であり，CodonViewer[1] で定めた．次に表 1 の出力確率 $P(x_Lx_Ry_Ly_R|t)$ は，2 本の RNA の位置の対 t において，1 本目，2 本目にそれぞれ塩基対 x_Lx_R, y_Ly_R が現れる確率を表し，

$$P(x_Lx_Ry_Ly_R|t) \simeq P^{pair}(y_Ly_R|t)P^{pair}(x_Lx_R|t) \times\frac{1}{2}\left[\frac{P(x_L, y_L|t)}{P(y_L|t)P(x_L|t)} + \frac{P(x_R, y_R|t)}{P(y_R|t)P(x_R|T)}\right]$$

と近似できる [12]．すなわち $P(x_Lx_Ry_Ly_R|t)$ は，次の 3 種の確率から算出できる．特定の位置 t で 1 本の RNA 配列中に塩基 $x \in \{u, c, g, a\}$ が出力される確率 $P(x|t)$，特定の位置 t で 1 本の RNA 配列中に塩基対 $x_Lx_R \in \{ug, ua, cg, gu, gc, au\}$ が出力される確率 $P^{pair}(x_Lx_R|t)$，および上述の $P(x, y|t)$．前二者の確率は標本となるデータが必要なため，本研究では無作為に 20 本の tRNA を抽出した．$P(x|t)$ は，20 本の全塩基に対して，塩基 x の個数を実際に数えることによって算出した．$P^{pair}(x_Lx_R|t)$ も同様に算出した．

3 実験

実験用の RNA ファミリーは，Rfam(Ver 9.1)[5] から取得した．このデータベースには，既に多重アラインメントを施された RNA ファミリーを多数登録してある．実験用に UPSK RNA，Prion，Corona_pk3 の 3 つを選択した (表 2)．この 3 つのファミリーはそれぞれ単純シュードノット構造 [1] をとる．ただし，Prion は量が膨大なため，その中から 10 本の RNA を無作為に抽出した．

表 3: 2 次構造予測の結果

ファミリー	適合率 [%]			再現率 [%]			実行時間 [sec]		
	平均値	最小値	最大値	平均値	最小値	最大値	平均値	最小値	最大値
UPSK RNA	50.9	16.7	75.0	53.2	14.3	85.7	5.9	5.6	6.6
Prion	46.7	0.0	90.0	34.1	0.0	75.0	68.5	52.1	130.0
Corona_ pk3	63.2	0.0	100.0	62.0	0.0	100.0	752.6	611.7	962.1

表 2: 使用する 3 つの RNA ファミリー

ファミリー	RNA 配列の長さ	実験に使用した本数
UPSK RNA	23	6
Prion	57	10
Corona_pk3	67	14

2 節で提案した Pair-SMCFG G_P に対して，SMCFG の CYK 型アルゴリズム [8] に基づいて 2 次構造予測アルゴリズムを実装した．ただし出力確率は 2.3 節で説明した方法により算出した．

Pair-SMCFG の CYK 型アルゴリズムとトレースバックアルゴリズムはすべて C 言語で実装し，CPU は Intel Core 2 Extreme CPU 3.00GHz，メモリは 4.00GB RAM の計算機で実験を行った．

上述の実験データに対して得られた結果を表 3 に示す．Prion ファミリーの適合率と再現率が若干低いが，これは，本手法で算出した出力確率では考慮されていない塩基対の存在によるものであると考える[2]．この 3 つのファミリーに対して，ファミリーへの文法の特化を行わないという同様の条件下で 1 本の 1 次構造からの 2 次構造予測をおこなったところ，適合率，再現率共に 1% 以下であった．また，Corona_pk3 の場合，上位約 3 分の 2 の適合率，再現率は非常に高く，それらの編集距離は比較的大きい．以上は比較解析により複数本の情報を同時に用いることの有効性を示している．

参考文献

[1] T. Akutsu, Dynamic Programming Algorithms for RNA Secondary Structure Prediction with Pseudoknots, *Descrete Applied Mathematics*, 104, 45-62 (2000).

[2] L. Cai, R. L. Malmberg and Y. Wu, Stochastic Modeling of RNA Pseudoknotted Structures: A Grammatical Approach, *Bioinformatics*, 19(1), i66–i73 (2003).

[3] D. H. Mathews and D. H. Turner, Dynalign: An algorithm for finding the secondary structure common to two RNA sequences, *JMB*, 317, 191–203 (2002).

[4] R. Durbin, S. R. Eddy, A. Krogh and G. Mitchison, *Biological Sequence Analysis*, Cambridge University Press (1998).

[5] S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy and A. Bateman, Rfam: An RNA family database, *Nucl. Acids Res.*, 33, 121–124 (2005).

[6] S. Henikoff and J. G. Henikoff, Amino acid substitution matrices from protein blocks, *Proc. Natl. Acad. Sci. USA*, 89, 10915–10919 (1992).

[7] 嵩, 関, 藤井, 一般化文脈自由文法と多重文脈自由文法, 信学論 *D*, J71-D(5), 758–765 (1988).

[8] Y. Kato, H. Seki and T. Kasami, RNA pseudoknotted structure prediction using stochastic multiple context-free grammar, *IPSJ Trans. Bioinform.*, 47(SIG17-TBIO1), 12–21 (2006).

[9] Y. Kato, T. Akutsu and H. Seki, A grammatical approach to RNA-RNA interaction prediction, *Pattern Recognition*, 42, 531–538 (2009).

[10] H. Matsui, K. Sato and Y. Sakakibara, Pair stochastic tree adjoining grammars for aligning and predicting pseudoknot RNA structures, *Bioinformatics*, 21(11), 2611–2617 (2005).

[11] E. Rivas and S. R. Eddy, The language of RNA: A formal grammar that includes pseudoknots, *Bioinformatics*, 16(4), 334–340 (2000).

[12] E. Rivas and S. R. Eddy, Noncoding RNA gene detection using comparative sequence analysis, *BMC Bioinformatics*, 2(8) (2001).

[13] Y. Sakakibara, Pair hidden Markov models on tree structures, *Bioinformatics*, 19(1), i232–i240 (2003).

[14] H. Seki, T. Matsumura, M. Fujii and T. Kasami, On multiple context-free grammars, *Theor. Comput. Sci.*, 88, 191–229 (1991).

[15] S. Seki and S. Kobayashi, A grammatical approach to the alignment of structure-annotated strings, *IEICE Trans. Inf. Syst.*,E88-D(12), 2727–2737 (2005).

[16] Y. Uemura, S. Hasegawa, S. Kobayashi and T. Yokomori, Tree adjoining grammars for RNA structure prediction, *Theor. Comput. Sci.*, 210, 277–303 (1999).

[1] http://www.symplus.co.jp/index.php

[2] $\{ug, ua, cg, gu, gc, au\}$ 以外の塩基対には出力確率値 0 を割り当てている．