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

2007－MPS－67 (31)
2007－BIO－11 (31)
2007／12／21

相互作用RNA２次構造予測 —形式文法によるアプローチ—

加藤 有己† 阿久津 達也† 関 浩之‡

† 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
‡ 奈良先端科学技術大学院大学 情報科学研究科

概要 遺伝子発現の転写後調節に係わるRNA間の相互作用が注目されている．これまで動的計画法に基づいて相互作用するRNAの２次構造予測が行われてきたが，形式文法による手法は提案されていなかった．本稿では，多重文脈自由文法（MCFG）に基づいてRNA間相互作用をモデル化する方法を提案する．まず，MCFGの確率的拡張モデルに対して，確率最大の導出木を計算する多項式時間の構文解析アルゴリズムを与える．これはキッシングヘアピンループを含む結合２次構造予測に適用される．また，実験によって提案手法がDPに基づく既存研究と同等以上の性能を上げることを示す．

Secondary Structure Prediction of Interacting RNAs: A Grammatical Approach

Yuki Kato†, Tatsuya Akutsu† and Hiroyuki Seki‡

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

Abstract Much attention has been paid to RNA-RNA interaction involved in posttranscriptional regulation of gene expression. Although there have been a few studies on secondary structure prediction of interacting RNAs based on dynamic programming (DP) algorithm, no grammar-based approach has been proposed. This paper provides a new modeling for RNA-RNA interaction based on multiple context-free grammar (MCFG). We present a polynomial time parsing algorithm for finding the most likely derivation tree for the stochastic version of MCFG, which is applicable to joint secondary structure prediction including kissing hairpin loops. Also, tests on prediction using the proposed method have shown that our approach is comparable to an existing work based on DP.

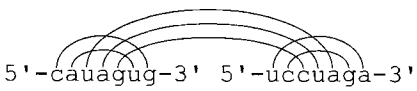
1 Introduction

In recent years, there has been much interest in antisense RNA regulation as well as RNA interference (RNAi) [4]. Antisense RNAs, which form specific structure, act via base complementarity on their target mRNAs that encode proteins of important functions, so that the translation is inhibited at the posttranscriptional level. Most of the naturally occurring RNA-RNA interactions have been found in bacteria, including CopA-CopT (antisense-target, resp.) interaction in E. coli [12]. These antisense RNAs are not fully complementary to their targets where intermolecular bonds alternate with intramolecular bonds. In particular, many loop-loop interactions have been observed, which are called *kissing hairpin loops* (see Figure 1 (a)). In this paper, we focus on this kind of *joint* structure formed by two interacting RNA molecules.

There have so far been a few studies that apply dynamic programming (DP) techniques to the joint secondary structure prediction problem based on free-energy models. Andronescu et al. [3] provide a prediction tool for interacting RNAs called PairFold, which uses Mfold [22] for pseudoknot-free structure prediction and thus cannot predict any kissing hairpin loop. Pervouchine [15] presents an extended algorithm of the Nussinov algorithm [13] to handle two interacting RNA sequences with kissing hairpins, and Alkan et al. [2] perform interaction prediction using several energy models based on the Pervouchine algorithm and its extentions.

Prediction algorithms for RNA secondary structure can be divided into two types: namely, DP algorithms based on free-energy models and DP algorithms based on parsing algorithms for formal grammars. As for the grammar-based approach, context-free grammars (CFGs) have been widely used for analyzing pseudoknot-free structure [7, 8, 17]. However, it is mathematically proved that the expressive power of CFGs is not sufficient for describing pseudoknotted structure, and several grammars whose expressive power is greater than that of

(a) A kissing hairpin loop



(b) Arc representation of (a)

Figure 1: A joint secondary structure

CFGs have been proposed [10, 16, 19] (also see [1]). The relationship between the expressive power of each of these grammars has been compared [6, 10].

This paper proposes a new method for RNA-RNA interaction prediction based on *multiple context-free grammars* (MCFGs), which are natural extensions of CFGs. The expressive power of MCFGs is known to be greater than that of CFGs. MCFGs have already been used for modeling single RNA secondary structure with pseudoknots [10, 11]. To the best of the authors' knowledge, this paper provides the first result on a formal grammar-based prediction method for interacting RNA structure prediction. The main idea of the proposed method is similar to that of the PairFold algorithm. We align given two interacting RNA sequences both in 5′-3′ directions. This pair of sequences may contain crossing pairs caused by intermolecular (*external*) bonds located between intramolecular (*internal*) bonds (see Figure 1 (b)). Then, the sequence is analyzed by a parsing algorithm of MCFG, which can recognize those crossing pairs.

The rest of this paper is organized as follows. In the next section, we formally define RNA-RNA interaction prediction problem. In sections 3 and 4, we introduce MCFGs and their subclass for modeling interacting RNAs called RNA-RNA interaction grammar (RIG) respectively. Then, we present a prediction method based on the stochastic version of RIG in the section 5. The results on some prediction tests are shown in section 6.

2 RNA-RNA Interaction Prediction Problem

We will use standard notions and notations on sequences (or strings), formal grammars and languages in this paper. Let ε denote the empty sequence. Let Σ be a finite alphabet. For a sequence $w \in \Sigma^*$, let $|w|$ denote the number of symbols appearing in w, which is called the length of w. First, we define joint secondary structure between two interacting RNA sequences. Let $\Sigma = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$.

Definition 1. (Joint secondary structure)
For two RNA sequences $a = a_1 \cdots a_n \in \Sigma^*$ ($|a| = n$) in 5′-3′ direction and $b = b_1 \cdots b_m \in \Sigma^*$ ($|b| = m$) in 3′-5′ direction, let R_α ($\alpha \in \{a, b\}$) denote a set of position pairs (i, j) in each sequence that satisfies the following conditions:

- $1 \leq i < i+1 < j \leq |\alpha|$,
- $\forall (i,j), (i',j') \in R_\alpha;\ i = i' \iff j = j'$.

Also, let R_{ab} denote a set of position pairs (k, l) between a and b that satisfies the following conditions:

- $(\exists i;\ (i,k) \in R_a$ or $(k,i) \in R_a)$ and $(\exists j;\ (j,l) \in R_b$ or $(l,j) \in R_b) \Longrightarrow (k,l) \notin R_{ab}$,
- $\forall (k,l), (k',l') \in R_{ab};\ k = k' \iff l = l'$.

Then, $R = (R_a, R_b, R_{ab})$ is called a *joint secondary structure* between a and b.

Actually, R represents a set of hydrogen-bonded base pairs such as Watson-Crick pairs (A-U and G-C pairs) and a wobble pair (G-U pair). Each pair in R_α ($\alpha \in \{a, b\}$) is said to be *internally bonded*, while each pair in R_{ab} is called *externally bonded*. RNAs are likely to fold into structure with the lowest free energy, and thus the structure prediction problem is often translated into an energy minimization problem. For simplicity of description, we assume that the score function is the number of base pairs, which is to be maximized[1], and the base pairs consist of only Watson-Crick pairs. Note that this assumption does not change the essence of the algorithm presented in this paper. Then, RNA-RNA interaction prediction problem is defined as follows:

[1] The maximization problem is equivalent to the minimization problem where the signs of scores s are inverted.

Definition 2. (RNA-RNA interaction prediction problem)

Input: two RNA sequences $a = a_1 \cdots a_n$ and $b = b_1 \cdots b_m$.

Output: a joint secondary structure $R = (R_a, R_b, R_{ab})$ between a and b that maximizes the following score:

$$\sum_{(i,j)\in R_a} s(a_i, a_j) + \sum_{(i,j)\in R_b} s(b_i, b_j) + \sum_{(k,l)\in R_{ab}} s(a_k, b_l).$$

The score function s is defined as follows:

$$s(\xi_1, \xi_2) = \begin{cases} 1 & (\{\xi_1, \xi_2\} = \{\mathrm{A}, \mathrm{U}\} \text{ or } \{\mathrm{C}, \mathrm{G}\}), \\ -\infty & (\text{otherwise}). \end{cases}$$

To make the problem simple, we impose two constraints on a joint secondary structure R. It is known that these constraints are satisfied by many example RNA-RNA complexes that have been observed.

Condition 1. (Pseudoknot free)

1. R includes no *internal pseudoknots*. That is, $\forall (i,j), (i',j') \in R_\alpha$ $(i < i', \alpha \in \{a, b\})$; $i < i' < j < j'$ does not hold.
2. R includes no *external pseudoknots*. That is, $\forall (k,l), (k',l') \in R_{ab}$ $(k < k')$; $l' < l$ does not hold.

In this paper, we propose an appropriate class of grammars that can describe joint secondary structure satisfying Condition 1. This class of grammars is a subclass of multiple context-free grammars, which are introduced in the next section.

3 Multiple Context-Free Grammar

A *multiple context-free grammar* (MCFG) [9, 18] is a 5-tuple $G = (N, T, F, P, S)$, where N is a finite set of nonterminals, T is a finite set of terminals, F is a finite set of *mcf-functions* defined below, P is a finite set of (production) rules defined below and $S \in N$ is the start symbol. For each $A \in N$, a positive integer denoted as $\dim(A)$ is given and A derives $\dim(A)$-tuples of terminal sequences. For the start symbol S, $\dim(S) = 1$. We say that f is an *mcf-function* if a nonnegative integer k and positive integers d_i $(0 \le i \le k)$ are given and f is a total function from $(T^*)^{d_1} \times \cdots \times (T^*)^{d_k}$ to $(T^*)^{d_0}$, where each component of the function value of f is defined as a concatenation of components of arguments of f and symbols in T. Each component of each argument of f can be used at most once to define the function value of f (see [9, 18] for details). Each rule in P has the form of $A_0 \to f[A_1, \ldots, A_k]$ where $A_i \in N$ $(0 \le i \le k)$ and $f : (T^*)^{\dim(A_1)} \times \cdots \times (T^*)^{\dim(A_k)} \to (T^*)^{\dim(A_0)} \in F$. If $k \ge 1$, the rule is called a nonterminating rule, and if $k = 0$, it is called a terminating rule. A terminating rule $A_0 \to f[\,]$ with $f^{[h]}[\,] = \beta_h$ $(1 \le h \le \dim(A_0))$ is simply written as $A_0 \to (\beta_1, \ldots, \beta_{\dim(A_0)})$.

We recursively define the relation $\stackrel{*}{\Rightarrow}$ by the following (L1) and (L2):

(L1) If $A \to \overline{\alpha} \in P$ $(\overline{\alpha} \in (T^*)^{\dim(A)})$, we write $A \stackrel{*}{\Rightarrow} \overline{\alpha}$.

(L2) If $A \to f[A_1, \ldots, A_k] \in P$ and $A_i \stackrel{*}{\Rightarrow} \overline{\alpha_i}$ $(1 \le i \le k)$, we write $A \stackrel{*}{\Rightarrow} f[\overline{\alpha_1}, \ldots, \overline{\alpha_k}]$.

Let $G = (N, T, F, P, S)$ be an MCFG. For $A \in N$, the set generated from A in G is defined as $L_A(G) = \{\overline{w} \in (T^*)^{\dim(A)} \mid A \stackrel{*}{\Rightarrow} \overline{w}\}$ and the language generated by G is defined as $L(G) = L_S(G)$.

Example 1. Let $G_1 = (N_1, T_1, F_1, P_1, S)$ be an MCFG, where $N_1 = \{S, A\}$, $T_1 = \{a, b\}$ and $P_1 = \{S \to J[A],\ A \to f_a[A] \mid f_b[A] \mid (\varepsilon, \varepsilon)\}$ where $\dim(S) = 1$, $\dim(A) = 2$, $J[(x_1, x_2)] = x_1 x_2$ and $f_\alpha[(x_1, x_2)] = (\alpha x_1, \alpha x_2)$ with $\alpha = a, b$. By (L1), $A \stackrel{*}{\Rightarrow} (\varepsilon, \varepsilon)$ since $A \to (\varepsilon, \varepsilon) \in P_1$. Since $f_a[(\varepsilon, \varepsilon)] = (a, a)$ and $f_b[(a, a)] = (ba, ba)$, we have $A \stackrel{*}{\Rightarrow} (a, a)$ and $A \stackrel{*}{\Rightarrow} (ba, ba)$ by (L2). Also by $S \to J[A]$, $S \stackrel{*}{\Rightarrow} J[(ba, ba)] = baba$. In fact, $L_A(G_1) = \{(w, w) \mid w \in \{a, b\}^*\}$ and $L(G_1) = \{ww \mid w \in \{a, b\}^*\}$.

Table 1: Production rules of SRIG

Rule set	Function	Transition prob.	Emission prob.
$A_v \to (\varepsilon, \varepsilon)$		1	1
$A_v \to J[A_y]^*$	$J[(x_1, x_2)] = x_1 x_2$	$t_v(y)$	1
$A_v \to SB_{1L}^{a_i}[A_y]$	$SB_{1L}^{a_i}[(x_1, x_2)] = (a_i x_1, x_2)$	$t_v(y)$	$e_v(a_i)$
$A_v \to SB_{1R}^{a_j}[A_y]$	$SB_{1R}^{a_j}[(x_1, x_2)] = (x_1 a_j, x_2)$	$t_v(y)$	$e_v(a_j)$
$A_v \to SB_{2L}^{b_k}[A_y]$	$SB_{2L}^{b_k}[(x_1, x_2)] = (x_1, b_k x_2)$	$t_v(y)$	$e_v(b_k)$
$A_v \to SB_{2R}^{b_l}[A_y]$	$SB_{2R}^{b_l}[(x_1, x_2)] = (x_1, x_2 b_l)$	$t_v(y)$	$e_v(b_l)$
$A_v \to IB_1^{a_i a_j}[A_y]$	$IB_1^{a_i a_j}[(x_1, x_2)] = (a_i x_1 a_j, x_2)$	$t_v(y)$	$e_v(a_i, a_j)$
$A_v \to IB_2^{b_k b_l}[A_y]$	$IB_2^{b_k b_l}[(x_1, x_2)] = (x_1, b_k x_2 b_l)$	$t_v(y)$	$e_v(b_k, b_l)$
$A_v \to EB^{a_i b_l}[A_y]$	$EB^{a_i b_l}[(x_1, x_2)] = (a_i x_1, x_2 b_l)$	$t_v(y)$	$e_v(a_i, b_l)$
$A_v \to W[A_y, A_z]$	$W[(x_{11}, x_{12}), (x_{21}, x_{22})] = (x_{11} x_{21}, x_{22} x_{12})$	$t_v(y, z)$	1

*Actually, this rule is excluded for the parsing algorithm.

A *stochastic* MCFG (SMCFG) $G = (N, T, F, P, S)$ is a probabilistic extension of MCFG. Each rule in P of an SMCFG has the form of $A_0 \xrightarrow{p} f[A_1, \ldots, A_k]$ where p is a real number with $0 < p \leq 1$ called the *probability* of this rule. The summation of the probabilities of the rules with the same left-hand side should be one. We define derivation trees for SMCFG as follows:

(S1) If $A \xrightarrow{p} \overline{\alpha} \in P$ ($\overline{\alpha} \in (T^*)^{\dim(A)}$), then the tree with a single node labeled $A : \overline{\alpha}$ is a derivation tree for $\overline{\alpha}$ with probability p.

(S2) If $A \xrightarrow{p} f[A_1, \ldots, A_k] \in P$ and $t_1, \ldots, t_k$ with the roots labeled $A_1, \ldots, A_k$ are derivation trees for $\overline{\alpha_1}, \ldots, \overline{\alpha_k}$ with probabilities $p_1, \ldots, p_k$, respectively, then the ordered tree with the root labeled $A : f$ that has $t_1, \ldots, t_k$ as (immediate) subtrees from left to right is a derivation tree for $f[\overline{\alpha_1}, \ldots, \overline{\alpha_k}]$ with probability $p \cdot \prod_{i=1}^{k} p_i$.

For $A \in N$, $\overline{\alpha} \in (T^*)^{\dim(A)}$ and q $(0 < q \leq 1)$, we write $A \stackrel{*}{\Rightarrow} \overline{\alpha}$ with probability q if q is the summation of the probabilities of derivation trees for $\overline{\alpha}$ with the root labeled A. The language generated by an SMCFG G is defined as $L(G) = \{w \in T^* \mid S \stackrel{*}{\Rightarrow} w \text{ with probability greater than } 0\}$.

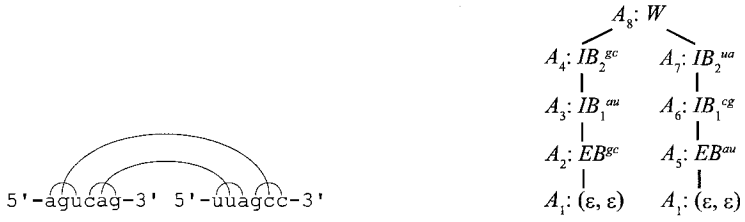
4 RNA-RNA Interaction Grammar

We introduce a subclass of MCFGs for modeling RNA-RNA interaction, which we call *RNA-RNA interaction grammars* (RIGs). The rules and functions of an RIG are shown in Table 1. The functions have mnemonic names, where SB, IB, EB and W stand for *single base*, *internal bond*, *external bond* and *wrapping* respectively. Note that a subscript of SB such as $1L$ indicates the position where a terminal symbol is concatenated. For example, $SB_{1L}^{a_i}$ denote that a_i is concatenated on the left end of the first component of the argument. An RIG has M different nonterminals denoted by $A_1, \ldots, A_M$, each of which uses the only one type of function. Let type(v) denote the name of a function that A_v uses. For example, we write type$(v) = SB_{1L}$ if $A_v \to SB_{1L}[A_y]$ is a rule. Exceptionally, we write type$(v) = E$ if $A_v \to (\varepsilon, \varepsilon)$ is a rule.

For $w = w_1 \cdots w_n \in \Sigma^*$, let w^{rev} denote the reverse of w, that is, $w^{\text{rev}} = w_n \cdots w_1$. We will use a particular RIG G such that $R = (R_a, R_b, R_{ab})$ is a joint secondary structure if and only if $A_v \stackrel{*}{\Rightarrow}_G (a, b^{\text{rev}})$ for a specific nonterminal symbol A_v in G. That is, if a pair of sequences (a, b) constitutes a joint secondary structure R, G derives a pair of sequences (a, b^{rev}) where both of the components a and b^{rev} are arranged in $5'$-$3'$ direction, and vice versa.

The following observations are important for designing the rules of RIG G:

- Each rule that uses a function SB (with appropriate super/subscripts) in Table 1 derives a single base. Rules that use functions IB and EB derive an internal bond and an external bond respectively. Precisely, IB_1 and IB_2 construct internal bonds in the first sequence and the second sequence (both in $5'$-$3'$ direction) respectively.



(a) A joint secondary structure

(b) The derivation tree

Figure 2: A joint secondary structure and its corresponding derivation tree

- Condition 1.1 states that there is no internal pseudoknot, which implies that functions SB and IB are sufficient for deriving secondary structures within each one of two sequences.
- Condition 1.2 states that there is no external pseudoknot, which implies that function EB is sufficient for deriving external bonds and function W is sufficient for concatenating two pairs of sequences $(a_1, {b_1}^{\mathrm{rev}})$ and $(a_2, {b_2}^{\mathrm{rev}})$, resulting $(a_1a_2, {b_2}^{\mathrm{rev}}{b_1}^{\mathrm{rev}}) = (a_1a_2, (b_1b_2)^{\mathrm{rev}})$.

Therefore, the rules in Table 1 are sufficient for deriving joint secondary structures that satisfy Condition 1. Although a mathematical proof for this claim is not difficult, it is omitted here due to limitation of the space, which will be given elsewhere.

Example 2. The pair of RNA sequences shown in Figure 2 (a) can be generated by RIG rules shown below:

$$A_1 \to (\varepsilon, \varepsilon), \quad A_2 \to EB^{gc}[A_1], \quad A_3 \to IB_1^{au}[A_2], \quad A_4 \to IB_2^{gc}[A_3],$$
$$A_5 \to EB^{au}[A_1], \quad A_6 \to IB_1^{cg}[A_5], \quad A_7 \to IB_2^{ua}[A_6], \quad A_8 \to W[A_4, A_7].$$

Note that the wrapping function W is applied as follows: after $A_4 \stackrel{*}{\Rightarrow} (agu, gcc)$ and $A_7 \stackrel{*}{\Rightarrow} (cag, uua)$, $A_8 \stackrel{*}{\Rightarrow} W[(agu, gcc), (cag, uua)] = (agucag, uuagcc)$. The derivation tree (the tree defined by (S1) and (S2), without probabilities) for the pair of sequences $(agucag, uuagcc)$ is shown in Figure 2 (b).

We extend RIG to a probabilistic model called stochastic RIG (SRIG) in order to predict RNA-RNA interaction. For each rule r of an RIG, two real values called *transition probability* p_1 and *emission probability* p_2 are specified as shown in Table 1. The probability of r is simply defined as $p_1 \cdot p_2$. In application, $p_1 = t_v(y)$ and $p_2 = e_v(a_i)$, etc. in Table 1 are parameters for the grammar, which are set by hand or by a training algorithm.

5 Prediction Algorithm

Let $G = (N, T, F, P, S)$ be an SRIG and (a, b) be a pair of input RNA sequences where $a = a_1 \cdots a_n$ ($|a| = n$) in 5′-3′ direction and $b = b_1 \cdots b_m$ ($|b| = m$) in 3′-5′ direction. Let $c = b^{\mathrm{rev}}$ for simplicity. The basic idea for RNA-RNA interaction prediction based on G is that we calculate the most likely derivation tree for (a, c). The most likely parse can be done by a CYK-style parsing algorithm described below. Let $\gamma_v(i, j; k, l)$ be the maximum log probability of a derivation subtree rooted at a nonterminal A_v for a pair of terminal subsequences $(a_i \cdots a_j, c_k \cdots c_l)$. The variable $\gamma_v(i, i-1; k, k-1)$ is defined as the maximum log probability for a pair of ε. We let $\mathcal{C}(v) = \{y \mid A_v \to f[A_y] \in P,\ f \in F\}$. For notational convenience, let Δ_v^{1L}, Δ_v^{1R}, Δ_v^{2L} and Δ_v^{2R} be the number of symbols generated to the left and right of the first component and to the left and right of the second component by A_v respectively (see Table 2). Also, we will use $e_v(a_i, a_j, c_k, c_l)$ for all emission probabilities. The parsing algorithm uses a five dimensional DP matrix to calculate γ, which leads to $\log P((a, c), \hat{\pi} \mid \theta)$ where $\hat{\pi}$ is the most likely derivation tree and θ is an entire set of probability parameters. The detailed description of the algorithm is shown in Figure 3. When the calculation terminates, we obtain $\log P((a, c), \hat{\pi} \mid \theta) = \gamma_v(1, n; 1, m)$, where A_v generates (a, c) with the highest probability. The time and space complexities of the algorithm are $O(M^2n^3m^3)$ and $O(Mn^2m^2)$ respectively. Note that if M is small or does not depend on input sequences, these complexities are $O(n^3m^3)$ in time and $O(n^2m^2)$ in space respectively. The optimal derivation tree can be constructed by a simple traceback procedure.

Table 2: The number of symbols generated by nonterminals

Type	SB_{1L}	SB_{1R}	SB_{2L}	SB_{2R}	IB_1	IB_2	EB
Δ_v^{1L}	1	0	0	0	1	0	1
Δ_v^{1R}	0	1	0	0	1	0	0
Δ_v^{2L}	0	0	1	0	0	1	0
Δ_v^{2R}	0	0	0	1	0	1	1

Initialization: **for** $i = 1$ **to** $n+1$; $k = 1$ **to** $m+1$; $v = 1$ **to** M:

$$\gamma_v(i, i-1; k, k-1) = \begin{cases} 0 & (\text{type}(v) = E), \\ -\infty & (\text{otherwise}). \end{cases}$$

Recursion: **for** $i = n$ **downto** 1; $j = i-1$ **to** n; $k = m$ **downto** 1; $l = k-1$ **to** m; $v = 1$ **to** M:

$$\gamma_v(i,j;k,l) = \begin{cases} -\infty & (\text{type}(v) = E), \\ -\infty & (\text{type}(v) = SB_{1L}, SB_{1R};\ j = i-1), \\ -\infty & (\text{type}(v) = SB_{2L}, SB_{2R};\ l = k-1), \\ -\infty & (\text{type}(v) = IB_1;\ j \le i+1), \\ -\infty & (\text{type}(v) = IB_2;\ l \le k+1), \\ -\infty & (\text{type}(v) = EB;\ j = i-1, l = k-1), \\ \max\limits_{y \in C(v)} \max\limits_{z \in C(v)} \max\limits_{i-1 \le p \le j} \max\limits_{k-1 \le q \le l} [\gamma_y(i,p;q+1,l) + \gamma_z(p+1,j;k,q) + \log t_v(y,z)] \\ \quad (\text{type}(v) = W), \\ \max\limits_{y \in C(v)} [\gamma_y(i + \Delta_v^{1L}, j - \Delta_v^{1R}; k + \Delta_v^{2L}, l - \Delta_v^{2R}) + \log t_v(y) + \log e_v(a_i, a_j, c_k, c_l)] \\ \quad (\text{otherwise}). \end{cases}$$

Figure 3: The parsing algorithm for SRIG

6 Experimental Results

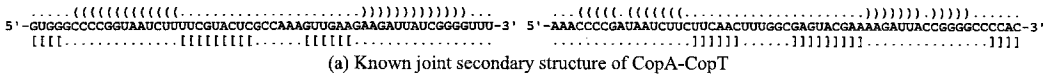
We performed tests on RNA-RNA interaction prediction using the parsing algorithm for G_{R2}. Pairs of RNA sequences taken as inputs for prediction are Tar-Tar* [5], DIS-DIS [14] and CopA-CopT [12], which are known to have kissing hairpin loop structures. In the experiments, we used a grammar model named "energy-based model (EBM), where rules were determined by taking global structure of kissing hairpin loop into consideration, and (transition) probabilities of rules for generating stacking base pairs were set by incorporating stacking energies at 37°C [21]. Note that this model can be regarded as a "generic" model in terms of ability to generate arbitrary kissing hairpin loop. We implemented the parsing algorithm in Java (version 1.6.0_02) on a machine with Dual-Core Intel Xeon processor 5160 3.00 GHz and 5.00 GB RAM. Since the parsing algorithm requires huge memory space due to the higher order DP matrix, we implemented the matrix as a hash table that stores only finite values of the log probabilities.

To evaluate prediction accuracy, we measured the sensitivity and specificity, which are the ratio of the number of correctly predicted base pairs to the total number of base pairs in the reference structure, and the ratio of the number of correctly predicted base pairs to the total number of predicted base pairs respectively. Prediction results are shown in Table 3 and Figure 4. In the figure, the upper parentheses denote internal bonds and the lower square brackets denote external bonds, and underlined base pairs indicate that they agree with correct base pairs.

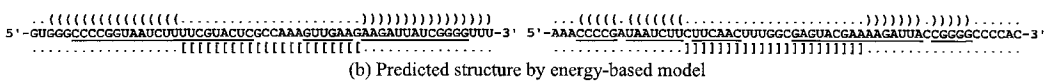
In addition, we constructed another grammar model named "profile-based model (PBM). The rules of PBM were determined by utilizing reference joint secondary structure, and the probabilities were estimated by the Laplace's rule. We compared the prediction accuracy of our models with that of three models named base pair energy model (BPEM), stacked pair energy model (SPEM) and loop energy model (LEM) presented in [2] for CopA-CopT complex (see Table 4). We calculated the F-measure F, which is the harmonic mean of sensitivity x and specificity y defined by $F = \frac{2xy}{x+y}$. As Table 4 shows, EBM is at least comparable to these three models in the same test set. Note that the number of sequences used for determining the consensus structure in PBM is only one. The experiment on PBM was performed to show that PBM achieves very high performance when

Table 3: Prediction accuracy for joint secondary structures

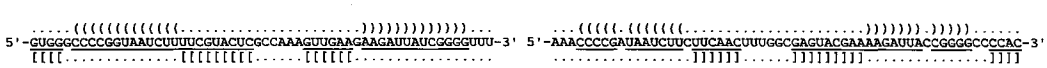
RNA-RNA complex (Reference)	n	m	Sensitivity [%]	Specificity [%]	CPU time [sec]
Tar-Tar* (Figure 1 in [5])	16	16	100.00	93.33	33.65
DIS-DIS (Figure 1 in [14])	35	35	78.57	78.57	540.18
CopA-CopT (Figure 2 in [20])	56	57	90.91	80.00	1281.77



(a) Known joint secondary structure of CopA-CopT



(b) Predicted structure by energy-based model



(c) Predicted structure by profile-based model

Figure 4: Joint secondary structure of CopA-CopT

the perfect information on the consensus structure is available. In the real setting, the performance of PBM will become lower but is expected to be still high.

7 Conclusion

We introduced a new modeling for RNA-RNA interaction based on multiple context-free grammar (MCFG). We then designed a polynomial time parsing algorithm for the specific subclass of MCFGs. Furthermore, we carried out some experiments on joint secondary structure prediction. Even if an RNA-RNA complex with internal/external pseudoknots is found, the MCFG-based method can be applied to this kind of sequence because of its expressive power, which deserves explicit emphasis.

Acknowledgments

The first author thanks JSPS Research Fellowships for Young Scientists for their generous financial assistance. This work was supported in part by Grant-in-Aid for Scientific Research from JSPS.

References

[1] Akutsu, T.: Recent Advances in RNA Secondary Structure Prediction with Pseudoknots, *Current Bioinformatics*, Vol. 1, No. 2, pp. 115–129 (2006).

[2] Alkan, C., Karakoç, E., Nadeau, J. H., Şahinalp, S. C. and Zhang, K.: RNA-RNA Interaction Prediction and Antisense RNA Target Search, *J. Comp. Biol.*, Vol. 13, No. 2, pp. 267–282 (2006).

Table 4: Comparison of accuracy for CopA-CopT

Model	**EBM**	**PBM**	BPEM	SPEM	LEM
Sensitivity [%]	90.91	100.00	45.45	95.45	86.36
Specificity [%]	80.00	100.00	37.04	76.36	84.44
F-measure [%]	85.11	100.00	40.82	84.84	85.39

[3] Andronescu, M., Zhang, Z. C. and Condon, A.: Secondary Structure Prediction of Interacting RNA Molecules, *J. Mol. Biol.*, Vol. 345, pp. 987–1001 (2005).

[4] Brantl, S.: Antisense-RNA Regulation and RNA Interference, *Biochimica et Biophysica Acta*, Vol. 1575, pp. 15–25 (2002).

[5] Chang, K.-Y. and Tinoco Jr, I.: The Structure of an RNA "Kissing" Hairpin Complex of the HIV TAR Hairpin Loop and its Complement, *J. Mol. Biol.*, Vol. 269, pp. 52–66 (1997).

[6] Condon, A., Davy, B., Rastegari, B., Zhao, S. and Tarrant, F.: Classifying RNA Pseudoknotted Structures, *Theor. Comp. Sci.*, Vol. 320, pp. 35–50 (2004).

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

[8] Eddy, S. R. and Durbin, R.: RNA Sequence Analysis Using Covariance Models, *Nuc. Acids Res.*, Vol. 22, No. 11, pp. 2079–2088 (1994).

[9] Kasami, T., Seki, H. and Fujii, M.: Generalized Context-Free Grammar and Multiple Context-Free Grammar, *IEICE Trans. Inf. & Syst.*, Vol. J71-D, No. 5, pp. 758–765 (1988) (in Japanese).

[10] Kato, Y., Seki, H. and Kasami, T.: On the Generative Power of Grammars for RNA Secondary Structure, *IEICE Trans. Inf. & Syst.*, Vol. E88-D, No. 1, pp. 53–64 (2005).

[11] Kato, Y., Seki, H. and Kasami, T.: RNA Pseudoknotted Structure Prediction Using Stochastic Multiple Context-Free Grammar, *IPSJ Trans. Bioinformatics*, Vol. 47, No. SIG17 (TBIO1), pp. 12–21 (2006).

[12] Kolb, F. A., Malmgren, C., Westhof, E., Ehresmann, C. Ehresmann, B., Wagner, E. G. and Romby, P.: An Unusual Structure Formed by Antisense-Target RNA Binding Involves an Extended Kissing Complex with a Four-Way Junction and a Side-by-Side Helical Alignment, *RNA*, Vol. 6, pp. 311–324 (2000).

[13] Nussinov, R., Pieczenik, G., Griggs, J. R. and Kleitman, D. J.: Algorithms for Loop Matchings, *SIAM Journal of Applied Mathematics*, Vol. 35, no. 1, pp. 68–82 (1978).

[14] Paillart, J.-C., Skripkin, E., Ehresmann, B., Ehresmann, C. and Marquet, R.: A Loop-Loop "Kissing" Complex is the Essential Part of the Dimer Linkage of Genomic HIV-1 RNA, *PNAS*, Vol. 93, pp. 5572–5577 (1996).

[15] Pervouchine, D. D.: IRIS: Intermolecular RNA Interaction Search, *Proc. 15th Intl. Conf. Genome Informatics (GIW2004)*, pp. 92–101 (2004).

[16] Rivas, E. and Eddy, S. R.: The Language of RNA: A Formal Grammar that Includes Pseudoknots, *Bioinformatics*, Vol. 16, No. 4, pp. 334–340 (2000).

[17] Sakakibara, Y., Brown, M., Hughey, R., Mian, I. S., Sjölander, K., Underwood, R. C. and Haussler, D.: Stochastic Context-Free Grammars for tRNA Modeling, *Nuc. Acids Res.*, Vol. 22, No. 23, pp. 5112–5120 (1994).

[18] Seki, H., Matsumura, T., Fujii, M. and Kasami, T.: On Multiple Context-Free Grammars, *Theor. Comp. Sci.*, Vol. 88, pp. 191–229 (1991).

[19] Uemura, Y., Hasegawa, A. Kobayashi, S. and Yokomori, T.: Tree Adjoining Grammars for RNA Structure Prediction, *Theor. Comp. Sci.*, Vol. 210, pp. 277–303 (1999).

[20] Wagner, E. G. H. and Flardh, K.: Antisense RNAs Everywhere?, *TRENDS in Genetics*, Vol. 18, No. 5, pp. 223–226 (2002).

[21] Zuker, M., Mathews, D. H. and Turner, D. H.: Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide, in *RNA Biochemistry and Biotechnology*, eds. Barciszewski, J. and Clark, B. F. C., Kluwer Academic Publishers (1999).

[22] Zuker, M.: Mfold Web Server for Nucleic Acid Folding and Hybridization Prediction, *Nuc. Acids Res.*, Vol. 31, No. 13, pp. 3406–3415 (2003).