



## Stochastic modeling of RNA pseudoknotted structures: a grammatical approach

Liming Cai<sup>1,\*</sup>, Russell L. Malmberg<sup>2</sup> and Yunzhou Wu<sup>1</sup>

<sup>1</sup>Department of Computer Science and <sup>2</sup>Department of Plant Biology, The University of Georgia, Athens, Georgia 30602, USA

Received on January 6, 2003; accepted on February 20, 2003

### ABSTRACT

**Motivation:** Modeling RNA pseudoknotted structures remains challenging. Methods have previously been developed to model RNA stem-loops successfully using stochastic context-free grammars (SCFG) adapted from computational linguistics; however, the additional complexity of pseudoknots has made modeling them more difficult. Formally a context-sensitive grammar is required, which would impose a large increase in complexity.

**Results:** We introduce a new grammar modeling approach for RNA pseudoknotted structures based on parallel communicating grammar systems (PCGS). Our new approach can specify pseudoknotted structures, while avoiding context-sensitive rules, using a single CFG synchronized with a number of regular grammars. Technically, the stochastic version of the grammar model can be as simple as an SCFG. As with SCFG, the new approach permits automatic generation of a single-RNA structure prediction algorithm for each specified pseudoknotted structure model. This approach also makes it possible to develop full probabilistic models of pseudoknotted structures to allow the prediction of consensus structures by comparative analysis and structural homology recognition in database searches.

**Availability:** Prototypes for the automated pseudoknot prediction algorithm are available upon request.

**Contact:** cai@cs.uga.edu; russell@plantbio.uga.edu

### INTRODUCTION

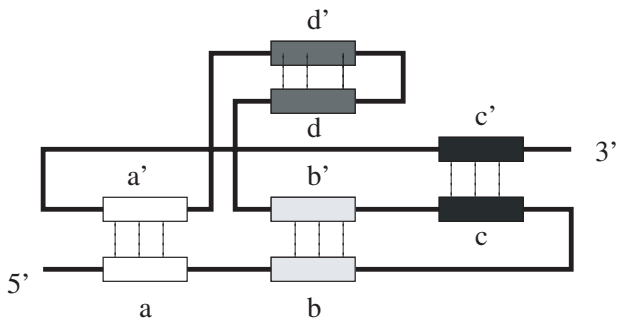
The structure of an RNA molecule is determined by interactions between pairs of nucleotides separated by short or long distances in the molecule. Such interactions can fold a sequence up into shapes such as stem-loops or more complicated ones called pseudoknots. Pseudoknots are functionally important in a number of RNAs, with roles in translation (Felden *et al.*, 2001; Zweib *et al.*, 1999), viral genome structure (Paillart *et al.*, 2002), and ribozyme active sites (Tanaka *et al.*, 2002). Computational determination of RNA structure from primary

sequence data includes a number of related problems: single-sequence structure prediction, multiple-sequence consensus structure prediction, and structural homology recognition in database searches (Durbin *et al.*, 1998).

Chomsky grammar systems from computational linguistics are ideal to model interactions of nucleotides in RNA molecules (Searls, 1992, 1999; Chomsky, 1956). Chomsky context-free rewriting rules (with probabilities) can be used to specify nested correlations of residues in an RNA sequence and therefore how the sequence is structurally formed. For RNA stem-loops these stochastic context-free grammars (SCFG) have provided a general modeling approach that permits effective construction of algorithmic solutions to the RNA structure determination problems (Brown, 2000; Durbin *et al.*, 1998; Eddy and Durbin, 1994; Holmes and Rubin, 2002; Knudsen and Hein, 1999; Sakakibara *et al.*, 1994). Thermodynamic approaches have also been used successfully (Zuker and Stiegler, 1981; Zuker, 1989). In particular, with SCFG modeling of stem-loops, not only can a single-sequence prediction algorithm be automatically generated for each specified stem-loop model, but also profiles can be developed for multiple alignment, consensus structure prediction, and structural homology recognition in database searches (Eddy and Durbin, 1994; Durbin *et al.*, 1998). In contrast, lacking general modeling approaches, most bioinformatic approaches to RNA pseudoknot structures have so far focused on specialized algorithms for solving the single-sequence structure prediction problem (Akutsu, 2000; Cary and Stormo, 1995; Lyngso and Pedersen, 2000; Tabaska *et al.*, 1998).

Ordinary RNA stem-loops can be viewed as nesting patterns of base-paired nucleotides; however, RNA pseudoknots require crossing patterns of base-pairings that cannot be modeled by SCFG (see Fig. 1). Formally pseudoknots require Chomsky context-sensitive grammars, which are much more complex than CFG, as well as being clumsy to implement. There have been some attempts to model pseudoknots using grammar systems in the spirit of SCFG. Brown and Wilson (1995) introduced pseudoknot modeling using an intersection of separate

\*To whom correspondence should be addressed.



**Fig. 1.** The 2nd pseudoknot in the consensus structure of tmRNAs from the  $\beta$ -proteobacteria.

SCFGs that can describe two crossing stems forming a pseudoknot. The intersecting grammar approach is heuristic and does not ensure optimality. Uemura *et al.* (1999) use tree adjoining grammars for pseudoknot modeling which appears to be too complicated to implement. Rivas and Eddy (2000), after devising a thermodynamic-based dynamic programming algorithm for predicting pseudoknots (Rivas and Eddy, 1999), presented a formal grammar to describe the legal structures identified by their prediction algorithm. This grammar is based on a number of auxiliary symbols used to reorder the strings generated by an otherwise context-free grammar. As described in (Rivas and Eddy, 2000), the grammar makes it possible to develop full probabilistic models of pseudoknots. The rules of reordering used in the grammar may be more difficult to implement than an SCFG.

In this paper, we introduce a new grammar modeling approach for RNA pseudoknots based on our previous work (Cai, 1995, 1996) on *parallel communicating grammar systems* (PCGSs). Introduced by Paun and Santean (1990), such a grammar system consists of a number of Chomsky grammars (called *components*) that rewrite in parallel and synchronously according to the communicating protocol; one component can query sequences produced by others and several components can make queries at the same time. PCGSs whose components are of a certain type are provably more powerful than a single Chomsky grammar of the same type. Theoretically, context-sensitive structures such as a pseudoknot can be generated by one CFG synchronized with a number of regular grammar components that describe a *crossing* double helix formed by two distant potential base pairing regions. Technically, the stochastic version of the parallel grammar models can be as simple as an SCFG with the CFG having special query symbols as nonterminals for potential base pairing regions that may form crossing helices.

In particular, our model allows simple specifications of

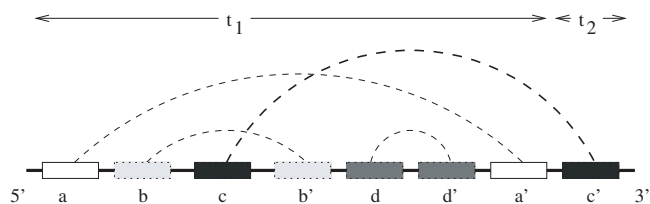
RNA pseudoknot models in almost the same way as an SCFG does for stem-loops. This leads to an automated dynamic programming algorithm for pseudoknotted structure prediction. Moreover, this model can take advantage of existing techniques and results developed for SCFG; potentially it can provide feasible solutions to training models and profiling pseudoknots for multiple alignment, consensus structure prediction, and structural homology recognition. PCGS for modeling pseudoknots may be more straightforward to implement than the previously discussed models of pseudoknots.

The paper is organized as follows. In the section **Parallel grammars to specify pseudoknots**, we describe how to specify pseudoknotted structures based on PCGS. In section **The stochastic form of the model**, we show technically the stochastic form of our grammar model can be as simple as an SCFG. The section **An automated algorithm for pseudoknot prediction** presents the automated dynamic programming-based algorithm for pseudoknotted structure prediction. Some preliminary testing results with the algorithm are given in section **Implementation and preliminary tests**.

## PARALLEL GRAMMARS TO SPECIFY PSEUDOKNOTS

Informally, a *parallel communicating grammar system* (Paun and Santean, 1990)  $\Gamma$  consists of more than one Chomsky grammars ( $G_0, G_1, \dots, G_k$ ) (called *components*, one of which,  $G_0$ , is called the *master*). These grammars can share an alphabet and a set of nonterminals. Additionally, there are some special nonterminals, called *query symbols*, for communication between the grammars. The derivation of the system is the rewriting of every grammar component in parallel and synchronously. Synchronization between the rewritings of the components is achieved by communications through queries. For instance, if query symbol  $Q_j$  is present in the string  $\omega_i$  produced by component  $G_i$ ,  $Q_j$  in  $\omega_i$  will be replaced in the next step by the string  $\omega_j$  currently produced by component  $G_j$ . Querying a string has priority over component-wise parallel derivations, provided the string being queried does not further contain query symbols nor nonterminal symbols that are *not derivable* in the component making the query. The communication protocol allows each grammar to return to its start symbol after being queried. The language produced by the system is the set of strings produced by the master grammar component (Paun and Santean, 1990; Cai, 1995).

To see how PCGS can be used to specify pseudoknots, we need to define some terminologies. Given an RNA sequence, a *base region* is a subsequence in the sequence. Two *base-paired* regions are two base regions that form a double helix. In such a case, each base region is said to



**Fig. 2.** As a *pseudoknotted structure* defined by Definition 3 for the pseudoknot in Figure 1, with  $c$  in  $t_1$  and  $c'$  in  $t_2$  being two potential regions.

contribute to the helix.

**DEFINITION 1.** Given an RNA sequence  $s$  and a subsequence  $t$  of  $s$ , a *potential region* in  $t$  is a base region such that (1) it contributes to some helix in  $s$ ; but (2) it does not contribute to any helix in  $t$ .

**DEFINITION 2.** Let  $s$  be an RNA sequence and  $t$  be a subsequence of  $s$ . Then  $t$  is a *P-structure* if it contains a potential region in it.  $t$  is called a *non-trivial P-structure* if  $t$  contains a potential region that lies *between* two base-paired regions in  $t$ .

**DEFINITION 3.** Let  $s$  be an RNA sequence.  $s$  is a *pseudoknotted structure* if  $s$  contains two non-overlapped P-structures  $t_1$  and  $t_2$  such that (1) either  $t_1$  or  $t_2$  is a non-trivial P-structure, and (2) the two potential regions contained (respectively) within  $t_1$  and  $t_2$  form a double helix (called a *crossing helix*).

As currently implemented, a P-structure itself cannot be pseudoknotted structure. The implemented pseudoknotted structures based on this kind of P-structure cover most of RNA pseudoknotted structures in the literature. For example, Figure 1 shows one of the pseudoknots contained in the consensus structure of tmRNAs from the  $\beta$ -proteobacteria family studied by Felden *et al.* (2001) and Figure 2 describes this pseudoknot based on Definition 3. To include more complex pseudoknotted structures, Definition 3 actually allows a P-structure to be pseudoknotted structures or equivalently, to contain more than one potential region. In particular, by induction on the number of potential region contained in a P-structure, it is not hard to observe the following:

**THEOREM 1.** *Definition 3 defines exactly all RNA pseudoknotted structures.*

Our canonical definition for pseudoknotted structures provides just a slightly different view on pseudoknots from those in (Akutsu, 2000; Brown and Wilson, 1995; Lyngso and Pedersen, 2000; Rivas and Eddy, 1999;

**Table 1.** Auxiliary grammar components specifying two base-paired regions

$G_1: S_1 \rightarrow Q_2$	$G_2: S_2 \rightarrow T$	$G_3: S_3 \rightarrow A$
$T \rightarrow T_1$	$T \rightarrow Q_3$	$S_3 \rightarrow C$
$T_1 \rightarrow Q_3$	$A \rightarrow Q_3u$	$S_3 \rightarrow G$
$A \rightarrow aQ_3$	$C \rightarrow Q_3g$	$S_3 \rightarrow U$
$C \rightarrow cQ_3$	$G \rightarrow Q_3c$	$S_3 \rightarrow H$
$G \rightarrow gQ_3$	$U \rightarrow Q_3a$	
$U \rightarrow uQ_3$		

**Table 2.** Parallel derivations of base-paired regions  $acg$  and  $cgu$

$S_1 \Rightarrow Q_2$	$S_2 \Rightarrow T$	$S_3 \Rightarrow A$
$\Rightarrow T$	$\Rightarrow S_2$	$\Rightarrow A$
$\Rightarrow T_1$	$\Rightarrow T$	$\Rightarrow A$
$\Rightarrow Q_3$	$\Rightarrow Q_3$	$\Rightarrow A$
$\Rightarrow A$	$\Rightarrow A$	$\Rightarrow S_3$
$\Rightarrow aQ_3$	$\Rightarrow Q_3u$	$\Rightarrow C$
$\Rightarrow aC$	$\Rightarrow Cu$	$\Rightarrow S_3$
$\Rightarrow acQ_3$	$\Rightarrow Q_3gu$	$\Rightarrow G$
$\Rightarrow acG$	$\Rightarrow Ggu$	$\Rightarrow S_3$
$\Rightarrow acgQ_3$	$\Rightarrow Q_3cgu$	$\Rightarrow H$
$\Rightarrow acgH$	$\Rightarrow Hcgu$	$\Rightarrow S_3$

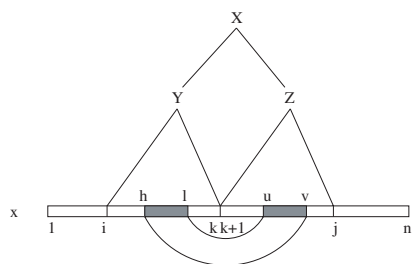
Uemura *et al.*, 1999). Nevertheless, it facilitates grammar description of pseudoknots and makes it convenient to model pseudoknotted structures based on PCGS.

We now show a PCGS example to model a pseudoknotted structure. Table 1 shows a portion of the PCGS (containing only the auxiliary regular components  $G_1$ ,  $G_2$  and  $G_3$  without the master component  $G_0$ ). In the example, we use capital letters for nonterminals to differentiate from nucleotides  $a$ ,  $c$ ,  $g$ , and  $u$ .

Table 2 demonstrates an example of parallel derivations of two base-paired regions  $acg$  and  $cgu$ . The synchronization between  $G_1$  and  $G_2$  is accomplished by production  $S_1 \rightarrow Q_2$  because  $G_1$  has to wait until  $G_2$  starts (deriving nonterminal  $T$ ) and  $T$  is copied to the current string of  $G_1$ . After  $G_2$  is queried, it returns to the start symbol  $S_2$ . Then  $G_1$  and  $G_2$  make queries to  $G_3$  at the same time. However, the same symbol copied to  $G_1$  and to  $G_2$  invokes different derivations to generate a base and its complement, eventually producing two base-paired regions in  $G_1$  and  $G_2$ , respectively.

A pseudoknot can be produced by a context-free master component  $G_0$  in addition to the above auxiliary regular components  $G_1$ ,  $G_2$ , and  $G_3$ . Essentially,  $G_0$  describes two non-overlapped P-structures, one of which is a non-trivial P-structure. These two P-structures contain two base-paired regions queried from  $G_1$  and  $G_2$ , respectively. Table 3 shows a part of simplified productions for  $G_0$





**Fig. 4.** The bottom-up computation of the maximum probability for pseudoknotted structure  $X$  from the maximum probability  $H(h, l, u, v)$  of forming a crossing helix and the maximum probabilities for substructures (P-structures)  $Y$  and  $Z$ .

where  $X_j$  is one of the nonterminals  $A, C, G$ , and  $U$  in  $G_1$  and  $Y_j$  is the corresponding (synchronized) nonterminal in  $G_2$ . Therefore,

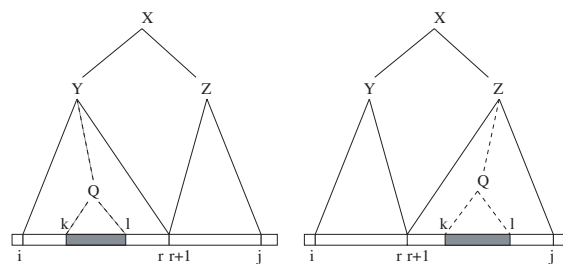
$$\begin{aligned} Pr(S_0 \Rightarrow^* s_1 r_1 s_2 r_2 s_3) \\ = Pr(S_0 \Rightarrow^* s_1 Q_1 s_2 Q_2 s_3) \prod_{j=1}^k (Pr(S_3 \rightarrow X_j) \\ Pr(X_j \rightarrow x_j Q_3 \& Y_j \rightarrow Q_3 y_j)) \end{aligned}$$

The product term of the above is actually the probability for a pairwise complementary alignment between the two base pairing regions  $r_1$  and  $r_2^{-1}$ . Therefore, the structure specification and probability computation with the PCGS pseudoknot model can be independent of the auxiliary grammar components. In other words, the stochastic version of our grammar model is simply the stochastic version of the CFG master component, as simple as an SCFG. The only difference between our stochastic grammar model and an SCFG is the query symbols used as nonterminals in specifying pseudoknots.

## AN AUTOMATED ALGORITHM FOR PSEUDOKNOT PREDICTION

As with standard SCFG models, our approach results in a system that can automatically generate a pseudoknot prediction algorithm for each specified pseudoknotted structure model. Technically, the model specification is the same as an SCFG except that query symbols can be used as nonterminals. To model the crossing helices represented by the query symbols, we require a  $5 \times 5$  probability matrix that describes the probability distribution in the crossing helices for all possible base pairs among the four nucleotides and gaps, which are included for the purpose of producing bulges.

The automated pseudoknot prediction algorithm is dynamic programming-based, resembling the CYK algorithm. Given an input sequence  $x[1..n]$ , the algo-



**Fig. 5.** The bottom-up computation of the maximum probability for P-structure  $X$  from the maximum probabilities for two substructures  $Y$  and  $Z$  (exactly one of them is a P-structure).

rithm essentially computes, for every nonterminal  $X$ , the maximum probability for every subsequence  $x[i..j]$  to admit the substructure specified by  $X$ . The algorithm distinguishes three categories of substructures: traditional stem-loops, pseudoknots, and P-structures. Some of the computation details for the algorithm are the following.

- (1) The computation for traditional stem-loops is done the same way as the CYK algorithm.
- (2) For pseudoknots, the computation is with the aid of an auxiliary function  $H$  for the maximum probability for which every pair of subsequences form a crossing double helix. ( $H$  can be computed separately by a pairwise complementary alignment algorithm). Specifically, for nonterminal  $X$ , the maximum probability for  $X$  to derive a pseudoknotted structure  $x[i..j]$  is

$$\begin{aligned} Pr(X, i, j) \\ = \max_{k, h, l, u, v, Y, Z} \{ H(h, l, u, v) Pr(Y, i, k, h, l) \\ Pr(Z, k + 1, j, u, v) Pr(X \rightarrow YZ) \} \end{aligned}$$

where  $Y$  and  $Z$  are substructures containing potential base pairing regions  $x[h..l]$  and  $x[u..v]$ , respectively.

- (3) The maximum probability for nonterminal  $X$  to derive P-structure subsequence  $x[i..j]$  containing a potential  $x[k..l]$  is defined as

$$Pr(X, i, j, k, l) = Pr(X \Rightarrow^* x[i..k-1] Q x[l+1..j])$$

Recursively

$$\begin{aligned} Pr(X, i, j, k, l) \\ = \max_{r, Y, Z} \{ \max_{r, Y, Z} \{ Pr(Y, i, r, k, l) Pr(Z, r + 1, j) \\ Pr(X \rightarrow YZ) \}, \\ \max_{r, Y, Z} \{ Pr(Y, i, r) Pr(Z, r + 1, j, k, l) \\ Pr(X \rightarrow YZ) \} \} \end{aligned}$$

The base cases are (a)  $i = j$  and  $Pr(X, i, j) = Pr(X \rightarrow x[i])$ , and (b)  $Pr(X, i, j, i, j) = 1$  for every  $X$  being a query symbol. The maximum probability for  $x$  to have the predicted structure is  $M[S_0, 1, n]$ .

As with the CYK algorithm, dynamic programming approaches can be used to compute the maximum probabilities  $Pr(X, i, j)$  and  $Pr(X, i, j, i, j)$  for every suitable nonterminal  $X$ , every subsequence  $x[i..j]$ , and every relevant potential region  $x[k..l]$  within the subsequence. Figures 4 and 5 illustrate the bottom-up computation of these probabilities.

In theory, as with most pseudoknot prediction algorithms developed by others, our algorithm has the worst-case costs  $O(n^6)$  for CPU time and  $O(n^4)$  for RAM space. We have used some implementation techniques to significantly reduce the CPU and memory consumptions by the algorithm (see next section and Wu *et al.* (2003) for details).

## IMPLEMENTATION AND PRELIMINARY TESTS

We have implemented a pseudoknot prediction system prototype in both C and C++ on the UNIX/Solaris platform of a SUN workstation with 64-bit instructions, 8 dual-processors, and 32 GB main memory and on the IRIX platform of an SGI Origin 2000 with 64 bit instructions, 16 processors, and 8 GB main memory. The input to the program is a specific structure model in the form of an SCFG (containing query nonterminals) and a  $5 \times 5$  base-pairing probability matrix for predicting pair-wise alignments (4 bases plus gaps). The ability to read an input SCFG from a file allows the program to predict structures from a range of models. The input SCFG is required to be in Chomsky Normal Form (CNF) (Chomsky, 1956) to ease the complexity of parsing. The prototype is available in both C and C++ via <http://128.192.4.4:2002>.

The implementation of dynamic programming for our model is memory and run-time intensive. The core matrix for the dynamic programming algorithm is sparsely populated, hence we have been able to improve memory allocation by use of a memory-mapping method. Similarly, we have begun work on a parallelized version of filling the critical matrices to take advantage of up to 16 processors (Wu *et al.*, 2003).

To evaluate the success of our approach, we needed a set of sequence data that was already aligned and annotated with respect to stem-loop and pseudoknot structures. The tmRNA database (<http://psyche.uthct.edu/dbs/tmRDB/tmRDB.html>) provided such a data set; tmRNA molecules have up to 4 pseudoknots in their structure. We downloaded 85 sequences that were aligned to each other and annotated. We constructed a phylogenetic tree of these sequences, then used this tree as the basis for dividing the data set in two, so that the two halves

**Table 4.** Some typical outcomes of the prediction test. Under each sequence is the correct pseudoknotted structure followed by the predicted structures. The correct predictions, as in sequence 43, plus nearly correct predictions, similar to sequence 29, totaled 69% of the total predictions

sequence 43

```
CGAGAAUGAGAGAAUCUCGUAAAACUUUC
AAAAA**QQQQQ**AAAAA*****QQQQQ
AAAAA**QQQQQ**AAAAA*****QQQQQ
```

sequence 29

```
AGUGUUGUAGACUAUAAAACUACUAGGUUUA
AAAAA**QQQQQQ**AAAAA*****QQQQQQ
AAAAA**QQQQQQ**AAAAA*****QQQQQQ
```

sequence 7

```
GUAUGAUUCCACCGUGGUUUUGCCAUAUGGAUCA
AAAAA**QQQQQQ**AAAAA*****QQQQQQ
AAAAA*****QQQQQQ**AAAAA*****QQQQQQ
```

sequence 11

```
CAUCUAGCGGAUGUAAAACGCCAGUUA
AAAAQQQQQAAAA*****QQQQQQQ
*QQQQQQQQQQ**QQQQQQQQQQ*
```

each sampled the evolutionary diversity of the data. For this study, we extracted pseudoknot 1 from each of the sequences. One set of 42 sequences was used to estimate the probabilities required for the program. We intended to use a second set of 43 sequences to evaluate the success or failure of the prediction algorithm. Within this evaluation set, however, 2 sequences were null for pseudoknot 1, and there were 5 pairs of sequences that were identical for pseudoknot 1; the total size of the evaluation set was thus 36 sequences.

Our program predicted a pseudoknot in 34 of the 36 sequences. For 7 of the 34, the algorithm correctly predicted the pseudoknot including structure and exact base-pairing regions. The remaining 27 sequences had the correct overall structure, but typically had errors in the length or position of a base-paired region. Of these 27, 18 were within 1, 2, or 3 changes of having the correct size and position of base-paired regions, in addition to overall structure. The correct plus nearly correct structures and base-pairing regions thus total  $(7 + 18)/36 = 69\%$ .

In the examples shown in Table 4, sequence 43 demonstrates an exactly correct prediction; sequence 29 shows a prediction that is structurally correct, but where the predicted base-pairing regions are either too short or too long; sequence 7 shows a prediction of a pseudoknot, but with large errors in the regions; sequence 11 demonstrates a failure to predict the structure.

Our next goal for experiment will be to turn this kind of test, training with one set of sequences then evaluating

results with another set, into more appropriate measure of prediction success, such as the sensitivity and specificity statistics used to compare gene prediction programs.

## CONCLUSION

We have introduced a new stochastic modeling for RNA pseudoknotted structures based on parallel communication grammar systems (PCGS). Our approach can specify crossing patterns of base-pairing nucleotides for RNA sequences, while avoiding context-sensitive rules, using a single context-free grammar synchronized with a number of simple regular grammars. Technically, the stochastic version of the model can be as simple as an SCFG. Compared with previous grammar models for RNA pseudoknots (Rivas and Eddy, 2000; Uemura *et al.*, 1999; Brown and Wilson, 1995), ours actually introduces only one new symbol and is very easy to implement.

Methods based on computational linguistics such as the model of Rivas and Eddy (2000) and ours can be effective in modeling RNA pseudoknotted structures. In contrast to other methods based on thermodynamics (Rivas and Eddy, 1999; Lyngso and Pedersen, 2000) and graph theoretics (Tabaska *et al.*, 1998), stochastic grammar models make it easy to describe specific RNA structures of the interest. In particular, as shown in the previous sections, our model allows the automatic generation of a single-RNA pseudoknot prediction algorithm for each specified pseudoknotted structure. Actually, applications of the stochastic grammar modeling should go beyond single-RNA structure prediction. Our approach makes it possible to develop full probabilistic models of pseudoknotted structures. In particular, stochastic grammars can be used to profile RNA sequences, providing feasible solutions to structural homology recognition in database searches and to consensus structure prediction through (semi)automated comparative analysis.

## ACKNOWLEDGEMENT

We thank the anonymous referees for helpful comments on the earlier version of this paper.

## REFERENCES

- Akutsu, T. (2000) Dynamic programming algorithms for RNA secondary prediction with pseudoknots. *Discrete Applied Mathematics*, **104**, 45–62.
- Brown, M.P. (2000) Small subunit ribosomal RNA modeling using stochastic context-free grammars. *Proc. Int. Conf. Intel. Syst. Mol. Biol.*, **56**, 57–66.
- Brown, M. and Wilson, C. (1995) RNA pseudoknot modeling using intersections of stochastic context-free grammars with applications to database search. In *Pacific Symposium on Biocomputing*. pp. 109–125.
- Cai, L. (1995) Computational complexity of PCGS with regular components. In *Proceedings of the 2nd International Conference on Development of Language Theory*. pp. 209–219.
- Cai, L. (1996) Computational complexity of linear PCGSs. *Computer and Artificial Intelligence*, **15**, 199–210.
- Cary, R.B. and Stormo, G.D. (1995) Graph-theoretic approach to RNA modeling using comparative data. In *Proceedings of International Conference on Intelligent Systems in Molecular Biology*. pp. 75–80.
- Chomsky, N. (1956) Three models for the description of languages. *IRE Transactions on Information Theory*, **2**, 113–124.
- Durbin, R., Eddy, S.R., Krogh, A. and Mitchison, J.G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.
- Eddy, S.R. and Durbin, R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, **22**, 2079–2088.
- Felden, B., Massire, C., Westhof, E., Atkins, J.F. and Gesteland, R.F. (2001) Phylogenetic analysis of tmRNA genes within a bacterial subgroup reveals a specific structural signature. *Nucleic Acids Res.*, **29**, 1602–1607.
- Holmes, I. and Rubin, D.H. (2002) Pairwise RNA structure comparison with stochastic context-free grammars. In *Pacific Symposium on Biocomputing*. pp. 191–203.
- Knudsen, B. and Hein, J. (1999) RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, **15**, 446–454.
- Lyngso, R.B. and Pedersen, C.N.S. (2000) RNA pseudoknot prediction in energy based models. *J. Comput. Biol.*, **7**, 409–428.
- Paillart, J.C., Skripkin, E., Ehresmann, B., Ehresmann, C. and Marquet, R. (2002) In vitro evidence for a long range pseudoknot in the 5'-untranslated and matrix coding regions of HIV-1 genomic RNA. *J. Biol. Chem.*, **277**, 5995–6004.
- Paun, Gh. and Santean, L. (1990) Further remarks on parallel communicating grammar systems. *Int. J. Comput. Math.*, **34**, 187–203.
- Rivas, E. and Eddy, S.R. (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, **285**, 2053–2068.
- Rivas, E. and Eddy, S.R. (2000) The language of RNA: A formal grammar that includes pseudoknots. *Bioinformatics*, **16**, 334–340.
- Sakakibara, Y., Brown, M., Hughey, R., Mian, I.S., Sjolander, K., Underwood, R.C. and Haussler, D. (1994) Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res.*, **22**, 5112–5120.
- Searls, D.B. (1992) The linguistics of DNA. *American Scientist*, **20**, 579–591.
- Searls, D.B. (1999) Formal language theory and biological macromolecules. *Series in Discrete Mathematics and Theoretical Computer Science*, **47**, 117–140.
- Tabaska, J.E., Cary, R.B., Gabow, H.N. and Stormo, G.D. (1998) An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics*, **8**, 691–699.
- Tanaka, Y., Hori, T., Tagaya, M., Sakamoto, T., Kurihara, Y., Katahira, M. and Uesugi, S. (2002) Imino proton NMR analysis of HDV ribozymes: nested double pseudoknot structure and Mg<sup>2+</sup> ion-binding site close to the catalytic core in solution. *Nucleic Acids Res.*, **30**, 766–774.

- Uemura, Y., Hasegawa, A., Kobayashi, Y. and Yokomori, T. (1999) Tree adjoining grammars for RNA structure prediction. *Theoretical Computer Science*, **210**, 277–303.
- Wu, Y., Cai, L. and Malmberg, R.L. (2003) Implementation improvements to a pseudoknot prediction algorithm, Manuscript.
- Zuker, M. and Stiegler, P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.
- Zuker, M. (1989) Computer prediction of RNA structure. *Meth. Enzymol.*, **180**, 262–288.
- Zweib, C., Wower, I. and Wower, J. (1999) Comparative sequence analysis of tmRNA. *Nucleic Acids Res.*, **27**, 2063–2071.