Modeling RNA Secondary Structures Using Tree Grammars

Satoshi KOBAYASHI
satoshi@cs.uec.ac.jp

Takashi YOKOMORI
yokomori@cs.uec.ac.jp

Department of Computer Science and Information Mathematics
The University of Electro-Communications
1-5-1, Chofugaoka, Chofu, Tokyo 182, Japan

Abstract

This paper proposes a grammatical tool, called tree adjunct grammar with tag for RNA (denoted by TAG^2_{RNA}), for representing secondary structures of RNAs, and shows some example TAG^2_{RNA} grammars for fairly complicated RNA secondary structures. We then demonstrate the appropriateness of the grammars for modeling RNA secondary structures by discussing its formal language and/or graph theoretic properties, including closure properties of TAG^2_{RNA} and graph planarity of secondary structures generated by TAG^2_{RNA}, the latter of which would provide a biologically reasonable constraint.

1 Introduction

Due to the rapid increase of biological knowledge that strongly supports that the structure of nucleotide sequences of RNAs plays an important role in the splicing process or the translation process of RNAs, it is of great interest to propose a computational model for representing the structure of RNAs or to pursue an efficient method for predicting RNA secondary structural features.

Conventionally, most methods developed so far for predicting the secondary structure of RNAs might be roughly classified into two categories : One is based upon phylogenetic comparative analysis for homologous RNAs using alignment techniques (e.g. [JOP89]) and the other is due to evaluating the optimal secondary structures with free energy folding (e.g. [JTZ90]). Rather recently, a new approach to this prediction problem has emerged in which grammatical models with probabilistic concepts are constructed during the process of analysing sample data of biological sequences ([Sak94][AM94]).

In this paper, we propose a grammatical tool, called tree adjunct grammar with tag for RNA (denoted by TAG^2_{RNA}), for representing secondary structures of RNAs, and demonstrate

小林聡、横森貴：電気通信大学情報工学科，〒 182 東京都調布市調布ケ丘 1–5–1

the appropriateness of the grammars for modeling RNA secondary structures by discussing its formal language and/or graph theoretic properties. Although our model TAG^2_{RNA} gets involved in no stochastic elements, it is closer to the third category above mentioned than others in that through the analysis of RNA sequences, a grammar TAG^2_{RNA} is built up, for modeling the secondary structures of the sample data and for predicting the secondary structures of the test data.

We formalize TAG^2_{RNA} as a *special* class of an *extended* model of TAGs (tree adjunct grammars), initially proposed by Joshi et al. ([JLT75]) to study the formal property of a natural language syntax. Generally, this tree grammar generates a tree structure that is a kind of derivation tree of a context-free grammar.

Using a biological example, we will illustrate how RNA secondary structures can be nicely modeled by our grammar, demonstrating the usefulness of TAG^2_{RNA}. Figure 1 pictures a part of the HIV-2 gag-pol region (from 1829) in RNA sequence forming a typical secondary structure. One can give the string representatin of the secondary structure as

$$w_1\,\mathtt{acccgg}\,w_2\,\mathtt{ccccuu}\,w_3\,\mathtt{gggg}\,w_4\,\mathtt{aagggg}\,w_5\,\mathtt{ccgggu}\,w_6\,\mathtt{cccc}\,w_7 \tag{1}$$

where $w_1 = \mathtt{uguccguccaaaaaauccug}$, $w_2 = \mathtt{gaa}$, $w_3 = \mathtt{ucuuc}$, $w_4 = \mathtt{cguug}$, $w_5 = \mathtt{ca}$, $w_6 = \mathtt{ucaaggcgu}$, $w_7 = \mathtt{gacugu}$. Further, let $x_1 = \mathtt{acccgg}$, $x_2 = \mathtt{ccccuu}$, $x_3 = \mathtt{gggg}$, then the complementary string of x_i is represented by $\overline{x}_i$, where we assume that $\overline{\mathtt{a}} = \mathtt{u}$, $\overline{\mathtt{u}} = \mathtt{a}$, $\overline{\mathtt{c}} = \mathtt{g}$, $\overline{\mathtt{g}} = \mathtt{c}$. (For example, $\overline{x}_1 = \mathtt{ugggcc}$.) An alternative representation of (1) is

$$w_1\,x_1\,w_2\,x_2\,w_3\,x_3\,w_4\,\overline{x}_2^R\,w_5\,\overline{x}_1^R\,w_6\,\overline{x}_3^R\,w_7 \tag{2}$$

where $\overline{x}^R$ denotes the reversal of $\overline{x}$.

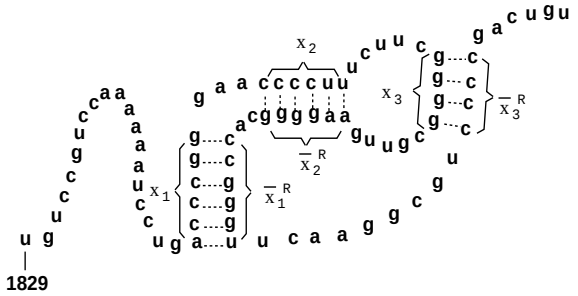

Figure 1: RNA Secondary Structure found in HIV-2 gag-pol Region

On the other hand, we can provide a generation process by a tree grammar (that we will call TAG^2_{RNA} and the details will be given later) in Figure 6 (see Appendix) where a successful derivation for the string represenation of (2) is illustrated. (Actually, the grammar given in Figure 4 can generate any structural string from a language $\{w_1\,x_1\,w_2\,x_2\,w_3\,x_3\,w_4\,\overline{x}_2^R\,w_5\,\overline{x}_1^R\,w_6\,\overline{x}_3^R\,w_7 \mid x_i, w_j \in \Sigma^+_{\mathrm{RNA}}\}$, where $\Sigma_{\mathrm{RNA}} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{u}\}$.)

Thus, one of the important features of TAG^2_{RNA} is the ability to represent *reversal crossing dependencies* as appeared in pseudoknotted structures in RNA sequences. Another interesting and important property of TAG^2_{RNA} is that any grammar in TAG^2_{RNA} always generates *planar* RNA secondary structures, which would provide a biologically reasonable restriction.

Σ^+ is the set of all finite length strings over Σ with length 1 or more.

We first introduce mathematical definitions and notations for RNA secondary structures in Section 2. Then we define an extended tree adjunct grammar, called tree adjunct grammar with tag (denoted by TAG^2), which has the ability to represent reversal crossing dependencies in Section 3. Section 4 introduces a definition of TAG^2_{RNA}, which is devised for special purpose of modeling RNA secondary structures. Section 5 presents formal language and/or graph theoretical properties of TAG^2_{RNA}, which concern closure properties of TAG^2_{RNA} and the planarity of secondary structures generated by TAG^2_{RNA}. Comparisons with some related works and the future directions of this work are also discussed.

2 Preliminaries

We first introduce a formal definition of RNA secondary structures as follows. Let $\Sigma = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$ be a finite alphabet, each of whose element represents a base in RNA sequences. For representing the Watson-Crick base pairing, we use a bar notation to represent the base complementarity, i.e., $\overline{\mathtt{a}} = \mathtt{u}$, $\overline{\mathtt{c}} = \mathtt{g}$, $\overline{\mathtt{g}} = \mathtt{c}$, $\overline{\mathtt{u}} = \mathtt{a}$.

A secondary structure T of an RNA sequence $w \in \Sigma^+$ with length k is a set of tuples (p, q) of positive integers such that $1 \leq p < q \leq k$. Each element of (p, q) in T represents a base pairing between the pth and the qth bases of w. The cardinality of a set T is denoted by $\#T$.

Let us consider a subset T_1 of T with $\#T_1 = k$, and for any positive integer i at most k, by (p_i, q_i), we denote the element of T_1 such that p_i is the ith smallest integer in $\{p \mid (p, q) \in T_1\}$. We say that (p_i, q_i) *is the ith element in* T_1. Then T_1 is said to be a *stacked pairs region of* T iff $k \geq 2$ and $1 \leq \forall i \leq k-1$ $(p_{i+1} = p_i + 1 \wedge q_{i+1} = q_i - 1)$ hold.

Let T be a secondary structure of an RNA sequence w with length k. Then we say that (p, q) and (r, s) in T are *crossing* iff either $p < r < q < s$ or $r < p < s < q$ holds. A secondary structure T is said to have *n-crossing property* iff there exists a subset T_1 of T with $\#T_1 = n \geq 2$ such that any distinct elements (p, q) and (r, s) in T_1 are crossing. Let T_1 and T_2 be stacked pairs regions of T with $\#T_1 = k \geq 2$ and $\#T_2 = m \geq 2$. By (p_i, q_i) and (r_i, s_i), we denote the ith element of T_1 and T_2, respectively. Then T_1 and T_2 are said to have *a reversal crossing dependency* iff either $p_k < r_1 < r_m < q_k < q_1 < s_m$ or $r_m < p_1 < p_k < s_m < s_1 < q_k$ holds.

Let T be a secondary structure of an RNA sequence w with length k. For any given secondary structure T of w, we consturuct a graph $G(T, w) =< V, E >$ as follows :

(1) $V = \{1, 2, ..., k\}$,

(2) $(i, j) \in E$ iff either $j = i + 1$ or $(i, j) \in T$ holds.

An RNA secondary structure T of w is said to be *planar* iff $G(T, w)$ is a planar graph.

Example 1 In Figure 1, the stacked pairs regions are denoted by pairs of variables, $(x_1, \overline{x}_1^R)$, $(x_2, \overline{x}_2^R)$, and $(x_3, \overline{x}_3^R)$. Further, the stacked regions denoted by $(x_1, \overline{x}_1^R)$ and $(x_3, \overline{x}_3^R)$ have *a reversal crossing dependency.*

3 Tree Adjunct Grammars with Tag

In this section, we introduce a grammatical device for generating trees, called *tree adjunct grammar with tag* (TAG^2, for short), which is an extension of tree asjunct grammar introduced by [JLT75]. The reason why we focus on TAG^2 is mainly due to the ability of TAG^2 to represent constraints such as reversal crossing dependencies which appears in RNA sequences.

s

Figure 2: An Example of TAG^2

Let V and Σ be alphabets such that $\Sigma \subset V$. The sets Σ and $V - \Sigma$ are called the set of *terminal symbols* and the set of *nonterminal symbols*, respectively. A terminal node of a tree is a frontier node, i.e., a leaf of the tree. A nonterminal node of a tree is a node other than the terminal nodes. The *yield* of a tree γ is the string consisting of labels of terminal nodes, and is denoted by $Y(\gamma)$. A *tree adjunct grammar with tag* (TAG^2, for short) G is a triple $G = (C, A, F)$, where F is a subset of $V - \Sigma$ (called the set of *final symbols*), and C and A are finite subsets of trees over V satisfying :

(1) If $\gamma \in C$, then $Y(\gamma) \in \Sigma^*$, the label of the root node in γ is S, and the labels of nonterminal nodes are in $V - \Sigma$.

(2) If $\gamma \in A$ and the label of the root node is T, then the labels of nonterminal nodes are in $V - \Sigma$ and $Y(\gamma)$ is in $\Sigma^* T \Sigma^* \bigcup \Sigma^* T \Sigma^*$. (Thus, the yield of γ has at least one terminal symbol.) The node whose label is correspoding to the unique T in the yield of γ is called the *foot node of* γ.

Further, elements in A and C may have some nonterminal nodes with an additional notation $*$, which are called *tagged nonterminal nodes*.

In this definition, S is called an initial symbol, and the sets C and A are called the sets of *center trees* and *adjunct trees*, respectively. The set $C \bigcup A$ is called the set of *basic trees* of G. The path from the root node to the foot node of an adjunct tree is called the *backbone* of the tree.

Let $\beta \in A$ and γ be some tree over V. Let p be a *tagged* nonterminal node in γ with the same label of the root node in β. Then we say β *is adjoinable to* γ *at* p *to produce* γ', where γ' is obatained by replacing the node p in γ by the adjunct tree β, with the subtree of γ bellow the node p to be adjoined to the foot node of β.

For any trees γ and γ', we write $\gamma \vdash \gamma'$ iff there exists $\beta \in A$ such that β is adjoinable to γ at p, for some p in γ, to produce γ'. Let $\vdash *$ be the reflective and transitive closure of $\vdash$. Then, the *tree set of* G, denoted by $\tau(G)$, is defined as the set of trees γ satisfying;

(1) there exists an $\alpha \in A$ such that $\alpha \vdash * \gamma$,

(2) all labels of tagged nonterminal nodes in γ are in F.

Finally, *a tree adjunct language L(G) by* G is defined by

$$L(G) = \{w \in \Sigma^* \mid w = Y(\gamma), \gamma \in \tau(G)\}.$$

Example 2 An example of TAG^2 grammar G_1 is presented in Figure2 (a), where $V = \{a, b, S, X\}$, $\Sigma = \{a, b\}$, $F = \{S, X\}$, and λ represents a null string. The tree adjunct language by G_1 is $\{w_1 \cdot w_2 \cdot w_1^R \cdot w_2^R \mid w_1, w_2 \in \Sigma^*\}$, which has reversal crossing dependencies. An example derivation of *ababaa* by G_1 is illustrated in Figure2 (b).

4 Specialized TAG^2 for Modeling RNA Secondary Structures

In this section, we introduce a special TAG^2, called tree adjunct grammar with tag for RNA (denoted by TAG$^2_{RNA}$), for modeling RNA secondary structures.

TAG$^2_{RNA}$ only allows five types of basic trees from TYPE1 to TYPE5 as is shown in Figure 3. TYPE1 is used as a center tree of TAG$^2_{RNA}$ and center trees of any TAG$^2_{RNA}$ grammar must be TYPE1. This tree is denoted by $T1$.

TYPE2 is used as adjunct trees for spreading out a couple of bases on both sides of the backbone. There are four classes $T2u[X, X]$, $T2u[X, Y]$, $T2d[X, X]$, $T2d[X, Y]$ of TYPE2 adjunct trees depending on whether the tagged nonterminal node is situated upward or downward of the branching point, and on whether the label of the tagged nonterminal node is the same as that of the root node or not. A lower case letter u (d) indicates that the tagged nonterminal node is located upward (downward) of the branching point. The first (second) argument in the brackets of the notation $T2*[*, *]$ tells us the label of the root (tagged nonterminal) node.

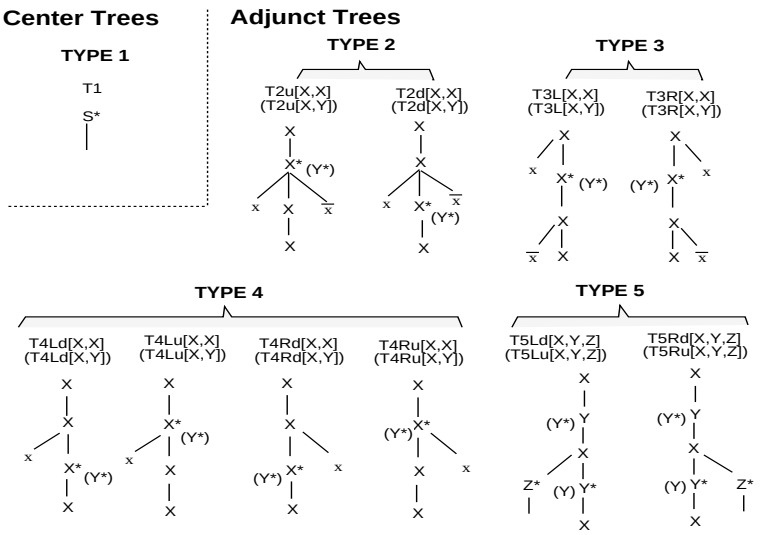

Figure 3: Types of Basic Trees Allowed in TAG$^2_{RNA}$

TYPE3 is used as adjunct trees for spreading out a couple of bases on a single side of the backbone. There are four classes $T3L[X, X]$, $T3L[X, Y]$, $T3R[X, X]$, $T3R[X, Y]$ of TYPE3 adjunct trees depending on whether both of the branches exist on the left hand side or the right hand side of the backbone, and on whether the label of the tagged nonterminal node is the same as that of the root node or not. An upper case letter L (R) represents that both of the branches are spreaded out on the left (right) hand side of the backbone. The first (second) argument in the brackets of the notation $T3*[*, *]$ tells us the label of the root (tagged nonterminal) node.

TYPE4 is used as adjunct trees for spreading out a base on a single side of the backbone. There are eight classes of TYPE4 adjunct trees depending on whether the branch exists on the left hand side or the right hand side of the backbone, on whether the tagged nonterminal node is situated upward or downward of the branching point, and on whether the label of the tagged nonterminal node is the same as that of the root node or not. An upper case letter L (R) represents that the branch is spreaded out on the left (right) hand side of the backbone. The lower case letter u (d) indicates that the tagged nonterminal node is located upward (downward) of the branching point. The first (second) argument in the brackets of the notation $T4**[*,*]$ tells us the label of the root (tagged nonterminal) node.

On the other hand, TYPE5 of adjunct trees are used for different purpose from that of the other types of adjunct trees. After adjoining this type of adjunct trees, the number of tagged nonterminal nodes is increased by one, and the additional tagged nonterminal node (Z^*) located away from the backbone is used for extending another side branch which generates other base pairings with cross dependencies. Four classes exist in TYPE5 adjunct trees depending on whether the side branch exists on the left hand side or right hand side of the backbone, and on whether the tagged nonterminal node is located upward or downward of the branching point. The upper case letters L and R , and the lower case letters u and d are used in a similar manner as in the notations defined above. The first, second, and third arguments in the brackets of the notation $T5**[*,*,*]$ denote the label of root node, that of the tagged nonterminal node on the backbone, and that of the tagged nonterminal node on the side branch, respectively.

Example 3 An example TAG^2_{RNA} grammar generating the double occurrences of reversal crossing dependencies found in HIV-2 gag-pol region in Figure 1 is shown in Figure 4. The derivation process of the grammar is also illustrated in Figure 6. (See Appendix.)

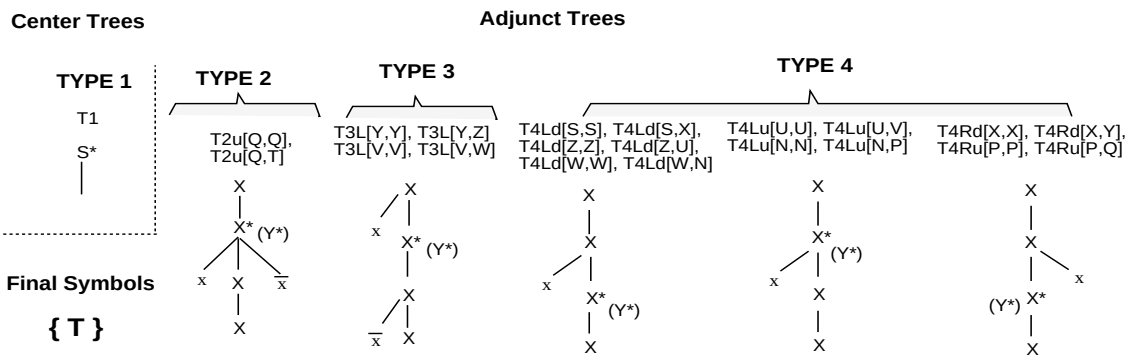

Figure 4: TAG^2_{RNA} Grammar for HIV-2 gag-pol Region shown in Figure 1

5 Basic Properties of TAG^2_{RNA}

5.1 Composition of RNAs with High Structural Complexity

In this subsection, we give a method for constructing a complicated RNA secondary structures by TAG^2_{RNA}. In the previous section, we demonstrate the ability of TAG^2_{RNA} to represent reversal crossing dependencies. Further, it is easy to see that TAG^2_{RNA} has the ability to represent other primitive secondary structures such as nested stacked pairs regions, paralel

arms, nested pseudoknots. Then, the question arises whether we can construct grammars for representing more and more complicated structures appeared in ribosomal RNAs or introns with high structural complexity.

The method for composing TAG^2_{RNA} grammars for such complicated secondary structures is based on the bottom-up construction of a target secondary structure from primitive structural features. We first decompose a target structure into pieces of primitive structural components whose grammar can be easily constructed. Then, the grammars for those primitive structures are composed into a grammar using language theoretical operation such as concatenation, insertion operations. Here, the insertion operation is defined as the operation of inserting a secondary structure T into another secondary structure T' at arbitrary position of T'. (Formal definition is omitted because of space constraint.)

The correctness of the method is shown in the following language theoretical property of TAG^2_{RNA}.

Property 1 The class of languages generated by TAG^2_{RNA} is closed under union, concatenation, insertion. □

An example construction of fairly complicated secondary structure found in group I intron from primitive secondary structures is shown in Figure 5.

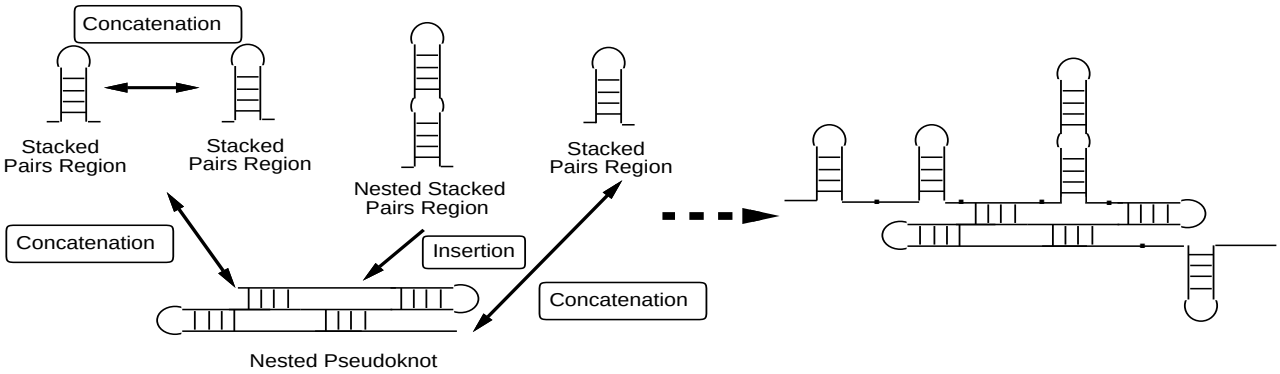

Figure 5: RNA Secondary Structure found in Group I Intron

5.2 On Planarity of RNA Secondary Structures

In this subsection, we discuss the planarity of secondary structures generated by a grammar in TAG^2_{RNA}. Following many researches on RNA secondary structures, we require that secondary structures of RNAs must be planar, which would be a boilogically and phisically reasonable constraint on the RNA secondary structure.

We begin with the next property.

Property 2 Any non-planar secondary structure has 3-crossing property.
(By the Kuratowski's characterization theorem for graph planarity. [Kur30]) □

This property gives us one of the sufficient conditions for finding out planar secondary structures. For example, pseudoknotted structures are planar since they have no 3-crossing properties. Furthermore, we have the next important result on the planarity of secondary structures generated by TAG^2_{RNA} grammar.

Property 3 Any secondary structure generated by a grammar in TAG^2_{RNA} has no 3-crossing property. □

Property 4 Any secondary structure generated by a grammar in TAG^2_{RNA} is planar.
(By Property 2 and Property 3.) □

This important property that TAG^2_{RNA} grammar always generates planar secondary structures would indicate a potential ability of TAG^2_{RNA} to be applied to RNA secondary structure prediction problem, since the use of TAG^2_{RNA} as a model of secondary structures would reduce the search space of possible structures to only planar ones.

6 Discussions

6.1 Related Works

Searl ([Sea93]) proposed a grammatical tool, called String Variable Grammar, for representing DNA or RNA sequences including its structural features, which has some relationship to our works. The grammar TAG^2_{RNA} is proposed especially for the purpose of modeling or predicting RNA secondary structures, while String Variable Grammar was proposed for modeling not only RNAs' but also DNAs' and proteins' structural features. We believe that the simplicity of the definition of TAG^2_{RNA} and its nice theoretical properties as a modeling tool for RNA secondary structures would provide us with efficient and effective methods for predicting or parsing secondary structures.

Recently Abe and Mamitsuka proposed an extended version of TAGs in stochastic configurations, called Stochastic Ranked Node Rewriting Grammars (SRNRG, for short) for the purpose of predicting protein secondary structures including β sheet regions.([AM94]) As is shown in their works, SRNRG can represent more complicated non-planar secondary structures, which can not be described with TAG^2_{RNA}. However, in spite of this weakness in representing a certain type of structural complexity, TAG^2_{RNA} has sufficient ability to represent RNA secondary structures as is shown in section 5.

A recent work on the stochastic prediction and identification of tRNAs secondary structures by Sakakibara et al. has close relationship to our works. ([Sak94]) They proposed stochastic version of context free grammars for modeling RNA secondary structures, and applied it to the prediction problem of tRNA secondary structures. However, their stochastic grammar can not represent reversal crossing dependencies, which can be elegantly captured by our TAG^2_{RNA}.

6.2 Future Works

We have just started to implement an integrated system assisting a user in solving computational problems concerning RNA secondary structures, including pattern matching, prediction and identification (classification) of RNA secondary structures.

In the current paper, we proposed an extended tree adjunct grammar, called tree adjunct grammar with tag for RNA (denoted by TAG^2_{RNA}), as a modeling language for RNAs, and showed its effectiveness with some theoretical analysis. In the developing system, TAG^2_{RNA} is used as a central model of RNA structural features.

However, there are some issues to be studied in future works. The first one is a theoretical characterization of RNA secondary structures which can be represented by TAG^2_{RNA}. What

kinds of secondary structures are beyond the descriptive capability of TAG^2_{RNA} is theoretically unclear in the current status of this work.

Further, we need to provide some convenient user interface for taking RNA secondary structures into the system. Gramatical representations of secondary structures would be a heavy task for users who are not fammiliar with the rudiments of the language theory. For this purpose, we must devise some graphical representations of secondary structures which can be automatically converted to TAGs representations.

As for the time complexity of parsing TAG^2, we have already known ([YK94]) that there exists some polynomial time parsing algorithm for TAG^2, whose proof is achieved by showing that TAL^2 is contained in the class of macro languages by Fisher that is known to be a member of the class PTIME([Fis68]). However, the proof is not a constructive one. Therefore, it remains open to be studied to find an efficient parsing algorithm for TAG^2 or TAG^2_{RNA}.

The prediction of secondary structures of RNAs or the identification of a class of RNAs using structural features are more difficult issues to be studied. However, we believe that the theoretical analysis of TAG^2_{RNA} with graph planarity property would help us overcome such difficulties.

References

[AM94] N. Abe and H. Mamitsuka. A New Method for Predicting Protein Secondary Structures Based on Stochastic Tree Grammars. In *Proc. of 11th International Conference on Machine Learning*, 1994, to appear.

[Fis68] M. J. Fisher. Grammars with macro-like productions. In *Proc. of 9th Annual IEEE Symposium on switching and Automata Theory*, pp.131-142, 1968

[JLT75] A. K. Joshi, L. S. Levy and M. Takahashi. Tree adjunct grammars. *Journal of Computer and System Sciences*, vol.10, pp.136-163, 1975

[JOP89] B. D. James, G. J. Olsen, and N. R. Pace. Phylogenetic comparative analysis of RNA secondary structure. *Methods in Enzymology*, vol.180, pp.227-239, 1989

[JTZ90] J. A. Jaeger, D. H. Turner and M. Zuker. Predicting optimal and suboptimal secondary structure for RNA. *Methods in Enzymology*, vol.183, pp.281-306, 1990

[Kur30] H. V. Kuratowski. Sur le problème des courbes gauches en topologie. *Fund. Math.*, vol.15, pp.271-283, 1930

[Sak94] Y. Sakakibara, M. Brown, R. C. Underwood, I. S. Mian and D. Haussler. Stochastic context-free grammars for modeling RNA. In *Proc. of 27th Hawaii International Conference on System Sciences*, vol.V, pp.284-293, 1994

[Sea93] D. B. Searls. The computational linguistics of biological sequences. In L. Hunter, editor, *Artificial Intelligence and Molecular biology : Chapter 2*, AAAI Press, pp. 47-120, 1993

[YK94] T. Yokomori and S. Kobayashi. DNA Evolutionary Linguistics and RNA Structure Modeling : A Computational Approach. Report CSIM 94-02, Dept. of Comput. Sci. and Inf. Math., Univ. of Elect.-Communi., 1994

Appendix

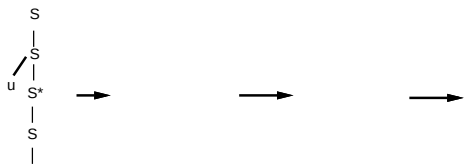

Figure 6: A Secondary Structure Generation by TAG^2_{RNA}
[Boldfaced lines indicate structures newly generated.]