

Pair Stochastic Tree Adjoining Grammars for Aligning and Predicting Pseudoknot RNA Structures

Hiroshi Matsui,

Kengo Sato,

Yasubumi Sakakibara

Keio University

Department of Biosciences and Informatics

3-14-1 Hiyoshi, Kohoku-ku, Yokohama, 223-8522, Japan

you@dna.bio.keio.ac.jp, satoken@bio.keio.ac.jp, yasu@bio.keio.ac.jp

Abstract

Motivation: Since the whole genome sequences for many species are currently available, computational predictions of RNA secondary structures and computational identifications of those non-coding RNA regions by comparative genomics become important, and require more advanced alignment methods. Recently, an approach of structural alignments for RNA sequences has been introduced to solve these problems. By structural alignments, we mean a pairwise alignment to align an unfolded RNA sequence into a folded RNA sequence of known secondary structure. Pair HMMs on tree structures (PHMMTSs) proposed by Sakakibara are efficient automata-theoretic models for structural alignments of RNA secondary structures, but are incapable of handling pseudoknots. On the other hand, tree adjoining grammars (TAGs) is a subclass of context-sensitive grammar, which is suitable for modeling pseudoknots. Our goal is to extend PHMMTSs by incorporating TAGs to be able to handle pseudoknots.

Results: We propose the pair stochastic tree adjoining grammars (PSTAGs) for modeling RNA secondary structures including pseudoknots and show the strong experimental evidences that modeling pseudoknot structures significantly improves the prediction accuracies of RNA secondary structures. First, we extend the notion of PHMMTSs defined on alignments of 'trees' to PSTAGs defined on alignments of "TAG (derivation) trees", which represent a top-down parsing process of TAGs and are functionally equivalent to derived trees of TAGs. Second, we modify PSTAGs so that it takes as input a pair of a linear sequence and a TAG tree representing a pseudoknot structure of RNA to produce a structural alignment. Then, we develop a polynomial-time algorithm for obtaining an optimal structural alignment by PSTAGs, based on dynamic programming parser. We have done several computational experiments for predicting pseudoknots by PSTAGs, and our computational experiments suggests that prediction of RNA pseudoknot struc-

tures by our method are more efficient and biologically plausible than by other conventional methods. The binary code for PSTAG method is freely available from our website at <http://www.dna.bio.keio.ac.jp/pstag/>.

1. Introduction

Secondary structures including pseudoknots of non-coding RNA molecules play important roles for their own functions such as catalytic functions [1]. Computational predictions of secondary structures from a primary RNA sequence are active research area in Bioinformatics, and further there are several theoretical or heuristic works [7, 12] to predict pseudoknot RNA structures by maximizing stacking base pairs or free energy minimizations. On the other hand, since the whole genome sequences for many species are currently available, computational identifications of non-coding RNA regions by comparative genomics become important [3, 11], and require precise local alignment or database search algorithms for detecting RNA sequences.

In this paper, we develop a novel method for structural alignments to align and predict pseudoknot RNA structures. The approach of structural alignment introduced by Sakakibara [13] is to calculate a pairwise alignment to align an unfolded RNA sequence into a folded RNA sequence of known secondary structure (as illustrated in Figure 1). An unfolded RNA sequence is going to be folded by the structural alignment into one single folded RNA sequence, and therefore the structural alignment is clearly different from the usual pairwise alignment only based on sequence homology.

Two important main features of structural alignment are (1) to predict a secondary structure for the unfolded RNA sequence and (2) to detect non-coding RNA regions with more sensitivity than simple homology. The second feature

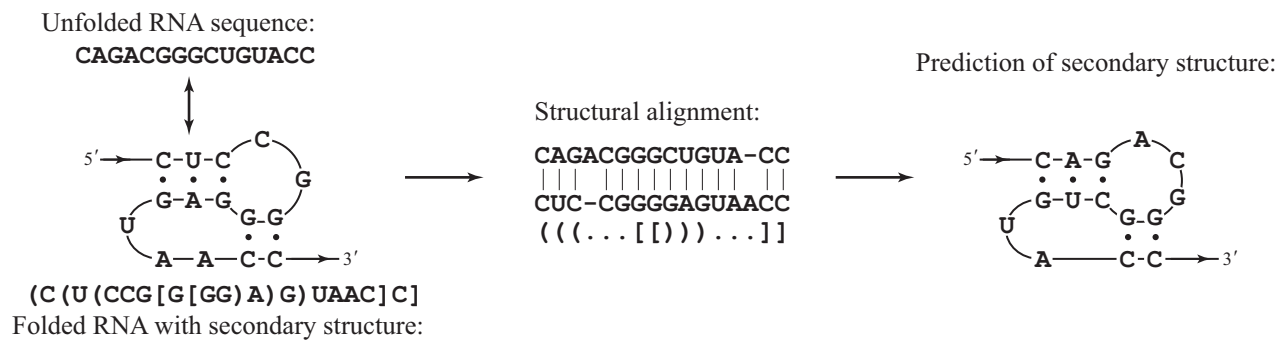


Figure 1. Structural pairwise alignment of an unfolded RNA sequence and a folded RNA.

is an obvious advantage compared with conventional methods only to predict RNA secondary structures. Recently, Sakakibara [13] has proposed *Pair HMMs on tree structures* (PHMMTSs), which is an extension of PHMMs [2] defined on alignments of trees [8], and applied PHMMTSs to the problem of structural alignments of RNA sequences. Compared with the profile stochastic context-free grammars (SCFGs) [14] for modeling RNA sequences, one advantage of structural alignments using PHMMTSs is that it does not require any training process and an enough number of sequences to estimate parameters of the grammars specific to a certain family of RNAs. This is practically important because there are in general a few annotated sequences available in most RNA families.

However, PHMMTSs are incapable of handling pseudoknots. What makes pseudoknots so difficult to be handled is the fact that modeling the pseudoknot structures of RNAs is beyond the generative power of context-free grammars, and inevitably involves in the hard complexity of context-sensitivity. In order to model pseudoknot RNA structures, we first employ special subclasses of *tree adjoining grammars* (TAGs), which have generative powers greater than context-free grammars and less than context-sensitive grammars. Uemura et al. [16] have introduced these subclasses to study pseudoknot structures. For example, Figure 2 (left) illustrates a derivation process of a TAG to produce a pseudoknot secondary structure (shown in Figure 2, right) for “(A (G [AC] U) U)”. Second, we extend the notion of PHMMTSs defined on alignments of trees to *pair stochastic TAGs* defined on alignments of “TAG (derivation) trees” which represent a pseudoknot structure of RNA and a top-down parsing process of TAGs. Third, we observe that a set of TAG trees where each node is labeled with an adjunct tree is recognized by a tree automaton, and combined with dynamic programming parser for TAGs, we reduce the problem of alignments of TAG trees to alignments by PHMMTS. Based on this theory, we develop a polynomial-time algorithm for obtaining an optimal structural alignment by PSTAGs.

Another advantage of our approach is that our method is model-based, that is, we build a rigorous formal-grammar model for aligning and predicting pseudoknot structures, while most conventional prediction methods for secondary or pseudoknot structures and alignment algorithms based on homology do not build any mathematical models.

We have tested the prediction accuracy of our PSTAG algorithm for three RNA families, Corona_pk3, HDV_ribozyme, Tombus_3_IV, which have pseudoknot annotations in Rfam [6]. The prediction performances by PSTAG are very well (more than 95% on average) and stable for all three RNA families. We compared the prediction performances of our PSTAG algorithm with the predictions of the PHMMTS method and of a standard alignment software “Clustal-W” [15]. The results clearly show that more accurate predictions can be obtained if we use grammatically more powerful methods. Further, the results of PSTAG predictions have suggested a new reliable structure which constitutes an additional internal loop in 3'-end for HDV_ribozyme.

2. Tree adjoining grammars and pseudoknot structures

For modeling pseudoknot structures of RNAs, we employ special subclasses of *tree adjoining grammars* (TAGs). Uemura et al. have first introduced these subclasses to study pseudoknot structures. Here, we briefly describe TAGs and the two subclasses, *simple linear TAGs* and *extended simple linear TAGs*. For more details, refer to the literature [16].

We first define notations for a ‘tree’ as follows. Let t be a tree defined over $V_N \cup V_T \cup \{\varepsilon\}$ where V_N and V_T are finite sets of nonterminal symbols and terminal symbols respectively, and ε is the empty sequence. The nodes in a tree t of size m are numbered from 1 to m according to the preorder where the root node is numbered 1. We denote $t(p) = A$ if $A \in V_N \cup V_T$ is the label of the node p . We define the yield of a tree t (denoted by $y(t)$) as the sequence obtained by concatenating the labels at leaf nodes of t traced from left

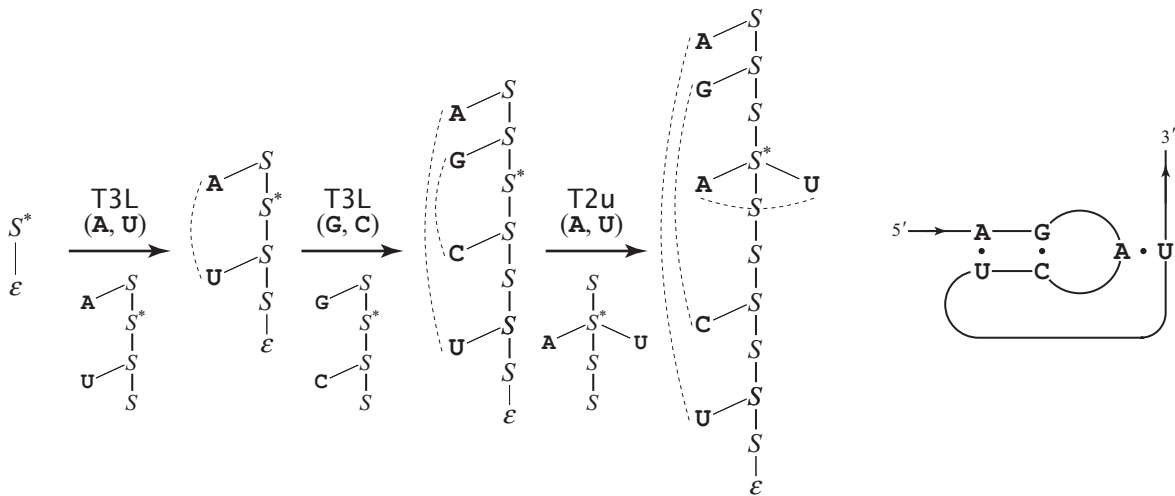


Figure 2. A derivation process to produce a typical pseudoknot structure for “(A(G[AC]U)U)”, which has crossing dependencies and cannot be modeled by any context-free grammars.

to right. Let θ denote the empty tree and λ denote the null label.

A *tree adjoining grammar* (TAG) [9] is a 5-tuple $G = (V_N, V_T, S, \mathcal{I}, \mathcal{A})$ where V_N and V_T are as mentioned above, S is the initial symbol, \mathcal{I} is a finite set of *initial trees* and \mathcal{A} is a finite set of *adjunct trees*. \mathcal{I} and \mathcal{A} satisfy the following conditions:

1. if $\alpha \in \mathcal{I}$, then $\alpha(1) = S$ and $y(\alpha) \in V_T^*$,
2. if $\beta \in \mathcal{A}$, then $\beta(1) = X \in V_N$ and $y(\beta) \in V_T^* X V_T^*$,

where V_T^* denotes the set of all finite sequences over V_T . The node whose label is $X \in V_N$ in the yield of an adjunct tree is called the *foot node*. The path of an adjunct tree from the root node to the foot node is called the *backbone*. All initial and adjunct trees are referred to as *elementary trees*.

We next define the adjoining operation over trees. Let γ be a tree with a label $X \in V_N$ at a node p . Let β be an adjunct tree with both root and foot nodes labeled with X . The tree γ' obtained by adjoining β at p in γ is shown in Figure 3. We call γ' a *derived tree* from γ . Further, we introduce the notion of *active node* to control derivation processes. We define that β is *adjoinable* to γ at p if and only if the node p is labeled with a nonterminal symbol and *active*. An active node is indicated by the mark $*$.

Let $D(\gamma)$ be the set of trees derived from an elementary tree $\gamma \in \mathcal{I} \cup \mathcal{A}$ by applying zero or more times of adjoining operations. Then, the tree set of a TAG G is defined as $\tau(G) = \{\gamma \mid \gamma \in D(\alpha), \alpha \in \mathcal{I}\}$. Finally, the language generated by G is $L(G) = \{w \in V_T^* \mid w = y(\gamma), \gamma \in \tau(G)\}$.

Uemura et al. [16] have introduced two subclasses of TAGs, called *simple linear TAGs* (SL-TAGs) and *extended simple linear TAGs* (ESL-TAGs), and designed parsing al-

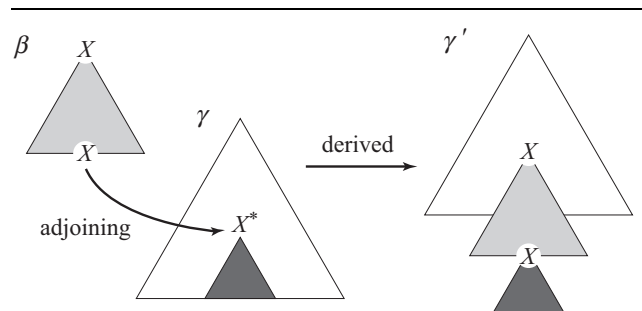


Figure 3. Adjoining operation in TAGs.

gorithms for those subclasses which run in time $O(n^4)$ and $O(n^5)$ respectively, where n is the length of an input string. Although the generative capability of ESL-TAGs is strictly less than TAGs, ESL-TAGs have enough capabilities for modeling RNA secondary structures including pseudoknots.

Let $G = (V_N, V_T, S, \mathcal{I}, \mathcal{A})$ be a TAG. An elementary tree is *simple linear* if all but one node in the tree are not active, and for an adjunct tree, the unique active node is on the backbone of the tree. A TAG G is *simple linear* (SL-TAG) if all elementary trees in G are simple linear. An adjunct tree is *semi-simple linear* if it has two active nodes, where one is on the backbone and the other is elsewhere. A TAG G is *extended simple linear* (ESL-TAG) if initial trees in G are simple linear and all adjunct trees in G are either simple linear or semi-simple linear.

Let β be a simple linear adjunct tree such that $\beta(1) = X$, $y(\beta) = a_1 \cdots a_i X a_{i+1} \cdots a_j$ and q is the

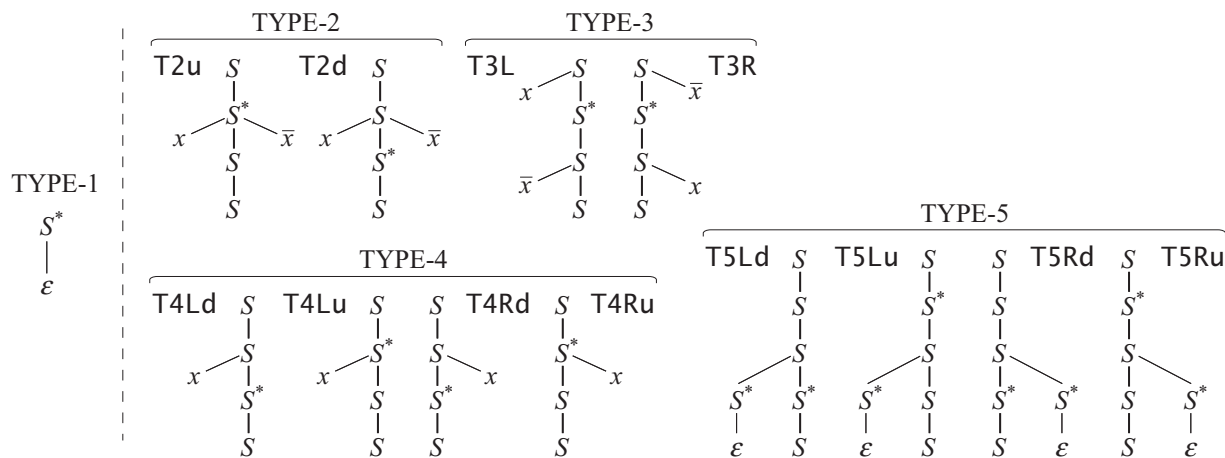


Figure 4. Types of initial trees and adjunct trees in ESL-TAGs for representing pseudoknots.

unique active node on the backbone labeled with Y^* . Further, let the yield of the subtree of β rooted at node q be $a_{i'} \cdots a_i X a_{i+1} \cdots a_{j'}$ for some i', j' ($1 \leq i' \leq i$, $i+1 \leq j' \leq j$). We denote the four subsequences of the yield $y(\beta)$ of an adjunct tree β by $LU(\beta) = a_1 \cdots a_{i'-1}$, $LD(\beta) = a_{i'} \cdots a_i$, $RD(\beta) = a_{i+1} \cdots a_{j'}$, and $RU(\beta) = a_{j'+1} \cdots a_j$.

We define a special form of ESL-TAG, denoted by $G_{RNA} = (V_N, V_T, S, \mathcal{I}, \mathcal{A})$, for representing RNA secondary structures including pseudoknots [16]. Let $V_T = \{A, C, G, U\}$, for representing four nucleotides of RNAs. $x, \bar{x} (\in V_T)$ represents a complementary base pairing, such as (A, U), (C, G) in Watson-Crick base pairing, and additionally (G, U) in wobble base pairing. Let $V_N = \{S\}$, that is, all nonterminal nodes are labeled with one nonterminal symbol S . Figure 4 shows all types of initial trees and adjunct trees in G_{RNA} . In G_{RNA} , every initial tree in \mathcal{I} is of the form of TYPE-1 tree, and every adjunct tree in \mathcal{A} is one of the forms of TYPE-2, TYPE-3, TYPE-4, or TYPE-5 tree. TYPE-2 and TYPE-3 are used to generate a base pair, while TYPE-4 is used to generate a unpaired base which does not form any base pairs. TYPE-5 is used to represent branching substructures. Again, confirm Figure 2 which illustrates a derivation process to produce a pseudoknot secondary structure for “(A (G [AC] U) U)”.

Let us consider a secondary structure of an RNA sequence $w = a_1 a_2 \cdots a_n (\in V_T^*)$, which is specified by a set T of base pairings (a_i, a_j) such that $1 \leq i < j \leq n$. The cardinality of a set T is denoted by $|T|$. Then we say that (a_i, a_j) and (a_k, a_l) in T are *crossing* if and only if either $i < k < j < l$ or $k < i < l < j$ holds. A secondary structure T is said to have *m-crossing property* if and only if there exists a subset T' of T with $|T'| = m \geq 2$ such that any distinct elements (a_i, a_j) and (a_k, a_l) in T'

are crossing. The class of pseudoknot structures defined by ESL-TAGs is *exactly* a class of secondary structures of 2-crossing properties.

3. Pair stochastic tree adjoining grammars

In this section, we describe our main method of PSTAGs and the related techniques.

3.1. TAG trees

We introduce a new representation, called *TAG trees*, which represents a process of how the adjunct trees of TAGs are put into by the adjoining operations to obtain a derived tree whose yield is an input sequence being parsed. Each node of a TAG tree is labeled with an adjunct tree of types shown in Figure 4 for G_{RNA} , and each edge is labeled with an active node in an adjunct tree at the parent node. The root node of a TAG tree is labeled with an adjunct tree which is adjoined to the initial tree. Figure 5 (a) and (b) illustrate some examples of TAG trees for parsing two structured sequences “(a (bB) A) (d [eD] E)” and “(bB) d (fF)”, where a pair of capital letter and small letter for a same symbol represent a base pair.

The TAG tree has two important properties: (1) Every TAG tree has a one-to-one correspondence to a derived tree. (2) The set of TAG trees of an ESL-TAG can be recognized by a tree automaton, and therefore, we can naturally extend the notion of *tree automata* to *TAG tree automata* defined on TAG trees.

3.2. TAG tree automata

A *TAG tree automaton* M is defined by a 5-tuple $M = (Q, \Sigma, \delta, q_0, F)$, where Q is a finite set of states, Σ is a fi-

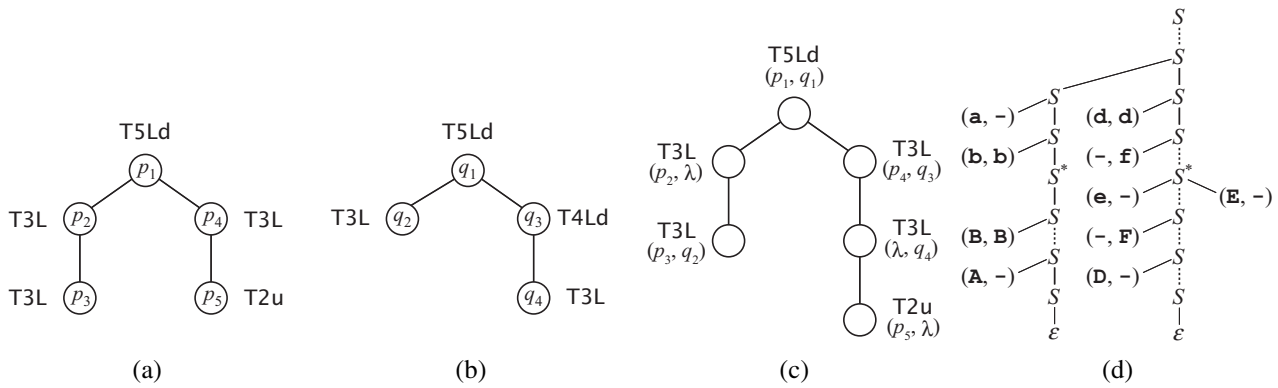


Figure 5. (a) TAG tree for “(a(bB)A)(d[eD]E)”. (b) TAG tree for “(bB)d(fF)”. (c) An alignment of the two TAG trees. (d) An alignment of two derived trees implied by (c).

nite set of labels (that is, adjunct trees), δ is a state transition function mapping a state and a label to a tuple of states: $Q \times \Sigma \rightarrow (Q \times \dots \times Q)$, $q_0 (\in Q)$ is the initial state, and $F (\subseteq Q)$ is a set of final states. A TAG tree automaton accepts a set of TAG trees.

3.3. Alignment of TAG trees

An alignment for a pair of trees is obtained by inserting nodes labeled with the null letter λ into each other’s trees such that two resulting trees have a same structure. A pairwise alignment of two TAG trees is equal to an alignment of trees except that a node in TAG trees is labeled with an adjunct tree. Therefore, a matching between two nodes in a pair of two TAG trees needs an alignment of two adjunct trees. Table 1 shows all possible pairs of adjunct-tree types allowed to be matched.

Figure 5(c) illustrates an alignment of two TAG trees (a) and (b) which represent two pseudoknots “(a(bB)A)(d[eD]E)” and “(bB)d(fF)”, and Figure 5(d) shows an alignment of two derived trees of ESL-TAG implied by the alignment of TAG trees. The resulting structural alignment for both sequences becomes as below:

sequence of (a) : abBAd-e-DE
sequence of (b) : -bB-df-F--
. ()

3.4. Pair stochastic tree adjoining grammars

We introduce *pair stochastic tree adjoining grammars* (PSTAGs) for modeling pairwise alignments of RNA sequences including pseudoknots, which are obtained by modifying stochastic TAG tree automata so that

PAIR	(x, \bar{x})	$(x, -)$	$(-, \bar{x})$
T2u	T2u	T4Lu	T4Ru
T2d	T2d	T4Ld	T4Rd
T3L	T3L	T4Ld	T4Lu
T3R	T3R	T4Ru	T4Rd

SINGLE	x	BRANCH	
T4Ld	T4Ld	T5Ld	T5Ld
T4Lu	T4Lu	T5Lu	T5Lu
T4Rd	T4Rd	T5Rd	T5Rd
T4Ru	T4Ru	T5Ru	T5Ru

Table 1. Pairs of adjunct trees allowed to be matched for alignments of TAG trees.

they emit pairwise alignments of TAG trees instead of emitting single TAG trees.

PSTAGs assign a probability to each alignment of TAG trees in the set of all possible alignments, and can provide the most likely alignment for the pair of TAG trees with the highest probability among the alternatives. The most likely alignment can be efficiently computed by using dynamic programming techniques.

3.5. Calculations of TAG trees

Given an RNA sequence with secondary structure annotation including pseudoknots, where the annotation is usually given by parentheses, the corresponding TAG tree is obtained by parsing the annotated RNA sequence with the ESL-TAG G_{RNA} .

3.6. Algorithm of PSTAGs

Let $w = a_1 a_2 \cdots a_n$ ($\in V_T^*$) be an unfolded RNA sequence of length n , T be a TAG tree of size m representing a folded RNA sequence of known pseudoknot structure, and $T[q]$ ($1 \leq q \leq m$) denote the subtree of T rooted at node q . Let $w[i, j, k, l]$ ($0 \leq i \leq j \leq k \leq l \leq n$) denote two subsequence $a_{i+1} \cdots a_j$ and $a_{k+1} \cdots a_l$ of w . We denote the children of q as q_1 and q_2 .

Here, we present the recurrence equations to calculate an optimal structural alignment between an unfolded RNA sequence w and a TAG tree T for a folded RNA sequence. The recurrence equations calculate an optimal structural alignment based on affine-gap model with three states: match state (M), insertion state (I) and deletion state (D).

$$P^M(w[i, j, k, l], T[q]) = \left\{ \begin{array}{l} \max_{\substack{i < r < j \\ i \leq s \leq r}} P_O^M(\text{T5Ld}, v(q)) \\ \quad \cdot \delta_{MX} \cdot P^X(w[r, j, k, l], T[q_1]) \\ \quad \cdot \delta_{MY} \cdot P^Y(w[i, s, s, r], T[q_2]), \\ \max_{\substack{i < r < j \\ r \leq s \leq j}} P_O^M(\text{T5Lu}, v(q)) \\ \quad \cdot \delta_{MX} \cdot P^X(w[i, r, k, l], T[q_1]) \\ \quad \cdot \delta_{MY} \cdot P^Y(w[r, s, s, j], T[q_2]), \\ \max_{\substack{k < r < l \\ r \leq s \leq l}} P_O^M(\text{T5Rd}, v(q)) \\ \quad \cdot \delta_{MX} \cdot P^X(w[i, j, k, r], T[q_1]) \\ \quad \cdot \delta_{MY} \cdot P^Y(w[r, s, s, l], T[q_2]), \\ \max_{X, Y \in \{M, I, D\}} \left\{ \begin{array}{l} \max_{\substack{k < r < l \\ r \leq s \leq r}} P_O^M(\text{T5Ru}, v(q)) \\ \quad \cdot \delta_{MX} \cdot P^X(w[i, j, r, l], T[q_1]) \\ \quad \cdot \delta_{MY} \cdot P^Y(w[k, s, s, r], T[q_2]), \\ \max_{\beta \in \mathcal{T}} P_O^M(\beta, v(q)) \cdot \delta_{MX} \cdot \\ \quad P^X(w[i - |LU(\beta)|, j + |LD(\beta)|, \\ \quad k - |RD(\beta)|, l + |RU(\beta)|], T[q_1]), \end{array} \right. \\ \text{where } \mathcal{T} \text{ is a set of simple-linear} \\ \text{adjunct trees of TYPE-2, TYPE-3} \\ \text{and TYPE-4 shown in Figure 4,} \end{array} \right.$$

$$P^I(w[i, j, k, l], T[q]) = \max_{\substack{x \in \{I, M\} \\ \beta \in \mathcal{T}}} P_O^I(\beta, \lambda) \cdot \delta_{IX} \cdot P^X(w[i - |LU(\beta)|, \\ j + |LD(\beta)|, k - |RD(\beta)|, l + |RU(\beta)|], T[q]),$$

where \mathcal{T} is a set of simple-linear adjunct trees of TYPE-4 shown in Figure 4,

$$P^D(w[i, j, k, l], T[q]) = \max_{X \in \{D, M\}} P_O^D(\varepsilon, v(q)) \cdot \delta_{DX} \cdot P^X(w[i, j, k, l], T[q_1]),$$

$$P^X(\varepsilon, \theta) = 1, \quad \text{for } X \in \{M, I, D\},$$

where δ_{XY} for $X, Y \in \{M, I, D\}$ denotes the state transition probability from state X to state Y , $P_O^X(\alpha, \beta)$ denotes the emission probability for the pair of labels α, β at

state X , and $v(q)$ denotes the label attached at node p in the tree T .

An optimal structural alignment between a sequence w and a TAG tree T is obtained by calculating $\max_{0 \leq i \leq n} \{\tau_M \cdot P^M(w[0, i, i, n], T[1]), \tau_I \cdot P^I(w[0, i, i, n], T[1]), \tau_D \cdot P^D(w[0, i, i, n], T[1])\}$ for some predefined initial probabilities τ_M, τ_I, τ_D .

The efficient computations of the above recurrence equations can be achieved by using dynamic programming techniques and combined with an efficient parser for TAGs.

3.7. Computational complexity

The computational cost to execute PSTAGs for structural alignments is theoretically of the same order as the computational complexity to parse an input sequence with TAGs. More precisely, the time complexity to run a PSTAG for an input pair of an unfolded sequence of length N and a TAG tree of size M comprised of m branch nodes and n other nodes ($M = m + n$) is $O(KnN^4 + KmN^5)$, where K is the number of states in the PSTAG ($K = 3$ for affine gap). The space complexity for PSTAGs is $O(KMN^4)$.

4. Experimental results

In our experiment to perform structural alignment tests, we have randomly chosen one RNA sequence annotated with a known pseudoknot structure from a RNA family in the database, and structurally aligned all other ‘unfolded’ RNA sequences without annotations into the selected ‘folded’ RNA sequence. We calculated the specificity, that is, the fraction of base pairs predicted by our PSTAG algorithm that agreed with the trusted annotation, and the sensitivity, that is, the fraction of base pairs specified by the trusted annotation that agreed with the predictions of our PSTAG algorithm. Further, in order to remove the dependency of the prediction results on the selected folded RNA sequence, we have executed the cross-validation and calculated the average for all cases.

4.1. Data

The datasets were taken from the RNA families database ‘‘Rfam’’ at Sanger Institute [6] and a collection of RNA pseudoknots ‘‘PseudoBase’’ at Leiden university [17]. The RNA sequences in Rfam are aligned and annotated with secondary structures by using the Covariance Model (CM) method [4], which is a SCFG-based method for modeling RNA sequences. Among 176 RNA families in Rfam (version 5.0), seven RNA families have pseudoknot annotations. The pseudoknot structures entered in PseudoBase are biologically reliable.

name	ave. length	number of seqs.	description
Corona_pk3	62.9 (62 – 64)	14	Coronavirus 3' UTR pseudoknot
HDV_ribozyme	89.1 (87 – 91)	15	Hepatitis delta virus ribozyme
Tombus_3_IV	91.2 (89 – 92)	18	Tombusvirus 3' UTR region IV

Table 2. Three RNA families for experiments from Rfam 5.0.

	specificity [%]		sensitivity [%]	
	average (\pm std. dev.)	worst	average (\pm std. dev.)	worst
Corona_pk3	95.5 \pm 5.0	72.2	94.6 \pm 5.0	72.2
HDV_ribozyme	95.6 \pm 5.1	81.5	94.1 \pm 5.6	81.5
Tombus_3_IV	97.4 \pm 6.0	76.0	97.4 \pm 6.0	76.0

Table 3. Prediction results of PSTAG for three RNA families.

	time [sec]		memory [MB]	
	average (\pm std. dev.)	max	average (\pm std. dev.)	max
Corona_pk3	25.8 \pm 1.2	28.0	(4.33 \pm 0.17) $\times 10^2$	474
HDV_ribozyme	177 \pm 10	197	(2.23 \pm 0.12) $\times 10^3$	2471
Tombus_3_IV	214 \pm 17	257	(2.70 \pm 0.10) $\times 10^3$	2815

Table 4. CPU time and memory usage of PSTAG.

For matching a pair of base pairs, we used a score matrix proposed by Gorodkin et al. [5] instead of probabilistic log-odds scores. Gorodkin's score matrix is defined to be the sum of matrices for alignment scores and base-pairing scores so that both of which can be independent of each other, and have been found to work well on their algorithm although each parameter seems to be determined ad-hoc. Therefore, we could obtain more effective score matrix tailored for our method by machine learning techniques such as EM estimation.

4.2. Test of prediction accuracies of PSTAG

We have tested the prediction accuracy of our PSTAG algorithm for three RNA families, Corona_pk3, HDV_ribozyme, Tombus_3_IV, which have pseudoknot annotations in Rfam (Table 2). The results are shown in Table 3. Corona_pk3 constitutes a simple and typical pseudoknot structure, and HDV_ribozyme constitutes a rather complex structure. Tombus_3_IV has one branching secondary structure which requires more powerful analyses. The prediction performances by PSTAG are very well and stable for all three RNA families.

The CPU time and memory usage for our PSTAG algorithm is shown in Table 4. All experiments were done on a

machine with Intel Pentium4 2.80 GHz processor and 4 GB RAM. The most serious problem is that our PSTAG consumes huge memory space.

4.3. Performance comparisons: three methods

We compared the prediction performances of our PSTAG algorithm with the predictions of the PHMMTS method and of a standard alignment software "Clustal-W" [15]. We tested the three methods for an RNA of HDV_ribozyme in PseudoBase with the reliable annotation about pseudoknot structures.

Pair HMMs on tree structures (PHMMTSs), which are defined on alignments of trees [8], is an extension of pair hidden Markov models [2] and based on the theory of stochastic context-free grammars. By introducing a technique for parsing sequences with context-free grammars, PHMMTSs are applied to the problem of structural alignments for RNA secondary structures without pseudoknots.

Clustal-W only calculates the sequence homologies and does not take into account the pseudoknot structure of the selected RNA sequence when it makes alignments.

The results of comparisons are shown in Table 5. The results clearly show that more accurate predictions can be obtained if we use grammatically more powerful methods.

HDV_ribozyme (EMBL no.: X04451/1-89)

▷ The trusted pseudoknot structure annotated in PseudoBase

```
(((((.....[[[[(((.....))]]))....(((((((.....(((.....))..))))))....]]]....  
GGGUCGGCAUGGCAUCUCCACCUCUCGCGGUCGACCUGGGCAUCCGAAGGAGGACGCACGUCCACUCGGAUGGCUAAGGGAGAGCCA
```

▷ Prediction by PSTAG

```
.(((.....[[[. [[(((.....))]]))....(((((((.....(((.....))..))))))....]]]..]  
GGGUCGGCAUGGCAUCUCCACCUCUCGCGGUCGACCUGGGCAUCCGAAGGAGGACGCACGUCCACUCGGAUGGCUAAGGGAGAGCCA
```

▷ Prediction by PHMMTS

```
.(((.....[[[. [[(((.....))]]))....(((((((.....(((.....))..))))))....]]]..]  
GGGUCGGCAUGGCAUCUCCACCUCUCGCGGUCGACCUGGGCAUCCGAAGGAGGACGCACGUCCACUCGGAUGGCUAAGGGAGAGCCA
```

▷ Prediction by Clustal-W

```
.(((.....[[[. [[(((.....))]]))....(((((((.....(((.....))..))))))....]]]..]  
GGGUCGGCAUGGCAUCUCCACCUCUCGCGGUCGACCUGGGCAUCCGAAGGAGGACGCACGUCCACUCGGAUGGCUAAGGGAGAGCCA
```

Figure 6. Comparisons of predicted secondary structures among three methods: PSTAG, PHMMTS and Clustal-W.

HDV_ribozyme (EMBL no.: L22063/691-781)

▷ The secondary structure annotated in Rfam

```
....(((((((.....[[[[[[(((.....))]]))....(((((((.....(((.....))..))))))....]]]]]....  
GUGGCCGGCAUGGCCCCAGCCUCUCGUCGCGCGGCCUGGGCAACGAUCCGAGGGAGCUACUCUCGAGAAUCGGCAAUGGGGCCCC
```

▷ Prediction by PSTAG

```
.(((.....[[[. [[(((.....))..))))....(((((((.....(((.....))..))))....]]]....  
GUGGCCGGCAUGGCCCCAGCCUCUCGUCGCGCGGCCUGGGCAACGAUCCGAGGGAGCUACUCUCGAGAAUCGGCAAUGGGGCCCC
```

Figure 7. An improvement of secondary structure by PSTAG.

method	specificity [%]	sensitivity [%]
PSTAG	88.9	96.0
PHMMTS	46.4	52.0
Clustal-W	25.9	28.0

Table 5. Comparisons of prediction accuracies among three methods.

Hence, this experiment illustrates the trade-off between the computational and resource costs and the prediction accuracy.

The detailed comparisons are shown in Figure 6, and the correctly predicted structures by each method are indicated with the mark .

4.4. A new secondary structure suggestion

Our third experiment is that with the reliable pseudoknot structures for HDV_ribozyme in PseudoBase, we structurally re-aligned all RNA sequences of the HDV_ribozyme family in Rfam. The predictions of PSTAG significantly improved the secondary structure annotations of Rfam for

HDV_ribozyme. In our estimations, the predictions of PSTAG improved about 25% base pairs in the annotations of Rfam.

An example to exhibit some significant difference between the Rfam annotation and the PSTAG prediction is shown in Figure 7. Some undesirable base pairs, which are indicated with the mark ^, are annotated in Rfam. Our PSTAG predictions improved these undesirable base pairs, and produced a more stable secondary structures than the annotation in Rfam.

Further, the results of PSTAG predictions have suggested a new structure which constitutes an additional internal loop in 3'-end of HDV_ribozyme, as indicated with the mark in the above and also indicated with the arrow ↖ in Figure 8.

5. Related works and discussions

There does not exist any other structural alignment approach to align and predict pseudoknot RNA structures.

In non-comparative approach, there are several theoretical or heuristic works [7, 12] to predict pseudoknot RNA structures for a single RNA sequence by maximizing stacking base pairs or free energy minimizations. Ruan et al. [12]

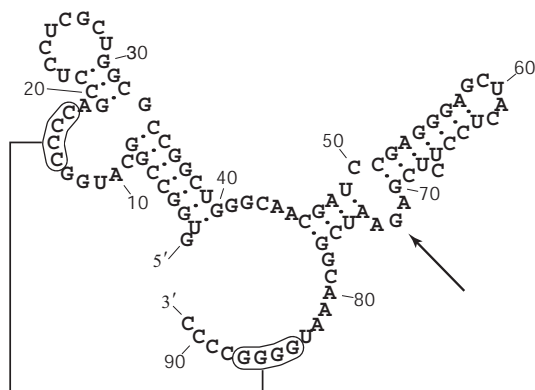


Figure 8. A new secondary structure predicted by PSTAG.

	HDV_ribozyme		TMV	
	specificity	sensitivity	specificity	sensitivity
PSTAG	88.9	96.0	92.0	92.0
ILM	100.0	82.4	80.0	80.0

Table 6. Comparisons of prediction accuracies between ILM and PSTAG.

have recently proposed a simple but effective heuristic method, called iterated loop matching (ILM), for pseudoknot predictions, and shown very high performance results compared with other existing methods. While their approach is completely different from ours, we have done some performance comparisons to verify the effectiveness of our method. Table 6 shows comparisons of pseudoknot predictions between ILM and our PSTAG for HDV_ribozyme and a “tobamovirus” TMV. The RNA sequence of HDV_ribozyme in PseudoBase was structurally aligned by PSTAG into a folded RNA sequence with structure annotation selected from Rfam, and the sequence of TMV was structurally aligned by PSTAG into the folded RNA sequence of “sunn-hemp mosaic virus” CcTMV whose data were retrieved from the paper of van Belkum et al. [18]. Note that the sequence homology between the sequence of HDV_ribozyme in PseudoBase and the selected sequence from Rfam is 65.1% and the sequence homology between TMV and CcTMV is only 26.0%. Our PSTAG exhibits comparable performances for prediction accuracies of pseudoknot structures.

Besides predictions of RNA secondary structures including pseudoknots, another important feature of our approach

is to search and detect non-coding RNA regions on genome. Klein and Eddy [10] have shown an interesting direction to search non-coding RNA sequences in a structural alignment approach using SCFGs. They have developed a local alignment program, called RSEARCH, to search a database for finding structurally homologous RNA sequences, and compared the performance with a well-known BLAST program. Our next research problem is to develop a *local structural alignment* algorithm based on PSTAG for searching a database. The PSTAG-based local alignment algorithm has an advantage to take pseudoknot RNA structures into account and must outperform other database-search methods on both sensitivity and specificity.

References

- [1] E. Dam, K. Pleij, and D. Draper. Structural and functional aspects of RNA pseudoknots. *Biochemistry*, 31(47):11665–11676, Dec. 1992.
- [2] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis*. Cambridge University Press, 1998.
- [3] S. R. Eddy. Noncoding RNA genes and the modern RNA world. *Nature Reviews Genetics*, 2(12):919–929, Dec. 2001.
- [4] S. R. Eddy and R. Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Research*, 22(11):2079–2088, June 1994.
- [5] J. Gorodkin, L. J. Heyer, and G. D. Stormo. Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Research*, 25(18):3724–3732, Sept. 1997.
- [6] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy. Rfam: an RNA family database. *Nucleic Acid Research*, 31(1):439–441, Jan. 2003. <http://www.sanger.ac.uk/Software/Rfam/>.
- [7] S. Jeong, M.-Y. Kao, T.-W. Lam, W.-K. Sung, and S.-M. Yiu. Predicting RNA secondary structures with arbitrary pseudoknots by maximizing the number of stacking pairs. *Journal of Computational Biology*, 10(6):981–995, 2003.
- [8] T. Jiang, L. Wang, and K. Zhang. Alignment of trees — an alternative to tree edit. *Theoretical Computer Science*, 143(1):137–148, 1995.
- [9] A. K. Joshi, L. S. Levy, and M. Takahashi. Tree adjunct grammars. *Journal of Computer and System Sciences*, 10(1):136–163, Feb. 1975.
- [10] R. J. Klein and S. R. Eddy. RSEARCH: Finding homologs of single structured RNA sequences. *BMC Bioinformatics*, 4(1):44, Sept. 2003.
- [11] E. Rivas and S. R. Eddy. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2(8):10, Oct. 2001.
- [12] J. Ruan, G. D. Stormo, and W. Zhang. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics*, 20(1):58–66, Jan. 2004.
- [13] Y. Sakakibara. Pair hidden Markov models on tree structures. *Bioinformatics*, 19(1):i232–i240, July 2003.

- [14] Y. Sakakibara, M. Brown, R. Hughey, I. S. Mian, K. Sjölander, R. C. Underwood, and D. Haussler. Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Research*, 22(23):5112–5120, Nov. 1994.
- [15] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673–4680, Nov. 1994. <http://www.ebi.ac.uk/clustalw/>.
- [16] Y. Uemura, A. Hasegawa, S. Kobayashi, and T. Yokomori. Tree adjoining grammars for RNA structure prediction. *Theoretical Computer Science*, 210(2):277–303, Jan. 1999.
- [17] F. H. D. van Batenburg, A. P. Gulyaev, C. W. A. Pleij, J. Ng, and J. Oliehoek. PseudoBase: a database with RNA pseudoknots. *Nucleic Acids Research*, 28(1):201–204, Jan. 2000. <http://wwwbio.leidenuniv.nl/~Batenburg/PKB.html>.
- [18] A. van Belkum, J. P. Abrahams, C. W. A. Pleij, and L. Bosch. Five pseudoknots are present at the 204 nucleotides long 3' noncoding region of tobacco mosaic virus RNA. *Nucleic Acids Research*, 13(21):7673–7686, Nov. 1985.