BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 11 2005, pages 2611–2617
doi:10.1093/bioinformatics/bti385

Sequence analysis

Pair stochastic tree adjoining grammars for aligning and predicting pseudoknot RNA structures

Hiroshi Matsui, Kengo Sato and Yasubumi Sakakibara*

Keio University, Department of Biosciences and Informatics, 3-14-1 Hiyoshi, Kohoku-ku,
Yokohama 223-8522, Japan

Received on December 7, 2004; revised on March 7, 2005; accepted on March 8, 2005
Advance Access publication March 15, 2005

ABSTRACT

Motivation: Since the whole genome sequences of many species have been determined, computational prediction of RNA secondary structures and computational identification of those non-coding RNA regions by comparative genomics become important. Therefore, more advanced alignment methods are required. Recently, an approach of structural alignment for RNA sequences has been introduced to solve these problems. Pair hidden Markov models on tree structures (PHMMTSs) proposed by Sakakibara are efficient automata-theoretic models for structural alignment of RNA secondary structures, although PHMMTSs are incapable of handling pseudoknots. On the other hand, tree adjoining grammars (TAGs), a subclass of context-sensitive grammars, are suitable for modeling pseudoknots. Our goal is to extend PHMMTSs by incorporating TAGs to be able to handle pseudoknots.

Results: We propose pair stochastic TAGs (PSTAGs) for aligning and predicting RNA secondary structures including a simple type of pseudoknot which can represent most known pseudoknot structures. First, we extend PHMMTSs defined on alignment of ‘trees’ to PSTAGs defined on alignment of ‘TAG trees’ which represent derivation processes of TAGs and are functionally equivalent to derived trees of TAGs. Then, we develop an efficient dynamic programming algorithm of PSTAGs for obtaining an optimal structural alignment including pseudoknots. We implement the PSTAG algorithm and demonstrate the properties of the algorithm by using it to align and predict several small pseudoknot structures. We believe that our implemented program based on PSTAGs is the first grammar-based and practically executable software for comparative analyses of RNA pseudoknot structures, and, further, non-coding RNAs.

Availability: The source code of PSTAG and its web application are available at http://phmmts.dna.bio.keio.ac.jp/pstag/

Contact: yasu@bio.keio.ac.jp

1 INTRODUCTION

Secondary structures including pseudoknots of non-coding RNA molecules play important roles for their own functions such as catalytic functions (Dam *et al.*, 1992). Thus, the computational prediction of pseudoknot RNA structures from primary RNA sequences has become an active research area in bioinformatics, and further there are several theoretical or heuristic works to predict pseudoknot RNA structures such as by maximizing stacking base pairs or free energy minimizations (Abrahams *et al.*, 1990; Cary and Stormo, 1995; Gultyaev *et al.*, 1995; van Batenburg *et al.*, 1995; Rivas and Eddy, 1999; Lyngsø and Pedersen, 2000; Ieong *et al.*, 2003; Ruan *et al.*, 2004). On the other hand, since the whole genome sequences of many species have been determined, computational identification of non-coding RNA regions by comparative genomics become important. Therefore, more advanced methods such as precise algorithms for database search are required for detecting non-coding RNA regions.

Recently, Sakakibara (2003) proposed pair hidden Markov models on tree structures (PHMMTSs) which is an extension of pair HMMs. The PHMMTSs are defined on alignment of trees based on stochastic context-free grammars (SCFGs), and applied to the problem of structural alignment of RNA secondary structures. The approach of structural alignment is to calculate a pairwise alignment to align an unfolded RNA sequence into a folded RNA sequence of known secondary structure (as illustrated in Fig. 1). Thus, an unfolded RNA sequence will be folded by the structural alignment into a single folded RNA sequence, and hence the structural alignment is clearly different from the usual pairwise alignment only based on sequence homology. Two important features of structural alignment are (1) to predict secondary structures for primary RNA sequences and (2) to detect non-coding RNA regions with more sensitivity than sequence homology. The second feature is obviously an advantage compared with conventional methods which can only predict RNA secondary structures.

However, PHMMTSs are incapable of handling pseudoknots because modeling pseudoknot RNA structures is beyond the generative power of context-free grammars, thus inevitably involves in the hard complexity of context sensitivity.

In this paper, we propose a novel method for structural alignment to align and predict RNA secondary structures including pseudoknots. For modeling pseudoknot RNA structures, we first employ special subclasses of tree adjoining grammars (TAGs), which are more generative than context-free grammars but less than context-sensitive grammars. Second, we extend PHMMTSs defined on the alignment of trees to pair stochastic TAGs (PSTAGs) defined on the alignment of ‘TAG trees’ which can represent the derivation process of TAGs for pseudoknot RNA structures. Thus, by combining it with a parsing algorithm for TAGs, we can solve the alignment problem of TAG trees by an efficient dynamic programming algorithm of PSTAGs for obtaining an optimal structural alignment including pseudoknots.

2 TAGs FOR PSEUDOKNOT STRUCTURES

For modeling pseudoknot RNA structures, we employ special subclasses of TAGs, which were introduced to study pseudoknot

*To whom correspondence should be addressed.

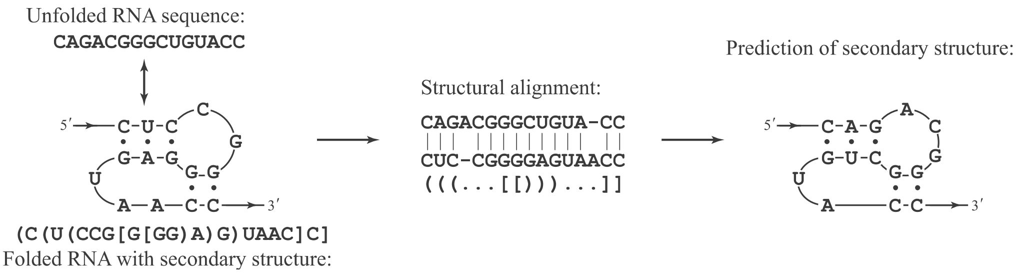

Fig. 1. A structural alignment of an unfolded RNA sequence and a folded RNA.

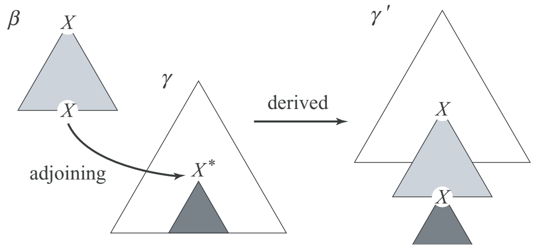

Fig. 2. An adjoining operation in TAGs.

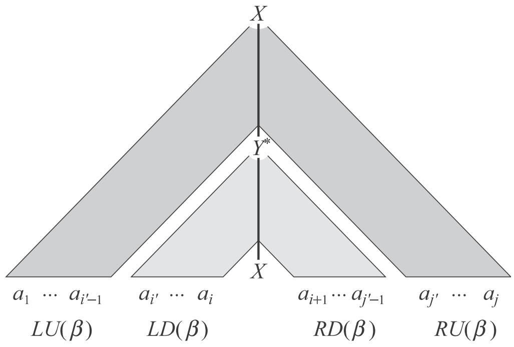

Fig. 3. Decomposition of $\mathcal{Y}(\beta)$.

structures by Uemura *et al*. (1999). In this section, we briefly describe TAGs and the two subclasses, simple linear TAGs (SL-TAGs) and extended simple linear TAGs (ESL-TAGs). For more details, refer to Uemura *et al*. (1999).

Let t be a tree which is a rooted indirected acyclic graph, each node of which is labeled with $X \in V_\mathrm{N} \cup V_\mathrm{T} \cup \{\varepsilon\}$, where V_N is a finite set of non-terminal symbols, V_T is a finite set of terminal symbols and ε is the empty sequence. Nodes in a tree t of the size m are numbered from 1 to m according to the preorder where the root node of t is numbered 1. By $t(p) = A$, we denote that the node p of t is labeled with $A \in V_\mathrm{N} \cup V_\mathrm{T}$. A yield of a tree t, which is denoted as $\mathcal{Y}(t)$, is defined as a concatenating sequence of labels at leaf nodes of t traced from left to right.

A TAG, introduced by Joshi *et al*. (1975), is specified by a 5-tuple $G = (V_\mathrm{N}, V_\mathrm{T}, S, \mathcal{I}, \mathcal{A})$, where $S \in V_\mathrm{N}$ is an initial symbol, $\mathcal{I}$ is a finite set of initial trees and $\mathcal{A}$ is a finite set of adjunct trees. $\mathcal{I}$ and $\mathcal{A}$ must satisfy the following conditions:

1. if $\alpha \in \mathcal{I}$, then $\alpha(1) = S$ and $\mathcal{Y}(\alpha) \in {V_\mathrm{T}}^*$;
2. if $\beta \in \mathcal{A}$, then $\beta(1) = X \in V_\mathrm{N}$ and $\mathcal{Y}(\beta) \in {V_\mathrm{T}}^* X {V_\mathrm{T}}^*$,

where ${V_\mathrm{T}}^*$ denotes a set of all finite sequences over V_T. A foot node of an adjunct tree α is the node which has a label of $X \in V_\mathrm{N}$ in $\mathcal{Y}(\alpha)$. Each path of an adjunct tree from the root node to a foot node is called a backbone. All of the initial trees and adjunct trees are referred to as elementary trees. Then, adjoining operations over trees in TAGs are defined as follows. Let γ be a tree such that $\gamma(p) = X \in V_\mathrm{N}$ and β be an adjunct tree such that $\beta(1) = X$, and one of the foot nodes is also labeled X. An adjoining operation is to derive γ' from γ and β such that β is adjoining γ at the node p as shown in Figure 2. We call γ' a derived tree from γ. A node p of γ with a label $X \in V_\mathrm{N}$ is active if and only if there exists $\beta \in \mathcal{A}$ which can adjoin γ at p such as the node indicated by * in Figure 2.

Uemura *et al*. (1999) introduced two subclasses of TAGs, SL-TAGs and ESL-TAGs, and parsing algorithms of them which run in time $O(n^4)$ and $O(n^5)$, respectively, for an input sequence of the length n.

An initial tree t_initial is simple linear if t_initial has one active node exactly. Similarly, an adjunct tree t_adjunct is simple linear if t_adjunct has one active node on its backbone exactly. Then, a TAG G is a simple linear TAG if all of the elementary trees in G are simple linear. An adjunct tree t_adjunct is semi-simple linear if t_adjunct has two active nodes exactly, where one is on its backbone and the other is elsewhere. Then, a TAG G is an extended simple linear TAG if initial trees in G are simple linear and all of the adjunct trees in G are either simple linear or semi-simple linear.

Let β be a simple linear adjunct tree such that $\beta(1) = X$, $\mathcal{Y}(\beta) = a_1 \cdots a_i X a_{i+1} \cdots a_j$ and q is the active node of β labeled with Y^*, and the yield of a subtree of β rooted at the node q be $a_{i'} \cdots a_i X a_{i+1} \cdots a_{j'}$ for i', j' $(1 \leq i' \leq i,\ i+1 \leq j' \leq j)$. We decompose $\mathcal{Y}(\beta)$ into four subsequences as $LU(\beta) = a_1 \cdots a_{i'-1}$, $LD(\beta) = a_{i'} \cdots a_i$, $RD(\beta) = a_{i+1} \cdots a_{j'}$ and $RU(\beta) = a_{j'+1} \cdots a_j$, as shown in Figure 3. For any sequence $w \in {V_\mathrm{T}}^*$, we denote the length of w as $|w|$. Note that for empty sequence ε, $|\varepsilon| = 0$.

For representing RNA secondary structures including pseudoknots, we define a special case of ESL-TAGs, denoted by $G_\mathrm{RNA} = (V_\mathrm{N}, V_\mathrm{T}, S, \mathcal{I}, \mathcal{A})$, in which $V_\mathrm{T} = \{\mathbf{A}, \mathbf{C}, \mathbf{G}, \mathbf{U}\}$ for representing four kinds of nucleotides and $V_\mathrm{N} = \{S\}$, that is, S is the only non-terminal symbol. G_RNA uses the following forms of elementary trees: a form for initial trees of TYPE-1 and forms for adjunct trees of TYPE-2, 3, 4 and 5 as shown in Figure 4. The forms of TYPE-2 and TYPE-3 are used for generating base pairs, the form of TYPE-4 is used for generating unpaired bases and the form of TYPE-5 is used for representing branching structures. For instance, Figure 5 illustrates a derivation process to produce a pseudoknot RNA secondary structure

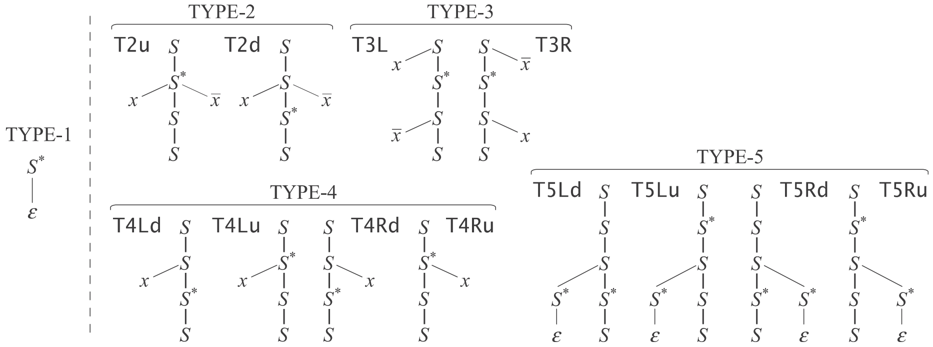

Fig. 4. Forms of initial trees and adjunct trees in ESL-TAGs for representing pseudoknot RNA structures. $\overline{x} \in V_{\mathrm{T}}$ represents a complementary base of $x \in V_{\mathrm{T}}$ such as (**A**, **U**), (**C**, **G**) in the Watson–Crick base pairing and (**G**, **U**) in the wobble base pairing.

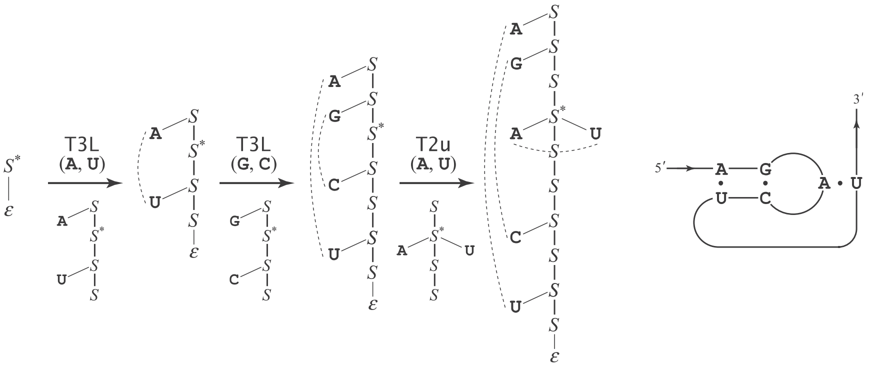

Fig. 5. A derivation process to produce a typical pseudoknot structure for '**(A(G[AC)U)U]**', which cannot be modeled by any context-free grammars because of a crossing dependency.

for '**(A(G[AC)U)U]**' by G_{RNA}, where two kinds of parentheses, '**( )**' and '**[ ]**', indicate base pairs.

Let us consider a secondary structure T of an RNA sequence $w = a_1a_2 \cdots a_n \in V_{\mathrm{T}}{}^*$, which is a set of base pairs (a_i, a_j) such that $1 \le i < j \le n$. Then, we say that (a_i, a_j) and (a_k, a_l) in T is crossing if and only if either $i < k < j < l$ or $k < i < l < j$. A secondary structure T has *m-crossing* property if and only if there exists a subset T' of T with $|T'| = m \ge 2$ such that any pair of (a_i, a_j) and (a_k, a_l) in T' is crossing, where $|T|$ is the number of base pairs included in T. ESL-TAGs have the ability to generate a simple type of pseudoknot RNA structures with exactly 2-crossing property, which can represent most known pseudoknot structures but cannot represent all of them.

3 PSTAGs

In this section, we propose PSTAGs for aligning and predicting RNA structures including pseudoknots.

3.1 TAG trees

We introduce a TAG tree which represents the derivation process of TAGs, that is, what order of adjoining trees could be adjoining to induce a derived tree whose yield is an input sequence. Each node of TAG trees is labeled with an adjunct tree, and each edge means an adjoining operation on an active node at its parent. The root node of a TAG tree is labeled with an adjunct tree which adjoins an initial tree. Figure 6(a) and (b) illustrate some examples of TAG trees for parsing two structured sequences, '**(a(bB)A)(d[eD)E]**' and '**(bB)d(fF)**'.

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

An alignment for a pair of trees is obtained by inserting some null nodes labeled with λ into each other such that two resulting trees have the same topology. Since each node of TAG trees represents an adjunct tree, an alignment of two TAG trees requires matches between adjunct trees. In the case of TYPE-4 and TYPE-5, two adjunct trees of exactly the same form can be matched. On the other hand, adjunct trees of TYPE-2 and TYPE-3 are allowed to be matched as shown in Table 1.

For instance, Figure 6 illustrates that two TAG trees (a) and (b) are aligned into an alignment of TAG trees (c) which corresponds to a derived tree (d). First, since both nodes p_1 and q_1 in the TAG tree (a) and (b) are of the same form T5Ld, they are matched into the node (p_1, q_1) in the aligned TAG tree (c). Similarly, either node p_2 or node p_3 will be matched with q_2, and a null node is inserted to make both TAG trees of the same topology. The nodes p_4 of the form T3L and q_3 of the form T4Ld are aligned into the node (p_4, q_3) of the form T3L. Then, the nodes p_5 of the form T2u and q_4 of the form T3L cannot be matched, and thus null nodes are inserted. Consequently,

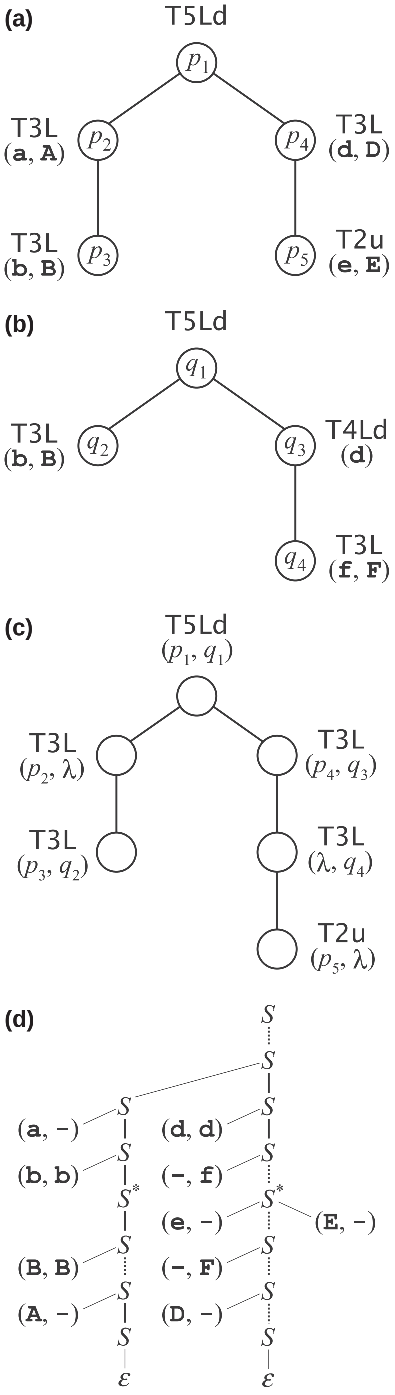

Fig. 6. (**a**) A TAG tree for '`(a(bB)A)(d[eD)E]`'. (**b**) A TAG tree for '`(bB)d(fF)`'. (**c**) An alignment of the two TAG trees. (**d**) An alignment of two derived trees implied by (c).

the resulting structural alignment is obtained as below:

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

Table 1. Any pair of adjunct trees shown in the table can be matched in aligning two TAG trees

(T2u, T2u)	(T2d, T2d)	(T3L, T3L)	(T3R, T3R)
(T2u, T4Lu)	(T2d, T4Ld)	(T3L, T4Ld)	(T3R, T4Ru)
(T2u, T4Ru)	(T2d, T4Rd)	(T3L, T4Lu)	(T3R, T4Rd)
(T4Ld, T4Ld)	(T4Lu, T4Lu)	(T4Rd, T4Rd)	(T4Ru, T4Ru)
(T5Ld, T5Ld)	(T5Lu, T5Lu)	(T5Rd, T5Rd)	(T5Ru, T5Ru)

3.2 Algorithm of PSTAGs

Given an RNA sequence with annotations of secondary structures including pseudoknots, where the annotations are usually given by parentheses, a TAG tree is obtained by parsing the annotated RNA sequence with the ESL-TAG G_{RNA}. We call it a skeletal tree.

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

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

$$
\begin{aligned}
&P^M(w[i,j,k,l],T[q])\\
&= \max_{X,Y\in\{M,I,D\}} \begin{cases}
\displaystyle\max_{\substack{i<r\leq j\\ i\leq s\leq r}} P_O^M(\mathsf{T5Ld}, v(q))\\
\qquad \cdot\delta_{MX}\cdot P^X(w[r,j,k,l],T[q_1])\\
\qquad \cdot\delta_{MY}\cdot P^Y(w[i,s,s,r],T[q_2]),\\
\displaystyle\max_{\substack{i\leq r<j\\ r\leq s\leq j}} P_O^M(\mathsf{T5Lu}, v(q))\\
\qquad \cdot\delta_{MX}\cdot P^X(w[i,r,k,l],T[q_1])\\
\qquad \cdot\delta_{MY}\cdot P^Y(w[r,s,s,j],T[q_2]),\\
\displaystyle\max_{\substack{k\leq r<l\\ r\leq s\leq l}} P_O^M(\mathsf{T5Rd}, v(q))\\
\qquad \cdot\delta_{MX}\cdot P^X(w[i,j,k,r],T[q_1])\\
\qquad \cdot\delta_{MY}\cdot P^Y(w[r,s,s,l],T[q_2]),\\
\displaystyle\max_{\substack{k<r\leq l\\ k\leq s\leq r}} P_O^M(\mathsf{T5Ru}, v(q))\\
\qquad \cdot\delta_{MX}\cdot P^X(w[i,j,r,l],T[q_1])\\
\qquad \cdot\delta_{MY}\cdot P^Y(w[k,s,s,r],T[q_2]),\\
\displaystyle\max_{\beta\in\mathcal{T}} P_O^M(\beta, v(q))\cdot\delta_{MX}\cdot\\
\qquad P^X(w[i-|LU(\beta)|, j+|LD(\beta)|,\\
\qquad\qquad k-|RD(\beta)|, l+|RU(\beta)|],\ T[q_1]),
\end{cases}
\end{aligned}
$$

where $\mathcal{T}$ is a set of simple-linear adjunct trees of TYPE-2, TYPE-3 and TYPE-4 defined in Figure 4,

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

Table 2. The result of predicting base pairs including pseudoknots by PSTAG for three RNA families in Rfam[a]

	Length of sequences	Number of sequences	Specificity (%)	Sensitivity (%)	Time (s)	Memory (MB)
Corona_pk3	62.9	14	95.5 ± 5.0	94.6 ± 5.0	25.8 ± 1.2	$(4.33 \pm 0.17) \times 10^2$
HDV_ribozyme	89.1	15	95.6 ± 5.1	94.1 ± 5.6	177 ± 10	$(2.23 \pm 0.12) \times 10^3$
Tombus_3_IV	91.2	18	97.4 ± 6.0	97.4 ± 6.0	214 ± 17	$(2.70 \pm 0.10) \times 10^3$

[a] Each value in columns of specificity, sensitivity, time and memory represents average and standard deviation of them with respect to the number of sequences. CPU time and memory usage are on a machine with Intel Pentium 4 2.80 GHz processor and 4 GB RAM.

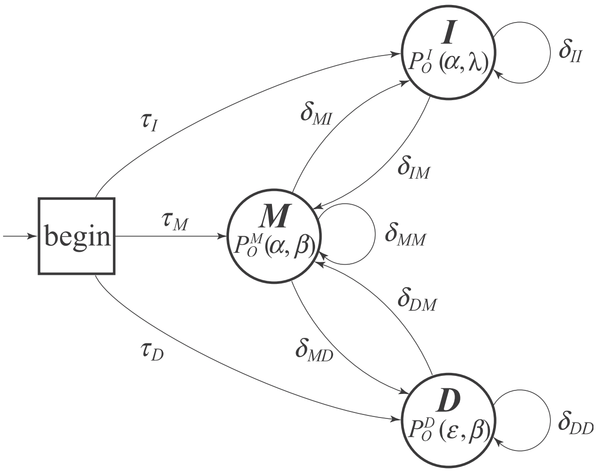

Fig. 7. A state transition diagram of PSTAG for affine gap alignments.

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

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

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

where δ_{XY} for $X, Y \in \{M, I, D\}$ denotes the probability of state transition from the state X to the state Y, $P_O^X(\alpha,\beta)$ denotes the probability of emission for a pair of adjunct trees α and β at the state X, $v(q)$ denotes an adjunct tree labeled at a node q in the tree T, and θ denotes the empty tree. A state transition diagram among three states M, I and D for affine gap alignment is given in Figure 7. An optimal structural alignment between a sequence w and a skeletal tree T is obtained by calculating

$$\max_{0\le i\le n} \begin{cases} \tau_M \cdot P^M(w[0,i,i,n],T[1]), \\ \tau_I \cdot P^I(w[0,i,i,n],T[1]), \\ \tau_D \cdot P^D(w[0,i,i,n],T[1]) \end{cases}$$

for some predefined initial probabilities τ_M, τ_I, τ_D.

Our implementation of PSTAG employs a non-stochastic score matrix proposed by Gorodkin *et al.* (1997) instead of the full stochastic model described above, because each score in Gorodkin's matrix is essentially identical to the probabilistic log-odds score approximated by their round number.

An efficient algorithm for calculating the above recurrence equations can be implemented by using dynamic programming techniques. The computational complexity of executing PSTAGs for structural alignment is the same order as that of parsing an input sequence with ESL-TAGs theoretically. More precisely, time complexity to run a PSTAG for an input pair of an unfolded sequence of the length N and a skeletal tree of the size M with 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 the affine gap model), and space complexity is $O(KMN^4)$.

4 EXPERIMENTAL RESULTS

To confirm our method, we performed some experiments using a certain RNA family in the database. We first randomly chose an RNA sequence annotated with a known pseudoknot structure and parsed it into a skeletal tree, then aligned all the other 'unfolded' RNA sequences in the family into the selected 'folded' skeletal tree without using annotations of them. We evaluated the results of our experiments by specificity and sensitivity, that is, the rate of correctly predicted base pairs by the method to all predicted base pairs, and the rate of correctly predicted base pairs to all of the trusted base pairs in the database, respectively. Further, in order to remove the dependency of the prediction results on the selected folded RNA sequence, we performed cross-validation and calculated the average for all cases.

The datasets used in our experiments were taken from RNA families database 'Rfam' at Sanger Institute (Griffiths-Jones *et al.*, 2003; http://www.sanger.ac.uk/Software/Rfam/) and a collection of RNA pseudoknots 'PseudoBase' at Leiden University (van Batenburg *et al.*, 2000; http://wwwbio.leidenuniv.nl/∼Batenburg/PKB.html). RNA sequences in Rfam are aligned and annotated with secondary structures by using the covariance model (CM) method (Eddy and Durbin, 1994). Among 176 RNA families in Rfam (version 5.0), 7 RNA families have pseudoknot annotations which are unreliable because CM is based on profile SCFGs for modeling RNA sequences which cannot deal with pseudoknots. On the other hand, the annotations of pseudoknot RNA structures in PseudoBase are biologically reliable.

First, we evaluated the accuracy of predicting base pairs by the PSTAG algorithm for three RNA families, Corona_pk3, HDV_ribozyme and Tombus_3_IV, which have pseudoknot annotations in Rfam. Corona_pk3 and HDV_ribozyme constitute simple pseudoknot structures which can be analyzed by an SL-TAG, whereas Tombus_3_IV has one branching secondary structure involving a pseudoknot which requires an ESL-TAG. The results in Table 2 show that PSTAG can predict accurate structural alignments for all three RNA families.

```
>The trusted pseudoknot structure annotated in PseudoBase
(((((((........[[[[(((.......))))))))))...(((((((....((((....))))..)))))))......]]]].....
GGGUCGGCAUGGCAUCUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGAGGACGCACGUCCACUCGGAUGGCUAAGGGAGAGCCA
>Prediction by PSTAG
.((((((....[[..[[[[(((.......)))))))))....(((((((....(((((..)))))..)))))))......]]]].]]..
GGGUCGGCAUGGCAUCUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGAGGACGCACGUCCACUCGGAUGGCUAAGGGAGAGCCA
>Prediction by PHMMTS
.((((((....[[[.[.[[(((.......)))))))))).......((((((((((((((..)))))))).)).))))....]]].]]]...
GGGUCGGCAUGGCAUCUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGAGGACGCACGUCCACUCGGAUGGCUAAGGGAGAGCCA
>Prediction by Clustal-W
..((((((...[[..[[[[(((.......)))))))))).......(((.(((((((((.(...))))))))))).).....]]]]]]...
GGGUCGGCAUGGCAUCUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGAGGACGCACGUCCACUCGGAUGGCUAAGGGAGAGCCA
```

Fig. 8. The detailed comparison for HDV_ribozyme (PKB76) in PseudoBase by secondary structure prediction among three methods: PSTAG, PHMMTS and Clustal-W. Correctly predicted structures by each method are indicated with the mark '_'.

```
>The secondary structure annotated in Rfam
...((((((...[[[[[[(((.........))))))))))).......(((((((((((((...).)))))).)).)))).....]]]]]]...
GUGGCCGGCAUGGCCCCAGCCUCCUCGCUGGCGCCGGCUGGGCAACGAUCCGAGGGAGCUACUCCUCUCGAGAAUCGGCAAAUGGGGCCCC
>Prediction by PSTAG                              ^     ^      ^      ^
.(((((((.....[[[[.(((.........))).))))))).....((((.(((((((.....)))).)))...))))......]]]]....
GUGGCCGGCAUGGCCCCAGCCUCCUCGCUGGCGCCGGCUGGGCAACGAUCCGAGGGAGCUACUCCUCUCGAGAAUCGGCAAAUGGGGCCCC
```

Fig. 9. PSTAG improved about 25% base pairs in the annotations of HDV_ribozyme (RF00094 for the above example) in Rfam by using the annotation of HDV_ribozyme (PKB76) in PseudoBase. Undesirable base pairs annotated in Rfam are indicated with the mark '^', whereas an additional internal loop suggested by PSTAG is indicated with the mark '_'.

Table 3. The accuracy of predicting base pairs for HDV_ribozyme (PKB76) in PseudoBase by PSTAG, PHMMTS and Clustal-W

	Specificity (%)	Sensitivity (%)
PSTAG	88.9	96.0
PHMMTS	46.4	52.0
Clustal-W	25.9	28.0

Second, we compared the prediction accuracy of the PSTAG algorithm with that of PHMMTS and of the standard alignment software 'Clustal-W' (Thompson *et al*., 1994; http://www.ebi.ac.uk/clustalw/) by an RNA of HDV_ribozyme in PseudoBase with reliable annotations about pseudoknot structures. In this experiment, PHMMTS ignores annotations of some stacked base pairs with crossing dependency due to the lack of generative power for pseudoknots. Similarly, Clustal-W ignores any structural annotations due to lack of generative power for secondary structures. Each row of Table 3 shows the accuracy of predicting base pairs by PSTAG, PHMMTS and Clustal-W, respectively. Figure 8 shows the detailed comparison among the three methods, in which correctly predicted structures by each method are indicated by the mark '_'. Obviously, PSTAG succeeded in predicting both '**( )**' and '**[ ]**' base pairs, PHMMTS can predict only "**( )**" base pairs and Clustal-W can predict a few structural annotations. These results indicate that more grammatically powerful the method used, the more accurate the predictions obtained. However, a more grammatically powerful method would consume lager CPU time and memory space generally.

In the third experiment, we structurally re-aligned all the RNA sequences of the HDV_ribozyme family in Rfam, which are unreliable regarding pseudoknots, into reliable pseudoknot structures of HDV_ribozyme in PseudoBase by PSTAG. As a result, PSTAG significantly improved about 25% base pairs in Rfam for HDV_ribozyme, which are undesirable in comparison to PseudoBase. For example, there are some significant differences between the annotation in Rfam and prediction by PSTAG as shown in Figure 9, where some undesirable base pairs, indicated with the mark '**^**', are annotated in Rfam. Therefore, PSTAG can predict more stable secondary structures on these undesirable base pairs than the annotations in Rfam.

In addition, the predictions by PSTAG have some suggestion of a new structure, which constitutes an additional internal loop in the 3′-end of HDV_ribozyme, as indicated with the mark '_' in Figure 9 and also indicated with the arrow '↖' in Figure 10.

5 RELATED WORK

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

In non-comparative approaches, there are several theoretical or heuristic works to predict pseudoknot RNA structures for a single RNA sequence by maximizing stacking base pairs or free energy minimizations (Abrahams *et al*., 1990; Cary and Stormo, 1995; Gultyaev *et al*., 1995; van Batenburg *et al*., 1995; Rivas and Eddy, 1999; Lyngsø and Pedersen, 2000; Ieong *et al*., 2003; Ruan *et al*., 2004). Ruan *et al*. (2004) recently proposed a simple but effective heuristic method, called iterated loop matching (ILM), for predicting pseudoknot structures, and showed high performance results compared with other existing methods. Although their approach is completely different from ours, we compared PSTAG with ILM to confirm the effectiveness of our approach. Table 4 shows comparisons between ILM and PSTAG in predicting pseudoknot structures for HDV_ribozyme and a 'tobamovirus' TMV. In this experiment, PSTAG aligned unfolded RNA sequences of HDV_ribozyme in Rfam into a folded RNA sequence of HDV_ribozyme with structural annotations in PseudoBase, and similarly, aligned unfolded sequences of TMV in Rfam into a folded RNA sequence of 'sunn-hemp mosaic virus' CcTMV whose structure has been determined by van Belkum *et al*. (1985). Note that sequence

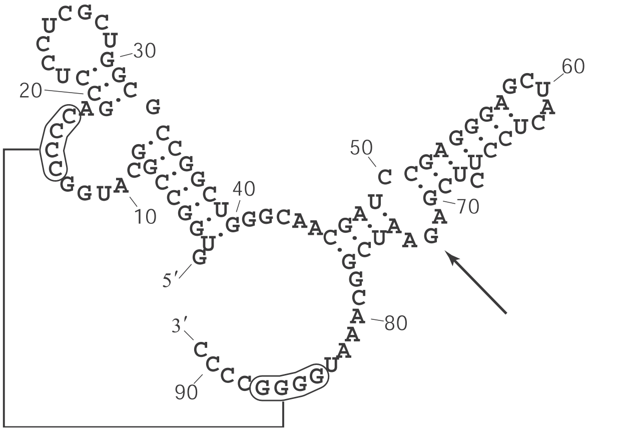

Fig. 10. A new structure suggested of HDV_ribozyme (RF00094) by PSTAG. An additional internal loop is indicated with the arrow '↖'.

Table 4. Comparisons of prediction accuracies between PSTAG and ILM

	HDV_ribozyme		TMV	
(%)	Specificity	Sensitivity	Specificity	Sensitivity
PSTAG	88.9	96.0	92.0	92.0
ILM[a]	100.0	82.4	80.0	80.0

[a] The results of ILM are cited from Ruan *et al*. (2004).

homology between the sequences of HDV_ribozyme in Rfam and the selected sequence of HDV_ribozyme in PseudoBase is 65.1% on average and that between TMV in Rfam and CcTMV is only 26.0%. This result exhibits comparable performances with ILM for prediction accuracy of pseudoknot structures, and further suggests that structural alignment by PSTAG does not require so much sequence homology between an unfolded sequence and a folded sequence.

Another important feature of our approach is searching and detecting non-coding RNA regions on genome. Klein and Eddy (2003) have shown an interesting direction to search non-coding RNA regions in a structural alignment approach based on SCFGs. They have developed a local alignment program, called RSEARCH, to search a database for finding structurally homologous RNA sequences, and compared performances with a well-known BLAST program. Our approach using PSTAG enables us to develop a more accurate database search method which takes a type of pseudoknot RNA structures into account.

REFERENCES

Abrahams,J.P. *et al.* (1990) Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. *Nucleic Acids Res.*, **18**, 3035–3044.

Cary,R.B. and Stormo,G.D. (1995) Graph-theoretic approach to RNA modeling using comparative data. In *Proceedings of the 3rd International Conference on Intelligent Systems for Molecular Biology*. American Association for Artificial Intelligence Press, Robinson College, Cambridge, pp. 75–80.

Dam,E. *et al.* (1992) Structural and functional aspects of RNA pseudoknots. *Biochemistry*, **31**, 11665–11676.

Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, **22**, 2079–2088.

Gorodkin,J. *et al.* (1997) Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Res.*, **25**, 3724–3732.

Griffiths-Jones,S. *et al.* (2003) Rfam: an RNA family database. *Nucleic Acid Res.*, **31**, 439–441.

Gultyaev,A.P. *et al.* (1995) The computer simulation of RNA folding pathways using a genetic algorithm. *J. Mol. Biol.*, **250**, 37–51.

Ieong,S. *et al.* (2003) Predicting RNA secondary structures with arbitrary pseudoknots by maximizing the number of stacking pairs. *J. Computat. Biol.*, **10**, 981–995.

Joshi,A.K. *et al.* (1975) Tree adjunct grammars. *J. Comput. Syst. Sci.*, **10**, 136–163.

Klein,R.J. and Eddy,S.R. (2003) RSEARCH: finding homologs of single structured RNA sequences. *BMC Bioinform.*, **4**, 44.

Lyngsø,R.B. and Pedersen,C.N.S. (2000) RNA pseudoknot prediction in energy-based models. *J. Computat. Biol.*, **7**, 409–427.

Rivas,E. and Eddy,S.R. (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, **285**, 2053–2068.

Ruan,J. *et al.* (2004) An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics*, **20**, 58–66.

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

Thompson,J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Uemura,Y. *et al.* (1999) Tree adjoining grammars for RNA structure prediction. *Theoret. Comput. Sci.*, **210**, 277–303.

van Batenburg,F.H.D. *et al*. (1995) An APL-programmed genetic algorithm for the prediction of RNA secondary structure. *J. Theoret. Biol.*, **174**, 269–280.

van Batenburg,F.H.D. *et al*. (2000) PseudoBase: a database with RNA pseudoknots. *Nucleic Acids Res.*, **28**, 201–204.

van Belkum,A. *et al.* (1985) Five pseudoknots are present at the 204 nucleotides long 3′ noncoding region of tobacco mosaic virus RNA. *Nucleic Acids Res.*, **13**, 7673–7686.