IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 20, 2021, accepted January 28, 2021, date of publication February 9, 2021, date of current version March 2, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3058263

Classifying Conserved RNA Secondary Structures With Pseudoknots by Vector-Edit Distance

LIYU HUANG[1,3], QINGFENG CHEN[2,4], YONGJIE LI[2], AND CHENG LUO[1,5]

[1]School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China
[2]School of Computer, Electronic and Information, Guangxi University, Nanning 530004, China
[3]Information Network Center, Guangxi University, Nanning 530004, China
[4]Department of Computer Science and Information Technology, La Trobe University, Melbourne, VIC 3086, Australia
[5]Xingjian College of Science and Liberal Arts, Guangxi University, Nanning 530004, China

Corresponding author: Qingfeng Chen (qingfeng@gxu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Project 61963004 and Project 61751314, in part by the Key Project of Natural Science Foundation of Guangxi under Grant 2017GXNSFDA198033, and in part by the Key Research and Development Plan of Guangxi under Grant AB17195055.

ABSTRACT Secondary structures have been proved to relate with the great functional diversity of RNA. There have been many studies to predict and compare the RNA secondary structures. However, fast and accurate comparison of RNA secondary structures with arbitrary pseudoknots is a challenging issue due to the hidden but important structural properties, such as the distribution of stems and branches. In this paper, we construct a novel RNA secondary structure model called modified adjoining grammars binary tree (BTMG_{CSP}). It can not only represent the complex RNA secondary structure including arbitrary pseudoknots intuitively, but also reserve RNA secondary structure properties. Further, we propose vector-edit distance to measure the structure similarity between BTMG_{CSP} trees converted from RNA sequences and their secondary structures for classifying conserved stem pattern. The experimental results show that our method substantially reduces the memory and time consumption in contrast to previous algorithms, such as $O((n/k)^2)$ and $O(n/k)$ for time and space complexity, respectively. In particular, the AUC value of our method achieves 0.949 in PseudoBase.

INDEX TERMS Secondary structures, pseudoknots, binary tree, adjoining grammar, edit distance.

I. INTRODUCTION

In the early 1990s, non-coding RNAs (ncRNAs) were revealed to perform non-coding function like catalysis and regulation in biological systems. These findings help to understand how cellular functions evolved from RNA-based origins. Thus, there have been considerable efforts to study regulatory characteristics of ncRNAs, such as ribosomal RNA modification [1]–[3], gene expression regulation [4]–[7] and muscle differentiation control [8], [9]. Especially, researches unveil that ncRNAs are involved in abroad range of human diseases recently [10]–[13]. The functions of ncRNAs have been proved to be primarily determined by their 3D structure. In other words, ncRNAs with high structure similarity are likely to exhibit similar functions. The 3D structure of ncRNA can be inferred from its tertiary structure, which is formed by folding its secondary structure. Compared with the complex RNA tertiary structures, it is easier to infer their potential functions by measuring secondary structure similarity [14], [15]. Therefore, the study of RNA secondary structure is crucial to understand the function and regulation of RNA transcripts [16], [17].

Studies have shown that the structure of ncRNA is more conservative than sequence in biological evolution [18]–[20]. Most researches in this field focus on structural prediction and comparison. The former relying on thermodynamic parameters [21], [22] mainly includes experimental analysis [23] and computational prediction [24]. Many methods of RNA secondary structure comparison have been proposed, and new ones are still emerging [25]. These methods can be classified into alignment problem and edit problem [26]. Given two structures R_1 and R_2, the alignment problem aims to identify a consensus structure R_c, which minimizes the total edit cost from them to R_c. For example, Sakakibara defined Hidden Markov Models on tree structures (PHMMTSs) based on a random context-free grammar and applied it to the alignment of RNA secondary structures [27]. In contrast, the edit problem uses a series of predefined

The associate editor coordinating the review of this manuscript and approving it for publication was Emre Koyuncu.

edit operations to calculate the minimum cost (distance) of transferring R_1 to R_2. There have been many researches on this issue, such as constructing RNA secondary structure models with ordered trees to calculate the edit distance or similarity score between them [28]. The heuristic method was used to compute the edit distance between two RNA structures [29]. Obviously, unlike alignment problem, edit problem only needs to calculate the minimum cost of the modification from R_1 to R_2, neither to find the consistent structure R_c, nor to calculate the total edit cost from R_1 and R_2 to R_c.

Most aforementioned methods do not take into account highly conserved pseudoknots that actually perform varied functions [30]. Further, the alignment of RNA structure containing arbitrary pseudoknots is NP-hard [31]. Therefore, the study of pseudoknots has received more attention. Uemura *et al.* [32] proposed two special subclasses of linear tree adjoining grammars (TAGs), simple linear TAGs (SL-TAGs) and extended simple linear TAGs (ESL-TAGs), and applied them to model and predict RNA secondary structure with pseudoknots. It is effective for some simple short pseudoknots. Based on these two special TAGs subclasses, PSTAGs was proposed for comparison and prediction of RNA secondary structures [33]. It can handle most simple type pseudoknots, but fails to handle the secondary structures with asymmetrical structures.

The stem graph, which is a directed complete graph, was used to model the stem pattern [34]. The conserved stem pattern was found by calculating the minimum cost of error-correcting graph matching (ECGM) between graphs, and the alignment of the base pairs in RNA secondary structures with pseudoknots was achieved. Nevertheless, this method only considered three structural relationships (nested, parallel and pseudoknotted) between stems. The structural properties within stems and branches are overlooked. This impacts on finding conserved stem patterns. For example, the sequence and secondary structure of potato leafroll virus (PKB43) and foot-and-mouth disease virus (PKB284) are UUUAAAUGGGCAAGCGGCACCGUCCGCCAAAACAAACGG and::::::::::::((((::[[[[)))):::::::]]]], and ACCGCCUACCCCGGCGUUAACGGGGAACAA and::((((::[[[[)))) :::::]]]]:::::, respectively. Their transformed stem graphs are exactly the same. The minimum cost of ECGM between two graphs is zero. They actually belong to the Viral ribosomal frameshifting signals class and Other Viral 5'-UTR class, respectively. Chen *et al.* proposed nested stack, parallel stack and intersected stack to represent the nested, parallel and pseudoknotted relationships in RNA secondary structures [35]. They considered the branched structure, and used the occurrence frequency of base pairs in the database as weight to calculate the vector distance between structures. To some extent, this method improves the accuracy of the classification of RNA secondary structures, but it only calculates the distance of their similar substructures and assumes that the number of branches between the structures is the same. Recently, to handle structures with arbitrary pseudoknots, ASPRAlign used special representations of secondary structures, called Algebraic RNA Trees and Structural RNA Trees, based on algebraic operators, to compare RNA secondary structures [36]. And time complexity is quadratic, while all the other alignment methods have a worst-case time complexity more than quadratic as shown in Table 1 of [37]. However, neither could the complex RNA secondary structure with more than two intersected relationships be represented, nor did the structural properties of the branches be considered in the above RNA Trees. These two shortcomings impact the usefulness and efficiency of this approach.

To overcome the above limitations, we construct a modified adjoining grammars binary tree for the complex RNA secondary structure with arbitrary pseudoknots called BTMG_{CSP} tree, based on the ideas of TAGs. And we propose a novel vector-edit distance based on BTMG_{CSP} tree to measure the similarity between RNA secondary structures. The experimental results demonstrate that our method is useful for querying, aligning or classifying pseudo-knotted conserved RNA structures. The main contributions of this paper are as follows:

- The model of the complex RNA secondary structure with arbitrary pseudoknots. A novel RNA secondary structure model called modified adjoining grammars binary tree for conserved stem pattern (BTMG_{CSP}) is proposed. It can not only represent the complex RNA secondary structure including arbitrary pseudoknots intuitively, but also reserve RNA secondary structure properties. It allows an effective comparison of secondary structures because we can compute the similarity of secondary structures based on their most informative structural features.
- New structural similarity strategy. We proposed vector-edit distance to measure the structure similarity between BTMG_{CSP} trees. Because our algorithm considers the node properties of RNA conserved stem pattern, but also takes the distribution of the nodes and branches into account, the correctness of classification and comparison of RNA secondary structures is improved as shown in our experiments.
- Space and time efficiency. Our method substantially reduces the memory and time requirements w.r.t. previous strategy computation algorithms since BTMG_{CSP} tree is an ordered binary tree. The worst-case time and space complexity reduce to $O((n/k)^2)$ and $O(n/k)$, respectively, where n is the number of nucleotides of the longer structure, k is the average of the numbers of the base pairs in the conserved stem patterns. Although our strategy algorithm considers more structural features, it is as efficient as the strategy algorithm in ASPRAlign (quadratic in the input size).

The remaining paper is organized as follows. Section II and Section III describe the construction of RNA complex secondary structure model and the similarity measure based vector-edit distance for (BTMG_{CSP}) trees, respectively. Experimental results are presented and discussed in Section IV. The conclusions and future work are stated in Section V.

II. PRELIMINARIES

Many folding methods have been developed for predicting RNA secondary structure. Although the accuracy of non-pseudoknot RNA secondary structure prediction method has been significantly improved, there are some shortcomings in the identification of conserved stem pattern owing to overlooking the structural properties within stems and branches. Moreover, the conserved structural motifs with arbitrary pseudoknots can further assist in detecting similar regulatory functions of non-coding RNAs. This section explores the construction of RNA complex secondary structure model.

A. TAGs FOR RNA SECONDARY STRUCTURE

The definition and operation of the adjacency tree method are briefly introduced below.

Definition 1 (Tree Adjoining Grammars (TAG)): Given a five-tuple $G = (V_N, V_T, I, A, S)$, where V_N is a finite set of nonterminal symbols, V_T is a finite set of terminal symbols, I is a finite set of initial trees, and A is a finite set of auxiliary trees. S is an initial symbol, and also a special non-terminal term, namely $S \in V_N$.

The basics of TAG formalism is as follows.

- Elementary trees are all the trees without any operations in $I \cup A$.
- Derivation tree is a tree composed of any two trees in $I \cup A$.
- Adjoining $(*)$ labelled for the non-terminal symbol of the leaf node in auxiliary tree A, and the node with $(*)$ is called the foot node.
- Substitution$(\downarrow)$ labelled for the non-terminal symbol of the leaf node in initial tree I.
- Derivation is a process to yield the derivation tree by using the two operations, Adjoining and Substitution.

Adjoining is the process of creating a new tree by inserting the auxiliary tree β into an arbitrary tree α (initial tree, auxiliary tree, or derivation tree), which is usually used in modelling RNA structures. The definition is as follows.

Let α be a tree including node n, where node n represents the nth node that traverses α tree by preorder traversal. The root of the tree is the first node, and the label on n is X. That is, $\alpha(n) = X \in V_N$.

Definition 2 (Adjoining): Let β be an auxiliary tree whose root node and foot node p on the boundary are also labelled as X, that is, $\beta(1) = \beta(p) = X$. By adjoining β and tree α at node n, a tree γ can be obtained, as shown in Fig. 1. Here, γ is called a derivation tree of α. The node n labelled $X \in V_N$ on the tree α is active, and if and only if $\beta \in V_N$, the tree β can be adjoined at the node n of tree α.

Two special subclasses of linear TAGs (SL-TAGs and ESL-TAGs) were proposed by Uemura *et al.* [32]. Let β be a simple linear auxiliary tree with active node n at p, where $\beta(0) = X$, and $Y(\beta) = \alpha_1 \cdots \alpha_i X \alpha_{i+1} \cdots \alpha_j$. Therefore, $Y(\beta)$ can be divided into two major parts, $\mathrm{L}(\beta) = \alpha_1 \cdots \alpha_i$ and $\mathrm{R}(\beta) = \alpha_{i+1} \cdots \alpha_j$. Obviously, there must exist i' and j' such that $\alpha_{i'} \cdots \alpha_i X \alpha_{i+1} \cdots \alpha_{j'} = Y(\beta/p)$, where

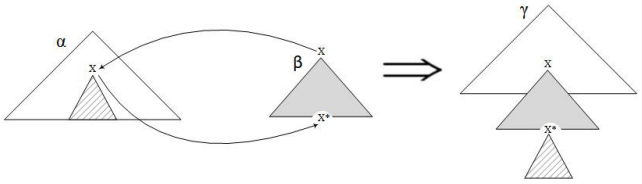


FIGURE 1. An adjoining operation in TAGs.

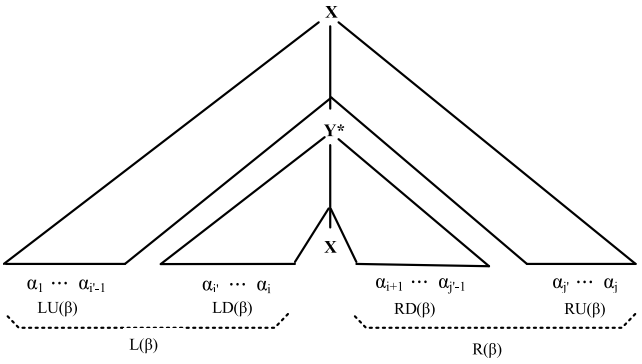


FIGURE 2. Decomposition of Y (β).

$1 \le i' \le i, i+1 \le j' \le j$, β/p is a subtree of β at p. Then, the left and right parts are further divided by i' and j', respectively. And we have $\mathrm{LU}(\beta) = \alpha_1 \cdots \alpha_{i'-1}$, $\mathrm{LD}(\beta) = \alpha_{i'} \cdots \alpha_i$, $\mathrm{RD}(\beta) = \alpha_{i+1} \cdots \alpha_{j'-1}$, $\mathrm{RU}(\beta) = \alpha_{j'} \cdots \alpha_j$, as shown in Fig. 2.

On the basis of the ESL-TAGs, pair stochastic TAGs (PSTAGs) was proposed by Matsui *et al.* [33]. It is better than ESL-TAGs in representing secondary structures including pseudoknots. However, both ESL-TAGs and PSTAGs represented RNA secondary structure models from the sequence. The sequence numbers on the right must be greater than the left. If a subsequence is nested inside a pseudoknot, they cannot be expressed. As in [32], [33], they can represent the secondary structure of RNA sequence (A(G[AC)U)U], but fail to describe the secondary structure of the sequence (A(G[AC)U]U). The reason is that the base U (sequence number is 5) of the pseudoknot pair [AU] should be on the right side of the backbone tree, and the base U (sequence number 6) of (AU) will be on the left side of its backbone tree. This contradicts the decomposition idea of $Y(\beta)$. Furthermore, they cannot represent complex RNA secondary structure with more than two intersected relationships. For example, the sequence and secondary structure of coli alpha operon mRNA are UGUGCGUUUCCAUUUGAGUAUCCUGAAAACGGGCUUUUCAGCAUGGAACGUACAUAUUAAAUAGUAGGAGUGCAUAGUGGCCCGUAUAGC AGGCAUUAACAUUCCUGA and (((((((.(((((…..[[[[….[[[[….{{ {{.)))))))))))).....]]]].....]]]]..........}} }}, respectively. To illustrate this further, the first pair of RNA stem used to simplify the conserved structure, (U1 (U9 [C22 [C30 {U38 A48) A54) G84] G93] A108}, where the number on the right side of the base indicates its position in the RNA sequence. If such a structure is

represented by ESL-TAGs or PSTAGs, the base {U38 would be located below [C30 on the left side, and the base A108} would be located above the base G93] on the right side. However, [C30 should be on the left side of the auxiliary tree T2u ([C30 G84]), and G93] should be on the right side of the auxiliary tree T2d ([C22 G93]). Therefore, the base pair of the second intersected relationship {U38 and A108} cannot be described on the backbone tree, and make it impossible to be represented by ESL-TAGs or PSTAGs.

B. BTMG$_{RNA}$ FOR PSEUDOKNOT STRUCTURES

To intuitively represent the complex RNA secondary structures containing arbitrary pseudoknots, a novel model is developed for RNA secondary structures with arbitrary pseudoknots by adapting and extending tree adjoining grammars. And modified tree adjoining grammars for RNA secondary structures (MG$_{RNA}$) is defined as follow.

*Definition 3 (***MG$_{RNA}$***):* $MG_{RNA} = (V_N, V_T, I, E, S)$, where $V_T = (A, C, G, U)$ is the four bases of RNA, and $V_N = S$, that is the only non-terminal symbol. The initial tree I is composed of an active non-terminal S^* and an empty string ε. The auxiliary tree of E = { T2uN, T2dN, T3d, T4d, T4u, T5d, T5u}, where the letter u and d mean the upper adjacent and the lower adjacent of a given node, respectively. T2uN denotes the first base pair of the N*th* intersected relationship. T2dN denotes the nested base pair of the N*th* intersected relationship. T3d denotes a general paired base, namely stem. T4d and T4u represent the adjacency of unpaired bases, and T5d and T5u represent parallel and nested branch structures, respectively.

Since MG$_{RNA}$ is not decomposed from the RNA sequence, there is no left and right part, and each of the auxiliary trees only represents a way of adjacency in RNA secondary structure. Therefore, the auxiliary tree of MG$_{RNA}$ is different from ESL-TAGs or PSTAGs. The initial tree I and auxiliary tree E in MG$_{RNA}$ are shown in Fig. 3. If the auxiliary tree in MG$_{RNA}$ is viewed as a node of a tree, a RNA secondary structure can be represented as a binary tree of MG$_{RNA}$ (BTMG$_{RNA}$ for short). An instance graph of BTMG$_{RNA}$ for RNA secondary structure (A(G[AC)U]U)(U[C[A{UU]G]A}A) is shown in Fig. 4. As explained in Section II-A, this structure cannot be described by ESL-TAGs and PSTAGs. In addition, PSMAlign, which used the directed complete graph to represent the stem pattern, only simply shows the relationship between two stem patterns, but fails to describe the multiple intersected relationships [34].

C. BTMG$_{RNA}$ FOR CONSERVED STEM PATTERN

Since the secondary structures of ncRNAs is more conserved than their sequences, similar secondary structures would share a common conserved stem patterns. In this paper, we focus on conserved secondary structure of ncRNAs. Based on MG$_{RNA}$ in Section II-B, a novel BTMG$_{RNA}$ tree is defined for Conserved Stem Pattern (BTMG$_{CSP}$ for short).

*Definition 4 (***BTMG$_{CSP}$***):* Let BTMG$_{CSP}$ = (V_N, V_T, I, Σ, S), where V_N, V_T, S and I are the same as MG$_{RNA}$,

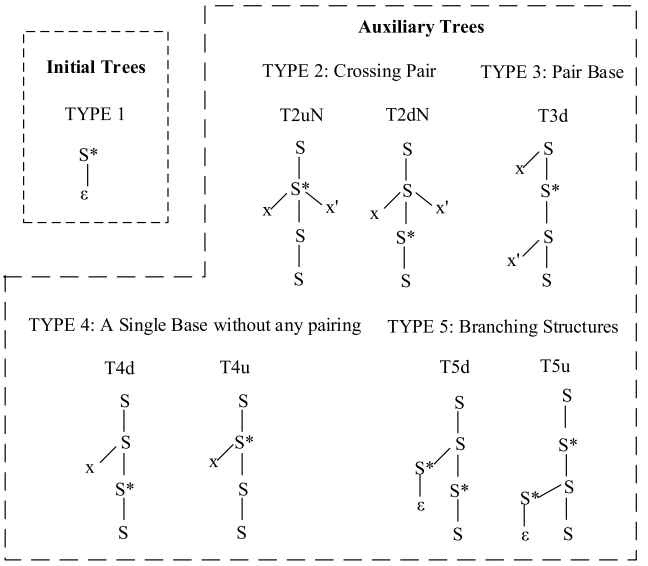


FIGURE 3. Forms of elementary trees in MG$_{RNA}$ for representing RNA pseudoknot.

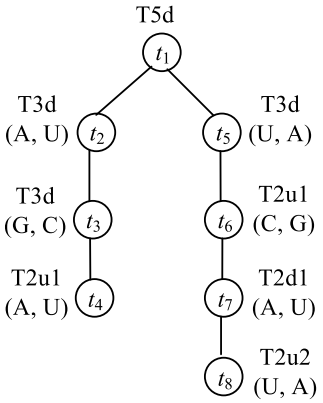


FIGURE 4. BTMG$_{RNA}$ for (A(G[AC)U]U)(U[C[A{UU]G]A}A).

Σ = {stem loop, pseudoknot, multi loop}, namely Σ ={T2uN, T2dN, T3d, T5d, T5u}.

Algorithm 1 provides a detailed description for converting RNA sequence and its secondary structure into BTMG$_{CSP}$ tree. In this algorithm, two intersected relationships are considered. Three stacks S_{paren}, $S_{bracket}$ and S_{brace} are initialized for storing the three structure types of base pairs in RNA dot-bracket secondary structures, respectively(line 1). The properties of $Q(i)$ and $S(i)$ are pushed into stack S_{paren}, $S_{bracket}$ or S_{brace} according to the type of brackets. If $S(i)$ is left paren/bracket/brace or $S(i)$ is the dot on the right side of the left paren/bracket/brace or $S(i)$ is the dot on the left side of the right paren/bracket/brace(lines 4-6). Pushing the dot into the stacks is to distinguish two successive stem patterns, and extract their properties, such as their lengths. While matching the right paren, right bracket or right brace, the corresponding stack is popped out and the node properties are added into the node list N_{paren}, $N_{bracket}$ or N_{brace}(lines 7-13). And then, binary tree $\boldsymbol{T}$ is constructed from the list

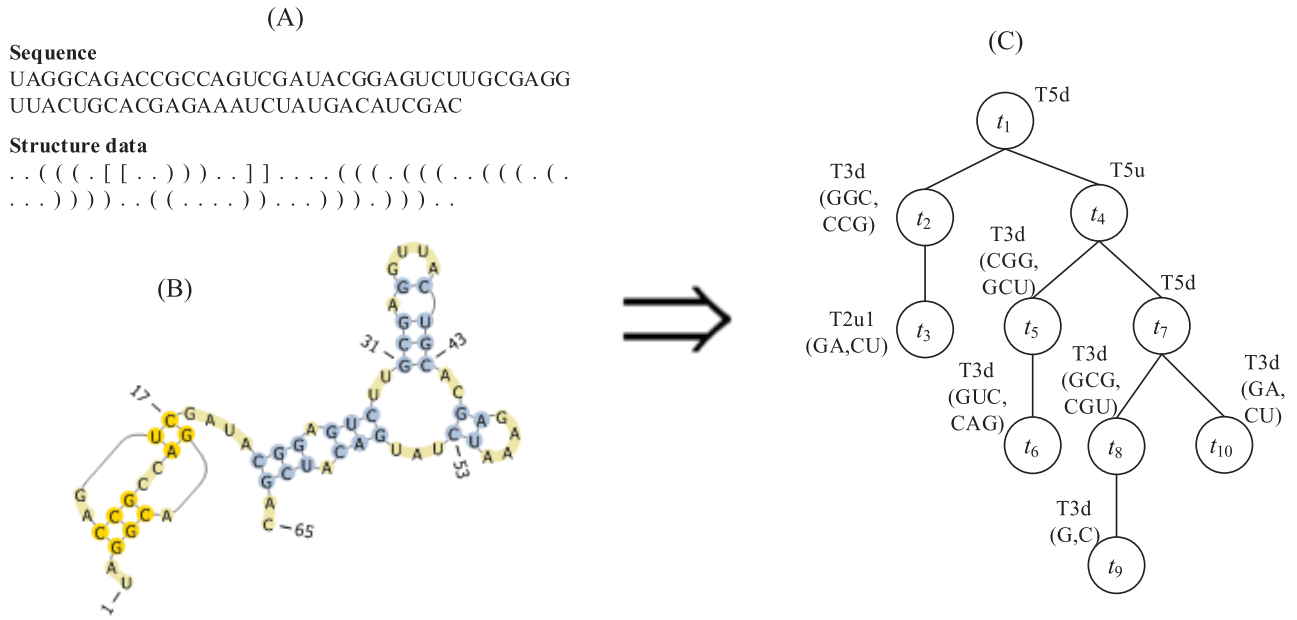


FIGURE 5. The sequence of RNA bases is indicated by dot-bracket (A), the structure diagram drawn by PseudoViewer (B), and its $BTMG_{CSP}$ (C).

N_{paren}. The nodes in $N_{bracket}$ and N_{brace} are inserted into the first tree branch that has intersected relationship with them in order(lines 16-18). Therefore, the binary tree $\boldsymbol{T}$ for RNA conserved stem pattern, that is $BTMG_{CSP}$ tree, is obtained.

Fig. 5 shows the dot-bracket representation of an RNA base sequence (A), the structure diagram drawn by PseudoViewer 3.0 [38] (B), and its $BTMG_{CSP}$ grammars tree (C).

$BTMG_{CSP}$ tree presents the RNA conserved stem pattern. In order to improve the validity and accuracy of the structure alignment/classification, the node properties are saved on the tree node as needed, including the important structural properties of base pair (stem and pseudoknot), such as the adjacency type, terminal position and length of base pair, as shown in Fig. 6.

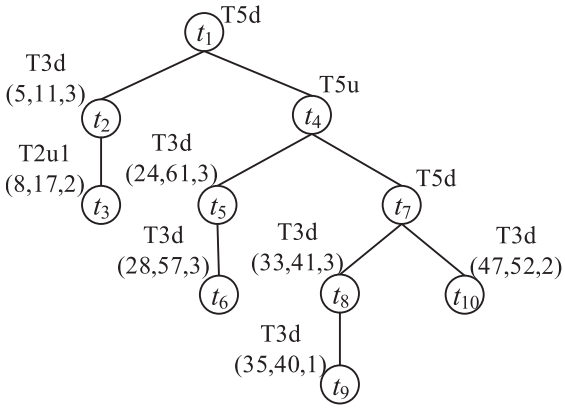


FIGURE 6. The node properties of $BTMG_{CSP}$ in Fig. 5.

III. SIMILARITY MEASURE FOR $BTMG_{CSP}$

After transforming RNAs into $BTMG_{CSP}$ trees, the distance of RNAs is transferred to compute the distances between their $BTMG_{CSP}$ trees. A common effective way of calculating tree similarity is to calculate the distance of their most similar subtrees and the cost for finding the similar subtrees.

$BTMG_{CSP}$ tree is obtained by parsing the annotated RNA sequence, called an RNA conserved tree.In the study of the conserved structure of RNA secondary structure, the similarity based on interval distance has been proved to be an effective way for identifying the conserved structure of RNA [35]. Edit distances have also been proved to be an effective method for calculating edit costs in conserved stem pattern without pseudoknots [39]. Based on these two distance strategies, vector distance and edit distance are defined in this section, and the vector-edit distance method is proposed to calculate the similarity between RNA secondary structures with arbitrary pseudoknots. This finds the most similar subtrees by calculating the minimal vector distance between $BTMG_{CSP}$ trees, and the cost for finding the most similar subtrees is the edit distance during the process. The distance between $BTMG_{CSP}$ trees is regarded as a function of the vector distance and the edit distance.

A. VECTOR DISTANCE

The similarity (distance) between $BTMG_{CSP}$ trees is used to determine whether two RNA secondary structures are similar. Thus, the most similar subtree needs finding by the minimal vector distance between them. According to [35], the distance between the vectors $a = [a_1, a_2]$ and $b = [b_1, b_2]$ is

Algorithm 1 Construct BTMG_{CSP} Tree

Input: RNA sequence(Q) and its dot-bracket secondary structure(S)

Output: BTMG_{CSP} tree $\boldsymbol{T}$

/∗Take two intersected relationships for instance here, there are three types of brackets parentheses, brackets, and braces ∗/
Initial stacks S_{paren}, $S_{bracket}$ and S_{brace}
for i = 1; $i \leq$ length(Q); i++ **do**
  if $S_i =$ ' (' / ' [' / ' { ' or
    $S_i =$ '. ' and $S_{i-1} =$ ' (' / ' [' / ' { ' or
    $S_i =$ '. ' and $S_{i+1} =$ ') ' / '] ' / ' } ' **then**
      // p_i denotes the properties of Q_i and S_i
      push (S_{paren} / $S_{bracket}$ / S_{brace}, p_i)
  else if $S(i) =$') ' / '] ' / ' } ' **then**
      $P =$ pop (S_{paren} / $S_{bracket}$ / S_{brace}),
      if $P =$'. ' **then**
        Calculate the length of base pair *len*,
        $P =$ pop (S_{paren} / $S_{bracket}$ / S_{brace})
      end if
      Insert P and *len* into the node list N_{paren}, $N_{bracket}$ or N_{brace}
  end if
end for
Build binary tree BTMG_{CSP} $\boldsymbol{T}$ from N_{paren}
The nodes in $N_{bracket}$ and N_{brace} are inserted into the branch of tree $\boldsymbol{T}$ in order
Return BTMG_{CSP} tree $\boldsymbol{T}$

defined as in (1):

$$d(a,b) = \begin{cases} 0 & a = b \\ max(|a_1 - b_1|, |a_2 - b_2|) & a \subset b \text{ or } b \subset a \\ H(a,b) \times \left(1 - \frac{O(a,b)}{|c_a - c_b|}\right) & a_1 < b_1 \leq a_2 < b_2 \text{ or } \\ & b_1 < a_1 \leq b_2 < a_2 \\ H(a,b) \times \left(1 - \frac{O(a,b)}{|c_a + c_b + 1|}\right) & otherwise \end{cases} \tag{1}$$

where $r_a = (a_2 - a_1)/2$, $r_b = (b_2 - b_1)/2$, $c_a = a_1 + r_a$, $c_b = b_1 + r_b$ and

$$O(a,b) = \begin{cases} 2r_a & a \subset b \\ 2r_b & b \subset a \\ |c_a - c_b| + r_a + r_b & else \end{cases} \tag{2}$$

$H(A,B) = max\{h(A,B), h(B,A)\}$ is the Hausdorff distance. $h(A,B) = max\{min\, d(a,b)\}$ denotes the maximum distance from the nearest point of set A to set B, and $h(B,A) = max\{min\, d(b,a)\}$ denotes the maximum distance from the nearest point of set B to set A. In this paper, $H(a,b)$ means $h(a,b)$ instead of $max\{h(A,B), h(B,A)\}$, and the vector is the terminal position of base pair. For example, in Fig. 6, the terminal position (8, 17) of the leftmost leaf node t_3 is the vector [8, 17].

B. EDIT DISTANCE

In many cases, neither the number of branches nor the number of nodes within the branch is equal between RNA secondary structures. While comparing RNA secondary structures or predicting their function, it is obviously insufficient to calculate only the vector distances for finding their most similar substructures. It is necessary to compute the cost of finding the similar substructures. In [39], conserved edit distances without pseudoknots have been given. In this subsection, an edit distance of RNA secondary conserved structure with arbitrary pseudoknots is proposed.

BTMG_{RNA} tree is a linear ordered tree. Suppose linear trees T_1 and T_2 represent two RNA secondary structure trees, respectively. t_1 and t_2 are the nodes within T_1 and T_2, respectively. The node edit operation (t_1, t_2) between T_1 and T_2 is defined as an evolutionary event on the RNA secondary structure. The concept of (t_1, t_2) is described in Tabel 1. If t_1 and t_2 are base pair nodes or unpaired base nodes, and $t_1 \neq t_2$, such as, the label of t_1 is base A, and the label of t_2 is base G, then relabeling (t_1, t_2) denotes the evolutionary event for relabeling A to G. $(t_1, \varnothing)$ called deletion, $(\varnothing, t_2)$ called insertion. If t_1 is a base pair node, and t_2 is an unpaired base node, then (t_1, t_2) denotes altering. If t_1 is an unpaired base node, and t_2 is a base pair node, then (t_1, t_2) denotes completion. If t_1 is a base pair node, and t_2 is a pair of unpaired base nodes, then (t_1, t_2) denotes the arc-breaking. If t_1 is a pair of unpaired base nodes, while t_2 is base pair node, then (t_1, t_2) represents the arc-creation. In the seven edit operations, insertion, altering, and arc-breaking are the symmetric operations of deletion, completion, and arc-creation, respectively. The cost of the symmetric operation is the same. Therefore, the cost of these edit operations (t_1, t_2) is defined by (3). If (t_1, t_2) is relabeling, and the nodes t_1 and t_2 have the same label, that is $t_1 = t_2$, then $\delta_{rb}(t_1, t_2) = 0$. Otherwise, the cost of (t_1, t_2) depends on the label of the nodes in t_1 and t_2. Hence, the cost of an operation depends on its nature and the labels of the

TABLE 1. The description of the seven edit operations.

Edit operation	Description
Relabeling (t_1, t_2)	Both t_1 and t_2 denote base pair or unpaired base
Deletion $(t_1, \varnothing)$	t_1 denotes base pair or unpaired base
Insertion $(\varnothing, t_2)$	t_2 denotes base pair or unpaired base
Altering (t_1, t_2)	t_1 denotes a base pair, and t_2 denotes an unpaired base
Completion (t_1, t_2)	t_1 denotes an unpaired base, and t_2 denotes a base pair
Arc-breaking (t_1, t_2)	t_1 denotes a base pair, and t_2 denotes a pair of unpaired bases
Arc-creation (t_1, t_2)	t_1 denotes a pair of unpaired bases, and t_2 denotes base pair

involved nodes.

$$\delta(t_1, t_2) = \begin{cases} 0, & (t_1, t_2) \text{ is relabeling, and } t_1 = t_2 \\ \delta_{rb}(t_1, t_2), & (t_1, t_2) \text{ is relabeling of base, and } t_1 \neq t_2 \\ \delta_{rbp}(t_1, t_2), & (t_1, t_2) \text{ is relabeling of base pair, and } t_1 \neq t_2 \\ \delta_{idb}(t_1, t_2), & (t_1, t_2) \text{ is insertion or deletion of base} \\ \delta_{ids}(t_1, t_2), & (t_1, t_2) \text{ is insertion or deletion of stem} \\ \delta_{idp}(t_1, t_2), & (t_1, t_2) \text{ is insertion or deletion of pseudoknot} \\ \delta_{aco}(t_1, t_2), & (t_1, t_2) \text{ is altering or completion} \\ \delta_{acb}(t_1, t_2), & (t_1, t_2) \text{ is arc-creation or arc-breaking} \end{cases} \tag{3}$$

C. SIMILARITY MEASURE BASED VECTOR-EDIT DISTANCE

It is observed that RNA conserved trees are not ordinary trees, but ordered binary trees. Therefore, they can be compared in order. Calculating the similarity between two trees is generally done in two steps. One is to calculate the similarity of the branched structure of BTMG_{CSP} tree, that is, the minimum vector distance and the edit cost. The second is to calculate the similarity of the BTMG_{CSP} tree, that is the total minimum vector distances of the branches, and the total edit cost for obtaining the total minimum vector distances.

For RNA conserved tree BTMG_{CSP}, there are three edit operations: insert a node, delete a node and merge two nodes. However, merging is arc-breaking or arc-creation for base pairs (i.e. stem or pseudoknots). And the edit cost of deleting or inserting the nodes within the branches is the same. Thus, the cost of arc-breaking or arc-creation of nodes and the cost of deleting redundant nodes are considered in the branches. Let two BTMG_{CSP} trees be T_1 and T_2, and the ith branch of T_1 is B_{1i}, and the jth branch of T_2 is B_{2j}. The computation of the minimum vector distance and edit cost of B_{1i} and B_{2j} can be seen in (4) and (5), respectively.

$$d(B_{1i}, B_{2j}) = \sum_{1 \leq i'}^{n_{B_{1i}}} \min_{1 \leq j' \leq n_{B_{2j}}} \left\{ d\left(t_{ii'}^{B_{1i}}, t_{jj'}^{B_{2j}}\right) \right\} \tag{4}$$

where $n_{B_{1i}} = |B_{1i}|$ denotes the node number of branch B_{1i}, $n_{B_{2j}} = |B_{2j}|$ denotes the node number of branch B_{2j}, and $n_{B_{1i}} \leq n_{B_{2j}}$. $t_{ii'}^{B_{1i}}$ and $t_{jj'}^{B_{2j}}$ are the ith node of B_{1i} and the jth node of B_{2j}, respectively.

$$c\left(B_{1i}, B_{2j}\right) = \sum \delta\left(t_{jj'}^{D}, \varnothing\right) + \sum \delta\left(t_{jj'}^{U}, t_{j(j'+1)}^{U}\right) \tag{5}$$

where $t_{jj'}^{D}$ denotes the node in branch B_{2j} that need to be deleted, and $\delta\left(t_{jj'}^{D}, \varnothing\right)$ is the cost of deleting node $t_{jj'}^{D}$. If $t_{jj'}^{D}$ is stem, $\delta\left(t_{jj'}^{D}, \varnothing\right) = \delta_{ids}\left(t_{jj'}^{D}, \varnothing\right) \times len_{t_{jj'}^{D}}$, and if $t_{jj'}^{D}$ is pseudoknot, $\delta\left(t_{jj'}^{D}, \varnothing\right) = \delta_{idp}\left(t_{jj'}^{D}, \varnothing\right) \times len_{t_{jj'}^{D}}$. $t_{jj'}^{U}$ and $t_{j(j'+1)}^{U}$ are the two nodes that need to be merged in the branch B_{2j}. $\delta\left(t_{jj'}^{U}, t_{j(j'+1)}^{U}\right)$ denotes the arc-creation/arc-breaking cost of merging two nodes, and $\delta\left(t_{jj'}^{U}, t_{j(j'+1)}^{U}\right) = \delta_{acb}\left(t_{jj'}^{U}, t_{j(j'+1)}^{U}\right) \times \left| len_{t_{jj'}^{U}} - len_{t_{j(j'+1)}^{U}} \right|$.

Note that the branches of the BTMG_{CSP} tree are ordered subtrees. Therefore, $\min_{1 \leq j' \leq n_{B_{2j}}} \left\{ d\left(t_{ii'}^{B_{1i}}, t_{jj'}^{B_{2j}}\right) \right\}$ can be computed just by traversing BTMG_{CSP} tree in order. Namely, if $\min_{1 \leq j' \leq n_{B_{2j}}} \left\{ d\left(t_{i1}^{B_{1i}}, t_{jj'}^{B_{2j}}\right) \right\} = d\left(t_{i1}^{B_{1i}}, t_{j2}^{B_{2j}}\right)$, then $\min_{1 \leq j' \leq n_{B_{2j}}} \left\{ d\left(t_{i2}^{B_{1i}}, t_{jj'}^{B_{2j}}\right) \right\}$ can be computed from the next node $t_{j3}^{B_{2j}}$ in the B_{2j} branch. In this way, there is no backtracking during traversal or comparison of BTMG_{CSP} trees. Thereby, it can improve the efficiency.

According to the vector distance of the branches, we can easily find out the most similar subtrees of two BTMG_{CSP} trees, and compute the edit cost for finding the similar subtrees. Thus, the minimum vector distance and the total edit cost between them are obtained, as shown in (6) and (7), respectively.

$$d\left(T_1, T_2\right) = \sum_{1 \leq i \leq n_1} \min_{1 \leq j \leq n_2} \{d(B_{1i}, B_{2j})\} \tag{6}$$

where $n_1 \leq n_2$, and n_1 and n_2 are the branch numbers of T_1 and T_2, respectively.

$$Cost\left(T_1, T_2\right) = \sum \delta\left(B_{2j}^{D}, \varnothing\right) + \sum c\left(B_{1i}, B_{2j}^{min}\right) \tag{7}$$

where B_{2j}^{D} denotes the branch in T_2 that need to be deleted. B_{2j}^{min} denotes the branch that has the minimum vector distance with T_2.

The distance of two BTMG_{CSP} trees includes the distance(vector distance) of the most similar subtree structure and the cost(edit distance) of finding this similar substructure. Therefore, the distance between the two BTMG_{CSP} trees is defined in (8).

$$D\left(T_1, T_2\right) = \left(d\left(T_1, T_2\right) + Cost\left(T_1, T_2\right)\right)/2 \tag{8}$$

The pseudocode describes how to calculate the similarity strategy for BTMG_{CSP} trees in Algorithm 2. There are three distances to be calculated: i) the distance of nodes; ii) the distance of branches; and iii) the distance of trees. Because BTMG_{CSP} trees are ordered binary trees, so both the calculation and comparison are in order. Firstly, given the BTMG_{CSP} trees T_i and T_j, for the branches B_{ik} in T_i and B_{jl} in T_j, the vector distance between the nodes $t_{kk'}^{B_{ik}}$ and $t_{ll'}^{B_{jl}}$, $d\left(t_{kk'}^{B_{ik}}, t_{ll'}^{B_{jl}}\right)$ can be obtained by (1) (lines 2-5). Then the similar nodes with smaller distance can be found by comparing the vector distances of the adjacent nodes. The sum of the vector distances of the similar nodes, that is the minimum vector distance $d(B_{ik}, B_{jl})$ of the branches B_{ik} and B_{jl}, can be calculated by (4) (line 6). While comparing the vector distances of the adjacent nodes, the nodes with lager vector distance would be merged or deleted, according their lengths. And the cost of these edit operations $c(B_{ik}, B_{jl})$ is obtained by (5) (line 7). Further, the most similar subtree of T_i and T_j can be identified,

Algorithm 2 Calculate the Distances Between RNAs

Input: RNAs sequences(QS) and their dot-bracket secondary structure(SS)

Output: The distances between RNAs $D(i, j)$

1: Convert RNAs to BTMG_{CSP} trees by Algorithm 1
2: **for** BTMG_{CSP} trees T_i and T_j **do**
3: //Calculate the branch distance between T_i and T_j
4: **for** the branches B_{ik} in T_i and B_{jl} in T_j **do**
5: Calculate the vector distance $d\left(t_{kk'}^{B_{ik}}, t_{ll'}^{B_{jl}}\right)$ between the nodes $t_{kk'}^{B_{ik}}$ and $t_{ll'}^{B_{jl}}$ in order by Equation (1)
6: Calculate the vector distance $d(B_{ik}, B_{jl})$ between the branches B_{ik} and B_{jl} by Equation (4)
7: Calculate the edit cost $c(B_{ik}, B_{jl})$ by Equation (5)
8: **end for**
9: //Calculate the distance between T_i and T_j
10: Calculate the vector distance between $d(T_i, T_j)$ in order by Equation (6)
11: Calculate the edit cost $Cost(T_i, T_j)$ between T_i and T_j by Equation (7)
12: Calculate the trees distance $D(T_i, T_j)$ by Equation (8)
13: **end for**
14: Return the distances $D(i, j)$ of RNAs

according to the branches with the smaller vector distance. So the vector distance $d(T_i, T_j)$ of T_i and T_j is the sum of the vector distance of these branches calculated by (6) (line 10). And the cost of finding the similar subtree $Cost(T_i, T_j)$ includes two parts as (7): i) the cost of deleting the redundant branches, ii) the cost of the edit operations of the branches in the similar tree (line 11). Finally, the vector-edit distance $D(T_i, T_j)$ of the BTMG_{CSP} trees T_i and T_j can be obtained by (8) (line 12). The vector-edit distances of BTMG_{CSP} trees, that is the distances of RNAs, are used to classify or compare the RNAs.

IV. EXPERIMENTAL RESULTS

To verify the performance of the BTMG_{CSP} tree similarity strategy, the classical ncRNAs dataset with non-pseudoknots and the pseudoknots database PseudoBase are used in experiments. Firstly, ncRNAs sequences and their dot-bracket secondary structures are converted to BTMG_{CSP} trees by Algorithm 1. Secondly, the distances of BTMG_{CSP} trees are calculated by Algorithm 2. Then, ncRNAs are classified and compared according to the distances of BTMG_{CSP} trees.

A. EXPERIMENT USING THE CLUSTER DATA OF ncRNA

The experiment is conducted by using a subset of the selected high scoring structures (https://www.tbi.univie.ac.at/papers/SUPPLEMENTS/ncRNA/lists/selection.html) in a typical ncRNAs dataset with non-pseudoknots [40]. Since the consensus secondary structure of the dataset is computed by RNAalifold. It computes the most likely structure and base pair probabilities for each base pairs. This structure is the results of the multiple sequence alignment, and there will be a certain deviation. Therefore, only the original RNA

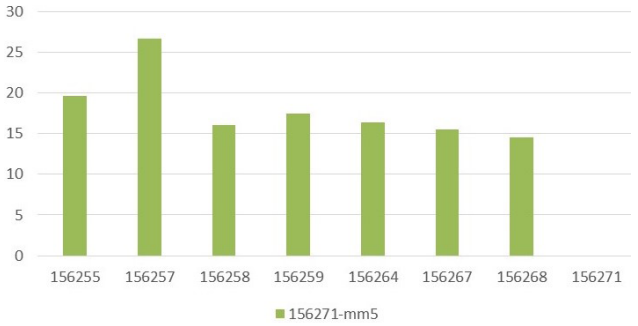


FIGURE 7. The distances from RNA mm5 in structure 156271 to each structure in Cluster 86486.

sequences and their secondary structures in the data set are selected for experiment.

The experiment aims to determine the class of the known structure clusters in ncRNAs dataset that the query specie secondary structure may belong to, according to the distance from the query RNA secondary structure to each known structure, namely classifying them into the smallest distance structure. However, each structure contains several species of RNA sequences and their secondary structures. If one specie RNA secondary structure in a known structure is very similar to the query specie secondary structure, and their distance is 0, it means that the distance from the query specie secondary structure to the known structure is 0. Therefore, to ensure the validity of similarity, a geometric mean method is applied to compute the distance from the query specie RNA secondary structure to the known structure.

Firstly, Cluster 86486 containing 8 structures is selected for analysis. If the mouse RNA sequence in Structure #156271 is used as the query species secondary structure, the distances from this mouse RNA to all the structures of Cluster 86486 is obtained by the proposed similarity strategy. In Fig. 7, the distances are 19.58, 26.64, 16.07, 17.45, 16.4, 15.51, 14.56, 0. It is observed that the distance from this mouse RNA to structure #156271 is 0, because the distance between two BTMG_{CSP} trees converted by the mouse RNA secondary structure and rat RNA secondary structure in structure # 156271 is $D\left(T_{mouse}, T_{156271-rn3}\right) = 0$. Therefore, the query mouse RNA should belong to Structure #156271. This is consistent with the database. Further, the average geometric distances from each species secondary structures in a specific structure to the structure within a cluster are almost the same, as shown in Fig. 8. This not only verified that the RNA secondary structures of the species in the same structure are very similar, but also demonstrated the distances of species in the same structure to other structures are also similar, which is one of the important features in the ncRNAs dataset.

In addition, RNAz Cluster 16165, 25547, 58284, 61845, 80001, 86486 and 113047 from Cluster A-F, which includes sequences of up to eight species: human (hg17), chimp (panTro1), mouse (mm5), rat (rn3), dog (canFam1), chicken (galGal2), zebrafish (danRer1) and fugu (fr1), are selected for experimental validation. The correction rate of classification

TABLE 2. The representation model and structural alignment for $BTMG_{CSP}$ and other methods.

Method	Pseudoknots	Properties of node	Properties of branch	Time complexity
ESL-TAGs	some, simple, short	little	×	$O(N^5)$
PSTAGs	most, simple	little	×	$O(N^5)$
PSMAlign	all, simple	little	×	$O(n^4)$
ASPRAlign	all, simple	little	×	$O(n^2)$
$BTMG_{CSP}$	all, simple/complex	most	✓	$O((n/k)^2)$

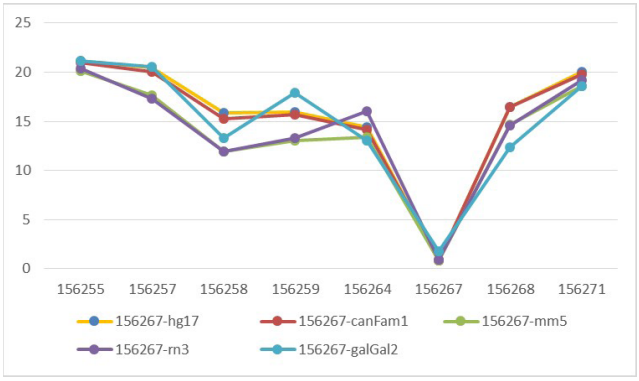


FIGURE 8. The geometric average distance from each species secondary structure in structure 156267 to each structure in Clsuter 86486.

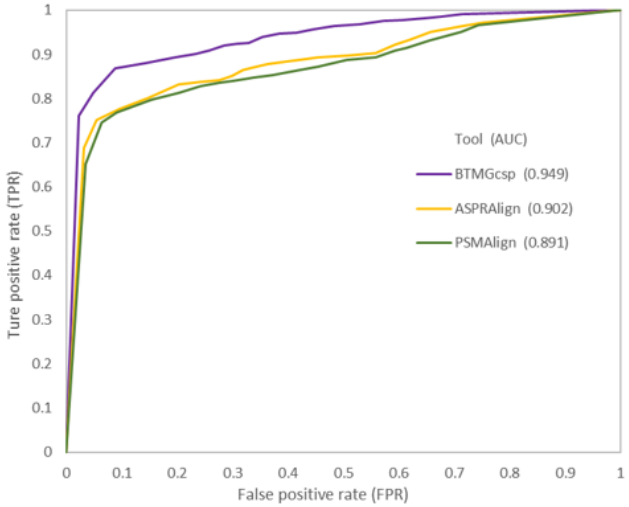


FIGURE 9. ROC curves of PseudoBase of PSMAlign, ASPRAlign and $BTMG_{CSP}$.

according to our similarity strategy is up to 92.5%. For the 438 Structures and 1545 RNAs included in the selection of high scoring structures, the classification accuracy rate can reach up to 86.6%.

B. EXPERIMENT USING PseudoBase

PseudoBase is a typical database of sequences, structures and functions related to RNA pseudoknots (http://www.ekevanbatenburg.nl/PKBASE/PKB.HTML) [41]. So far, the database contains 393 RNA sequences with pseudoknots, and their secondary structures with dot-bracket notation. PseudoBase is different from the previous ncRNAs dataset. To query the classification of an RNA, it is necessary to calculate the distance from the query RNA to all classes in the database. And the query RNA belongs to the class which has the minimum distance.

To avoid the dependence of the classification results on the dataset, the cross-validation is performed by calculating the average of the results in all cases to verify the validity and feasibility of the proposed model and similarity strategy. The RNA in PseudoBase was randomly divided into 10 subsets. One of them was used as the query RNA sequence structure set S, and the rest as the reference sequence structure set R, $S \cap R = \varnothing$. For each RNA in S, we need to calculate the distance to the RNA in the reference set R and find out the RNA with the smallest distance from the query RNA in each class of the reference set. If there are several reference RNAs with the smallest distance from the query RNA in the same class, the reciprocal of the number of these reference RNAs is used as the weight of the query RNA for computing the distance from the query RNA to the class. The calculated distance is the score of the RNA, and then the reference RNA is classified according to the threshold. By continuously changing the threshold, a series of the true positive rate (TPR) and the false positive rate (FPR) are calculated. The receiver operating characteristic (ROC) curve is plotted by all the TPR and FPR, in which TPR is plotted on the Y axis, and FPR is plotted on the X axis. An important performance indicator of the classifier is the area under the ROC curve (AUC). The larger the AUC value is, the better the classifier is. From Fig. 9, we could see that the AUC of $BTMG_{CSP}$ was 0.949, Its performance better than PSMAlign (0.891) and ASPRAlign (0.902) in PseudoBase. Thus, the accuracy of $BTMG_{CSP}$ in classification and comparison of RNA secondary structures with arbitrary pseudoknots was verified.

In addition, since both ESL-TAGs [32] and PSTAGs [33] are modeled based on RNA sequences, their time complexity is $O(N^5)$, where N is the length of the RNA sequence. PSMAlign only focused on the alignment of the base pairs in RNA secondary structures [34], thus its algorithm complexity is less than ESL-TAGs and PSTAGs. However, it uses the directed complete graph with n nodes to represent the stem pattern, where n is the number of RNA base pairs. The number of the edges is $n(n-1)$, and the time complexity of traversing the directed complete graph is $O(n^2)$. Thus, the time complexity of the alignment algorithm based on the base pair of RNA secondary structure using PSMAlign is $O(n^4)$. Recently, ASPRAlign reduces the time complexity to $O(n^2)$ [36]. Because the ordered binary tree $BTMG_{CSP}$ proposed in this paper, The computational complexity of executing $BTMG_{CSP}$ for structural alignment is the same order as that

of parsing an input sequence with ASPRAlign theoretically. More precisely, the range of the depth of the ordered binary tree BTMG_{CSP} is $log_2(n/k)$ to (n/k), where k is the average of the numbers of the base pairs in the conserved stem patterns. The time complexity of comparing two BTMG_{CSP} trees is $O((log_2(n/k))^2)$ to $O((n/k)^2)$. Therefore, the worst-case time complexity is $O((n/k)^2)$ and space complexity is $O(n/k)$.

Summarizing, the representation model and structural alignment of RNA secondary structure for BTMG_{CSP} and other methods are shown as in Table 2

V. CONCLUSION AND DISCUSSION

The function of a particular RNA molecule in an organic system is primarily determined by its structure. However, the existing methods for understanding RNA function by comparing the complex RNA secondary structure including arbitrary pseudoknots are time consuming and expensive. The secondary structures of RNA containing arbitrary pseudoknots are more complicated, which is not conducive to model the RNA structure. Further, thousands of RNA secondary structures are generated by high-throughput detection techniques. So it is necessary to design an efficient similarity strategy to compare complex RNA secondary structures. In this paper, an improved RNA secondary structure grammar tree MG_{RNA} was proposed to model the complex RNA secondary structures. Since the similar secondary structures will share a common conserved stem pattern, BTMG_{CSP} tree represented a conserved stem pattern of RNA is proposed based on MG_{RNA}, which is an ordered binary tree. A high-efficiency similarity strategy based on vector-edit distance is offered to calculate the distance between two BTMG_{CSP} trees. Finally, the effectiveness and feasibility of our method are proved in the comparison of RNA secondary structures.

Our model not only visually and succinctly represents the complex RNA secondary structure containing any type of pseudoknots, but also preserves RNA secondary structural properties. The vector-edit distance method based on can efficiently and accurately compare the complex RNA secondary structures, and naturally tend to classify the corresponding cluster structures in a way that reflects the known secondary structure families. Thus, it is easy to classify the members of RNA secondary structure family that are expected to perform related functions into the same cluster. This is useful for RNA annotation, structure-based phylogeny, homology searches in databases, and identification of new families in RNA populations.

In our future work, this can be extended to compare RNA based on sequence and secondary or tertiary structures. In fact, in addition to pseudoknots, our framework can be adapted to deal with other motifs, such as G4 motifs, sarcin-ricin, kink twist, and so on. Thus, a rich global RNA view can be produced by combining sequence and structural features. It can also be used as a heuristic to help biologists understand their functional roles and biological implications.

REFERENCES

[1] K. C. Baldridge and L. M. Contreras, "Functional implications of ribosomal RNA methylation in response to environmental stress," *Crit. Rev. Biochem. Mol. Biol.*, vol. 49, no. 1, pp. 69–89, Jan. 2014.

[2] S. Higa-Nakamine, T. Suzuki, T. Uechi, A. Chakraborty, Y. Nakajima, M. Nakamura, N. Hirano, T. Suzuki, and N. Kenmochi, "Loss of ribosomal RNA modification causes developmental defects in zebrafish," *Nucleic Acids Res.*, vol. 40, no. 1, pp. 391–398, Jan. 2012.

[3] D. C. Shippy and A. A. Fadl, "RNA modification enzymes encoded by the gid operon: Implications in biology and virulence of bacteria," *Microbial Pathogenesis*, vol. 89, pp. 100–107, Dec. 2015.

[4] J. Yi, S. Li, C. Wang, N. Cao, H. Qu, C. Cheng, Z. Wang, L. Wang, and L. Zhou, "Potential applications of polyphenols on main ncRNAs regulations as novel therapeutic strategy for cancer," *Biomed. Pharmacotherapy*, vol. 113, May 2019, Art. no. 108703.

[5] M. U. Kaikkonen, M. T. Y. Lam, and C. K. Glass, "Non-coding RNAs as regulators of gene expression and epigenetics," *Cardiovascular Res.*, vol. 90, no. 3, pp. 430–440, Jun. 2011.

[6] M. K. Elbashir, M. Ezz, M. Mohammed, and S. S. Saloum, "Lightweight convolutional neural network for breast cancer classification using RNA-seq gene expression data," *IEEE Access*, vol. 7, pp. 185338–185348, Dec. 2019.

[7] M. K. Atianand and K. A. Fitzgerald, "Long non-coding RNAs and control of gene expression in the immune system," *Trends Mol. Med.*, vol. 20, no. 11, pp. 623–631, Nov. 2014.

[8] M. R. Alexander and G. K. Owens, "Epigenetic control of smooth muscle cell differentiation and phenotypic switching in vascular development and disease," *Annu. Rev. Physiol.*, vol. 74, no. 1, pp. 13–40, Mar. 2012.

[9] M. Cesana, D. Cacchiarelli, I. Legnini, T. Santini, O. Sthandier, M. Chinappi, A. Tramontano, and I. Bozzoni, "A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA," *Cell*, vol. 147, no. 2, pp. 358–369, Oct. 2011.

[10] H. Kagami, T. Akutsu, S. Maegawa, H. Hosokawa, and J. C. Nacher, "Determining associations between human diseases and non-coding RNAs with critical roles in network control," *Sci. Rep.*, vol. 5, no. 1, p. 14577, Oct. 2015.

[11] S. Alaimo, R. Giugno, and A. Pulvirenti, "NcPred: NcRNA-disease association prediction through tripartite network-based inference," *Frontiers Bioeng. Biotechnol.*, vol. 2, pp. 1–8, Dec. 2014.

[12] M. Esteller, "Non-coding RNAs in human disease," *Nature Rev. Genet.*, vol. 12, no. 12, pp. 861–874, Nov. 2011.

[13] Q. Chen, D. Lai, W. Lan, X. Wu, B. Chen, Y.-P.-P. Chen, and J. Wang, "ILDMSF: Inferring associations between long non-coding RNA and disease based on multi-similarity fusion," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Aug. 20, 2019, doi: 10.1109/TCBB.2019.2936476.

[14] S. A. Mortimer, M. A. Kidwell, and J. A. Doudna, "Insights into RNA structure and function from genome-wide studies," *Nature Rev. Genet.*, vol. 15, no. 7, pp. 469–479, Jul. 2014.

[15] Q. Chen, C. Lan, B. Chen, L. Wang, J. Li, and C. Zhang, "Exploring consensus RNA substructural patterns using subgraph mining," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 5, pp. 1134–1146, Sep. 2017.

[16] R. Bonasio and R. Shiekhattar, "Regulation of transcription by long non-coding RNAs," *Annu. Rev. Genet.*, vol. 48, no. 1, pp. 433–455, Nov. 2014.

[17] D. Bhartiya and V. Scaria, "Genomic variations in non-coding RNAs: Structure, function and regulation," *Genomics*, vol. 107, nos. 2–3, pp. 59–68, Mar. 2016.

[18] L. E. Vandivier, S. J. Anderson, S. W. Foley, and B. D. Gregory, "The conservation and function of RNA secondary structure in plants," *Annu. Rev. Plant Biol.*, vol. 67, no. 1, pp. 463–488, Apr. 2016.

[19] B. Panwar, A. Arora, and G. P. Raghava, "Prediction and classification of ncRNAs using structural information," *BMC Genomics*, vol. 15, no. 1, pp. 1–13, Feb. 2014.

[20] N. J. Reiter, C. W. Chan, and A. Mondragón, "Emerging structural themes in large RNA molecules," *Current Opinion Struct. Biol.*, vol. 21, no. 3, pp. 319–326, Jun. 2011.

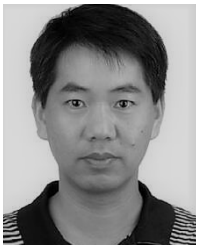
[21] D. H. Turner and D. H. Mathews, "NNDB: The nearest neighbor parameter database for predicting stability of nucleic acid secondary structure," *Nucleic Acids Res.*, vol. 38, pp. D280–D282, Jan. 2010.

[22] C. B. Do, D. A. Woods, and S. Batzoglou, "CONTRAfold: RNA secondary structure prediction without physics-based models," *Bioinformatics*, vol. 22, no. 14, pp. e90–e98, Jul. 2006.

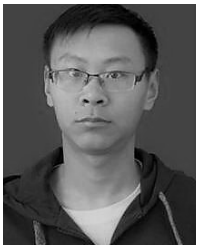
[23] L. G. Scott and M. Hennig, "RNA structure determination by NMR," in *Methods in Molecular Biology*, vol. 278, M. Consens and G. Navarro, Eds. Totowa, NJ, USA: Humana Press, 2008, pp. 29–61.

[24] Z. Li, J. Zhu, X. Xu, and Y. Yao, "RDense: A protein-RNA binding prediction model based on bidirectional recurrent neural network and densely connected convolutional networks," *IEEE Access*, vol. 8, pp. 14588–14605, 2020.
[25] S. Schirmer, Y. Ponty, and R. Giegerich, "Introduction to RNA secondary structure comparison," in *RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods*, vol. 1097, J. Gorodkin and W. L. Ruzzo, Eds. Totowa, NJ, USA: Humana Press, Dec. 2014, pp. 247–273.
[26] A. Denise and P. Rinaudo, "Optimisation problems for pairwise RNA sequence and structure comparison: A brief survey," in *Transactions on Computational Intelligence XIII*, vol. 8342, N.-T. Nguyen and H. A. Le-Thi, Eds. Berlin, Germany: Springer, 2014, pp. 70–82.
[27] Y. Sakakibara, "Pair hidden Markov models on tree structures," *Bioinformatics*, vol. 19, pp. i232–i240, Jul. 2003.
[28] T. Jiang, L. Wang, and K. Zhang, "Alignment of trees—An alternative to tree edit," *Theor. Comput. Sci.*, vol. 143, no. 1, pp. 137–148, Jul. 1995.
[29] M. Möhl, S. Will, and R. Backofen, "Fixed parameter tractable alignment of RNA structures including arbitrary pseudoknots," in *Combinat. Pattern Matching*, vol. 5029, P. Ferragina and G. M. Landau, Eds. Berlin, Germany: Springer, 2008, pp. 69–81.
[30] C. A. Theimer, C. A. Blois, and J. Feigon, "Structure of the human telomerase RNA pseudoknot reveals conserved tertiary interactions essential for function," *Mol. Cell*, vol. 17, no. 5, pp. 671–682, Mar. 2005.
[31] T. Jiang, G. Lin, B. Ma, and K. Zhang, "A general edit distance between RNA structures," *J. Comput. Biol.*, vol. 9, no. 2, pp. 371–388, Apr. 2002.
[32] Y. Uemura, A. Hasegawa, S. Kobayashi, and T. Yokomori, "Tree adjoining grammars for RNA structure prediction," *Theor. Comput. Sci.*, vol. 210, no. 2, pp. 277–303, Jan. 1999.
[33] H. Matsui, K. Sato, and Y. Sakakibara, "Pair stochastic tree adjoining grammars for aligning and predicting pseudoknot RNA structures," *Bioinformatics*, vol. 21, no. 11, pp. 2611–2617, Jun. 2005.
[34] J. K. H. Chiu and Y.-P. P. Chen, "Pairwise RNA secondary structure alignment with conserved stem pattern," *Bioinformatics*, vol. 31, no. 24, pp. 3914–3921, Aug. 2015.
[35] Q. Chen, Y.-P.-P. Chen, and C. Zhang, "Interval-based similarity for classifying conserved RNA secondary structures," *IEEE Intell. Syst.*, vol. 31, no. 3, pp. 78–85, May 2016.
[36] M. Quadrini, L. Tesei, and E. Merelli, "ASPRAlign: A tool for the alignment of RNA secondary structures with arbitrary pseudoknots," *Bioinformatics*, vol. 36, no. 11, pp. 3578–3579, Mar. 2020.
[37] C. J. K. Ho and Y. P. Phoebe, "A comprehensive study of RNA secondary structure alignment algorithms," *Brief. Bioinform.*, vol. 18, no. 2, pp. 291–305, Mar. 2017.
[38] Y. Byun and K. Han, "PseudoViewer3: Generating planar drawings of large-scale RNA structures with pseudoknots," *Bioinformatics*, vol. 25, no. 11, pp. 1435–1437, Jun. 2009.
[39] V. Guignon, C. Chauve, and S. Hamel, "An edit distance between RNA stem-loops," in *String Processing and Information Retrieval*, vol. 3772, M. Consens and G. Navarro, Eds. Berlin, Germany: Springer, 2005, pp. 335–347.
[40] S. Washietl, I. L. Hofacker, M. Lukasser, A. Hüttenhofer, and P. F. Stadler, "Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome," *Nat. Biotechnol.*, vol. 23, no. 11, p. 1383, Nov. 2005.
[41] F. H. D. van Batenburg, "PseudoBase: A database with RNA pseudoknots," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 201–204, Jan. 2000.

LIYU HUANG received the B.E. degree in information and computing science from Guangxi University for Nationalities, Nanning, China, in 2005, and the M.S. degree in computer software and theory from Guangxi Normal University, Guilin, China, in 2010. She is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. She is also an Engineer with the Information Network Center, Guangxi University, Nanning. Her current research interests include bioinformatics and data mining.



QINGFENG CHEN received the B.Sc. and M.Sc. degrees in mathematics from Guangxi Normal University, China, in 1995 and 1998, respectively, and the Ph.D. degree in computer science from the University of Technology Sydney, in September 2004. He is currently a Professor with Guangxi University, China, and a hundred talent program of Guangxi. He has published 40 refereed papers and two monographs by Springer, including the IEEE TRANSACTION ON KNOWLEDGE AND DATA ENGINEERING and *Data Mining and Knowledge Discovery*. His research interests include bioinformatics, data mining, and artificial intelligence. He is a co-chair for several international conferences. He has been serving as an Associate Editor for *Engineering Letters*. He is invited to be a guest editor of two special issues for Current Protein & Peptide Science.



YONGJIE LI is currently pursuing the M.S. degree with the School of Computer, Electronic, and Information, Guangxi University, China. His research interests include machine learning and data mining.

CHENG LUO is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. She is also a Lecturer with the Xingjian College of Science and Liberal Arts, Guangxi University, Nanning, China. Her research interests include bioinformatics and neural networks.