Discovery of Structural and Functional Features in RNA Pseudoknots

Qingfeng Chen and Yi-Ping Phoebe Chen, *Senior Member*, *IEEE*

Abstract—An RNA pseudoknot consists of nonnested double-stranded stems connected by single-stranded loops. There is increasing recognition that RNA pseudoknots are one of the most prevalent RNA structures and fulfill a diverse set of biological roles within cells, and there is an expanding rate of studies into RNA pseudoknotted structures as well as increasing allocation of function. These not only produce valuable structural data but also facilitate an understanding of structural and functional characteristics in RNA molecules. PseudoBase is a database providing structural, functional, and sequence data related to RNA pseudoknots. To capture the features of RNA pseudoknots, we present a novel framework using quantitative association rule mining to analyze the pseudoknot data. The derived rules are classified into specified association groups regarding structure, function, and category of RNA pseudoknots. The discovered association rules assist biologists in filtering out significant knowledge of structure-function and structure-category relationships. A brief biological interpretation to the relationships is presented, and their potential correlations with each other are highlighted.

Index Terms—RNA pseudoknots, stem, loop, association rule mining, PseudoBase, H-pseudoknot, function, structure, partition.

1 Introduction

Accurately predicting the functions of biological macromolecules is one of the biggest challenges in functional genomics. Whereas the protein folding problem is difficult because the local secondary and nonlocal tertiary contacts both contribute to the stability of the final native folds in RNA, it is the secondary structure (base-pairing interactions) that has more influences on the final fold rather than tertiary contacts. Thus, studies of structural information on RNA can be an alternative to understand structure-function relationships in biology.

RNA molecules play a central role in a number of biological functions within cells, from the transfer of genetic information from DNA to protein, to enzymatic catalysis. To fulfill this range of functions, a simple linear nucleotide string of RNA including uracil, guanine, cytosine, and adenine, forms a variety of complex three-dimensional structures. One of the most prevalent structures adopted by RNA molecules is a commonly-occurring structural motif known as the pseudoknot that was first recognized in the turnip yellow mosaic virus in 1982 [25].

A pseudoknot is an RNA structure that involves base pairing between a loop, formed by an orthodox secondary structure, and some region outside this loop [37]. Although several distinct folding topologies of pseudoknots exist, the simplest or classical pseudoknot is the H-pseudoknot (Fig. 1a) that is formed by the pairing of a region in the hairpin loop with the bases outside the hairpin. The H-pseudoknot minimally consists of two stems (*stem* 1 and *stem* 2) and two loops (*loop* 1 and *loop* 2). If the two stems form a quasi-consecutive helix, the base stacking at the junction becomes possible; otherwise, an additional loop (*loop* 3 in Fig. 1b) might be created. The single stranded loop regions may contain hundreds of nucleotides and often interact with adjacent stems, and hence, a relatively simple fold can yield complex and stable RNA structures. Moreover, due to the variation of the lengths of loops and stems and their base composition, as well as each other's interactions, pseudoknots show a diverse set of roles in biology. Thus, a comprehensive understanding of the functions of RNA molecules requires knowledge of their structures.

RNA pseudoknots are viewed as essential elements of the topology of many structural RNAs such as ribosomal RNAs or ribozymes. They have been found in most organisms and comprise functional domains within ribozymes, self-splicing introns, ribonucleoprotein complexes, viral genomes, and many other biological systems [1], [10]. The in-depth knowledge of a psuedoknot's structure provides a better insight into understanding their psuedoknot's functions in varied organisms.

Studies of molecular RNA previously focused on the prediction of RNA secondary structures [4], [41] by using comparative RNA analysis [15], [26], [27] or approximating the free energy of any given structure [35]. Many techniques using the Tinoco model [19], [40] or thermodynamics-based energy minimization algorithm [41] such as well-accepted parameters of Turner et al. [14], aim to discover the structure of optimal score.

Regardless of continued work to increase the prediction accuracy [5], [18], [32] by improving on the parameters, the

- *Q. Chen is with the Faculty of Science and Technology, Deakin University, VIC 3125, Australia. E-mail: qifengch@deakin.edu.au.*
- *Y.-P.P. Chen is with the Faculty of Science and Technology, Deakin University, 221 Burwood Highway, VIC 3125, Australia, and with the ARC Centre of Excellence in Bioinformatics, Melbourne, Australia. E-mail: phoebe@deakin.edu.au.*

Manuscript received 8 May 2008; revised 25 Sept. 2008; accepted 12 Nov. 2008; published online 25 Nov. 2008.
Recommended for acceptance by J. Dix.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2008-05-0247.
Digital Object Identifier no. 10.1109/TKDE.2008.231.

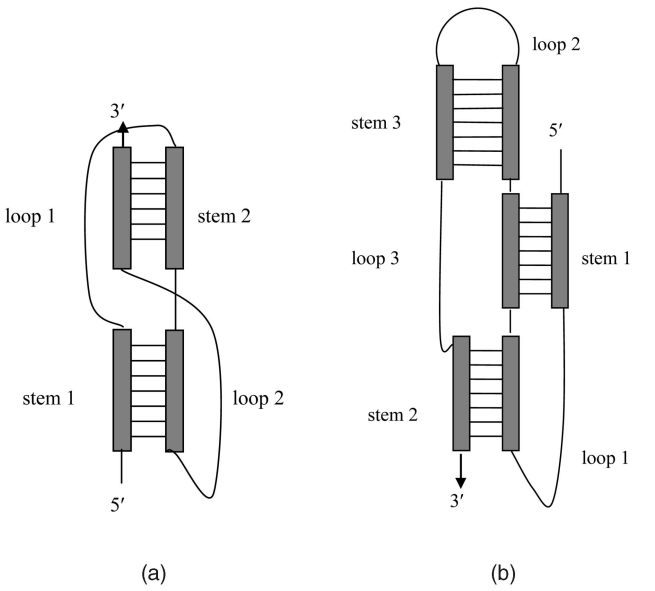

Fig. 1. RNA pseudoknot architecture. (a) Classical H-type pseudoknot fold. (b) Three-loops RNA pseudoknot fold.

accuracy of predictions using the Tinoco model can never reach 100 percent and the constraint on nested secondary structures is applied. However, the problem of RNA structure prediction can be attributed not only to inaccurate thermodynamic parameters but also pseudoknot formation. Although more than 95 percent of the base pairs do not contain any pseudoknots at all, nearly all RNA molecules have one or more pseudoknots. As a result, there have been multiple known algorithms to predict RNA pseudoknots, such as heuristic modelling [11] and RNA sampler [38] for generating accurate structural information in RNA molecules.

PseudoBase [23] is the only online database containing structural, functional, and sequence data of RNA pseudoknots, allowing us to investigate deeply into structure-function relationships in RNA molecules. Unfortunately, the analysis of this valuable data set is underdeveloped due to the difficulty in modeling and complexity in computing. It is difficult to predict structural and functional features of pseudoknots by only analyzing individual experimental results. Thus, association rule mining that has been successfully used to discover valuable information from a large amount of data [39] can be used to analyze this data set.

Recently, many applications using data mining have been reported in analyzing various biological data sets [6], [7], [8]. Most of them [16], [31], however, show limitations in handling the data with multivalued variables including categorical multivalued valuables (such as color {red, blue, green}) and quantitative multivalued variables (such as weight {[40, 50], [50, 75]}). A model is proposed in [39] to identify quantitative association rules, in which the domain of multivalued variables is partitioned into intervals. An association rule is represented as $X \Rightarrow Y$ along with a conditional probability matrix $M_{Y|X}$ according to Bayesian rules.

In the present study, we develop a framework to identify potential top-k covering rule groups in RNA pseudoknots, including the relationships of structure-function and structure-category and significant ratios of stems and loops. The relationships are captured by using an intuitive conditional probability matrix. It allows users to regulate k and the $minsupp$ threshold and compare between rules in the same group. The domain of quantitative attributes is divided by a novel point-based partition schema. Further, the performance evaluation demonstrates the advantages of our miner in handling high dimensional data. The results by 0.1 (*minsupp*) in contrast to the results by 0.2 are presented in the analysis. The identified distributions (sizes and nucleotide composition) of stems and loops indicate the interactions between loops and stems and account for the reproduction of a variety of pseudoknots. In particular, the identified ratios imply the role of pseudoknots in the promotion of function efficiency. We also observe that there is discipline of sizes and base composition (stems and loops) in specific organisms. A brief interpretation of the obtained rules is presented. Furthermore, a deep study builds connections between these rules and enhances the understanding of structure-function relationships in RNA pseudoknots of various organisms.

2 Motivations

Traditional association rule mining has been widely and successfully used to identify frequent patterns from general data sets. However, it is unfit for the data that contains multivalued variables [39]. It has been argued that the former mining approach depended on two thresholds and a conditional probability matrix can be helpful for association studies due to its impressive expressiveness. However, if the item variable X impacts on variable Y at only a few point values, item-based association mining and quantitative association rule mining may be more appropriate and efficient than this method.

The previous techniques can only identify rules among simple variables, such as *tea* $\rightarrow$ *sugar* or *state* $\rightarrow$ *united*. They have limitations in discovering rules among multivalued variables from large databases and for representing them. For example, in Fig. 1, *stem* and *loop* are categorical multivalued variables. They have a range of categories of {*stem* 1, *stem* 2} and {*loop* 1, *loop* 3}, respectively. The sizes of stems and loops are quantitative multivalued variables that are represented as a collection of intervals such as {(0, 1], (1, 2], (2, 3]}. The size distribution of stems and loops are discrepant. For example, the size of *stem* 1 in PseudoBase varies between 0 and 22 only, whereas the size of *loop* 3 is between 0 and 890. Thus, it is necessary to generate a common partition. Therefore, this urges us to develop new methods to address the relationships among these multivalued variables.

Usually, we may obtain a number of rules in traditional association rule mining. However, it is not easy to sort those rules that are ranked higher than the others. Furthermore, some interesting rules might be missed or redundant rules were generated due to an inappropriate threshold. Thus, this paper extends and adapts traditional association rule mining by representing a rule as the form of $X \rightarrow Y$ in conjunction with a probability matrix $M_{Y|X}$ in terms of Bayesian rules.

This captures the relationship that the presence of X results in the occurrence of Y. $M_{Y|X}$ is defined as

$$M_{Y|X} \triangleq P(Y = y|X = x) = \begin{pmatrix} p(y_1|x_1) & p(y_2|x_1) & \dots & p(y_n|x_1) \\ p(y_1|x_2) & p(y_2|x_2) & \dots & p(y_n|x_2) \\ & \dots & & \\ p(y_1|x_m) & p(y_2|x_m) & \dots & p(y_n|x_m) \end{pmatrix},$$

where $p(y_j|x_i)$ represents the conditional probability, $i = 1, 2, \dots, m$, and $j = 1, 2, \dots, n$. x_i and y_j represent a categorical item and a quantitative item, such as *stem* 1 and a size interval (1, 2], respectively. The matrix comprises a group of association rules that correspond to a specified characteristic relationship of RNA pseudoknots. Each column consists of a subgroup of the rules corresponding to a categorical attribute variable x_i. Thus, this assists us in extracting the most significant rules in each subgroup separately rather than identifying the rules from the whole group.

A high-dimensional dataset can result in many redundant rules and long mining process [8], and makes it difficult to sort out interesting information from databases. These challenges block the analysis of the pseudoknot data. To address this critical problem, we propose a novel mining method to identify the most significant top-*k* covering rule groups. As mentioned in [8], it is easier and semantically clearer to choose *k* than minimum confidence. Moreover, it avoids missing interesting rules and generating too many redundant rules. A natural alternative to our model is to set different values of *k*, and compare each other's results.

3 Materials and Methods

Pseudoknot Data. Suppose S_1, S_2, L_1, L_2, and L_3 represent *stem* 1, *stem* 2, *loop* 1, *loop* 2, and *loop* 3, respectively; A, G, C, and U represent base *adenine*, *guanine*, *cytosine*, and *uracil*, respectively; the abbreviation *vr*, *vt*, *vf*, *v3*, *v5*, *vo*, *rr*, *mr*, *tm*, *ri*, *ap*, *ot*, and *ar*, denote *viral ribosomal readthrough signals*, *viral tRNA-like structures*, *viral ribosomal frameshifting signals*, *other viral 3′-UTR*, *other viral 5′-UTR*, *viral others*, *rRNA*, *mRNA*, *tmRNA*, *ribozymes*, *aptamers*, *artifical molecules* and *others*, respectively; and *ss*, *tc* and *fs* represent *self-splicing*, *translation control* and *viral frameshifting*, respectively. Let uppercase X and Y be multivalued attribute valuables, lowercase x and y be items, $p(X)$ be the probability of some event X and $p(Y|X)$ be the conditional probability of some event Y, given the occurrence of some other event X, and *minsupp* be the minimum support in the context.

The data here is collected from PseudoBase that includes the whole pesudoknot data from the publications in Medline, and can be reached at http://wwwbio.LeidenUniv.nl/~Batenburg/PKB.html. Originally, each pseudoknot is recorded by 12 data items, such as *PKB number*, *EMBL number* and *reference*, whereas some of them are not useful for data mining application. Thus, only *organism*, *RNA type* and *bracket view of structure* are considered in this paper. Furthermore, the structural information is classified by two stems and three loops, including their corresponding nucleotide sequence and size.

TABLE 1
An Example of Pseudoknot Data

C	S_1	*Sequence*	S_2	*Sequence*	L_1	*Sequence*
Vt	3	CCC	6	UCCUGC	2	CC
V3	3	CCU	5	GUCUC	1	U
V3	3	CUU	4	GGCU	1	U
Vf	6	GGGGGG	3	GCG	5	ACUUA

After removing eight unusual pseudoknots, seven reductant pseudoknots [1], and five pseudoknots that have loop lengths ≥ 200, a data set consisting of 225 H-pseudoknots is obtained. Within the 225 unique H-pseudoknots, 170, or 76 percent, have $L_2 = 0$; 22, or 10 percent have $L_2 = 1$; 8, or 4 percent have $L_2 = 2$; 1, or 0.4 percent has $L_2 = 3$. In particular, *loop* 2 is ignored here since the most studied type of pseudoknot is with coaxial stacking of stems so that *loop* 2 is absent.

Table 1 presents an example of the structures, classes, and functions of pseudoknots in PseudoBase. Each row in the table represents an RNA pseudoknot. The nonnegative integers denote the number of nucleotides. It can be seen at http://www.deakin.edu.au/~qifengch/rna/pesudoknot/pseudoknot.zip in more details.

Partition of Attributes. Suppose {*class*, *function*, *stem*, *loop*, *base*, *ratio*, *length*} denotes the domain of attributes of PseudoBase. The first six elements are viewed as categorical attributes, and the last one is a quantitative attribute. We thus propose a novel partition in conjunction with the properties of pseudoknot data and top-*k* rule groups. A categorical attribute has a number of categories, such as hair color including *blonde*, *brown* and *black*. According to the specification in PseudoBase, the partition of domain of categorical attributes *class*, *function*, *stem*, *loop*, *base*, and *ratio* is defined as follows:

- $class = \{vr, vt, vf, v3, v5, vo, rr, mr, tm, ri, ap, ot, ar\}$,
- $function = \{ss, tc, fs\}$,
- $stem = \{S_1, S_2\}$,
- $loop = \{L_1, L_2, L_3\}$,
- $base = \{A, C, G, U\}$, and
- $ratio = \{S_1/S_2, L_1/L_3, S_1/L_1, S_2/L_1, S_1/L_3, S_2/L_3\}$.

Unlike the categorical attributes, the domain of quantitative attribute has to be partitioned into intervals. The partition usually needs to determine 1) the number of intervals and 2) the size of each interval. Although PseudoBase provides an initial partition of base length for each stem and loop, they are actually inconsistent with each other. For example, in the initial partition, (14, 15] is included in *stem* 1, *stem* 2, and *loop* 1 but not in *loop* 3. Nevertheless, as mentioned in [39], a unified partition is required to generate the probability matrix.

The equidepth partitioning model proposed by Agrawal [31] is an alternative method for causality mining. The number of partitions is defined as $Number\ of\ Intervals = \frac{2n}{m(K-1)}$, in which n represents the number of quantitative attributes, m represents the minimum support, and K represents the partial completeness level. However, it is inappropriate for sparse data sets and might include much unnecessary information. Thus, traditional partition of

TABLE 2
Distribution of Stem Sizes of Pseudoknots

Stem 1	*Number*	*Stem* 1	*Number*	*Stem* 2	*Number*
0	0	10	5	0, 1, 2	0
1	0	11	7	3	5
2	0	12	3	4	33
3	77	13	6	5	66
4	42	14	6	6	69
5	24	16	3	7	36
6	14	17	3	8	9
7	8	18	3	9	5
8	10	19	3	10	1
9	10	22	1	33	1

variables like equal interval width and equal frequency might lead to inaccurate or uninteresting results of data mining.

Definition 3.1. *Suppose a quantitative attribute y is divided into a set of intervals $\{y_1, \ldots, y_n\}$ (called base intervals) using the categorical item x_i such that for any base interval y_j, y_j consists of a single value for $1 \leq j \leq n$.*

- $|y_i| = 1$, $1 \leq i \leq n$, *and*
- *for* $\forall l \neq k$, *and* $1 \leq l, k \leq n$, $y_l \cap y_k = \emptyset$.

Suppose $y_{1i}, \ldots$ and y_{mi} represent the partition using the categorical item x_i in ascending order of their maximum sizes. The partition starts from the categorical item with the minimum of maximum sizes, and integrates it with the next one until all items are gone through. The partition using x_i is defined as $\{(y_{1i}, max(y_{2i})], \ldots, (max(y_{m-1i}), max(y_{mi})]\}$. Table 2 presents the distribution of sizes of *stem* 1 and *stem* 2 of pseudoknots in PseudoBase.

The initial partition for each attribute variable actually comprises a collection of ranges. It simply includes one element in each range since each point-value may imply an important structural feature of pseudoknots. For example, according to Definition 3.1, the partition starts from stem 1 and is presented in Table 2 as $Y_1 = \{0, (0, 1], (1, 2], (2, 3], (3, 4], (4, 5], (5, 6], (6, 7], (7, 8], (8, 9], (9, 10], (10, 11], (11, 12], (12, 13], (13, 14], (14, 15], (15, 16], (16, 17], (17, 18], (18, 19], (19, 20], (20, 21], (21, 22]\}$. Nevertheless, it is observed that (19, 20] and (20, 21] are not recorded with *stem* 1. They will be combined with (21, 22] if we cannot find them in the partition of other attribute variables. Thus, the final partition needs to consider all attribute variables and integrate their partitions together.

Definition 3.2. *Suppose $Y_i = \{y_{1i}, \ldots, y_{mi}\}$ and $Y_{i+1} = \{y_{1i+1}, \ldots, y_{ni+1}\}$ are two adjacent partitions. Let $Y = \emptyset$. The integration of them is defined as*

- $Y = Y \cup x$, *if* $x \in Y_i \cup Y_{i+1}$, *and* $|x| = 1$;
- $Y = Y \cup x \cup \cdots \cup x_c$, *if* $|x| = 0$, $|x_c| = 1$, $x \in Y_i$ *and* $x \notin Y_{i+1}$, $max(x) < max(x_c) \leq max(y_{mi})$; *and*
- $Y = Y \cup x \cup \cdots \cup x_c$, *if* $|x| = 0$, $|x_c| = 1$, $x \in Y_{i+1}$ *and* $x \notin Y_i$, $max(x) < max(x_c) \leq max(y_{ni+1})$;

where x_c is the closest point value to x that includes one element in the range, and max() represents the function of maximum. There might be more than one categorical item that has the equivalent maximum size with one another. In this extreme case, it will be reported to the user, rather than selecting them randomly. Suppose $Y = \{y_1, \ldots, y_k\}$ is the final partition after integration. The partition y_k might be a large interval that includes unfrequent or missing structures (no record). For example, there is just one record of stem 1 of size 22 and there is no record of stem 1 from size 20 to 21 at all. In that case, it is reasonable to combine these point values to a partition instead of listing all of them one by one.

In a similar manner, we can obtain the partition for stem 2 as $Y_2 = \{0, (0, 1], (1, 2], (2, 3], (3, 4], (4, 5], (5, 6], (6, 7], (7, 8], (8, 9], (9, 10], (10, 11], (11, 12], (12, 13], (13, 14], (14, 15], \ldots, (31, 32], (32, 33]\}$, where 33 denotes the maximum size of *stem* 2. This will be integrated with the partition of stem 1 in terms of Definition 3.2. As a result, the integrated partition of Y_1 and Y_2 is $\{0, (0, 1], (1, 2], (2, 3], (3, 4], (4, 5], (5, 6], (6, 7], (7, 8], (8, 9], (9, 10], (10, 11], (11, 12], (12, 13], (13, 14], (14, 15], (15, 16], (16, 17], (17, 18], (18, 19], (19, 22], (22, 33]\}$.

The partition scheme of *length* in this study adopts the point-based decomposition of quantitative attributes. In the similar way, we can generate partition for loop lengths. In comparison, the values of ratio attributes are positive real numbers rather than integers. Thus, the condition $|y_i| = 1$ in Definition 3.1 needs to be changed to $|y_i| = 1$ or $|y_i| = 0.5$. Accordingly, $|x| = 1$ and $|x_c| = 1$ in Definition 3.2 are changed to $|x| = 1$ and $|x_c| = 1$ or $|x| = 0.5$ and $|x_c| = 0.5$. These aim to avoid missing interesting knowledge.

Generation of rule groups. Based on the partitioned variables, we then work out the conditional probabilities for X and Y in the probability matrix below. Therefore, we can determine the conditional probability of $Y = y_i$, given $X = x_i$, as $p(y_i|x_i) = p(x_i|y_i) * p(y_i)/p(x_i)$.

For example, x and y represent *stem* 1 of pseudoknots, and the size interval (3, 4] of *stem* 1, respectively. By Table 2, we have $n = 225$ and $p(x = stem1) = 225/225 = 1$. Additionally, the number of pseudoknots containing *stem* 1 with four nucleotides is equal to 42, and we have $p(y = (3, 4] \wedge x = stem\ 1) = 42/225 = 0.19$. In the same way, we have $p(y = (3, 4] | x = stem\ 1) = p(y = (3, 4] \wedge x = stem\ 1)/p(x = stem\ 1) = 0.19$. As a result, we are able to compute the entire conditional probabilities of *stem* 1, namely [$p(y_1$ |*stem* 1) $p(y_2$ | *stem* 1) ... $p(y_n|$ *stem* 1)], where y_j denotes the jth size interval by partition.

In a similar manner, the conditional probabilities of *stem* 2, *loop* 1, and *loop* 3, can be computed. Thus, we have

$$M_{Y|X} = \begin{pmatrix} p(y_1|stem\ 1) & p(y_2|stem\ 1) & \ldots & p(y_n|stem\ 1) \\ p(y_1|stem\ 2) & p(y_2|stem\ 2) & \ldots & p(y_n|stem\ 2) \\ p(y_1|loop\ 1) & p(y_2|loop\ 1) & \ldots & p(y_n|loop\ 1) \\ p(y_1|loop\ 3) & p(y_2|loop\ 3) & \ldots & p(y_n|loop\ 3) \end{pmatrix}.$$

We can generate other matrixes in terms of different associations. As mentioned above, there must be enough point pairs (x_i, y_j) in the conditional probability matrix $M_{Y|X}$ that satisfy the conditions of valid rules. In contrast to traditional *minimum confidence*, this paper uses a flexible way to allow users to have the ability to control the number of rules in each rule group.

Suppose $M_{Y|X}$ corresponding to an association ***AS*** consists of a set of rows $\{r_1, \ldots, r_n\}$. Let $A = \{A_1, \ldots, A_m\}$ be the complete set of antecedent items of ***AS***, and

TABLE 3
Top-2 Rule Group of *Stem* 1/*Loop* 3 and *Stem* 2/*Loop* 3

Ratio	Interval	Percentage
Stem 1/*Loop* 3	[1, 1.5]	34
Stem 1/*Loop* 3	[0.5, 1)	32
Stem 2/*Loop* 3	[0, 0.5)	22
Stem 2/*Loop* 3	[2, 2.5)	19

$C = \{C_1, \ldots, C_k\}$ be the complete set of consequent items of ***AS***, then each row r includes an antecedent item from A and a set of consequent items from C. As a mapping between rows and items, given a row r_i, we define ***PS*** (Point-pairs Support Set) as the set of point-pairs whose conditional probabilities are not equal to zero, namely $PS(x) = \{(x, y_j) | y_j \in C, p(y_j|x) \neq 0\}$.

Definition 3.3 (Rule group). *Let $G_x = \{x \rightarrow C_j \mid (x, C_j) \in PS(x)\}$ be a rule group with an antecedent item x and consequent support set C.*

It is observed that the rules from different rule groups might have different supports and confidences. Moreover, there might be different numbers of valid rules derived from different groups. The top-*k* covering rule groups are thus applied to encapsulate the most significant association of the data set while enabling users to control the number of rules in a convenient manner.

Definition 3.4. *Let R_i: $X \rightarrow Y_i$ and R_j: $X \rightarrow Y_j$ be two valid rules with respect to a given categorical item X. **Top-k covering rule group** is the subset of the union of rule groups where $1 \leq k \leq k_{max}$ and k_{max} is the upper bound of the number of rules we would like to find. A rule is of interest if, and only if, it is in the **top-k covering rule group**. R_i is ranked higher than R_j if $p(Y_i|X) > p(Y_j|X)$.*

Example 3.1. In Table 2, we have $k_{max} = 21$ due to 21 intervals of *stem* 1. As a result, we have **top-1 covering rule group** = {*stem* 1 $\rightarrow$ (2, 3], *stem* 2 $\rightarrow$ (5, 6]} and **top-2 covering rule group** = {*stem* 1 $\rightarrow$ (2, 3], *stem* 1 $\rightarrow$ (3, 4], *stem* 2 $\rightarrow$ (5, 6], *stem* 2 $\rightarrow$ (4, 5]}. The rule *stem* 1 $\rightarrow$ (2, 3] is given higher ranking than *stem* 1 $\rightarrow$ (3, 4] due to its higher support in the conditional probability matrix.

4 Results

To identify top-*k* covering rule group, we need to set up the values of parameters. According to the assumed associations, $k_{max1} = 46$ is the maximum of *k* of the size domain. According to Definition 3.1, we have $k_{max2} = 8$ for the association of base composition. In a similar way, we have $k_{max3} = 27$ for the association of ratios between stems or loops. We can obtain different numbers of rules by regulating the values of $k_1 \leq k_{max1}$, $k_2 \leq k_{max2}$, and $k_3 \leq k_{max3}$. The interesting rules will be determined by k_1, k_2, or k_3 in combination with the specified minimum support.

After partition, we need to construct the matrix in terms of the conditional probability of point pairs, such as (*stem* 1, (0, 1]). Each association AS_i consists of $|A| \times |C|$ initial rules, in which $|A|$ and $|C|$ represent the size of the antecedent item set and consequent item set of AS_i, respectively. Nevertheless, not all of them have prominent statistical significance. Moreover, searching in a large number of uninteresting or redundant rules may result in excessive and expensive computation. Thus, top-*k* covering rule groups are used to search for the dominant rules and brought into comparison with other rules in the same group.

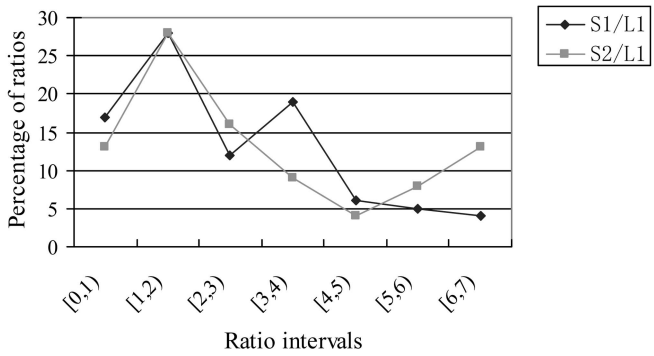

Fig. 2. The distribution of stem 1/loop 1 and stem 2/loop 1.

In practice, we may need to vary *minsupp* and *k* in terms of different associations. For simplicity, we only discuss the results by $k = 4$ in this paper. Moreover, given *k*, we compare the difference in case of varied *minsupp*.

By comparison, we observe that there is no sharp drop in rule output when assigning the *minsupp* from 0.1 to 0.2. Thus, the corresponding results by 0.1 in contrast to the results by 0.2 are selected in the following analysis. Based on the selected *k*, there are 13 rules in AS_1 (sizes of stems and loops) and 16 rules in AS_2 (base composition of stems and loops). Moreover, AS_3 (classes and sizes), AS_4 (classes and base composition), AS_5 (functions and sizes), AS_6 (function and base composition), and AS_7 (ratios of stems and loops) consist of subassociations in terms of different classes and functions of pseudoknots.

Table 3 presents a random example of the significant ratios between stems and *loop* 3. The distribution of *stem* 1/*loop* 1 and *stem* 2/*loop* 1, *stem* 1/*loop* 3 and *stem* 2/*loop* 3, and *stem* 1/*stem* 2 and *loop* 1/*loop* 3 is shown in Fig. 2, Fig. 3, and Fig. 4, respectively. Such structure features have not been reported before and may play an important role in prompting the efficiency of functions.

Moreover, we also identify some new correlations that have been unknown previously. Table 4 shows the rule groups about functions, including self-splicing, translation

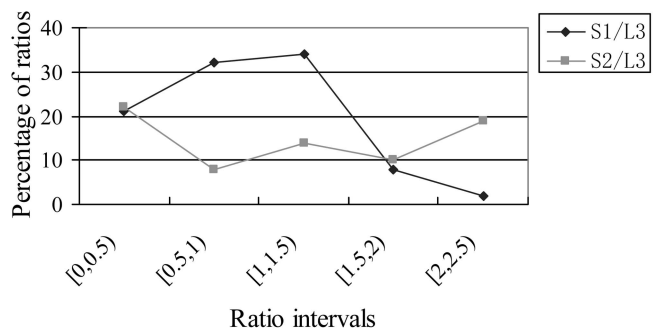

Fig. 3. The distribution of stem 1/loop 3 and stem 2/loop 3.

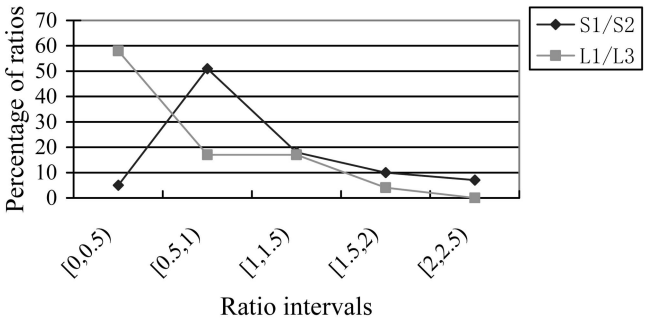

Fig. 4. The distribution of stem 1/stem 2 and loop 1/loop 3.

control, and frameshifting, and Table 5 presents part of the rule groups about pseudoknot classes.

The derived rules assist us in understanding the structure-function relationship in pseudoknots. The rule groups not only confirm the previously observed results (AS_1) in [1] but also discover interesting pseudoknot properties such as AS_3, AS_4, and AS_7 that have not been reported before.

The details can be seen in the following interpretation. As for the other rule groups, such as the rules regarding pseudoknot functions, the details can be reached by http://www.deakin.edu.au/~qifengch/rna/pseudoknot/causality.zip.

Performance evaluation. In contrast to data mining, there are some inherent limitations to Bayesian methods, including computational complexity and the quality and extent of the prior beliefs. It is only useful as this prior knowledge is reliable. Thus, the Bayesian method is assumption-driven in the sense that a hypothesis is formed and validated against the data. However, the learning of prior belief is an NP-complete problem in case of enormous dataset. The structure-function correlations are usually hidden in pseudoknot data with multivalued variables. These prevent us from obtaining reliable prior knowledge. Furthermore, some associations are not obvious (undetermined) and might be ignored from the assumption. This may result in missing interesting knowledge of RNA pseudoknots. Therefore, we turn to association rule mining, a data-driven method, in terms of the available pseudoknot data and the potential structure-function correlations commented on by our collaborators.

All tests reported herein were performed on a 1.86 GHz Intel Core(TM)2 PC. The parameters including *applicability*, *Top-k rule group*, *minsupp*, *frequent patterns* and *CPU Times*(s) are selected as the comparison metric, so as to assess the efficiency of algorithms while using the same data set. The comparison is implemented among three related algorithms. Although there may be other algorithms such as FPtree [17] to identify associations, they are not included since they are inappropriate to identify rule groups.

Our miner (kTOP) extends the Local Causal Discovery (LCD) method [9] to discover association rules among multivalued variables from PseudoBase. Moreover, we adapt the proposed method in [39] using top-*k* covering rule group instead of enumerating all potential correlations. This is able to avoid not only the huge number of rules owing to the high-dimensional pesudonkot data set, but also a long mining process due to large number of rules. Table 6 shows a performance comparison between our miner and algorithms LCD [9] and PPM (Probability Partition Matrix) [39]. In the comparison, we identify the

TABLE 4
Rule Groups Regarding Pseudoknot Functions, Size, and Base Composition

Functions	size of stems	size of loops	base composition of stems	base composition of loops
self-splicing	stem 1 = 7, 60%	loop 1 = 3, 20%	$0.2 <$ adenine of stem $1 \leq 0.3$, 40%	$0.2 <$ adenine of loop $1 \leq 0.3$, 60%
	stem 2 = 8, 40%	loop 3 = 7, 20%	$0.3 <$ adenine of stem $2 \leq 0.4$, 40%	$0.2 <$ cytosine of loop $1 \leq 0.3$, 60%
			$0.1 <$ cytosine of stem $1 \leq 0.2$, 40%	
			$0.2 <$ cytosine of stem $2 \leq 0.3$, 40%	
			$0.2 <$ guanine of stem $1 \leq 0.3$, 40%	$0.2 <$ guanine of loop $1 \leq 0.3$, 40%
			$0.3 <$ guanine of stem $2 \leq 0.4$, 40%	$0.1 <$ guanine of loop $3 \leq 0.2$, 40%
			$0.1 <$ uracil of stem $1 \leq 0.2$, 50%	$0.2 <$ uracil of loop $1 \leq 0.3$, 40%
			$0.3 <$ uracil of stem $2 \leq 0.4$, 40%	
translation control	stem 1 = 3, 39%	loop 1 = 1, 35%	$0.2 <$ adenine of stem $1 \leq 0.3$, 21%	$0.3 <$ adenine of loop $1 \leq 0.4$, 11%
	stem 2 = 5, 32%	loop 3 = 3, 28%	$0.1 <$ adenine of stem $2 \leq 0.2$, 38%	$0.4 <$ adenine of loop $3 \leq 0.5$, 21%
			$0.3 <$ cytosine of stem $1 \leq 0.4$, 23%	$0.1 <$ cytosine of loop $3 \leq 0.2$, 16%
			$0.1 <$ cytosine of stem $2 \leq 0.2$, 30%	
			$0.3 <$ guanine of stem $1 \leq 0.4$, 29%	$0.7 <$ guanine of loop $1 \leq 1$, 14%
			$0.1 <$ guanine of stem $2 \leq 0.2$, 32%	$0.1 <$ guanine of loop $3 \leq 0.2$, 21%
			$0.3 <$ uracil of stem $1 \leq 0.4$, 29%	$0.7 <$ uracil of loop $1 \leq 1$, 25%
			$0.1 <$ uracil of stem $2 \leq 0.2$, 29%	$0.3 <$ uracil of loop $1 \leq 0.4$, 31%
frameshifting	stem 1 = 5, 28%	loop 1 = 2, 44%	$0.1 <$ adenine of stem $1 \leq 0.2$, 16%	$0.4 <$ adenine of loop $1 \leq 0.5$, 24%
	stem 2 = 4, 28%		$0.1 <$ adenine of stem $2 \leq 0.2$, 20%	$0.7 <$ adenine of loop $3 \leq 1$, 20%
			$0.1 <$ cytosine of stem $1 \leq 0.2$, 28%	$0.4 <$ cytosine of loop $1 \leq 0.5$, 32%
			$0.4 <$ cytosine of stem $2 \leq 0.5$, 36%	$0.1 <$ cytosine of loop $3 \leq 0.2$, 44%
			$0.7 <$ guanine of stem $1 \leq 1$, 32%	$0.4 <$ guanine of loop $1 \leq 0.5$, 20%
			$0.2 <$ guanine of stem $2 \leq 0.3$, 28%	$0.1 <$ guanine of loop $3 \leq 0.2$, 36%
			$0 <$ uracil of stem $1 \leq 0.1$, 20%	$0.1 <$ uracil of loop $3 \leq 0.2$, 28%
			$0.1 <$ uracil of stem $2 \leq 0.2$, 32%	

TABLE 5
Rule Groups Regarding Pseudoknot Classes, Size, and Base Composition

Classes	size of stems	size of loops	base composition of stems	base composition of loops
other viral 3′-UTR	stem 1 = 3, 47%	loop 1 = 1, 52%	$0.2 <$ adenine of stem $1 \leq 0.3$, 23%	$0.5 <$ adenine of loop $3 \leq 0.6$, 30%
	stem 2 = 6, 46%	loop 3 = 3, 45%	$0.1 <$ adenine of stem $2 \leq 0.2$, 39%	$0.1 <$ cytosine of loop $3 \leq 0.2$, 25%
			$0.3 <$ cytosine of stem $1 \leq 0.4$, 33%	
			$0.1 <$ cytosine of stem $2 \leq 0.2$, 40%	
			$0.3 <$ guanine of stem $1 \leq 0.4$, 37%	$0.7 <$ guanine of loop $1 \leq 1$, 25%
			$0.1 <$ guanine of stem $2 \leq 0.2$, 37%	$0.1 <$ guanine of loop $3 \leq 0.2$, 16%
			$0.3 <$ uracil of stem $1 \leq 0.4$, 35%	$0.7 <$ uracil of loop $1 \leq 1$, 37%
			$0.3 <$ uracil of stem $2 \leq 0.4$, 29%	$0.3 <$ uracil of loop $3 \leq 0.4$, 53%
viral tRNA like structure	stem 1 = 3, 65%	loop 1 = 3, 41%	$0.2 <$ adenine of stem $1 \leq 0.3$, 12%	$0.3 <$ adenine of loop $1 \leq 0.4$, 18%
	stem 2 = 5, 41%	loop 3 = 3, 43%	$0.1 <$ adenine of stem $2 \leq 0.2$, 29%	$0.3 <$ adenine of loop $3 \leq 0.4$, 33%
			$0.7 <$ cytosine of stem $1 \leq 1$, 33%	$0.3 <$ cytosine of loop $1 \leq 0.4$, 12%
			$0.3 <$ cytosine of stem $2 \leq 0.4$ 24%	$0.3 <$ cytosine of loop $3 \leq 0.4$, 16%
			$0.3 <$ guanine of stem $1 \leq 0.4$, 31%	$0.4 <$ guanine of loop $1 \leq 0.5$, 12%
			$0.1 <$ guanine of stem $2 \leq 0.2$, 33%	$0.2 <$ guanine of loop $3 \leq 0.3$, 10%
			$0.3 <$ uracil of stem $1 \leq 0.4$, 33%	$0.3 <$ uracil of loop $1 \leq 0.4$, 16%
			$0.3 <$ uracil of stem $2 \leq 0.4$, 31%	$0.3 <$ uracil of loop $1 \leq 0.4$, 29%

rules regarding the lengths of stems and loops using a dataset OPMV from PseudoBase at http://www.deakin.edu.au/~qifengch/rna/pseudoknot/comparison.zip. Note that the number of rules in Table 6 include all possible rules in theory. Some of them can be pruned if the minimum support or k is applied.

The comparison shows that kTOP has better performance than LCD and PPM methods, and can still have a short process for a small minimum support. In Table 6, the number of obtained frequent patterns from kTOP is 16 in comparison with the number (varied from 19 to 49) of PPM even using a small minimum support. The derived patterns assist in understanding structure-function relationships in RNA pseudoknots. The relevance of the obtained rules to the problems that need to be solved are described in the late interpretation and Section 5. Thus, kTOP assists biologists in sorting out the most significant or interesting biological knowledge. From the observation, both LCD and PPM show limitations in high-dimensional data, which may lead to the long process and huge number of rules even with rather high minimum support and confidence threshold. kTOP has almost the same running time as PPM in case of the low-dimensional data and a low-level minimum support within [0.01, 0.05], but shows an acceleration when a high-dimensional dataset is used.

As mentioned above, this paper uses a point-based decomposition for quantitative attributes in contrast to the optimization-based partition of PPM [39]. The latter has to find the bad quantitative items that result in missing valid rules, decompose these item variables, and compose the good item variables. This is complex because it aims to find an optimized partition for the domain of all attributes (*categorical* and *quantitative*). However, the categorical attributes in pseudoknot data are already partitioned. Thus, the partition of categorical attributes in PMM should be ignored to deal with the pseudoknot data. However, even so, PMM may generate reductant rules or miss interesting rules. In terms of the personnel data set at a university [39], the domain of *Education* can be divided into {*Doctor*, *Master*, *UnderMaster*} or {*Doctor*, *Master*}. We use the same configuration of *minsupp* = 0.6. In either cases, kTOP can obtain the same results by regulating k. However, the rule *Education* = *Doctor* $\rightarrow$ *2,100* $\leq$ *Salary* $<$ 3,500 is removed by PMM but reserved by our miner. The top-k rules enable a flexible comparison between rules in the same group.

The miner kTOP requires users to specify the minimum support threshold and the number of top covering groups, k, only. Such improvement is useful because it is not easy to select an appropriate confidence threshold while the choice of k is semantically clear. It provides users the flexibility to control the output and balance between two extremes [8]. Usually, only a rule from each row can be obtained by the rule induction algorithms like a decision tree, which could miss interesting rules. And, too many redundant rules covering the same rows can be found by traditional association rule mining algorithms. Moreover, our method includes some extra processes to facilitate the identification of association rules. The experiments found a number of interesting rules regarding structure-function relationship of pseudoknots. Most of them were unknown previously. These can benefit in the understanding of the occurring structure motifs in RNA, such as RNA folding, and a

TABLE 6
Performance Comparison in Identification of Rules

Miner	Data set	Applicability	Top-k rule group	minsupp	Frequent patterns	CPU Time(s)
kTOP	*OPMV*	*Yes*	Yes	[0.01, 0.05]	16	[4.5, 380]
LCD	*OPMV*	*No*	No	N/A	N/A	N/A
PPM	*OPMV*	*Yes*	No	[0.01, 0.05]	[19, 49]	[9.5, 380]

TABLE 7
Selected Rules from the Above Rule Groups

Association rules
1. *stem* 1 $\rightarrow$ *stem* 1 = 3 with support 34%
2. *stem* 1 $\rightarrow$ *stem* 1 = 4 with support 19%
3. *stem* 2 $\rightarrow$ *stem* 2 = 6 with support 31%
4. *loop* 1 $\rightarrow$ *loop* 1 = 1 with support 32%
5. *loop* 3 $\rightarrow$ *loop* 3 = 3 with support 24%
6. *stem* 1 $\rightarrow$ 0.2 $<$ *adenine* $\leq$ 0.3 with support 19%
7. *stem* 1 $\rightarrow$ 0.3 $<$ *guanine* $\leq$ 0.4 with support 26%
8. *stem* 1 $\rightarrow$ 0.3 $<$ *cytosine* $\leq$ 0.4 with support 23%
9. *stem* 1 $\rightarrow$ 0.3 $<$ *uracil* $\leq$ 0.4 with support 26%
10.*loop* 1 $\rightarrow$ 0.7 $<$ *uracil* $\leq$ 1 with support 23%
11. *loop* 3 $\rightarrow$ 0.3 $<$ *adenine* $\leq$ 0.4 with support 20%
12. *mRNA* $\rightarrow$ *stem* 1 = 6 with support 22%
13. *mRNA* $\rightarrow$ *stem* 2 = 7 with support 44%
14. *mRNA* $\rightarrow$ 0.1 $<$ *adenine* in *stem* 1 $\leq$ 0.2 with support 75%
15. *mRNA* $\rightarrow$ 0.1 $<$ *guanine* in *stem* 1 $\leq$ 0.2 with support 44%
16. *mRNA* $\rightarrow$ 0.3 $<$ *cytosine* in *stem* 1 $\leq$ 0.4 with support 33%
17. *mRNA* $\rightarrow$ 0.3 $<$ *uracil* in *stem* 1 $\leq$ 0.4 with support 33%
18. *translation control* $\rightarrow$ *stem* 1 = 3 with support 41%
19. *translation control* $\rightarrow$ 0.2 $<$ *adenine* in *stem* 1 $\leq$ 0.3 with support 21%

number of RNA functions, such as ribosomal frameshifting, translation control, and splicing.

Interpretation. Table 7 presents a subset of the originally derived association rules by using 0.1 (*minimum support*) and 4 (*number of top covering rule*). The rules are selected from rule groups and have dominant support in each subgroup. For example, *rules* 1 and 2 are from AS_1 and present the top-2 rules in the subgroup with respect to the length of *stem* 1; and the rules 6, 7, 8, and 9 are from AS_2 and indicate the most significant rules of the subgroups regarding *adenine*, *guanine*, *cytosine* and *uracil*, respectively.

These rules not only present that in most of simple pseudoknots their stems and loops favor different numbers of nucleotides and different base compositions, but also indicate that potential associations may exist between category and pseudoknot structure, and between function and pseudoknot structure. Moreover, several significant ratios regarding stems and loops are reported. For example, rule 12 shows that in most of cases, the number of nucleotides of pseudoknot of *mRNA* (messenger RNA that is transcribed from a DNA template, and carries coding information to the sites of protein synthesis: the ribosomes) may peak at six base pairs. In a similar manner, the remaining rules in Table 7 can be interpreted. The rules, in fact, unveil the structural features of RNA pseudoknots and potential structure-function relationship.

The rules about AS_1 and AS_2 demonstrate previous work in a more comprehensive and accurate way. Especially, the rules (AS_3, AS_4, AS_5, AS_6, AS_7) that were unknown previously will be highlighted, and specific comparisons will be conducted between stems, between loops, between different classes, and between different functions, respectively.

Moreover, this paper provides a novel facility to predict some potential correlations by combining several association rules together, which can be left for biologists to examine in the future experiments. By doing so, it is able to generate new biological knowledge. Some recent studies [21], [33] also mention such information but do not provide semantically clear interpretation for the potential correlations.

Looking at *rule* 1, *rule* 3, *rule* 4, and *rule* 5, there are discrepant leading numbers of nucleotides between *stem* 1 and *stem* 2 and between loop 1 and loop 3. These characteristics (asymmetry) may arise from the difference in tertiary interactions between stems and loops [33]. The difference of the sizes of stems and loops, as well as the types of interaction between them, mean that pseudoknots represent a structurally diverse group. It is necessary that they play diverse roles in biology such as forming the catalytic core of various ribozymes [24] and self-splicing introns [3], and altering gene expression of many viruses by inducing ribosomal frameshifting [30]. The generated leading rule is a novel point of this paper because this assists in not only understanding the properties of stems and loops, but also providing an intuitive and quantified comparison to their difference. *Rule* 2 can be a supplement to demonstrate the difference between *stem* 1 and *stem* 2. Looking at rules 6-11, there is apparent bias of base composition in the loops of H-pseudoknots. The facts of *adenine*-rich in *loop* 3 and *uracil*-rich in *loop* 1 are coherent with results of [1], [2], [21].

The remaining rules in Table 7 are novel and can be classified into two categories in terms of different purposes. *Rules* 12 and 13 describe the correlations between pesudoknot categories and the size of stems. *Rules* 14, 15, 16, and 17 describe the associations between pseudoknot classes and the base composition in *stem* 1. Especially, the associations between size and class, the associations between base composition and class, and the ratios between stems or loops have not been reported by previous pseudoknot studies.

Looking at *rule* 12 and *rule* 13, the pseudoknots of *mRNA* favor six base pairs in *stem* 1, but peak at seven base pairs in *stem* 2. Such rules can be viewed as a secondary evidence in determining pseudoknots' categories, predicting the

size distribution of specific class of pseudoknots and understanding the association between structure and function. Looking at *rules* 14, 15, 16, and 17, they show that *stem* 1 of *mRNA* has a high percentage of *adenine* rather than *cytosine*, *guanine* and *uracil*.

In a similar way, we can predict the size distribution and base composition for other pseudoknot categories, such as *other viral* $3'$-UTR and *viral tRNA-like structure* in Table 5. The pseudoknots of *other viral* $3'$-UTR favor three base pairs, six base pairs, one base pair and three base pairs in *stem* 1, *stem* 2, *loop* 1, and *loop* 3, respectively. Such rules can be viewed as a secondary evidence in determining pseudoknots' categories, predicting the size distribution of specific class of pseudoknots and understanding the association between structure and function. Looking at its dependencies regarding base composition, they show that stem 1 of *other viral* $3'$-UTR has a high percentage of guanine rather than adenine, cytosine, and uracil. Although the *other viral* $3'$-UTR has the same percentages of uracil and cytosine as guanine, the support of guanine in the dependency is a little higher than the percentage of uracil and cytosine in *stem* 1. Thus, we determine that the *stem* 1 of *other viral* $3'$-UTR is guanine-rich. The observation is consistent with reports that GC-rich *stem* 1 (many DNA sequences carry long stretches of repeated G and C which often indicate a gene-rich region) presents resistance to chemical cleavage. This makes *stem* 1 appear to be remarkably stable. On the other hand, there is a preference for the G in the $5'$ end of the stem [34] and a number of pseudoknots with G-rich stretch may be more effective in frameshifting [20]. Looking at the dependencies regarding base composition of loops, we cannot obtain the rules between L_1 and adenine and between L_1 and cytosine due to insufficient support from the current data set.

Looking at the dependencies of *viral tRNA-like structure*, it also peaks at three base pairs of *stem* 1 and three base pairs of *loop* 3 as *other viral* $3'$-UTR, whereas it favors three base pairs of *loop* 1 and five base pairs of *stem* 2. As to the base composition of *viral tRNA-like structure*, it has a high percentage of cytosine of stem 1, high percentage of uracil of *stem* 2, high percentage of uracil of *loop* 1, and high percentage of adenine of *loop* 3. As mentioned above, *stem* 1 is stabilized due to abundant G-C base pairs. A stable pseudoknot structure is important for both aminoacylation and transcription. Moreover, GC-rich *stem* 1 rather than A-U rich may increase the transcription efficiency. It was reported that the mutation in *stem* 1 by changing specific G-C base pair into an A-U base pair reduced the transcription efficiency [12]. These features may help explain the reports of flexible tertiary contacts between stems and loops. Thus, the results in this paper not only discover the structural properties of RNA pseudoknots in specific organisms, but also aid in understanding structure-function relationships in RNA molecules.

Looking at the last two rules in Table 7, most *stem* 1 in a pseudoknot that plays a role in translation regulation usually, has three base pairs by *rule* 18. This may indicate that efficient translation control depended upon the presence of a close three base pair; pseudoknots with a shorter or longer stem 1 were either nonfunctional or had reduced translational efficiency. *Rule* 19 represents the percentage of adenine of *stem* 1 in a pseudoknot for translational regulation peaks at 20 percent to 30 percent. In comparison with the compositions of guanine, uracil, and cytosine in *stem* 1, such *stem* 1 has a high percentage of uracil.

These observations also indicate that RNA pseudoknots are critical for specific protein binding. A number of proteins bind to a pseudoknot in its mRNA, which result in autoregulation [29]. In the rule groups of mRNA, we can see GC-rich is prevalent. Usually, the major loop is likely to be flexible. However, the stable structures with a flexible major loop also indicate the possibility that they can fold in a precise pattern when in contact with a protein. This may imply a motif in the pseudoknot that may show interaction with specific mRNA. For example, the CUGGG motif in the human prion pseudoknot that was also found in the loop of HIV TAR RNA has been proved to interact with human prion mRNA [28]. Moreover, the structural flexibility (flexible loop and neutral interaction) at helical junctions due to U-rich *loop* 1 and A-rich *loop* 3 may be important for proper telomerase function and regulation of protein binding.

In particular, Fig. 2, Fig. 3, and Fig. 4 present novel and significant ratios of stems and loops, which may have relation to functions. We observe that the ratio of S_1/L_1 peaks at the interval [1, 2). Its number decreases in the consequent intervals. This phenomenon can be seen in both frameshifting-related and translation control-related RNA pseudknots. We also observe similar discipline with respect to S_2/L_1, S_1/L_3, and S_2/L_3. As we know, the folding of a RNA pseudoknot requires that loops span the helix of stems. If we altered the length of *stem* 1 or loop 1, it is possible that the consequent change in ratio of stem length to stem helix length may have an effect on function efficiency. A further understanding of these ratios needs to be demonstrated in future biological experiments.

Furthermore, we can predict some novel correlations from the obtained association rules. For example, as for *rule* 18, if we find any pseudoknot whose *stem* 1 peaks at three base pairs, we may predict its functions according to *rule* 18 in Table 7. Thus, the newly generated association rules can be used to complement the prediction of pseudoknots' functions. We may also predict the function of pseudoknots in terms of the rules like *rule* 19. For example, if we find a pseudoknot whose *stem* 1 is cytosine-rich and favors 20 percent to 30 percent of adenine, it may be translational regulation relevant. In practice, we may need to consider the composition of other bases together to enhance its reliability. The experimental results demonstrate that our approach not only can discover meaningful biological patterns but also can facilitate the analysis for biologists by purposely controlling the number of interesting patterns.

5 Conclusion and Discussion

As an important functional structure, the pseudoknot is more highly constrained by nonlocal base pairs and presents specific three-dimensional geometries. Such nonlocal contacts make pseudoknot problem NP-complete. A number of pseudoknot algorithms [13], [26] have been developed but searched for only a subset of pseudoknots.

Many algorithms have been developed to identify correlations. However, they may either generate a large number of rules or miss interesting rules due to inappropriate threshold. This may prevent us from selecting the most significant knowledge. It is inflexible for users to find the top-ranked rules in a specific group and compare with each other. In particular, unlike general data, the genome data contain not only the sequence data but also structural information. Thus, it is important to develop methods to address these critical issues.

We have focused our attention here on the interpretation of the most significant rules in each specified rule group. If more rules are considered together, a further understanding of pseudoknot's structure and function can be achieved. Moreover, we may need to seek more data to support some rules with lightly weak support from current pseudoknot data. We did not touch the biased base composition at the end (3′ side) of loop 3 and at the start (5′ side) of loop 3. It may be an interesting problem to interpret the tertiary interactions between loops and the grooves helices, such as why a loop interacts more with the minor groove of a stem, or why a loop interacts less with the major groove of a stem. The rules about function indicate that a shorter or longer stem or loop, or a stem or loop with irregular base composition may make pesudoknots nonfunctional or have reduced function efficiency. We attempt to generate new knowledge by the combination of rules. Extending this idea to more complex and more realistic scenarios is therefore desirable, but it would require a large data set and evaluation of their soundness.

This paper aims to analyze increasingly available RNA pseudoknot data and identifies interesting patterns from PseudoBase. The obtained rule groups reveal the structural properties of pseudoknots and imply potential structure-function and structure-class relationships in RNA molecules. Moreover, the interpretation of rules demonstrates their significance in the sense of biology.

ACKNOWLEDGMENTS

The work reported in this paper was partially supported by the Australian Research Council's Discovery Project under Grant DP0559251. The authors would like to thank Professor F.H.D. van Batenburg (Leiden University) for his expertise in pseudoknots and Professor Wen-Hsiung Li (University of Chicago) for his time and useful feedback to increase the quality of this paper.

REFERENCES

[1] D.P. Aalberts and N.O. Hodas, "Asymmetry in RNA Pseudoknots: Observation and Theory," *Nucleic Acids Research,* vol. 33, no. 7, pp. 2210-2214, 2005.

[2] D.P. Aalberts, J.M. Parman, and N.L. Goddard, "Single-Strand Stacking Free Energy from DNA Beacon Kinetics," *Biophysical J.,* vol. 84, pp. 3212-3217, 2003.

[3] P.L. Adams, M.R. Stahley, A.B. Kosek, J. Wang, and S.A. Strobel, "Crystal Structure of a Self-Splicing Group I Intron with Both Exons," *Nature,* vol. 430, no. 6995, pp. 45-50, 2004.

[4] K. Bjarne and H. Jotun, "RNA Secondary Structure Prediction Using Stochastic Context-Free Grammars and Evolutionary History," *Bioinformatics,* vol. 15, no. 6, pp. 446-454, 1999.

[5] J.L. Chen and C.W. Greiger, "Functional Analysis of the Pseudoknot Structure in Human Telomerase RNA," *Proc. Nat'l Academy of Sciences USA,* vol. 102, no. 23, pp. 8080-8085, 2005.

[6] Q.F. Chen and Y.P.P. Chen, "Mining Frequent Patterns for AMP-activated Protein Kinase Regulation on Skeletal Muscle," *BMC Bioinformatics,* vol. 7, pp. 1-14, 2006.

[7] Q.F. Chen, Y.P.P. Chen, and C.Q. Zhang, "Detecting Inconsistency in Biological Molecular Databases using Ontology," *Data Mining and Knowledge Discovery,* vol. 15, pp. 275-296, 2007.

[8] G. Cong, K.L. Tan, K.H. Anthony, T. Xin, and X. Xu, "Mining Top-K Covering Rule Groups for Gene Expression Data," *Proc. 2005 ACM SIGMOD Int'l Conf. Management of Data,* pp. 670-681, 2005.

[9] G. Coope, "A Simple Constraint-Based Algorithm for Efficiently Mining Observational Databases for Causal Relationships," *Data Mining and Knowledge Discovery,* vol. 1, no. 2, pp. 203-224, 1997.

[10] W.S. David and E.B. Samuel, "Pseudoknots: RNA Structures with Diverse Functions." *PloS Biology,* vol. 3, no. 6, pp. 956-959, 2005.

[11] W.K. Dawson, K. Fujiwara, and G. Kawai, "Prediction of RNA Pseudoknots Using Heuristic Modeling with Mapping and Sequential Folding," *PLoS ONE,* vol. 2, no. 9, 2007.

[12] B.A.L.M. Deiman, R.M. Kortlever, and C.W.A. Pleij, "The Role of the Pseudoknot at the 3′ End of Turnip Yellow Virus RNA in Minus-Strand Synthesis by the Viral RNA-Dependent RNA Polymerse," *J. Virology,* pp. 5990-5996, 1997.

[13] R.M. Dirks and N.A. Pierce, "A Partition Function Algorithm for Nucleic Acid Secondary Structure Including Pseudoknots," *J. Computational Chemistry,* vol. 24, pp. 1664-1677, 2003.

[14] S. Freier, R. Kierzek, J. Jaeger, N. Sugimoto, M. Caruthers, T. Neilson, and D. Turner, "Improved Free-Energy Parameters for Predictions of RNA Duplex Stability," *Proc. Nat'l Academy of Sciences USA,* vol. 83, no. 24, pp. 9373-9377, 1986.

[15] P. Gardner and R. Giegeric, "A Comprehensive Comparison of Comparative RNA Structure Prediction Approaches," *BMC Bioinformatics,* vol. 5, pp. 140, 2004.

[16] J.W. Han, Y. Cai, and N. Cercon, "Data-Driven Discovery of Quantitative Rules in Relational Databases," *IEEE Trans. Knowledge and Data Eng.,* vol. 5, no. 1, pp. 29-40, Feb. 1993.

[17] J.W. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns Without Candidate Generation," *Proc. ACM SIGMOD Int'l Conf. Management of Data,* pp. 1-12, 2000.

[18] R.B. Lyngso and C.N. Pedersen, "RNA Pseudoknot Prediction in Energy-Based Models," *J. Computational Biology,* vol. 7, no. 3, pp. 409-427, 2000.

[19] R.B. Lyngso, M. Zuker, and C.N.S. Pederse, "Fast Evaluation of Internal Loops in RNA Secondary Structure Prediction," *Bioinformatics,* vol. 15, no. 6, pp. 440-445, 1999.

[20] S. Napthine, J. Liphardt, A. Bloys, S. Routledge, and I. Brierley, "The Role of RNA Pseidoknot Stem 1 Length in the Promotion of Efficient—1 Ribosomal Frameshifting," *J. Molecular Biology,* vol. 288, pp. 305-320, 1999.

[21] P.L. Nixon and D.P. Giedroc, "Energetics of a Strongly pH Dependent RNA Tertiary Structure in a Frameshifting Pseudoknot," *J. Molecular Biology,* vol. 296, pp. 659-671, 2000.

[22] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann Publishers, 1988.

[23] C.W. Pleij, K. Rietveld, and L. Bosch, "A New Principle of RNA Folding Based on Pseudoknotting," *Nucleic Acids Research,* vol. 13, no. 5, pp. 1717-1731, 1985.

[24] T. Rastogi, T.L. Beattie, J.E. Olive, and R.A. Collin, "A Long-Range Pseudoknot is Required for Activity of the Neurospora VS Ribozyme," *EMBO J.,* vol. 15, no. 11, pp. 2820-2825, 1996.

[25] K. Rietveld, R. Van Poelgeest, C.W. Pleij, J.H. Van Boom, and L. Bosch, "The tRNA-Like Structure at the 3′ Terminus of Turnip Yellow Mosaic Virus RNA. Differences and Similarities with Canonical tRNA," *Nucleic Acids Research,* vol. 10, pp. 1929-1946, 1982.

[26] E. Rivas and S.R. Edd, "The Language of RNA: a Formal Grammar that Includes Pseudoknots," *Bioinformatics,* vol. 16, no. 4, pp. 334-340, 2000.

[27] J. Ruan, G.D. Stormo, and W. Zhang, "An Iterated Loop Matching Approach to the Prediction of RNA Secondary Structures with Pseudoknots," *Bioinformatics,* vol. 20, no. 1, pp. 58-66, 2004.

[28] U. Scheffer, T. Okamoto, J.M.S. Forrest, P.G. Rytik, W.E.G. Muller, and H.C. Schrode, "Interaction of 68-kDa TAR RNA-Binding Protein and Other Cellular Proteins with Prion Protein-RNA Stem-Loop," *J. Neurovirology,* vol. 1, pp. 391-398, 1995.

[29] Y. Links Shamoo, A. Tam, W.H. Konigsberg, and K.R. Williams, "Translational Repression by the Bacteriophage T4 Gene 32 Protein Involves Specific Recognition of an RNA Pseudoknot Structure," *J. Molecular Biology,* vol. 232, no. 1, pp. 89-104, 1993.
[30] L.X. Shen and I. Tinoc, "The Structure of an RNA Pseudoknot that Causes Efficient Frameshifting in Mouse Mammary Tumor Virus," *J. Molecular Biology,* vol. 247, no. 5, pp. 963-978, 1995.
[31] R. Srikant and R. Agrawal, "Mining Quantitative Association Rules in Large Relational Tables," *Proc. 1996 ACM SIGMOD Int'l Conf. Management of Data,* pp. 1-12, 1996.
[32] D.W. Staple and S.E. Butche, "Pseudoknots: RNA Structures with Diverse Functions," *PLoS Biology,* vol. 3, no. 6, pp. 956-959, 2005.
[33] S.A. Strobe, "Biochemical Identification of A-Minor Motifs Within RNA Tertiary Structure by Interference Analysis," *Biochemical Soc. Trans.,* vol. 30, pp. 1126-1131, 2002.
[34] E.B. ten Dam, C.W.A. Pleij, and L. Bosch, "RNA Pseudoknots: Translational Frameshifting and Readthrough on Viral RNAs," *Virus Genes,* vol. 4, pp. 121-136, 1990.
[35] I. Tinoco, P.N. Borer, B. Dengler, M.D. Levine, O.C. Uhlenbeck, D.M. Crothers, and J. Grall, "Improved Estimation of Secondary Structure in Ribonucleic Acids," *Nature New Biology,* vol. 246, no. 150, pp. 40-41, 1973.
[36] F.H. van Batenburg, A.P. Gultyaev, and C.W. Plei, "PseudoBase: Structural Information on RNA Pseudoknots," *Nucleic Acids Research,* vol. 29, no. 1, pp. 194-195, 2001.
[37] F.H. van Batenburg, A.P. Gultyaev, C.W. Pleij, J. Ng, and J. Iliehoek, "PseudoBase: A Database with RNA Pseudoknots," *Nucleic Acids Research,* vol. 28, no. 1, pp. 201-204, 2000.
[38] X. Xu, Y. Ji, and G.D. Stormo, "RNA Sampler: A New Sampling Based Algorithm for Common RNA Secondary Structure Prediction and Structural Alignment," *Bioinformatics,* vol. 23, no. 15, pp. 1883-1891, 2007.
[39] C.Q. Zhang and S.C. Zhang, *Association Rule Mining: Models and Algorithms.* Springer-Verlag, 2002.
[40] M. Zuke, "On Finding all Suboptimal Foldings of an RNA Molecule," *Science,* vol. 244, no. 4900, pp. 48-52, 1989.
[41] M. Zuker and P. Stiegler, "Optimal Computer Folding of Large RNA Sequences Using Thermodynamics and Auxiliary Information," *Nucleic Acids Research,* vol. 9, no. 1, pp. 133-48, 1981.

Qingfeng Chen received the BSc and MSc degrees in mathematics from Guangxi Normal University, China, in 1995 and 1998, respectively, and the PhD degree in computer science from the University of Technology Sydney in September 2004. He is now a research fellow at the School of Information Technology, Deakin University, Australia. His research interests include bioinformatics, data mining, and artificial intelligence. He has published 30 refereed papers and one monograph by Springer, including international journals *Data Mining and Knowledge Discovery* and *BMC Bioinformatics*. One of his papers "Dealing with Inconsistent Secure Message" won the Best Paper Award at the Eighth Pacific Rim International Conference on Artificial Intelligence, New Zealand, 2004. He was invited to publish in the *Journal of Artificial Intelligence*. He has been serving as an associate editor for *Engineering Letters* and cochair for two international conferences.

Yi-Ping Phoebe Chen received the BInfTech degree (with first class honours) and the PhD degree in computer science from the University of Queensland. She is currently an associate professor (reader) at Deakin University, Melbourne, Australia. She is the director of the Bioinformatics Group, and the chief investigator of the ARC Centre in Bioinformatics and head of the Multimedia Stream. She is the steering committee chair of Asia Pacific Bioinformatics Conference (founder) and Multimedia Modelling. Her research interests include bioinformatics, multimedia databases and technology, visual query, Web information systems, machine learning, and data mining. She has been awarded 23 research grants (include 12 prestigious Australia Research Council (ARC) grants). She has published more than 130 refereed publications. These papers appeared in *Nucleic Acids Research*, *Current Drug Metabolism*, *BMC Genomics*, *BMC Bioinformatics*, *Data Mining and Knowledge Discovery*, *Information Systems*, *IEEE Multimedia*, the *ACM Transactions on Multimedia Computing*, *Communications and Applications*, etc. She is an associate editor for a number of journals. She is a senior member of the IEEE.