PseudoViewer: automatic visualization of RNA pseudoknots

Kyungsook Han[1,], Yujin Lee[1] and Wootaek Kim[2]*

[1]Department of Computer Science and Engineering, Inha University, Inchon 402-751, South Korea and [2]Department of Automation Engineering, Inha University, Inchon 402-751, South Korea

ABSTRACT

Motivation: Several algorithms have been developed for drawing RNA secondary structures, however none of these can be used to draw RNA pseudoknot structures. In the sense of graph theory, a drawing of RNA secondary structures is a tree, whereas a drawing of RNA pseudoknots is a graph with inner cycles within a pseudoknot as well as possible outer cycles formed between a pseudoknot and other structural elements. Thus, RNA pseudoknots are more difficult to visualize than RNA secondary structures. Since no automatic method for drawing RNA pseudoknots exists, visualizing RNA pseudoknots relies on significant amount of manual work and does not yield satisfactory results. The task of visualizing RNA pseudoknots by hand becomes more challenging as the size and complexity of the RNA pseudoknots increase.

Results: We have developed a new representation and an algorithm for drawing H-type pseudoknots with RNA secondary structures. Compared to existing representations of H-type pseudoknots, the new representation ensures uniform and clear drawings with no edge crossing for any H-type pseudoknots. To the best of our knowledge, this is the first algorithm for automatically drawing RNA pseudoknots with RNA secondary structures. The algorithm has been implemented in a Java program, which can be executed on any computing system. Experimental results demonstrate that the algorithm generates an aesthetically pleasing drawing of all H-type pseudoknots. The results have also shown that the drawing has high readability, enabling the user to quickly and easily recognize the whole RNA structure as well as the pseudoknots themselves.

Availability: All algorithms are available on request from the corresponding author.

Contact: khan@inha.ac.kr

**Keywords:** RNA structure; H-type pseudoknot; visualization, graph drawing.

*To whom correspondence should be addressed.

INTRODUCTION

An RNA pseudoknot is a tertiary structural element formed when bases of a single-stranded loop pair with complementary bases outside the loop. Pseudoknots are not only widely occurring structural motifs in all kinds of viral RNA molecules, but also responsible for several important functions of RNA. For example, pseudoknot structures present in coding regions can stimulate ribosomal frameshifting and translational read-through during elongation. Pseudoknots in noncoding regions can initiate translation either by being part of so-called internal ribosomal entry site (IRES) in the 5′ noncoding region or by forming translational enhancers in the 3′ noncoding region (Deiman and Pleij, 1997).

Currently, no automatic method for drawing RNA pseudoknot structures exists. Several computer programs are available for drawing RNA secondary structures (for example, De Rijk and De Wachter, 1997; Han *et al.*, 1999), but none of these can be used to draw RNA pseudoknots. In the sense of graph theory, a drawing of RNA secondary structures is a tree, whereas a drawing of RNA pseudoknots is a graph (and possibly a nonplanar graph) with inner cycles within a pseudoknot as well as possible outer cycles formed between a pseudoknot and other structural elements. Thus, drawing RNA pseudoknot structures is computationally more difficult than depicting RNA secondary structures.

RNA pseudoknots are often represented by adding line segments to the RNA secondary structure drawings to indicate base pairs formed by the RNA pseudoknots. Alternatively, RNA pseudoknots are drawn either manually by modifying RNA secondary structure drawings or from scratch. In either case, drawing RNA pseudoknots manually becomes more difficult and yields unsatisfactory results as the size and complexity of the drawings increase. One of the difficulties in drawing RNA structures is an overlapping of structural elements, which reduces the readability of the drawing. In most drawing programs of RNA secondary structures, the computational load is
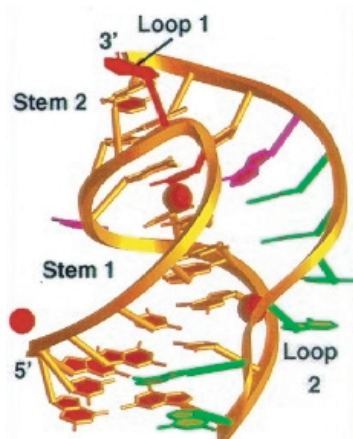
Fig. 1. Three-dimensional structure of a H-type pseudoknot, reprinted with Nature's copyright permission from (Su *et al.*, 1999). The pseudoknot is bent at the stem-stem junction and the two stems are coaxially stacked.

increased because of the work associated with removing the overlap of structural elements, performed either by an iterative process of the programs or with user intervention.

We have developed a new representation method and an algorithm for automatically drawing RNA secondary structures with H-type pseudoknots, and implemented the algorithm in a working program called PSEUDOVIEWER. To the best of our knowledge, this is the first algorithm for drawing RNA pseudoknot structures. We adopted two basic criteria when designing the algorithm: (1) overlapping of structural elements should be minimized to increase the readability of the drawing, and (2) not only the pseudoknots themselves but also the entire RNA structure containing them should be recognized quickly and easily. Experimental results demonstrate that the algorithm is capable of automatically producing a clear and aesthetically appealing drawing of RNA structures. The rest of this paper describes the algorithm and its experimental results.

REPRESENTATION OF H-TYPE PSEUDOKNOTS

In the classic type or H-type pseudoknots, bases in a hairpin loop pair with complementary bases outside the hairpin loop (see Figure 1 for a three-dimensional structure of a H-type pseudoknot). According to the broad definition of pseudoknots (Pleij, 1990), 14 types of topologically distinct pseudoknots are possible in principle. However, the most commonly occurring pseudoknots are of the H-type, where H stands for hairpin loop.

Figure 2 displays typical representations of H-type pseu-

doknots (Hilbers *et al.*, 1998). All H-type pseudoknots are drawn with edge crossings (drawings with edge crossings are said to be *nonplanar*). The edge crossings reduce the readability of the drawings and make it difficult to follow the RNA sequence from the 5′-end to 3′-end. The edge crossings, however, are inevitable in these drawings in order to stack the two stems coaxially. The coaxial stacking of the two stems has a biological meaning for the two stems of a pseudoknot to mimic a single stem, and has been confirmed by an NMR study (Hilbers *et al.*, 1998). However, the drawing of pseudoknots with RNA secondary structure describes a *topological* structure rather than a *geometric* structure. That is, the drawing of this type is intended to represent the connectivity relation between bases, so the drawing should focus on making the connectivity relation clear.

We propose a new method for representing all H-type pseudoknots uniformly and without edge-crossings. The drawings shown in Figure 3a–d represent exactly the same pseudoknots of Figures 2a–d. The drawing of Figure 3b is obtained by flipping stem 2 (enclosed in a box) with respect to a horizontal axis and by translating it horizontally by the stem width. The drawing of Figure 3c is obtained by flipping stem 2 with respect to a vertical axis. The drawing of Figure 3d is obtained by flipping stem 2 with respect to a horizontal axis and by translating it horizontally. The resulting drawings contain no edge-crossings (i.e. they are *planar*) and have similar shapes with exactly two inner cycles regardless of their types. Furthermore, it is much easier to follow the RNA sequence direction from the 5′-end to 3′-end. In the new representation, the two stems of a pseudoknot are not stacked coaxially, but are parallel and adjacent to each other.

ALGORITHM
Preliminaries

In RNA secondary structures, there are two types of structural elements:

- Stem (also called helix): double-stranded part, which is a contiguous region of base pairs.

- Regular loop: single-stranded part such as a hairpin loop, internal loop, bulge loop, multiple loop, or dangling end.

Since our algorithm draws pseudoknot structures as well as secondary structural elements, we considered additional structural elements as well:

- Pseudoknot: structural element formed by a pairing of bases in a regular loop with complementary bases outside the loop.
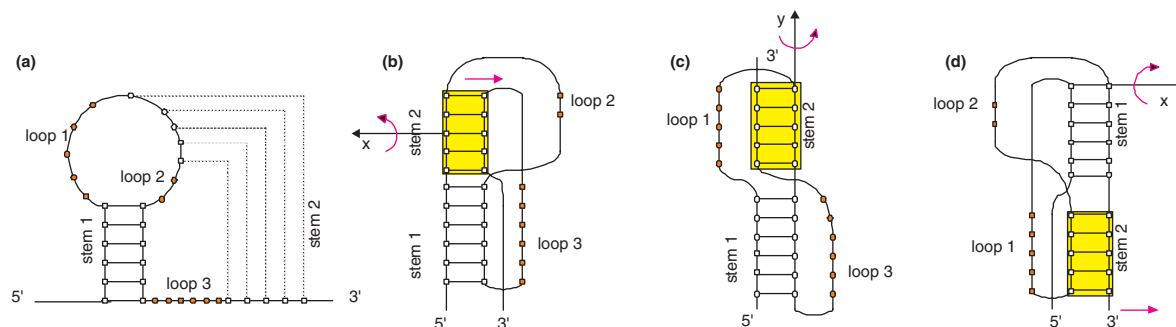
Fig. 2. Schematic representation of H-type pseudoknots, adapted from (Hilbers *et al.*, 1998). (a) General configuration. (b) Loop 1 is eliminated. (c) Loop 2 is eliminated. (d) Loop 3 is eliminated. The pseudoknot in (c) is the most abundant in natural RNA.
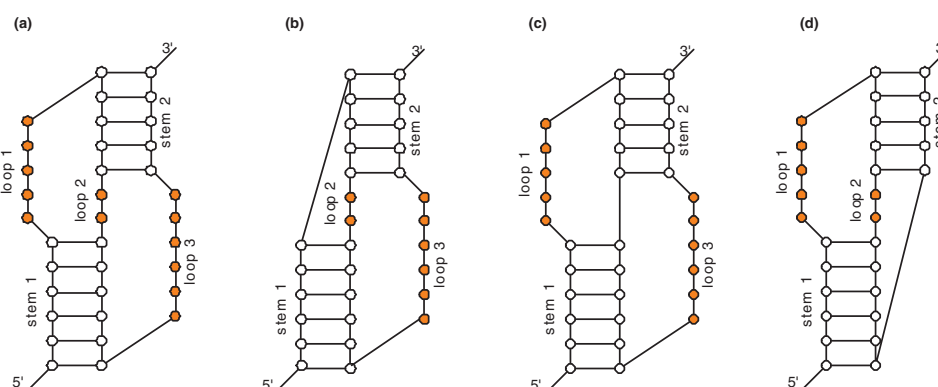
Fig. 3. New representation of H-type pseudoknots.

- Pseudoknot loop (PK loop): high-level loop that contains a pseudoknot as well as single-stranded part (see Figure 6 for an example).

As mentioned earlier, a drawing of RNA pseudoknots is a graph with *inner* cycles within a pseudoknot as well as possible *outer* cycles formed between a pseudoknot and other structural elements. What we call a "PK loop" represents the outer cycle. Given RNA pseudoknots and secondary structures, we represent the whole structure as a tree rather than as a graph (see Figure 4 for an example of the abstract tree). This is possible by representing both regular loops and PK loops as nodes of the tree. Edges of the tree represent stems of the secondary structure. The root node of the tree is the loop with the smallest starting base number. If there is a dangling end, we add artificial bases to pair the first and last bases. The artificial bases introduced in this step are not actually shown in the final drawing. A pseudoknot itself is not represented in the abstract tree; it is part of a node of the tree.

By hiding the inner cycles as well as the outer cycles in the nodes of the abstract tree, we can represent the whole, top-level RNA structure as a tree, making the drawing process simple. Loops of the tree are placed and drawn in increasing order of their depth values (the root node has the smallest depth value). The outline of both a regular loop and a PK loop is drawn in a circular shape.
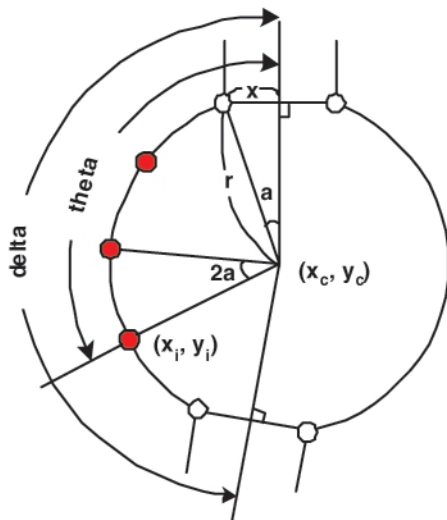
The algorithm of PSEUDOVIEWER is outlined as follows: (1) stems, regular loops, pseudoknots and PK loops are identified from the input structure data; (2) an abstract tree is constructed for representing the entire structure; (3) for each node and edge of the tree, its size and shape is determined; and (4) starting with the root node, each node and edge of the tree is positioned level by level by both translation and rotation. This section describes each step of the algorithm in detail.

The following notations are used in describing the algorithm.

- All angles are measured with respect to the positive y-axis.

- Units of angles are radians rather than degrees, unless specified otherwise.
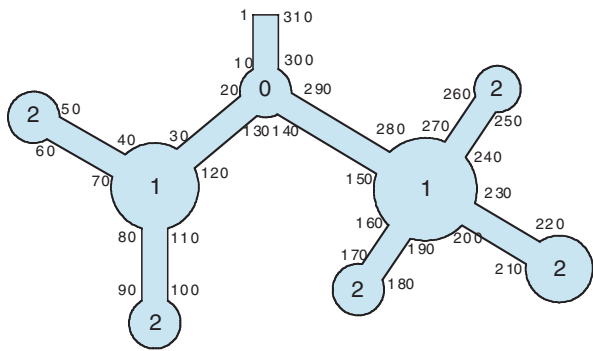
Fig. 4. An abstract tree for the whole structure, in which a node represents either a regular loop or a PK loop and an edge represents a stem. The node label represents the depth of the loop in the tree. The labels in each edge represent the starting and ending base numbers of the stem. For example, the edge label 1–10/300–310 indicates that bases 1–10 pair with bases 300–310 to form the stem.

Fig. 6. Components of a PK loop. Ws: width of a stem as well as the distance between adjacent bases, Lp: diagonal length of a pseudoknot, r: radius of a PK loop, α: angle between the diagonal direction and the stem direction of a pseudoknot

senting pseudoknots (van Batenburg *et al.*, 2001). The paring format describes pseudoknots as well as secondary structures in the following style. In this example, a pseudoknot is formed by base pairing $G_{56}CGGUU_{61}$ with $A_{74}GCCGC_{79}$.

```
          50            60            70
$    CGAGGGGCGGUUGGCCUCGUAAAAAGCCGC
%    ( ( ( ( ( ( : [ [ [ [ [ [ : : ) ) ) ) ) ) : : : : : ] ] ] ] ] ]
```

Given this input, PSEUDOVIEWER generates a drawing of the pseudoknot. The size of a pseudoknot is the diagonal length *PKLength* of a bounding box of the pseudoknot and is computed using equation 1:

$$PKLength = \sqrt{(x_{\max} - x_{\min})^2 + (y_{\max} - y_{\min})^2} \quad (1)$$

$x_{\min}, y_{\min}$: minimum x and y coordinates of the pseudoknot's bounding box

$x_{\max}, y_{\max}$: maximum x and y coordinates of the pseudoknot's bounding box

Regular loops

The outline of a regular loop is drawn in a circle shape. In Figure 5, the distance between the centers of adjacent bases of a regular loop is $2x$. Since a regular loop is considered to consist of several isosceles triangles, the vertical angle a of the isosceles triangle and the radius r of the regular loop can be computed using equations 2 and
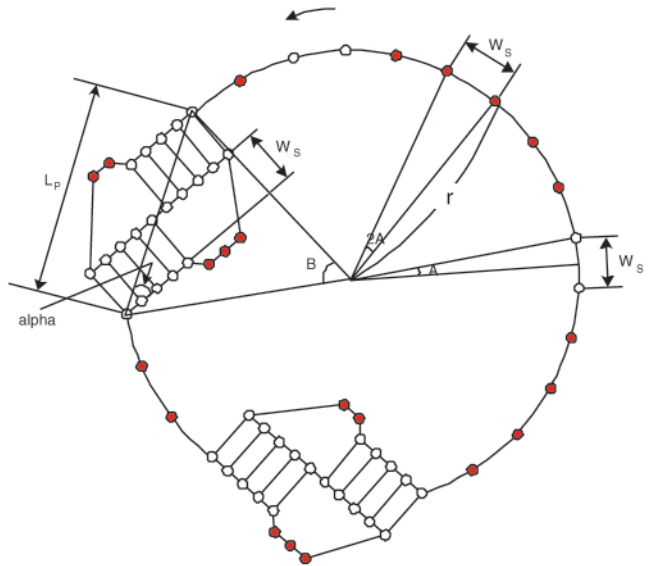

Fig. 5. The radius of a regular loop and the angle between adjacent stems and the angle between every base of the loop and the positive *y*-axis.

- Modulus operator (%) is used for computing angle values since angle values should be in the range $[0, 2\pi]$.

- In Figures 2, 3, 5, and 6, unfilled circles represent the bases of a stem while filled circles represent the bases in the single-stranded part.

Pseudoknots

The program PSEUDOVIEWER takes as input an ASCII file in pairing format, which is widely used for repre-

3, respectively:

$$a = \frac{1}{2} \cdot \frac{2\pi}{n_{lb}} = \frac{\pi}{n_{lb}} \qquad (2)$$

where n_{lb} is the number of bases on the regular loop and

$$r = \frac{x}{\sin a} \qquad (3)$$

Once we determine the vertical angle a of the isosceles triangle, we can calculate the angle δ between adjacent stems and the angle θ_i between the ith base of a regular loop and the positive y-axis. Let n_s be the number of stems connected to a regular loop, and n_b be the number of intervening bases between adjacent stems. Then

$$\delta = 2a(n_b + 2) \qquad (4)$$
$$\theta_i = (2i + 1)a, \quad i = 0, 1, 2, \ldots, n_{lb} - 1 \qquad (5)$$

Using the angles computed by equations 2–5, the bases of a regular loop can be positioned as follows:

$$x_i = -r \sin \theta_i + x_c \qquad (6)$$
$$y_i = r \cos \theta_i + y_c \qquad (7)$$

x_i, y_i: x and y coordinates of the ith base of a regular loop

x_c, y_c: x and y coordinates of the center of a regular loop

θ_i: angle between the ith base of a regular loop and the positive y-axis

r: radius of a regular loop

In the root node of the abstract tree, both the stem angle ϕ of the first stem and the startAngle of the loop are zero. The stem angles ϕ of other stems are measured with respect to the positive y-axis, and are calculated using equation 8:

$$\phi = D_{loop} + \pi + \delta \qquad (8)$$

where D_{loop} is the startAngle of the upper regular loop. Notice that in equation 8, π is added to make the stem point the upper regular loop. The startAngle of the lower regular loop is the same as the current stem angle. The distance D_{rl} between the centers of regular loops is computed using equation 9:

$$D_{rl} = r_c + (n_{ub} - 1) \times h_s + r_u \qquad (9)$$

r_c: radius of the current regular loop

r_u: radius of the upper regular loop

n_{ub}: number of bases of the upper stem

h_s: distance between adjacent base pairs of a stem

Pseudoknot loops

In order to handle a pseudoknot loop (PK loop) in a similar way to a regular loop, we should first determine the size of the PK loop. Since the outline of a PK loop is drawn in a circle shape, its radius can be computed from pseudoknots and bases contained in it. When a PK loop contains p pseudoknots and b bases, an inscribed polygon of the PK loop can be considered to consist of n ($= p + b$) inscribed isosceles triangles (see Figure 6). Therefore, equation 10 holds:

$$\sin a_i = \frac{x_i}{r} \quad \Rightarrow \quad a_i = \arcsin\left(\frac{x_i}{r}\right) \qquad (10)$$

Since there are total n isosceles triangles inside a PK loop, the following holds:

$$f = \sum_{i=1}^{n} \arcsin\left(\frac{x_i}{r}\right) - \pi = 0 \qquad (11)$$

Equation 11 is a monotonically decreasing function after a certain point, and can be solved by Newton's method to determine r. However Newton's method requires differentiation of equation 11, which requires a significant amount of computation as shown in equation 12.

$$\frac{df}{dr} = \sum_{i=1}^{n} \frac{-x_i}{r\sqrt{r^2 - x_i^2}} \qquad (12)$$

Instead of using Newton's method, we determine the radius r of a PK loop by incrementing r by a small step. Once we determine r, several angles associated with a PK loop can be computed. Let W_s be the width of a stem (which is also the distance between adjacent bases on a PK loop), L_p be the diagonal length of the pseudoknot region, x_e and y_e be the coordinates of the ending base of a pseudoknot, and x_s and y_s be the coordinates of the starting base of a pseudoknot (see Figure 6). Then:
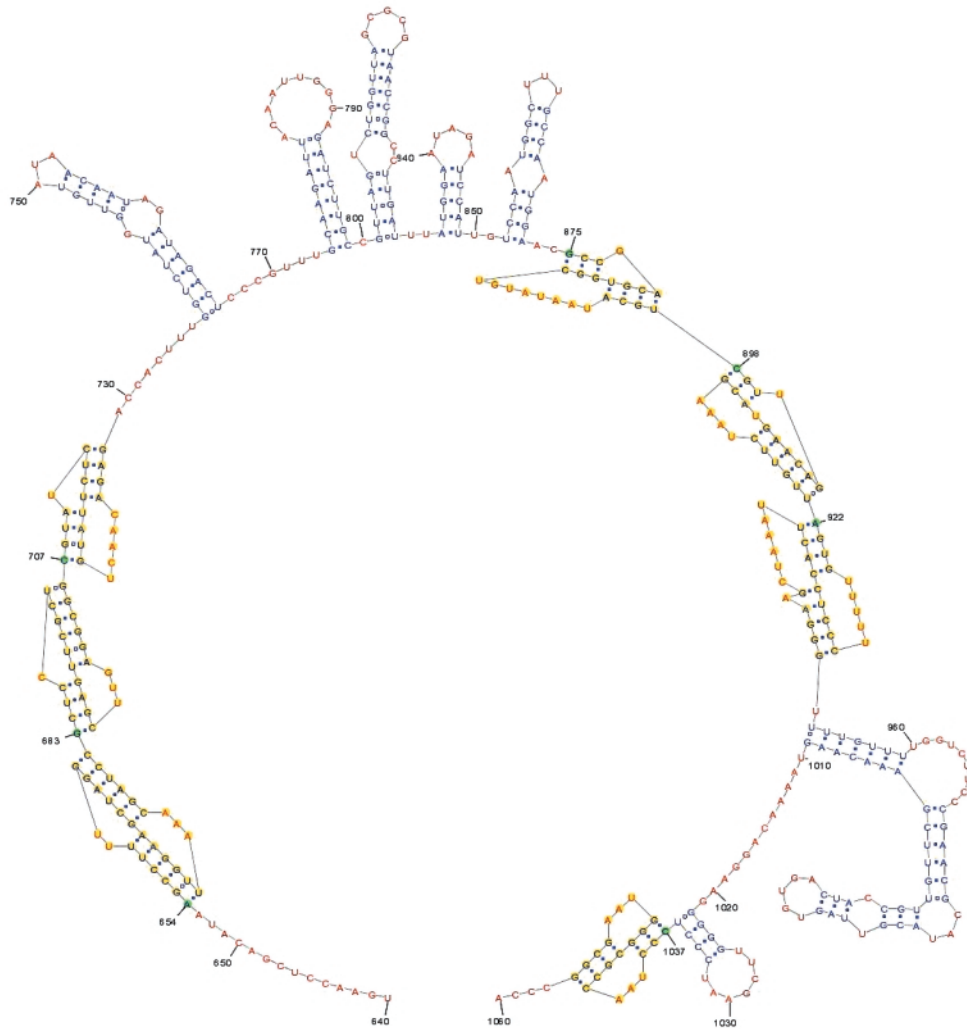
$$L_p = \sqrt{(x_e - x_s)^2 + (y_e - y_s)^2} \qquad (13)$$
$$\sin A = \frac{W_s/2}{r} \quad \Rightarrow \quad A = \arcsin\left(\frac{W_s}{2r}\right) \qquad (14)$$
$$\sin(B/2) = \frac{L_p/2}{r} \quad \Rightarrow \quad B = 2 \cdot \arcsin\left(\frac{L_p}{2r}\right) \qquad (15)$$

A pseudoknot in a PK loop can be positioned in a similar way to a stem. To do this, we first orient a pseudoknot in the positive y direction by rotating it through angle α counterclockwise about its corner. The angle of the pseudoknot is the angle of the first base angle of the pseudoknot added to $3\pi/2$:

$$\sin \alpha = \frac{2W_s}{L_p} \quad \Rightarrow \quad \alpha = \arcsin\left(\frac{2W_s}{L_p}\right). \qquad (16)$$

Fig. 7. Structure of STMV RNA 3′-UTR with 7 pseudoknots

IMPLEMENTATION

PSEUDOVIEWER is written in pure Java, so it is executable on any platform that supports Java. Figures 7 and 8 show the structures of satellite tobacco mosaic virus (STMV) RNA 3′-UTR (Gultyaev *et al.*, 1994) and *E. coli* tmRNA (Williams, 2000), respectively. tmRNA is a molecule that combines functions of tRNA and mRNA in trans-translation (Knudsen *et al.*, 2001). Bases are numbered in increments of 10. If a base number falls on a loop, it is shown in the drawing; otherwise, it is not shown to avoid overlaps. In addition to this, the starting base of each pseudoknot is shown in green background color with its base number. Pseudoknots are shown in yellow background color, and thus are easily distinguished from other structural elements. Filled circles represent the canonical base pairs (A-U or G-C), and

open circles represent the G-U wobble pairs. There is no overlapping of structural elements, and the PK loop with several pseudoknots takes on a circle shape similar to a regular loop. Unlike manual drawings of pseudoknots, the drawings generated by PSEUDOVIEWER are aesthetically pleasing and clear.

We have tested PSEUDOVIEWER on many RNA structures with H-type pseudoknots, and have not yet found the size limitation of the RNA imposed by the PSEUDOVIEWER algorithm. We suppose that the maximum size that can be visualized by PSEUDOVIEWER is limited by the computing power of the computer system and the bitmap size supported by the graphic device driver. The most time-consuming part of PSEUDOVIEWER is the computation of the radius of a PK loop since it involves adding the arcsin values of n inscribed triangles of a PK
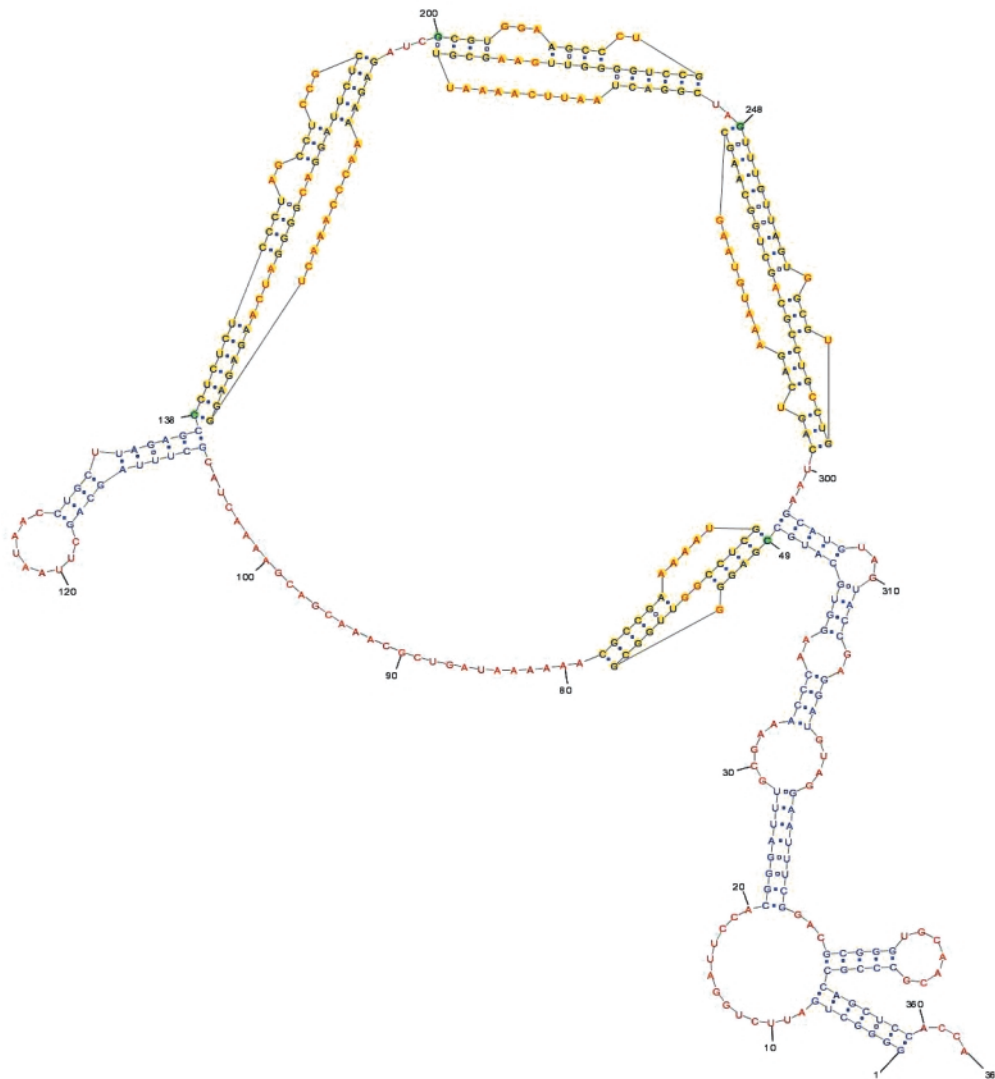
Fig. 8. Structure of *E. coli* tmRNA with 4 pseudoknots. 3 pseudoknots starting at 138, 200, and 248 contain more than 2 stems in each of them.

Table 1. Execution time of PSEUDOVIEWER

RNA	# bases	# pseudoknots	# inscribed triangles in PK loops	time (ms)
satellite tobacco necrosis virus 1	252	4	45	407
tobacco mosaic virus	214	4	45	422
E.coli tmRNA	363	4	49	516
satellite tobacco mosaic virus	421	7	87	1516
odontoglossum ringspot virus	419	8	86	1532
cyanophora paradoxa cyanelle tmRNA	291	1	153	2360

loop in each step of the iteration. When computing the radius of a PK loop, its inscribed triangle corresponds to either a pseudoknot or a pair of adjacent bases of the PK loop. This is corroborated by the execution time of PSEUDOVIEWER for 6 test cases (Table 1). The execution time depends on the number of pseudoknots and on the

number of inscribed triangles of PK loops rather than on the total number of bases.

DISCUSSION AND CONCLUSION

A drawing of RNA pseudoknots is a graph (possibly nonplanar) with inner cycles within a pseudoknot as well as possible outer cycles formed between a pseudoknot and other structural elements. Thus, generating a clear representation of RNA pseudoknot structures is computationally harder than RNA secondary structures. We have developed a new representation method and an algorithm for visualizing RNA pseudoknots as a two-dimensional drawing and have implemented the algorithm in a working program called PSEUDOVIEWER. The new representation produces uniform, clear drawings with no edge crossing for all H-type pseudoknots.

Given RNA pseudoknots and secondary structures, we represent the whole structure as a tree rather than as a graph by hiding the inner cycles as well as the outer cycles in the nodes of the abstract tree. Once the entire, top-level RNA structure is represented as a tree, loops are placed and drawn in increasing order of their depth values in the tree. Experimental results have shown that PSEUDOVIEWER is capable of generating an aesthetically pleasing planar drawing of H-type pseudoknots. The algorithm of PSEUDOVIEWER is the first one for automatically drawing RNA structures containing pseudoknots.

The development of PSEUDOVIEWER is not complete, and is being extended in several directions. First, it can currently visualize only H-type pseudoknots. It is being extended to handle other types of pseudoknots. Second, PSEUDOVIEWER will be made available as a web-based application program so that it can be executed anywhere using a web browser supporting Java. This is the reason that we have chosen Java as the implementation language. Finally, PSEUDOVIEWER currently represents pseudoknots as two-dimensional drawings. We would like to display pseudoknots as a three-dimensional drawing. Perhaps this will be the ultimate goal of our visualization work on RNA pseudoknots.

REFERENCES

Deiman,B.A.L.M. and Pleij,C.W.A. (1997) Pseudoknots: A vital feature in viral RNA. *Seminars Virol.*, **8**, 166–175.

De Rijk,P. and De Wachter,R. (1997) RnaVIz, a program for the visualization of RNA secondary structure. *Nucleic Acids Res.*, **25**, 4679–4684.

Gultyaev,A.P., van Batenburg,E. and Pleij,C.W.A. (1994) Similarities between the secondary structure of satellite tobacco mosaic virus and tobamovirus RNAs. *J. Gen. Virol.*, **75**, 2851–2856.

Han,K., Kim,D. and Kim,H.-J. (1999) A vector-based method for drawing RNA secondary structure. *Bioinformatics*, **15**, 286–297.

Hilbers,C.W., Michiels,P.J.A. and Heus,H.A. (1998) New developments in structure determination of pseudoknots. *Biopolymers*, **48**, 137–153.

Knudsen,B., Wower,J., Zwieb,C. and Gorodkin,J. (2001) tmRDB (tmRNA database). *Nucleic Acids Res.*, **29**, 171–172.

Pleij,C.W.A. (1990) Pseudoknots: a new motif in the RNA game. *Trends Biochem. Sci.*, **15**, 143–147.

Su,L., Chen,L., Egli,M., Berger,J.M. and Rich,A. (1999) Minor groove RNA triplex in the crystal structure of a ribosomal frameshifting viral pseudoknot. *Nat. Struct. Biol.*, **6**, 285–292.

van Batenburg,F.H.D., Gultyaev,A.P. and Pleij,C.W.A. (2001) PseudoBase: structural information on RNA pseudoknots. *Nucleic Acids Res.*, **29**, 194–195.

Williams,K.P. (2000) The tmRNA website. *Nucleic Acids Res.*, **28**, 168.