Predicting structures and stabilities for H-type pseudoknots with interhelix loops

Song Cao and Shi-Jie Chen

*RNA* 2009 15: 696-706 originally published online February 23, 2009
Access the most recent version at doi:10.1261/rna.1429009

Supplemental Material
http://rnajournal.cshlp.org/content/suppl/2009/02/23/rna.1429009.DC1.html

References
This article cites 81 articles, 31 of which can be accessed free at:
http://rnajournal.cshlp.org/content/15/4/696.full.html#ref-list-1

Email alerting service
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here

To subscribe to *RNA* go to:
http://rnajournal.cshlp.org/subscriptions

Copyright © 2009 RNA Society
Predicting structures and stabilities for H-type pseudoknots with interhelix loops

SONG CAO and SHI-JIE CHEN
Department of Physics and Department of Biochemistry, University of Missouri, Columbia, Missouri 65211, USA

ABSTRACT
RNA pseudoknots play a critical role in RNA-related biology from the assembly of ribosome to the regulation of viral gene expression. A predictive model for pseudoknot structure and stability is essential for understanding and designing RNA structure and function. A previous statistical mechanical theory allows us to treat canonical H-type RNA pseudoknots that contain no intervening loop between the helices (see S. Cao and S.J. Chen [2006] in Nucleic Acids Research, Vol. 34; pp. 2634–2652). Biologically significant RNA pseudoknots often contain interhelix loops. Predicting the structure and stability for such more-general pseudoknots remains an unsolved problem. In the present study, we develop a predictive model for pseudoknots with interhelix loops. The model gives conformational entropy, stability, and the free-energy landscape from RNA sequences. The main features of this new model are the computation of the conformational entropy and folding free-energy base on the complete conformational ensemble and rigorous treatment for the excluded volume effects. Extensive tests for the structural predictions show overall good accuracy with average sensitivity and specificity equal to 0.91 and 0.91, respectively. The theory developed here may be a solid starting point for first-principles modeling of more complex, larger RNAs.

INTRODUCTION
An RNA pseudoknot is formed when nucleotides in a loop base-pair with complementary nucleotides outside the loop. An H-type pseudoknot is formed by base-pairing between a hairpin loop and the single-stranded region of the hairpin. The structure consists of two helix stems (Fig. 1A, S₁, S₂) and two loops (Fig. 1A, L₁, L₂) as well as a possible third loop/junction (Fig. 1A, L₃) that connects the two helix stems. In most naturally occurring RNA pseudoknots, interhelix loop L₃ contains no more than 1 nucleotide (nt). For these canonical pseudoknot structures, helix stems S₁ and S₂ tend to stack coaxially (or partially coaxially) to form a quasicontinuous RNA helix in the three-dimensional space (3D) (Walter and Turner 1994; Chen et al. 1996; Cornish et al. 2005; Theimer et al. 2005). The coaxial stacking interaction can provide an essential stabilizing force for the structure.

The pseudoknot is a widespread motif in RNA structures (van Belkum et al. 1985; Perrotta and Been 1991; Tanner et al. 1994; Deiman et al. 1997; Ferré-D’Amaré et al. 1998; Su et al. 1999; Schultes and Bartel 2000) and plays a variety of structural and functional roles in RNAs. For instance, pseudoknots form the core structural motif in the central catalytic domain of human telomerase RNA (Chen et al. 2000; Comolli et al. 2002; Theimer et al. 2005). As another example, for many viruses, pseudoknots play indispensable roles in promoting ribosomal frameshifting, a mechanism used by a retrovirus to regulate retroviral genome expression (Brierley et al. 1989, 2007; Somogyi et al. 1993; Giedroc et al. 2000; Plant et al. 2003; Staple and Butcher 2005; Namy et al. 2006; Hansen et al. 2007; Cao and Chen 2008; Pennell et al. 2008). Mutations that strengthen or weaken pseudoknot (thermal or mechanical) stability can cause changes in ribosomal frameshifting efficiency (Cornish et al. 2005; Theimer et al. 2005). For these, and a vast number of other RNA-related problems, quantitative prediction of pseudoknot structure and its stability is essential in order to unveil the mechanisms of RNA functions and in order to design therapeutic strategies for the diseases. In the present study, we develop a rigorous statistical mechanical model to predict the structure and folding stability for general RNA pseudoknots.

There are two main approaches used to predict RNA structures: free-energy minimization and comparative sequence
virtual bond model for RNA nucleotides involves two bonds, P–C₄–P. To describe the base
accuracies are limited by the availability of reliable param-
et al. 2007; Shapiro et al. 2007; Bon et al. 2008), their
remarkable in their computational efficiency to treat large
Current pseudoknot structural prediction algorithms often
dynamic algorithms, they cannot guarantee finding the
distributions (McCaskill 1990; Chen and Dill 2000; Hofacker 2003).
For RNA pseudoknots, several lines of computational
have also been developed (Gultyaev et al. 1995; Rivas and Eddy 1999; Dirks and Pierce 2003; Reeder and Giegerich 2004; Ruan et al. 2004; Ren et al. 2005; Huang and Ali 2007; Chen et al. 2008; Metzler and Nebel 2008; Sperschneider and Datta 2008). Heuristic approaches (Ren et al. 2005) are computationally efficient, but unlike dynamic algorithms, they cannot guarantee finding the global free-energy minimum. Critical to an accurate structure prediction are the energy and entropy parameters. Current pseudoknot structural prediction algorithms often ignore the contribution of loop entropies (Ren et al. 2005) or use simplified (nonphysical) approximations (Dirks and Pierce 2003) for the loops. Although these models are remarkable in their computational efficiency to treat large RNA pseudoknots with hundreds of nucleotides (Mathews and Turner 2006; Reeder et al. 2006; Schuster 2006; Jossinet et al. 2007; Shapiro et al. 2007; Bon et al. 2008), their accuracies are limited by the availability of reliable param-
eters for the entropies of pseudoknotted loops. A set of rigorous entropy param-
eters, such as the one derived in the present study, would be highly desirable for reliable structure prediction (Cao and Chen 2006b). Indeed, a current emphasis for RNA pseudoknot prediction is how to include the thermody-
namic parameters, especially the loop entropy, in the dynamic algorithms (Zhang and Chen 2001; Ding 2006; Kopeikin and Chen 2006; Chen 2008; Chu and Herschlag 2008; Jabbari et al. 2008; Li et al. 2008; Zhang et al. 2008).
Using a virtual-bond-based RNA con-
formational model (termed the “Vfold” model) (Cao and Chen 2005, 2006a), we recently developed a physics-based theory to calculate the loop entropy and free energy for simple canonical H-type pseudoknots (Cao and Chen 2006b), namely, pseudoknots with no interhelix loop (Fig. 1A, L₃). For such canonical H-type pseudoknots, the two helix stems often form a quasicontinuous coaxially stacked helix. Central to the loop entropy calculation is the influence of the excluded volume between loop and helix. The effect of volume exclusion is sensitive to the stem and loop lengths. Here, we develop a new Vfold model to treat more complex pseudoknots that contain an interhelix loop (Fig. 1A, L₃). The development of such a more-general model is significant for two reasons. First, the general pseudoknots studied here form the structural basis for large RNA folds, which involve multiple loops between helices. Second, the interhelix loops considered here are biologi-
cally important. For example, it has been suggested that a large class of anti-HIV RNA aptamers form pseudoknots with interhelix loops (Burke et al. 1996) so that the aptamers can be flexible and prevent the rigid coaxial stacking between the helices.
This paper is organized as follows. We first present a new three-vector virtual-bond-based RNA conformational model. The development of the new virtual bond model is motivated by the need to explicitly include the base orientations in addition to the backbone configuration considered in the original Vfold model (Cao and Chen 2005). We then use the new conformational model to compute the loop entropies in different pseudoknot con-
texts. A key issue in the calculation is how to account for the excluded volume effects. The entropy parameters will then allow us to predict the lowest free-energy structure as well as the folding thermodynamics from the RNA sequence. Comparisons with other models for structural prediction show improved results from our new model. As an application of the theory, we also investigate the equilibrium unfolding pathway for an anti-HIV RNA pseudoknot aptamer (Burke et al. 1996), the Visna-Maedi

**FIGURE 1.** (A) An RNA pseudoknot with an interhelix loop L₃. (B) Traditional two-vector virtual bond model for RNA nucleotides involves two bonds, P–C₄–P. To describe the base orientation, we introduce a third virtual bond model, C₄–N₁ (pyrimidine) or C₄–N₉ (purine). (C) A virtual bond representation for a pseudoknot motif with S₁ = 8 bp, S₂ = 6 bp, L₁ = 4 nt, L₂ = 4 nt, and L₃ = 2 nt.
virus (VMV) pseudoknot (Pennell et al. 2008), and the 5′ coding region of the R2 retrotransposon (Hart et al. 2008). The anti-HIV and VMV pseudoknots contain 3-nt and 6-nt interhelix loop L₃, respectively.

**STRUCTURAL MODEL**

**A three-vector virtual bond model**

Because the torsional angles of the C–O bonds (Fig. 1B, ε, β) in the nucleotide backbone tend to adopt the trans isomeric state, Olson (Olson and Flory 1972; Olson 1980) proposed to use a two-vector virtual bond to represent nucleic structures (see Fig. 1A). We recently developed a virtual-bond-based RNA folding model (the Vfold model) for H-type RNA pseudoknots (Cao and Chen 2006b). In Vfold, we model the helix as an A-form RNA helix using the experimentally measured atomic coordinates. For loops, which can be flexible, we use the usual gauche⁺ (g⁺), trans (t), and gauche⁻ (g⁻) rotational isomeric states for polymers (Flory 1969) to sample backbone conformations. The fact that the three isomeric states can be exactly configured in a diamond lattice (Cao and Chen 2005, 2006b) suggests that we can effectively configure the loop conformations as random walks of the virtual bonds on a diamond lattice. We note that the rotameric nature of RNA backbone conformations also has been observed for the known RNA structures (Duarte and Pyle 1998; Murthy et al. 1999; Murray et al. 2003; Wadley et al. 1998; Pyle 1998; Murthy et al. 1999; Murray et al. 2003; Wadley 1998; Olson and Flory 1972; Olson 1980) for RNA molecules also suggested a rigid base orientation (see Fig. 1A). Specifically, we add the N₁ (for pyrimidine) or N₉ (for purine) atom to the original P–C₄ and C₄–P virtual bonds (Fig. 1B). From the PDB database (Michiels et al. 2001; Theimer et al. 2005) for RNA pseudoknots, we find that the distance (D₂CN) between N₁ (N₉) and C₄ atoms is close to 3.9 Å. In addition, we find that the torsion angle (χ) between plane Pᵣ–C₄–Pᵣ₊₁ and plane Pᵣ–C₄–N₁ (N₉) is close to the g⁻¹ isomeric state; see the distributions for D₂CN and the torsion angle in Figure 2. The localized distributions for the virtual bond C₄–N₁ (N₉) in Figure 2 suggest that C₄–N₁ (N₉) is quite rigid and can be configured in a diamond lattice. A previous study on RNA molecules also suggested a rigid base orientation ( Olson and Flory 1972).

Figure 1A shows a pseudoknot with an interhelix loop. We use the atomic coordinates of the A-form RNA helix to configure the helices (Arnott and Hukins 1972). The (r, θ, z) coordinates (in a cylindrical coordinate system) for the P, C₄, and N₁ (or N₉) atoms are (8.71 Å, 70.5 + 32.7i, –3.75 + 2.81i), (9.68 Å, 46.9 + 32.7i, –3.10 + 2.81i), and (7.12 Å, 37.2 + 32.7i, –1.39 + 2.81i) (i = 0, 1, 2, …)

**CONFORMATIONAL ENTROPY FOR PSEUDOKNOT WITH AN INTERHELIX LOOP**

For a given pseudoknot defined by the stem lengths (S₁, S₂) and the loop lengths (L₁, L₂, L₃), we enumerate all the possible (virtual bond) conformations in the 3D space. From the total number of the viable conformations Ω, we calculate the conformational entropy of the given pseudoknot as ΔS(S₁, S₂, L₁, L₂, L₃) = kᵦ ln Ω, where kᵦ = 1.99 cal/K is the Boltzmann constant. We choose different (S₁, S₂, L₁, L₂, L₃) values (i.e., different pseudoknots), compute the conformational entropy for each pseudoknot, and compile the results as a large table for pseudoknot conformational entropy parameters.

Compared to simple canonical H-pseudoknots with no interhelix junction (junction-free pseudoknots), the pseudoknots here are much more complicated because the interhelix loop (Fig. 1, L₃) between the two stems may be
flexible, causing variable relative orientations between the helices. The previous model for the junction-free pseudoknots is a special case for the model developed here. To compute the total conformational count $\Omega$ for a pseudoknot with an interhelix loop, we enumerate the possible orientations between the two helices and then enumerate the loop conformations for each helix orientation:

$$\Omega = \sum_{\text{helix orientation}} \Omega_{PK},$$

where $\Omega_{PK}$ is the number of conformations for a given helix orientation.

### Enumeration of helix orientations

The orientations of helices $S_1$ and $S_2$ are determined by the coordinates of the terminal nucleotides (Fig. 1A, $c_1$, $c_2$) of the loop $L_3$. To enumerate the relative orientations of the helices, we fix the $(P_k, C_4, N_1)$ or $(N_9, P_{j+1})$ coordinates for $c_j$, then enumerate the viable $(P_k, C_4, N_1)$ or $(N_9, P_{j+1})$ coordinates for $c_j$. Specifically, we enumerate the loop $L_3$ conformations as self-avoiding random walks of the virtual bonds in a diamond lattice. The number of possible coordinates of the terminal nucleotide $c_2$ (specifically, the coordinates of $P_k$, $C_4$, $N_1$, or $N_9$, and $P_{j+1}$ atoms of nucleotide $c_j$) is much smaller than the number of loop $L_3$ conformations (see Fig. 4A, below). Therefore, the number of helix orientations, as determined by the $c_j$ and $c_2$ positions/coordinates, increases with $L_3$ much more slowly than the number of loop $(L_3)$ conformations. For instance, the number of helix orientations grows as $73 \to 390 \to 1358 \to 3208 \to 6096 \to 10,272 \to 15,984$ for an increasing interhelix loop length $1 \to 2 \to 3 \to 4 \to 5 \to 6 \to 7$ nt. The slow growth of the number of helix orientations makes the exact enumeration of all the possible helix orientations computationally viable.

### Enumeration of loop conformations for a given helix orientation

For each relative orientation of the helices $S_1$ and $S_2$, we compute $\Omega_{PK}$ in Equation 1 by enumerating the conformations for loops $L_1$ and $L_2$ and loop $L_3$. The key is how to treat the excluded volume effect, i.e., the effect that different atoms cannot bump into each other.

We have two choices to treat the excluded volume effect here. We may fit the off-lattice helix onto the diamond lattice, then both helix and loop are configured in the same lattice, thus the volume exclusion effect can be conveniently treated in the lattice framework. Such an approach is computationally time-consuming because it requires off-lattice → on-lattice fitting for all the helices for each and every helix orientation. Given the large number of helix orientations that we enumerate (Equation 1), the excluded volume treatment based on the above lattice-fitting is highly inefficient.

Alternatively, we can take a different approach that avoids the off-lattice → on-lattice fitting procedure. The strategy of the alternative approach is to keep the off-lattice helix structure. To treat the mixed system with the on-lattice loop conformations and off-lattice helix structure, we introduce a cutoff distance $D_c$ such that atoms separated by a distance below the cutoff are considered to bump into each other, and the corresponding conformation is eliminated. Such a cutoff would allow us to treat the excluded volume effect in a unified framework, irrespective of the on-lattice or off-lattice representation of the conformations. We determine the value of the cutoff distance $D_c$ from the criterion that it gives the same entropy as the one calculated from the off-lattice → on-lattice fitting. We found that the optimal $D_c$ value is $2.8$ Å (see Fig. 3C,D). Therefore, in our entropy computation, we use $D_c = 2.8$ Å as the criterion for volume exclusion.

For a fixed helix–helix orientation, we enumerate the loop conformations through self-avoiding random walks in the diamond lattice. The excluded volume between helix and helix, helix and loop, and loop and loop are explicitly considered. The treatment here for the excluded volume effect is more rigorous than previous Gaussian chain-based models (Gultyaev et al. 1999; Isambert and Siggia 2000; Bon et al. 2008), which ignore the excluded volume effect.

Using the three-vector virtual bond conformational model developed here, we can test the strengths of the different excluded volume effects (helix–helix, helix–loop, and loop–loop) (see Fig. 4B). We find that the loop–helix excluded volume interaction is strong. In contrast, the loop–loop excluded volume effect is rather weak (Fig. 4B).
where $l$ is the loop length and $a$ and $b$ are fitted parameters; see Supplemental Tables S2 and S3 for the $a$ and $b$ parameters for different loop sizes.

**PARTITION FUNCTION**

In this section, we show how to use the recursive algorithm (Cao and Chen 2005, 2006b) to compute the partition function, from which all the thermodynamic properties of the system can be determined. Our partition function calculation is a sum over all the possible secondary and pseudoknotted structures, with and without an interhelix loop. A typical “prototype” structure contains internal/bulge loops in the helix stems (see Fig. 5). Our conformational ensemble also includes other structures that stem from the prototype structure. For example, for the pseudoknot in Figure 5, we enumerate different $L_1$ and $L_2$ loop structures by allowing the formation of possible secondary structures in loops $L_1$ and $L_2$. We also allow multidomain structures, where each domain is an independently folded pseudoknotted or secondary structure.

We denote the partition function for the ensemble of all the possible pseudoknotted structures by $C_3(a, b)$, where $a$ and $b$ denote the 5' and 3' terminal nucleotides, respectively, and the subscript 3 denotes a pseudoknot (tertiary) structure. A general pseudoknotted structure shown in Figure 5 is described by 11 parameters: the lengths of loops $(L_1, L_2, L_3)$, the lengths of upper and lower segments of the helices $(n_{11}, n_{12}, n_{21}, n_{22})$, and the lengths of the internal loops $(l_{11}, l_{12}, l_{21}, l_{22})$. Exhaustive enumeration of all the possibilities for these 11 parameters for an $n$-nt chain requires a computational time scale on the order of $n^{11}$, which would be computationally unfeasible for long pseudoknot sequences.

If the internal/bulge loops in the helix stems $S_1$ and $S_2$ are flexible, $S_1$ and $S_2$ would be floppy, causing complications in the calculations of the loop $(L_1$ and $L_2$) entropies. In order to make use of our entropy parameter table, which assumes rigid helix stems (Fig. 1A), we employ an approximation for the stems, as described below. The (internal/bulge) loop tends to leave or enter a helix in a right-handed helical way, resulting in rigid conformations for a short internal/bulge loop. Therefore, for the purpose of the loop entropy calculation, we treat an internal/bulge loop as an effective helix of length $S^\text{eff}$ ($S_1^\text{eff}$) as determined by the following equations (see Fig. 5):

$$S_1^\text{eff} = n_{11} + L_{12} + n_{12}, S_2^\text{eff} = n_{21} + L_{21} + n_{22}. \quad (2)$$

In this way, we can read the entropy directly from the entropy table $\Delta S(S_1, S_2, L_1, L_2, L_3)$ with $S_1$ and $S_2$ substituted by the effective helix lengths $S_1^\text{eff}$ and $S_2^\text{eff}$ for stems $S_1$ and $S_2$, respectively. For pseudoknots without internal/bulge loops in the stems, $S_1^\text{eff}$ and $S_2^\text{eff}$ are equal to
the lengths of the original helices. An internal/bulge loop often causes bending of the stem (S₁ or S₂). Through the above approximation, we replace a bent stem with a continuous helix for the purpose of loop entropy calculation. Our control tests show that the approximation causes minor errors: ≤5% and 15% in the entropy results with helix stems containing bulge loops of length ≤2 nt and 3 nt, respectively (see details in the Supplemental Material).

For a loop (L₁ or L₂) with nested helices (see Fig. 8A, below), we neglect the excluded volume effect from the nested helices and calculate the effective loop length as the number of helices plus the unpaired nucleotides outside the helices.

We separate out pseudoknot-containing structures from pseudoknot-free structures (secondary structures) in the partition function calculation. We compute the partition function \(C_s(a, b)\) for pseudoknot-containing structures from nucleotide \(a\) at the 5’ end to nucleotide \(b\) at the 3’-end by enumerating all the possible values of helix stem lengths \(S₁\) and \(S₂\) and loop lengths \(L₁, L₂, L₃\):

\[
C_s(a, b) = \sum_{S₁} \sum_{S₂} \sum_{L₁} \sum_{L₂} \sum_{L₃} e^{-\Delta G(S₁\text{eff}, S₂\text{eff}, L₁, L₂, L₃)/k_BT},
\]

where \(\Delta G(S₁\text{eff}, S₂\text{eff}, L₁, L₂, L₃)\) is the free energy for a given structure:

\[
\Delta G(S₁\text{eff}, S₂\text{eff}, L₁, L₂, L₃) = \Delta G_{\text{stem}}(S₁\text{eff}) + \Delta G_{\text{stem}}(S₂\text{eff}) - T\Delta S(S₁\text{eff}, S₂\text{eff}, L₁, L₂, L₃).
\]

We read out \(\Delta S(S₁\text{eff}, S₂\text{eff}, L₁, L₂, L₃)\) from the entropy table. \(\Delta G_{\text{stem}}(S₁\text{eff})\) and \(\Delta G_{\text{stem}}(S₂\text{eff})\) are the folding free energy of the respective stems. \(\Delta G_{\text{stem}}(S_{\text{eff}})\) for a stem \((S₁\text{ or } S₂)\) is computed from the local partition function for the stem:

\[
\Delta G_{\text{stem}}(S_{\text{eff}}) = -k_BT \ln \sum_{\text{internal/bulge loops}} e^{-\Delta G_{\text{stem}}/k_BT}.\]

Here in the sum for stems with a given \(S_{\text{eff}}\), we consider the presence and absence of an internal or bulge loop and the different sizes and positions of the loop. The free energy of the stem \(\Delta G_{\text{stem}}\) in the above equation is the sum of the free energies for the base stacks and the loop in the stem, as determined by the nearest-neighbor model (Serra and Turner 1995; Cao and Chen 2005).

With the internal loops replaced by the effective helices in the loop entropy calculation, the conformational entropy for a general structure shown in Figure 5 is only dependent on five (instead of 11) parameters: \(S₁\text{eff}, S₂\text{eff}, L₁, L₂,\) and \(L₃.

As shown in Equation 3, the computation for the partition function is now much more efficient, and the computational time scales as \(n^5\) instead of \(n^{11}\) for an \(n\)-nt chain.

Using the recursive algorithm in Cao and Chen (2006b), we compute the total partition function \(Q(a, b)\) for a chain from \(a\) to \(b\). From the conditional partition function \(Q_{ij}\) for all the conformations that contain base pair \((i, j)\) between nucleotides \(i\) and \(j\), we compute the base-pairing probability \(P_{ij}\):

\[
P_{ij} = Q_{ij}/Q_{\text{tot}}.
\]

Here \(Q_{\text{tot}}\) is the total partition function for all the possible structures. From \(P_{ij}\) for all the possible \((i, j)\)’s, we can

---

**TABLE 1.** The sensitivity (SE) values for the structures predicted from seven different models

<table>
<thead>
<tr>
<th>Sequence ID</th>
<th>Reference</th>
<th>Vfold</th>
<th>Hotknots</th>
<th>ILM</th>
<th>pknotsRE</th>
<th>STAR</th>
<th>Pknots-RG</th>
<th>NUPACK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bt-PrP</td>
<td>van Batenburg et al. (2000)</td>
<td>0.42</td>
<td>0.41</td>
<td><strong>0.83</strong></td>
<td>0.5</td>
<td>0.33</td>
<td>0.33</td>
<td>0.41</td>
</tr>
<tr>
<td>BW-YV</td>
<td>van Batenburg et al. (2000)</td>
<td>1</td>
<td>1</td>
<td>0.88</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Ec-PK-1</td>
<td>van Batenburg et al. (2000)</td>
<td>1</td>
<td>1</td>
<td>0.36</td>
<td>1</td>
<td>0.36</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Ec-PK-3</td>
<td>van Batenburg et al. (2000)</td>
<td>0.84</td>
<td>0.68</td>
<td>0.52</td>
<td>0.68</td>
<td>0.68</td>
<td>0.68</td>
<td>1</td>
</tr>
<tr>
<td>Ec-S15</td>
<td>van Batenburg et al. (2000)</td>
<td>1</td>
<td>1</td>
<td>0.58</td>
<td>0.94</td>
<td>0.58</td>
<td>0.76</td>
<td>0.88</td>
</tr>
<tr>
<td>HIVRT32</td>
<td>Tuerk et al. (1992)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.9</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>HIVRT322</td>
<td>Tuerk et al. (1992)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.9</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>HIVRT33</td>
<td>Tuerk et al. (1992)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.9</td>
<td>1</td>
</tr>
<tr>
<td>Hs-PrP</td>
<td>van Batenburg et al. (2000)</td>
<td><strong>0.45</strong></td>
<td>0</td>
<td>0.27</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LP-PK1</td>
<td>van Batenburg et al. (2000)</td>
<td>0.9</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.8</td>
</tr>
<tr>
<td>minimalBV</td>
<td>Giedroc et al. (2000)</td>
<td><strong>0.94</strong></td>
<td><strong>0.94</strong></td>
<td>0.88</td>
<td><strong>0.94</strong></td>
<td><strong>0.88</strong></td>
<td><strong>0.94</strong></td>
<td><strong>0.94</strong></td>
</tr>
<tr>
<td>MMTV</td>
<td>Giedroc et al. (2000)</td>
<td>1</td>
<td>1</td>
<td>0.81</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>MMTV-vpk</td>
<td>Giedroc et al. (2000)</td>
<td>1</td>
<td>1</td>
<td>0.54</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>pKAA-A</td>
<td>Giedroc et al. (2000)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SRV-1</td>
<td>Giedroc et al. (2000)</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>T2-gene32</td>
<td>van Batenburg et al. (2000)</td>
<td>1</td>
<td>1</td>
<td>0.58</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>T4-gene32</td>
<td>van Batenburg et al. (2000)</td>
<td>1</td>
<td>1</td>
<td>0.63</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Tr-LSU-P3P</td>
<td>van Batenburg et al. (2000)</td>
<td>0.85</td>
<td><strong>0.95</strong></td>
<td>0.8</td>
<td>0.9</td>
<td>0.6</td>
<td>0.85</td>
<td><strong>0.95</strong></td>
</tr>
<tr>
<td>Average</td>
<td><strong>0.91</strong></td>
<td>0.84</td>
<td>0.68</td>
<td>0.86</td>
<td>0.71</td>
<td>0.84</td>
<td>0.85</td>
<td>1</td>
</tr>
</tbody>
</table>

The tested sequences are adapted from Table 1 in Ren et al. (2005). Our Vfold model gives much improved sensitivity values for the 18 pseudoknot sequences. In the calculation, the temperature is 37°C. Bold numbers show the highest accuracy.
predict the stable structures and the equilibrium folding pathways.

**STRUCTURAL PREDICTIONS**

**Comparison with other models**

We measure the accuracy of structure predictions by two parameters: (1) the sensitivity parameter $SE$, defined as the ratio between the number of correctly predicted base pairs and number of the base pairs in the experimentally determined structure; and (2) the specificity parameter $SP$, defined as the ratio between the number of correctly predicted base pairs and the total number of predicted base pairs. Our tests for structural predictions indicate that the model developed here gives better results than other models that we have tested (see Tables 1, 2). Specifically, our model gives the highest $SE$ value for 15 sequences among the total 18 sequences, and the highest $SP$ value for 13 sequences. In addition, our model gives higher overall average $SE$ (0.91) and $SP$ (0.91) than other models.

**H-type pseudoknot**

LP-PK1 and SRV-1 are two H-type pseudoknots. LP-PK1 is a PK1 domain of *Legionella pneumophila* tmRNA (Zwieb et al. 1999), and Simian retrovirus type-1 (SRV-1) (ten Dam et al. 1995) forms a pseudoknot that promotes the ribosomal frameshifting. For the two H-type pseudoknots, our Vfold model gives the highest $SE$ value (see Fig. 6A,B; Table 1). We note that the ILM model gives a false prediction for the SRV-1 pseudoknot. The failure of the ILM model may be due to the fact that the model does not account for the loop entropy for pseudoknots.

**VMV pseudoknot**

From a recent biochemical study, Pennell and Brierley and colleagues found that the stimulatory RNA for VMV frameshifting forms a pseudoknot structure (Pennell et al. 2008) instead of a stem–loop structure. Moreover, the pseudoknot is quite unique because it contains a long interhelix loop. We perform the structural prediction for this 67-nt RNA. Figure 7A shows that the predicted structure agrees exactly with the experimental structure, with $SE$ and $SP$ values both equal to 1.

**R2 retrotransposon pseudoknot**

The 5′ header of the R2 retrotransposon controls the R2 protein binding and cleavage of the DNA target (Christensen TABLE 2. The specificity ($SP$) values for the predicted structures from seven different models

<table>
<thead>
<tr>
<th>Sequence ID</th>
<th>Length</th>
<th>Vfold</th>
<th>Hotknots</th>
<th>ILM</th>
<th>pknotsRE</th>
<th>STAR</th>
<th>Pknoots-RG</th>
<th>NUPACK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bt-PrP</td>
<td>45</td>
<td>0.33</td>
<td>0.38</td>
<td><strong>0.76</strong></td>
<td>0.5</td>
<td>0.26</td>
<td>0.26</td>
<td>0.38</td>
</tr>
<tr>
<td>BWYV</td>
<td>28</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Ec-PK1</td>
<td>30</td>
<td>0.92</td>
<td>1</td>
<td>0.44</td>
<td>1</td>
<td>0.5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Ec-PK4</td>
<td>52</td>
<td>1</td>
<td>0.58</td>
<td>0.92</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Ec-S15</td>
<td>67</td>
<td>0.73</td>
<td>0.47</td>
<td>0.64</td>
<td>0.62</td>
<td>0.68</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>HIVRT32</td>
<td>35</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>HIVRT322</td>
<td>35</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>HIVRT33</td>
<td>35</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Hs-PrP</td>
<td>45</td>
<td>0.5</td>
<td>0.27</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LP-PK1</td>
<td>30</td>
<td>0.9</td>
<td>1</td>
<td>0.71</td>
<td>0.83</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>minimalIBV</td>
<td>45</td>
<td><strong>0.94</strong></td>
<td>0.88</td>
<td>0.88</td>
<td><strong>0.94</strong></td>
<td>0.93</td>
<td><strong>0.94</strong></td>
<td><strong>0.94</strong></td>
</tr>
<tr>
<td>MMTV</td>
<td>34</td>
<td>1</td>
<td>0.91</td>
<td>0.81</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0</td>
</tr>
<tr>
<td>MMTV-vpk</td>
<td>34</td>
<td>0.92</td>
<td>0.91</td>
<td>0.54</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>1</td>
</tr>
<tr>
<td>pKA-A</td>
<td>36</td>
<td>1</td>
<td>0.92</td>
<td>0.92</td>
<td>0.92</td>
<td>0.92</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td>SRV-1</td>
<td>38</td>
<td>1</td>
<td>0.91</td>
<td>0</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>T2-gene32</td>
<td>33</td>
<td>1</td>
<td>1</td>
<td>0.7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>T4-gene32</td>
<td>28</td>
<td>1</td>
<td>0.87</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Tt-LSU-P3P7</td>
<td>65</td>
<td>0.85</td>
<td>1</td>
<td>0.69</td>
<td>0.85</td>
<td>0.75</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td><strong>0.91</strong></td>
<td>0.86</td>
<td>0.71</td>
<td>0.85</td>
<td>0.76</td>
<td>0.86</td>
<td>0.85</td>
</tr>
</tbody>
</table>

The tested sequences are adapted from Table 1 in Ren et al. (2005). Our Vfold model gives much improved specificity values for the 18 pseudoknot sequences. Bold numbers show the highest accuracy.

**FIGURE 6.** The predicted structures for three pseudoknots. (A) For LP-PK1, Hotknots (Ren et al. 2005), ILM (Ruan et al. 2004), pknotsRE (Rivas and Eddy 1999), STAR (Gultyaev et al. 1995), and pknots-RG (Reeder and Giegerich 2004) all give poor predictions for the structure ($SE = 0.5$). NUPACK (Dirks and Pierce 2003) gives a relatively high $SE$ value ($SE = 0.8$). Our Vfold model gives the highest accuracy with $SE = 0.9$. (B) For the SRV-1 pseudoknot, the ILM model fails to predict the native structure of the SRV-1 pseudoknot. (C) For the 70.8 anti-HIV aptamer, we predict a pseudoknot with a 3-nt interhelix loop. In the calculations, the temperature is 37°C for $A$ and $B$ and 20°C for $C$ according to the experimental condition (Held et al. 2006a,b). Also shown in the figures are the density plot for the base-pairing probability $P_{ij}$ (Equation 4). In the density plots, the horizontal and vertical axes denote the indices of the nucleotides $i$ and $j$.**
et al. 2006). Based on the NMR spectra and computational models (Hart et al. 2008), Hart and Turner and colleagues found a knotted structure in the 74-nt header of the R2 retrotransposon. In this study, we use the Vfold model developed here to predict the secondary structure for the 74-nt header. Figure 8A shows our predicted structure. The predicted structure is a pseudoknot with four stems. All four stems have been found in the experiments (Hart et al. 2008). The predicted structure shows a high accuracy with SE = 1.0 and SP = 1.0. In the calculation, we have added the base-stacking energy for the Watson-Crick base pairs between nucleotides 48CG49 and 62CG63 (see the dashed lines in Fig. 8A). This tertiary interaction has been confirmed in previous NMR measurement (Ferre´-D’Amaré et al. 1998).

Anti-HIV RNA aptamer

Recently experiments suggested that the interhelix loop may be essential for efficient ribosomal frameshifting (Brierley et al. 2008; Giedroc and Cornish 2008). Moreover, previous experimental studies on the anti-HIV RNA aptamer (Held et al. 2006a,b) suggested that the interhelix loop, which causes flexible helix orientations in the pseudoknot aptamers, may play an important functional role in accommodating aptamer binding to the HIV reverse transcriptase. For example, for an aptamer (labeled as 70.8 according to the notations used in the literature) (Held et al. 2006a,b), the proposed native structure contains a 3-nt interhelix loop. The predicted structure from our model (Fig. 6C), indeed, shows a 3-nt interhelix loop. The structure has a high accuracy of SE = 0.9 and SP = 1.0 if we treat the experimentally proposed structure (Held et al. 2006a,b) as the “experimental” structure.
stability against temperature increase. At $T = 80^\circ C$, both stems are unfolded. Thus, stems $S_1$ and $S_2$ have the comparable thermal stability.

VMV pseudoknot

A recent combined biochemical and NMR experiment (Pennell et al. 2008) showed that the VMV pseudoknot contains a 6-n interhelix loop. Our predicted unfolding pathway suggests that at $T = 80^\circ C$, stem $S_2$ is the first stem to be unzipped, and stem $S_1$ is the last one to be unzipped (see Fig. 7B,C). Our prediction agrees with the experimental finding (Pennell et al. 2008), which suggested that $S_1$ is the most stable and $S_2$ is disrupted at a temperature around 76.8°C.

R2 retrotransposon pseudoknot

The native structure of the R2 retrotransposon pseudoknot contains four stems (Fig. 8A). The structure shows high stability against temperature increase. At $T = 80^\circ C$, both stems are unfolded. Thus, stems $S_1$ and $S_2$ have the comparable thermal stability.

FOLDING THERMODYNAMICS

The 70.8 aptamer

As the temperature increases, stems $S_1$ and $S_2$ of the pseudoknot (see Fig. 6A) is disrupted at nearly the same temperature. At $T = 80^\circ C$, both stems are unfolded. Thus, stems $S_1$ and $S_2$ have the comparable thermal stability.

R2 retrotransposon pseudoknot

The native structure of the R2 retrotransposon pseudoknot contains four stems (Fig. 8A). The structure shows high stability against temperature increase. At $T = 80^\circ C$, stem $S_1$ is the first stem to be unfolded, resulting in an intermediate state that contains stem $S_3$ and a partially unfolded stem $S_2$. As the temperature is further increased to 90°C, stem $S_2$ becomes completely unzipped since the hairpin with $S_2$ is destabilized by the large loop. Stem $S_1$ is the most robust stem and is the last stem to be unzipped. The melting temperature for $S_3$ is $\sim$100°C.

SUMMARY

In summary, we have developed a new virtual-bond-based model (Vfold) for general RNA pseudoknots with inter-helix loops. The model allows an accurate treatment for the loop–helix excluded volume interactions and rigorous calculations for the conformational entropy for general pseudoknotted folds. Tests against other existing models suggest that this new model gives improved predictions for the native structures, with average sensitivity and specificity measures of the accuracy equal to 0.91 and 0.91, respectively. We attribute the improved accuracy to the rigorous conformational entropy parameters. For any given RNA sequence, the model enables predictions for not only the native structures, but also the folding stabilities and equilibrium folding pathways. Despite the success of this new model, it has several limitations that should be removed in future model development. First, the model does not treat possible noncanonical interactions such as base triple interactions between loops and stems and noncanonical base-pairing between loop nucleotides. These interactions can be biologically important for more complex pseudoknotted structures (Cornish et al. 2005; Theimer et al. 2005). Moreover, ions, especially Mg$^{2+}$ ions, can play an important role in loop entropy and the global folding stability of pseudoknots (Chen 2008; Tan and Chen 2008).

SUPPLEMENTAL MATERIAL

Supplemental material can be found at http://www.rnajournal.org.

ACKNOWLEDGMENTS

We thank Professor Donald H. Burke for useful discussions. The research was supported by NIH through grant GM063732 (to S.-J.C.). Most of the computations involved in this research were performed on the HPC resources at the University of Missouri Bioinformatics Consortium (UMBC).

Received October 21, 2008; accepted January 10, 2009.

REFERENCES

A recent combined biochemical and NMR experiment (Pennell et al. 2008) showed that the VMV pseudoknot contains a 6-n interhelix loop. Our predicted unfolding pathway suggests that at $T = 80^\circ C$, stem $S_2$ is the first stem to be unzipped, and stem $S_1$ is the last one to be unzipped (see Fig. 7B,C). Our prediction agrees with the experimental finding (Pennell et al. 2008), which suggested that $S_1$ is the most stable and $S_2$ is disrupted at a temperature around 76.8°C.

SUMMARY

In summary, we have developed a new virtual-bond-based model (Vfold) for general RNA pseudoknots with inter-helix loops. The model allows an accurate treatment for the loop–helix excluded volume interactions and rigorous calculations for the conformational entropy for general pseudoknotted folds. Tests against other existing models suggest that this new model gives improved predictions for the native structures, with average sensitivity and specificity measures of the accuracy equal to 0.91 and 0.91, respectively. We attribute the improved accuracy to the rigorous conformational entropy parameters. For any given RNA sequence, the model enables predictions for not only the native structures, but also the folding stabilities and equilibrium folding pathways. Despite the success of this new model, it has several limitations that should be removed in future model development. First, the model does not treat possible noncanonical interactions such as base triple interactions between loops and stems and noncanonical base-pairing between loop nucleotides. These interactions can be biologically important for more complex pseudoknotted structures (Cornish et al. 2005; Theimer et al. 2005). Moreover, ions, especially Mg$^{2+}$ ions, can play an important role in loop entropy and the global folding stability of pseudoknots (Chen 2008; Tan and Chen 2008).

SUPPLEMENTAL MATERIAL

Supplemental material can be found at http://www.rnajournal.org.

ACKNOWLEDGMENTS

We thank Professor Donald H. Burke for useful discussions. The research was supported by NIH through grant GM063732 (to S.-J.C.). Most of the computations involved in this research were performed on the HPC resources at the University of Missouri Bioinformatics Consortium (UMBC).

Received October 21, 2008; accepted January 10, 2009.

REFERENCES


Pseudoknots with interhelix loops


www.rnajournal.org 705


van Belkum, A., Abrahams, J., Pleij, C., and Bosch, L. 1985. Five pseudoknots are present at the 204 nucleotides long 3′ noncoding region of tobacco mosaic virus RNA. Nucleic Acids Res. 13: 7673–7686.


