RNA

A PUBLICATION OF THE RNA SOCIETY

Predicting structure and stability for RNA complexes with intermolecular loop–loop base-pairing

Song Cao, Xiaojun Xu and Shi-Jie Chen

Supplemental Material	http://rnajournal.cshlp.org/content/suppl/2014/04/03/rna.043976.113.DC1.html
P<P	Published online April 21, 2014 in advance of the print journal.
Creative Commons License	This article is distributed exclusively by the RNA Society for the first 12 months after the full-issue publication date (see http://rnajournal.cshlp.org/site/misc/terms.xhtml). After 12 months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.
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

Predicting structure and stability for RNA complexes with intermolecular loop–loop base-pairing

SONG CAO,[1,2,3] **XIAOJUN XU,**[1,2] and **SHI-JIE CHEN**[1,4]

[1]Department of Physics and Department of Biochemistry, University of Missouri, Columbia, Missouri 65211, USA

ABSTRACT

RNA loop–loop interactions are essential for genomic RNA dimerization and regulation of gene expression. In this article, a statistical mechanics-based computational method that predicts the structures and thermodynamic stabilities of RNA complexes with loop–loop kissing interactions is described. The method accounts for the entropy changes for the formation of loop–loop interactions, which is a notable advancement that other computational models have neglected. Benchmark tests with several experimentally validated systems show that the inclusion of the entropy parameters can indeed improve predictions for RNA complexes. Furthermore, the method can predict not only the native structures of RNA/RNA complexes but also alternative metastable structures. For instance, the model predicts that the SL1 domain of HIV-1 RNA can form two different dimer structures with similar stabilities. The prediction is consistent with experimental observation. In addition, the model predicts two different binding sites for hTR dimerization: One binding site has been experimentally proposed, and the other structure, which has a higher stability, is structurally feasible and needs further experimental validation.

Keywords: statistical mechanical model; folding thermodynamics; structure prediction

INTRODUCTION

Intermolecular loop–loop base-pairing is a widespread and functionally important tertiary structure motif in RNA. For example, intermolecular loop–loop interactions are found in complementary anticodon–anticodon pairs between different tRNAs (Eisinger 1971; VanLoock et al. 1999). Loop–loop interactions often facilitate dimerization reactions between RNA molecules (Jossinet et al. 1999; Kolb et al. 2000a,b). In humans, the loop–loop contact is important for the dimerization of human telomerase RNA (hTR). Dimer-destabilizing mutants can result in low telomerase activity and disease (Ly et al. 2003; Ren et al. 2003; Theimer and Feigon 2006). In bacteria, loop–loop interaction can regulate gene expression and affect replication and translation of the bacteria (Schmidt et al. 1995; Argaman and Altuvia 2000; Repoila et al. 2003; Bossi and Figueroa-Bossi 2007; Vogel and Wagner 2007). For example, OxyS RNA repression of fhlA translation in *Escherichia coli* through the formation of a stable loop kissing interaction is one well-documented case (Argaman and Altuvia 2000). In viruses, long-range loop–loop interactions can block translation of certain sequence fragments and affect viral replication. For instance, the R3.5 RNA can tightly bind to the T-shaped domain (TSD) in tomato bushy stunt virus (TBSV) (Miller and White 2006) and regulate gene expression. Moreover, in HIV-1 virus, the loop–loop kissing interaction is critical for one form of HIV-1 dimerization (Laughrea and Jette 1994; Muriaux et al. 1996; Paillart et al. 2004).

Current computational models for the prediction of RNA/RNA complex formation are mainly focused on secondary structures (Mathews et al. 1999; Zhang and Chen 2001; Dimitrov and Zuker 2004; Rehmsmeier et al. 2004; Andronescu et al. 2005; Cao and Chen 2006a; Muckstein et al. 2006; Dirks et al. 2007). In particular, the physics-based models can be classified into two categories: minimum free-energy methods and partition function methods. The algorithms based on the minimum free energy are extensions of algorithms from single-stranded RNAs to RNA/RNA complexes (Mathews et al. 1999; Andronescu et al. 2005). The partition function-based method uses Boltzmann-weighted statistics for the complete ensemble of secondary structures. By calculating the base-pairing probability for each nucleotide pair, the partition function method gives all the probable structures. Moreover, from the partition function, one can

predict the melting curves and folding thermodynamics from the sequence (Zhang and Chen 2001; Cao and Chen 2006a; Muckstein et al. 2006; Dirks et al. 2007).

The partition function is a sum over all the possible structures. For each structure, the free energy is determined by the free energies of the constituent helices and loops. The loop free-energy calculation directly impacts the result of the partition function and the base-pairing probabilities (RNA structure). The entropy (free-energy) parameters for simple loops (hairpin, bulge, and internal loops) have been determined from thermodynamic experiments (Serra and Turner 1995). However, because of conformational coupling between loops, the loop entropies are not additive for tertiary motifs such as loop–loop kissing contacts. Furthermore, thermodynamic experiments alone are not sufficient to provide individual loop entropies because of loop entropy nonadditivity and system complexity. We need a new model. In this article, we develop a model for structures with kissing loop complexes.

Most current folding algorithms for RNA/RNA complexes employ a virtual link to connect the two RNA molecules and convert the original two-RNA system into an effective one-RNA system. The algorithms then predict the structure of the equivalent one-RNA system at the secondary structure level. However, many biologically important RNA/RNA complexes involve intermolecular loop–loop contacts such that the effective one-RNA system goes beyond the secondary structural level. For example, two hairpin loops form the kissing loop complex for an HIV-1 dimer. Such a pseudoknotted fold cannot be treated by the existing secondary structure prediction algorithms (Bon and Orland 2011) for RNA/RNA complexes. Computational prediction of intermolecular loop–loop kissing interactions is a new challenge (Alkan et al. 2006; Busch et al. 2008; Chitsaz et al. 2009; Chen et al. 2009; Huang et al. 2009). Furthermore, most existing programs are devoted to finding the binding sites between two RNA molecules (Long et al. 2008; Huang et al. 2009). These programs cannot provide information about structures away from the binding site or the impact of the structures on the binding affinity. To predict the full structure and thermal stability of RNA/RNA complexes that involve intermolecular loop–loop kissing complexes, we need a new model. One of the challenges in such a new model is the calculation of the conformational entropy and the free energy.

Calculation of conformational entropy and free energy is a major challenge in the development of a new model. Entropy can be computed either from simulations methods such as molecular

dynamics simulations or from direct conformational enumeration. Molecular dynamics simulations has the advantage of accounting for atomistic force fields. Due to the complications of conformational sampling and force field, most RNA folding simulations are focused on relatively small systems such as RNA tetraloops and small kissing complexes (Chen and Garcia 2013; Kuhrova et al. 2013; Stephenson et al. 2013). A reliable computational sampling for flexible loop conformations with all-atom resolution requires special sampling technique (Schafer et al. 2001) and can take exceedingly long computational time. To overcome the difficulty requires the use of special simulation techniques to achieve reliable sampling for flexible loops. An alternative approach is to use a coarse-grained structure model to enumerate conformations. We previously developed the Vfold model for predicting secondary and pseudoknot structure within single-stranded RNA molecules (Cao and Chen 2005, 2006b, 2009). The model uses coarse-grained (virtual bond) conformations of RNA and predicts the conformational entropy for a given structure by allowing loops/junctions to fluctuate in three-dimensional space (see Fig. 1). In the Vfold model, RNA loop conformations are generated through random walks of virtual bonds on a diamond lattice. An advantage of the model for the loop entropy calculation is the ability to account for chain connectivity, the excluded volume effect, and the completeness of the conformational ensemble. Furthermore, the model provides the conformational entropy for loops with (mismatched) intraloop base pairs and accounts for the effect of intraloop base pairs. Studies by us and other groups show that an accurate entropy parameter improves the prediction of RNA secondary structures and thermodynamic stabilities (Andronescu et al. 2010;
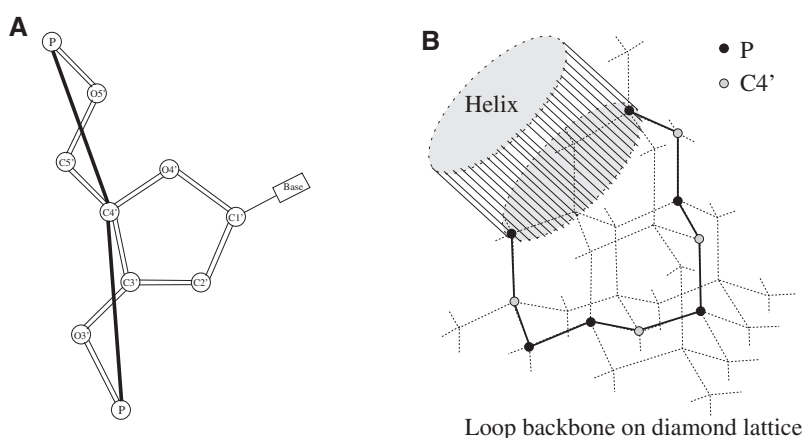
FIGURE 1. (*A*) The virtual bond model for RNA conformation. Each nucleotide is represented by two (virtual) bonds. The virtual bonds have bond length of ~3.9 Å and bond angles in the range of 90°–120°, as determined from the known RNA structures. As a coarse-grained representation, RNA virtual bond conformations can be configured on a diamond lattice with three equiprobable torsional angles (60°, 180°, 300°). (*B*) A schematic diagram of a hairpin (virtual bond) conformation. The loop conformation can be generated through self-avoiding walks of the virtual bonds on the diamond lattice. An advantage of the virtual-bond model is the ability to account for the excluded volume and chain connectivity.

Sperschneider and Datta 2010; Sperschneider et al. 2011). The results suggest that proper treatment of excluded volume, chain connectivity, completeness of the conformational ensemble, and intraloop mismatched base pairs can lead to notable improvement in the prediction of RNA folding (Cao and Chen 2005, 2006b, 2009). Recently, we used the Vfold model to calculate the entropy parameters and predict the structure and stability of the simple kissing interaction between two hairpin loops (Cao and Chen 2011). Tests against experimental data suggest that the entropy parameters for the formation of the kissing interaction may be reliable.

In this article, we develop a new method that can treat intermolecular base-pairing (kissing) between general loop types, such as hairpin loops, internal loops, and multibranched junctions. A main advancement in the current new model is a method for estimating the entropy parameters (Cao and Chen 2011) for structures with general loop–loop base-pairing. The model enables us to predict the structure and folding thermodynamics of complicated RNA/RNA complexes. Moreover, the essence of the model is to parse the structure prediction for the whole system into two steps. First, we identify the binding sites. Then, we calculate the base-pairing probabilities for a given binding mode. If the number of computational operations for steps 1 and 2 are t_1 and t_2, the total number of computational operations for predicting stable structures would scale as $t_1 + t_2$. In contrast, without using the above two-step procedure, the number of computational operations would be $t_1 \times t_2$, which is much larger than $t_1 + t_2$. Thus, the new computational model significantly improves computational efficiency. As shown below, the new model yields high sensitivity (SE) and specificity in the structure prediction for a variety of RNA/RNA complexes.

THEORY AND MODEL

Entropy of kissing loops

At the center of the statistical thermodynamics is the calculation of the partition function Q, which is the sum over all the possible structures s: $Q = \sum_s e^{-(\Delta H - T\Delta S)/k_B T}$, where ΔH and ΔS are the enthalpy and entropy parameters for the structure. $(\Delta H, \Delta S)$ for a helix stem can be calculated from the Turner rules (Serra and Turner 1995). A key problem in calculating the partition function is determining loop entropy parameters (ΔS) for the structures. In the present study, an RNA structure is defined by the base pairs (and loops and stems) in the structure. We call such a

structure a 2D structure. Figure 2A shows a schematic 2D structure for two RNA molecules bound through loop–loop base-pairing. The loop conformations are affected by the intermolecular helix and the stem–loops attached to the kissing loops. Exact enumeration of all loop conformations is computationally intractable because of the system's large size and complexity. One complication comes from the stem–loop substructures connected to the (multibranched) loop. The stem–loop substructures' impact on loop entropy is approximated by replacing the terminal base pairs (of the stems) by single nucleotides. Then, the effective loop lengths of l_1, l_2, l_3, or l_4 are equal to the sum of the number of unpaired nucleotides and the number of stem–loop substructures. With this loop entropy approximation, we can reduce the original complicated system in Figure 2A into a simple hairpin–hairpin kissing complex as shown in Figure 2B, which includes three helices (helix 1, helix 2, and helix 3) and four junctions/loops (l_1^{eff}, l_2^{eff}, l_3^{eff}, and l_4^{eff} in the figure). This approximation enables us to treat general kissing motifs. The approximation ignores the (weak) excluded volume interference between the stem–loop substructure and the loop. In the Supplemental Material, using simple test systems, we show the results of a series of tests that support the validity of the approximation.

In our previous work (Cao and Chen 2011), we developed a standard method to calculate the entropy parameter for a simple kissing hairpin system. In our method, we model the helices using the coordinates of an A-form helix. In a cylindrical coordinate system, the (r, θ, z) coordinates for the P, C_4, and N_1 (or N_9) in one strand are (8.71 Å, 70.5 + 32.7 i,
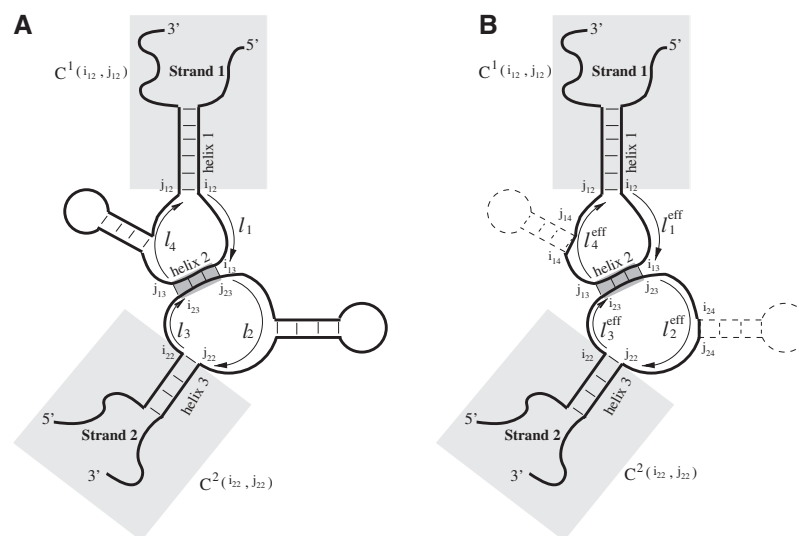
FIGURE 2. (*A*) A schematic diagram for the structure of an RNA/RNA complex that involves an intermolecular kissing interaction (see the region in dark gray). (*B*) In the calculation of the (kissing) loop entropy, the original RNA/RNA complex structure in *A* is converted to an effective hairpin-hairpin kissing system by reducing a stem–loop substructure to a single nucleotide. The first number in the subscript denotes the strand (1 or 2), and the second number denotes the nucleotide position in the respective strand.

−3.75 + 2.81 i), (9.68 Å, 46.9 + 32.7 i, −3.10 + 2.81 i), and (7.12 Å, 37.2 + 32.7 i, −1.39 + 2.81 i) ($i = 0,1,2,\ldots$) (Arnott et al. 1972), respectively. For the other strand, we only need to negate θ and z. Two atoms, P and C_4, are used to describe the backbone configurations, and N_1 (or N_9) is an additional atom used to describe the base orientation (Cao and Chen 2009). For the loops, three isomeric states (g^+, t, g^{-1}) are used to sample the backbone conformations (Flory 1969). Since the three isomeric states can be exactly configured in the diamond lattice (Cao and Chen 2009), we can enumerate the loop conformation through self-avoiding walks in the diamond lattice. By counting the total number of viable loop conformations (Ω), we can obtain the entropy change for the formation of the kissing loop ΔS as k_B In (Ω/Ω_{Coil}). Here Ω_{Coil} is the conformational count for the random coil chain. We note that hairpin loop kissing complexes often favor the formation of coaxial stacking interactions. The entropy parameters for the hairpin kissing complexes listed in Table 1 of the article by Cao and Chen (2011) are for hairpin kissing systems with three helices coaxially stacked.

A new computational model

The previously developed model (Cao and Chen 2011) can only treat simple kissing interaction between hairpin loops. For large, more complex structures of long RNA sequences, the previous method is not useful due to (1) lack of entropy parameters for more complex loop–loop contacts (such as the kissing interactions between two internal loops or multibranched junctions) and (2) the huge conformational sampling space for large structures. These problems motivate us to develop a new computational model.

Finding the binding sites

We first use a search algorithm, implemented in the Vfold model, to find the binding region between strand 1 and

strand 2. In order to identify the binding region, we need to determine the starting and ending nucleotides for the binding sites between strand 1 and strand 2 (see nucleotides i_{13}, j_{13}, i_{23}, and j_{23} in Fig. 2A). Here, the first number in the subscript for i and j denotes the strand (1 or 2), and the second number denotes the position of the nucleotide in the respective strand along the 5′-to-3′ direction. The conditional partition function $Q^{(M)}$ sums over all the conformations for the two RNAs bound at the given site (i_{13}, j_{13}, i_{23}, j_{23}) (denoted as mode "M"):

$$Q^{(M)} = \sum_{l_1^{eff}, l_2^{eff}, l_3^{eff}, l_4^{eff}} Q_1(i_{13}, \ j_{13}) \cdot Q_2(i_{23}, \ j_{23}) \cdot e^{-\Delta G_{kiss}/k_B T} \tag{1}$$

$$Q_1(i_{13}, \ j_{13}) = C^1(i_{12}, \ j_{12}) \cdot O^1(i_{12}, \ i_{13}, \ l_1^{eff}) \cdot O^1(j_{13}, \ j_{12}, \ l_4^{eff}) \cdot \tag{2}$$

$$Q_2(i_{23}, \ j_{23}) = C^2(i_{22}, \ j_{22}) \cdot O^2(i_{22}, \ i_{23}, \ l_3^{eff}) \cdot O^2(j_{23}, \ j_{22}, \ l_2^{eff}). \tag{3}$$

The physical meanings of the above equations are explained below. In Equation 1, the sum is over all the possible (2D) structures with intermolecular base-pairing at the binding site (i_{13}, j_{13}, i_{23}, j_{23}). The partition function for each structure is calculated as the product of partition functions for its constituent structural domains, as explained below. We use strand 1 for illustration. Strand 1 consists of four segments. From 5′ to 3′, these segments are as follows: the structure closed by the base pair (i_{12}, j_{12}) (Fig. 2A, upper shaded gray region), the open structure from i_{12} to i_{13}, the bound region from i_{13} to j_{13}, and the open structure from j_{13} to j_{12}. Here a "closed (open) structure" is defined as a structure whose terminal nucleotides are (not) base-paired. As shown by Equation 2, the partition function for strand 1, $Q_1(i_{13}, j_{13})$, is given by the product of the partition functions of the above segments.

1. $O^{1 \ or \ 2}(i, j, l^{eff})$ is the partition function for all the open conformations from nucleotides i to j with an effective loop length l^{eff} (see Fig. 2B), and the superscripts, 1 and 2, represent strands 1 and 2, respectively. The partition

TABLE 1. The sensitivity (SE) and positive predictive value (PPV) for structures predicted from three different models

Sequence ID	l_1/l_2	Our model SE/PPV	PairFold	IntaRNA	References
SL1/SL1 complex in HIV-1	35/35	1.0/1.0	1.0/1.0	0/0	Ulyanov et al. (2006)
Domain IIa/IIa-14t aptamer	19/36	0.96/1.0	0.35/0.53	0.48/0.55	Da Rocha Gomes et al. (2004)
Domain IV/IV-55	21/67	0.77/0.75	0.81/0.93	0.81/0.86	Aldaz-Carroll et al. (2002)
SL1/5-39	33/77	0.90/0.92	0.74/0.91	0.77/0.83	Aldaz-Carroll et al. (2002)
fhlA/OxyS (Site 1)	nucleotides −53 to 7/ nucleotides 64–109	0.90/0.82	0.70/0.58	0/0	Argaman and Altuvia (2000)
fhlA/OxyS (Site 2)	nucleotides 8 to 60/ nucleotides 1–63	0.90/0.88	0.78/0.84	0.71/0.69	Argaman and Altuvia (2000)
hTR/hTR complex	64/64	0.79/0.63	0.58/0.61	0.50/0.56	Ren et al. (2003)
RNA IIi/RNA Ii	19/21	0.85/1.0	0.35/0.41	0.85/0.94	Lee and Crothers (1998)
Average		0.88/0.87	0.66/0.73	0.52/0.55	

We used eight available RNA/RNA complexes to test the accuracies of model predictions. In the calculations, the ion concentration was assumed to be the standard 1 M Na$^+$ and the temperature is 25°C, which corresponds to room temperature in the experiments. For all the three models, the only input information required by the model is the sequences of the two chains.

function includes all of the secondary and pseudoknot structures. The partition function $O^{1 \text{ or } 2}(i, j, l^{\text{eff}})$ can be calculated from the recursive algorithm for a single-stranded RNA (Cao and Chen 2005).

2. $C^{1 \text{ or } 2}(i, j)$ is the partition function for all of the conformations closed by intrachain base pairs (i, j) (Cao and Chen 2006a); see the shaded gray regions in Figure 2.

3. ΔG_{kiss} is the free-energy change upon the formation of the kissing loop complex. ΔG_{kiss} has two components: the free-energy change ΔG_{bp} for intermolecular base-pairing between the loops and the entropy change ΔS due to loop formation and conformational restriction. $\Delta G_{\text{kiss}} = \Delta G_{\text{bp}} - T\Delta S$. Here, ΔG_{bp} can be estimated from the base-pairing/stacking free energies as given by the Turner rules (Mathews et al. 1999) and the nucleation free energy for the complex formation, which is strand concentration dependent (Cao and Chen 2006a). ΔS is the entropy decrease caused by the formation of the kissing complex. As a crude approximation, ΔS can be extracted from the precalculated parameters for kissing loops (Table 1 in Cao and Chen 2011). The entropy parameter for the formation of kissing loop complex is dependent on the length of the interloop helix (the dark gray region in the figure) and the effective lengths of four junctions/loops (l_1^{eff}, l_2^{eff}, l_3^{eff}, and l_4^{eff}) shown in Figure 2B. We allow the formation of a bulge loop in the interloop helix, so the helix length is $min(j_{13} - i_{13} + 1, j_{23} - i_{23} + 1)$. The minimum stem length at the binding site is one base stack, i.e., two consecutive base pairs.

From the partition function, we can compute the probability of binding at $(i_{13}, j_{13}, i_{23}, j_{23})$:

$$P^{(M)} = \frac{Q^{(M)}}{Q_1^{(tot)} Q_2^{(tot)}}, \tag{4}$$

where $Q_1^{(tot)}$ and $Q_2^{(tot)}$ are the total partition functions for strands 1 and 2, respectively. The mode M with the maximum probability $P^{(M)}$ is the most probable binding mode.

Predicting the structures for the bound complex

For the predicted binding mode M, we calculate the probability p_{ij} for nucleotides i and j to form a base pair. From the p_{ij} distribution for all the nucleotides, we can determine the (2D) structures of the RNA complex.

We assume only one binding site exists for an RNA/RNA complex, such as the nucleotides from i_{13} to j_{13} in strand 1 and the nucleotides from i_{23} to j_{23} in strand 2. Based on the approximation, we can use the truncated sequences from nucleotide i_{13} to nucleotide j_{13} for strand 1 and from nucleotide i_{23} to nucleotide j_{23} for strand 2 to predict the intermolecular kissing base pairs (Cao and Chen 2006a).

The base-pairing probability p_{ij} for the intramolecular interactions can be calculated from the conditional partition function $Q_{ij}^{(M)}$ for all the conformations that contain (i, j)

base pair and the total partition function $Q^{(M)}$ for all the possible conformations for the given binding mode M:

$$p_{ij} = \frac{Q_{ij}^{(M)}}{Q^{(M)}}, \tag{5}$$

where the conditional partition function $Q_{ij}^{(M)}$ can be calculated from Equations 1 through 3 with the constraint that nucleotides (i, j) form a base pair.

From the partition functions, the base-pairing probabilities, and the predicted binding sites at the different temperatures and RNA concentrations, we can predict the structure and the equilibrium folding pathway.

For a given sequence, the Vfold model (Cao and Chen 2005, 2006b, 2009, 2011) uses a recursive algorithm to enumerate all the possible secondary and pseudoknotted structures. To partially account for the sequence-dependent intraloop interactions, the model enumerates all the possible arrangements of mismatched base stacks within a loop. The formation of mismatched base stacks in a loop is sequence dependent and can cause a significant reduction in the loop entropy. For a given set of intraloop base stacks, the reduced loop entropy can be determined from the Vfold model. For practical use of the entropy parameters, we have systematically calculated and tabulated the entropic parameters for the different types of loops, including hairpin, internal, bulge, H-type pseudoknot, and hairpin-hairpin kissing loops (Cao and Chen 2005, 2006b, 2009, 2011). For a more detailed explanation about the Vfold algorithm, see the Supplemental Material.

RESULTS AND DISCUSSION

Structure prediction

Computational time

We use two sequence fragments from the hepatitis C virus (HCV) genome to show the computational time (Romero-López and Berzal-Herranz 2009). The lengths of the two strands are 154 and 136 nucleotides (nt), respectively. Figure 3 shows the dependence of the computational time on the sequence length l. The nonlinear scaling relationship shown in Figure 3 is due to the nonlinear increase of the number of structures with the RNA sequence length. The computer resource we used was a Dell PC desktop with dual cores (Intel Xeon 5150 [2.66-GHz] processor). The result shows that it is computationally feasible (about ≤100 h) to predict the complex structure of two RNA binding using the current computational model for total sequence length $l_1 + l_2 \leq 140$ nt, where l_1 and l_2 denote the lengths of strands 1 and 2, respectively.

Comparison of our model with other existing models

In Table 1, we measure the accuracy of the model prediction by two parameters, the SE and the positive predictive value
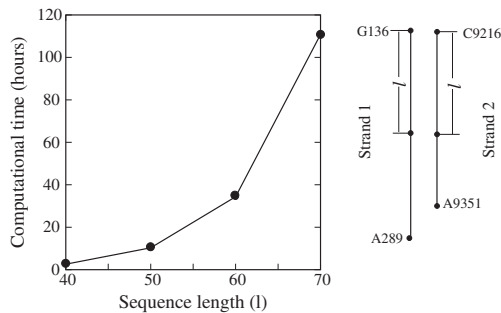
FIGURE 3. The computational time for predicting the optimal binding sites (the *left* panel). In the test calculations, we used two strands with equal length selected from the HCV genome (Romero-López and Berzal-Herranz 2009). The *x*-axis is the length of each strand defined in the *right* panel, and the *y*-axis is the computational time (in hours). Due to the rapid increase of the number of possible secondary structures with the sequence length, our current model can only treat medium-size RNA/RNA complex.

(PPV), which are defined as the ratios between the correctly predicted base pairs and the total number of base pairs in the experimental and in the predicted structures, respectively. The benchmark test results in Table 1 show that our model yields improved predictions compared to other models (PairFold [Andronescu et al. 2005] and IntaRNA [Busch et al. 2008; Chitsaz et al. 2009]). PairFold was originally developed to predict secondary structures, not kissing complex-

es for RNA/RNA binding. IntaRNA (Chitsaz et al. 2009) can predict the loop–loop kissing complex structures but is based on a simplified thermodynamic model for these tertiary interactions and cannot account for the entropy contributions from kissing loops. We attribute the improvement of our predictions to the ability to treat intermolecular loop–loop kissing base pairs and the use of physics-based entropy parameters for the kissing interactions.

IIa/IIa-14t and SL1/5-39 complexes

IIa-14t and 5–39 are two aptamers selected from SELEX (systematic evolution of ligands by exponential enrichment) (Da Rocha Gomes et al. 2004). The two aptamers tightly bind to the IIa domain and SL1 domain of HCV, respectively. The 2D structures of IIa/IIa-14t (Da Rocha Gomes et al. 2004) and SL1/5-39 complexes (Aldaz-Carroll et al. 2002) were determined from biochemical experiments. We predicted the structures using the two-step procedure described in the *Theory and Model*. Figure 4, A and B, shows the predicted binding site $(i_{13}, j_{13}, i_{23}, j_{23}) = (8,12,27,31)$ for the IIa/IIa-14t complex. Based on the predicted binding site, we further predicted the intramolecular base pairs for both the aptamer IIa-14t and the domain IIa of HCV. Figure 4, C and D, shows the calculated probability for the intramolecular base pairs. From the base-pairing probability and the binding sites, we predicted the structure of the IIa/IIa–14t complex (see Fig.
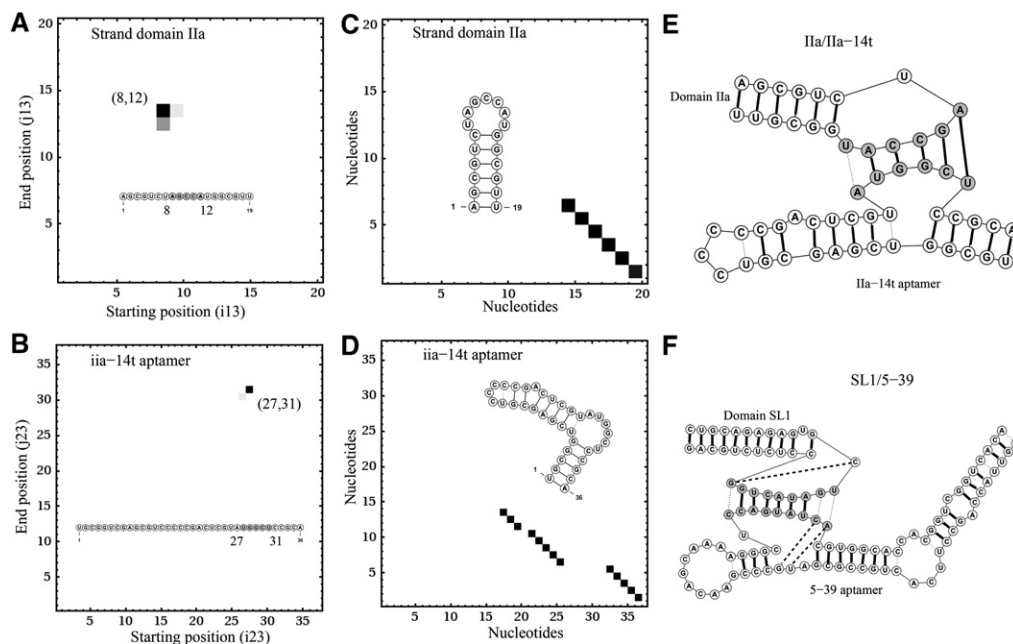
FIGURE 4. The predicted binding positions (starting and ending nucleotides) at room temperature for the IIa/IIa-14t complex in strand IIa (*A*) and IIa-14t aptamer (*B*). The *insets* in *A* and *B* highlight the most probable binding sites for the IIa/IIa-14t complex. The density plots for the base-pairing probabilities and the corresponding (predicted) secondary structures for the domain IIa (*C*) and IIa-14t aptamer (*D*), with the most probable binding mode shown in *A* and *B*. The predicted secondary structures for the IIa/IIa-14t complex (*E*) and SL1/5-39 complex (*F*), respectively. In the complex structures, the thick lines denote the correctly predicted base pairs, the thick dashed lines denote the missed native base pairs, and the thin dotted lines denote false predictions. The base pairs in blue lines are noncanonical base pairs that are considered in our model (not included in the SE and PPV calculations).

A

Binding site 1

B
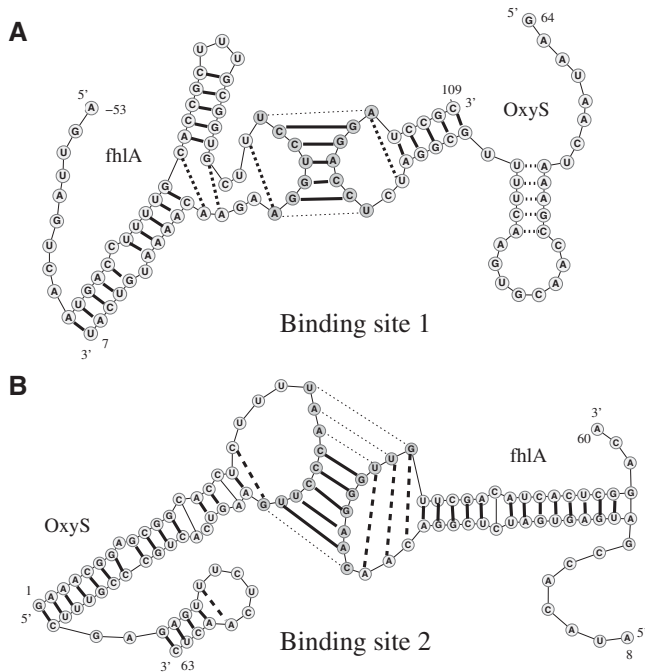
Binding site 2

FIGURE 5. The predicted secondary structures with the alternative binding sites of fhlA/OxyS complex at room temperature. (*A*) Binding site 1; (*B*) binding site 2. The legend for the lines of the different base pairs is the same as that used in Figure 4. Since our model can treat only one binding site for a complex structure, each of the two binding sites shown here was predicted separately.

4E). The predicted structure agrees well with the experimentally determined structure with (SE, PPV) equal to (0.96, 1.0). In a similar way, we predicted the structure of the SL1/5-39 complex (see Fig. 4F), which is also in good agreement with the experimentally determined structure (see Fig. 8A in Aldaz-Carroll et al. 2002) with (SE, PPV) equal to (0.90, 0.92).

fhlA/OxyS complex

OxyS can regulate the gene expression of bacterial *E. coli* by binding to a short sequence in gene fhlA. According to the experimentally proposed structure (Argaman and Altuvia 2000), two binding sites exist for the fhlA/OxyS complex. As described in *Theory and Model*, our theory assumes there is only one binding site. Nevertheless, the model correctly predicted the two binding sites (separately). Figure 5, A and B, shows the predicted structures for fhlA/OxyS complex. The predicted intramolecular and intermolecular base pairs are in good agreement with the experiment. Furthermore, we find that the number of the predicted intermolecular base pairs is less than that in the experimentally determined structure. We note that the intermolecular base pairs in the experimental structure is deduced according to the complementary base-pairings between fhlA and OxyS, which does not account for the spatial chain connectivity in the three-dimen-

sional structure. In contrast, our model accounts for the chain connectivity effect, which may contribute to the difference between our predicted structure and the experimental structure.

SL1/SL1 complex in HIV virus

According to experimental evidence, the full-length SL1 domain is involved in a dimerization state in vitro (Russell et al. 2004). The SL1 dimerization is important for HIV packaging. In addition, the stable structure of the SL1/SL1 complex is a linear dimer, which is different from the kissing complex determined by previous NMR measurement for the truncated short SL1 sequence (Laughrea and Jette 1994; Muriaux et al. 1996; Ennifar et al. 2001; Paillart et al. 2004; Ulyanov et al. 2006). Our previous theoretical study (Cao and Chen 2011) on the truncated SL1 sequence also showed that the linear dimer and the kissing complex coexist at room temperature. However, the previous model cannot be used to predict the complex structure of the whole SL1 domain. The present new model can predict the complex structure for the whole SL1 domain. Figure 6, A and B, shows the predicted binding sites at room temperature. Two different binding sites (I, II) exist for the SL1/SL1 complex (see Fig. 6). In Figure 6C, we show the base-pairing information of the two predicted structures. The two structures correspond to the linear dimer and the kissing complex, respectively. The predicted populational ratio of the structures with site I and site II is (5:1) at room temperature. The lowest free-energy state for the SL1/SL1 complex is the linear dimer structure. The predicted native structure is consistent with NMR
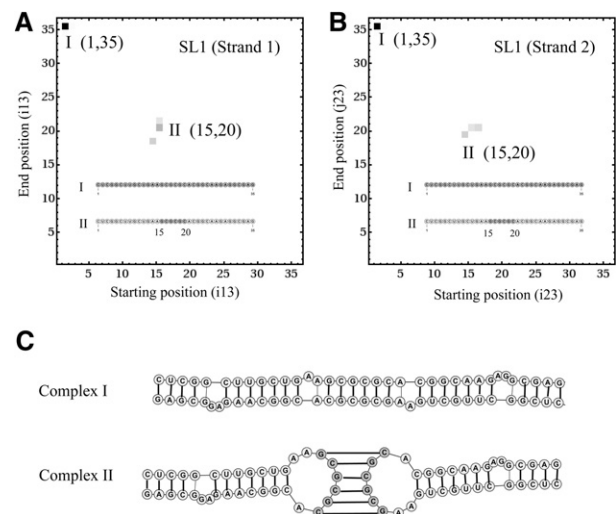
FIGURE 6. The predicted binding sites at room temperature for the SL1/SL1 complex in strand 1 (*A*) and strand 2 (*B*). The *insets* highlight the top two most probable binding sites. (*C*) The predicted secondary structures for the SL1/SL1 complex. Complex I is a linear dimer, while complex II is a kissing complex. The predicted ratio of stability for the complex I and complex II at room temperature is 5:1.

experimental data (Ulyanov et al. 2006), which indicates that linear dimer structure of SL1–SL1 complex is thermodynamically more stable than the kissing dimer form. Based on the partition function calculations, we found that the free-energy difference between the linear dimer and the kissing complex is ≈1 kcal/mol. The small free-energy difference suggests that the two structures can coexist in thermodynamic equilibrium and can possibly interconvert with the change of temperature and solution conditions (Weixlbaumer et al. 2004; Kim and Shapiro 2013).
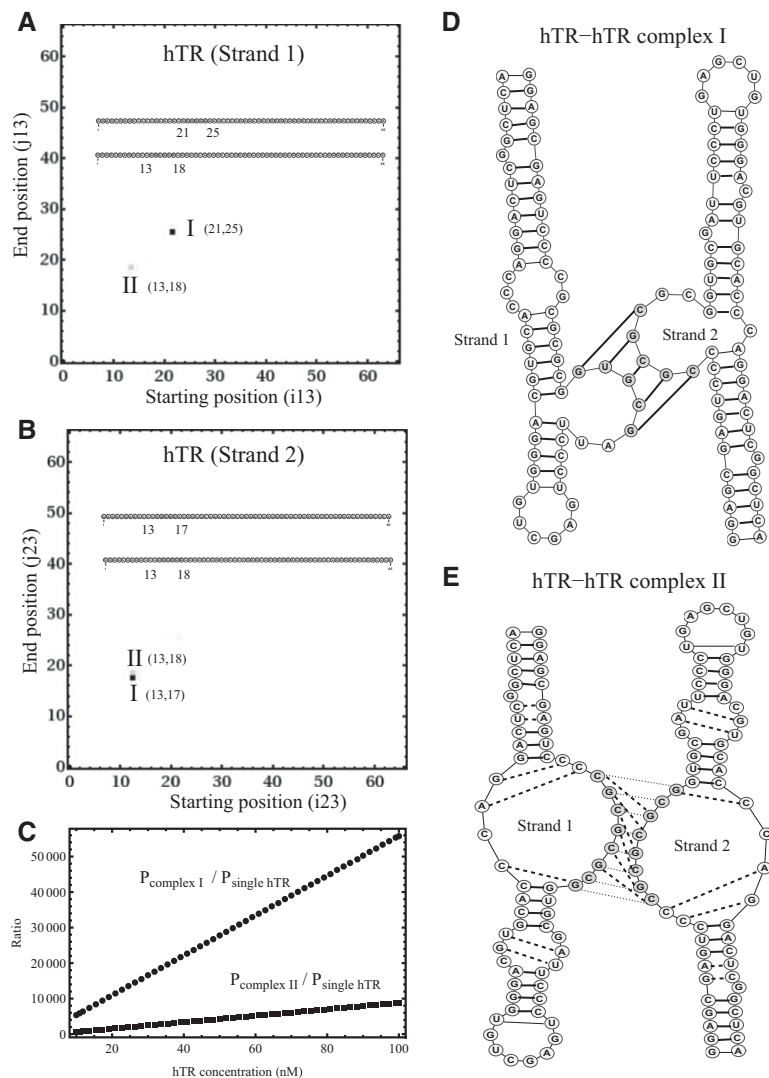
hTR/hTR complex

From the biochemical and functional analysis, Ren et al. (2003) found that hTR can form a dimer at the J7b/8a loop in domain hTR$_{380-444}$. Figure 7, A and B, shows the predicted binding sites. Similar to the SL1/SL1 complex, there are two binding sites for hTR/hTR complex. Our structure prediction showed that both binding modes correspond to kissing complexes (see Fig. 7D,E). According to the calculation, structure I is slightly more stable than structure II. Structure II is consistent with the structure proposed based on the experiment (Ren et al. 2003). For structure I, loop C393GCGC397 in strand 1 forms base pairs with loop G401UGCG405 of strand 2. We estimated the population of the kissing complex structures (I and II) in the cellular condition. According to the experiment (Ren et al. 2003), the hTR concentration is ≈10–100 nM in the cellular condition. Figure 7C shows the calculated ratios between complexes I and II and the single-stranded hTR at different hTR concentrations C_T. As we can see from Figure 7C, both ratios ([complex-I]/[single-hTR] and [complex-II]/[single-hTR]) are linearly dependent on the hTR concentration with slopes of 560/C_T (nM) and 91/C_T (nM), respectively. In the cellular condition, the population of complexes I and II are dominant over the population of monomeric hTR, suggesting that hTR is in the dimer state at the physiological condition. The prediction is consistent with the experimental hypothesis (Ren et al. 2003). However, quantitative validation of the predicted populational distribution shown in Figure 7 requires further thermodynamic measurements.

Folding thermodynamics

From the temperature dependence of the partition function $Q(T)$, we can compute the heat capacity melting curve $C(T)$ for a given sequence: $C(T) = \frac{\partial}{\partial T}[k_B T^2 \frac{\partial}{\partial T} ln Q(T)]$, where the total partition function Q is a sum over the unbound and bound systems:

$$Q(T) = Q_1 \cdot Q_2 + e^{-\Delta G_{associate}/k_B T} \cdot Q_{12}. \qquad (6)$$

Here, $\Delta G_{associate}$ is dependent on the RNA concentration C_T: $\Delta G_{associate} = \Delta G_{init} - k_B T$ ln $(C_T/4)$. We choose ΔG_{init} to be

FIGURE 7. The predicted binding positions for the hTR/hTR complex in strand 1 (*A*) and strand 2 (*B*) at room temperature. There exist two binding sites (I, II) with comparable stabilities. The predicted complex structures: complex I (*D*) and complex II (*E*), correspond to the two binding sites shown in *A* and *B*. Complex I is slightly more stable than complex II based on our calculation. However, complex II is the native-like complex structure as suggested by the experiment (Ren et al. 2003); see *E* with the same legend as that used in Figure 4. (*C*) The ratio between the population of complex I or II and the population of the single-stranded hTR at the different hTR concentrations (nM). The data show that the complexes are dominant for hTR concentration in the range [10 nM, 100 nM] (cellular condition) (Ren et al. 2003).

4.1 kcal/mol according to the experiment result (Serra and Turner 1995; Zuker 2003). In order for the two strands to form a complex, the binding free energy (affinity) must exceed the free-energy cost $\Delta G_{associate}$ for the association of the two strands. Increasing the temperature or decreasing the strand concentration would result in the dissociation of the complex.

To further validate the entropy parameters for the kissing motif (Cao and Chen 2011) and the two-step procedure in our model, we calculate the melting curves of XYMAL, an RNA hairpin kissing complex, at different RNA concentrations (Lorenz et al. 2006). To compare with the experimental results, we use the same solution condition as the experimental condition (1 M NaCl solution condition and 1–6 μM of RNA strand concentration) (Lorenz et al. 2006). The theoretical predictions (Fig. 8A) show two transitions. The first transition is ≈60°C and is concentration dependent, while the second transition is at a higher temperature and is concentration independent. Our calculation further indicates that the low- and high-temperature transitions correspond to the dissociation of the kissing complex and the unfolding of monomeric hairpins, respectively. The overall melting profiles for the different RNA concentration levels agree with the experimental results (shown in fig. 2C of Lorenz et al. 2006). Furthermore, the concentration dependence of melting temperatures (first peaks in Fig. 8A) also agree the experimental data (see Fig. 8B). The theory-experiment test suggests the validity of our entropy model for the kissing complex.

IIa/IIa-14t complex

Supplemental Figure S2, a through c, in the Supplemental Material shows the percentage of the IIa/IIa-14t complex at different temperatures and different strand concentrations. As the strand concentration increases, the population is dominated by the complex form. We find that the IIa/IIa-14t complex is quite stable for strand concentration ≥10 nM at 37°C. As the temperature increases, IIa/IIa-14t is destabilized. The IIa/IIa-14t complex is completely unfolded at 80°C for the strand concentration ranging from 1 nM to 1000 nM (see Supplemental Fig. S2c).

fhlA/OxyS complex

From the predicted structures, we find two stable kissing binding sites for the fhlA/OxyS complex (see Fig. 5). Supplemental Figure S2, d through f, in Supplemental Material shows the fractional population of the complex at the different temperatures and the different strand concentrations. The calculation shows that binding site 1 is more stable than binding site 2, which may be due to the fact that site 1 contains four G-C base pairs while site 2 has only three G-C base pairs. The calculation also shows that both site 1 and site 2 are completely unzipped at high temperature (80°C).

CONCLUSIONS

In summary, we have developed a new computational model that can predict general intermolecular loop–loop base-pairing between two RNAs. Tests against other models show that this new model can provide improved predictions for the structure and stability of RNA/RNA complexes. Moreover, the model can predict not only the global minimum free-energy structure but also the possible alternative structures (local minima on the free-energy landscape). Many biological functions are related to the suboptimal (alternative) structure, suggesting the importance of the viable alternative secondary structures. For example, we found two distinct binding interactions for SL1/SL1 complex in HIV, which correspond to the linear dimer and the kissing dimer. The linear dimer is a simple duplex structure, while the kissing dimer is a more complex structure. Moreover, the physics-based
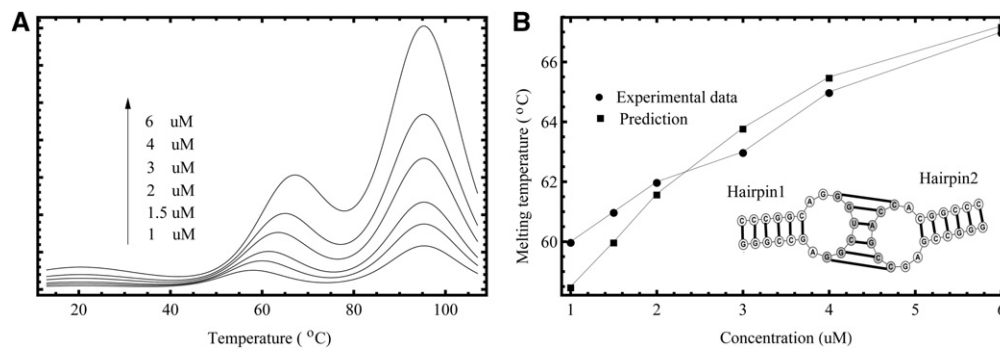
FIGURE 8. (*A*) The predicted melting curves for XYMAL, an RNA hairpin kissing complex (Lorenz et al. 2006), at different RNA concentrations. For each melting curve, the first peak (≈60°C) corresponds to dissociation of the kissing complex, and the second peak (≈95°C) corresponds to the unfolding of a monomeric hairpin. (*B*) Comparison between the theoretical prediction and the experimental data for the melting temperatures (first peak) for different RNA concentrations. The experimental results are from the study by Lorenz et al. (2006). (*Inset*) The 2d structure of XYMAL.

model enables us to predict the folding stability of the complicated RNA/RNA complexes at the different temperatures and strand concentrations.

The current model offers a theoretical framework for further systematic development of the method. For example, the current model assumes canonical loop–loop base-pairing. In many RNA/RNA complexes, loop–loop contacts are formed by noncanonical base pairs. With the proper thermodynamic parameters, we can extend the present framework of the theory to treat more complex structures with noncanonical loop–loop contacts. Furthermore, the model shows success for RNA/RNA complexes with a single loop–loop kissing site. Future development of the model should include structures with simultaneous multiple kissing interactions, such as those in the whole fhlA/OxyS complex (Argaman and Altuvia 2000). In addition, the current model can only treat the medium-size RNA/RNA complex. As shown in Figure 3, the computational time for the prediction of a complex between two 70-nt RNAs is ≈110 h. In order to treat a large RNA/RNA complex, such as those found in the systems of gene expression (Miller and White 2006; Busch et al. 2008; Chitsaz et al. 2009), we need to develop a computationally more efficient algorithm.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

REFERENCES

Aldaz-Carroll L, Tallet B, Dausse E, Yurchenko L, Toulme JJ. 2002. Apical loop–internal loop interactions: a new RNA–RNA recognition motif identified through in vitro selection against RNA hairpins of the hepatitis C virus mRNA. *Biochemistry* **41:** 5883–5893.

Alkan C, Karakoc E, Nadeau JH, Sahinalp SC, Zhang K. 2006. RNA–RNA interaction prediction and antisense RNA target search. *J Comput Biol* **13:** 267–282.

Andronescu M, Zhang ZC, Condon A. 2005. Secondary structure prediction of interacting RNA molecules. *J Mol Biol* **345:** 987–1001.

Andronescu MS, Pop C, Condon AE. 2010. Improved free energy parameters for RNA pseudoknotted secondary structure prediction. *RNA* **16:** 26–42.

Argaman L, Altuvia S. 2000. fhlA repression by OxyS RNA: Kissing complex formation at two sites results in a stable antisense-target RNA complex. *J Mol Biol* **300:** 1101–1112.

Arnott S, Hukins DW, Dover SD. 1972. Optimised parameters for RNA double-helices. *Biochem Biophys Res Commun* **48:** 1392–1399.

Bon M, Orland H. 2011. TT2NE: a novel algorithm to predict RNA secondary structures with pseudoknots. *Nucleic Acids Res* **39:** e93.

Bossi L, Figueroa-Bossi N. 2007. A small RNA downregulates LamB maltoporin in *Salmonella*. *Mol Microbiol* **65:** 799–810.

Busch A, Richter AS, Backofen R. 2008. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics* **24:** 2849–2856.

Cao S, Chen S-J. 2005. Predicting RNA folding thermodynamics with a reduced chain representation model. *RNA* **11:** 1884–1897.

Cao S, Chen S-J. 2006a. Free energy landscapes of RNA/RNA complexes: with applications to snRNA complexes in spliceosomes. *J Mol Biol* **357:** 292–312.

Cao S, Chen S-J. 2006b. Predicting RNA pseudoknot folding thermodynamics. *Nucleic Acids Res* **34:** 2634–2652.

Cao S, Chen S-J. 2009. Predicting structures and stabilities for H-type pseudoknots with interhelix loops. *RNA* **15:** 696–706.

Cao S, Chen S-J. 2011. Structure and stability of RNA/RNA kissing complex: with application to HIV dimerization initiation signal. *RNA* **17:** 2130–2143.

Chen AA, Garcia AE. 2013. High-resolution reversible folding of hyperstable RNA tetraloops using molecular dynamics simulations. *Proc Natl Acad Sci* **110:** 16820–16825.

Chen HL, Condon A, Jabbari H. 2009. An $O(n^5)$ algorithm for MFE prediction of kissing hairpins and four-chains in nucleic acids. *J Comput Biol* **16:** 803–815.

Chitsaz H, Salari R, Sahinalp SC, Backofen R. 2009. A partition function algorithm for interacting nucleic acid strands. *Bioinformatics* **25:** i365–i373.

Da Rocha Gomes S, Dausse E, Toulme JJ. 2004. Determinants of apical loop-internal loop RNA–RNA interactions involving the HCV IRES. *Biochem Biophys Res Commun* **322:** 820–826.

Dimitrov RA, Zuker M. 2004. Prediction of hybridization and melting for double-stranded nucleic acids. *Biophys J* **87:** 215–226.

Dirks RM, Bois JS, Schaeffer JM, Winfree E, Pierce NA. 2007. Thermodynamic analysis of interacting nucleic acid strands. *SIAM Rev* **49:** 65–88.

Eisinger J. 1971. Complex formation between transfer RNA's with complementary anticodons. *Biochem Biophys Res Commun* **43:** 854–861.

Ennifar E, Walter P, Ehresmann B, Ehresmann C, Dumas P. 2001. Crystal structures of coaxially stacked kissing complexes of the HIV-1 RNA dimerization initiation site. *Nat Struct Biol* **8:** 1064–1068.

Flory PJ. 1969. *Statistical mechanics of chain molecules.* Wiley, New York.

Huang FWD, Qin J, Reidys CM, Stadler PF. 2009. Partition function and base pairing probabilities for RNA–RNA interaction prediction. *Bioinformatics* **25:** 2646–2654.

Jossinet F, Paillart JC, Westhof E, Hermann T, Skripkin E, Lodmell JS, Ehresmann C, Ehresmann B, Marquet R. 1999. Dimerization of HIV-1 genomic RNA of subtypes A and B: RNA loop structure and magnesium binding. *RNA* **5:** 1222–1234.

Kim T, Shapiro BA. 2013. The role of salt concentration and magnesium binding in HIV-1 subtype-A and subtype-B kissing loop monomer structures. *J Biomol Struct Dyn* **31:** 495–510.

Kolb FA, Malmgren C, Westhof E, Ehresmann C, Ehresmann B, Wagner EG, Romby P. 2000a. An unusual structure formed by antisense-target RNA binding involves an extended kissing complex with a four-way junction and a side-by-side helical alignment. *RNA* **6:** 311–324.

Kolb FA, Engdahl HM, Slagter-Jager JG, Ehresmann B, Ehresmann C, Westhof E, Wagner EG, Romby P. 2000b. Progression of a loop–loop complex to a four-way junction is crucial for the activity of a regulatory antisense RNA. *EMBO J* **19:** 5905–5915.

Kuhrova P, Banas P, Best RB, Sponer J, Otyepka M. 2013. Computer folding of RNA tetraloops? Are we there yet? *J Chem Theory Comput* **9:** 2115–2125.

Laughrea M, Jette L. 1994. A 19-nucleotide sequence upstream of the 5′ major splice donor is part of the dimerization domain of human immunodeficiency virus 1 genomic RNA. *Biochemistry* **33:** 13464–13474.

Lee AJ, Crothers DM. 1998. The solution structure of an RNA loop–loop complex: the ColE1 inverted loop sequence. *Structure* **6:** 993–1005.

Long D, Chan CY, Ding Y. 2008. Analysis of microRNA–target interactions by a target structure based hybridization model. *Pac Symp Biocomput* **2008:** 64–74.

Lorenz C, Piganeau N, Schroeder R. 2006. Stabilities of HIV-1 DIS type RNA loop–loop interactions *in vitro* and *in vivo*. *Nucleic Acids Res* **34:** 334–342.

Ly H, Xu L, Rivera MA, Parslow TG, Blackburn EH. 2003. A role for a novel "*trans*-pseudoknot" RNA–RNA interaction in the functional dimerization of human telomerase. *Genes Dev* **17:** 1078–1083.

Mathews DH, Burkard ME, Freier SM, Wyatt JR, Turner DH. 1999. Predicting oligonucleotide affinity to nucleic acid targets. *RNA* **5:** 1458–1469.

Miller WA, White KA. 2006. Long-distance RNA-RNA interactions in plant virus gene expression and replication. *Annu Rev Phytopathol* **44:** 447–467.

Muckstein U, Tafer H, Hackermuller J, Bernhart SH, Stadler PF, Hofacker IL. 2006. Thermodynamics of RNA–RNA binding. *Bioinformatics* **22:** 1177–1182.

Muriaux D, Fosse P, Paoletti J. 1996. A kissing complex together with a stable dimer is involved in the HIV-1$_{Lai}$ RNA dimerization process *in vitro*. *Biochemistry* **35:** 5075–5082.

Paillart JC, Shehu-Xhilaga M, Marquet R, Mak J. 2004. Dimerization of retroviral RNA genomes: an inseparable pair. *Nat Rev Microbiol* **2:** 461–472.

Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R. 2004. Fast and effective prediction of microRNA/target duplexes. *RNA* **10:** 1507–1517.

Ren X, Gavory G, Li H, Ying L, Klenerman D, Balasubramanian S. 2003. Identification of a new RNA–RNA interaction site for human telomerase RNA (hTR): structural implications for hTR accumulation and a dyskeratosis congenita point mutation. *Nucleic Acids Res* **31:** 6509–6515.

Repoila F, Majdalani N, Gottesman S. 2003. Small non-coding RNAs, co-ordinators of adaptation processes in *Escherichia coli*: the RpoS paradigm. *Mol Microbiol* **48:** 855–861.

Romero-López C, Berzal-Herranz A. 2009. A long-range RNA–RNA interaction between the 5 and 3 ends of the HCV genome. *RNA* **15:** 1740–1752.

Russell RS, Liang C, Wainberg MA. 2004. Is HIV-1 RNA dimerization a prerequisite for packaging? Yes, no, probably? *Retrovirology* **1:** 23.

Schafer H, Daura X, Mark AE, Gunsteren WF. 2001. Entropy calculations on a reversibly folding peptide: Changes in solute free energy cannot explain folding behavior. *Proteins* **43:** 45–56.

Schmidt M, Zheng P, Delihas N. 1995. Secondary structures of *Escherichia coli* antisense micF RNA, the 5′-end of the target ompF mRNA, and the RNA/RNA duplex. *Biochemistry* **34:** 3621–3631.

Serra MJ, Turner DH. 1995. Predicting thermodynamic properties of RNA. *Methods Enzymol* **259:** 242–261.

Sperschneider J, Datta A. 2010. DotKnot: pseudoknot prediction using the probability dot plot under a refined energy model. *Nucleic Acids Res* **38:** e103.

Sperschneider J, Datta A, Wise MJ. 2011. Heuristic RNA pseudoknot prediction including intramolecular kissing hairpins. *RNA* **17:** 27–38.

Stephenson W, Asare-Okai PN, Chen AA, Keller S, Santiago R, Tenenbaum SA, Garcia AE, Fabris D, Li PT. 2013. The essential role of stacking adenines in a two-base-pair RNA kissing complex. *J Am Chem Soc* **135:** 5602–5611.

Theimer CA, Feigon J. 2006. Structure and function of telomerase RNA. *Curr Opin Struct Biol* **16:** 307–318.

Ulyanov NB, Mujeeb A, Du Z, Tonelli M, Parslow TG, James TL. 2006. NMR structure of the full-length linear dimer of stem-loop-1 RNA in the HIV-1 dimer initiation site. *J Biol Chem* **281:** 16168–16177.

VanLoock MS, Easterwood TR, Harvey SC. 1999. Major groove binding of the tRNA/mRNA complex to the 16 S ribosomal RNA decoding site. *J Mol Biol* **285:** 2069–2078.

Vogel J, Wagner EG. 2007. Target identification of small noncoding RNAs in bacteria. *Curr Opin Microbiol* **10:** 262–270.

Weixlbaumer A, Werner A, Flamm C, Westhof E, Schroeder R. 2004. Determination of thermodynamic parameters for HIV DIS type loop–loop kissing complexes. *Nucleic Acids Res* **32:** 5126–5133.

Zhang WB, Chen S-J. 2001. A three-dimensional statistical mechanical model of folding double-stranded chain molecules. *J Chem Phys* **114:** 7669–7681.

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **6:** 3406–3415.