

A domain-based model for predicting large and complex pseudoknotted structures

Song Cao and Shi-Jie Chen*

Department of Physics and Department of Biochemistry University of Missouri; Columbia, MO USA

Keywords: human telomerase RNA (hTR), human immunodeficiency virus (HIV), internal ribosome entry site (IRES), hepatitis delta virus (HDV), Pseudoknots, structural predictions, large RNAs

Pseudoknotted structures play important structural and functional roles in RNA cellular functions at the level of transcription, splicing and translation. However, the problem of computational prediction for large pseudoknotted folds remains. Here we develop a domain-based method for predicting complex and large pseudoknotted structures from RNA sequences. The model is based on the observation that large RNAs can be separated into different structural domains. The basic idea is to first identify the domains and then predict the structures for each domain. Assembly of the domain structures gives the full structure. The use of the domain-based approach leads to a reduction of computational time by a factor of about $\sim N^2$ for an N-nt sequence. As applications of the model, we predict structures for a variety of RNA systems, such as regions in human telomerase RNA (hTR), internal ribosome entry site (IRES) and HIV genome. The lengths of these sequences range from 200-nt to 400-nt. The results show good agreements with the experiments.

Introduction

Pseudoknot is an important RNA motif that frequently occurs in many biologically significant RNAs, including viral RNAs, such as influenza virus,¹ hepatitis delta virus (HDV),²⁻⁵ turnip yellow mosaic virus (TYMV),⁶ tobacco mosaic virus (TMV),⁷ beet western yellow virus (BWYV),^{8,9} and the intergenic region (IGR) of cricket paralysis virus-like (CrPV-like) internal ribosome entry site (IRES).¹⁰ Moreover, pseudoknots are found in most large RNAs, such as rRNAs (rRNA).^{11,12} These pseudoknots often involve long-range tertiary interactions between distant domains of RNA. Pseudoknots play a variety of structural and functional roles that are essential in biological processes.^{13,14} For example, a pseudoknot is found to be a vital component for promoting efficient frameshifting.¹⁴⁻¹⁷ Disruption of the pseudoknot structure is found to decrease the frameshifting efficiency. Other examples include the essential functional roles of pseudoknotted structures in CR1/CR2/CR3 domain in telomerase activity¹⁸ and IRES. In the ribosome-binding domain of IRES from a Dicistroviridae intergenic region,¹⁹ biochemical studies revealed a well-conserved pseudoknot structure which is functionally important.

Several computational methods, including the heuristic methods,²⁰⁻²² the dynamic programming algorithms,²³⁻²⁵ the simulational methods²⁶⁻²⁸ and the integer programming method,²⁹ have led to many useful predictions for RNA secondary structures including pseudoknots. However, the currently available dynamic programming algorithms based on the physical conformational entropy parameters³⁰⁻³² can only treat pseudoknots of limited chain length, primarily due to the rapidly increasing computational

time for the conformational enumeration/sampling. Heuristic methods are computationally efficient. However, they do not guarantee that the predicted structure has the lowest free energy. Moreover, the heuristic method is often based on simplified approximations for the loop entropies in a pseudoknot,²² which may cause inaccuracy.

Many of the current pseudoknot folding algorithms are focused on simple H-type pseudoknots.⁴⁶ A canonical H-type pseudoknot consists of two stems and two loops, where each loop spans across a helix stem. With the extensive computational studies in recent years,³⁰⁻³⁸ prediction of simple H-type pseudoknot is becoming more and more reliable. For example, a recently developed virtual bond-based RNA folding model (called “*Vfold*” model) allows direct calculation of pseudoknot loop entropy parameters and can predict H-type pseudoknots with an accuracy around 90%.³² Inclusion of the entropy parameters in the *Vfold* model distinguishes it from the other existing computational models.^{27,31,39-45} For large pseudoknotted RNAs that go beyond the H-type pseudoknots, however, the structure prediction remains a challenge. One of the bottleneck is lacking loop entropy parameters for large and complex pseudoknots.^{34,35}

In order to predict the large and complex pseudoknot structure, in this study, we employ the virtual bond model (*Vfold*) to calculate and tabulate loop entropy parameters with the different structural contexts. To treat long RNA chains, we develop a domain-based method to decompose the large RNA structure into weakly coupled structural domains. Compared with the other domain-based algorithms for the secondary structures,^{47,48} the algorithm reported here has the advantage of being able to

*Correspondence to: Shi-Jie Chen; Email: chenshi@missouri.edu
Submitted: 04/26/11; Revised: 10/14/11; Accepted: 10/21/11
<http://dx.doi.org/10.4161/rna.92.18488>

treat pseudoknot structures. For each domain, which has a shorter chain length, we can apply our *Vfold* model to predict the stable structures. The use of the domain-based method causes a reduction of the computational time by a factor of about N^2 for an N -mer chain. As an application of the domain-based model, we predict the stable structures for the 210-nt catalytic domain of human telomerase RNA, the 192-nt IGR of CrPV-like IRES and several functional regions of the HIV genome. The results show a good theory-experiment agreement. We aim to predict all the stable as well as the alternative structures. For example, as we will show, for the hTR domain, we found that a long hairpin-like structure has the comparable thermodynamic stability as the hTR pseudoknot. In the predicted hairpin-like structure, we find nucleotides in single-stranded region that are used by telomerase reverse transcriptase (TERT) to synthesize telomere. We note that a previous experiment⁴⁹ suggested a hairpin structure that coexists with the pseudoknot structure. Our results echo the experimental finding.

Results

Domain-based approach to structure prediction. Our entropy parameters give the free energy for a given structure. However, the calculation of a partition function involves a vast ensemble of graphs, and, for an N -nt chain, the structural calculation from the base pairing probability distribution requires the computation for $N(N-1)/2$ partition functions Q_{ij} (see Eq. 1 in Materials and Methods). The *Vfold* model can predict the structure for the pseudoknotted structure with sequence length ~ 100 nts. For longer sequence, the method becomes computationally limited. In the following, we develop a domain-based partition function theory in order to efficiently tackle larger and more complex RNA structures. Our basic strategy is to factorize the large complex structures into domains that are computationally more tractable. To illustrate the method, we use a graph to represent a structure where the horizontal straight line in the graph represents the backbone chain and a curved link is added between two pairing nucleotides; see Figure 1. According to the presence/absence of cross-linked base pairs (curved links) in the graph, we classify two types of structural domains: pseudoknotted (C_{PK} in Fig. 1) and non-pseudoknot (C_S in Fig. 1) domains. The partition function for the non-pseudoknot domains can be calculated using the computationally efficient dynamic programming algorithm while the calculation for the pseudoknotted domains is

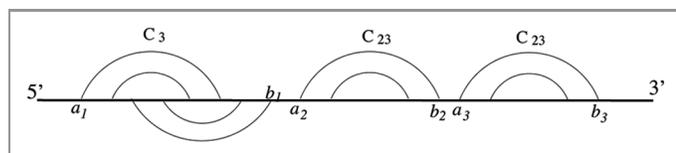


Figure 1. An RNA pseudoknot structure contains several smaller domains such as $C_{PK}(a_1, b_1)$ and $C_S(a_2, b_2)$, where C_S represents a structure in which boundary nucleotides a_2 and b_2 form base pair with each other and $C_{PK}(a_1, b_1)$ is an H-type pseudoknot structure, in which a_1 and b_1 form base pairs with other nucleotides.

computationally demanding. The essence of our approach is to recursively reduce a structure into domains until irreducible domains, namely, pseudoknotted domains, are reached.

First, we use partition functions to identify the domain boundaries such as (a_1, b_1) , (a_2, b_2) and (a_3, b_3) in Figure 1.

(1) We first identify the 5' boundary of each domain. For each nucleotide a_i ($1 \leq a_i \leq N$), we consider the ensemble (= the macrostate z in Eq. 1) of all the structures with a_i as the 5' boundary of a domain. For each nucleotide a_i , we use Eq. One to calculate the probability for a_i to be the 5' boundary nucleotide of a domain. The a_i 's of the large probabilities are identified as the 5' boundary nucleotides.

(2) Similarly, for each 5' boundary nucleotide found above, we can identify the 3' boundary nucleotide. Specifically, for each 5' boundary nucleotide a_i identified above, we consider the ensemble (= the macrostate z in Eq. 1) of all the structures that contain a domain "p" bound by the nucleotides a_i and b_i ($> a_i$). For each nucleotide b_i , we use Eq. One to calculate the probability for the formation of a domain bound by (a_i, b_i) . The (a_i, b_i) pairs with large probabilities are identified as the domain boundaries.

Applying the above method to hTR (Fig. 2), we found G17 and G209 as the 5' and 3' boundaries, respectively (see Fig. 2A). In the above computation, we distinguish the pseudoknotted and the non-pseudoknot domains. Using the recursive algorithm,³⁰ we can identify all the 5' boundaries in the efficient (one-time) partition function calculation for the chain. We include the pseudoknot structures in the calculation of the partition function. For M identified 5' boundaries, we need to perform M times partition function calculations to find all the 3' boundaries. Usually M is much smaller than the sequence length N . Therefore, this approach can effectively reduce the original $N(N-1)/2$ partition function calculations into a much smaller number ($\sim M$) of calculations.

Second, we recursively reduce the domains. If the sequence gives a continuous train of Watson-Crick base pairs (= a helix) that starts from the boundary base pair (see the continuous helix from base pair G17-G209 to U32-U194 in Fig. 2), we identify the helix and remove it. This would result in a shorter chain which goes from nucleotides 33 to 193 (Fig. 2B). We further apply the above domain-reduction procedure to this shorter chain. Continuous iteration of the above procedure will lead to shorter and shorter chains (Fig. 2C). The domain reduction process stops when we reach a pseudoknotted domain, which is irreducible, or the domain size is small enough (less than 130 nt) for an efficient calculation using the *Vfold* model. For hTR in Figure 2, the irreducible domain is bound by nucleotides 63 and 184 (see Fig. 2D).

Next, for the irreducible domains, using the entropy parameters and partition function theory (see Eq. 1), we predict the structures from the base pairing probability; see Figure 2D for the predicted structure of the pseudoknotted domain.

The domain-based methodology developed here has three advantages over other pseudoknot models.^{20,22-25} First, for a long sequence, the model can explicitly account for the entropy parameters for the formation of the pseudoknots. Second, the

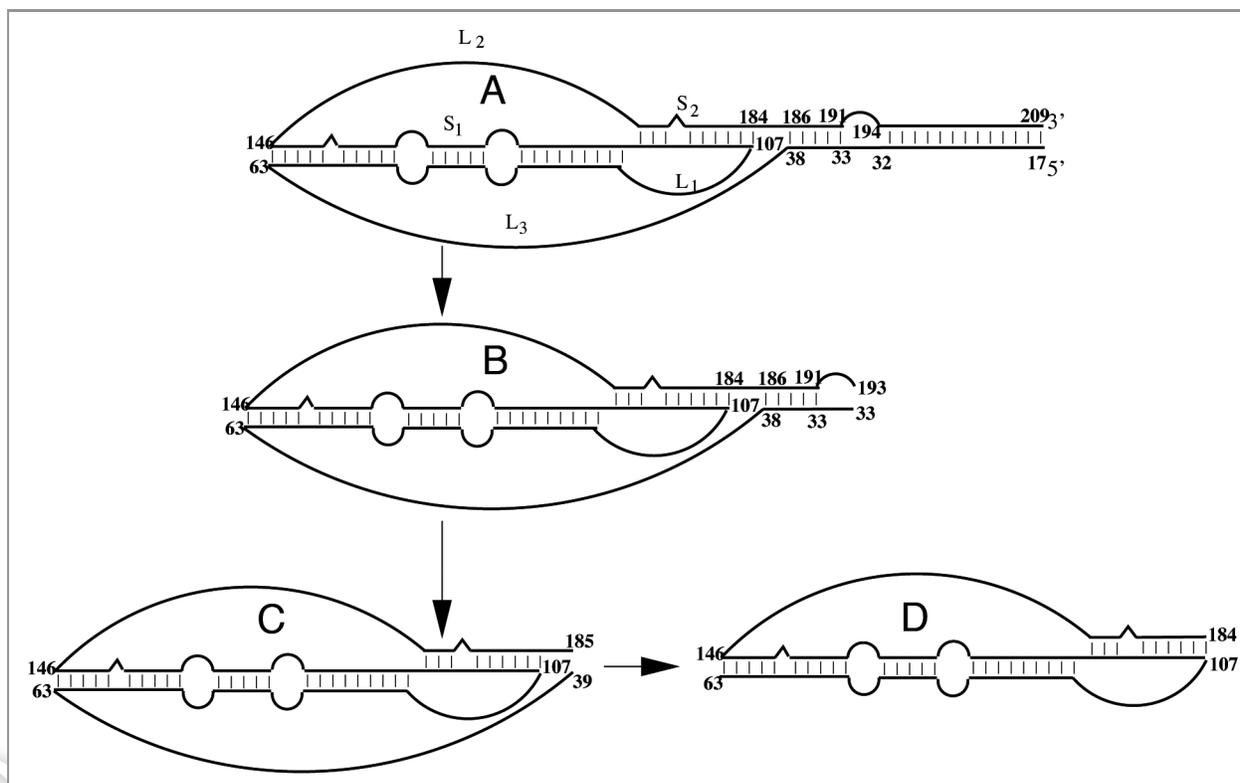


Figure 2. A schematic figure for the domain-reduction method to predict the stable structure of an 210-nt hTR domain. (A) In the first step, using the partition algorithm we find the boundary nucleotides for the domains; see $(a_1, b_1), (a_2, b_2)$ and (a_3, b_3) in Figure 1. Through this step, we find a domain closed by the base pair (G17, G209). (B) In the second step, we find the continuous helix from base pair G17-G209 to U32-U194 and we truncate the sequence by deleting the helix from G17-G209 to U32-U194. We repeat the domain-reduction process to identify the pseudoknot or closed secondary structure domain for the truncated sequence. We identify a domain closed by base pair (G33, C191). We also find another continuous helix from G33-C191 to U38-C186. (C) We truncated the sequence by deleting this helix. (D) We continue the process until we reach a pseudoknotted domain or a domain of less than 130 nts. For example, we find the short pseudoknot domain with 122 nts from G63 to U184. In the last step, we use our free energy/partition function model to predict the detailed structure for the final minimum domain. Assembly of the structure domains gives the stable structure for the 210-nt hTR domain.

model is based on the complete conformation ensemble. Third, based on base pairing probabilities, the model enables prediction for all the stable and metastable states. This is important for large RNAs which can fold into multiple alternative structures.

Figure S1 shows the workflowchart for predicting the structure from the sequence. For a given sequence, if the length is < 100 nt, we run the Vfold model directly, otherwise, we run the domain-based algorithm reported in this article. In the domain-based method, for each irreducible pseudoknot domain, we run the Vfold model to predict the structure.

Comparison with other models. To measure the accuracy of the predicted structures, we introduce the sensitivity (SE) and the positive predictive value (PPV) parameters, which are defined as following:

$$SE = \frac{\text{Number of correctly predicted base pairs}}{\text{Total number of the base pairs in the experimentally determined structure}}$$

$$PPV = \frac{\text{Number of correctly predicted base pairs}}{\text{Total number of the base pairs in the predicted structure}}$$

(SE, PPV) = (1, 1) and (0, 0) correspond to the perfectly accurate and the completely failed predictions, respectively. A low SE and/or PPV parameter indicates a high rate of false negative (positive) in the prediction. In Table 1, we tested the model accuracy on nine different pseudoknot structures, such as VMV,⁵⁰ SARS-CoV,⁵¹ HDV,⁵² HDV-anti,⁵³ TMV.L,⁵⁴ TYMV,⁵⁵ IGR IRES,¹⁰ Tetrahymena telomerase⁵⁶ and Human telomerase RNAs,⁵⁷ and compared the results with six of the existing state of the art models, i.e., HotKnots,²² ProbKnot,³⁷ pknotsRE,²³ STAR,²⁶ pknots-RG²⁵ and NUPACK.²⁴ Our model shows significant improvements with (SE, PPV) = (0.92, 0.88). We note that the present model also gives good predictions for non-pseudoknotted structures. For example, Table S1 of reference 58 gives a (SE, PPV) value close to (1.0, 1.0) for 12 hairpin structures.

Genomic and anti-genomic HDV pseudoknot. Both genomic and antigenomic RNAs of HDV contain a self-cleavage domain with sequence length about 90 nt. Biochemical and structural studies²⁻⁵ indicate that the domain folds into a pseudoknotted structure. We apply the above free energy model and the

Table 1. The sensitivity SE (upper) and positive predictive value PPV (lower) values for seven different pseudoknot models. We used nine complex pseudoknot structures such as VMV, SARS-CoV, HDV, anti-HDV, TMV-L and TYMV, IGR IRES, Tetrahymena telomerase and Human telomerase as the test sequences. Overall the *Vfold* model gives better predictions for the structures of these sequences

Sequence ID	length	<i>Vfold</i>	HotKnots	ProbKnot	pknotsRE	STAR	pknots-RG	NUPACK
VMV	67	1.0	0.5	0.5	1.0	0.5	0.5	1.0
SARS-CoV	71	0.96	0.92	0.69	1.0	0.73	0.92	0.85
HDV	87	0.9	0.4	0.43	0.46	0.6	0.96	0.63
HDV-anti	91	1	0.16	0.29	0.41	0.62	0.16	0.41
TMV.L	84	0.96	0.52	0.56	0.52	0.64	0.8	0.52
TYMV	86	0.96	0.72	0.88	0.72	0.88	0.76	0.44
IGR IRES	192	0.81	0.74	0.66	0.50	0.47	0.59	failed to run
Tetrahymena telomerase	159	0.68	0.61	0.66	0.45	0.39	0.66	failed to run
Human telomerase	210	0.98	0.70	0.76	0.48	0.48	0.54	failed to run
Average		0.92	0.58	0.60	0.62	0.59	0.65	
VMV	67	1.0	0.41	0.41	0.61	0.41	0.41	0.64
SARS-CoV	71	1.0	0.86	0.78	1.0	0.9	0.92	0.96
HDV	87	0.93	0.44	0.42	0.46	0.7	0.93	0.61
HDV-anti	91	0.78	0.14	0.21	0.31	0.6	0.14	0.32
TMV.L	84	0.96	0.61	0.78	0.59	0.69	0.83	0.61
TYMV	86	1	0.78	0.88	0.78	0.88	0.79	0.5
IGR IRES	192	0.76	0.73	0.63	0.67	0.51	0.71	failed to run
Tetrahymena telomerase	159	0.60	0.53	0.50	0.37	0.45	0.57	failed to run
Human telomerase	210	0.87	0.55	0.60	0.32	0.38	0.42	failed to run
Average		0.88	0.56	0.58	0.57	0.61	0.64	

domain-based method to predict the native structure for HDV pseudoknot. The theory gives satisfactory predictions with (SE, PPV) = (0.9, 0.93) for genomic HDV and (SE, PPV) = (1.0, 0.78) for the anti-genomic HDV. The predicted structure (Fig. 3) is nearly identical to the experimentally determined structure³ except that the P2 stem is slightly longer in the predicted structure. In the calculation, we have used our calculated parameters for the loop entropy⁵⁹ and the empirical free energy parameters for base stacking (in the helices) and the coaxially stacking.^{60,61} The model gives improved predictions as compared with other models (see Table 1). Since the different models are based on the same free energy parameters (for the helices), we attribute the improvement in the structure prediction to the use of more accurate entropies (for the loops).

TYMV and TMV pseudoknot. The 3' end of Turnip yellow mosaic virus (TYMV) RNA forms a tRNA-like structure. The pseudoknotted structure in the 3' end is vital for viral transcription.⁶ For a 86-nt truncated TYMV sequence,²² our model predicts the native structure with (SE, PPV) equal to (0.96, 1.0). Here we allow coaxial stacking between domain I and domain II (as shown in the NMR measurement).⁶² The predicted structure (Fig. S2) nearly exactly agrees with the experimentally measured structure.⁶ Similarly, the model predicts the native structure for tobacco mosaic virus-tomato strain (TMV-L) RNA (see Table 1)⁷ with (SE, PPV) is equal to (0.96, 0.92).

CR1/CR2/CR3 domain of the wild-type hTR. According to the experimental result,⁵⁷ there exists a pseudoknot structure in the CR1/CR2/CR3 domain, which has a >20-bp long stem. The large end-end distance of the stem requires a long loop to connect the two ends of the stems. Conformational enumeration for loop entropy calculation is only viable for stem length less than 12 bps.³⁰ Thus, we develop a method to estimate the loop entropy across a long helix stem (see the **Supplementary Material**).

Previous studies on the functional role of hTR suggest the possible presence of a functionally important molecular switch between a hairpin-like structure and the hTR pseudoknot.^{63,64} These experiments are based on pseudoknot-forming small portion of the whole domain. The whole 210-nt pseudoknot domain of hTR is much more complex than a simple H-type pseudoknot.⁶⁵ The free energy model developed in the present study allows us to predict the structures for the whole domain.

NMR and thermodynamic experiments show the presence of multiple loop-stem tertiary interactions in hTR pseudoknot.⁵⁶ To treat loop-stem interactions, we previously developed a model to fit the empirical loop-stem base triples from thermodynamic data.⁵⁸ The parameters were derived from simple H-type pseudoknots. It is not clear whether the parameters would be applicable to more complex structures such as the whole hTR pseudoknot domain. For the present large domain structure, to simplify the calculation, we assume a total of 4 kcal/mol free

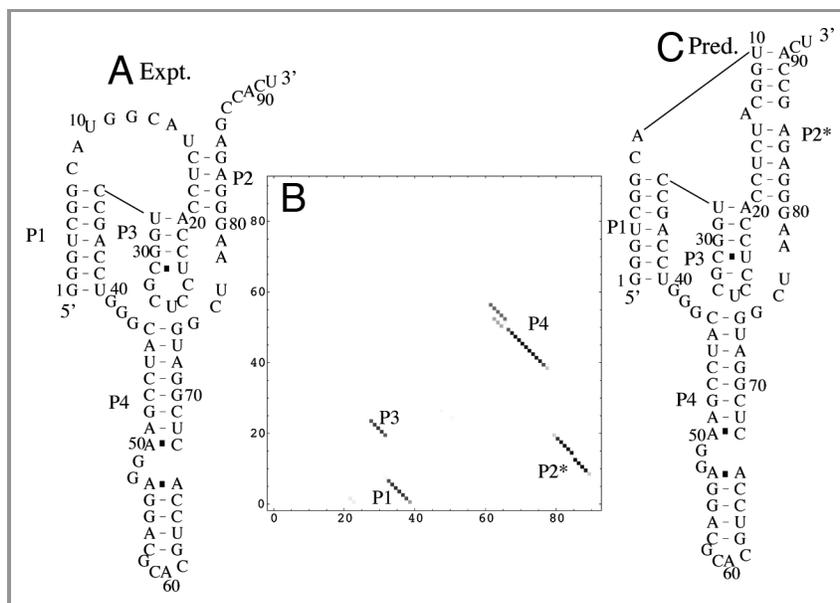


Figure 3. The predicted structure at $t = 37^{\circ}\text{C}$ for antigenomic HDV ribozymes. Our *Vfold* model gives a much more accurate prediction with (SE, PPV) equal to (1.00, 0.78) than other models.

energy contribution from the tertiary interactions. The NMR study shows three A·(A-U) base triples between loop 1 and stem 2 in hTR pseudoknot.¹⁸ Our previous calculation⁵⁸ shows that a single U·(A-U) base triple contributes about 1.3 kcal/mol to the free energy at room temperature. Thus, the estimated 4 kcal/mol is close to the energy contribution from three A·(A-U) base triples. Our theory predicts four stable structures I, II, III and IV (see Fig. 4). The structures have similar thermodynamic stabilities. In the following, using structure I as an illustrative example, we show how the structures are predicted.

In the first step, using the domain-reduction method (see Materials and Methods), we predict a structural domain from G17 to G209, as denoted by the closing base pair [17, 209] (see Figs. 2, 4 and 5). Our calculation shows that the probability for the formation of the [17, 209] domain is 0.94. We then remove the continuous helix from base pair G17-G209 to G31-U195 (see Fig. 4A). For the resultant truncated sequence from nucleotides 32 to 194, the domain-reduction procedure leads to four alternative domain structures (see Fig. 5 and I, II, III and IV in Fig. 4). Structures III and IV each contains two sub-domains closed by base pairs [33, 180] and [181, 191] for III and by [35, 179] and [181, 191] for IV. Structures II, III and IV (see Fig. 4) all contain a stable domain [56, 151] (see Fig. S3b for the detailed structural information for the domain). The high similarity between structures II, III and IV suggests that we may classify them into one (stem-loop) structure. Structure I contains a pseudoknotted domain (denoted as PK in Fig. 5) bound by nucleotides 63 and 184 (see Fig. S3a). The predicted structure I agrees exactly with the structure derived from phylogenetic analysis.^{56,57,65}

Our predicted structural distribution (Figs. 4 and 5) for the whole pseudoknot domain of hTR shows coexisting pseudoknot (Fig. 4-I) and stem-loop structure (Fig. 4-II, III and IV). We note that the long stem-loop structure has been detected in a

previous biochemical analysis.⁴⁹ It is important to note that the (highly conserved) template sequence (see Fig. 4) in the predicted pseudoknotted and stem-loop structures is single-stranded and thus is directly accessible for the synthesis of telomere with telomerase reverse transcriptase (TERT).

Our previous kinetic study⁶⁶ for the folding kinetics of the small truncated pseudoknot showed that the hairpin structure is kinetically accessible. Our present study on the whole domain further shows the presence of a (thermodynamically stable) stem-loop structure. The result would facilitate our understanding of the functional relevance of the conformational switch⁶³ between stem-loop structure and pseudoknot in telomerase activity.

Structure of IGR IRES. An internal ribosome entry site (IRES) is a sequence domain inside a mRNA, which can recruit the ribosomes for the protein synthesis.^{67,68} Computational prediction of the IRES tertiary structure is limited by the inability to treat the pseudoknotted structure.^{69,70} With the domain-based method, we can now predict the stable structure for IGR IRES. We applied the domain-reduction method and found three domains bound by nucleotides [1, 70], [72, 143] and [147, 190], respectively (see Fig. 6A). Further domain-reduction calculations for the [72, 143] give two sub-domains [81, 94] and [97, 137] (see Fig. 6B). The theoretical predictions show good agreement with the experiments with (SE, SP) values equal to (0.84 and 0.77).¹⁰ We correctly predict the domains I and III. However, the current form of our theory does not treat the long range kissing interaction between residues C139-G143 and C36-G40.¹⁰

Structure of HIV-1 5'UTR. The structure of the HIV-1 5' untranslated region (UTR) is still not fully studied. Until now, there is no consistent structural model for the 5' UTR. Previous computational and experimental studies confirmed the different hairpin domain such as the *trans*-activation region (TAR) and the polyadenylation signal hairpin [poly(A)] in the 5' UTR.

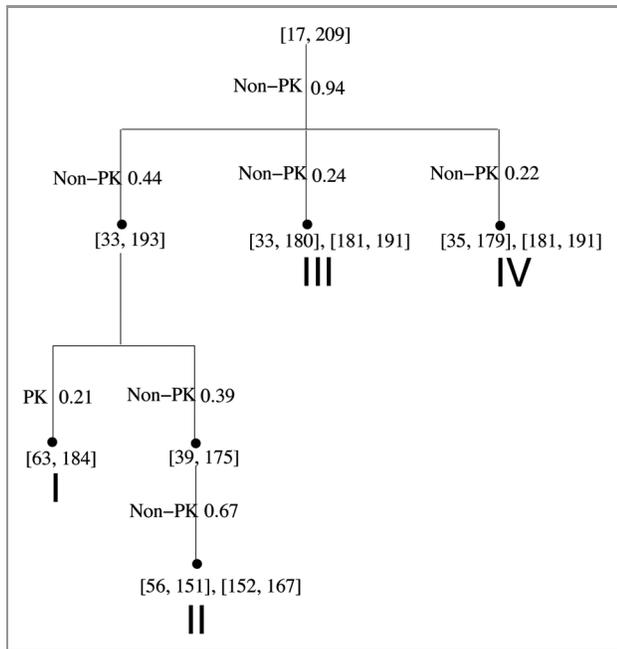


Figure 5. The predicted stable domain of hTR. There are four stable structures (I, II, III and IV) at room temperature. The non-PK and PK represent the non-pseudoknot and pseudoknot domains, respectively. The corresponding value is the probability for finding each domain.

the experiment. For HIV 5' UTR, the model does not predict pseudoknotted domains for the 363-nt truncated sequence. The structure predicted by the model is in accordance with the recent experimental results.⁷¹

For the hTR sequence, our model predicts a long stem-loop that coexists with a pseudoknotted structure. Based on the predicted structures, we propose that the telomerase activity may be connected to the metastability of the structure and the conformational switch between the pseudoknotted structure and the stem-loop structure. The result is consistent with earlier experimental reports on the hTR pseudoknot.^{49,63,64} It was proposed in these experimental studies that the conformational switch between the hairpin and the pseudoknot may be important for the hTR activity. Our theoretical predictions here support the experimental findings. We note that our results here are based on the calculations for the large hTR domains instead of the truncated short sequence of the model system studied in the experiments.¹⁸ The theoretical prediction here may be useful to facilitate understanding of the structure-function relationship for hTR.

Although this newly developed model can predict the large and complex pseudoknotted structure based on a physical model for RNA folding free energy, the model cannot treat loop-stem tertiary interaction between a large loop and a long stem. The previous work for estimating the loop-stem tertiary interaction is restricted to a short loop and a short stem.⁵⁸ For a long loop, we need to develop an ab initio model to evaluate the energetics for loop-stem tertiary interactions. Moreover, the current form of the model cannot give the free energies for structures involving long-range tertiary interactions such as the pseudoknots found in

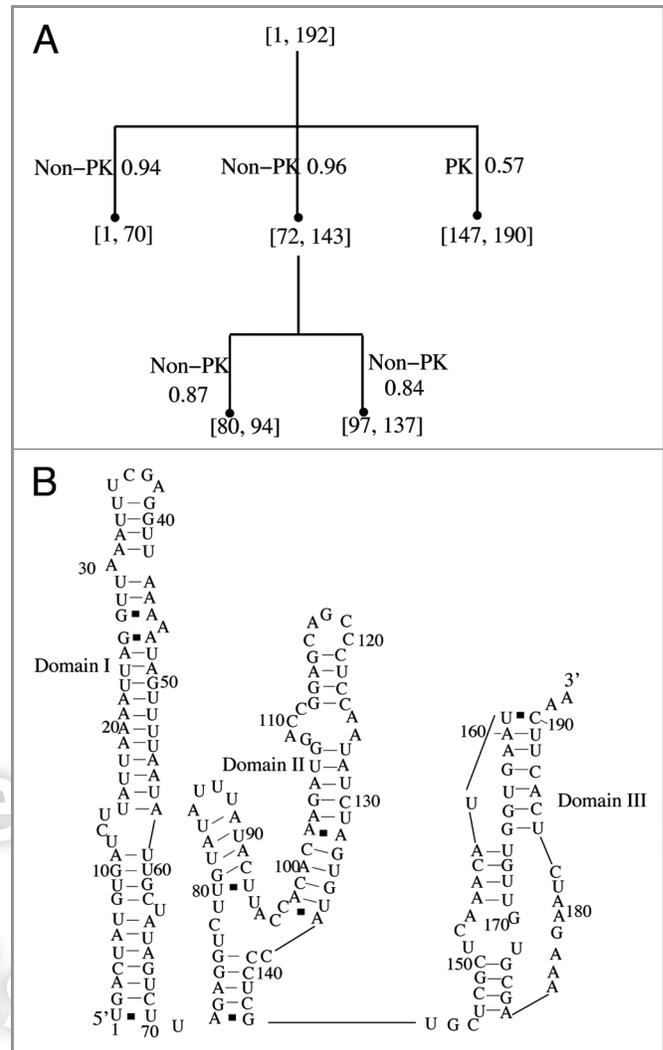


Figure 6. (A) The predicted stable structural domain of IGR IRES at room temperature. Also shown in the figure are the probabilities for forming each domain. (B) The predicted structure for the domains. We correctly predict domain I and domain III. Domain III is a pseudoknot structure. For domain II, the model cannot treat the long-range tertiary interaction between residues C139-G143 and C36-G40. The SE and SP values for the predicted structure are 0.84 and 0.77, respectively. In (a), the non-PK and PK represent the non-pseudoknot and pseudoknot domains, respectively.

large rRNAs. As a result, the model cannot predict pseudoknots formed by inter-domain long-range contacts in large RNAs.^{11,12} Furthermore, we can extend our algorithm to include the pseudoknot motif in the prediction of sRNA and their targets in bacteria. The pseudoknot motif is usually neglected in the existing algorithms.⁷⁸

Materials and Methods

The free energy of folding for a structure s is calculated from $\Delta G_s = \Delta G_{helix} - T\Delta S_{loop}$, where T is the temperature, ΔG_{helix} and ΔS_{loop} are the free energy and the entropy for the helix stems and the loops, respectively. In order to account for the intraloop

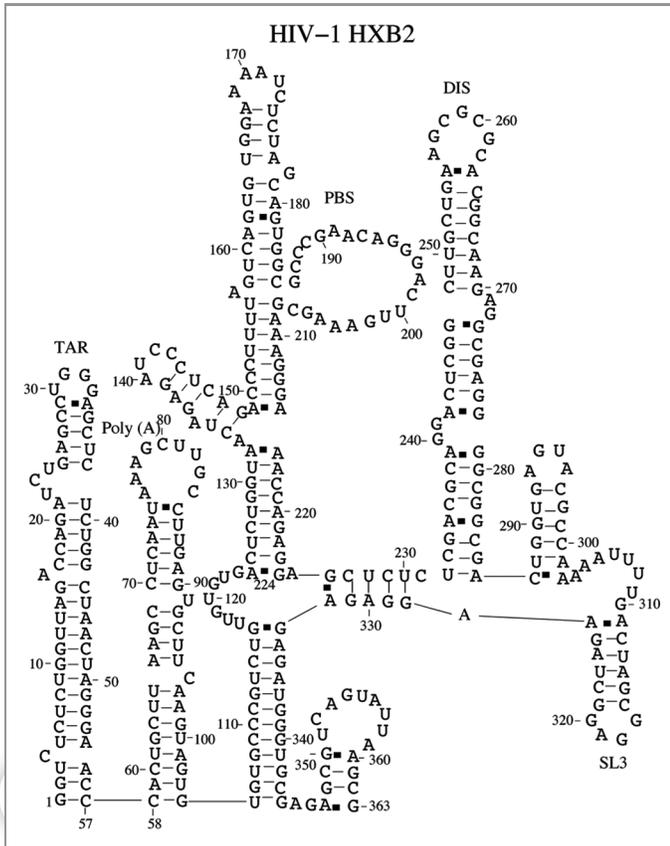


Figure 7. The predicted native structure for HIV-1 HXB2 sequence at room temperature. We correctly predict the TAR, Poly(A), PBS, DIS and SL3 domains. In addition, we find that nucleotides 105–116 form base pairings with nucleotides 333–344. The long-range interaction has been found in HIV-1 NL4–3 sequence.⁷¹

base stacks, we define an irreducible loop as a loop that contains no stable base stacks, including canonical and mismatched base stacks. Here a mismatched base stack is formed by a Watson-Crick (WC) base pair stacking on a non-WC base pair. The purpose of using irreducible loops is 2-fold. First, for a given (regular, non-irreducible) loop, it allows us to enumerate the conformations through the different assignments of the (intra-loop) mismatches.⁷⁹ The model gives (nonzero) loop enthalpy as the ensemble (Boltzmann) averaged sum of the enthalpies of the intraloop mismatches. Second, for a given irreducible loop, the simplicity of the conformations (without stable base stacks) makes the conformational count possible. We note that the conformational entropy for an irreducible loop can only come from the theoretical calculation. We predict the structures from the probability distribution for the formation of the different base pairs. Consider a macrostate z defined as the ensemble of all the structures, the folding free energy ΔG_z for the state can be computed from the partition function

$$Q_z = \sum_s e^{-\Delta G_s/k_B T}, \Delta G_z = -k_B T \ln Q_z \quad (1)$$

where k_B is the Boltzmann constant and the sum is over all the conformations “ s ” accessible to the macrostate z . The ratio between Q_z and the total partition function gives the probability p_z for the formation of the macrostate z . If macrostate z is the ensemble of structures that contain a specific base pair (i,j) , p_z would be the probability of forming base pair (i,j) . From the base pairing probability distribution for the different base pairs (i,j) , we deduce the structure.

The key step in the above procedure is the evaluation of the free energy ΔG_s for a given structure s . While the free energy for the helix stems ΔG_{helix} can be approximately evaluated from the nearest neighbor model,⁶⁰ the evaluation of the conformational entropy ΔS_{loop} for the loops/junctions requires a model. With the virtual bond-based RNA folding model (*Vfold* model), we develop a method to compute the entropy parameters.

We evaluate the entropy as $k_B \ln \Omega$, where Ω is the number of accessible conformations. We compute Ω through direct enumeration of chain (loop) conformations within the given structural context. Based on the rotameric properties of RNA backbone, we enumerate RNA conformations using a 3-vector virtual bond conformational representation for each nucleotide (*Vfold* model).³² We model a helix stem as an A-form helix using the experimentally measured atomic coordinates.⁸⁰ The (r, θ, z) coordinates (in a cylindrical coordinate system) for the P, C₄, and N₁ (or N₉) atoms are $(8.71 \text{ \AA}, 70.5 + i32.7, -3.75 + i2.81)$, $(9.68 \text{ \AA}, 46.9 + i32.7, -3.10 + i2.81)$, and $(7.12 \text{ \AA}, 37.2 + i32.7, -1.39 + i2.81)$ ($i=0,1,2,\dots$),⁸⁰ respectively. For the other strand, we need to negate θ and z to obtain the coordinates in cylindrical coordinate system. For a loop/junction, we generate the conformations of the loop through self-avoiding walks in the diamond lattice by using the *Vfold* model.⁵⁹ By counting the number of viable conformations of the loop, we calculate the entropy parameter of the loop.

In the present theory, we neglect the loop-stem tertiary interaction in order to keep the computational efficiency for the prediction of large (pseudoknotted) structures. Loop-stem tertiary contacts can contribute a non-zero enthalpy to the loop free energy. Our previous studies show that the loop-stem tertiary interactions do not change the native folded secondary structures for many H-type pseudoknots and only affect the stability of these pseudoknots.^{30,58} The benchmark test (see Table 1) shows that the current pseudoknot model based on explicit entropy calculation can indeed give good predictions for large pseudoknotted secondary structures.

Energetics parameters for the complex pseudoknot structure. We use hTR pseudoknot¹⁸ (see Fig. 8A) to illustrate the strategy. The structure contains three stems S_1, S_2 and S_3 and four loops L_1, L_2, L_3^1 and L_3^2 . We aim to evaluate and tabulate the entropy parameters for general structures with the different helix and loop lengths. Exhaustive computer enumeration of all the possible conformations for all the possible helix and loop lengths is not viable. To circumvent the problem, we first reduce the original structure so the entropy calculation becomes tractable.

1. The effect of stem S_3 on the conformations of loop L_3 ($= L_3^1 + L_3^2$) is mainly through the base pair (a, b) . Therefore, to

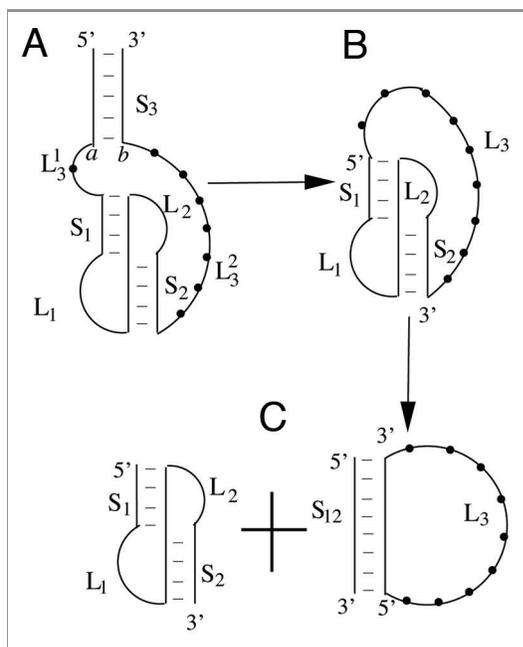


Figure 8. (A) A schematic diagram for a complex pseudoknot structure. (B) The simplified pseudoknot with stem S_3 replaced by a phantom nucleotide. (C) We divide the pseudoknot structure in (B) into two separate simple structures: a structure with stem S_{12} and loop L_3 and an H-type pseudoknot with stems S_1 and S_2 and loops L_1 and L_2 .

evaluate the number of viable conformation for loop L_3 , we can replace stem S_3 with the base pair (a, b) . Furthermore, we replace the (rigid) base pair (a, b) with a single nucleotide (see Fig. 8B) and accordingly, the length of loop L_3 is equal to $L_3^1 + L_3^2 + 1$.

2. Neglecting the excluded volume interaction between loop L_3 and loops L_1 and L_2 , we calculate the entropy parameter for loop L_3 and loops L_1 and L_2 separately. In the calculation, we account for the presence of the helices by considering the volume exclusion from the helix stems (see Fig. 8C). The entropy parameters are functions of the helix and loop lengths. The

entropy parameters for loops L_1 and L_2 are dependent on the stem lengths S_1 and S_2 and can be adopted from Table 2 of reference 30. For the entropy parameter of loop L_3 , we use a long stem S_{12} to represent stems S_1 and S_2 . The length of stem S_{12} is equal to the sum of the length of S_1 and the length of S_2 . The approximation is based on the fact that stems S_1 and S_2 tend to coaxially stack and form a quasi-continuous helix.⁶⁴

The entropy parameters for loop L_3 depend on the length of stem S_{12} . Figure 9A shows the distribution of the end-end distance, which is defined as the distance between the P atom in the 5' starting nucleotide of the loop L_3 and the P atom in the 3' end of the loop L_3 . The end-end distance increases as the sum (S_{12}) of the length of stem S_1 and the length of stem S_2 increases. Table 2 shows the entropy parameters for loop L_3 . In the calculation, the maximum loop length for a computationally viable calculation is 14 nts. larger loops, we extrapolate the entropy parameters using following formula:

$$\Delta S = a \ln(l - l_{min} + 1) + bl + c \quad (2)$$

where l_{min} is the minimal loop length L_3 in Table 2. We fit the parameters a , b and c based on the entropy values for shorter loops listed in Table 2. Table 3 shows the fitted parameters a , b and c .

The total loop entropy of hTR pseudoknot ($\Delta S(S_1, S_2, S_3, L_1, L_2, L_3^1, L_3^2)$) is given by the following formula:

$$\Delta S(S_1, S_2, S_3, L_1, L_2, L_3^1, L_3^2) = \Delta S(S_{12}, L_3) + \Delta S(S_1, S_2, L_1, L_2) \quad (3)$$

where $\Delta S(S_{12}, L_3)$ is the entropy for loop L_3 and $\Delta S(S_1, S_2, L_1, L_2)$ is the loop entropy for the reduced H-type pseudoknot (see Fig. 8C). The loop entropy parameter for the H-type pseudoknot can be read out from the entropy table reported in the previous studies.³⁰

Our previous models^{30,32} only treat simple H-type pseudoknot structures. In the current study, we extend the Vfold model to deal with more complex pseudoknot motif, such as a pseudoknot embedded in a hairpin loop ($a \rightarrow b$ in Fig. 8). The current model

Table 2. The entropy parameters ($\Delta S/k_B$) for the formation of loop L_3 in the pseudoknot motif of Figure 8B, where L is the length of loop L_3 and S_{12} is the length sum of stem S_1 and stem S_2

S_{12}/L	1	2	3	4	5	6	7	8	9	10	11	12	13	14
4	-	-	-	-	-	-	-12.8	-12.4	-12.5	-12.5	-12.7	-12.7	-12.9	-12.9
5	-	-	-	-	-	-	-	-12.8	-12.8	-12.7	-12.8	-12.8	-12.9	-12.9
6	-	-	-	-	-	-	-	-15.5	-14.0	-13.8	-13.7	-13.6	-13.6	-13.5
7	-	-	-	-	-	-	-	-14.2	-14.2	-14.0	-14.0	-13.9	-13.9	-13.9
8	-	-	-	-	-	-	-	-14.5	-13.8	-13.4	-13.3	-13.2	-13.2	-13.3
9	-	-	-	-	-	-	-	-	-14.8	-13.8	-13.5	-13.4	-13.4	-13.4
10	-	-	-	-	-	-	-	-	-14.9	-13.8	-13.5	-13.4	-13.4	-13.4
11	-	-	-	-	-	-	-	-	-14.8	-14.2	-14.1	-14.0	-14.0	-14.0
12	-	-	-	-	-	-	-	-	-16.6	-15.4	-14.9	-14.6	-14.5	-14.4
13	-	-	-	-	-	-	-	-	-	-17.0	-16.1	-15.5	-15.3	-15.0
14	-	-	-	-	-	-	-	-	-	-	-18.3	-17.1	-16.5	-16.0
15	-	-	-	-	-	-	-	-	-	-	-	-22.8	-21.1	-19.9

Table 3. The fitted parameters for the loop entropies in Equation 1

stem length	4	5	6	7	8	9	10	11	12	13	14	15
a	0.66	0.20	2.27	0.14	1.49	1.96	2.13	1.09	2.19	1.51	1.77	2.73
b	-0.23	-0.08	-0.44	0.013	-0.28	-0.43	-0.48	-0.24	-0.35	-0.11	-0.06	-0.06
c	-11.1	-12.2	-11.8	-14.3	-12.3	-10.9	-10.6	-12.6	-13.4	-15.9	-17.7	-22.1

allows us to calculate and tabulate the entropy parameters for such complex pseudoknotted motifs. Furthermore, we develop a domain-based approach to treat pseudoknot-forming long sequences. With this new method, we can easily treat the

sequence length with 400 nts. However, in the original Vfold model,^{30,32} it is difficult to predict the pseudoknot structure even for a 100-nt sequence (see Fig. 9B).

Computational time. The recursive process of domain-reduction quickly converges to the irreducible domain, thus, the procedure is computationally efficient. The computational time is limited only by the size of the pseudoknotted domain. For each irreducible pseudoknotted domain, the computer time for calculating the partition function scales as $O(N^{11})$.³⁰ For the RNA secondary structure, the computer time scales as $O(N^6)$.⁵⁹ Therefore, the total computational efficiency is limited by the calculations for the pseudoknotted domains.

The computer time $t(L)$ for searching for the 5' domain boundary nucleotides (using Eq. 1) grows algebraically with the sequence length L (see Fig. 9B). For a sequence of 400 nt, it takes less than 4 h. The required computer time for the present approach is significantly shorter than the original Vfold model, which can only treat sequences less than 100 nt (see Fig. 9B). It takes about the same computational time to find the 3' boundary nucleotides for the domains. For a large RNA with M domains, the total computer time is approximately $2M t(L)$. For a system with $M = 5$ domains, it takes about 40 h to predict the full structure for a 400-nt sequence. The domain-based method is computationally feasible for large RNAs. In addition, we compared the computational times for our model and other existing models. We found that our model is slower than HotKnots, ProbKnot, STAR and pknots-RG and faster than pknotsRE and NUPACK.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed

Acknowledgments

This research was supported by NIH through grant GM063732 and NSF through grants MCB0920411 and MCB0920067. Most of numerical calculations involved in this research were performed on the HPC resources at the University of Missouri Bioinformatics Consortium (UMBC).

Supplemental Materials

Supplemental materials can be found at:
www.landesbioscience.com/journals/rnabiology/article/18488

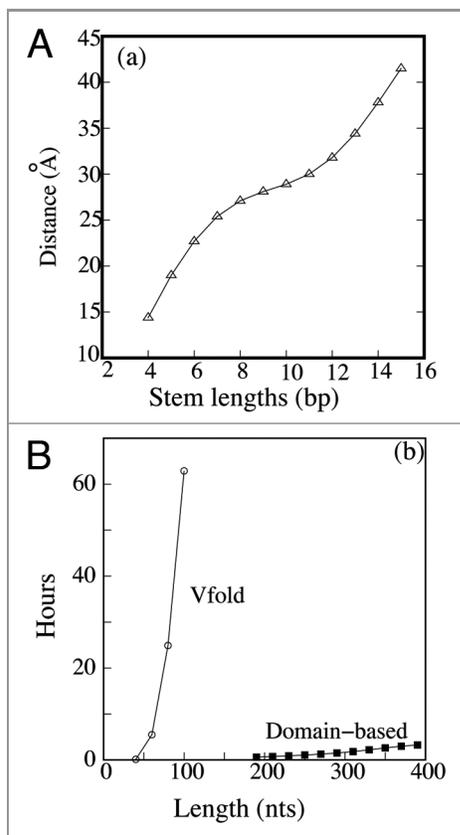


Figure 9. (A) The end-to-end distance for a loop (see L_3 in Fig. 8) with the different stem lengths (S_{12}). The length of stem S_{12} is the sum of the stem S_1 length and the S_2 length. The end-to-end distance of a loop is defined as the distance between the P atom at the 5' end of loop L_3 and the P atom at the 3' end of loop L_3 . (B) The comparison of the computational time between the original Vfold model (square) for predicting pseudoknot structure and the new domain-based algorithm for the calculation of the probability for a specific nucleotide to be the 5' boundary of a domain (circle). The computer resource that we use is Dell EM64T cluster with the Intel (R) Xeon(R) 5150 (2.66 GHz) processor.

References

- Gulyaev AP, Olsthoorn RCL. A family of non-classical pseudoknots in influenza A and B viruses. *RNA Biol* 2010; 7:125-9; PMID:20200490; <http://dx.doi.org/10.4161/rna.7.2.11287>
- Perrotta AT, Been MD. A pseudoknot-like structure required for efficient self-cleavage of hepatitis delta virus RNA. *Nature* 1991; 350:434-6; PMID:2011192; <http://dx.doi.org/10.1038/350434a0>
- Ferré-D'Amaré AR, Zhou KH, Doudna JA. Crystal structure of a hepatitis delta virus ribozyme. *Nature* 1998; 395:567-74; PMID:9783582; <http://dx.doi.org/10.1038/26912>
- Tanner NK, Schaff S, Thill G, Petit-Koskas E, Crain-Denoyelle AM, Westhof E. A three-dimensional model of hepatitis delta virus ribozyme based on biochemical and mutational analyses. *Curr Biol* 1994; 4:488-98; PMID:7922369; [http://dx.doi.org/10.1016/S0960-9822\(00\)00109-3](http://dx.doi.org/10.1016/S0960-9822(00)00109-3)
- Schultes EA, Bartel DP. One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science* 2000; 289:448-52; PMID:10903205; <http://dx.doi.org/10.1126/science.289.5478.448>
- Deiman BA, Kortlever RM, Pleij CW. The role of the pseudoknot at the 3' end of turnip yellow mosaic virus RNA in minus-strand synthesis by the viral RNA-dependent RNA polymerase. *J Virol* 1997; 71:5990-6; PMID:9223489
- van Belkum A, Abrahams JP, Pleij CW, Bosch L. Five pseudoknots are present at the 204 nucleotides long 3' noncoding region of tobacco mosaic virus RNA. *Nucleic Acids Res* 1985; 13:7673-86; PMID:3934645; <http://dx.doi.org/10.1093/nar/13.21.7673>
- Su L, Chen L, Egli M, Berger JM, Rich A. Minor groove RNA triplex in the crystal structure of a ribosomal frameshifting viral pseudoknot. *Nat Struct Biol* 1999; 6:285-92; PMID:10074948; <http://dx.doi.org/10.1038/6722>
- Giedroc DP, Theimer CA, Nixon PL. Structure, stability and function of RNA pseudoknots involved in stimulating ribosomal frameshifting. *J Mol Biol* 2000; 298:167-85; PMID:10764589; <http://dx.doi.org/10.1006/jmbi.2000.3668>
- Costantino D, Kieft JS. A performed compact ribosome-binding domain in the cricket paralysis-like virus IRES RNAs. *RNA* 2005; 12:332-43; <http://dx.doi.org/10.1261/rna.7184705>
- Poot RA, Pleij CWA, van Duin J. The central pseudoknot in 16S ribosomal RNA is needed for ribosome stability but is not essential for 30S initiation complex formation. *Nucleic Acids Res* 1996; 24:3670-6; PMID:8871543; <http://dx.doi.org/10.1093/nar/24.19.3670>
- villa A, Farley-viril J, Tappich WE. Pseudoknot in the central domain of small subunit ribosomal RNA is essential for translation. *Proc Natl Acad Sci USA* 1994; 23:11148-52.
- Chen G, Chang KY, Chou MY, Bustamante C, Tinoco I, Jr. Triplex structures in an RNA pseudoknot enhance mechanical stability and increase efficiency of -1 ribosomal frameshifting. *Proc Natl Acad Sci U S A* 2009; 106:12706-11; PMID:19628688; <http://dx.doi.org/10.1073/pnas.0905046106>
- Giedroc DP, Cornish PV. Frameshifting RNA pseudoknots: structure and mechanism. *Virus Res* 2009; 139:193-208; PMID:18621088; <http://dx.doi.org/10.1016/j.virusres.2008.06.008>
- Plant EP, Jacobs KL, Harger JW, Meskauskas A, Jacobs JL, Baxter JL, et al. The 9-A solution: how mRNA pseudoknots promote efficient programmed -1 ribosomal frameshifting. *RNA* 2003; 9:168-74; PMID:12554858; <http://dx.doi.org/10.1261/rna.2132503>
- Namy O, Moran SJ, Stuart DI, Gilbert RJ, Brierley I. A mechanical explanation of RNA pseudoknot function in programmed ribosomal frameshifting. *Nature* 2006; 441:244-7; PMID:16688178; <http://dx.doi.org/10.1038/nature04735>
- Cao S, Chen S-J. Predicting ribosomal frameshifting efficiency. *Phys Biol* 2008; 5:016002; PMID:18367782; <http://dx.doi.org/10.1088/1478-3975/5/1/016002>
- Theimer CA, Blois CA, Feigon J. Structure of the human telomerase RNA pseudoknot reveals conserved tertiary interactions essential for function. *Mol Cell* 2005; 17:671-82; PMID:15749017; <http://dx.doi.org/10.1016/j.molcel.2005.01.017>
- Pfingsten JS, Costantino DA, Kieft JS. Structural basis for ribosome recruitment and manipulation by a viral IRES RNA. *Science* 2006; 314:1450-4; PMID:17124290; <http://dx.doi.org/10.1126/science.1133281>
- Ruan J, Stormo GD, Zhang W. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics* 2004; 20:58-66; PMID:14693809; <http://dx.doi.org/10.1093/bioinformatics/btg373>
- Witwer C, Hofacker IL, Stadler PF. Prediction of consensus RNA secondary structures including pseudoknots. *IEEE/ACM Trans Comput Biol Bioinform* 2004; 1:66-77; PMID:17048382; <http://dx.doi.org/10.1109/TCBB.2004.22>
- Ren J, Rastegari B, Condon A, Hoos HH. HotKnots: heuristic prediction of RNA secondary structures including pseudoknots. *RNA* 2005; 11:1494-504; PMID:16199760; <http://dx.doi.org/10.1261/rna.7284905>
- Rivas E, Eddy SR. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J Mol Biol* 1999; 285:2053-68; PMID:9925784; <http://dx.doi.org/10.1006/jmbi.1998.2436>
- Dirks RM, Pierce NA. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J Comput Chem* 2003; 24:1664-77; PMID:12926009; <http://dx.doi.org/10.1002/jcc.10296>
- Reeder J, Giegerich R. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinformatics* 2004; 5:104; PMID:15294028; <http://dx.doi.org/10.1186/1471-2105-5-104>
- Gulyaev AP, van Batenburg FHD, Pleij CWA. The computer simulation of RNA folding pathways using a genetic algorithm. *J Mol Biol* 1995; 250:37-51; PMID:7541471; <http://dx.doi.org/10.1006/jmbi.1995.0356>
- Kasprzak W, Bindewald E, Shapiro BA. Structural polymorphism of the HIV-1 leader region explored by computational methods. *Nucleic Acids Res* 2005; 33:7151-63; PMID:16371347; <http://dx.doi.org/10.1093/nar/gki1015>
- Xayaphoummine A, Bucher T, Isambert H. Kinofold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acids Res* 2005; 33(Web Server issue):W605-10; PMID:15980546; <http://dx.doi.org/10.1093/nar/gki447>
- Poolap U, Kato Y, Akutsu T. Prediction of RNA secondary structure with pseudoknots using integer programming. *BMC Bioinformatics* 2009; 10(Suppl 1):S38; PMID:19208139; <http://dx.doi.org/10.1186/1471-2105-10-S1-S38>
- Cao S, Chen S-J. Predicting RNA pseudoknot folding thermodynamics. *Nucleic Acids Res* 2006; 34:2634-52; PMID:16709732; <http://dx.doi.org/10.1093/nar/gkl346>
- Chen S-J. RNA folding: conformational statistics, folding kinetics, and ion electrostatics. *Annu Rev Biophys* 2008; 37:197-214; PMID:18573079; <http://dx.doi.org/10.1146/annurev.biophys.37.032807.125957>
- Cao S, Chen S-J. Predicting structures and stabilities for H-type pseudoknots with interhelix loops. *RNA* 2009; 15:696-706; PMID:19237463; <http://dx.doi.org/10.1261/rna.1429009>
- Zhang J, Dundas J, Lin M, Chen R, Wang W, Liang J. Prediction of geometrically feasible three-dimensional structures of pseudoknotted RNA through free energy estimation. *RNA* 2009; 15:2248-63; PMID:19864433; <http://dx.doi.org/10.1261/rna.1723609>
- Andronescu MS, Pop C, Condon AE. Improved free energy parameters for RNA pseudoknotted secondary structure prediction. *RNA* 2010; 16:26-42; PMID:19933322; <http://dx.doi.org/10.1261/rna.1689910>
- Sperschneider J, Datta A. DotKnot: pseudoknot prediction using the probability dot plot under a refined energy model. *Nucleic Acids Res* 2010; 38:e103; PMID:20123730; <http://dx.doi.org/10.1093/nar/gkq021>
- Wiebe NJP, Meyer IM. Transat-A method for detecting the conserved helices of functional RNA structures, including transient, pseudo-knotted and alternative structures. *PLOS Comput Biol* 2010; 6:1-13; <http://dx.doi.org/10.1371/journal.pcbi.1000823>
- Bellaousov S, Mathews DH. ProbKnot: fast prediction of RNA secondary structure including pseudoknots. *RNA* 2010; 16:1870-80; PMID:20699301; <http://dx.doi.org/10.1261/rna.2125310>
- Sperschneider J, Datta A, Wise MJ. Heuristic RNA pseudoknot prediction including intramolecular kissing hairpins. *RNA* 2011; 17:27-38; PMID:21098139; <http://dx.doi.org/10.1261/rna.2394511>
- McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* 1990; 29:1105-19; PMID:1695107; <http://dx.doi.org/10.1002/bip.360290621>
- Nussinov R, Jacobson AB. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc Natl Acad Sci U S A* 1980; 77:6309-13; PMID:6161375; <http://dx.doi.org/10.1073/pnas.77.11.6309>
- Tan RKZ, Petrov AS, Harvey SC. YUP: A molecular simulation program for coarse-grained and multiscaled models. *J Chem Theory Comput* 2006; 2:529-40; <http://dx.doi.org/10.1021/ct050323r>
- Bevilacqua PC, SantaLucia J, Jr.. The biophysics of RNA. *ACS Chem Biol* 2007; 2:440-4; PMID:17649965; <http://dx.doi.org/10.1021/cb7001363>
- Mathews DH, Moss WN, Turner DH. Folding and finding RNA secondary structure. *Cold Spring Harb Perspect Biol* 2010; 2:a003665; PMID:20685845; <http://dx.doi.org/10.1101/cshperspect.a003665>
- Westhof E, Masquida B, Jossinet F. Predicting and modeling RNA architecture. *Cold Spring Harb Perspect Biol* 2011; 3:a003632; PMID:20504963; <http://dx.doi.org/10.1101/cshperspect.a003632>
- Rother K, Rother M, Boniecki M, Puton T, Bujnicki JM. RNA and protein 3D structure modeling: similarities and differences. *J Mol Model* 2011; 17:2325-36; PMID:21258831; <http://dx.doi.org/10.1007/s00894-010-0951-x>
- Gulyaev AP, van Batenburg FH, Pleij CW. An approximation of loop free energy values of RNA H-pseudoknots. *RNA* 1999; 5:609-17; PMID:10334330; <http://dx.doi.org/10.1017/S135583829998189X>
- Jaeger JA, Turner DH, Zuker M. Improved predictions of secondary structures for RNA. *Proc Natl Acad Sci U S A* 1989; 86:7706-10; PMID:2479010; <http://dx.doi.org/10.1073/pnas.86.20.7706>
- Dotú I, Lorenz WA, Van Hentenryck P, Clote P. RNA Structural Segmentation. *Pac Symp Biocomput* 2010; 57-68.

49. Antal M, Boros E, Solymosy F, Kiss T. Analysis of the structure of human telomerase RNA in vivo. *Nucleic Acids Res* 2002; 30:912-20; PMID:11842102; <http://dx.doi.org/10.1093/nar/30.4.912>
50. Pennell S, Manktelow E, Flatt A, Kelly G, Smerdon SJ, Brierley I. The stimulatory RNA of the Visna-Maedi retrovirus ribosomal frameshifting signal is an unusual pseudoknot with an interstem element. *RNA* 2008; 14:1366-77; PMID:18495941; <http://dx.doi.org/10.1261/rna.1042108>
51. Plant EP, Pérez-Alvarado GC, Jacobs JL, Mukhopadhyay B, Hennig M, Dinman JD. A three-stemmed mRNA pseudoknot in the SARS coronavirus frameshift signal. *PLoS Biol* 2005; 3:e172; PMID:15884978; <http://dx.doi.org/10.1371/journal.pbio.0030172>
52. Isambert H, Siggia ED. Modeling RNA folding paths with pseudoknots: application to hepatitis delta virus ribozyme. *Proc Natl Acad Sci U S A* 2000; 97:6515-20; PMID:10823910; <http://dx.doi.org/10.1073/pnas.110533697>
53. Ferré-D'Amaré AR, Zhou K, Doudna JA. Crystal structure of a hepatitis delta virus ribozyme. *Nature* 1998; 395:567-74; PMID:9783582; <http://dx.doi.org/10.1038/26912>
54. van Belkum A, Abrahams JP, Pleij CW, Bosch L. Five pseudoknots are present at the 204 nucleotides long 3' noncoding region of tobacco mosaic virus RNA. *Nucleic Acids Res* 1985; 13:7673-86; PMID:3934645; <http://dx.doi.org/10.1093/nar/13.21.7673>
55. Deiman BALM, Kortlever RM, Pleij CWA. The role of the pseudoknot at the 3' end of turnip yellow mosaic virus RNA in minus-strand synthesis by the viral RNA-dependent RNA polymerase. *J Virol* 1997; 71:5990-6; PMID:9223489
56. Theimer CA, Feigon J. Structure and function of telomerase RNA. *Curr Opin Struct Biol* 2006; 16:307-18; PMID:16713250; <http://dx.doi.org/10.1016/j.sbi.2006.05.005>
57. Chen JL, Blasco MA, Greider CW. Secondary structure of vertebrate telomerase RNA. *Cell* 2000; 100:503-14; PMID:10721988; [http://dx.doi.org/10.1016/S0092-8674\(00\)80687-X](http://dx.doi.org/10.1016/S0092-8674(00)80687-X)
58. Cao S, Giedroc DP, Chen S-J. Predicting loop-helix tertiary structural contacts in RNA pseudoknots. *RNA* 2010; 16:538-52; PMID:20100813; <http://dx.doi.org/10.1261/rna.1800210>
59. Cao S, Chen S-J. Predicting RNA folding thermodynamics with a reduced chain representation model. *RNA* 2005; 11:1884-97; PMID:16251382; <http://dx.doi.org/10.1261/rna.2109105>
60. Serra MJ, Turner DH. Predicting thermodynamic properties of RNA. *Methods Enzymol* 1995; 259:242-61; PMID:8538457; [http://dx.doi.org/10.1016/0076-6879\(95\)59047-1](http://dx.doi.org/10.1016/0076-6879(95)59047-1)
61. Walter AE, Turner DH. Sequence dependence of stability for coaxial stacking of RNA helices with Watson-Crick base paired interfaces. *Biochemistry* 1994; 33:12715-9; PMID:7522562; <http://dx.doi.org/10.1021/bi00208a024>
62. Kolk MH, van der Graaf M, Wijmenga SS, Pleij CWA, Heus HA, Hilbers CW. NMR structure of a classical pseudoknot: interplay of single- and double-stranded RNA. *Science* 1998; 280:434-8; PMID:9545221; <http://dx.doi.org/10.1126/science.280.5362.434>
63. Comolli LR, Smirnov I, Xu L, Blackburn EH, James TL. A molecular switch underlies a human telomerase disease. *Proc Natl Acad Sci U S A* 2002; 99:16998-7003; PMID:12482936; <http://dx.doi.org/10.1073/pnas.262663599>
64. Theimer CA, Finger LD, Trantírek L, Feigon J. Mutations linked to dyskeratosis congenita cause changes in the structural equilibrium in telomerase RNA. *Proc Natl Acad Sci U S A* 2003; 100:449-54; PMID:12525685; <http://dx.doi.org/10.1073/pnas.242720799>
65. Chen JL, Greider CW. Telomerase RNA structure and function: implications for dyskeratosis congenita. *Trends Biochem Sci* 2004; 29:183-92; PMID:15082312; <http://dx.doi.org/10.1016/j.tibs.2004.02.003>
66. Cao S, Chen S-J. Biphasic folding kinetics of RNA pseudoknots and telomerase RNA activity. *J Mol Biol* 2007; 367:909-24; PMID:17276459; <http://dx.doi.org/10.1016/j.jmb.2007.01.006>
67. Pelletier J, Sonenberg N. Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. *Nature* 1988; 334:320-5; PMID:2839775; <http://dx.doi.org/10.1038/334320a0>
68. Jang SK, Kräusslich HG, Nicklin MJ, Duke GM, Palmenberg AC, Wimmer E. A segment of the 5' nontranslated region of encephalomyocarditis virus RNA directs internal entry of ribosomes during in vitro translation. *J Virol* 1988; 62:2636-43; PMID:2839690
69. Otto GA, Puglisi JD. The pathway of HCV IRES-mediated translation initiation. *Cell* 2004; 119:369-80; PMID:15507208; <http://dx.doi.org/10.1016/j.cell.2004.09.038>
70. Baird SD, Turcotte M, Korneluk RG, Holecik M. Searching for IRES. *RNA* 2006; 12:1755-85; PMID:16957278; <http://dx.doi.org/10.1261/ma.157806>
71. Wilkinson KA, Gorelick RJ, Vasa SM, Guex N, Rein A, Mathews DH, et al. High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states. *PLoS Biol* 2008; 6:e96; PMID:18447581; <http://dx.doi.org/10.1371/journal.pbio.0060096>
72. Watts JM, Dang KK, Gorelick RJ, Leonard CW, Bess JW, Jr., Swanstrom R, et al. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* 2009; 460:711-6; PMID:19661910; <http://dx.doi.org/10.1038/nature08237>
73. Clever JL, Miranda D, Jr., Parslow TG. RNA structure and packaging signals in the 5' leader region of the human immunodeficiency virus type 1 genome. *J Virol* 2002; 76:12381-7; PMID:12414982; <http://dx.doi.org/10.1128/JVI.76.23.12381-12387.2002>
74. Russell RS, Liang C, Wainberg MA. Is HIV-1 RNA dimerization a prerequisite for packaging? Yes, no, probably? *Retrovirology* 2004; 1:23; PMID:15355057; <http://dx.doi.org/10.1186/1742-4690-1-23>
75. Damgaard CK, Andersen ES, Knudsen B, Gorodkin J, Kjems J. RNA interactions in the 5' region of the HIV-1 genome. *J Mol Biol* 2004; 336:369-79; PMID:14757051; <http://dx.doi.org/10.1016/j.jmb.2003.12.010>
76. Paillart JC, Dettenhofer M, Yu XF, Ehresmann C, Ehresmann B, Marquet R. First snapshots of the HIV-1 RNA structure in infected cells and in virions. *J Biol Chem* 2004; 279:48397-403; PMID:15355993; <http://dx.doi.org/10.1074/jbc.M408294200>
77. Berkhout B, Ooms M, Beerens N, Huthoff H, Southern E, Verhoef K. In vitro evidence that the untranslated leader of the HIV-1 genome is an RNA checkpoint that regulates multiple functions through conformational changes. *J Biol Chem* 2002; 277:19967-75; PMID:11896057; <http://dx.doi.org/10.1074/jbc.M200950200>
78. Backofen R, Hess WR. Computational prediction of sRNAs and their targets in bacteria. *RNA Biol* 2010; 7:33-42; PMID:20061798; <http://dx.doi.org/10.4161/rna.7.1.10655>
79. Lu ZJ, Turner DH, Mathews DH. A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation. *Nucleic Acids Res* 2006; 34:4912-24; PMID:16982646; <http://dx.doi.org/10.1093/nar/gkl472>
80. Arnott S, Hukins DWL, Dover SD. Optimised parameters for RNA double-helices. *Biochem Biophys Res Commun* 1972; 48:1392-9; PMID:5077825; [http://dx.doi.org/10.1016/0006-291X\(72\)90867-4](http://dx.doi.org/10.1016/0006-291X(72)90867-4)
81. Watts JM, Dang KK, Gorelick RJ, Leonard CW, Bess JW, Jr., Swanstrom R, et al. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* 2009; 460:711-6; PMID:19661910; <http://dx.doi.org/10.1038/nature08237>