Journal reference: RNA ; Published in Advance October 28, 2009, doi:10.1261/rna.1723609

Prediction of Geometrically Feasible Three Dimensional Structures of Pseudoknotted RNA through Free Energy Estimation

Jian Zhang[1,2], Joseph Dundas[1], Ming Lin[3], Rong Chen[3], Wei Wang[2], Jie Liang[1,*]
[1]*Department of Bioengineering, University of Illinois at Chicago, Chicago, Illinois 60607, USA;*
[2]*National Laboratory of Solid State Microstructure, Nanjing University, Nanjing 21093, China;*
[3]*Department of Statistics, Rutgers University, Piscataway, New Jersey 08854-8019, USA* *

Accurate free energy estimation is essential for RNA structure prediction. The widely used Turner's energy model works well for nested structures. For pseudoknotted RNAs, however, there is no effective rule for estimation of loop entropy and free energy. In this work we present a new free energy estimation method, named pk3D (for PseudoKnot predictor in three dimensional space), which goes beyond the Turner's model. Our approach is based on the physical consideration that the loop entropy is determined primarily by the loop length, the end-to-end distance of the connected stems, and the interference from nearby structures. We first generate a list of candidate secondary structures using an approximation algorithm that assembles all possible low energy stem regions without conflicting assignment of nucleotides to multiple stems. For each candidate secondary structure, we then systematically search possible arrangements of helical stems in space, and assess the overall free energy of each arrangement for identification of the optimal spatial structure. Here we calculate the loop entropy as the fraction of number of loop conformations with respect to the number of random coil of the same length. It is estimated using a 6-state discrete RNA chain model and sequential Monte Carlo sampling. Our approach treats nested and pseudoknotted structures alike in one unifying physical framework, regardless how complex the RNA structures are. We first test the ability of pk3D in selecting native structures from a large number of decoys for a set of 43 pseudoknotted RNA molecules, with length ranging from 23 to 113. We find that pk3D performs slightly better than the Dirks and Pierce extension of the Turner's rule. We then test pk3D for blind secondary structure prediction, and find that pk3D gives the best sensitivity and comparable positive predictive value (PPV, related to specificity) in predicting pseudoknotted RNA secondary structures, when compared with other methods, including HotKnots, pknotsRG, NUPACK and ILM. A unique strength of pk3D is that it also generates spatial arrangement of structural elements of the RNA molecule. Comparison of three-dimensional structures predicted by pk3D with the native structure measured by NMR or X-ray experiments shows that the predicted spatial arrangement of stems and loops is often similar to that found in the native structure. These close-to-native structures can be used as starting points for further refinement to derive accurate three-dimensional structures of RNA molecules, including those with pseudoknots.

Keywords: RNA structure prediction, Pseudoknots, RNA free energy estimation, RNA structure modeling, RNA secondary and tertiary structure.

Running title: Prediction of RNA structures with pseudoknots

I. INTRODUCTION

Biological functions of RNA range from carrying genetic information, participating in protein synthesis, catalyzing biochemical reactions, regulating gene expressions, to acting as a structural molecule in cellular organelles [1]. To understand how RNA molecules perform these tasks, knowledge of the three-dimensional structures of RNA is often required. Although the most reliable sources of RNA structural information are experimental measurements from X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy, experimental structures of RNAs are technically challenging to obtain and are costly in both time and efforts. As a result, knowledge of RNA structures lags far behind that of RNA sequences. Computational prediction of RNA structures therefore can provide an alternative source of information for gaining biological insights.

Prediction of secondary structures of small and non-pseudoknotted RNAs has been very successful [2]. Predicted secondary structures of RNA molecules can provide valuable information, as they could reveal the functions of RNA molecules [3], help understanding RNA folding as RNAs often fold hierarchically [4]. They can also be used for RNA comparison [5], for predicting three dimensional RNA structures [6]. There are two general strategies in predicting RNA secondary structures. The most successful one is through comparative sequence analysis [7, 8], which utilizes homology information and incorporates many complex factors in determining RNA structures implicitly. However, the approach of comparative sequence analysis requires the availability of many related RNA sequences, thus is not always feasible. The

*Author to whom correspondence should be addressed. Jie Liang, Fax: (312)413-2018, Email: jliang@uic.edu

other approach is through minimization of free energies (MFE). This is based on the thermodynamic hypothesis, which states that the conformation with the lowest free energy is the native RNA structure [9].

The most widely used secondary structure prediction programs are based on the second strategy. Among these, Mfold [10], RNAfold [11], and RNAstructure [12] are based on dynamic programming and guarantee the generation of a structure with the lowest free energy, within the accuracy limitations of the free energy rules employed, and with the condition that the RNA secondary structures are nested and contain no pseudoknot. When pseudoknotted RNAs are excluded, Mfold, for example, can correctly predict an average of 69% of known base-pairs in a test involving a large data set containing tRNA, 5S, 16S, 32S, Group I, II intron, RNase P and SRP RNAs [2].

However, prediction of RNA secondary structure with pseudoknots is far more challenging. For example, when the test set includes pseudoknotted RNAs, the accuracy of Mfold prediction deteriorates to $54\% - 68\%$ [13–19]. The difficulties are two fold. First, there are exponentially many ways that pseudoknots can form. It has been shown that the general problem of predicting RNA secondary structures containing pseudoknots is NP hard [22, 23]. Methods based on dynamic programming cannot solve this problem, aside from a few special cases [14, 15]. Second, there is no known effective free energy model that can describe accurately the free energy of loops in pseudoknotted RNA molecules.

Recently, several methods have been designed to predict the secondary structures of RNAs with pseudoknots. These include NUPACK [13], pknotsRE [14], pknotsRG-mfe [15], ILM [16], TdFold [17], STAR [18], HotKnots [19], FlexStem [20], and Kinefold [21]. These either employed approximation algorithms and generated predicted structures that were within certain approximation ratio with the optimal structure [13–17], or employed empirical algorithms that were more stochastic in nature [18–21]. The average accuracy were improved to $76\% - 80\%$ for small pseudoknots (< 150 nucleotides). For large pseudoknots, the problem was still very challenging, and the accuracy ranged from 36% to 55%.

All these methods represented development of new algorithms in generating candidate secondary structures. They all employed the well-established Turner's energy rule and its modifications [2, 24, 25] to estimate the free energy of pseudoknotted RNA secondary structures [13]. However, the lack of progress in calculating free energies associated with pseudoknotted loops posed a significant limit on what these methods could achieve.

There has been a long line of researches studying the free energy rules of pseudoknots, with the entropic cost as the main focus, since the enthalpic contribution can be accurately obtained by using Turner's energy rule. Gultyaev *et al.* adjusted parameters of equations derived from polymer theory such that they were consistent with known data on pseudoknots. They have compiled a table of recommended free energy values for H-type pseudoknots with different stem and loop lengths [26]. Using the Gaussian chain approximation and neglecting the excluded volume effects, Aalberts *et al.* developed a model to estimate the free energy of pseudoknots of ABAB-type [27]. Isambert and Siggia treated pseudoknots in two stages by modeling short-scale structures as "net" and large-scale structures as "Gaussian crosslinked gel" in their Kinefold method [21]. They calculated the short-scale conformational entropy analytically from the Gaussian chain model, and obtained the large-scale entropy by algebraic integrations. According to the authors, the excluded volume effects were incorporated crudely by adjusting the value of an exponent [21]. Based on Rivas and Eddy's work [14], Dirks and Pierce developed a free energy rule for pseudoknots using a phenomenological linear equation, in which the coefficients were trained by using known data of pseudoknots [13]. This model has become the standard for pseudoknot free energy calculation and has been often regarded as an extension to the Turner's energy rule, due to its Turner-style formulation as well as the linear functional form necessary for dynamic programming algorithm. Although this free energy rule is used frequently, it is not realistic as it does not model the important excluded volume effect.

Another important work on modeling RNA free energies was a constraint generation method presented by Andronescu *et al.* [28], which employed an iterative scheme to train hundreds of free energy parameters on large sets of structural and thermodynamic data. Based on the parameters optimized by this method, significant improvements in prediction accuracy over the other methods have been achieved. Aside from these thermodynamical measurement based methods, there were alternative probabilistic methodologies for modeling RNA secondary structures. Among them, CONTRAfold was based on conditional log-linear models and generalized upon stochastic context-free grammars by using discriminative training and feature-rich scoring [29]. The work of CONTRAfold method demonstrated that statistical learning procedures provide an effective alternative to physics-based approach in deriving parameters for RNA secondary structure prediction.

Despite these successes, calculating free energy of loop regions in RNA pseudoknots is still an unsolved problem. In our opinion, this is intrinsically a three-dimensional problem. For loops embedded in pseudoknots, commonly used secondary structural features such as number of base pairs bordering the pseudoknots, loop asymmetricity, and penalty for overlapping pseudoknots do not necessarily capture the most relevant information. It is necessary to develop new free energy rule based on three dimensional spatial models rather than models solely restricted to secondary structures. Based on self-avoiding random walks of chain conformations on lattice models,

Lucas and Dill, and Kopeikin and Chen developed theoretical models of pseudoknots and simple RNA tertiary folds [30–32] . These models have been used in the studies of folding stability, thermal transitions, and general shape of the free energy landscape of RNA folding. A limitation of these models was that they were lattice-based and cannot represent realistic RNA conformations. The first *ab-initio* free energy model for realistic H-type pseudoknots that was easy to implement was the Vfold model developed by Cao and Chen [33–35]. Since the loop conformation depends on the nearby helical stems through chain connectivity and stem-loop volume exclusion, a template was first constructed from experimentally measured atomic coordinates of RNA stem structures. The number of conformations of loops was then enumerated by generating self-avoiding walks on a diamond lattice that connect the stem-ends. An important development in these works was the three-dimensional templates created for estimating the loop entropy of pseudoknotted RNA structures, as they already contain rich information. The Vfold method worked very well for studying RNA thermodynamics and for RNA structure predictions [33–35]. However, although it is not difficult to build templates for H-type pseudoknots and create a look-up table for entropic costs of forming stem-loop structures with different lengths, it is not feasible to do so for all possible pseudoknots.

It is our goal here to go beyond previous approaches and develop a general framework for computing free energies of RNA molecules with arbitrary secondary structures, including those with complex pseudoknots. Our approach is based on considerations of the spatial nature of RNA molecules, and is not restricted to any specific type of pseudoknot, such as the H-type pseudoknot, but is applicable to all types of pseudoknots. Physically, the entropy of a loop of a specific length is determined to a large extent by the end-to-end distance and the spatial interference from nearby stems or loops. This is true for loops of all nature, regardless whether it is a hairpin loop, an internal or bulge loop, a multi-branch loop, or a pseudoknotted loop [31, 32, 35, 36]. From this consideration and our previous work [36], we have developed an efficient and accurate method to calculate the loop entropy of RNA structures with pseudoknots. By first growing multiple RNA chains in three-dimensional space, our method searches among all possible arrangements of helical stems for the optimal three-dimensional structure. The loop entropy of RNA structures in each spatial arrangement of helical stems is then computed by accurately estimating the fraction of number of loop conformations with respect to number of random coils of the same length based on a 6-state discrete RNA chain model [36]. We call our method pk3D (for PseudoKnot predictor in 3D). Our method is feasible because the constraints from chain connectivity and the avoidance of geometric collisions allow early pruning of a vast number of unlikely spatial arrangements, which occur early in the branches of a search tree. As a result, the actual number of spatial arrangement is relatively small.

The pk3D method takes secondary structure candidates of given RNA sequences as input and computes their free energies using more realistic physical methods; it also outputs the approximate shapes of the corresponding three-dimensional structures. We note that our pk3D and previous methods for predicting RNA tertiary structure such as the MC-Fold and MC-Sym pipeline [37] are different: First, the MC-Fold and MC-Sym pipeline use a statistical potential function, whereas ours is a physics-based potential function; Second, our focus is not to predict exact tertiary structures using pk3D; instead, we aim to develop the pk3D method for constructing approximate three-dimensional shapes of a given secondary structure, and for accurate calculation of its free energy. The approximate three-dimensional shapes generated by the pk3D method can be further used by additional structural refinement methods to obtain accurate and more detailed conformations; Third, the pk3D method is computationally much faster than the MC-Fold and MC-Sym pipeline, which is essential for large scale studies of RNA molecules.

Since the pk3D method is designed to treat pseudoknot of any complexity, we need a candidate list as input that maximizes the diversity of pseudoknot topology. All previous methods for prediction of secondary structure were not particularly designed for this purpose. Therefore we have developed a method, called pk2D (for PseudoKnot predictor in 2D), to create candidate secondary structures of RNAs for given sequences. The pk2D program first uses dynamic programming based local alignment to create a pool of helical stems, and then employs an approximation algorithm that can identify a large number of close-to-optimal solutions of stem combinations without conflicts, which is a well-known NP-complete problem [22, 23]. Details of both pk2D and pk3D are given in the section of Materials and Methods.

The remainder of this paper is organized as follows: First, we discuss the performance of pk2D, namely, its ability to generate good candidate list for further evaluation; Second, we then test the ability of pk3D in selecting native structures from decoy structures; Third, we make blind predictions and compare pk3D with several widely used programs which can predict RNA secondary structures with pseudoknots; Finally, in the last part of the Results and Discussions section, we show with several examples the performance of pk3D in generating approximate three dimensional shapes of the predicted RNA structures. We summarize the paper with the Conclusion section. The data set and detailed algorithms are given in the Materials and Methods section, and can be downloaded from `gila.bioengr.uic.edu/lab/tools/pk3d/`.

II. RESULTS AND DISCUSSIONS

Our overall goal is to develop a new approach to assess free-energy of loops in pseudoknotted RNA molecules and to generate coarse three-dimensional structures. We first examine how this method can aid in prediction of RNA secondary structures with pseudoknots.

A. Generating candidate secondary structures with pseudoknots: Performance of pk2D

The accuracy of our secondary structure prediction method depends on the quality of the candidate list, which in this work is created by a specially designed program called pk2D. pk2D can generate a long list of candidate secondary structures with enriched diversity in pseudoknot topology.

To assess the quality of the candidate structures created by pk2D and to explore the optimal number of candidate structures for prediction of correct spatial fold, we first test the performance of pk2D. For each sequence in a testing set of a total of 43 pseudoknotted RNA molecules (see the Materials and Methods section), we create a pool of stems using local alignment through dynamic programming. These stems are then processed by the pk2D program, which generates secondary structures consisting of stems without conflicts. We take the top m structures from these secondary structures, which are ranked by the sum of stem free energies using the Turner's rule. Note that in pk2D, the free energy contribution from loops is neglected temporarily, as it will be treated comprehensively using a physical framework in pk3D. We find that on average, more than 95% of the structures in the candidate list generated by pK2D contain at least one pseudoknot. We then compare the candidates with the true known native structure, and calculate the sensitivity and the positive predictive value (PPV) of each candidate. The structure closest to the true native structure is then identified (see the section of Materials and Methods).

On average, the best candidate is ranked 23rd by pk2D. The average sensitivity of the best structure in the candidate list for this set of RNA molecules are 0.92, 0.93, 0.93, 0.95 for $m = 30$, 60, 100 and 500, respectively, and the PPV of the best structure are 0.85, 0.86, 0.87, 0.90, respectively. These results show that the quality of the candidate list is adequate even though only the stem free energy is accounted for in pk2D. For further structure prediction, we find a candidate list of size $30 - 50$ is sufficient.

B. Selecting native structure from candidate set by free energy with improved loop entropy method: Performance of pK3D

The free energy of an RNA loops, especially that in pseudoknots, is difficult to evaluated. In pk3D, the free energy of loops is estimated based on a physical model on the assumption that the loop entropy is determined to a large extent by the end-to-end distance and the spatial interference of nearby stems. This assumption is reasonable and is applicable to all nested and pseudoknotted loops, regardless of its complexity. It provides a unifying framework for the treatment of loop entropy. Details of the pk3D method are given in the section of Materials and Methods.

We first test the ability of pk3D in selecting native secondary structures out of the other candidate structures (called decoys). For each of the 43 sequences in the data set, we take the top m secondary structures output by pk2D as decoys. These decoys have diverse structures, about 95% of them are pseudoknotted. We then manually insert the native structure into this list and use the Turner's energy rule and pk3D respectively to further evaluate the free energies. When using the Turner's energy rule, we use the Dirks and Pierce extension for calculating the loop entropy of pseudoknots [13]. The optimal structure with the lowest free energy in each case is then compared with the native structure, and the corresponding sensitivity and PPV are calculated. The results averaged over the whole test set are shown in Table I.

As shown in Table I, the performance in selecting native structures from decoys for this test set is slightly better using pk3D ($\sim$ 2 percent improvement) than using the Turner's energy rule. Although the improvement is modest, this result is promising, considering that the current form of free energy rule in pk3D is very simple, as the entropy of both nested and pseudoknotted loops are indexed by only two parameters, *i.e.*, the loop length and the end-to-end distance. Since pk3D is intrinsically a spatial method, it can easily incorporate more complex factors such as the docking of loops onto nearby helices, which would be impossible for methods based on secondary structures.

C. Secondary structure prediction: Comparing pk3D with other methods

In this test, we make a blind secondary structure prediction using pk3D and compare its performance with those of other methods, including HotKnots [19], pknotsRG [14], NUPACK [13], and ILM [16]. For each given RNA sequence, we first create the pool of stem regions, which is used to generate m number of candidate secondary structures by pk2D ranked by the sum of stem energy. We then apply the pk3D algorithm to further

evaluate the free energies of these candidates. The secondary structure with the lowest free energy is predicted to be the native one.

The overall accuracy of our prediction depends on the size of the input candidate list for pk3D. It has been shown that by carefully calibrating the size of sampled space, the accuracy of secondary structure prediction can be improved [20]. Here the evaluation of the free energy stops when pk3D finds 30 "valid" secondary structures, or else finishes evaluation of all of the top $m = 500$ candidates. The secondary structure is assumed to be "valid" if a corresponding three-dimensional structure satisfying all chain constraints and free of geometric collisions is found. This strategy of selecting proper candidate size was tested and was found to give the best performance for the current set of RNA sequences.

The optimal structure with the lowest free energy is then compared with the true native structure, and the corresponding sensitivity and PPV are calculated. The prediction results by pk3D and by several other methods are listed in Table II. Overall, pk3D gives the best sensitivity and comparable PPV. On average, the pk3D method has a sensitivity $6 - 7\%$ higher than that of pknotsRG and NUPACK, and 10% higher than that of HotKnots. In terms of PPV, HotKnots and pknotsRG have the best performance, although pk3D's PPV is quite comparable. In both cases, ILM has poor performance, possibly because we were unable to supply the best parameters to the ILM algorithm.

We also compare our results with a recently published new model (Vfold) for predicting structures of general H-type pseudoknots with inter-helix loops [35]. The Vfold model computes the conformational entropy and folding free-energy based on complete conformational ensemble and rigorous treatment for the excluded volume effects. In a test for 18 H-type pseudoknots, the model gave an average value of 0.91 for both sensitivity and PPV, about 5 percent higher than the other methods including Hotknots, ILM, pknotsRE, STAR, pknots-RG and NUPACK. The 18-pseudoknot testing set used is a subset of what is used in this study (Table II). We calculate the performance of pk3D on this subset of 18-pseudoknotted RNA molecules and find that the average sensitivity and PPV are 0.89 and 0.84, respectively. Our result lags behind that of Vfold slightly, although this is expected, as our method is developed for general pseudoknots of arbitrary complexity, whereas the Vfold method is currently restricted to H-type pseudoknots. It is interesting to note that Vfold and pk3D perform very similarly on this test set, with the difference mostly from Hs-PrP. For Hs-PrP, Vfold gives a sensitivity of 0.45 and a PPV of 0.5, whereas pk3D fails and gives two zeroes. For another RNA molecule, Bt-PrP, Vfold and pk3D give the same sensitivity (0.42 for both) and very similar PPV (0.33 versus 0.31). Except these two cases, both methods give very high sensitivity and PPV (usually close to 1.0) for the remaining pseudoknotted RNAs . The similar performance of Vfold and pk3D is understandable, since they calculate the loop entropy based on the same physical consideration, *i.e.*, the loop entropy is determined primarily by the loop length, the end-to-end distance, and the interference from nearby structures.

We discuss in the following sections details of free energy evaluation of pk3D using several specific examples.

tmRNA-Ec-PK4: Importance of spatial arrangement of stems

The RNA molecule tmRNA-Ec-PK4 contains a H-type pseudoknot. However, it has in addition a 1×1 internal loop embedded within each of its two stems (Fig. 1). As shown in Table II, pk3D predicts exactly the true native structure, with both sensitivity and PPV value of 1.00. The prediction by NUPACK is also at 100% accuracy, but HotKnots, pknotsRG and ILM miss more than 1/3 of the base pairs, with the sensitivity of prediction at 0.68 and a PPV ranging between 0.81 and 1.00.

We have examined all of the 500 candidate structures generated by pk2D and found that more than 95% of them contain pseudoknots. The native structure is captured automatically by pk2D and is within this candidate list. All of the 499 candidates other than the native structure are easily recognized by pk3D as spatially infeasible, and are therefore ruled out immediately. Only the true native secondary structure has a feasible spatial arrangement with a reasonable free energy. In this case, the prediction of the native structure is simple, as it is sufficient to examine spatial feasibility of stems and there is no need to calculate the loop free energy in detail. This example illustrates the important role of the spatial arrangement of double helices and its geometric constraints in reduction of the feasible space of RNA structures.

mRNA-Hs-PrP: Importance of candidate structures

The native structure of this sequence is a simple H-type pseudoknot (Fig 2a). However, despite of the simplicity of its native structure, all of the tested methods except ILM fail in this test, with 0.0 in both sensitivity and PPV. The ILM method also gives a poor prediction with very low sensitivity (0.36) and PPV (0.25).

For pk3D, the problem is that the native structure is not among the list of candidates, and all of the 500 candidates are readily rejected as spatially infeasible, since the specific combination of stem and loop lengths in each cannot be satisfied without violation of geometric constraints when modelled in three dimensional space. Even the best candidate closest to the native structure has an uncommon pseudoknot and over-estimates the number of base-pairings (about 35% over-estimation, Fig. 2b). The important fact is, when the native structure is manually inserted into the candidate list, it is correctly identified by pk3D as the only spatially feasible secondary struc-

ture, with a reasonable free energy.

For both tmRNA-Ec-PK4 and mRNA-Hs-PrP, when the native structures are present in the candidate lists, they are both selected as the only spatially feasible secondary structures among all candidates. The blind prediction by pk3D of tmRNA-Ec-PK4 is successful but not mRNA-Hs-PrP. Here the bottle-neck is not free-energy estimation of loops in pseudoknotted RNAs, rather, it is the generation of a candidate list that includes the native structure. These two examples show that the pk3D algorithm is sensitive to the quality of the candidate list, and the improvement in its generation has the promise to significantly increase the overall accuracy of pk3D prediction.

TMV.R and HDV: More complex structures

The native structure of the 3' terminal region of the tobacco mosaic virus RNA (TMV.R) contains two pseudoknots, one is a simple H-type, and the other is a H-type with a long hairpin and a 1×5 internal loop embedded within (Fig. 3a). The structure of hepatitis delta virus genomic ribozyme (HDV) is an H-type pseudoknot with two embedded substructures, a simple hairpin and an imperfect hairpin with a bulge (Fig. 3b). pk3D performs similarly for TMV.R and HDV as in the previous two cases, that is, all candidates generated by pk2D are deemed as infeasible, as the native structures are not captured by pk2D thus not included in the candidate list. When the native structures of TMV.R and HDV are inserted manually into the corresponding candidate list, both are found as the only spatially feasible structures and correct predictions are made, with 100% prediction accuracy in both cases.

D. RNA spatial arrangement and three-dimensional structure prediction

The approach of pk3D towards secondary structure prediction is spatial in nature. First, stem regions with some stability are combinatorically assembled, second, spatial considerations are enforced in the form of loop entropy estimation, and the vast majority of candidate structures with stems compatible by secondary structure but spatially infeasible are eliminated. Here the loop entropy calculation in pk3D is fundamentally different from that in the Tuner's energy rule. As we are only at the very beginning of understanding the governing principles of RNA three-dimensional structures, it is premature to adopt Turner-style empirical rules and inventing additional phenomenological equations. Instead, pk3D builds spatial models of stem regions, rejects infeasible candidates, and searches among feasible arrangements of stems for the optimal one and numerically estimates the loop entropy by calculating the fraction of closed loops with respect to random coils of the same length based on the sequential Monte Carlo algorithm [36]. With this strategy, each feasible secondary structure will be automatically assigned a representative spatial arrangement of helices, containing coarse grain information of its tertiary structure. In essence, pk3D is a secondary structure predictor that also generates tertiary information, even though it only gives a coarse-grained shape of the three-dimensional conformation of the RNA molecule, due to the discrete nature of the state model used in the algorithm. This coarse-grained spatial conformation is still very useful: not only it makes it possible to estimate the pseudoknotted loop entropy, but also can serve as a starting point for further structural refinement.

Here we use several examples to describe how pk3D predicts the general shape of the native conformations. These examples are not selected from the 43-sequence testing set, since it is difficult to find the corresponding Protein Data Bank (PDB) structures from these sequences to compare our predictions with. Instead, we directly obtain from the PDB data bank several typical RNA structures, including nested and pseudoknotted with varying complexity, extract the sequences, and use the combination of pk2D and pk3D to predict their native structures.

Predicting the spatial structure of a H-type pseudoknot within the gene 32 mRNA of bacteriophage T2

The PDB structure 2TPK contains a simple H-type pseudoknot within the gene 32 mRNA of bacteriophage T2. Figs. 4a and 4b show its native secondary and tertiary structures as derived from NMR experiments, respectively. Its secondary structure is predicted by pk3D correctly with an accuracy of 1.0. The predicted tertiary structure (Fig 4c) shares the major common structural features with the native structure. These include: 1) the two helices are co-axially stacked on each other and run continuously in space, forming a long quasi-continuous helix, which helps to stabilize the overall RNA structure; and 2) the nucleotide A8 constitutes a loop of only one nucleotide, which connects the far two ends of the two helices, rendering a typical H-type pseudoknot structure.

Similar results are obtained using pk3D for the pseudoknot of SRV-1 RNA involved in ribosomal frameshifting (PDB 1E95), and the P2B-P3 pseudoknot from human telomerase RNA (PDB 1YMO). Both are simple H-type pseudoknots but have different loop lengths (these figures are similar and therefore not shown).

The acceptor arm of TYMV tRNA-like structure

The pseudoknotted T arm and acceptor arm of the tRNA-like structure of turnip yellow mosaic virus (TYMV, PDB 1A60) is a combination of a H-type pseudoknot and a coaxially stacked hairpin structure. Its sec-

ondary structure is predicted correctly by pk3D with an accuracy of 1.0. By comparing the NMR measured structure with the predicted spatial structure (Fig 5b and 5c, respectively), it can be seen that the overall spatial relationship between three helices is the same: the two pseudoknotted helices are co-axially stacked on each other, on the top of them is also stacked the hairpin helix. These three helices assemble into a very long quasi-continuous helix in both measured and predicted spatial structures.

PDB 1S9S-A core encapsidation signaling RNA of the Moloney Murine Leukemia virus

The 101-nucleotide molecule of core encapsidation signal of the moloney murine leukemia virus (PDB 1S9S) is important for efficient genome packaging. It is a non-pseudoknotted RNA containing six helices. Among these helices, five are co-axially stacked and form a single long quasi-continuous helix. The first helix, on the other hand, is flexible and is connected to the other helices via a flexible five-nucleotide loop. The secondary structure, the first spatial model of the NMR structure, and the structure predicted by pk3D are shown in Fig 6a, 6b, and 6c, respectively.

The spatial arrangement of the last five helices in the measured and predicted structures are in general agreement. The position of the first helix in the predicted structure is different from that of the NMR measured model. This is due to the highly flexible nature of the loop that connects the first helix with the rest of the molecule. In fact, NMR measurement shows that the first helix has very flexible positions, and the different models of the structure as deposited in the PDB databank show that this helix can be located in a wide range of positions with respect to the other helices [38]. The spatial position of the first helix predicted by pk3D is well within the experimentally measured range.

The orientation of the fifth helix (colored in magenta) is different for the experimentally measured and computationally predicted structures. In the NMR-measured structure, this helix is loosely stacked on the fourth helix (colored in yellow), and the central axes of these two helices form a large angle, apparently due to the intervening 4-nucleotide bulge (G62 to A65, Fig 6a and 6b) between these two helices. In contrast, in the predicted structure, the fifth and the fourth helices are co-axially stacked closely together. This discrepancy is due to the fact that there is a competition between the favorable co-axial stacking energy bonus and the loop entropy. A flush co-axially stacked structure is favored by the stacking energy bonus, whereas the loosely stacked structure as given by the NMR experiment is consistent with the larger entropy of the bulge loop. This disparity between measured and calculated structures indicates the need to model accurately the delicate balance between these two energetic factors. Our algorithm will likely improve once the balance between these two factors is fine tuned.

Protein S15 binding fragment of 16S rRNA

In bacterial ribosomes, protein S15 binds to 16S rRNA and forms a key element required by the assembly of the small subunit of ribosome. This element is also important for the inter-subunit association. The relevant rRNA fragment is a non-pseudoknotted structure with five helices (Fig 7a, PDB 1DK1). An interesting feature of this structure is that it has a three-way junction, which is constrained by a conserved base triple and the associated stacking interactions, and is locked into place by magnesium ions and side chains from bound protein (Fig. 7b) [39].

The predicted spatial structure of this RNA molecule reproduces the spatial positions of all helices correctly, except helix-1 and helix-2 (colored in magenta and green, respectively, Fig 7c). The position of helix-1 in the predicted structure is only roughly correct, whereas helix-2 locates far away from its correct position. This discrepancy is not surprising, since the junction region involving helix-1 and helix-2 is stabilized by the combination of a base triple, magnesium ions, and bound protein [39]. In addition, the chain is locally parallel in the junction region [39]. None of these factors are accounted for explicitly in our model, nor in any other RNA models used in existing RNA structure predictors.

Hepatitis delta virus (HDV) ribozyme precursor

The hepatitis delta virus (HDV) ribosome precursor is among the most complicated pseudoknotted RNA molecules discovered so far. Its secondary structure and three-dimensional structure obtained from X-ray crystallography (PDB 1SJ3) are shown in Fig 8a and 8b, respectively. Its tertiary structure contains two long quasi-continuous double helices. The first consists of two coaxially stacked short helices, helices P1 and P1.1. The second also contains two coaxially stacked helices, P2 and P3. The two long quasi-continuous helices are parallel to each other, and each resembles locally a H-type pseudoknotted structure. That is, each appears as a long quasi-continuous pseudoknotted helix formed by one long continuous strand and two separate shorter strands. In addition, a hairpin structure (P4) is loosely stacked on helix P1.1, rendering an extended long helix (P1+P1.1+P4) [40].

The pk3D correctly selects the native secondary structure as the optimal one out of the top 100 candidates and at the same time predicts the spatial arrangement of the helices and the loops. Note that in this example, pk2D fails to find the native secondary structure automatically as a candidate. Therefore we manually insert the native secondary structure into the candidate list created by pk2D, to test the ability of pk3D in identifying the native structure out of decoys and the ability to predict

three dimensional shapes. The overall predicted tertiary conformation is shown in Fig 8c. It can be seen that the structure generated by pk3D contains most features of the X-ray structure: the two quasi-continuous long helices are reproduced, and they consist of the same two shorter coaxially stacked helices as in the experimentally determined structure (P1+P1.1, P2+P3, respectively); plus, these two long helices are parallel to each other, also consistent with experiments.

There are subtle but important structural features of this RNA molecule that are correctly predicted by pk3D. In HDV ribozyme precursor, there are three 0-length loops in the secondary structure (see the Materials and Methods section for the detailed definition of 0-length loops). Among these, loop-1 connects helices P1 and P1.1, loop-2,3 connects helices P2 and P3, and loop-1,3 connects helices P1 and P3 (Fig 8a). Pk3D correctly selects the first two interfaces and coaxially stacks the corresponding helices; it also correctly leaves the interface around loop-1,3 alone without stacking helices P1 and P3. The overall result is the formation of two long quasi-continuous helices (P1+P1.1, and P2+P3) connected by a simple loop-1,3 of zero length (Fig 8c), in excellent agreement with the X-ray structure.

The two nucleotides (C21 and C22) within the hairpin loop of helix P3 extend out and form base pairs with G38 and G39, thus forming the helix P1.1. It is likely that this interaction is important to hold the overall RNA structure together. Pk3D program correctly reproduces this structural feature as well.

There is a structural aspect that pk3D does not predict correctly, i.e., the orientation of helix P4. According to the X-ray structure, helix P4 is positioned on the top of helix P1.1, stabilized by a base triple (G61-C44-C41), a non-canonical base-pair (A43-G62), and the nearby metal ion [40]. The existence of base triples, non-canonical base-pairs, and interactions between nucleotides and metal ions are not considered explicitly in our current model, therefore this failure is not surprising. In pk3D, the connection between P4 and P1.1, and between P4 and P2 are modeled as a simple 4-nucleotide loop and a 5-nucleotide loop, respectively (Fig 8a). These two loops lead to much freedom for positioning and orienting helix P4.

Summary of the performance of pk3D in tertiary structure prediction

Overall, the pk3D algorithm can be used to generate approximate spatial arrangements of helices and loops, in addition to estimating free energies of loops and predicting native secondary structures. The overall tertiary shape is often very similar to the native structure of RNA molecule, regardless whether pseudoknots are present or not. It works especially well when the involved loops are short. These short loops impose significant constraints on the number of feasible conformations, as seen in the predicted structure of the hepatitis delta virus ribozyme precursor (PDB 1SJ3). Nevertheless, when the tertiary structure involves significant contributions from non-regular elements such as base-triples, non-canonical base pairs, metal ions, or bound proteins, the pk3D algorithm usually fails to produce an accurate three-dimensional structure. In fact, these complex factors present great challenges to all current efforts in predicting RNA structures.

Although in such cases the specific positions and orientations of the helices involved are inaccurate, the general arrangement of the overall structures may still be correct, as is the case of the core encapsidation signaling RNA (1S9S), protein S15 binding fragment of 16S rRNA (1DK1), and the hepatitis delta virus ribozyme precursor (1SJ3). Although the three dimensional shapes provided by pk3D are approximate in nature, they can be fairly close-to-native. These structures can be very useful, for example, in providing the initial seed conformations for further structure refinement. This task can be performed by using all-atom MD simulation packages such as AMBER, Charmm, or Gromacs. It is expected that these close-to-native conformations would lead to significant speed-up in structure predictions.

E. Timing information

For the testing set of 43 small RNA pseudoknots used in this study, the computation time used by pk3D for searching optimal arrangement for each candidate secondary structure usually finishes within $10^{-3} - 10^{-2}$ sec on an AMD Opteron-256 CPU. However, the time complexity of the pk3D algorithm is not directly related to the length of the RNA chain. Rather, it is determined by the total length of the loops in the "link", as defined in the section Materials and Methods. For the structures with many long links, the computation time may take up to several tens of seconds to finish. This limits the current version of pk3D program to small RNAs (< 150 nucleotides), since longer chains are likely to have larger links. We expect that this problem can be solved by dividing a large "link" into several small segments based on the loop length pattern and treating them separately. These structural fragments are inner-connected only by short loops, hence are more rigid and likely to fold separately. Further development of pk3D for long chain RNA molecules will be in our future work.

III. CONCLUSION

Calculating the free energy of RNA loops, especially that of pseudoknotted loops, is an important unsolved problem. We have developed a novel method called pk3D to address this problem. Our method is based on

the physical consideration that the entropy of an RNA loop is largely determined by its loop length, the end-to-end distance of the helices connected by the loop, and the steric interference from nearby helices. To calculate the loop entropy, our method searches among all possible spatial arrangements of helical stems for the optimal structure, and then estimates the number of loop conformations for the optimal structure using a 6-state discrete model and the Sequential Monte Carlo method. The excluded volume effect is explicitly treated by the algorithm. Our method treats both nested and pseudoknotted loops within a unifying physical framework, regardless how complex the pseudoknot might be.

We have also developed the pk2D method, which finds approximately optimal combination of low energy non-conflicting helical stems for a given sequence. The list of secondary structures created by pk2D are used as input to pk3D for further free energy evaluation. The spatial arrangement of stems and loops with the lowest free energy is then predicted to be the native conformation of the RNA sequence.

We have tested the performance of pk2D/3D on a data set of 43 small RNA molecules with pseudoknots. The quality of the list of candidate secondary structures created by pk2D is good: on average, the best candidate among the list has a sensitivity higher than 90% and a PPV close to 90% in terms of base-pairs. In this test, the ability for pk3D to select the native structure from a large number of decoys is slightly better than that of the extended Turner's free energy rule. In the blind test of predicting the secondary structures of these 43 RNA molecules from sequences, pk3D is found to have the best prediction results in terms of sensitivity and comparable PPV of correctly predicted base-pairs, when compared with several existing pseudoknots prediction methods.

Perhaps the most important contribution of this work is that pk3D can frequently produce generally useful coarse three-dimensional model of the native RNA structure. We found that the arrangement of stems and loops in three-dimensional space is generally similar to that of the native structure. This rough three-dimensional model is useful as a starting point for further structural refinement, as it provides a close-to-native physical structure. It is expected that refinements starting from pK3D predicted structures will benefit significantly as a result of accelerated folding and packing.

The weak point of the current version of pk3D is that it does not perform well for long sequences, since the computation time is determined by the total length of the loops in the "links", whose size is likely to increase with the chain length. However, as discussed earlier, this problem can be solved by dividing the links into several smaller segments or domains. With this simplification, the algorithm can in addition incorporate aspects of kinetic folding of RNA molecules, which is likely to be very important for large RNAs [41].

IV. MATERIALS AND METHODS

A. Definition of pseudoknots and H-type pseudoknots

An RNA structure is called pseudoknotted if it contains interleaved stem regions. Formally, if we denote a base pair as an ordered pair of positions of upstream and downstream position (i, j), where $i < j$, a structure is non-pseudoknotted if and only if for all pairs (i, j) and (k, l), nowhere the relationship $i < k < j < l$ holds; otherwise, the structure is called pseudoknotted [19].

We also need to define the H-type pseudoknot explicitly to facilitate discussions in the text. H-type pseudoknot in this study is referred to as a structure formed by base-pairing between a hairpin loop and the exterior loop of another hairpin. It consists of two helical stems and two loops as well as a possible third loop/junction that connects the two helical stems [35]. An H-type pseudoknot may also contain embedded sub-structures such as hairpins, internal loops, bulges or their combinations. Here the H-type pseudoknot is solely defined in terms of its secondary structure, regardless whether the involved helical stems form quasi-continuous helices in three-dimensional space. Our definition of the H-type pseudoknot is more general than that used in [35]. However, it should be emphasized that the pk2D and pk3D algorithms are not restricted to H-type pseudoknots and their more general versions. In our study, pseudoknots of arbitrary complexity are treated in one unifying physical framework.

B. Data set

All of our testings are based on a data set of 43 pseudoknotted structures unless indicated otherwise. They are taken from the 50 small pseudoknotted RNAs in reference [17], with 7 sequences removed. These are removed because all of them have a long unstructured $5'$ loop of length > 13 nucleotides, which is unstable and can easily form internal structures within itself, or with another part of the RNA molecule. The corresponding native structures resolved in the databank may exist as a result of additional interactions with proteins or other molecules to form stable complexes. Therefore, it is not appropriate to include these sequences in testing the performance of structure prediction of single chain RNAs. These 7 sequences are: frameshifting-EIAV, frameshifting-PLRV-S, viral-tRNA-like-APLV, viral-tRNA-like-CGMMV, viral-tRNA-like-ORSV-S1, viral-tRNA-like-OYMV, and viral-tRNA-like-SBWMV1.

C. Generating the pool of stem regions

For a given sequence S with nucleotides ordered from $5'$ to $3'$, we use the Smith-Waterman dynamic programming algorithm for local alignment to align it with its reverse sequence S', with the same nucleotides ordered from $3'$ to $5'$. Both Watson-Crick and wobble base pairings are considered matches. This generates all possible stems with energy scores below a given threshold.

D. Generate candidate secondary structures with pk2D

All of the computed stable stems are processed by the algorithm pk2D, which finds solutions of multiple compatible stems with overall low energies. These solutions provided by pk2D form a list of candidate secondary structures, each containing several non-conflicting stems. That is, none nucleotide in any stem appears in another stem of the same secondary structure. Finding non-conflicting stems with overall lowest energy is a known difficult NP-complete problem [22, 23]. Numerous methods have been developed to address this challenging problem, including dynamic programming based methods [13–15], heuristic methods [16, 19, 20], Monte Carlo method [21], methods based on genetic algorithm [18], and the strategy that reformulates this problem into a combinatorial graph problem and solving it using the technique of tree decomposition [17].

Here we use an approximation algorithm to generate candidate secondary structures. The problem of finding all consistent sets of candidate stems can be formulated as an Integer Programming (IP) problem. The objective of the IP problem is to minimize the sum of the free energy contributed by each selected stem. The constraints of the IP problem are formulated to encode the non-conflicting condition between the selected stems. We have adapted a method originally developed for protein structural alignment in [42] to solve this IP problem. This approach is based on an approximation algorithm for scheduling split interval graphs [43].

An approximate solution of the IP problem can be found by an iterative process, where the IP problem is first relaxed into a Linear Programming (LP) problem and solved using the commercial LP package BPMPD [44]. A conflict graph $G = (V, E)$ is then constructed where vertex v_i represents stem i. An edge $e_{i,j}$ is drawn between two vertices v_i and v_j if the two corresponding stems are in conflict. Each vertex is assigned three values. The first value is the free energy of the stem σ_i. The second value x_i is equal to the corresponding output from the LP solution. The third value is called the *local conflict number:*

$$\alpha(i) = \sum_{\{v_j|(i,j)\in E\}\cup v_i} x_j.$$

The vertex k with the minimum local conflict number α_k is then identified. A new free energy score $\sigma_i^{new} = \sigma_i - \sigma_k$ is then assigned to v_k and to all vertices that share an edge with v_k. After this update, all vertices with $\sigma \leq 0$ are pushed onto a stack S, and are removed from further consideration. A new LP problem is subsequently formulated using the remaining vertices. This process is iterated until all vertices have been pushed onto the stack.

Vertices on the stack are then continuously popped. The first popped vertex forms the candidate set of compatible stems. A subsequently popped vertex is then inserted into each existing candidate set if it does not cause conflict with stems already in the set. In addition, a new candidate set is formed with the currently popped vertex as the sole member. This process is repeated until all vertices are popped from the stack.

The above procedure is another version of the maximum weight independent set problem [42, 45]. The set created by the first popped vertex is guaranteed to be within a factor of 1/2.89 from the optimal solution.

The candidate secondary structures generated by pk2D are ranked by the sum of stem energies, estimated by the standard Turner's energy rules. Note that the free energy contributions from loops are neglected temporarily at this stage, as they will be treated within a physical framework in pk3D. The top 500 secondary structures serve as the set of candidates for more accurate further free energy estimation.

E. Generating RNA conformations by growth using pk3D

The free energies of the secondary structures provided by pk2D are further estimated by incorporating loop entropies. Pk3D calculates the free energy for each candidate secondary structure based on the corresponding 3D conformations, which are generated automatically using a growth method. The procedures are as follows:

1) Finding the "link". For each secondary structure, its "link" is defined as the set of $n-1$ loops that connects all the n stems in the secondary structure, such that the overall length of the loops belonging to the "link" is the shortest. Recall that a stem may have as many as four loops. An example is shown in Fig. 10, which depicts the secondary structure of the hepatitis delta virus (HDV) ribozyme precursor and its corresponding graph representation. Note that in our method we have generalized the concept of loops to include the special case when two stems are directly connected; the junction between these two stems is called a 0-length loop. The

reason for this generalization is that, although its length is zero, it can lead to different relative spatial orientations between two involved helices, similar to the loops of non-zero length. The 0-length and non-zero-length loops are therefore treated in the same way in pk3D. In Fig. 10, the loops plotted as red thick curves define the "link", which contains three 0-length loops and one loop of length 4. The overall length of the loops contained in the "link" is therefore 4. The usage of "link" is inspired by the work described in [32].

Since the number of stems (or vertices in the graph representation) in the RNA pseudoknots is small (usually 4-8), we use a heuristic method to find the "link" for a given secondary structure. Although a heuristic method cannot guarantee an optimal solution, it works well in practice for small graphs. We first select the shortest loop, so the two stems at both ends are connected as one component. We then iteratively select the next shortest loop if it adds a new stem to the set(s) of components that have been generated so far. This is repeated until $n-1$ loops are chosen, and all components are merged into one. For the candidate secondary structures generated for this data set of 43 RNA molecules, n is typically between 4 and 8.

2) Generating three-dimensional structures. After the link is identified, we generate the corresponding three-dimensional structures or conformations of the RNA molecule. This is accomplished by sequentially enumerating all possible conformations for loops (typically with length < 7) in the link using a 6-state discrete model [36] and by spatially arranging the connected stems in all possible 6 orientations, followed by selecting those that are physically plausible (*e.g.*, without sharp turns). We start with the shortest loop. After spatially adding the connecting stem to this loop at all possible orientations (typically 2 or 4 feasible out of all 6), we continue by generating all possible conformations for the loop of the next shortest length.

During this process, whenever a new stem is added to a partial conformation, or when two partial components are merged, we examine if steric collisions occur and if the lengths of loops connecting this new stem to stems already added are sufficiently long to accommodate them spatially. For example, a loop length of 0 or 1 cannot extend in space to connect stems that are distant from each other. This examination usually rules out thousands of infeasible conformations .

This process is repeated until all stems are added and merged into one component. Altogether, we have theoretically $O[(\prod_i l_i^{\alpha}) \cdot K^{n-1}]$ number of spatial conformations, where l_i is the length of the i-th loop, and α is the scaling exponent for the number of loops with loop length l_i, which is estimated to be between 3 and 5 [36]. K is the number of the possible orientations of a stem when connected to an end of a loop. Typically, $K = 2$ or $K = 4$, and $n-1$ is the number of loops in the link. In practice, because of the strong constraints of the loops and excluded volume effect, the number of feasible conformations is substantially smaller than this bound. Among the candidate secondary structures, about 90% are found to be spatially infeasible after simple examinations, and are eliminated from further considerations at very early stages. The overall procedure of pk2D is shown in Fig. 9a.

3) Speeding-up with look-up tables. In the current implementation, the number of conformations for each loop is solely determined by the loop length. To improve the runtime efficiency, we pre-compute the conformations of loops with length between 0 and 7, recording the starting and end positions of each loop in a look-up table for loop-conformations. This table can be further improved by introducing sequence-dependent information, which will be in our future work.

F. Further free energy estimation

For each spatial conformation constructed by pk3D, its free energy is calculated as the sum of free energy of stems and loops. The sub-structures are divided into two groups and treated differently in pk3D. The first group contains stem, hairpin, bulge and short internal loops ($\leq$ 4 nucleotides). The free energies of these sub-structures are calculated straightforwardly using the Turner's energy rules. The second group consists of all the other loops, including the longer internal loops, all multibranch loops and all pseudoknotted loops, whose physical models are explicitly constructed by pk3D through sequential Monte Carlo sampling. Based on the examination of the top 500 candidates for each of 43 sequences used in this study, we find that $75-80\%$ loops belong to the second group. Simple hairpin, bulge and short internal loops only count for $\sim 20\%$. Therefore, the majority of the loops are treated by our new free energy calculation scheme. We also did another test in which both groups are treated by our new scheme. We found that the performance is still good, although there is a slight deterioration. The reason is that the loop entropy calculated by our new scheme is close to the empirical value from the Turner's energy rule, as shown in our previous works [36]. In the current implementation, we treat two groups differently.

The model for loop free energy calculation in pk3D is based on several physical assumptions. First, the enthalpy component is assumed to be zero. Second, as a first order approximation, the loop entropy is assumed to be determined by its length and the end-to-end distance, which is determined by the connected helices. The volume exclusion effect between monomers within a loop is also explicitly accounted for. To improve the runtime efficiency, an entropy table is pre-built, with each entry indexed by a loop length and an end-to-end distance. Therefore the entropy of a loop, regardless whether it belongs to a pseudoknotted or nested structure, is in-

dexed by these two parameters. The entropy value is calculated by estimating the fraction of number of conformations of closed loop over number of conformations of random coil of the same length using a 6-state discrete model through sequential Monte Carlo sampling. This approach is described in more detail in our previous publications [36, 46–49]. This strategy of calculating loop entropy treats both nested and pseudoknotted loops in an unified physical framework, regardless how complex the structures are.

The excluded volume effect between helices, and that between nucleic acids within one loop are accounted for explicitly. Nevertheless, the excluded volume effect between stems and loops are not fully considered. The only reason for this is to improve performance. Although it can be accounted for explicitly and rigorously in our model, its incorporation will significantly decrease the runtime efficiency, as it is difficult to pre-build a template and hence a look-up table for all different types of combinations of loops and stems, and therefore they have to be computed on the fly. In our current implementation, this excluded volume effect is partially represented by the sampled three-dimensional structural models when building the entropy look-up table, although detailed excluded volume effect specific to individual candidate secondary structure are not taken into account yet. The excluded volume effect between different loops are of minor contribution to the overall free energy, hence can be ignored without much consequence [35].

Coaxial stacking can make important contributions to the overall stability of RNA molecules. An advantage of pk3D is that this effect can be modeled in a straightforward fasion, since the spatial arrangements among stems are generated explicitly. Specifically, when the head of stem A is close to the end of stem B, we find the location of the center $\boldsymbol{c}_A$ and the normal vector $\boldsymbol{n}_A$ of the base plane at the head of stem A, and the similarly defined $\boldsymbol{c}_B$ and $\boldsymbol{n}_B$ for stem B. Define $\Delta\boldsymbol{c} = \boldsymbol{c}_B - \boldsymbol{c}_A$ as the inter-center vector pointing from $\boldsymbol{c}_A$ to $\boldsymbol{c}_B$. Stems A and B are coaxially stacked if the length of $\Delta\boldsymbol{c} < 6$ Å, and at the same time the angle between $\boldsymbol{n}_A$ and $\boldsymbol{n}_B$, between $\Delta\boldsymbol{c}$ and $\boldsymbol{n}_A$, and between $\Delta\boldsymbol{c}$ and $\boldsymbol{n}_B$ are all smaller than 30 degrees. These criteria are slightly different from that proposed by Tyagi and Mathews [50]. When two stems are found to be coaxially stacked, the stacking free energy bonus is calculated using the parameter from the Turner's rule as if the helix were uninterrupted.

For each secondary structure, pk3D generates many three-dimensionally feasible conformations and these conformations may have different stacking patterns in the loop/junction region. When multiple conformations are found, the conformation with the lowest estimated free energy is chosen as the representative three-dimensional structure, whose free energy is assigned to the corresponding secondary structure. The secondary structure in the candidate list with lowest assigned free energy is then predicted to be the native secondary structure of the RNA molecule, and the associated three-dimensional conformation is assumed to be the native conformation. This predicted tertiary conformation could serve as a good starting point for further structure refinement.

The overall procedure of pk3D is shown in Fig. 9b.

G. Calculation of the sensitivity and PPV

We use RP (real positive) to denote the number of base pairs in the real structure; TP (true positive) the number of correctly predicted base pairs; FP (false positive) the number of predicted base pairs that do not exist in the real structure. The sensitivity (SE) of the prediction of an algorithm is defined as TP/RP, and the positive predictive value (PPV) is defined as TP/(TP+FP), and the F value as $2 \cdot SE \cdot PPV/(SE + PPV)$ [2]. We use the term "accuracy" to summarize performance in both sensitivity and PPV. The terms "lowest SE" and "lowest PPV" are defined as the SE and PPV of the structure that has the lowest calculated free energy, respectively. The terms "best SE" and "best PPV" are the SE and PPV of the candidate that is closest to true native structure, respectively, with the "closeness" measured by the F value.

V. ACKNOWLEDGEMENTS

This work is supported by grants from the US National Science Foundation (DMS-0800183 and DMS-0800257), National Institute of Health (GM68958 and GM079804), and Office of Naval Research (N00014-06-1-0100). J.Z. and W.W. acknowledge the National Natural Science Foundation of China (90403120, 10504012, and 10704033) and the National Basic Research Program of China (2006CB910302). This research is supported in part by the Project of Knowledge Innovation Program (PKIP) of Chinese Academy of Sciences, Grant No. KJCX2.YW.W10. We also acknowledge Shanghai Supercomputer Center for providing computing resources.

[1] Gesteland, R.F., Cech, T.R., Atkins, J.F. (Eds.) The RNA world, 3rd ed. Cold Spring Harbo, NY: Cold Spring Harbor Press.

[2] Mathews, D.H., Sabina, J., Zuker, M., Turner, D.H.

1999. J. Mol. Biol. 288: 911-940.

[3] Gonzalez-Diaz, H., Aguero-Chapin, G., Varona, J., Molina, R., Delogu, G., Santana, L., Uriarte, E., Podda, G. 2007. 2D-RNA-coupling numbers: A new computational chemistry approach to link secondary structure topology with biological function. J. Comput. Chem. 28: 1049-1056.

[4] Brion, P., Westhof, E. 1997. Hierarchy and dynamics of RNA folding. Annv. Rev. Biophys. Biomol. Struct. 26: 113-137.

[5] Gan, H.H., Pasquali, S., Schlick, T. 2003. Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. Nucl. Acid. Res. 31: 2926-2943.

[6] Shapiro, B.A., Yingling, Y.G. Kasprzak, W., Bindewald, E. 2007. Bridging the gap in RNA structure prediction. Curr. Opin. Struct. Biol. 17: 157-165.

[7] Doshi, K.J., Cannone, J.J. Cobaugh, C.W., Gutell, R.R. 2004. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. BMC Bioinformatics 5: 105.

[8] Gardner, P.P., Giegerich, R. 2004. A comprehensive comparison of comparative RNA structure prediction approaches. BMC Bioinformatics 5: 140.

[9] Tinoco, I., Uhlenbeck, O.C., Levine, M.D. 1971. Estimation of Secondary Structure in Ribonucleic Acids. Nature 230: 363-367.

[10] Zuker, M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 31: 3406-3415.

[11] Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, S., Tacker, M., Schuster, P. 1994. Fast Folding and Comparison of RNA Secondary Structures. Monatshefte f. Chemie 125: 167-188.

[12] Mathews, D.H., Disney, M.D., Childs, J.L., Schroeder, S.J., Zuker, M., Turner, D.H. 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc. Natl. Acad. Sci. USA 101: 7287-7292.

[13] Dirks, R.M., Pierce, N.A. 2003. A partition function algorithm for nucleic acid secondary structure including pseudoknots. J. Comput. Chem. 24: 1664-1677.

[14] Rivas, E., Eddy, S.R. 1999. A dynamic programming algorithm for RNA structure prediction including pseudoknots. J. Mol. Biol. 285: 2053-2068.

[15] Reeder, J., Steffen, P., Giegerich, R. 2007. pknotsRG: RNA pseudoknot folding including near-optimal structures and sliding windows. Nucleic Acids Res. 35: W320-W324.

[16] Ruan, J. Stormo, G.D., Zhang, W. 2004. An interated loop matching approach to the prediction of RNA secondary structures with pseudoknots. Bioinformatics 20: 58-66.

[17] Zhao, J., Malmberg, R.L., Cai, L. 2008. Rapid ab initio prediction of RNA pseudoknots via graph tree decomposition. J. Math. Biol. 56: 145-159.

[18] van Batenburg, F.H.D., Gultyaev, A.P., Pleij, C.W.A. 1995. An APL-programmed genetic algorithm for the prediction of RNA secondary structure. J. Theor. Biol. 174: 269-280.

[19] Ren, J., Rastegari, B., Condon, A., Hoos, H.H. 2005. HotKnots: Heuristic prediction of RNA secondary structures including pseudoknots. RNA 11: 1494-1504.

[20] Chen, X., He, S., Bu, D., Zhang, F., Wang, Z., Chen, R., Gao, W. 2008. FlexStem: improving predictions of RNA secondary structures with pseudoknots by reducing the search space. Bioinformatics 24: 1994-2001.

[21] Isambert, H., Siggia, E.D. 2000. Modeling RNA folding paths with pseudoknots: Application to hepatitis delta virus ribozyme. Proc. Natl. Acad. Sci. USA 97: 6515-6520.

[22] Lyngsø, R.B., Pedersen, C.N.S. 2000. RNA pseudoknot prediction in energy-based models. J. Comput. Biol. 7: 409-427.

[23] Akutsu, T. 2000. Dynamic programming algorithms for RNA secodnary structure prediction with pseudoknots. Discrete Appl. Math. 104: 45-52.

[24] Serra, M.J., Turner, D.H., and Freier, S.M. 1995. Predicting thermodynamic properties of RNA. Meth. Enzymol. 259: 243-261.

[25] Xia, T., SantaLucia Jr, J., Burkard, M.E., Kierzek, R., Schroeder, S.J., Jiao, X., Cox, C., Turner, D.H. 1998. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry 37: 14719-14735.

[26] Gultyaev, A.P., van Batenburg, F.H., Pleij, C.W. 1999. An approximation of loop free energy values of RNA H-pseudoknots. RNA 5: 609-617.

[27] Aalberts, D.P., Hodas, N.O. 2005. Asymmetry in RNA pseudoknots: observation and theory. Nucleic Acids Res. 33: 2210-2214.

[28] Andronescu, M., Condon, A., Hoos, H.H., Mathews, D.H., Murphy, K.P. 2007. Efficient parameter estimation for RNA secondary structure prediction. Bioinformatics 23: i19-i28.

[29] Do, C.B., Woods, D.A., Batzoglou, S. 2006. CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics 22: e90-e98.

[30] Lucas, A., Dill, K.A. 2003. Statistical mechanics of pseudoknot polymers. J. Chem. Phys. 119: 2414-2421.

[31] Kopeikin, Z., Chen, S.J. 2005. Statistical thermodynamics for chain molecules with simple RNA tertiary contacts. J. Chem. Phys. 122: 094909.

[32] Kopeikin, Z., Chen, S.J. 2006. Folding thermodynamics of pseudoknotted chain conformations. J. Chem. Phys. 124: 154903.

[33] Cao, S., Chen, S.J. 2006. Predicting RNA pseudoknot folding thermodynamics. Nucleic Acid Res. 34: 2634-2652.

[34] Chen, S.J. 2008. RNA folding: conformational statistics, folding kinetics, and ion electrostatics. Annv. Rev, Biophys. 37: 197-214.

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

[36] Zhang, J., Lin, M., Chen, R., Wang, W., Liang, J. 2008. Discrete state model and accurate estimation of loop entropy of RNA secondary structures. J. Chem. Phys. 128: 125107.

[37] Parisien, M., Major, F. 2008. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. Nature. 452: 51-55.

[38] D'Souza, V., Dey, A., Habib, D., Summers, M.F. 2004. NMR Structure of the 101-nucleotide Core Encapsidation Signal of the Moloney Murine Leukemia Virus. J. Mol. Biol. 337: 427-442.

[39] Nikulin, A., Serganov, A., Ennifar, E., Tishchenko, S.,

Nevskaya, N., Shepard, W., Portier, C., Garber, M., Ehresmann, B., Ehresmann, C., Nikonov S., Dumas, P. 2000. Crystal structure of the S15-rRNA complex. Nature Structural Biology 7: 273-277.

[40] Ke, A., Zhou, K., Ding, F., Cate, J.H.D., Doudna, J.A. 2004. A conformational switch controls hepatitis delta virus ribozyme catalysis. Nature 429: 201-205.

[41] Flamm, C., Hofacker, I.L. 2008. Beyond energy minimization: approaches to the kinetic folding of RNA. Monatsh Chem 139: 447-457.

[42] Dundas, J., Binkowski, T., DasGupta, B., Liang, J. 2007. Topology independent protein structural alignment. BMC Bioinformatics 8: 388.

[43] Bar-Yehuda, R., Halldorsson, M.M., Naor, J., Shacknai, H., Shapira, I. 2002. Scheduling split intervals. 14th ACM-SIAM Symposium on Discrete Algorithms 732-741.

[44] Meszaros, C. 1996. Fast Cholesky factorization for interior point methods of linear programming. Comp Math Appl 31: 49-51.

[45] Binkowski, T.A., DasGupta, B., Liang, J. 2004. Order independent structural alignment of circularly permuted proteins. Conf Proc IEEE Eng Med Biol Soc 4: 2781-2784.

[46] Lin, M., Lu, H.M., Che, R., Liang, J. 2008. Generating properly weighted ensemble of conformations of proteins from sparse or indirect distance constraints. J. Chem. Phys. 129: 094101.

[47] Lin, M., Chen, R., Liang, J. 2008. Statistical geometry of lattice chain polymers with voids of defined shapes: Sampling with strong constraints. J. Chem. Phys. 128: 084903.

[48] Zhang, J.F., Lin, M., Chen, R., Liang, J., Liu, J. 2007. Monte Carlo sampling of near-native structures of proteins with applications. Proteins 66: 61-68.

[49] Zhang, J.F., Chen, R., Tang, C., Liang, J. 2003. Origin of scaling behavior of protein packing density: A sequential Monte Carlo study of compact long chain polymers. J. Chem. Phsy. 118: 6102-6109.

[50] Tyagi, R., Mathews, D.H. 2007. Predicting helical coaxial stacking in RNA multibranch loops. RNA 13: 939-951.

[51] Byun, Y., Han, K. 2006. PseudoViewer: Web application and web service for visualizing RNA pseudoknots and secondary structures. Nucleic Acids Research 34: W416-W422.

[52] Batenburg, F.H.D., Gultyaev, A.P., Pleij, C.W.A., Ng, J., Oliehoek, J. 2000. Pseudobase: a database with RNA pseudoknots. Nucl. Acids Res. 28: 201-204.

Tables

TABLE I: The ability of the extended Turner rule and pk3D to select the native structure from decoys.

	Turner + DP *		pk3D	
decoy size (m)	Sensitivity	PPV	Sensitivity	PPV
30	0.90	0.87	0.93	0.90
60	0.90	0.87	0.93	0.89
100	0.89	0.87	0.92	0.89
500	0.87	0.85	0.89	0.84

*Note: DP denoted the Dirks & Pierce's extension to the Turner's energy rule to account for pseudoknots [13].

Figure Captions:

Fig. 1. The secondary structure of the native state of tmRNA-Ec-PK4. Secondary structures in this paper are drawn using PSEUDOVIEWER [51] unless otherwise indicated.

Fig. 2. The native and predicted secondary structures of mRNA-Hs-Prp pseudoknotted RNA. (a) The native structure taken from the PseudoBase [52]. (b) The best candidate secondary structure in the candidate list that is most similar to the native secondary structure. Compared with the native structure, it overestimates the number of base pairs ($SE = 1, PPV = 0.65$) and has an uncommon pseudoknotted structure.

Fig. 3. The native secondary structures of (a) the 3' terminal region of the tobacco mosaic virus (TMV.R) and (b) the hepatitis delta virus genomic ribozyme (HDV), respectively.

Fig. 4. The secondary and tertiary structures of a fragment of gene 32 mRNA of bacteriophate T2. (a) The native secondary structure. (b) The native tertiary structure, taken from the PDB databank (PDB 2tpk). (c) The tertiary structure predicted by pk3D. Note that in both (b) and (c), loops are shown in yellow dashed lines to illustrate chain connectivity and to facilitate structural comparison. The orange spheres show the position of A8 nucleotide, which forms a short loop of length 1. All tertiary structures are drawn using Pymol.

Fig. 5. The native secondary structure and tertiary structure of the T arm and acceptor arm of the tRNA-like structure of turnip yellow mosaic virus. (a) The native secondary structure, (b) The native tertiary structure taken from the PDB databank (PDB 1A60). (c) The tertiary structure predicted by pk3D, which reproduces correctly the spatial arrangement of three stems, as well as their coaxial stacking observed in the experimental structure. The helices in (b) and (c) are colored using the same scheme. The loops are shown in yellow dashed lines to illustrate the chain connectivity.

Fig. 6. The native secondary structure and tertiary structure of the core encapsidation signaling RNA of the molney murine leukemia virus. (a) The native secondary structure. (b) The native tertiary structure taken from the first NMR model in the PDB databank (PDB 1S9S) (c) The tertiary structure predicted by pk3D. The helices are numbered from 1 to 6, in the direction from 5'-end to 3'-end, and are colored in the order of red, green, blue, yellow, magenta, and cyan, respectively. The orange spheres in (b) show the positions of the phosphorus atoms in the first nucleotide G1 for all of the first 20 models given by NMR experiments. The diverse positions demonstrate the experimentally observed large flexibility of the first helix. The gray segment at the upper right region in (b) connecting the fourth and fifth helices shows the conformation of the bulge loop G62-A65 determined by experiments.

Fig. 7. The native secondary structure and tertiary structure of the protein S15 binding fragment of 16S rRNA. (a) Its native secondary structure. (b) Its native tertiary structure as observed in X-ray crystallography (PDB 1DK1). The helices are numbered from 1 to 5 in the direction from 5' end to 3' end, and are colored in the order of magenta, green, yellow, blue, and cyan, respectively. The three-way junction is colored in red, which contains a base-triple and the metal binding site. In addition, local strands in this junction are in parallel direction instead of the canonical antiparallel direction. (c) The tertiary structure predicted by pk3D, plotted in the same color code as in (b). The spatial arrangement of the helices 3-5 are predicted correctly. The overall position of helix 1 is only roughly correct, tilted at a different angle.

Fig. 8. The native secondary structure and tertiary structure of the hepatitis delta virus (HDV) ribozyme precursor. (a) Its native secondary structure. (b) The tertiary structure (PDB 1SJ3). (c) The tertiary structure predicted by pk3D. The helices in (b) and (c) are colored in the same code. The loops are shown in orange dashed lines to illustrate chain connectivity. The structure generated by pk3D contains most features of the X-ray structure.

Fig. 9. The flow chart of pk2D and pk3D.

Fig. 10. The secondary structure of the hepatitis delta virus ribozyme precursor and its graph representation. The vertices in the graph correspond to helical stems and the edges correspond to loops; the weight of edges are set to the length of the corresponding loops. Note that this graph is not a metric graph.

TABLE II: The accuracy of prediction of secondary structures of 43 small pseudoknotted RNA molecules using five different algorithms.

		Lowest Sensitivity					Lowest PPV				
Sqeuence	Length	pk3D	HotKnots	pknotsRG	NUPACK	ILM	pk3D	HotKnots	pknotsRG	NUPACK	ILM
NGF-L6	48	1	1	0.65	1	0.94	1	1	0.69	1	1
BWYV	28	1	1	1	1	1	0.89	0.89	0.89	0.89	0.89
Rr-ODCanti	70	0.82	0.65	0.65	1	1	0.54	0.52	0.5	0.63	0.65
HDV	87	0.7	0.4	0.97	0.63	0.87	0.68	0.46	0.94	0.61	0.7
HDV-anti	91	0.92	0.17	0.17	0.42	0.71	0.65	0.14	0.14	0.32	0.5
HIVRT322	35	1	1	1	1	0.55	1	1	1	1	1
HIVRT32	35	1	1	1	1	0.91	1	1	1	1	1
HIVRT33	35	1	1	1	0.91	0.91	1	1	1	1	1
minimalIBV	45	1	0.94	0.94	0.94	0.94	0.94	0.89	0.94	0.94	0.89
MMTV	34	1	1	1	0.45	0.91	0.92	0.92	0.92	0.5	0.91
MMTV-vpk	34	1	1	1	0.91	0.91	0.92	0.92	0.92	1	0.91
mRNA-Bt-PrP	45	0.42	0.42	0.33	0.42	0	0.31	0.42	0.27	0.42	0
mRNA-Ec-alpha	108	0.79	0.46	0.46	0.46	0.54	0.54	0.31	0.29	0.31	0.28
mRNA-Ec-S15	67	0.94	1	0.76	0.88	0.88	0.73	0.74	0.68	0.71	0.68
mRNA-Hs-PrP	45	0	0	0	0	0.36	0	0	0	0	0.25
mRNA-T4-gene32	28	1	0.64	1	1	0.91	1	1	1	1	1
pKA-A	36	1	1	1	1	0.92	0.92	0.92	0.92	0.92	0.92
Bp-PK2	90	1	0.79	0.79	1	0.79	0.91	0.85	0.74	0.91	0.72
HDV-It-ag	89	0.92	0.16	0.16	0.4	0.68	0.68	0.14	0.14	0.32	0.49
satRPV	73	0.77	0.59	0.82	0.59	0.23	0.71	0.68	0.86	0.68	0.25
Tt-LSU-P3-P7	65	0.84	0.95	0.85	0.95	0.8	0.73	1	1	1	0.69
Sc-18S-PKE21-7-8	51	0.89	0.5	0.5	0.5	0.56	0.89	0.53	0.53	0.53	0.56
SRV-1	38	1	1	1	1	0	0.92	0.92	0.92	0.92	0
T4-gene32	31	1	1	1	1	0.91	1	1	1	1	1
T.the-telo	35	0.67	0.58	0.67	0.67	0.33	0.67	1	0.89	0.89	0.44
tmRNA-Ec-PK1	30	1	1	1	1	1	1	1	1	1	1
tmRNA-Ec-PK4	52	1	0.68	0.68	1	0.68	1	1	1	1	0.81
tmRNA-Lp-PK1	30	0.9	0.5	0.5	0.8	0.5	0.9	1	1	1	0.71
TMV.L	84	0.88	0.54	0.83	0.54	0.46	0.81	0.65	0.83	0.65	0.44
TMV.R	105	0.5	0.68	0.68	0.53	0.56	0.4	0.74	0.74	0.55	0.61
TYMV	86	0.84	0.72	0.76	0.44	0.52	0.72	0.78	0.79	0.5	0.46
BSBV1-UPD-PKc	24	1	1	1	1	0.67	1	1	1	1	1
BSBV3-UPD-PKc	24	0.67	1	0.67	0	0.67	1	1	1	0	1
BVQ3-UPD-PKb	33	0.78	0.56	1	1	0.56	0.7	1	1	1	0.5
PSLVbeta-UPD-PK1	23	0.62	0.62	0.62	0.62	0.62	1	1	1	1	1
PSLVbeta-UPD-PK3	35	1	1	1	1	1	0.92	0.92	0.92	0.92	0.92
SBRMV1-UPD-PKb	27	1	0.7	1	1	0.7	1	1	1	1	1
STMV-UPD2-PK3	24	1	1	1	1	0.75	0.89	0.8	0.89	0.89	0.75
TMV-L-UPD-PK3	32	0.88	0.5	1	1	0.38	0.7	1	1	1	0.3
PSIV-IRES	47	0.86	0.64	0.64	0.93	0.36	0.8	0.69	0.69	1	0.42
AMV3	113	0.64	0.87	0.87	0.69	0.87	0.66	0.89	0.89	0.68	0.83
BSMVbeta	96	0.45	0.74	0.84	0.71	0.94	0.42	0.79	0.81	0.67	0.85
CGMMV-PKbulge	69	0.43	0.83	0.65	0.61	0.61	0.43	1	0.68	0.61	0.64
Average	53	0.84	0.74	0.78	0.77	0.68	0.79	0.80	0.80	0.77	0.70

*Note: in some cases (such as TMV.R and HDV), the pk3D program rules out all the candidate secondary structures, since none are spatially feasible because of chain constraints or geometrical collisions. In these cases, the structure with the lowest free energy ranked by pk2D is taken as the native secondary structure.

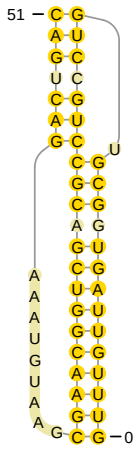

Fig. 1. The secondary structure of the native state of tmRNA-Ec-PK4. Secondary structures in this paper are drawn using pseudoviewer unless otherwise indicated.

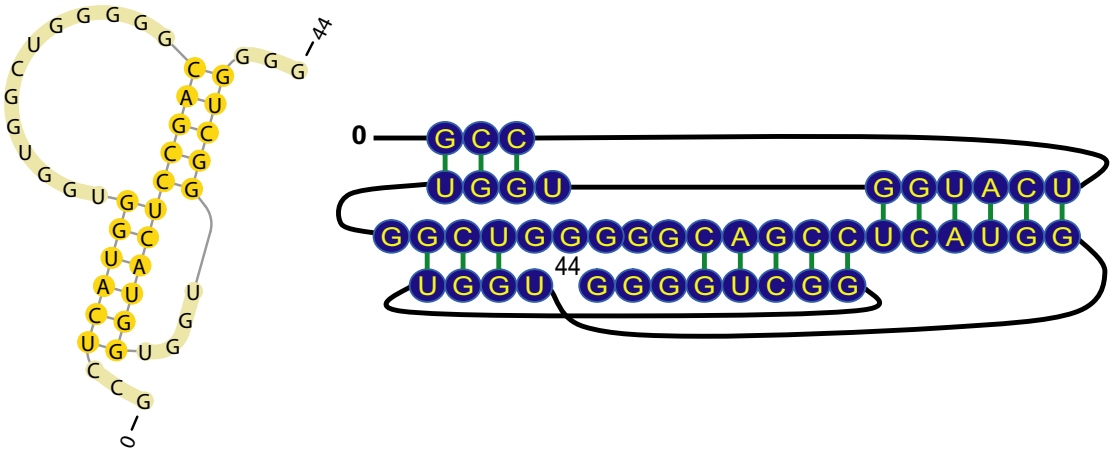

Fig. 2. The native and predicted secondary structures of mRNA-Hs-Prp pseudoknotted RNA. (a) The native structure taken from the PseudoBase [52]. (b) The best candidate secondary structure in the candidate list that is most similar to the native secondary structure. Compared with the native structure, it overestimates the number of base pairs (SE = 1, PPV = 0.65) and has an uncommon pseudoknotted structure.

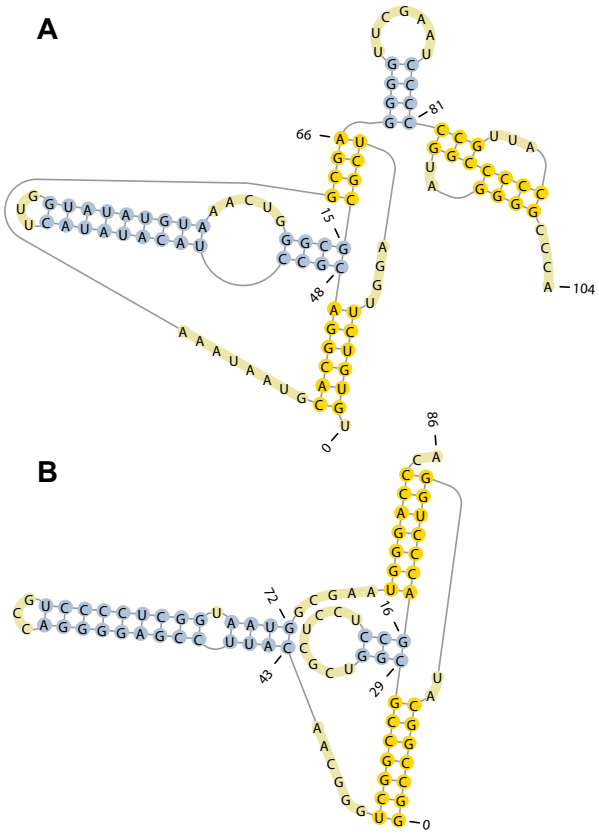

Fig. 3. The native secondary structures of (a) the 3’ terminal region of the tobacco mosaic virus (TMV.R) and (b) the hepatitis delta virus genomic ribozyme (HDV), respectively.

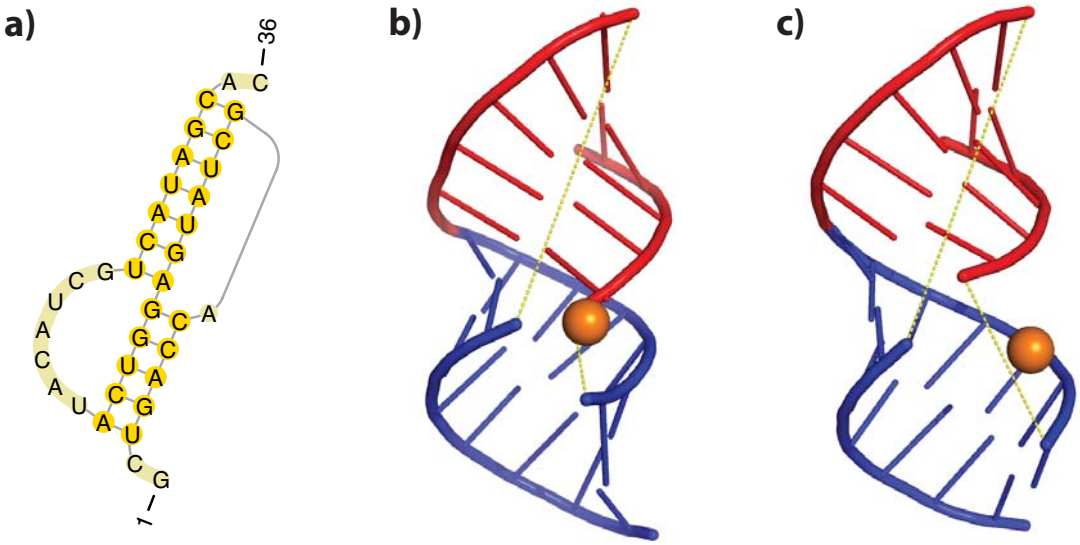

Fig. 4. The secondary and tertiary structures of a fragment of gene 32 mRNA of bacteriophate T2. (a) The native secondary structure. (b) The native tertiary structure, taken from the PDB databank (PDB 2tpk). (c) The tertiary structure predicted by pk3D. Note that in both (b) and (c), loops are shown in yellow dashed lines to illustrate chain connectivity and to facilitate structural comparison. The orange spheres show the position of A8 nucleotide, which forms a short loop of length 1. All tertiary structures are drawn using Pymol.

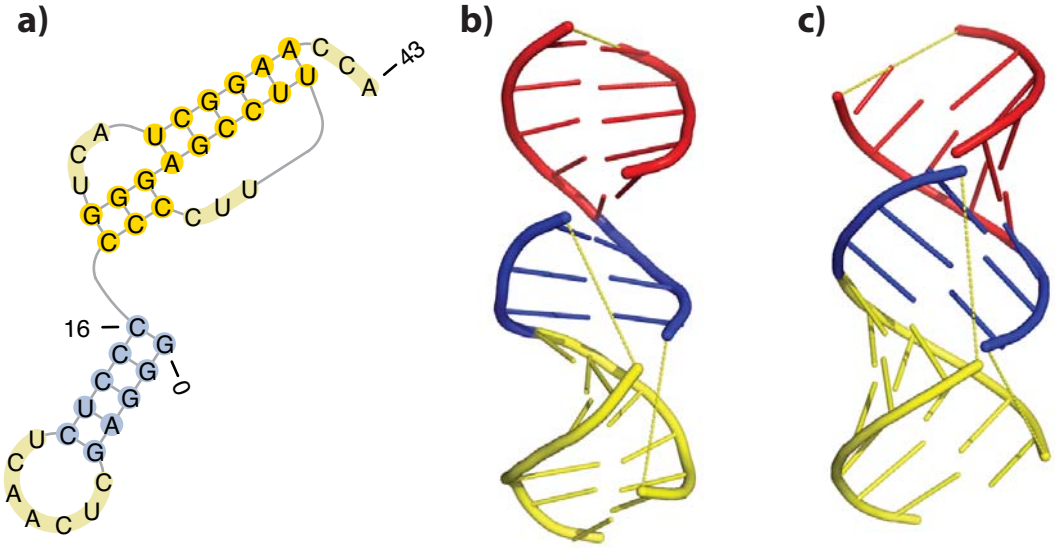

Fig. 5. The native secondary structure and tertiary structure of the T arm and acceptor arm of the tRNA-like structure of turnip yellow mosaic virus. (a) The native secondary structure, (b) The native tertiary structure taken from the PDB databank (PDB 1A60). (c) The tertiary structure predicted by pk3D, which reproduces correctly the spatial arrangement of three stems, as well as their coaxial stacking observed in the experimental structure. The helices in (b) and (c) are colored using the same scheme. The loops are shown in yellow dashed lines to illustrate the chain connectivity.

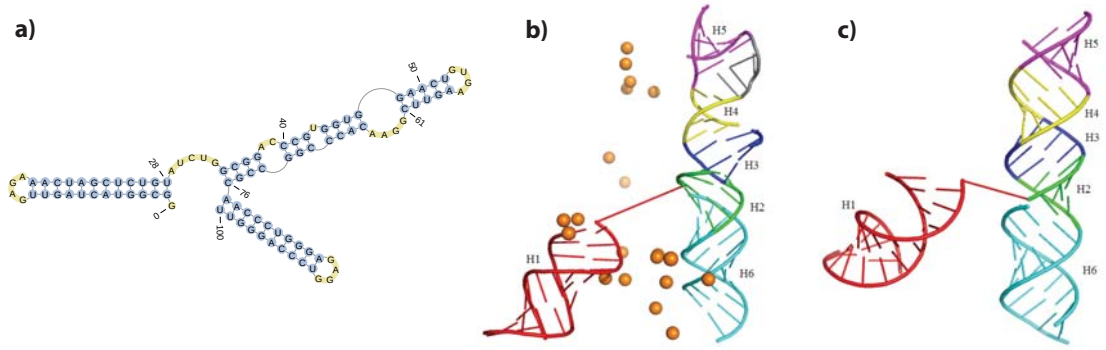

Fig. 6. The native secondary structure and tertiary structure of the core encapsidation signaling RNA of the molney murine leukemia virus. (a) The native secondary structure. (b) The native tertiary structure taken from the first NMR model in the PDB databank (PDB 1S9S) (c) The tertiary structure predicted by pk3D. The helices are numbered from 1 to 6, in the direction from 5'-end to 3'-end, and are colored in the order of red, green, blue, yellow, magenta, and cyan, respectively. The orange spheres in (b) show the positions of the phosphorus atoms in the first nucleotide G1 for all of the first 20 models given by NMR experiments. The diverse positions demonstrate the experimentally observed large flexibility of the first helix. The gray segment at the upper right region in (b) connecting the fourth and fifth helices shows the conformation of the bulge loop G62-A65 determined by experiments.

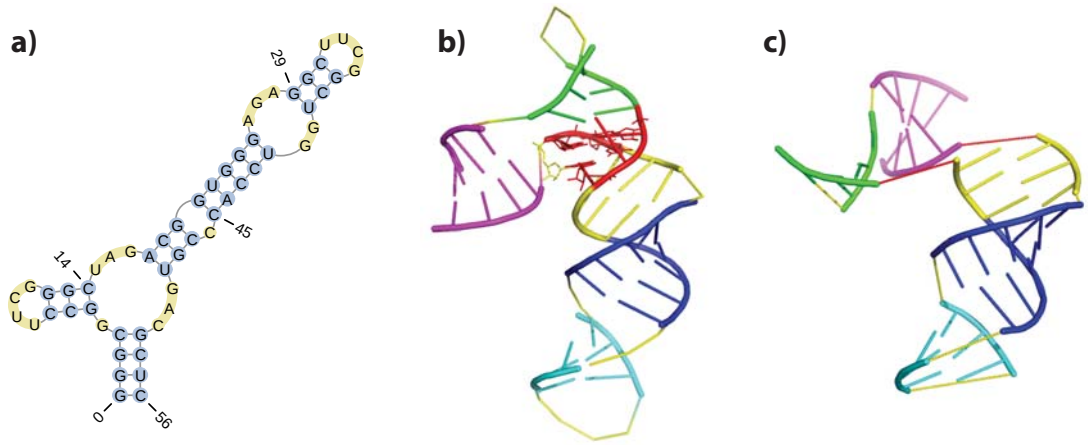

Fig. 7. The native secondary structure and tertiary structure of the protein S15 binding fragment of 16S rRNA. (a) Its native secondary structure. (b) Its native tertiary structure as observed in X-ray crystallography (PDB 1DK1). The helices are numbered from 1 to 5 in the direction from 5' end to 3' end, and are colored in the order of magenta, green, yellow, blue, and cyan, respectively. The three-way junction is colored in red, which contains a base-triple and the metal binding site. In addition, local strands in this junction are in parallel direction instead of the canonical antiparallel direction. (c) The tertiary structure predicted by pk3D, plotted in the same color code as in (b). The spatial arrangement of the helices 3-5 are predicted correctly. The overall position of helix 1 is only roughly correct, tilted at a different angle.

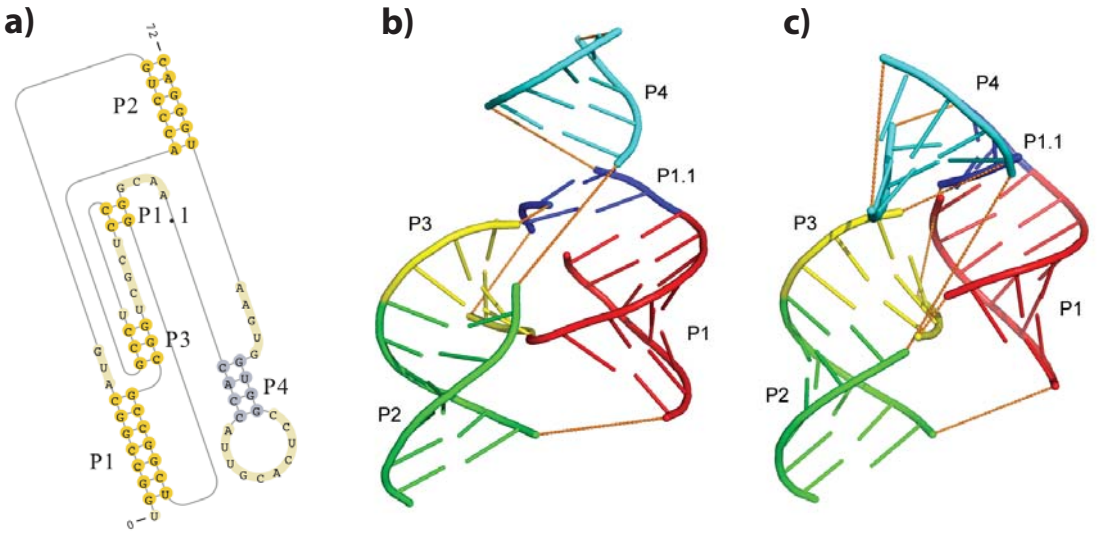

Fig. 8. The native secondary structure and tertiary structure of the hepatitis delta virus (HDV) ribozyme precursor. (a) Its native secondary structure. (b) The tertiary structure (PDB 1SJ3). (c) The tertiary structure predicted by pk3D. The helices in (b) and (c) are colored in the same code. The loops are shown in orange dashed lines to illustrate chain connectivity. The structure generated by pk3D contains most features of the X-ray structure.

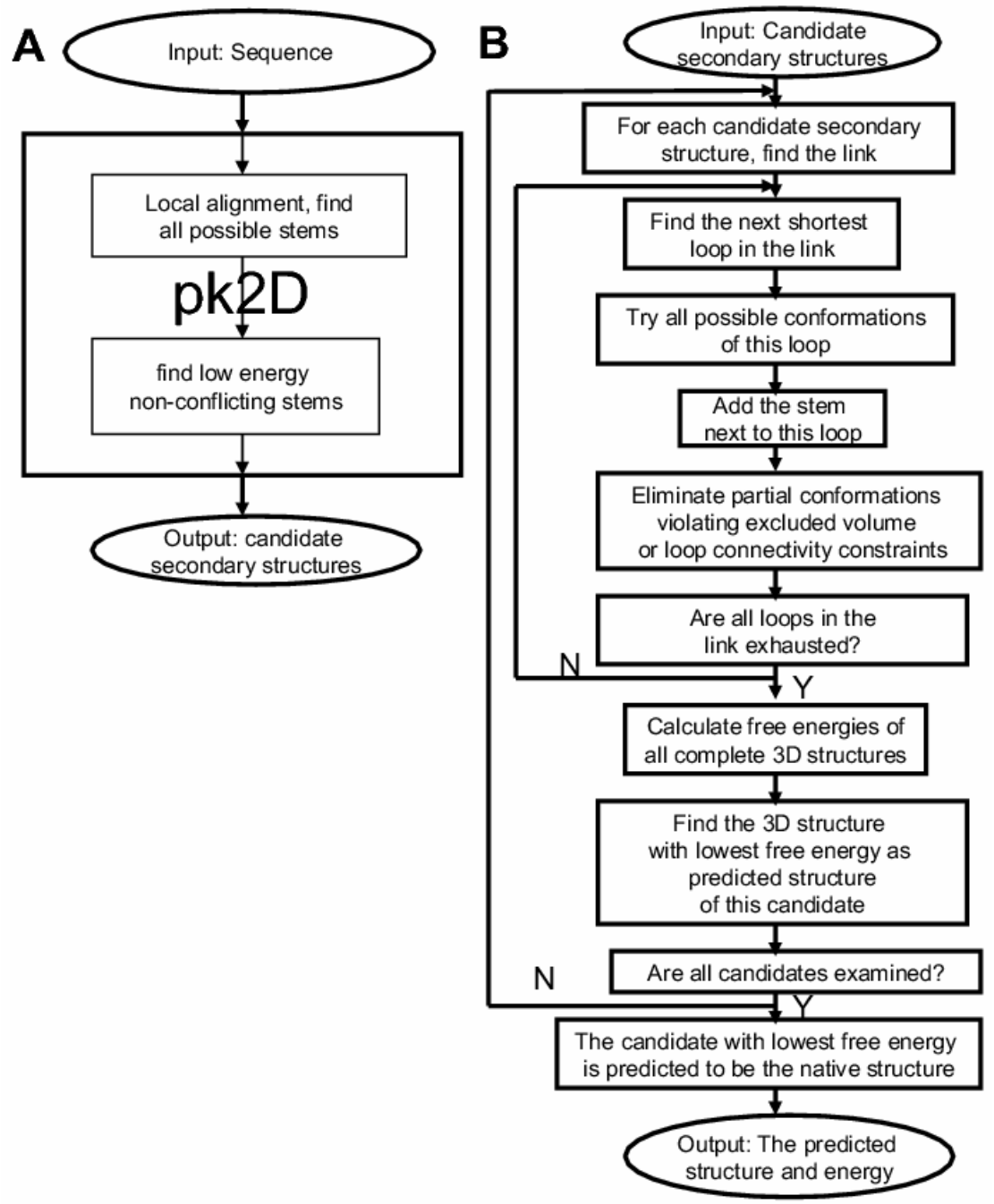

Fig. 9. The flow chart of pk2D and pk3D.

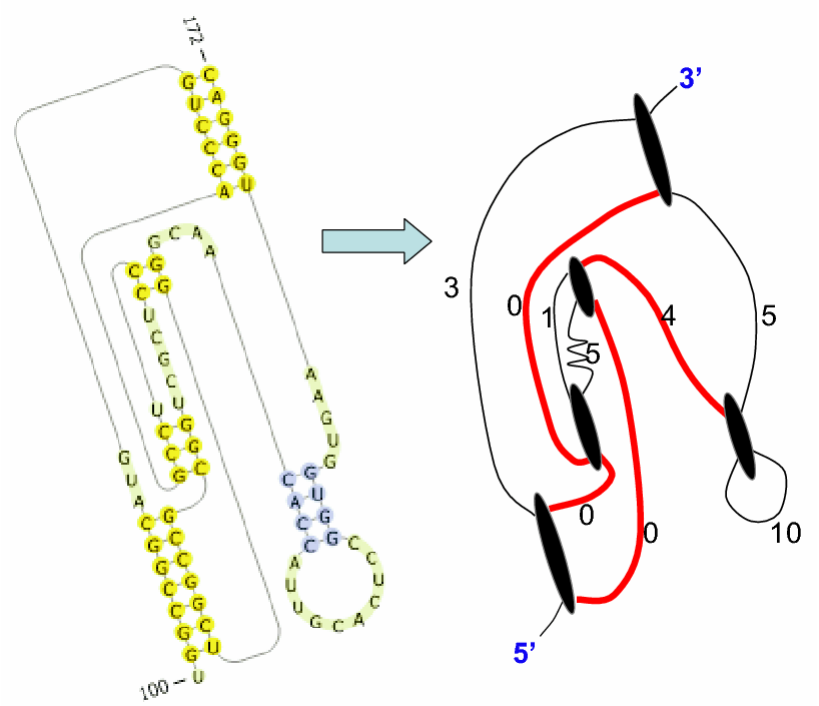

Fig. 10. The secondary structure of the hepatitis delta virus ribozyme precursor and its graph representation. The vertices in the graph correspond to helical stems and the edges correspond to loops; the weight of edges are set to the length of the corresponding loops. Note that this graph is not a metric graph.

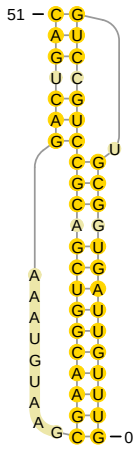

Fig. 1. The secondary structure of the native state of tmRNA-Ec-PK4. Secondary structures in this paper are drawn using pseudoviewer unless otherwise indicated.

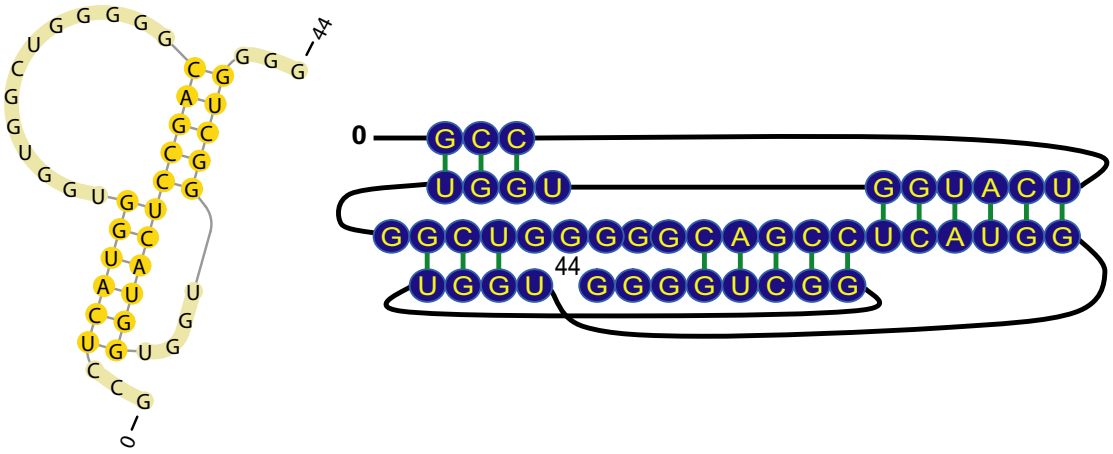

Fig. 2. The native and predicted secondary structures of mRNA-Hs-Prp pseudoknotted RNA. (a) The native structure taken from the PseudoBase [52]. (b) The best candidate secondary structure in the candidate list that is most similar to the native secondary structure. Compared with the native structure, it overestimates the number of base pairs (SE = 1, PPV = 0.65) and has an uncommon pseudoknotted structure.

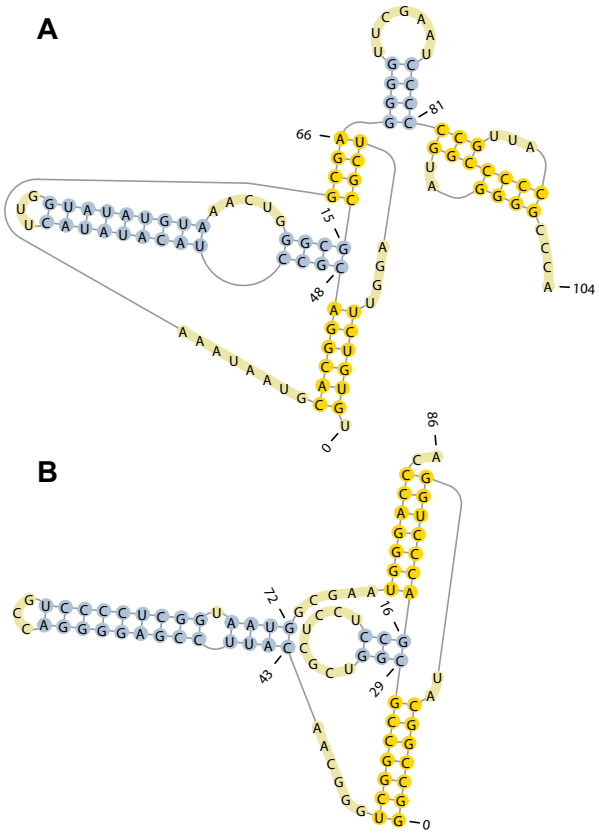

Fig. 3. The native secondary structures of (a) the 3’ terminal region of the tobacco mosaic virus (TMV.R) and (b) the hepatitis delta virus genomic ribozyme (HDV), respectively.

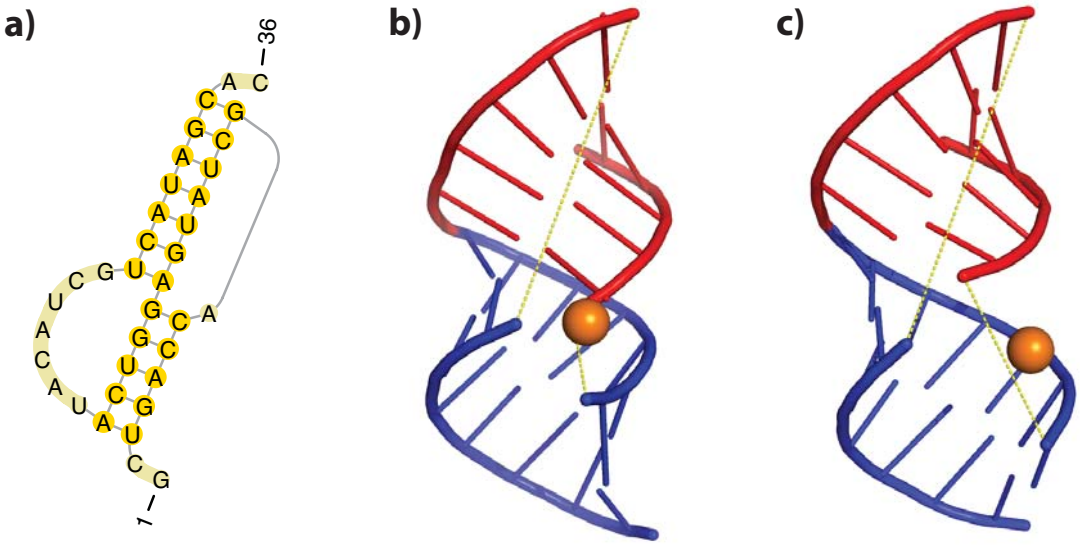

Fig. 4. The secondary and tertiary structures of a fragment of gene 32 mRNA of bacteriophate T2. (a) The native secondary structure. (b) The native tertiary structure, taken from the PDB databank (PDB 2tpk). (c) The tertiary structure predicted by pk3D. Note that in both (b) and (c), loops are shown in yellow dashed lines to illustrate chain connectivity and to facilitate structural comparison. The orange spheres show the position of A8 nucleotide, which forms a short loop of length 1. All tertiary structures are drawn using Pymol.

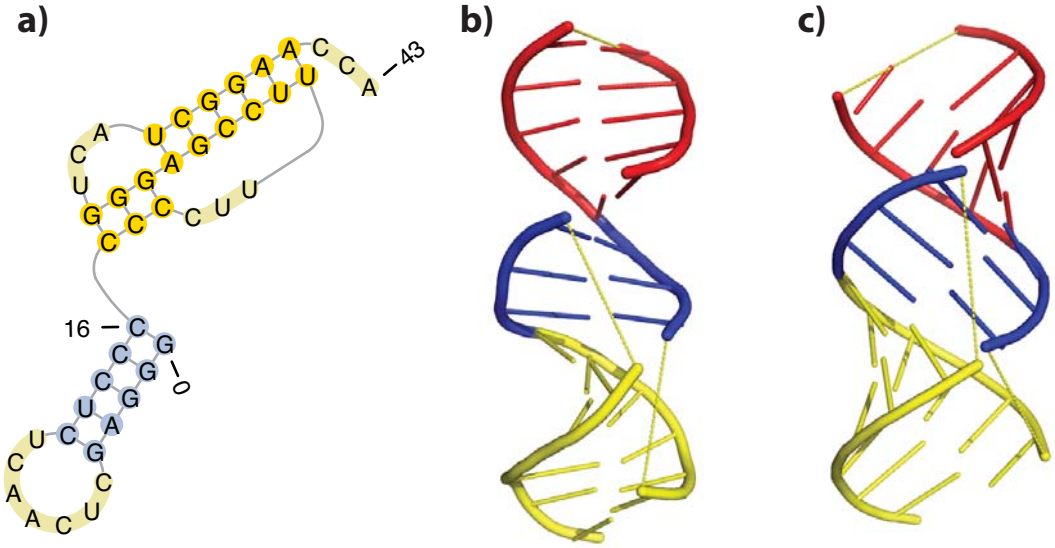

Fig. 5. The native secondary structure and tertiary structure of the T arm and acceptor arm of the tRNA-like structure of turnip yellow mosaic virus. (a) The native secondary structure, (b) The native tertiary structure taken from the PDB databank (PDB 1A60). (c) The tertiary structure predicted by pk3D, which reproduces correctly the spatial arrangement of three stems, as well as their coaxial stacking observed in the experimental structure. The helices in (b) and (c) are colored using the same scheme. The loops are shown in yellow dashed lines to illustrate the chain connectivity.

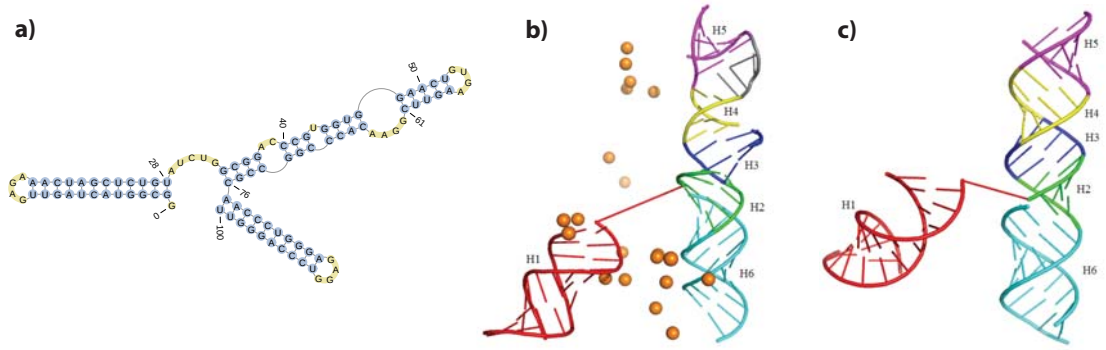

Fig. 6. The native secondary structure and tertiary structure of the core encapsidation signaling RNA of the molney murine leukemia virus. (a) The native secondary structure. (b) The native tertiary structure taken from the first NMR model in the PDB databank (PDB 1S9S) (c) The tertiary structure predicted by pk3D. The helices are numbered from 1 to 6, in the direction from 5'-end to 3'-end, and are colored in the order of red, green, blue, yellow, magenta, and cyan, respectively. The orange spheres in (b) show the positions of the phosphorus atoms in the first nucleotide G1 for all of the first 20 models given by NMR experiments. The diverse positions demonstrate the experimentally observed large flexibility of the first helix. The gray segment at the upper right region in (b) connecting the fourth and fifth helices shows the conformation of the bulge loop G62-A65 determined by experiments.

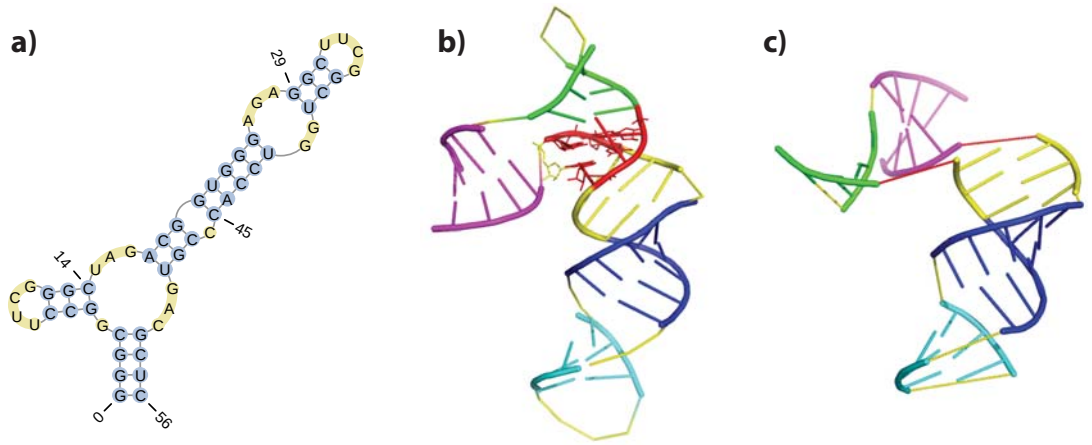

Fig. 7. The native secondary structure and tertiary structure of the protein S15 binding fragment of 16S rRNA. (a) Its native secondary structure. (b) Its native tertiary structure as observed in X-ray crystallography (PDB 1DK1). The helices are numbered from 1 to 5 in the direction from 5' end to 3' end, and are colored in the order of magenta, green, yellow, blue, and cyan, respectively. The three-way junction is colored in red, which contains a base-triple and the metal binding site. In addition, local strands in this junction are in parallel direction instead of the canonical antiparallel direction. (c) The tertiary structure predicted by pk3D, plotted in the same color code as in (b). The spatial arrangement of the helices 3-5 are predicted correctly. The overall position of helix 1 is only roughly correct, tilted at a different angle.

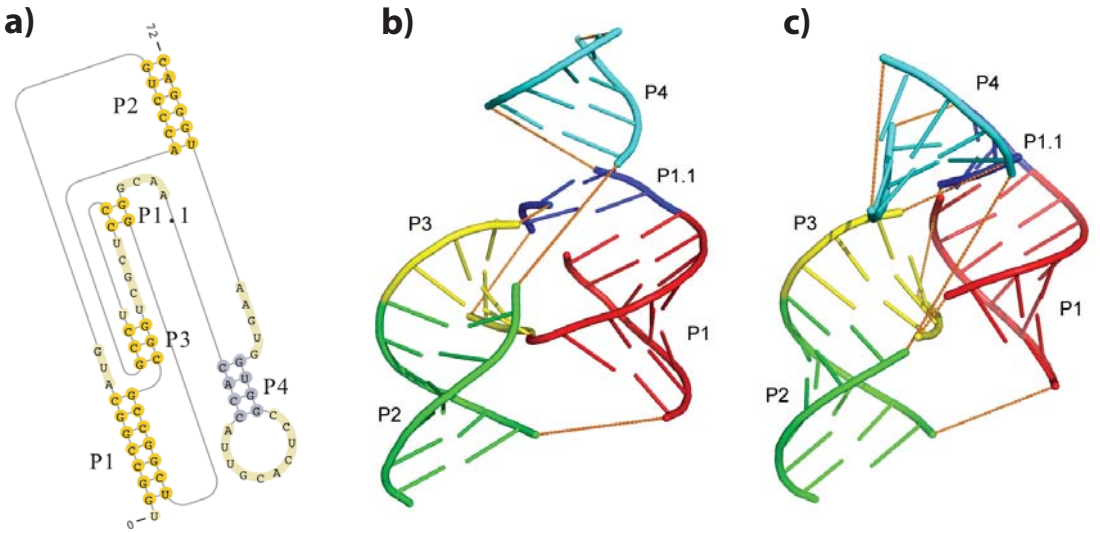

Fig. 8. The native secondary structure and tertiary structure of the hepatitis delta virus (HDV) ribozyme precursor. (a) Its native secondary structure. (b) The tertiary structure (PDB 1SJ3). (c) The tertiary structure predicted by pk3D. The helices in (b) and (c) are colored in the same code. The loops are shown in orange dashed lines to illustrate chain connectivity. The structure generated by pk3D contains most features of the X-ray structure.

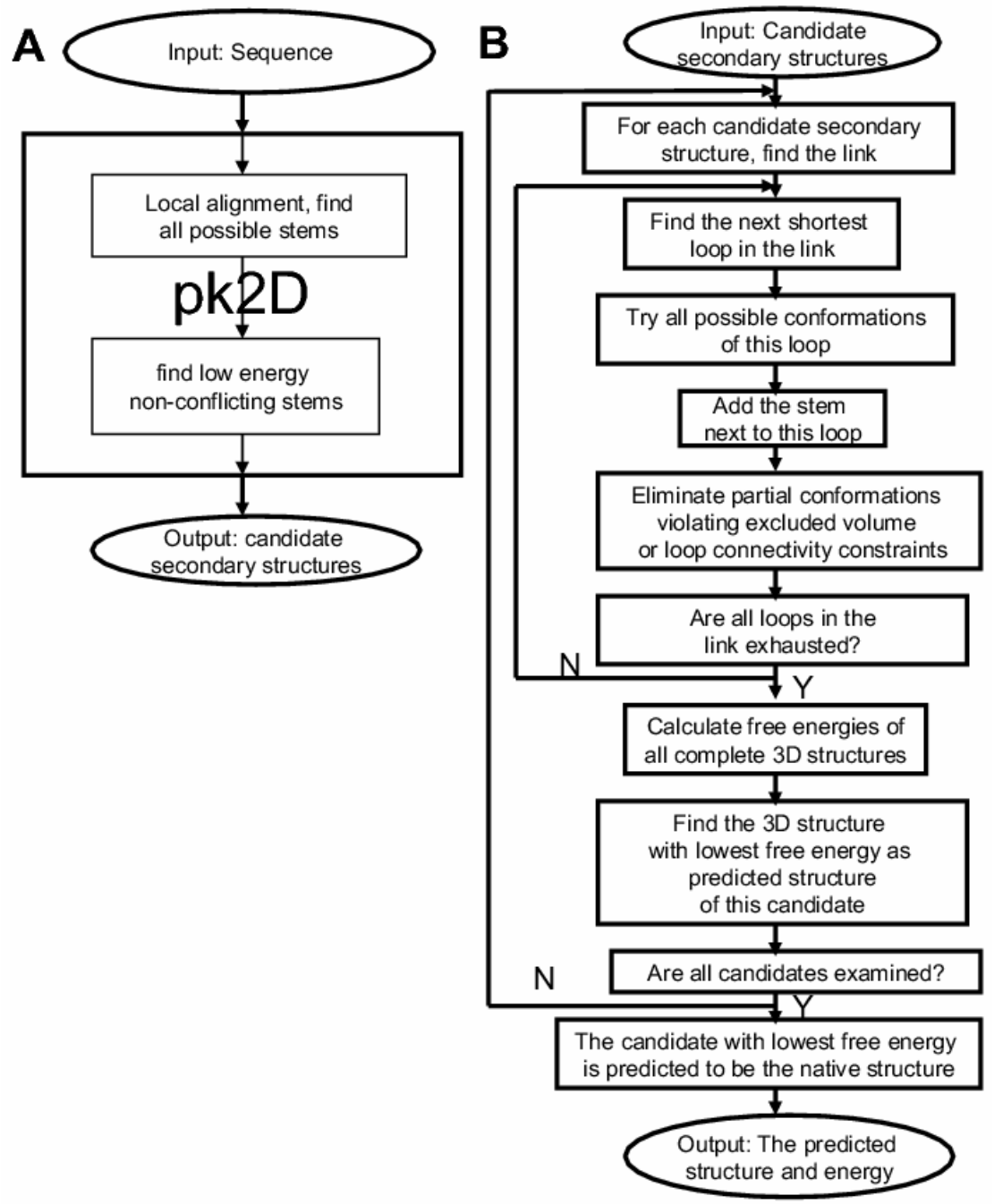

Fig. 9. The flow chart of pk2D and pk3D.

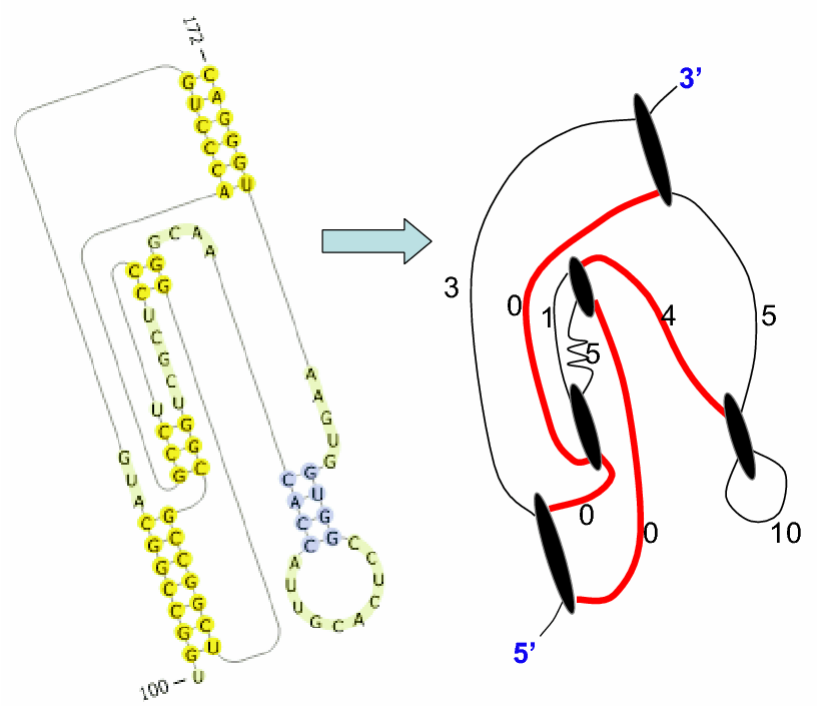

Fig. 10. The secondary structure of the hepatitis delta virus ribozyme precursor and its graph representation. The vertices in the graph correspond to helical stems and the edges correspond to loops; the weight of edges are set to the length of the corresponding loops. Note that this graph is not a metric graph.