*Published online 20 May 2019*

*Nucleic Acids Research, 2019, Vol. 47, No. 11* ***5563–5572***
*doi: 10.1093/nar/gkz427*

# Limits in accuracy and a strategy of RNA structure prediction using experimental information

**Jian Wang [1], Benfeard Williams, II[2], Venkata R. Chirasani[1], Andrey Krokhotin[2], Rajeshree Das[3] and Nikolay V. Dokholyan [1,2,4,5,6,*]**

[1]Department of Pharmacology, Penn State University College of Medicine, Hershey, PA 17033, USA, [2]Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27514, USA, [3]Weinberg College of Arts and Sciences, Northwestern University, Evanston, IL 60208, USA, [4]Department of Biochemistry and Molecular Biology, Penn State University College of Medicine, Hershey, PA 17033, USA, [5]Department of Chemistry, Penn State University, University Park, PA 16802, USA and [6]Department of Biomedical Engineering, Penn State University, University Park, PA 16802, USA

Received March 13, 2019; Revised May 03, 2019; Editorial Decision May 06, 2019; Accepted May 08, 2019

## ABSTRACT

**RNA structural complexity and flexibility present a challenge for computational modeling efforts. Experimental information and bioinformatics data can be used as restraints to improve the accuracy of RNA tertiary structure prediction. Regarding utilization of restraints, the fundamental questions are: (i) What is the limit in prediction accuracy that one can achieve with arbitrary number of restraints? (ii) Is there a strategy for selection of the minimal number of restraints that would result in the best structural model? We address the first question by testing the limits in prediction accuracy using native contacts as restraints. To address the second question, we develop an algorithm based on the distance variation allowed by secondary structure (DVASS), which ranks restraints according to their importance to RNA tertiary structure prediction. We find that due to kinetic traps, the greatest improvement in the structure prediction accuracy is achieved when we utilize only 40–60% of the total number of native contacts as restraints. When the restraints are sorted by DVASS algorithm, using only the first 20% ranked restraints can greatly improve the prediction accuracy. Our findings suggest that only a limited number of strategically selected distance restraints can significantly assist in RNA structure modeling.**

## INTRODUCTION

Biological functions of RNA molecules rely on a variety of complex 3D conformations, such as pseudoknots and non-canonical base pairs. Over the past few decades, high-resolution structures (1–3) of these complex conformations have significantly advanced our understanding of RNA structure and function. However, the gap between high-resolution RNA tertiary structures solved experimentally and newly discovered functional RNAs is immense due to lack of efficient structural biology tools for solving RNA tertiary structures. Although computer-based methods have progressed substantially in the past decade for RNA structure modeling (4–21), based on the results of an RNA tertiary structure prediction competition, RNA-Puzzles (9–11), these methods are not yet adequate to model large RNA structures with complex architectures. Predicting structures of large RNA molecules with non-canonical interactions has become increasingly difficult due to the longer time scales needed to sample vast conformational landscapes. Nonetheless, the efficiency and accuracy of secondary (22–24) and tertiary structure prediction algorithms (4–11,14–21,25) can be substantially improved by combining with structural restraints derived from either sequence co-evolution analysis data (12,26,27) or chemical detection data (e.g. SHAPE-MaP (28) and RING-MaP (29)) or hydroxyl radical detection (30) or mutation and mapping methods (31). In this process, researchers have efficiently developed advanced methodologies for RNA structure prediction or modeling by amalgamating experimental and computational techniques (4,12,27,30,32). SimRNA (15,33) supports user-designated distance restraints that represent any type of pairwise interaction as long as it can be defined in terms of any pair of the five atoms (P, C4′ and N1, C2, C4 for pyrimidines or N9, C2, C6 for purines) or a virtual point at the middle of the base. FARFAR (20,34) integrates NMR $^{1}$H chemical shift data with Rosetta *de novo* modeling to consistently produce high-resolution RNA structures. 3dRNA (12,35) utilizes sequence coevolution analysis results to guide the structure optimization procedure after an initial fragment assembly procedure. NAST

*To whom correspondence should be addressed. Tel: +1 717 531 5177; Email: dokh@psu.edu

(36) supports known or predicted tertiary contacts and additionally, it can use residue-resolution experimental data such as hydroxyl radical to filter the generated decoy structures. Vfold (6) is a statistical mechanics-based RNA folding model that can predict both RNA 2D and 3D structures. RNAComposer (16) is a knowledge-based RNA modeling method employing fully automated fragment assembly. ModeRNA (17) is a comparative RNA modeling method capable of handling 115 different nucleotide modifications. We have previously developed a platform iFoldRNA (http://iFoldRNA.DokhLab.org) (5,14) for RNA tertiary structure modeling using physical force field combined with different types of experimental data, including those obtained from NMR (4) or FRET (37).

While the incorporation of restraints could efficiently improve the prediction accuracy, additional restraints impose significant burden on computational cost, especially when the number of restraints is extremely large. Furthermore, unlike protein modeling (38), the incorporation of restraints in RNA structure prediction may not always lead to an increase of the prediction accuracy due to kinetic traps in over-constrained models. There is currently no strategy available for selecting suitable restraints that contribute to improving the accuracy of RNA structure prediction. Such strategy would have significantly reduced experimental burden for structural characterization of RNA molecules by focusing on determining experimental information that would result in most accurate RNA structures. Here, we ask two critical questions: (i) What is the limit in accuracy that one can achieve with arbitrary number of pair-wise distance restraints? (ii) Is there a strategy for selection of the minimal number of restraints that would result in the best structural model of RNA?

We utilize the Gō model (39) to mimic pair-wise distance restraints. In Gō model, two nucleotides attract or repel each other if they are or not in proximity in the native state. Gō potential effectively biases formation of native contacts and disfavors non-native ones, thus allowing proteins (38,40-50) and RNA molecules (4,30,51,52) to navigate the energy landscape (53–56) to their native state by reducing frustrations in the free energy landscape (57). Utilization of the Gō model in simulations ensures consistency of the modeled RNA molecules with their native structures without any specific force field biases. We have found that (4,30,51,58–61) RNA native structures can be reproduced with high fidelity using experimental-derived restraints through Gō model in discrete molecular dynamics (DMD) simulations. Using a fraction of native contacts as restraints in Gō model allows us to directly interrogate prediction accuracy as a function of the fraction of utilized restraints, and address the first question.

To address the second question, we develop a restraints-sorting algorithm based on distance variation allowed by secondary structure (DVASS), which ranks restraints by the distance variation between two residues when the given secondary structure is formed. Using the distance variation metric, we measure the importance of restraints on RNA tertiary structure prediction, that is, how much a constraint would improve the prediction accuracy. The DVASS algorithm is based on purely geometrical considerations and does not rely on molecular dynamics simulations, allowing rapid evaluation of restraints' importance. Our results show that DVASS can effectively rank all restraints, and the high-ranking restraints could greatly improve the structure prediction, while the low-ranking restraints result in insignificant accuracy improvement. Without using DVASS, approximately 40–60% of all the restraints are needed to maximally improve the prediction accuracy. After sorting the restraints by DVASS, if merely the 20% top ranked sorted restraints are employed, the improvement of the prediction accuracy is comparable to the improvement of the prediction accuracy obtained by using 60% unsorted restraints. Our findings suggest that due to the rugged free energy landscape of RNA, constraining the molecule in simulations beyond 60% of the total number of restraints is not beneficial for accurate structure prediction. Our new algorithm reduces the burden of determining pair-wise distance restraints to 20%, thus offering a strategy for integration of experimental and computational workflows for RNA structure modeling. The implementation of DVASS algorithm could be downloaded from http://dokhlab.org/dokhlab/download/dvass.tar.gz or https://bitbucket.org/dokhlab/dvass.

## MATERIALS AND METHODS

### iFoldRNA

iFoldRNA utilizes a coarse-grained three-bead RNA model (5), in which each bead represents a phosphate, sugar or nucleobase. The prediction is based on the discrete molecular dynamics (DMD) (51,60,62) engine implemented in Dokholyan Lab. Base-pairing information is incorporated in the simulation as an additional potential. A collection of RNA molecules are subject to replication exchange molecular dynamics at different temperatures to enhance conformation sampling (63). After the DMD simulation, the 100 lowest energy structures are selected and clustered according to the RMSD between the selected pair of structures. The centroid of the resulting cluster is preserved for all-atom reconstruction. If hydroxyl radical reactivity probing (30) or NMR data (4) are available, an additional force field is applied to effectively bias the RNA to the native structure. The reconstruction of an all-atom model from the coarse grained model is performed by replacing each of the three-bead nucleotides with a rotamer of the corresponding nucleotide selected at random from all-atom structures available in Protein Data Bank (64).

### Restraints definition

Base pairs between nucleotides in the form of Watson–Crick interactions are the general form of restraints used in RNA structure prediction approaches. Additional interactions exist such as base stacking and non-canonical base pairs such as reverse base pair interactions and Hoogsteen base pairs (65). These interactions together form the basis for the contacts defining RNA tertiary structures. We measure various canonical and non-canonical interactions seen in published RNA structures in the Protein Data Bank (64) to define a native contact reaction coordinate for our DMD simulations (Figure 1A). We find that the optimal definition for a native contact in RNA is between any distance

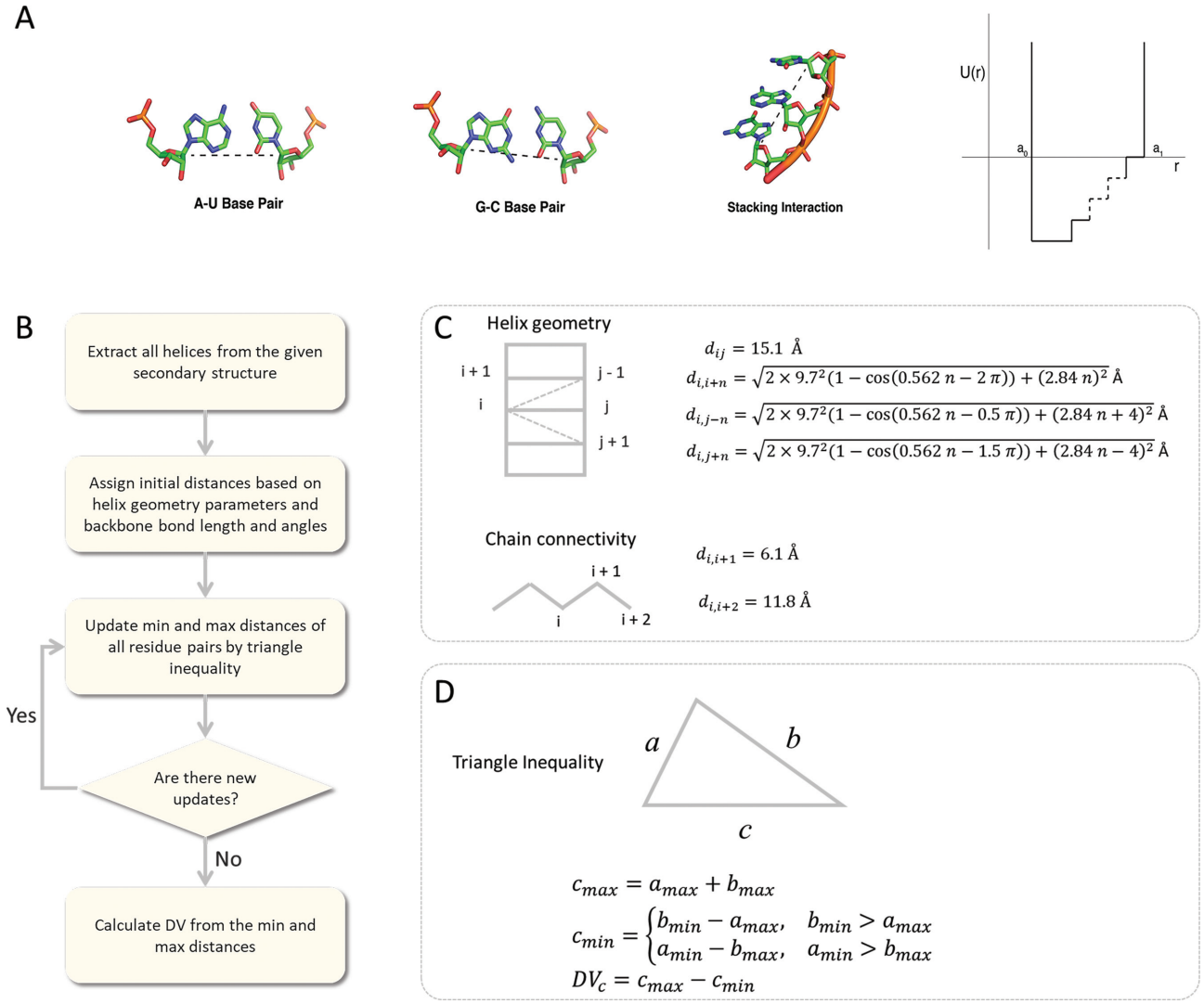


**Figure 1.** Design of RNA Native Contacts, Gō Potential Restraints and the DVASS Algorithm. (**A**) Distances between C1′ atoms in RNA fall under 15 Å for many common base interactions such as A-U Watson–Crick base pairs (10.2 Å) and G-C Watson–Crick base pairs (10.7 Å) or stacking interactions with adjacent nucleotides (5.4–10.7 Å). Native restraints are modeled using a step-wise function in DMD simulations. Potential energy values for the well reach −1.0 kcal/mol/K and the walls of the restraints span from 3.75 to 15 Å. (**B**) The workflow of DVASS algorithm. (**C**) The helix geometry parameters and the chain connectivity parameters. (**D**) The triangle inequality.

less than or equal to 15 Å between C1′ atoms in any two nucleotides. The C1′ atom, located in the sugar ring, is between the nucleobase and phosphate group and provides a central and consistent marker for defining contacts of various canonical and non-canonical RNA interactions. We can then use this distance measurement to implement Gō restraints that reinforce native contacts with an energy bonus defined by a step function (Figure 1A).

### Implementation of RNA Gō potential

The potential is designed as step functions constraining distances between beads of the iFoldRNA three bead model. Both attractive and repulsive restraints are tested and implemented to either limit beads within or outside of the 3.75–15 Å range. The restraints are only applied to pairs of nucleotides with sequence numbers $i$ and $j \geq i + 2$. The energy well for inter-unit interactions is assumed to be identical for all interacting pairs and is assigned a value of 1.0 kcal/mol/K. The step-function also contains a single well over the distance range allowing for flexibility within the system during folding simulations.

### Computational modeling using Gō restraints

We perform RNA coarse-grained DMD simulations, consisting of three pseudo-atoms per nucleotide representing base, sugar and phosphate groups. The native contacts and Gō restraints are generated using a custom Python script that analyzes the atom distances within a model from X-ray crystal structures or the lowest energy model from NMR ensembles downloaded from the Protein Data Bank. We incorporate the native contacts into the DMD simulations as attractive potentials and base pair (A•U, G•C Watson–Crick pairs and G•U wobble base pairs) restraints into the state file before running replica exchange DMD simulations

for 500 000 time units (∼25 ns) at temperatures of 0.2, 0.225, 0.25, 0.27, 0.3, 0.333, 0.367 and 0.4 kcal/mol•kB. We perform a clustering analysis on the lowest energy models that satisfy the native contacts from the coarse-grained trajectories. Clustering analysis is performed using an RMSD-based hierarchical clustering algorithm, OC (66), to select the final structural model.

### Thermodynamic study of the DMD simulation

We derive the potential of mean force (PMF) as follows:

$$\mathrm{PMF}(\mathrm{RMSD}, E) = -k_b T \ln(W(\mathrm{RMSD}, E)) + C$$

where $E$ is the iFoldRNA energy, $k_b$ is the Boltzmann constant, $T$ is the temperature (K), $W$ is a function that defines the probability of a given pair of RMSD and the iFoldRNA energy, and the constant $C$ sets the lowest PMF value at any given temperature to be zero.

### The DVASS algorithm

Using only the secondary structure as the input, the DVASS algorithm estimates and outputs distance variation values for all pairs of residues (or the given residue pairs list provided by the user). Distance variation of two residues is the difference between the maximum and minimum distances between the two residues among all possible structures in the conformational space. By forming the secondary structure, the distance between any two residues in helix regions is supposed to be nearly fixed because helices in RNA structures ordinarily possess similar structures without significant disparity. Based on distances between C1′ atoms in residues, we compiled a set of parameters (Figure 1C) depicting characteristic local structures, such as helix, generally assuming a canonical double-helical form, and backbone bond, typically featuring fixed values of both bond length and bond angle. The distance between two successive residues in the backbone, calculated by the distance between C1′ atoms in the two residues, is around 6.1 Å. The distance between two residues in the backbone separated by one residue is ∼11.2 Å. Based on this distance information derived from the secondary structure, we could infer the maximum and minimum distances of all other residue pairs by conducting the steps in the workflow in Figure 1B. A distance variation is then defined as the difference between the minimum possible distance and the maximum possible distance. The algorithm is implemented by iteratively applying the triangle inequality (Figure 1D). If we know the minimum and the maximum distances between residue $a$ and residue $b$, and the minimum and maximum distances between residue $b$ and residue $c$, then we could infer the minimum and maximum distances between residue $a$ and residue $c$.

We show in Supplementary Figure S1 an example of the calculation of distance variations given an artificial secondary structure: '((((....((((...)))))..((((()))))..))))'. Initially, we construct a matrix ($N \times N$), where $N$ is the number of residues. Entries in the upper triangle of the matrix are all assigned 999 (Å) and entries in the lower triangle are assigned 6.1 (Å), which is the distance between C4′ atoms in two adjacent residues in RNA. The pair-wise distances in the chain and the three helix regions (Supplementary Figure S1F) are determined based on the formula in Figure 1C and 1D, and they are then assigned to the corresponding positions in the matrix (Supplementary Figure S1A). The triangle inequality is then iteratively applied in the matrix. In each step, all entries in the matrix are updated by applying triangle inequality to all possible ternary tuples. The procedure is converged after five steps when there are no changes in the matrix (Supplementary Figure S1B, C and D). The distance variations are then calculated by subtracting the lower triangle from the upper triangle (Supplementary Figure S1E).

Once we attain all distance variations, the next step is to cluster them since we observed that certain distance variations are inter-correlated. Suppose residue $a$ and residue $b$ have a fairly large distance variation, while residue $a$ and residue $c$ have an extremely small distance variation, then the distance variation between residue $c$ and residue $b$ is likely to be large. The clustering algorithm is based on such a straightforward observation described above. Initially, each of the distance variations is allotted to a distinct cluster, then we randomly pick out a distance variation and then assign the distance of the corresponding residue pair a definite value, which renders the selected distance variation to be 0. We then merge this distance variation and those distance variations that undergo drastic change (larger than a customized cut-off) into one single cluster. By iteratively conducting this procedure, we cluster all the distance variations. The selection of the cut-off is crucial to the resulted number of clusters. A small cut-off may result in a plethora of clusters, while a large cut-off will only engender a smattering of clusters.

### Prediction of the number of contacts

We extend the DVASS algorithm to predict the number of contacts given the secondary structure. First, we likewise assign some initial distances by the knowledge of RNA helix geometry parameter and backbone conformation (Figure 1C). We then iteratively utilize the triangle inequality (Figure 1D) to deduce the distance variations of all the residue pairs. We traverse all the distance variations to find out the one that has the largest value and then assign a definite value to the distance between the two residues pertaining to the corresponded distance variation. Thus, the new distance variation of this residue pair is 0. We then again iteratively utilize the triangle inequality to update all the distance variations. We repeat this process 50 times and then count the number of residue pairs that have the minimum distance <15.2 Å, which is trained in a 22 RNAs dataset (Supplementary Table S1) and slightly larger than the aforementioned cut-off value 15 Å. This number serves as the final predicted number of contacts.

## RESULTS

### Dataset generation

We perform simulations of RNA folding using native contact restraints in coarse-grained DMD simulations (iFoldRNA) to evaluate the relationship between the prediction accuracy and the number of restraints imposed. We select

22 RNAs (Supplementary Table S1) from PDB database as the test set. We choose RNA structures from PDB database in such a way that the lengths of the 22 RNAs vary from 22 to 233 nt, and they have various types of loops, including hairpin loops (3OVA, 2ZY6), internal loops (2PXV, 4QVI), junction loops (2N3R, 3RG5) and pseudo-knots (2M8K, 3L1V). Subsequently, we extract all contacts from their native structures. The contacts between two nucleotides are defined by a distance less than or equal to 15 Å between corresponding C1′ atoms. This definition covers commonly occurring native contacts, such as Watson–Crick base pairs, and non-canonical interactions crucial for proper modeling of complex RNA. Each of the contacts could be used in the structure prediction as a constraint.

### Correlation between the number of restraints and the prediction accuracy

To test how the prediction accuracy changes with the number of restraints, we vary percentages (0%, 5%, 10%, 20%, 40%, 60%, 80% and 100%) of the total native contacts restraints used in simulation to determine the tertiary structures of the 22 RNAs. For each set of restraints, we randomly select a corresponding subset of restraints and perform 20 different simulation attempts. The restraints in different simulation attempts are different. For each of the RNAs, we calculate the average RMSD and the corresponding standard deviation of each of the constraint sets.

From our simulations, we find that the RMSD decreases steadily with the rise in the percentage of restraints from 0 to 40% (Figure 2A). The RMSD shows no noticeable decrease when the restraints percentage increases from 40% to 60% (Figure 2A). The RMSD increases when more than 60% restraints are imposed on the RNA model, suggesting that the molecule becomes kinetically trapped in local minima. Using 100% native contacts restraints to model RNA results in a slightly higher RMSD of the predicted structure than that using 60% restraints (Figure 2A) for some but not all studied RNA. For some of the RNAs (such as 2L1V, Figure 2C), using 100% restraints results in an RMSD as low as that of 60% restraints, while for other RNAs (such as 3LA5, Figure 2B), 100% restraints result in a much higher RMSD than that of 60% restraints. Kinetic traps associated with RNA folding are likely related to the complexity of RNA structure. For example, the secondary structure of 3LA5 is a complex three-way junction (Figure 2D); correspondingly, the RMSD of the predicted native structure significantly increases when more than 60% restraints are imposed (Figure 2B). The secondary structure of 2L1V is relatively simple since it contains only two short helices (Figure 2E), although it is a pseudo-knot structure; correspondingly, the RMSD of the predicted native structure does not increase when more than 60% restraints are imposed (Figure 2C).

### The folding landscape by using Gō model

We select a three-way junction from the Varkud Satellite Ribozyme (67) (PDB ID: 2MTJ) as a case to interrogate how a Gō model constrains the folding landscape during simulations. The simulations with no restraints show a single free energy minimum (Figure 3A), which is referred to as a distal-native state (DN). The simulations with 40% restraints (Figure 3B) show both a DN state and a near-native state (NN), suggesting that the imposed restraints are capable of shifting the sampling space of the simulations toward the NN states. The RMSD of the NN state is ∼3.37 Å (Figure 3D), and that of the DN state is ∼9.96 Å (Figure 3E). Almost no energy barrier exists between the two states, suggesting that the DN and NN states could be inter-converted readily. The simulations with 80% restraints (Figure 3C) show separated NN state and DN state, indicating that the two states are hard to inter-convert due to the kinetic barriers formed by the additional restraints. Once the structure situates in one state, it is prone to being trapped due to the higher free energy barrier (white dashed box in Figure 3C). The dissection of the contacts in the native structure, the NN and the DN structures (Figure 3F) suggests that a portion of the contacts (red dashed box in Figure 3F) do not readily form even though they are imposed as restraints. The existence of the free energy barrier hinders the conversion of the structure to a near-native state, in which all contacts are formed.

### DVASS strategy for restraints selection

We have demonstrated that the usage of ∼40–60% restraints yields optimum structures in RNA structure prediction. We will utilize the DVASS algorithm to further reduce the minimum required number of restraints. We hypothesize that the extent of distance variation is directly correlated with its ability to improve the prediction accuracy of RNA tertiary structure. We test our hypothesis and subsequently the potency of DVASS using the same dataset of 22 RNAs. Initially, we extract all native contacts from these RNAs as restraints, calculate distance variation between every two bases by using DVASS and sort all the restraints according to distance variation. The larger the distance variation, the higher the corresponding constraint is ranked. Subsequently, we divide the ranked restraints into five groups, where each group consists of the top 20%, 20–40%, 40–60%, 60–80% and 80–100% restraints, respectively. In addition, we consider a control group with no restraints. We perform 20 predicting attempts for each of the groups, and the average RMSDs are shown in Figure 4A. The average normalized RMSD reaches a minimum when the top 20% ranked restraints are employed, and RMSD gradually increases when low ranked restraints are used (Figure 4A). The positive correlation (Figure 4A) between the average normalized RMSD, standing for the prediction accuracy, and the rank of restraints, standing for the extent of distance variation, attests to our hypothesis that the extent of distance variation is correlated with its ability to improve the prediction accuracy. The RMSDs of 3RG5 (Figure 4E) generated by using different groups of restraints present the same trend as the overall average normalized RMSD. It is intriguing to find that by using only the top 20% ranked restraints, we could achieve an RMSD as low as that obtained by using 60% unsorted restraints (Figure 4C and E).

We use 3RG5 as an example to demonstrate how differently ranked restraints are typically distributed in the RNA structure. The secondary structure of 3RG5 (Figure 4D) is a 5-way junction encompassing a pseudo-knot between

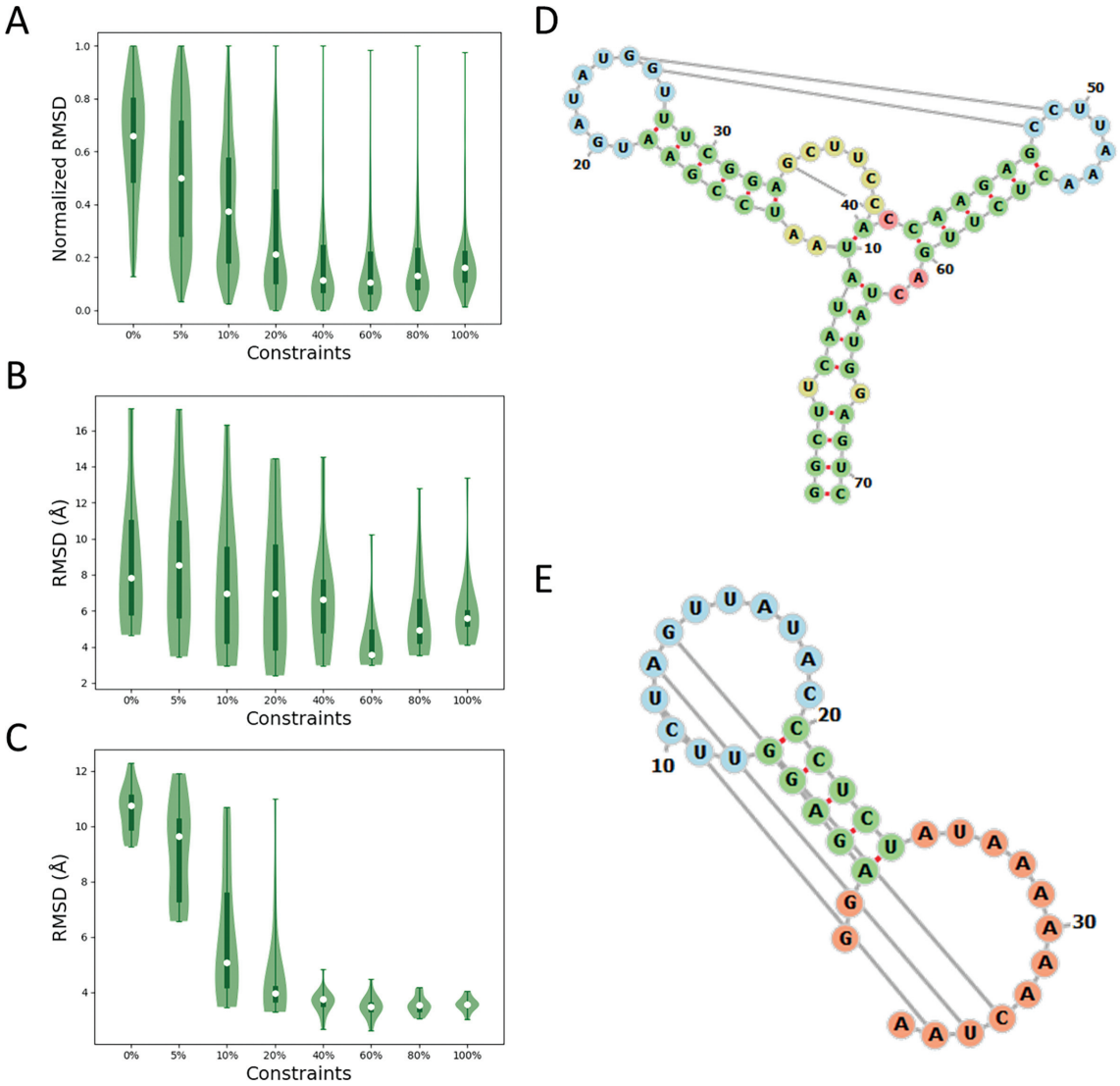


**Figure 2.** Relationship between the normalized RMSD and the percentage of imposed restraints. (**A**) The average normalized RMSD of the 22 RNAs when using 0%, 5%, 10%, 20%, 40%, 60%, 80% and 100% restraints, respectively. (**B**) The RMSD of 3LA5 when using different fractions of restraints. (**C**) The RMSD of 2L1V when using different fractions of restraints. (**D**) The secondary structure of 3LA5. (**E**) The secondary structure of 2L1V.

residue 20 and 70. Six different residue pairs are colored by red, green, light green, pink, blue and cyan, respectively. The red residue pair [1, 37] is predicted to have the largest distance variation. Residues 1 and 37 are located at the very end of the two longest stems in the junction. The blue [7, 79] and cyan [20, 69] residue pairs are predicted to have the lowest distance variations. Residues 7 and 79 are spaced by only one residue, which makes sense of having a low distance variation. Although the location of residue 20 and residue 69 is similar to that of residues 1 and 37, which are also located at the very end of two stems, the existence of the pseudo-knot [20, 70] makes the two residues spatially close to each other. Apparently, a high-ranked restraint such as the one in residue pair [1, 37] affects the spatial arrangement of the two longest stems, while a low-ranked restraint in residue pair [7, 79] or [20, 69] only affects local structures, and restraints in other residue pairs colored in Figure 4D affect the structure of 3RG5 to various extent.

## Predict the number of native contacts

Although using 20% top ranked restraints is demonstrated to be sufficient to obtain a satisfactory prediction accuracy, the percentage is relative to the total number of native contacts. In most cases we lack *a priori* knowledge of the tertiary structure, thus we cannot calculate the total number of native contacts to further derive the minimum needed number of restraints for optimum prediction accuracy. Hence, it is necessary to estimate the number of native contacts given only the sequence and the secondary structure. Therefore, we extend the DVASS algorithm further (see 'Materials and Methods' section) to predict the number of contacts from RNA secondary structure. We test our prediction ability in 22 RNAs based on their secondary structural data. Our analyses provide a direct correlation between the predicted number of contacts and the number of native contacts (Figure 4B), thereby validating our methodology.

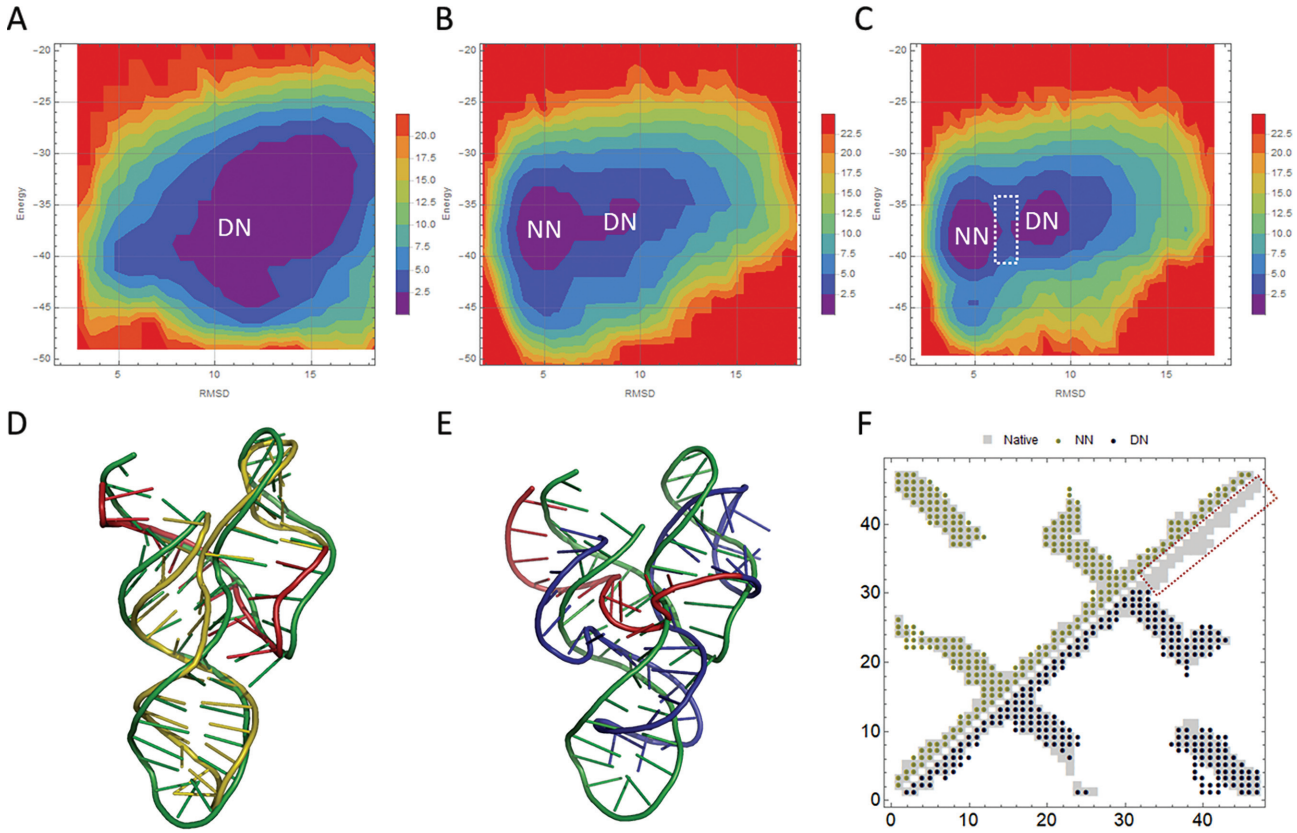


**Figure 3.** Thermodynamic analyses of a three-way junction in Varkud Satellite Ribozyme. (**A, B** and **C**) The free energy landscapes of 2MTJ obtained by imposing no restraints (A), 40% restraints (B) and 80% restraints (C), respectively. NN refers to the near-native structure, and DN refers to the distal-native structure. The white box refers to a free energy barrier that impedes the inter-conversion of the NN and DN states. The landscapes are derived from the potential of mean force of RMSD and iFoldRNA energies. The color bar represents the relative Helmholtz free energy in kcal/mol. (**D**) Overlap of the crystal structure (green) of 2MTJ and the NN structure (olive) with an RMSD of 3.37 Å. The red color refers to the red boxed region in (F). (**E**) Overlap of the crystal structure (green) of 2MTJ and the DN structure (blue) with an RMSD of 9.96 Å. The red color refers to the red boxed region in (F). (**F**) The comparison of the contacts in the native structure (grey square), the NN structure (olive circle) and the DN structure (blue circle) of 2MTJ.

## DISCUSSION

We find that the reduction in the number of restraints does not negatively influence the prediction accuracy, while utilization of 100% restraints does not necessarily improve the accuracy of RNA tertiary predictions. Hence, we can choose only a fraction of the restraints (40–60%) to reduce the overhead caused by restraints. The application of DVASS algorithm can further reduce the number of restraints to be imposed by outlining crucial restraints for structure prediction. Through our analyses, we observed that a portion of restraints does not play a significant role in improving the prediction accuracy (Figures 3D and 4A). In even worse cases (Figure 4E), imposing the lowest ranked (80–100%) restraints result in higher RMSD than that produced by not using any constraint. This finding sheds light on the phenomenon that imposing 100% restraints does not necessarily lead to better prediction accuracy, because some restraints play negligible or even negative role on improving the prediction accuracy. Hence, we can safely ignore such restraints to unload computational burden and to improve the speed of calculation. Therefore, using only a subset of all possible restraints that are beneficial for structure prediction is a rational strategy to reduce the time overhead and improve the prediction accuracy.

We observe a strong correlation between distance variation and the capacity of restraints to improve the prediction accuracy (Figure 4A), which could be rationalized from a conformational space perspective. Suppose all distances between all residues are given definite fixed values, the conformational space would thus contain only one structure. If the distance between two certain residues is no longer fixed, rather it varies in a range, the conformational space then extends to some allowed volume. The larger is the distance variation, the larger is the allowed conformational space, the smaller is the probability of finding the native structure upon sampling, the more likely the prediction accuracy will be low. Therefore, large distance variations strongly affect prediction accuracy compared to that of small distance variations.

The DVASS algorithm, described in the 'Materials and Methods' section, is not limited to RNA. We initially extract all the helices from the 22 RNA structures and subsequently obtain inter-residue distances from the helices (Figure 1C). In addition, based on the chain connectivity of RNA strands, we acquire distance information between adjacent residues. Then, we employ DVASS to derive the maxi-

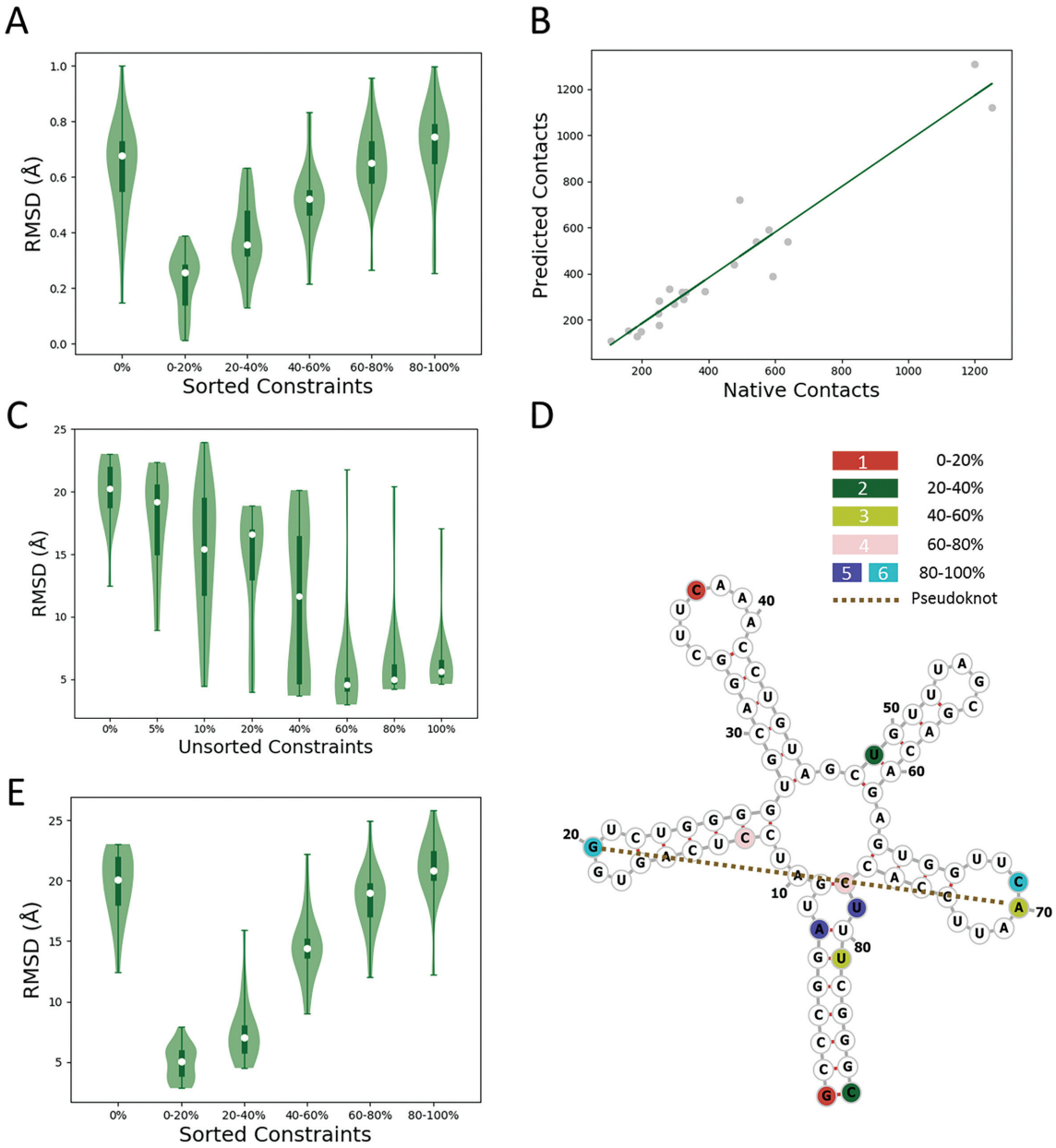


**Figure 4.** DVASS Algorithm Test Results in the 22 RNAs Dataset. (**A**) The average normalized RMSD of the 22 RNAs when using 0%, 0–20%, 20–40%, 40–60%, 60–80% and 80–100% top ranked restraints. (**B**) Predicted number of contacts versus the number of native contacts. The green line refers to the linear regression of the correlation between the predicted and native contacts, which is: $N_{predicted} = 0.989 \times N_{native} - 13.9$, where $N_{predicted}$ is the predicted number of contacts and $N_{native}$ is the genuine number of contacts. The Pearson Correlation Coefficient is 0.962. The *P*-value is $3.54 \times 10^{-12}$. (**C**) The RMSD of 3RG5 obtained by imposing 0%, 5%, 10%, 20%, 40%, 60%, 80% and 100% unsorted restraints. (**D**) The secondary structure of 3RG5. Restraints in residue pairs colored by red, green, light green, pink, blue and cyan are predicted to be located in top 0–20%, 20–40%, 40–60%, 60–80%, 80–100% and 80–100% ranked area, respectively. The brown dashed line refers to the pseudoknot. (**E**) The average RMSD of 3RG5 obtained by imposing 0%, 0–20%, 20–40%, 40–60%, 60–80% and 80–100% top ranked restraints.

mum and minimum distances between other residues in the RNA structure. Thus, the DVASS algorithm is essentially utilized to deduce other distance restraints from a given set of distance restraints. Hence, it is apparent that the DVASS algorithm can be not only used to predict the importance of RNA restraints, but also technically used to derive the importance of restraints used in protein structure prediction.

Gō model was initially proposed in protein folding research (68). While Gō model is computationally efficient, its application has been limited by several known shortcomings. First, the details of the computational folding kinetics mechanisms are not always exactly the same as experiments (69–72). Second, the simulated folding temperature in the Gō model appears to show a strong dependence on the number of native contacts and the number of residues (53). The utilization of Gō model in our work is not aimed at studying the details of RNA folding kinetics mechanisms, but rather on the thermodynamic outcomes—the predic-

tion of the native structural ensemble of RNAs. Hence, these shortcomings of Gō model have no ramification on our work. Interestingly, the formation of kinetic traps for some RNAs (Figure 3) may, in some cases, be attributed to Gō model artifacts. Kinetic traps might be the reason behind the anti-correlation between larger than optimal number of restraints and the extent of the improvement of RNA tertiary structure prediction accuracy. The utilization of DVASS algorithm circumvents this shortcoming to some extent by significantly reducing the number of restraints to decrease the odds of forming kinetic traps. Hence, amalgamating Gō model and DVASS algorithm provides an avenue of efficiently and accurately predicting RNA tertiary structure. In the future, we plan to propose more sophisticated strategies to more confidently subdue the shortcomings of Gō model. For example, an ostensibly reasonable and viable method is to tune the order in which the restraints are applied in the simulation. It is well known that the assembly of ribosomal RNA (rRNA) is a complex process with multiple assembly steps at different locations within the cell (73,74). If we can determine how the order of the formation of contacts affects the folding of rRNA, we may be able to computationally mimic the folding process or even predict the structure of ribosomal RNA efficiently, which is not currently feasible.

## DATA AVAILABILITY

The DVASS algorithm, the 22 RNAs dataset and the distance variations of the 22 RNAs can be downloaded from: http://dokhlab.org/dokhlab/download/dvass.tar.gz or https://bitbucket.org/dokhlab/dvass

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## FUNDING

We gratefully acknowledge the G. Thomas Passananti endowment and U.S. National Institutes of Health [R01 GM123238–01, R01 GM064803–12]. Funding for open access charge: National Institutes of Health [R01 GM123238–01, R01 GM064803–12].
*Conflict of interest statement.* None declared.

## REFERENCES

1. Huang,L. and Lilley,D.M.J. (2018) Structure and ligand binding of the SAM-V riboswitch. *Nucleic Acids Res.*, **46**, 6869–6879.
2. Peselis,A. and Serganov,A. (2018) ykkC riboswitches employ an add-on helix to adjust specificity for polyanionic ligands. *Nat. Chem. Biol.*, **14**, 887–894.
3. Huang,L., Wang,J. and Lilley,D.M.J. (2016) A critical base pair in k-turns determines the conformational class adopted, and correlates with biological function. *Nucleic Acids Res.*, **44**, 5390–5398.
4. Williams Benfeard,I.I., Zhao,B., Tandon,A., Ding,F., Weeks,K.M., Zhang,Q. and Dokholyan,N.V. (2017) Structure modeling of RNA using sparse NMR constraints. *Nucleic Acids Res.*, **45**, 12638–12647.
5. Sharma,S., Ding,F. and Dokholyan,N.V. (2008) iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics*, **24**, 1951–1952.
6. Cao,S. and Chen,S.-J. (2011) Physics-based de novo prediction of RNA 3D structures. *J. Phys. Chem. B*, **115**, 4216–4226.
7. Shi,Y.Z., Wang,F.H., Wu,Y.Y. and Tan,Z.J. (2014) A coarse-grained model with implicit salt for RNAs: predicting 3D structure, stability and salt effect. *J. Chem. Phys.*, **141**, 105102.
8. Kerpedjiev,P., Zu Siederdissen,C.H. and Hofacker,I.L. (2015) Predicting RNA 3D structure using a coarse-grain helix-centered model. *RNA*, **21**, 1110–1121.
9. Miao,Z., Adamiak,R.W., Antczak,M., Batey,R.T., Becka,A.J., Biesiada,M., Boniecki,M.J., Bujnicki,J.M., Chen,S.-J.J., Cheng,C.Y. *et al.* (2017) RNA-Puzzles Round III: 3D RNA structure prediction of five riboswitches and one ribozyme. *RNA*, **23**, 655–672.
10. Miao,Z., Adamiak,R.W., Blanchet,M.-F.M-F., Boniecki,M., Bujnicki,J.M., Chen,S.-J., Cheng,C., Chojnowski,G., Chou,F.-C., Cordero,P. *et al.* (2015) RNA-Puzzles Round II: assessment of RNA structure prediction programs applied to three large RNA structures. *RNA*, **21**, 1066–1084.
11. Cruz,J.A., Blanchet,M.F., Boniecki,M., Bujnicki,J.M., Chen,S.J., Cao,S., Das,R., Ding,F., Dokholyan,N.V., Flores,S.C. *et al.* (2012) RNA-Puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction. *RNA*, **18**, 610–625.
12. Wang,J., Mao,K., Zhao,Y., Zeng,C., Xiang,J., Zhang,Y. and Xiao,Y. (2017) Optimization of RNA 3D structure prediction using evolutionary restraints of nucleotide–nucleotide interactions from direct coupling analysis. *Nucleic Acids Res.*, **45**, 6299–6309.
13. Zhao,Y.J.Y., Wang,J., Chen,X.W., Luo,H.T., Zhao,Y.J.Y., Xiao,Y. and Chen,R.S. (2013) Large-scale study of long non-coding RNA functions based on structure and expression features. *Sci. China Life Sci.*, **56**, 953–959.
14. Krokhotin,A., Houlihan,K. and Dokholyan,N.V. (2015) iFoldRNA v2: Folding RNA with constraints. *Bioinformatics*, **31**, 2891–2893.
15. Boniecki,M.J., Lach,G., Dawson,W.K., Tomala,K., Lukasz,P., Soltysinski,T., Rother,K.M. and Bujnicki,J.M. (2016) SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Res.*, **44**, e63.
16. Popenda,M., Szachniuk,M., Antczak,M., Purzycka,K.J., Lukasiak,P., Bartol,N., Blazewicz,J. and Adamiak,R.W. (2012) Automated 3D structure composition for large RNAs. *Nucleic Acids Res.*, **40**, e112.
17. Rother,M., Rother,K., Puton,T. and Bujnicki,J.M. (2011) ModeRNA: a tool for comparative modeling of RNA 3D structure. *Nucleic Acids Res.*, **39**, 4007–4022.
18. Parisien,M. and Major,F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, **452**, 51–55.
19. Das,R. and Baker,D. (2007) Automated de novo prediction of native-like RNA tertiary structures. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 14664–14669.
20. Das,R., Karanicolas,J. and Baker,D. (2010) Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat. Methods*, **7**, 291–294.
21. Zhao,Y., Huang,Y., Gong,Z., Wang,Y., Man,J. and Xiao,Y. (2012) Automated and fast building of three-dimensional RNA structures. *Sci. Rep.*, **2**, 734.
22. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
23. Sato,K., Hamada,M., Asai,K. and Mituyama,T. (2009) CentroidFold: a web server for RNA secondary structure prediction. *Nucleic Acids Res.*, **37**, 277–280.
24. Puton,T., Kozlowski,L.P., Rother,K.M. and Bujnicki,J.M. (2013) CompaRNA: a server for continuous benchmarking of automated methods for RNA secondary structure prediction. *Nucleic Acids Res.*, **41**, 4307–4323.
25. Wang,J., Zhao,Y., Zhu,C. and Xiao,Y. (2015) 3dRNAscore: a distance and torsion angle dependent evaluation function of 3D RNA structures. *Nucleic Acids Res.*, **43**, e63.
26. Morcos,F., Pagnani,A., Lunt,B., Bertolino,A., Marks,D.S., Sander,C., Zecchina,R., Onuchic,J.N., Hwa,T. and Weigt,M. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, E1293–E1301.
27. De Leonardis,E., Lutz,B., Ratz,S., Cocco,S., Monasson,R., Schug,A. and Weigt,M. (2015) Direct-Coupling Analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. *Nucleic Acids Res.*, **43**, 10444–10455.
28. Smola,M.J. and Weeks,K.M. (2018) In-cell RNA structure probing with SHAPE-MaP. *Nat. Protoc.*, **13**, 1181–1195.

29. Homan,P.J., Favorov,O.V., Lavender,C.A., Kursun,O., Ge,X., Busan,S., Dokholyan,N.V. and Weeks,K.M. (2014) Single-molecule correlated chemical probing of RNA. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 13858–13863.
30. Ding,F., Lavender,C.A., Weeks,K.M. and Dokholyan,N.V. (2012) Three-dimensional RNA structure refinement by hydroxyl radical probing. *Nat. Methods*, **9**, 603–608.
31. Cheng,C.Y., Kladwang,W., Yesselman,J.D. and Das,R. (2017) RNA structure inference through chemical mapping after accidental or intentional mutations. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, 9876–9881.
32. Krokhotin,A., Mustoe,A.M., Weeks,K.M. and Dokholyan,N.V. (2017) Direct identification of base-paired RNA nucleotides by correlated chemical probing. *RNA*, **23**, 6–13.
33. Magnus,M., Boniecki,M.J., Dawson,W. and Bujnicki,J.M. (2016) SimRNAweb: a web server for RNA 3D structure modeling with optional restraints. *Nucleic Acids Res.*, **44**, W315–W319.
34. Sripakdeevong,P., Cevec,M., Chang,A.T., Erat,M.C., Ziegeler,M., Zhao,Q., Fox,G.E., Gao,X., Kennedy,S.D., Kierzek,R. *et al.* (2014) Structure determination of noncanonical RNA motifs guided by 1H NMR chemical shifts. *Nat. Methods*, **11**, 413–416.
35. Wang,J. and Xiao,Y. (2017) Using 3dRNA for RNA 3-D structure prediction and evaluation. *Curr. Protoc. Bioinform.*, **57**, 5.9.1–5.9.12.
36. Jonikas,M.A., Radmer,R.J., Laederach,A., Das,R., Pearlman,S., Herschlag,D. and Altman,R.B. (2009) Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA*, **15**, 189–199.
37. Cole,D.I., Legassie,J.D., Bonifacio,L.N., Sekaran,V.G., Ding,F., Dokholyan,N.V. and Jarstfer,M.B. (2012) New Models of Tetrahymena Telomerase RNA from Experimentally Derived Constraints and Modeling. *J. Am. Chem. Soc.*, **134**, 20070–20080.
38. Chen,Y., Ding,F. and Dokholyan,N.V. (2007) Fidelity of the protein structure reconstruction from inter-residue proximity constraints. *J. Phys. Chem. B*, **111**, 7432–7438.
39. Ueda,Y., Taketomi,H., Go,N., Ueda,Y. and Gō,N. (1975) Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effects of specific amino acid sequence represented by specific inter-unit interactions. *Int. J. Pept. Protein Res.*, **7**, 445–459.
40. Chen,Y., Ding,F., Nie,H., Serohijos,A.W., Sharma,S., Wilcox,K.C., Yin,S. and Dokholyan,N.V. (2008) Protein folding: then and now. *Arch. Biochem. Biophys.*, **469**, 4–19.
41. Ding,F. and Dokholyan,N.V. (2005) Simple but predictive protein models. *Trends Biotechnol.*, **23**, 450–455.
42. Dobson,C.M. (2003) Protein folding and misfolding. *Nature*, **426**, 884–890.
43. Nicholls,A., Sharp,K.A. and Honig,B. (1991) Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins Struct. Funct. Bioinforma.*, **11**, 281–296.
44. Dill,K.A., Bromberg,S., Yue,K., Chan,H.S., Ftebig,K.M., Yee,D.P. and Thomas,P.D. (1995) Principles of protein folding—a perspective from simple exact models. *Protein Sci.*, **4**, 561–602.
45. Dill,K.A. and MacCallum,J.L. (2012) The protein-folding problem, 50 years on. *Science*, **338**, 1042–1046.
46. Clementi,C. (2008) Coarse-grained models of protein folding: toy models or predictive tools? *Curr. Opin. Struct. Biol.*, **18**, 10–15.
47. Camacho,C.J. and Thirumalai,D. (1993) Kinetics and thermodynamics of folding in model proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **90**, 6369–6372.
48. Onuchic,J.N. and Wolynes,P.G. (2004) Theory of protein folding. *Curr. Opin. Struct. Biol.*, **14**, 70–75.
49. Klimov,D.K. and Thirumalai,D. (2000) Mechanisms and kinetics of β-hairpin formation. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 2544–2549.
50. Chen,Y., Campbell,S.L. and Dokholyan,N.V. (2007) Deciphering protein dynamics from NMR data using explicit structure sampling and selection. *Biophys. J.*, **93**, 2300–2306.
51. Ding,F., Sharma,S., Chalasani,P., Demidov,V.V., Broude,N.E. and Dokholyan,N.V. (2008) Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA*, **14**, 1164–1173.
52. Gherghe,C.M., Leonard,C.W., Ding,F., Dokholyan,N.V. and Weeks,K.M. (2009) Native-like RNA tertiary structures using a sequence-encoded cleavage agent and refinement by discrete molecular dynamics. *J Am Chem Soc*, **131**, 2541–2546.
53. Chavez,L.L., Onuchic,J.N. and Clementi,C. (2004) Quantifying the roughness on the free energy landscape: entropic bottlenecks and protein folding rates. *J. Am. Chem. Soc.*, **126**, 8426–8432.
54. Chen,S.-J. and Dill,K.A. (2000) RNA folding energy landscapes. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 646–651.
55. Oliveberg,M. and Wolynes,P.G. (2005) The experimental survey of protein-folding energy landscapes. *Q. Rev. Biophys.*, **38**, 245–288.
56. Nussinov,R. and Wolynes,P.G. (2014) A second molecular biology revolution? The energy landscapes of biomolecular function. *Phys. Chem. Chem. Phys.*, **16**, 6321–6322.
57. Bryngelson,J.D. and Wolynes,P.G. (1987) Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. U.S.A.*, **84**, 7524–7528.
58. Proctor,E.A. and Dokholyan,N.V. (2016) Applications of Discrete Molecular Dynamics in biology and medicine. *Curr. Opin. Struct. Biol.*, **37**, 9–13.
59. Ding,F., Tsao,D., Nie,H. and Dokholyan,N.V. (2008) Ab initio folding of proteins with All-Atom discrete molecular dynamics. *Structure*, **16**, 1010–1018.
60. Proctor,E.A., Ding,F. and Dokholyan,N.V. (2011) Discrete molecular dynamics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, **1**, 80–92.
61. Dokholyan,N.V., Buldyrev,S.V., Stanley,H.E. and Shakhnovich,E.I. (1998) Discrete molecular dynamics studies of the folding of a protein-like model. *Fold. Des.*, **3**, 577–587.
62. Shirvanyants,D., Ding,F., Tsao,D., Ramachandran,S. and Dokholyan,N.V. (2012) Discrete molecular dynamics: an efficient and versatile simulation method for fine protein characterization. *J. Phys. Chem. B*, **116**, 8375–8382.
63. Sugita,Y. and Okamoto,Y. (1999) Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, **314**, 141–151.
64. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
65. Hoogsteen,K. (1963) The crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine. *Acta Crystallogr.*, **16**, 907–916.
66. Barton,G.J. (2002) *OC-A cluster analysis program*. Univ. Dundee, Scotland.
67. Bonneau,E. and Legault,P. (2014) Nuclear magnetic resonance structure of the III–IV–V three-way junction from the Varkud Satellite ribozyme and identification of magnesium-binding sites using paramagnetic relaxation enhancement. *Biochemistry*, **53**, 6264–6275.
68. Go,N. (1983) Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.*, **12**, 183–210.
69. Clementi,C., Nymeyer,H. and Onuchic,J.N. (2000) Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.*, **298**, 937–953.
70. Paci,E., Vendruscolo,M. and Karplus,M. (2002) Validity of Gō models: comparison with a solvent-shielded empirical energy decomposition. *Biophys. J.*, **83**, 3032–3038.
71. Daggett,V. and Fersht,A. (2003) The present view of the mechanism of protein folding. *Nat. Rev. Mol. Cell Biol.*, **4**, 497–502.
72. Kaya,H. and Chan,H.S. (2002) Towards a consistent modeling of protein thermodynamic and kinetic cooperativity: how applicable is the transition state picture to folding and unfolding? *J. Mol. Biol.*, **315**, 899–909.
73. Mulder,A.M., Yoshioka,C., Beck,A.H., Bunner,A.E., Milligan,R.A., Potter,C.S., Carragher,B. and Williamson,J.R. (2010) Visualizing ribosome biogenesis: parallel assembly pathways for the 30S subunit. *Science*, **330**, 673–677.
74. Calidas,D. and Culver,G.M. (2011) Interdependencies govern multidomain architecture in ribosomal small subunit assembly. *RNA*, **17**, 263–277.