Journal of
Chemical Information
and Modeling

Article

pubs.acs.org/jcim

Prospects for Tertiary Structure Prediction of RNA Based on Secondary Structure Information

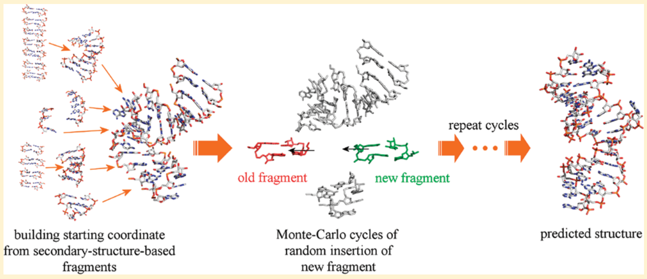
Satoshi Yamasaki,*[†] Shugo Nakamura,[‡] and Kazuhiko Fukui[†]

[†]Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan

[‡]Department of Biotechnology, Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan

Supporting Information

ABSTRACT: We developed a method, called RNA Assembler using Secondary Structure Information Effectively (RASSIE), for predicting RNA tertiary structures using known secondary structure information. We attempted a fragment assembly-based method that uses a secondary structure-based fragment library. For several typical target structures such as stem-loops, bulge-loops, and 2-way junctions, our method provided numerous good quality candidate structures in less computational time than previously proposed methods. By using a high-resolution potential energy function, we were able to select good predicted structures from candidate structures. This method of efficient conformational search and detailed structure evaluation using high-resolution potential is potentially useful for the tertiary structure prediction of RNA.


repeat cycles
old fragment
new fragment
building starting coordinate from secondary-structure-based fragments
Monte-Carlo cycles of random insertion of new fragment
predicted structure

1. INTRODUCTION

The role of functional RNA is important for several essential biological functions, such as the regulation of gene expression and degradation of mRNA. Many kinds of functional RNA have already had their functional roles confirmed in biochemical experiments but few of their tertiary structures have so far been determined. The structural complexity and flexibility of RNA tends to make its structural determination exceptionally difficult.[1] However, tertiary structures are needed to perform detailed analyses, such as using molecular or docking simulations, of RNA functions.

This has prompted several researchers to attempt to develop computational methods for predicting the tertiary structures of RNA molecules from their nucleotide sequences. The Fragment Assembly of RNA (FARNA), which is one of the major functions of the ROSETTA software package developed by Das and Baker[2] and Das et al.,[3] is one of the most successful methods for predicting RNA tertiary structures. FARNA is an extension of their fragment assembly methods for protein structure prediction that has demonstrated high efficiency. In this method, Metropolis Monte Carlo cycles are performed, replacing the structure of a short nucleotide fragment with a new one from a predefined structural library. The advantages of this method are this conformational searching algorithm and the scoring functions used in the cycle.

MC-Fold and MC-Sym, developed by Parisien et al., is also an excellent method for predicting RNA secondary structures by using small RNA tertiary structure fragment units.[4] They successfully used this method to improve the performance of secondary structure prediction and also demonstrated that their method provided a candidate tertiary structure by connecting fragment structures in ways that are consistent with the predicted secondary structure. There are a few other methods for predicting the tertiary structures of RNAs,[5,6] but there is still room for progress in overcoming the limitations of prediction accuracy and size of RNA.

One means of improving prediction accuracy is to use secondary structure information. The secondary structure of RNA can be determined by single-point mutagenesis, cross-linking, or comparing sequences. Moreover, studies predicting the secondary structure of RNA have been proceeding for many years, and now many high-performance tools for secondary structure prediction of RNA are available.[7−9] The accuracy of secondary structure information has been improving, so it is very important to know how to use that information when modeling or predicting tertiary structures.

Some of the tertiary structure prediction methods developed previously can use the secondary structure information as input. In FARNA of ROSETTA (version 3.1), the input secondary structure information is converted to the constraints between base pairs. Energy functions then serve to fill these constraints during the model generation step. However, the conformational spaces, which are searched during fragment insertion steps, are not narrowed down in this process, regardless of the input of secondary structure information, and many kinds of tertiary

Received: July 22, 2011
Published: January 13, 2012

structures were generated that were not matched to input secondary structures.

In MC-Fold and MC-Sym,[4] many types of structure classes (nucleic cyclic motifs (NCMs)) are defined according to backbone linkage, type of base-stacking, and type of base-pairing. Parisien et al. select numerous NCM fragment structures from known structure databases and calculate the probability density of each NCM class to derive their score. Their classification was appropriate, but detailed classification resulted in several empty or rarefaction classes, making it difficult to predict structures that include such empty or rarefaction NCM classes. This problem can be solved if the number of known structures increases sufficiently, but this will take time.

We developed a method, called RNA Assembler using Secondary Structure Information Effectively (RASSIE), for predicting RNA tertiary structures using known secondary structure information. A fragment library is generated based on secondary structure information, and fragments from the library are assembled to model tertiary structures. As shown in Figure 1, we used three types of secondary structures: helix-duplet, hairpin loop, and single-stranded loop. A total of 557 known RNA structures were disassembled, and a secondary structure-based fragment library was derived on the basis of the secondary structures. Tertiary structure prediction was conducted by assembling these fragment structures, like a puzzle, following the blueprint of the input secondary structures. Metropolis Monte Carlo cycles of insertions of secondary structure-based fragments with several low-resolution scoring functions served to focus on the optimum structure. We were able to predict the tertiary structure of several typical RNA structures by using this technique. We also conducted a structural evaluation using the high-resolution potential energy function to assess and improve the predicted results.

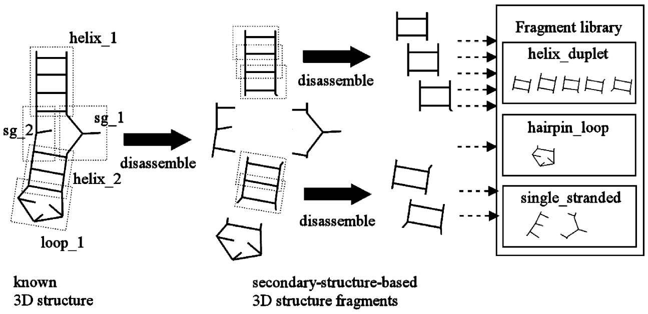


Figure 1. Simplified diagram illustrating construction of fragment library. Known structures were disassembled into three types of secondary structure-based fragments.

2. METHODS

2.1. Deriving a Secondary Structure Fragment Library Used in RASSIE.

We used 557 nonredundant RNA structures to derive a secondary structure-based fragment library. The process for deriving a secondary structure fragment library is summarized in Figure 1. We first used the RNA STRAND database (http://www.rnasoft.ca/strand/)[10] using the searching options of "non-redundant sequences only" and "RNA length less than 1000" as our source of tertiary structures. We obtained 626 RNA sets. Next, entries that included noncanonical residues were manually excluded from this set. Additionally, a large ribosomal subunit structure, 1JJ2,[11] was added to this set as being typical of structures of more than 1000 nucleotides in length. Finally, we derived 557 nonredundant RNA sets. These sets included both the RNA in free state and the RNA in complex with protein or small ligands. The coordinates of all these RNAs were derived from the Protein Data Bank.[12]

Three types of secondary structures, "helix", "loop" (hairpin loop), and "SG" (single-stranded region), assigned in the RNAML files generated using the RNAView program,[13,14] were used for classification of the fragments. Each tertiary structure was disassembled into fragments according to assigned secondary structure names and classified into three categories: helix fragments, loop fragments, and SG fragments. For loop and SG fragments, neighboring (±1) nucleotides were also included in the fragment data for use in overlap width. The helix fragments were again disassembled into several duplet fragments, which we called "helix-duplet fragments." For example, the helix fragment of $d(AATT)_2$ was disassembled into three helix-duplet fragments: AA-TT, AT-AT, and TT-AA (N_1N_2−N_3N_4 implies the helix-duplet of strands 5′-N_1N_2-3′ and 5′-N_3N_4-3′). Using these procedures, we derived 13802 helix-duplet fragments, 595 loop fragments, and 1870 SG fragments from the initial 557 nonredundant RNA set. All these fragment structures in each category were again categorized into

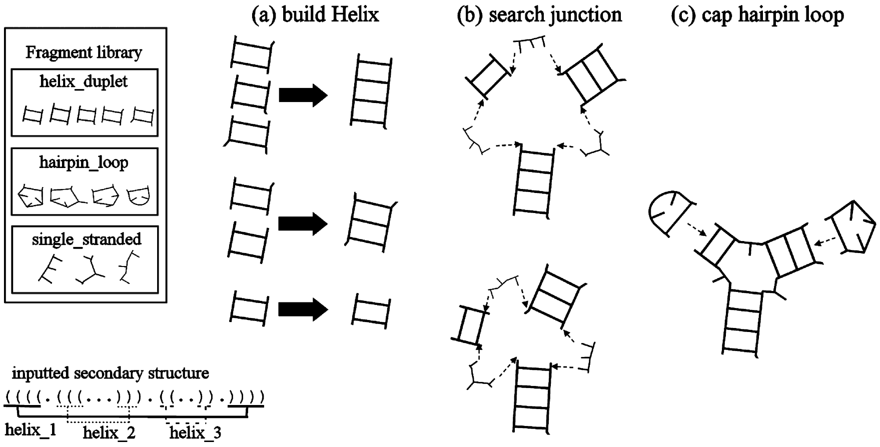


Figure 2. Flowchart of initial building stage.

sequence-based subcategories described by “R” (purine) or “Y” (pyrimidine).

2.2. Detailed Description of Conformational Sampling Process of RASSIE: Generating Tertiary Structure by Assembling Secondary Structure-Based Fragment Structures. We describe the process of generating one candidate structure in RASSIE, which was achieved by assembling the secondary structure fragments according to a given nucleotide sequence and secondary structure information. This process can be divided into two stages: initial building and fragment insertion. First, we briefly describe the flow and algorithm of these stages.

2.2.1. Initial Building Stage. The flowchart of the initial building stage is shown in Figure 2. This stage starts from building the helix region by assembling the helix duplet fragments (Figure 2a). For the helix of d(RYR)$_2$, for example, two helix-duplet fragments, RY-RY and YR-YR, were randomly selected from the helix-duplet fragment library. These contiguous helix-duplet fragments were connected by superimposing the overlapped base pairs (Y·R base pairs). If the root-mean-square deviation (rmsd) of the overlapped base pairs was more than the threshold (1.0 Å), the fragment was rejected, and another fragment was again randomly selected from the library. The rmsd during structure sampling was calculated using P, C3, C4, C5, O3, O5, O4, C1, C2, and N9 (for purines) or N1 (for pyrimidines). For longer helix fragments, the fragment could be extended by iterating this process from one helix end to the other.

If necessary, helix fragments were connected by the SG fragments according to their known secondary structure (Figure 2b). This was achieved by superimposing the overlapped base pairs of SG and helix fragments. Many sets of SG fragments derived from the SG fragment library were tried to connect the helix fragments iteratively, while rmsd values of each overlapped bases became less than the threshold. The threshold we used was 5.0 Å.

Finally, the loop fragments were assembled with the helix fragments. This was achieved by superimposing the loop-closing base pairs of the loop fragment on the base pair of the edge of the helix fragment (Figure 2c). The rmsd of the overlapped base pairs was also checked (threshold = 1.0 Å). Once the overall structure was derived, the steric clashes among all atom−atom pairs were checked. If an atom pair close to less than 1.0 Å was found, the overall structure was reset, and the process started back to the “build helix” process. At this stage, no scoring function was used to evaluate the intermediate structures. The criteria for determining whether the intermediate structure was accepted or rejected were “the rmsd between overlapped base pairs” and “with or without any steric clash”.

2.2.2. Fragment Insertion Stage. The structures derived at the initial building stage were optimized at the fragment insertion stage. The brief diagram of this stage is shown in Figure 3. A randomly selected secondary structure fragment region was replaced with a new fragment derived from the secondary structure fragment library. When a loop or SG fragment was selected, the previous fragment was replaced with a new one in the same manner as in the initial building stage. When the helix-duplet fragment was selected, the helix fragment was separated from the selected helix-duplet fragment and reassembled with a new one. The energies of the previous and new structures were calculated using the energy functions described in the next section, and the new structure was accepted or rejected using the Metropolis criterion. These Metropolis Monte Carlo cycles were repeated hundreds or thousands of times. After this fragment insertion stage, one candidate structure was derived.

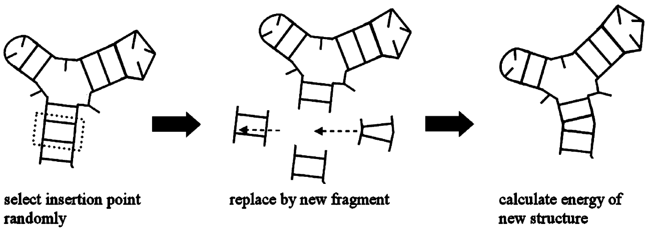

select insertion point randomly
replace by new fragment
calculate energy of new structure

Figure 3. Simplified diagram showing fragment insertion.

2.2.3. RASSIE Energy Functions. We used the following five energy functions to evaluate the intermediate structures of the fragment insertion stage (Figure 3). The summation of these energy values becomes the total energy and it is optimized during the fragment insertion process.

1. Radius of Gyration. This function achieves sufficient compactness of the RNA molecule. Using the heavy atoms of the RNA molecule, the radius of gyration (R_g) was calculated from the following equation:

$$R_g = \sum m_i(r_i - R_C)^2/M \tag{1}$$

where m_i is the mass of i-th atom, r_i is its coordinates, R_C is the coordinates of the center of mass, and M is the mass of the molecule. The energy of 1 kT/ Å, proportional to R_g, was given.

2. Base Pairing. This function achieves a good relative configuration between bases. It is based on the spatial distribution of A, C, G, and T bases around a nucleotide. A similar method was used in a study evaluating the direct interaction energy between DNA and binding proteins.[15] The definition of a coordinate system for calculating this spatial distribution is as follows. The origin of this coordinate system is the geometric center of the heavy atoms of the base. The X-axis is parallel to the N9−C4 (for purines) or N1−C2 (for pyrimidines) vectors and starts from the origin. The Y-axis is perpendicular to the X-axis and antiparallel to the vertical line from N7 (for purines) or C5 (for pyrimidines) to the X-axis. The Z-axis, in the right-hand reference frame, is the vector normal to the $X-Y$ plane. We considered an 18 Å × 18 Å × 6 Å rectangular box around the origin divided into grid cells with a grid interval of 2 Å. When the base origin of another base is within the grid cells, the energy, $-kT \ln P$ (P is the probability distribution of A, C, G, and T at that grid cell), is given. The probability distribution of each grid cell was calculated using the 557 nonredundant data set described in Section 2.1.

3. Base Stacking. This function optimizes the stacking configuration between two bases. The same coordinate system as the base pairing potential was considered. An energy of −1 kT is given to each base pair whose base origin is within the range of $(x^2 + y^2)^{1/2} < 4$ Å and 3 Å < $|z|$ < 6 Å. The idea behind this potential energy function can be found in the literature on FARNA.[2]

4. Steric Clash. This function penalizes steric clashes between several selected atoms on each nucleotide. Five sugar−phosphate backbone atoms (C1, C2, C3, C5, and P) and four base atoms (N6, N7, N9, and C2 for adenine; O6, N7, N9, and N2 for guanine; N4, O2, N1, and C5 for cytosine; O4, O2, N1, and C5 for uracil) were selected as representative atoms. The values of “steric radii” of these representative atoms were inferred from the third smallest distance between two representative atoms observed in the 557 nonredundant data set. If the distance between two representative atoms of candidate structure is less than their steric radii, an energy penalty of 1 kT is given. The idea behind this potential energy function can also be found in the literature on FARNA.[2]

5. Linkage of Backbone. This function rules out unnaturally stretched P−O3′ and P−O5′ bonds. Because fragment assembly is achieved by superimposing overlapped bases, these bonds, which connect the fragments, are sometimes too long or too short. To avoid these unusual P−O bonds, an energy penalty (E_{po}) is applied using the following function

$$E_{po} = [\max(0, |l - 1.60| - 0.20) - 1.0] \times 10.0 \times kT \qquad (2)$$

where l is length of P−O bonds, "1.60" is the equilibrated bond length of P−O3′ and P−O5′ derived from the AMBER parmbsc0 parameter set,[23] "0.20" is the tolerance and then E_{po} takes the minimum (-1.0×10.0) when l is within 1.60 ± 0.20, and "10.0" is a weight factor to match the order of the energy of this function with those of other functions.

2.3. Benchmarking.

2.3.1. Overview of Prediction. The flowchart for making a prediction is summarized in Figure 4.

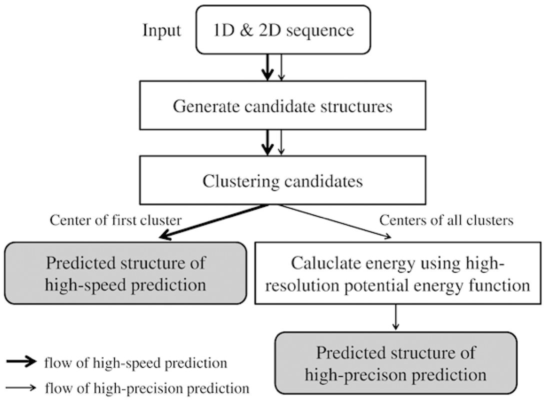

Input
1D & 2D sequence
Generate candidate structures
Clustering candidates
Center of first cluster
Centers of all clusters
Predicted structure of high-speed prediction
Caluclate energy using high-resolution potential energy function
Predicted structure of high-precison prediction
flow of high-speed prediction
flow of high-precision prediction

Figure 4. Flowchart of two protocols: high-speed prediction (thick arrows) and high-precision prediction (thin arrows).

We tested two protocols for this study. The first is "high-speed prediction using low-resolution energy functions and structure clustering", which is denoted with thick arrows in Figure 4 (we simply call this protocol "high-speed prediction"). In this protocol, many candidate structures are first generated on the basis of the input nucleotide sequence and secondary structure information. Next, the probability of each candidate structure is evaluated using the clustering method. Finally, the centers of the largest clusters are shown as "predicted structures" of this protocol. The details of each process are described in Sections 2.3.2 and 2.3.3, and the results are shown in Sections 3.1 and 3.2.

The second protocol is "high-precision prediction using high-resolution energy functions", which is denoted with thin arrows in Figure 4 (we simply call this protocol "high-precision prediction"). We conducted an additional and more time-consuming analysis for all the cluster centers derived in the second process of the first protocol using high-resolution potential energy functions to improve the results and to spotlight where to improve in RASSIE. The details of this process are described in Section 2.3.4, and the results are shown in Section 3.3. The difference between the two protocols was only how we found the best structure among many cluster centers.

2.3.2. Targets and Parameters for Sampling Candidate Structures. To evaluate the performance of RASSIE, six typical RNA structures were selected as targets (1KKA,[16] 1QWA,[17] 2F88,[18] 1CQ5,[19] 1L2X[20] and 1DK1[21]) (Table 1). These targets were selected as typical "simple stem-loops" (1KKA), "one-bulge loops" (1QWA), "2-way junctions" (2F88), "2 × 2-way junctions" (1CQ5), "pseudo-knots" (1L2X), and "3-way junctions" (1DK1). Among these structures, 1KKA, 1QWA, 2F88, and 1L2X had also been used as targets in a previous study.[2] These RNA structures were excluded from our data set when we derived the fragment library or the energy functions.

We used RASSIE on these six targets. To determine the number of insertion cycles of RASSIE, we tested several values and found that 500 cycles were sufficient. For each target, we derived 5000 candidate structures.

For comparison, we also performed tertiary structure prediction using FARNA of the ROSETTA 3.1 program,[2] one of the most well-known and effective programs for RNA tertiary structure prediction. FARNA can also use secondary structure information as input, with some constraints added according to the type of information. Therefore, we compared the results derived from RASSIE, FARNA without secondary structure information, and FARNA with secondary structure information (which we call "FARNA2D"). The built-in energy minimization function of FARNA was not activated. We then set the numbers of insertion cycles of FARNA and FARNA2D to 5000 to make the computational time for generating one candidate structure roughly equal to RASSIE for comparing performance of the prediction. A total of 5000 candidate structures were also derived for each target by using these

Table 1. Summary of Six Target Structures and Computational Times for Calculation[a]

					CPU time (sec/candidate)		
	method	length	base pairs	I-loops	RASSIE	FARNA	FARNA2D
1KKA	NMR	17	7	0	5.1	3.9	4.4
1QWA	NMR	21	9	1	3.1	4.6	5.5
2F88	NMR	34	13	2	4.3	7.5	9.1
1CQ5	NMR	43	15	4	9.2	9.4	11.6
1L2X	X-ray	28	(5)	0	3.7	6.1	6.8
1DK1	X-ray	57	21	5	64.4	12.2	15.7

[a] Method of structure determination (method), nucleotide length (length), number of base pairs (base pairs), number of internal-loops (I-loops), and CPU time for generating one candidate are summarized. Number of base pairs of 1L2X is in parentheses because it was not equal to that observed in native structure (see Section 3.1).

Table 2. Summary of Prediction Results with Values in bold Being Lowest rmsd for RASSIE and FARNA2D in Each Selecting Method

							selecting method					
	RASSIE			FARNA2D			center of 1st cluster		lowest rmsd		lowest AMBER	
	Th^{clst}[a]	M^{1st}[b]	N^{clst}[c]	Th^{clst}[a]	M^{1st}[b]	N^{clst}[c]	RASSIE (Å)	FARNA2D (Å)	RASSIE (Å)	FARNA2D (Å)	RASSIE (Å)	FARNA2D (Å)
1KKA	1.0	109	101	1.0	116	107	**5.10**	5.26	**3.98**	4.17	**4.72**	5.29
1QWA	1.5	123	111	1.5	121	41	**4.29**	5.25	**2.40**	4.16	**3.20**	6.35
2F88	2.0	106	152	2.5	87	22	**3.86**	4.97	**2.45**	3.44	**2.89**	3.56
1CQ5	2.5	70	121	8.0	109	222	7.41	**5.90**	5.95	**5.90**	**7.01**	11.22
1L2X	1.0	84	66	4.5	100	81	14.64	**13.50**	–	–	–	–
1DK1	5.0	100	116	9.5	114	203	18.32	**14.01**	–	–	–	–

[a] Th^{clst}: Threshold for clustering (Å). [b] M^{1st}: Number of cluster members of first cluster. [c] N^{clst}: Number of clusters.

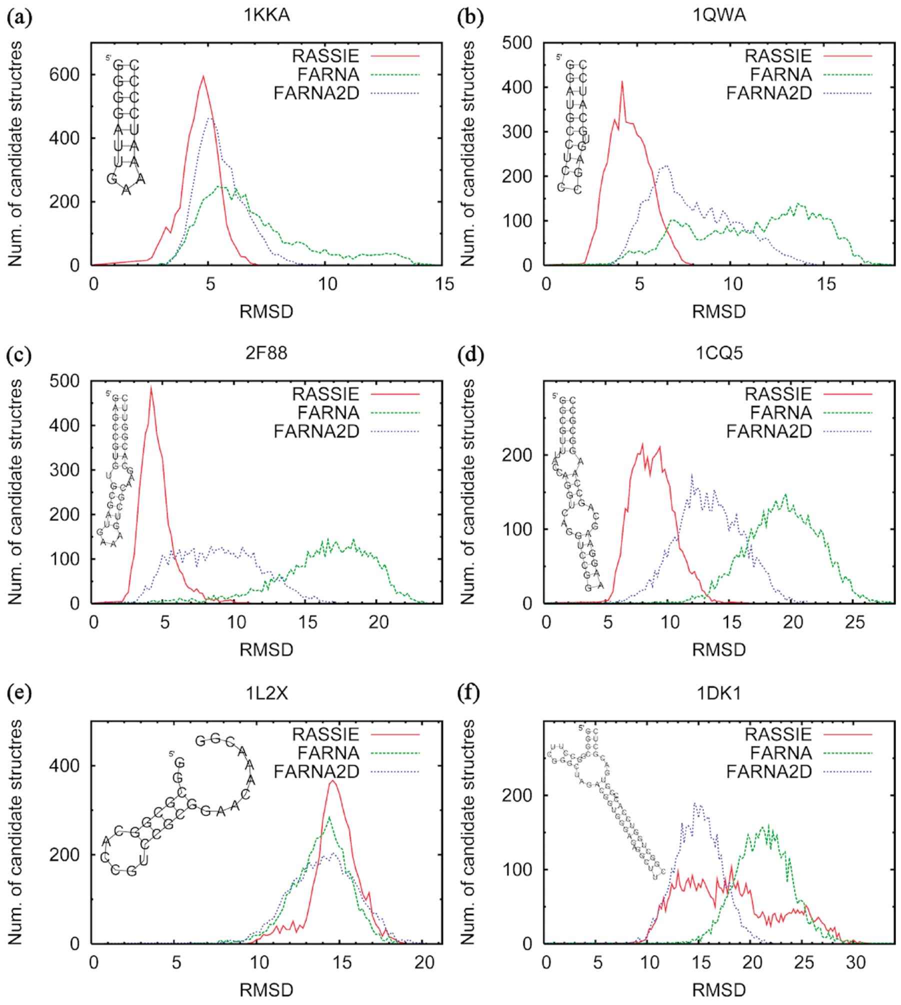


Figure 5. Distributions of rmsd (Å) from 5000 candidates to native structures for six targets: (a) 1KKA, (b) 1QWA, (c) 2F88, (d) 1CQ5, (e) 1L2X, and (f) 1DK1 (bin size = 0.2 Å). Nucleotide sequences and secondary structures used for input are also shown at top left of these plots.

methods. The details of structural feature and the computational time for generating one candidate of six targets are summarized in Table 1.

2.3.3. Clustering Candidate Structures. The 5000 candidate structures derived in Section 2.3.2 were clustered according to pairwise rmsd in a similar way to the method used in ROSETTA program.[22] We tested several values for the clustering threshold at intervals of 0.5 Å, and we adopted one in which the largest cluster contained about 100 members. Clusters containing fewer than five members were excluded from the analysis. The parameters for clustering are listed in Table 2. The center of the first cluster was used as the "predicted structure" of high-speed prediction in each method. Notice that the number of insertion cycles and the number of derived candidate structures of the two FARNA-based methods in this study was fewer than those in the original study.[2] We could not run the FARNA method as they had because of the limitations of our computer resources.

The rmsd used for clustering and structure evaluation was calculated using sugar−phosphate backbone atoms and several representative atoms in bases (P, C3′, C4′, C5′, O3′, O5′, O4′, C1′, C2′, N1, C4, and C5).

2.3.4. Evaluation Using High-Resolution Potential Energy Functions. To assess and improve the results, the potential energy of AMBER parmbsc0[23] using the GB/SA implicit solvent model[24−26] was calculated for each cluster center. The 500 steps of conjugate gradient energy minimization were performed using the sander program within AMBER11,[27] and the energy of the final step of this minimization was used as the AMBER energy. The initial step length was 0.01 Å, and the minimization halted when the rms of the Cartesian elements of the gradient went below 1.0×10^{-4} kcal/mol Å. If the AMBER energy of a cluster center was too high to allow energy minimization, the structure was excluded from the analysis.

2.3.5. Creation of Larger Data Set for Prediction. We performed tertiary structure prediction of other 217 RNA targets to evaluate the versatility of our method. To derive this set, we again used the RNA STRAND database with the searching options of "non-redundant sequences only", "length: less than or equal to 100″, "number of molecules in complex: less than or equal to 1", and "number of pseudoknots per molecule: less than or equal to 1". The targets that included modified bases or missing residues/atoms were manually excluded from the set, and we derived 217 target structures. This set included four 3-way junctions, six 4-way junctions, one 5-way junction, and 30 pseudoknot structures. The RNA-protein or RNA-peptide complex structures were also included. Structure sampling was performed using RASSIE in the same manner as the previous six targets. The structures at the center of the first cluster were derived for each target, and the rmsd values of the native structures were calculated.

3. RESULTS AND DISCUSSION

We conducted a tertiary structure prediction for six typical targets using RASSIE, FARNA, and FARNA2D and examined the advantages of using secondary structure information. Each secondary structure we used was a native secondary structure, not a predicted secondary structure, to avoid unnecessary discussion of the accuracy of secondary structure prediction. The nucleotide sequences and the secondary structures drawn using RNAplot of the Vienna RNA package[28] are shown at the top left of the plots in Figure 5.

3.1. rmsd Distribution of 5000 Candidate Structures. As described in Section 2.3.2, we derived 5000 candidate structures per method and per target. Figure 5 shows the distributions of rmsd from 5000 candidate structures to the native structures for the six targets to compare the sampling efficiency of the three methods. For the first four targets—1KKA, 1QWA, 2F88, and 1CQ5—the candidate structures derived from RASSIE showed a larger distribution in the lower rmsd area than those derived using FARNA and FARNA2D. Even when the same secondary structure information was used, RASSIE achieved a more efficient conformational search than FARNA2D. The most distinctive advantage of RASSIE came from our method of creating a fragment library.

The fragment structures used in FARNA are short and single stranded. The structural features of one fragment are represented by the set of their torsional angles. This simple implementation provided an effective and smart conformational search algorithm. However, some structural features are not reflected in the fragment structure, especially the base−base pairing conformations. The superior scoring functions used in FARNA can recapture base pairing conformations but may require many insertion cycles. The 5000 insertion cycles we used for FARNA were considered to be too short. In fact, in the original study of FARNA, 50000 insertion cycles were conducted to derive one candidate structure.[2]

On the other hand, the fragment structures used in RASSIE were secondary structure-based fragments. The fragment structures derived from hairpin loops or internal loops were also single stranded, but their structural features, such as torsional angles, bond length, and whole length of fragment, were directly implemented. The fragment structures derived from the helix region had the structural feature of a base-paired duplet. Any length of helix could be rebuilt by connecting the helix duplet fragments. Regions determined as a helix in the input file would never become nonhelix structures during the initial building and fragment insertion stages. RASSIE worked under the strong restraints imposed by the secondary structure information; not only by the scoring functions but also by the fragment structures. RASSIE thus performed a great deal of efficient sampling of the first four targets, which consisted mainly of helix stems.

For the last two targets, 1L2X and 1DK1, RASSIE did not reproduce near-native structures. These targets are hard problem for current version of RASSIE. The 1L2X target is a pseudoknot structure that includes a triplex structure. In RASSIE, triplex (or multiplex) structures were not given their own structural class. Therefore, we approximated this target as "stem loop with long single-stranded loop region at 3′ end" and modified the secondary structure of this target from "..(((((..[[[.))))).......]]]" to "..(((((......)))))..........". We expected that the constraint of the energy functions might possibly rebuild the original triplex structure, but this attempt did not go well. This result suggests the need for a secondary structure fragment class for multiplex structures or another conformational search algorithm for long single-stranded loop regions.

The 1DK1 target is a branched stem structure that consists of three branched stems connected by three single-stranded loop fragments. RASSIE recaptured each stem-loop structure but rarely found a suitable trio of single-stranded loop fragments that could join these stem-loops without any steric exclusion. A detailed description of the difficulty in predicting this target is shown in Figure S1 of the Supporting Information. To solve this problem, we could use two approaches. The first approach was lowering the threshold of a steric clash in the initial building stage and lowering the weight factor of "steric clash" and "linkage of backbone" energies in the fragment insertion stage. However, this approach resulted in only a mass of "broken" candidate structures with severe steric exclusion. The second approach was iterating the initial building stage, until the structure without any steric exclusion was derived. This approach requires much computational time but successfully provides the candidate structures without any steric exclusion. We selected the second approach and obtained the results shown in Figure 5 and Tables 1 and 2. We believe that a shortage of structural variations in the SG fragments make it difficult to perform structure sampling of such branched stem structures. The more the tertiary structure of RNA is known, the smaller the shortage by increasing known structures or conducting computational simulations.

3.2. Detailed Analysis of Predicted Structures of High-Speed Prediction. Figure 6 shows the predicted structures of high-speed prediction (the cluster center of the first cluster) with rmsd to the native structures for the six targets derived using RASSIE and FARNA2D. For the first three targets,

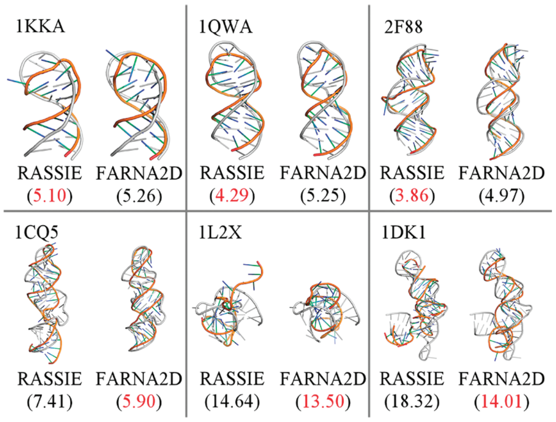


Figure 6. Predicted structures of high-speed prediction (cluster center of 1st cluster) for six targets derived using RASSIE (left) and FARNA2D (right). Native structures are shown superimposed on predicted structures (in gray). rmsd (Å) between predicted and native structures are also in parentheses. Values in red are lower rmsd of two methods.

RASSIE predicted closer-to-native structures than FARNA2D. It also solved such small and simple structures in several seconds per candidate structure (Table 1), and more than half of the candidates were near-native structures. FARNA2D also solved them as fast as RASSIE, but many candidate structures were non near-native (Figure 5 a−c). RASSIE worked well for the stem-loop structures that include one or two internal loops such as these three candidates.

As shown by the results for 1CQ5, however, RASSIE did not work well for the targets that include three or more internal loops. The predicted structure derived using FARNA2D had lower rmsd to the native structure, but the rmsd distribution of the candidate structures was much broader (Figure 5d). Therefore, we must set the clustering threshold at 8.0 Å to achieve a clustering size of about 100 for the first cluster (Table 2). We believe that RASSIE is superior in terms of the rmsd distribution of the candidate structures.

Figure 7 shows the closest-to-native structures, which we call "lowest rmsd", among the cluster centers for first four targets derived using RASSIE and FARNA2D. For the first three targets, RASSIE again derived closer-to-native structures than FARNA2D. For 1CQ5, the closest-to-native structure derived using RASSIE had low rmsd to the native structure as well as those derived using FARNA2D. These structures were more near-native but ranked lower in the clustering than those shown in Figure 6, which were the centers of the first cluster (and the predicted structures of high-speed prediction). The candidate structures were optimized for low-resolution energy function during the fragment insertion stage, and then the structures of the first cluster were the most probable structures for low-resolution energy function, which was normally a little different from the true energy function. This difference in energy function might appear in the difference between "the centers of the first cluster" and "lowest rmsd" structures. We should discuss about the validation and the improvement of the energy function.

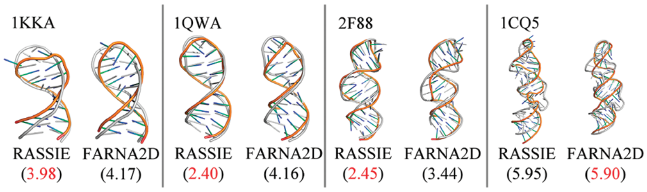


Figure 7. Closest-to-native structures among cluster centers for six targets derived using RASSIE (left) and FARNA2D (right). Native structures are shown superimposed on candidate structures (in gray). rmsd (Å) between closest-to-native structures and native structures are also in parentheses. Values in red are lower rmsd of two methods.

Figure 8 is a scatter plot of the energies of the RASSIE energy function and that of FARNA2D against the rmsd from the 5000 candidate structures to the native structures of these four targets. The correlation coefficients between the RASSIE energy and the rmsd were −0.01, 0.16, 0.20, and −0.05. These low correlations suggest room for improvement. The correlation coefficients between the FARNA2D energy and rmsd were higher; however, we found many candidate structures with low FARNA2D energy but high rmsd. The energy functions used in both methods are called "low-resolution" energy functions, which evaluate several structural features such as radius-of-gyration, base pairing, and steric clash. We believe that such low-resolution energy functions might still not be enough for the structural evaluation of RNA. Therefore, we need to perform the evaluation using high-resolution energy functions that evaluate all atom−atom interactions of the candidate structures.

3.3. Prediction Using High-Resolution Potential Energy Function.

To obtain information on improving RASSIE energy functions, we conducted an evaluation using a high-resolution potential energy function, which is described as "high-precision prediction" in Figure 4. We used the AMBER parmbsc0 force field and the GB/SA implicit solvent model to conduct this analysis. For 1KKA, 1QWA, 2F88, and 1CQ5, the AMBER energy of the cluster centers derived from RASSIE and FARNA2D were calculated and plotted against rmsd (Figure 9). For these four targets, the rmsd values of the candidate structure with the lowest AMBER energy (red triangle) were better than those selected using the clustering method (green cross). When we selected the predicted structure according to the AMBER energy, we found closer-to-native structures. Therefore, the cluster centers with the lowest AMBER energy were adopted as "predicted structures of high-precision prediction", and the structures and rmsd values of the native structures are shown in Figure 10. These results suggest the value of evaluating using a high-resolution potential energy function. We also found several non-native structures that had low energies comparable to native structures, indicating the need for improving the current high-resolution potential energy function.

Interestingly, many of the candidate structures derived using FARNA2D had too high of an AMBER energy to run the energy calculation. Some of the candidate structures derived using RASSIE also had high AMBER energies, but they were much fewer than when using FARNA2D. Therefore, RASSIE requires less computational time and can generate many near-native candidate structures. This is another advantage of RASSIE.

3.4. Evaluation of Versatility Using Larger Data Set.

We performed tertiary structure prediction of 217 RNA targets to evaluate the versatility of our method. Structure sampling was performed using RASSIE in the same manner as the previous six targets. Note that all the pseudoknot targets were approximated as "stem loop with long single-stranded loop region at 3′ end". The structures of the center of the first cluster

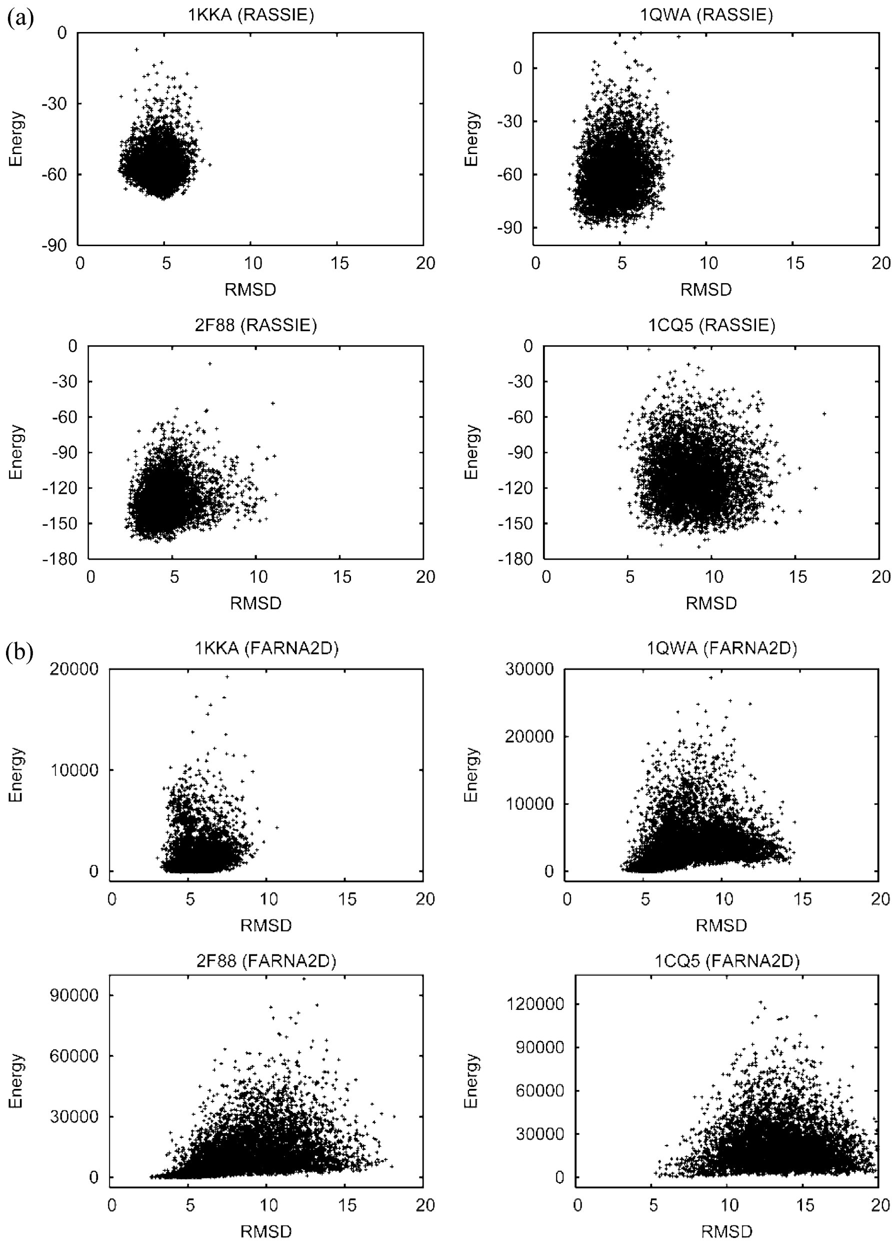


Figure 8. Scatter plot of energy (kcal/mol) for (a) RASSIE and (b) FARNA2D against rmsd (Å) from 5000 candidate structures to the native structures.

were derived for each target and the rmsd values of the native structures were calculated. Table S1 of the Supporting Information and Figure 11 show the results of this prediction. For the simple stem-loops, bulge-loops, and 2-way junctions, RASSIE successfully predicted low rmsd structures.

For more complicated targets (3-way junctions, 4-way junctions, 5-way junctions, and pseudoknots) RASSIE could not predict structures. The prediction of multijunctions often resulted in “no candidate structure” because the conformational search process shown in Figure S1 of the Supporting Information could not find a good set of the pieces of structural puzzle. Moreover, the prediction of the targets that include long hairpin loops or long SG loops also failed due to shortage of such long hairpin/SG loops in the fragment library.

From these results, we concluded that the stem-loop, bulge-loop, and 2-way junctions that include less than 6−7 loops (note that 163 of the 217 targets meet this requirement) can be accurately predicted using RASSIE. As is the case with 1L2X and 1DK1, the remaining 54 targets, which included long loop regions and multibranched or multiplexed structures, were difficult to predict for current version of RASSIE. In the future, we will improve the method by adding the protocol to model and store a variety of long hairpin/SG loop structures in the fragment library.

4. CONCLUSION

We investigated two strategies for achieving higher performance of the tertiary structure prediction of RNA. The first strategy was the implementation of base-paired information into the fragment structure library. This strategy was successfully implemented in RASSIE and decreased the computational time and attained better candidate structures than previous methods. These advantages were especially evident when reproducing simple base-paired stem loops. To

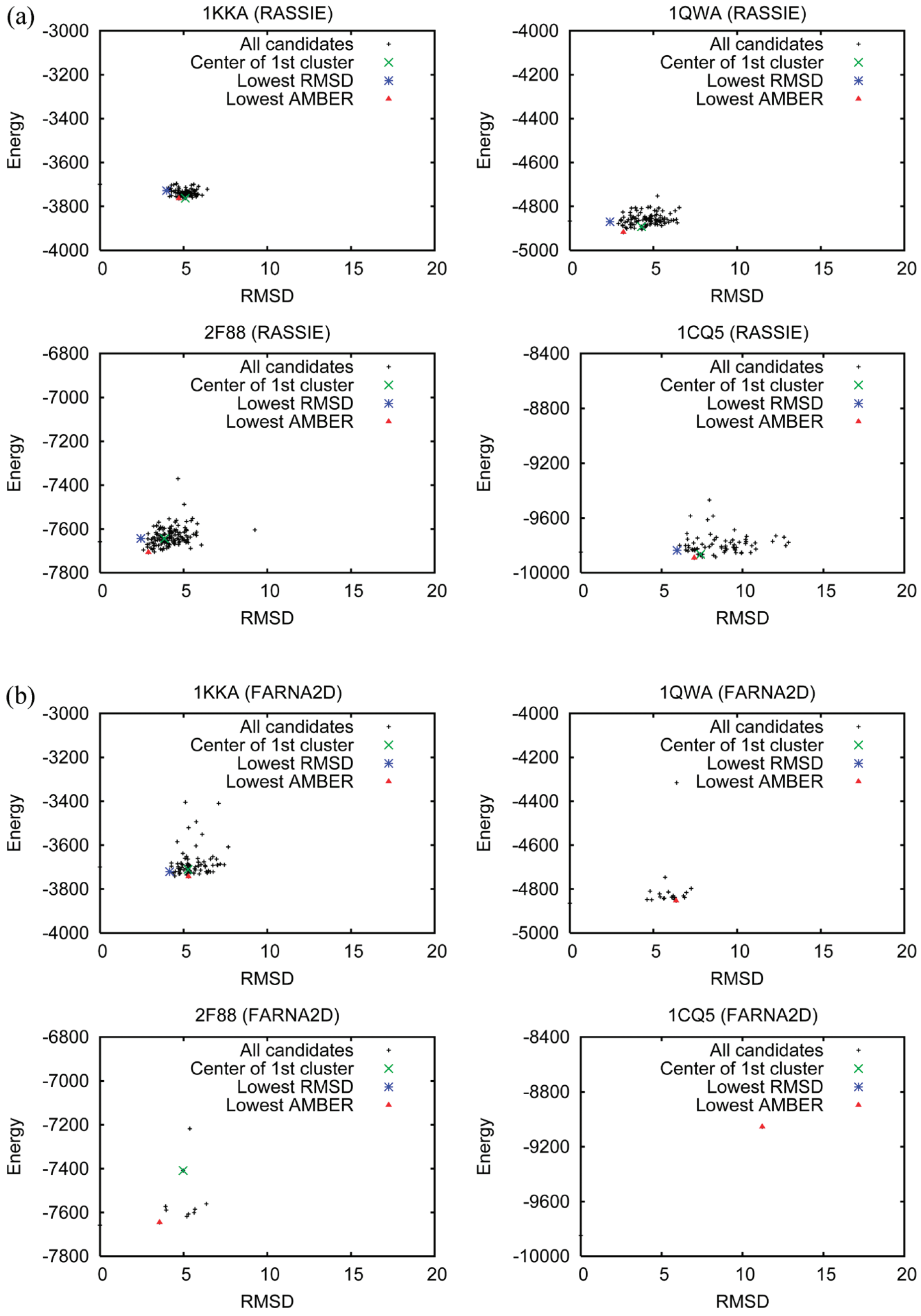


Figure 9. Scatter plots of AMBER energy (kcal/mol) of cluster centers of 5000 candidate structures derived using (a) RASSIE and (b) FARNA2D against rmsd (Å) to native structures. Center of first cluster (= predicted structure of high-speed prediction), structure with lowest rmsd, and structure with lowest AMBER energy (= predicted structure of high-precision prediction) are plotted with green crosses, blue asterisks, and red triangles, respectively. Note that some cluster centers do not appear in this figure because they had extremely larger AMBER energy than native structures, which are plotted on the *Y*-axis.

succeed with more complicated structures, a first step should be to improve the range of the secondary structure-based fragment library to cover several complicated structural segments such as long loops, triplexes, or branched stems.

Our second strategy was a structural evaluation using high-resolution potential energy functions. We in fact often found closer-to-native structures when using the high-resolution potential energy function. If we could set aside more computational time, this procedure would be helpful for prediction. As RASSIE can decrease the computational time to derive candidate structures, we can set aside much more computational time to this evaluation.

Moreover, using the native secondary structure, we demonstrated that using a secondary structure-based fragment library can contribute to improved accuracy of tertiary structure prediction. Secondary structures can now be efficiently determined with biochemical experiments or secondary structure prediction tools. We believe the assembly of secondary structure-based fragments and the evaluation of candidate structures using a high-resolution potential energy function, as demonstrated in this paper, represents a major step forward in tertiary structure prediction studies based on secondary structure information.

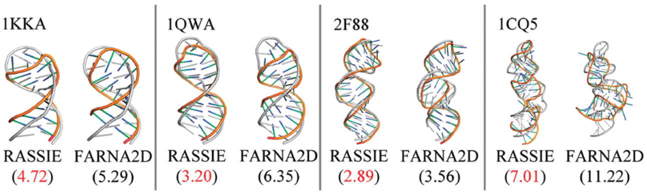


Figure 10. Predicted structures of high-precision prediction (cluster centers with lowest AMBER energy) for four targets derived using RASSIE. Native structures are shown superimposed on predicted structures (in gray). rmsd (Å) between predicted and native structures are also in parentheses. Values in red are lower rmsd of two methods.

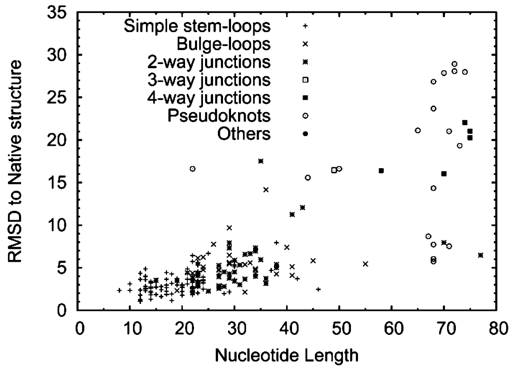


Figure 11. Scatter plot of rmsd values of native structures of first cluster centers of each target against nucleotide length.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Figure S1 shows the detailed description of difficulty in predicting branched structures. Table S1 shows the results of predicting 217 targets discussed in Section 3.4. The rmsd values from the predicted structure to the native structure, nucleotide length, structural type, and length of longest hairpin/SG loop of target are summarized in this table. This information is available free of charge via the Internet at http://pubs.acs.org

■ AUTHOR INFORMATION

Corresponding Author

*E-mail. s.yamasaki@aist.go.jp.

Funding

This study was supported by a Grant-in-Aid for Young Scientists (B) (22700317).

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Shapiro, B. A.; Yingling, Y. G.; Kasprzak, W.; Bindewald, E. Bridging the gap in RNA structure prediction. *Curr. Opin. Struct. Biol.* **2007**, *17*, 157−165.

(2) Das, R.; Baker, D. Automated de novo prediction of native-like RNA tertiary structures. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 14664−14669.

(3) Das, R.; Karanicolas, J.; Baker, D. Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat. Methods* **2010**, *7*, 291−294.

(4) Parisien, M.; Major, F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* **2008**, *452*, 51−55.

(5) Frellsen, J.; Moltke, I.; Thiim, M.; Mardia, K. V.; Ferkinghoff-Borg, J.; Hamelryck, T. A probabilistic model of RNA conformational space. *PLoS Comput. Biol.* **2009**, *5*, e1000406.

(6) Sharma, S.; Ding, F.; Dokholyan, N. V. iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics* **2008**, *24*, 1951−1952.

(7) Hamada, M.; Kiryu, H.; Sato, K.; Mituyama, T.; Asai, K. Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics* **2009**, *25*, 465−473.

(8) Zuker, M.; Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* **1981**, *9*, 133−148.

(9) Do, C. B.; Woods, D. A.; Batzoglou, S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* **2006**, *22*, E90−E98.

(10) Andronescu, M.; Bereg, V.; Hoos, H. H.; Condon, A. RNA STRAND: The RNA secondary structure and statistical analysis database. *BMC Bioinf.* **2008**, *9*, 340.

(11) Klein, D. J.; Schmeing, T. M.; Moore, P. B.; Steitz, T. A. The kink-turn: A new RNA secondary structure motif. *Embo J.* **2001**, *20*, 4214−4221.

(12) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(13) Leontis, N. B.; Westhof, E. Geometric nomenclature and classification of RNA base pairs. *RNA* **2001**, *7*, 499−512.

(14) Yang, H. W.; Jossinet, F.; Leontis, N.; Chen, L.; Westbrook, J.; Berman, H.; Westhof, E. Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res.* **2003**, *31*, 3450−3460.

(15) Kono, H.; Sarai, A. Structure-based prediction of DNA target sites by regulatory proteins. *Proteins* **1999**, *35*, 114−131.

(16) Cabello-Villegas, J.; Winkler, M. E.; Nikonowicz, E. P. Solution conformations of unmodified and A(37)N(6)-dimethylallyl modified anticodon stem-loops of *Escherichia coli* tRNA(Phe). *J. Mol. Biol.* **2002**, *319*, 1015−1034.

(17) Finger, L. D.; Trantirek, L.; Johansson, C.; Feigon, J. Solution structures of stem-loop RNAs that bind to the two N-terminal RNA-binding domains of nucleolin. *Nucleic Acids Res.* **2003**, *31*, 6461−6472.

(18) Seetharaman, M.; Eldho, N. V.; Padgett, R. A.; Dayie, K. T. Structure of a self-splicing group II intron catalytic effector domain 5: Parallels with spliceosomal U6 RNA. *RNA* **2006**, *12*, 235−247.

(19) Schmitz, U.; Behrens, S.; Freymann, D. M.; Keenan, R. J.; Lukavsky, P.; Walter, P.; James, T. L. Structure of the phylogenetically most conserved domain of SRP RNA. *RNA* **1999**, *5*, 1419−1429.

(20) Egli, M.; Minasov, G.; Su, L.; Rich, A. Metal ions and flexibility in a viral RNA pseudoknot at atomic resolution. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 4302−4307.

(21) Nikulin, A.; Serganov, A.; Ennifar, E.; Tishchenko, S.; Nevskaya, N.; Shepard, W.; Portier, C.; Garber, M.; Ehresmann, B.; Ehresmann, C.; Nikonov, S.; Dumas, P. Crystal structure of the S15-rRNA complex. *Nat. Struct. Biol.* **2000**, *7*, 273−277.

(22) Bonneau, R.; Tsai, J.; Ruczinski, I.; Chivian, D.; Rohl, C.; Strauss, C. E. M.; Baker, D. Rosetta in CASP4: Progress in ab initio protein structure prediction. *Proteins* **2001**, *Suppl 5*, 119−126.

(23) Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham, T. E.; Laughton, C. A.; Orozco, M. Refinenement of the AMBER force field for nucleic acids: Improving the description of alpha/gamma conformers. *Biophys. J.* **2007**, *92*, 3817−3829.

(24) Tsui, V.; Case, D. A. Theory and applications of the generalized Born solvation model in macromolecular simulations. *Biopolymers* **2001**, *56*, 275−291.

(25) Schaefer, M.; Van Vlijmen, H. W. T.; Karplus, M. Electrostatic contributions to molecular free energies in solution. *Adv. Protein Chem.* **1998**, *51*, 1−57.

(26) Weiser, J.; Shenkin, P. S.; Still, W. C. Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO). *J. Comput. Chem.* **1999**, *20*, 217−230.

(27) Case, D. A.; Darden, T. A.; T.E. Cheatham, I.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S.

R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. *AMBER11*; University of California, San Francisco, 2010.

(28) Hofacker, I. L.; Fontana, W.; Stadler, P. F.; Bonhoeffer, L. S.; Tacker, M.; Schuster, P. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* **1994**, *125*, 167−188.