Supporting Information Appendix

Namhee Kim [*], **Christian Laing** [†], **Shereef Elmetwaly** [*], **Segun Jung** [*], **Jeremy Curuksu** [*], **and Tamar Schlick** [*] [‡] [§]

[*]Department of Chemistry, New York University, 100 Washington Square East, New York, NY 10003,[†]Department of Biology, Mathematics and Computer Science, Wilkes University, 84 West South Street, Wilkes-Barre, PA 18766,[‡]Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012, and [§]To whom correspondence should be addressed: schlick@nyu.edu

Submitted to Proceedings of the National Academy of Sciences of the United States of America

Methods

Overview. Our computational approaches involve two stages: estimation of knowledge-based statistical potentials and Monte Carlo/ Simulated Annealing (MC/SA) sampling of 3D graphs.

In the first stage, we develop knowledge-based statistical potentials based on 3D graph representations of non-redundant solved RNAs and statistical analysis of their 3D geometrical features, from parts to whole, including sizes of helices/hairpins/bulges/junctions, bending and torsion angles between two helices of internal loops, and radii of gyration of the entire RNAs. Three steps are involved in this process of potential development:

(i) Non-redundant sets of solved RNAs are translated to 3D graphs and linked to their 2D structures. See below for detailed translation rules of 3D graphs and RNA dataset.

(ii) To determine overall helical arrangements, we measure details of local and global geometries and correlate these 3D geometries to 2D structural information. The local geometrical descriptors for RNA include sizes for each building block (helices, hairpins, internal loops, and junctions) and local inter-helical angles (bending and torsion angles). The global geometrical measure of the radius of gyration describes the overall compactness in 3D. We quantify these measures using 3D graphs and relate them to the available 2D information. See below for a coordinate system for 3D graphs, mathematical formulas for measures, and the resulting statistics.

(iii) Based on resulting statistics of 3D geometries linked to 2D information, knowledge-based statistical potentials of bending, torsion angles and radii of gyration are calculated and extrapolated by polynomial expansion to handle unrepresented regions of experimental data. See below for the detailed refinement procedures and resulting statistical potentials.

In the second stage, to build native-like structures from 2D structures guided by preferred conformations, we employ hierarchical MC/SA sampling approaches where the objective junction is the combination of knowledge-based statistical potentials computed from the first stage. The MC/SA consists of three steps: (i) set-up of initial graphs given a 2D structure by assignment of weighted edges and vertices to the different families of 2D structures (helices, hairpins, internal loops, junctions) and using size measures and junction prediction; (ii) MC/SA sampling of RNA 3D graphs based on two types of moves (restricted and random pivot moves) guided by the knowledge-based potentials; (iii) analysis of resulting sampled graphs using RMSD and by clustering analysis. Detailed procedures are described below.

RNA dataset for statistical analysis of RNA geometries. We collect 781 non-redundant high-resolution PDB structures for 3D geometry analysis. We generate 2D structures corresponding to these PDB structures using results from three 2D extraction tools – RNAView [1], MC-Annotate [2], and FR3D [3]. From these RNAs, we obtain 1181 hairpins (with 0-20 nucleotides [nt] in loop regions), 2118 internal loops (0-20 nt in each side of bulges), and 244 junctions (3- and 4-way junctions).

RNA 3D graph representation. The original RAG representation defines planar tree graphs where unpaired regions are vertices and helices are edges [4,5]. To represent RNA helical arrangements given a 2D structure before MC/SA, we refine our RAG tree graph by defining additional vertices at helix ends and edges connecting loop vertices to proximal helix-end vertices. We also add a vertex for an internal loop with at least one nt: in RAG, internal loops with at least two nucleotides are represented as vertices. Regarding helices, like RAG, we represent each helix having at least two consecutive canonical Watson-Crick or wobble GU base pairs by two vertices which are connected by one edge: if there is only one base pair, it is considered as two single strands with one nt for each side. Two vertices represent two terminal base pairs denoting a helix (Figure S1A).

We further extend these refined RAG planar tree graph to nonplanar graph in 3D space for MC moves by defining a coordinate system for vertices representing helix ends or center of hairpin, internal loop, or junction (Figures 2A and S1B). To make comparable 3D models with solved structures, the Cartesian coordinates for each vertex are defined by the origin of each terminal base pair. We utilize an algorithm to calculate the origin of a base pair proposed by Schlick [6], where the origin is defined as a translation (4 Å) of the projection from the midpoint of C8 atom for purine and C6 atom for pyrimidine to the line connecting C1$'$ atoms for both purine and pyrimidine bases. The edge connecting these two vertices forms the global axis of an A-form RNA helix (Figure 2A). The Cartesian centroid of an internal loop is defined as an average of coordinates of two adjacent vertices representing two proximal helix ends. Extending this definition, the centroid of an N-way junction is an average of coordinates of N adjacent vertices for N helix ends (see Figure 2A for illustration for 3-way and 4-way junctions). Since hairpins has only one helix connected to them, we define the centroid of a hairpin end as an average of C1$'$ atoms of all unpaired bases of a hairpin loop. We connect each vertex representing the centroid to adjacent vertices representing the proximal helix ends. For example, an N-way junction has N edges connecting the junction centroid to N vertices representing helix ends. Similarly, the number of edges connected to the centroid vertices of hairpin, dangling ends, and internal loop are 1, 1, and 2, respectively. Figure S1B shows three examples of solved structures and their 3D graph representations.

Size measures of RNA structural elements. To evaluate RNA 3D geometrical features correlated with 2D information, we quantify the sizes of helices, hairpins, internal loops, and junctions, using solved RNAs, and estimate them using the number of bases or base pairs of each 2D element.

We formulate the helix length (S_{helix}) as the distance connecting the two base pair origins at the helix ends. According to Ref. [6], the axial rise parameter – the vertical distance along the double helix axis between adjacent base pairs – is determined as 2.87 Å for an A-form helix. Thus, we set S_{helix}(Å) to 2.87*(n-1), where n is the number

Reserved for Publication Footnotes

of base pairs. Correspondingly, the length of edge is scaled by the number of base pairs.

We determine the hairpin size (S_{hairpin}) as the distance between the hairpin centroid and the origin of the last base pair of the hairpin. From our analysis of 1181 hairpins with 1 to 20 nt of the non-redundant dataset of solved RNA structures, we identify a strong linear relation between the sequence length of hairpin, N_{hairpin} (nt), and hairpin size: S_{hairpin} (Å) = 0.66*N_{hairpin}+4.23, where the coeffecent of determination $R^2 = 1 - \sum(\text{actual value} - \text{estimated value})^2 / \sum(\text{actual value} - \text{mean (actual value)})^2 = 0.7$. As R^2 is close to 1, the linear regression fits the actual data better [7]. Thus, the total length of an edge for hairpins is scaled by the number of bases in the hairpin.

Similarly, the size of an internal loop (S_{internal}) is defined as the distance between the centroid of the internal loop and the origin of one of the two base pairs adjacent to the centroid (distance between v1 and C in Figure 2A). For the 2188 internal loops analyzed, we find a linear relationship for the internal loop size: S_{internal}(Å) = 2.44+1.16*L+0.21*R (R^2 = 0.8), where L and R are the nucleotide numbers in the two portions of the loop where L≤R.

For junctions, we calculate the distances between coaxial helices (s0) in all three-way and four-way junctions (Figure S2). We find a linear relationship for the coaxial helical distance s0 (Å) = 2.75*N_{coaxial}+ 3.91 (R^2 = 0.84), where N_{coaxial} is the number of nucleotides of single strands between coaxial helices (e.g., residues colored green in Fig. S2). To locate non-coaxial helices, we calculate additional distances between perpendicular (noted as s1), diagonal (s2), and parallel (s3) helices, which are averaged as 19.95, 21.17, and 20.48 Å, respectively. We also calculate distance between non-connected perpendicular helices in the cL 4-way junction family, which is averaged as 19.95 Å(see s4 in Fig. S2). With the application of RNAJAG [8], this statistical analysis allows us to set up an initial planar tree graph (a revised RAG graph which we embed in 3D) of junctions.

Bending and torsion angles of inter-helices. To determine the orientations between two helices of internal loops, we formulate bending and torsion angles of inter-helices (Fig. 2B). The bending angle is defined as the angle between two consecutive helices (vectors $\nu 1$ and $\nu 3$ in Fig. 2B) connected by two single stranded regions, L and R, where L≤R, of an internal loop:

$$\text{Bending angle } \theta = cos^{-1}(|\nu 1 \cdot \nu 3|/|\nu 1||\nu 3|). \quad [1]$$

The torsion angle is defined by the dihedral angle between two consecutive helices (vectors $\nu 1$ and $\nu 3$ in Fig. 2B) connected by two single strand regions, L and R, where L≤R, of an internal loop along $\nu 2$ which is the bulge loop connecting two base pair origins of two helix ends proximal to a bulge (vector $\nu 2$ in Fig. 2B):

$$\text{Torsion angle } \tau = sign(n1 \cdot \nu 2)(cos^{-1}(|n1 \cdot n3|/|n1||n3|), \quad [2]$$

where $sign(x)$ is 1 if x is positive and -1 if x is negative, and $n1$ and $n3$ are the normal vectors of planes spanned by $\nu 1 \nu 2$ and $\nu 2 \nu 3$, respectively. The torsion angles determine the angular orientations of the bend angles, and thus, the two variables of bending and torsion angles describe local arrangements of two helices adjacent to an internal loop in 3D space.

We calculate the bending and torsion angles of 2188 internal loops obtained from the 781 non-redundant RNA set that we collect. To link these angles to 2D information, all internal loops are classified by L and R, where L≤R. Our analysis reveals a strong correlation between these bend and torsion angles and L/R. We observe a positive correlation of the bend angle with the loop size while the correlation of torsion angles with the loop size is negative. Figure S3 shows typical distributions (L/R = 0/1, 0/2, 0/3+, and 1/1 are taken as examples for illustrative purpose) of bending and torsion angles for symmetric groups and asymmetric groups. As shown in the case of L/R = 1/1, the bending angles for symmetric internal loops show narrow distributions, with a strong preference for small bending angles ($\sim 30°$) and large torsion angles ($180° \sim 240°[= -120°]$). For asymmetric cases (L/R = 0/1, 0/2, and 0/3+), we observe a high degree of flexibility in strongly asymmetric internal loops. For L/R = 0/1, the bending and torsion angles are centered around 30° and 160°, respectively, while those angles for L/R = 0/2 are 45° and 150°, respectively. When longer single strands exist on one side (e.g., L/R = 0/3+), the bending angles tend to increase while the torsion angles tend to decrease. As shown in the examples in Figure S3, two helices connected to one internal loop have similar helical shape with A-form RNA with bending angle to 0° and torsion angle close to 180°.

Since there are not many internal loops with L>6, all loops with at least six nt on either side are grouped together, resulting in the 27 following groups by L/R with L≤R as: 0/1, 0/2, 0/3, 0/4, 0/5, 0/6+, 1/1, 1/2, 1/3, 1/4, 1/5, 1/6+, 2/2, 2/3, 2/4, 2/5, 2/6+, 3/3, 3/4, 3/5, 3/6+, 4/4, 4/5, 4/6+, 5/5, 5/6+, and 6+/6+, where 6+ means greater than or equal to six. See Figure S4 for full histogram of the bending and torsion angles for each group of internal loops. Incidentally, the corresponding knowledge-based statistical potentials for bending and torsion preferences for internal loops are classified into 27 groups (see Figures S4 and S5).

Radii of gyration of 3D graphs. To capture the overall compactness of RNA in addition to local bending and torsion geometries, we formulate radius of gyration (R_g) measure. We define R_g of a graph as the root mean square distance of vertices representing all loop centers and helix ends (Fig. 2B):

$$R_g = \frac{\sum_{i=1}^{V} |V_i - \bar{V}|}{V}, \quad [3]$$

where V_i is the coordinates of vertex from $i = 1$ to V, and $\bar{V}$ is the average of all vertices of 3D graphs. We analyze the radii of gyration (R_g) of 781 solved structures in our RNA dataset along the sequence length (L) and the vertex number (V). The sequence length (L) and vertex number (V) range from 14 to 2633 nt and from 4 to 416 vertices, respectively. Among the 781 RNAs, 774 have less than 400 nt and 80 vertices and $R_g < 40$ Å. The other 7 represent 16S rRNAs and 23S rRNAs with around 1500 nt and 3000 nt, respectively, and R_g near 65 Å. Using correlation analysis, we find that R_g follows a logarithmic relationship with sequence length (L) and the vertex number (V):

$$R_g(L, V) = a * ln(L) + b * ln(V) + c, \quad [4]$$

where a, b, and c are fitted parameters. For our dataset, a= 8.58, b = 2.30, and $c = -19.50$. This formula shows that R_g increases logarithmically with the sequence length, but for a given sequence length, decreases logarithmically with the vertex number. Since the vertex number increases with the degree of branching of a graph, R_g decreases with increasing branches. Our corresponding scoring function uses the relationship above to reproduce the overall compactness of the RNA.

Knowledge-based potentials for 3D graphs. Our combined statistical potential is

$$\Delta G = \Delta G_{\text{internal}} + \Delta G_{R_g}, \quad [5]$$

$$\Delta G_{\text{internal}} = \sum_{i=1}^{\text{all internal loops}} (\Delta G(\theta_i) + \Delta G(\tau_i)), \quad [6]$$

$$\Delta G_{R_g} = |R - \bar{R}|, \quad [7]$$

where θ_i, τ_i, R and $\bar{R}$ are the bending and torsion angles of all i internal loops in a given RNA 2D structure, the radius of gyration of an RNA conformation, and the preferable radius of gyration given the length of RNA, respectively.

Specifically, for $\Delta G_{\text{internal}}$, we categorize internal loop families by the lengths of single strands L/R, L$\leq$R, where L and R are the nucleotide length of single strands of internal loops. Note that we have the 27 groups by L/R with L$\leq$R as: 0/1, 0/2, 0/3, 0/4, 0/5, 0/6+, 1/1, 1/2, 1/3, 1/4, 1/5, 1/6+, 2/2, 2/3, 2/4, 2/5, 2/6+, 3/3, 3/4, 3/5, 3/6+, 4/4, 4/5, 4/6+, 5/5, 5/6+, and 6+/6+, where 6+ means greater than or equal to six (see Figures S4 and S5). For each L/R category, we develop the bending and torsion angle statistical potentials in three steps. First, we partition the angles (θ) every 45° (for bending angles which are $0^\circ \leq \theta < 180^\circ$, partition number $|P| = 4$ with partitions P_i, i=1,2,3, and 4 while for torsion angles which are $-180^\circ \leq \tau < 180^\circ$, partition number $|P| = 8$ with partitions P_i, i=1,2,. . . ,8). The probability for a given angle (θ or τ in P_i) is then:

$$Pr(\theta) = N_i/N, \qquad \textbf{[8]}$$

where N_i is the number of internal loops that have an angle in P_i among all N internal loops. The probability that such an angle would form randomly is: $P_{\text{random}} = 1/|P| = 1/4$ for bending angles or $1/8$ for torsion angles. We apply Boltzmann statistics to the internal loop angles, so that the free energy for the angle becomes:

$$\Delta G(\theta) = -k_b T ln(Pr(\theta)/P_{\text{random}}), \qquad \textbf{[9]}$$

where k_b is the Boltzmann factor (8.31J/K) and T is the temperature (300K). Smoothing functions are needed since data have gaps missing values due to limited experimental data (see Fig. S5).

Set-up of 2D structure to initial graph. To form an initial 3D graph, we use a 2D structure in a BPSEQ format as an input. The information in the BPSEQ file is parsed to capture all the aspects of paired and unpaired bases of a given 2D structure (Fig. 1). Then, we determine topology and Cartesian coordinates of an initial graph in two steps. First, we label all loops and involved helices following the order of $5'$ to $3'$ end, hairpins, internal loops, and junctions, and determine a tree graph topology from the 2D structure by translating helices and loops defined as edges and vertices based on our graph definition. Second, we determine the geometry of an initial graphs by fixing the Cartesian coordinates of each vertex after adding the scaled edge lengths by size measures in one direction. If junctions are present, we determine the coordinates of junction vertices (one for the junction loop center and 2N for N-way helices) using the RNAJAG program [8].

MC/SA sampling: pivot moves, steric clash removal, and score minimization. We use two types of pivot moves of a graph based on range of angle degrees: (1) reciprocally decreasingly restricted angle ranges along MC steps ($\sim$ 1/Step) from 360° to 10° (restricted moves) and (2) 360° (random moves). With random degrees within given range of angles, all vertices linked to a randomly selected internal loop by a randomly-selected helix which is rotated along randomly-selected one of three axis (x, y, z axis). For each move, we identify a steric clash to eliminate conformations before scoring. A steric clash in an RNA graph is defined when a minimum distance between any two edges of an RNA 3D graph is less than 1 Å. We score each graph conformation without steric clashes by our statistical potentials (Eq [5]). For each step j, graph moves are selected at random within given range of angles to transform the old conformation g_{j-1} into a new conformation g_j. A score $E(g_j)$ is assigned to each conformation g_j and used to determine acceptance or rejection. Specifically, if the score $E(g_j)$ for a new conformation g_j is lower than that of the old conformation (i.e., $E(g_{j-1}) \geq E(g_j)$), the new graph conformation g_j is accepted. If the score $E(g_j)$ is higher, the simulated annealing sampling proceeds: the move is accepted with probability $P(j) = 2^{E_j/T_j}$, where $E_j = E(g_j) - E(g_{j-1})$ and the decreasing system temperature $T_j = c/log_2(1 + j/s)$ where s is the total MC step and $c = 1/4log_2(10)$ (for restricted moves) or $c = 1/log_2(10)$ (for random moves). The rejected probability is $1 - P(j)$. We run our program implemented in C++ on the local Mac computer (2x2.26 GHz Quad-Core Intel Xeon processor with 8GB memory) and the computational time is less than 20 minutes for 10^4 MC steps.

Comparison between graphs by RMSD. To evaluate our predicted graph structure, we compare our sampled graphs to the reference graph translated from the solved structures or predicted by other computational tools. We superimpose the two graphs by translating both graphs into the origin and rotate one graph using singular value decomposition [9]. To compare the difference of global helical arrangements, we use graph-based root mean square deviation (RMSD). The graph RMSD measures the average distance of vertices between two superimposed 3D graphs:

$$\text{RMSD} = \frac{\sum_{i=1}^{V} |V_i - W_i|}{V}, \qquad \textbf{[10]}$$

where V_i and W_i ($1 \leq i \leq V$) are vertices in our graph and the reference graph after they are aligned, respectively, and V is the total number of vertices.

Clustering of sampled RNA graphs. To assign representative graphs of clusters with similar geometries, we further cluster sampled conformations by MC scores and graph RMSD from reference graph (e.g., graph translated from solved structure or lowest-scored graph). We use the k-means algorithm, which partitions the sampled graphs into k groups so that the sum of squares between the assigned cluster centers and each point is minimized [10]. To validate the clustering, we calculate the silhouette width SC, for each sampled graph g_j, $SC_{g_j} = (b_{g_j} - a_{g_j})/max(a_{g_j}, b_{g_j})$, where a_{g_j} is the average distance from g_j to other members in its group and b_{g_j} is the minimum distance from g_j to other cluster centers. The standard measure of clustering, SC, can vary -1 (poor clustering) to $+1$ (good clustering). Typically, $SC > 0.4$ indicates good clustering. We cluster into 5 groups since SC for all 30 RNAs clustering is greater than 0.5 and give best prediction results (see Table S2). The average SC indicates how well separated the clusters are as well as how well cohered each cluster within the group. We select the graph from these 5 clusters as canddiate for procedure P3 by the lowest-scored representative.

1. Yang H, et al. (2003) Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res* 31(13):3450–3460.
2. Gendron P, Lemieux S, Major F (2001) Quantitative analysis of nucleic acid three-dimensional structures.*J Mol Biol* 308(5):919–936.
3. Petrov AI, Zirbel CL, Leontis NB (2011) WebFR3D–a server for finding, aligning and analyzing recurrent RNA 3D motifs. *Nucleic Acids Res* 39:W50–55.
4. Fera D, et al. (2004) RAG: RNA-As-Graphs web resource. *BMC Bioinformatics* 5:88.
5. Izzo JA, Kim N, Elmetwaly S, Schlick T (2011) RAG: an update to the RNA-As-Graphs resource. *BMC Bioinformatics* 12:219.
6. Schlick T (1988) A modular strategy for generating starting conformations and data-structures of polynucleotide helices for potential-energy calculations. *J Comp Chem* 9(8):861–889.
7. Draper NR and Smith H (1998) *Applied Regression Analysis.* Wiley, New York, NY.
8. Laing C, Jung S, Kim N, Elmetwaly S, Zahran M, Schlick T (2013) Predicting helical topologies in RNA junctions as tree graphs. *PLoS One*, 8(8):e71947.
9. Kabsch W (1976) A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crytal Physics, Diffraction, Theoretical and General Crystallography* 32(5):922–923.
10. Kaufman L, Rousseeuw PJ (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York, NY.

Table S1. List of 30 RNAs from the PDB database. The function of each structure is listed along with protein-binding property, sequence length, maximum degree of branches up to 4-way junction (4WJ), and numbers of internal loops (IL), junctions (J), and pseudoknots (PK).

PDB	Function	Protein-Binding	Seq. Length (nt)	Branching Degree	# of IL	# of J	PK
1RLG	box C/D RNA	Yes	25	IL	1	0	0
1OOA	NF-κB aptamer	Yes	29	IL	1	0	0
2IPY	iron-responsive element	Yes	30	IL	1	0	0
2OZB	snRNA	Yes	33	IL	2	0	0
1MJI	5S rRNA	Yes	34	IL	1	0	0
2HW8	mRNA	Yes	36	IL	1	0	0
1I6U	rRNA fragment	Yes	37	IL	2	0	0
1F1T	MG aptamer	No	38	IL	2	0	0
1ZHO	mRNA	Yes	38	IL	1	0	0
1S03	mRNA	Yes	47	IL	3	0	0
1XJR	viral RNA	No	47	IL	4	0	0
1U63	mRNA	Yes	49	IL	2	0	0
2PXB	SRP	Yes	49	IL	2	0	0
2OIU	RNA ligase	No	51	3WJ	1	1	0
1MZP	rRNA fragment	Yes	55	IL	2	0	0
2HGH	5S rRNA	No	55	3WJ	1	1	0
1DK1	rRNA fragment	Yes	57	3WJ	2	1	0
1MMS	rRNA fragment	Yes	58	3WJ	1	1	0
1D4R	SRP	No	58	IL	3	0	0
1KXK	group II intron	No	70	IL	3	0	0
1SJ4	HDV ribozyme	No	73	4WJ	2	0	1
1P5O	HCV IRES	No	77	IL	5	0	1
3D2G	A. Thaliana TPP riboswitch	No	77	3WJ	2	1	0
2HOJ	E. Coli TPP riboswitch	No	79	3WJ	2	1	0
2GDI	E. Coli TPP riboswitch	No	80	3WJ	3	1	0
2GIS	SAM riboswitch	No	94	4WJ	2	1	1
1LNG	SRP	Yes	97	3WJ	3	1	0
2LKR	U2/U6 snRNA	No	111	3WJ	4	1	0
1MFQ	SRP	Yes	128	3WJ	4	1	0
1GID	group I intron P4-P6	No	158	3WJ	6	1	0

Table S2. Graph results for 30 test RNAs. RMSD between reference graphs from solved structures and our sampled graphs by two MC/SA protocols (restricted and random pivot moves) – initial, lowest RMSD (P1), lowest score (P2) and lowest cluster representative (P3, only for random moves) and the correlation coefficient between RMSD and score (r) are shown. In comparison to predictions by MC-Sym, FARNA, and NAST, best RMSDs are indicated in bold. Best results for P3 and other tools are highlighted in gray.

PDB	Length	Initial	MC/SA (restricted moves) (Å)		MC/SA (random moves) (Å)			Pearson's	Other Tools (Å)		
ID	(nt)	(Å)	P1	P2	P1	P2	P3	r	MC-Sym	FARNA	NAST
1RLG	25	4.28	**2.20**	**4.18**	**2.20**	**4.00**	**4.17**	0.43	5.97	6.31	5.94
1OOA	29	3.72	**2.46**	**3.93**	**2.44**	**2.59**	**3.57**	0.93	4.12	8.46	6.23
2IPY	30	2.42	1.95	2.38	1.96	2.29	2.91	0.96	**1.22**	2.09	3.47
2OZB	33	7.30	**3.17**	6.95	**2.82**	6.57	5.45	0.46	5.52	**4.27**	5.34
1MJI	34	6.24	**2.47**	**2.48**	**2.47**	**3.28**	**3.15**	0.91	5.33	5.08	N/A
2HW8	36	8.32	**2.07**	**5.60**	**2.11**	6.35	**5.40**	0.16	8.37	6.19	5.85
1I6U	37	3.01	**1.37**	**2.56**	**1.47**	**2.58**	**2.52**	0.86	2.88	5.85	4.25
1F1T	38	2.98	**1.97**	**2.84**	**1.86**	**2.99**	**2.68**	0.83	3.01	6.07	4.83
1ZHO	38	8.74	**2.33**	6.50	**2.24**	7.46	7.24	0.20	7.91	**5.75**	8.09
1S03	47	4.56	2.05	4.72	1.93	3.74	3.23	0.79	**1.73**	4.67	6.57
1XJR	47	6.82	**3.74**	**6.18**	**3.48**	**5.45**	**4.25**	0.82	6.84	9.72	9.21
1U63	49	10.39	**2.74**	**7.89**	**2.55**	**7.83**	**6.01**	0.31	14.22	14.82	N/A
2PXB	49	4.71	**1.92**	**4.26**	**1.30**	5.07	**3.85**	0.89	4.84	5.52	5.04
2OIU	51	6.23	**2.83**	**6.14**	**2.79**	7.43	7.72	0.67	**6.40**	14.55	N/A
1MZP	55	8.47	**3.04**	7.20	**2.58**	6.94	6.74	0.55	14.09	11.70	**6.14**
2HGH	55	5.24	**4.18**	**5.89**	**4.17**	**5.93**	**7.16**	0.84	13.98	11.58	7.64
1DK1	57	10.32	**4.09**	**6.23**	**3.97**	**7.16**	**6.43**	0.56	8.14	15.59	9.47
1MMS	58	10.13	**4.25**	**10.79**	**4.48**	**10.97**	**10.35**	−0.29	18.00	18.31	11.13
1D4R	58	4.30	**2.93**	8.64	**3.69**	8.13	**7.27**	0.74	N/A	7.33	N/A
1KXK	70	6.21	**2.77**	5.24	**2.58**	**4.02**	5.29	0.88	**4.70**	7.21	7.04
1SJ4	73	13.63	**5.87**	12.28	**5.45**	12.08	11.14	0.13	N/A	**7.10**	N/A
1P5O	77	14.64	**5.20**	9.55	**4.33**	8.33	9.44	0.82	**6.69**	9.38	9.14
3D2G	77	13.39	**4.08**	15.74	**3.62**	17.24	18.34	−0.42	**10.97**	16.67	N/A
2HOJ	79	15.40	**5.09**	**14.93**	**4.70**	**15.83**	**13.01**	−0.47	16.34	17.64	N/A
2GDI	80	13.73	**6.13**	17.98	**4.25**	15.25	17.90	−0.44	13.81	19.11	**12.90**
2GIS	94	15.34	12.95	13.43	12.47	17.87	17.43	−0.01	19.04	**12.33**	N/A
1LNG	97	6.20	**4.48**	**13.43**	**4.98**	**12.51**	**14.56**	0.22	17.29	19.18	27.98
2LKR	111	16.78	**11.32**	30.89	**12.57**	19.31	16.90	0.17	**15.47**	25.42	16.35
1MFQ	128	7.78	**6.68**	**12.08**	**6.77**	**9.28**	**10.55**	0.63	35.28	16.48	27.76
1GID	158	46.11	**14.56**	29.56	**10.62**	28.63	28.24	0.42	N/A	**27.13**	61.03

(N/A: program fails.)

Table S3. RMSD of the lowest cluster representative graph using the number of clusters from 2 to 6 of MC/SA based on random moves with 10^4 MC steps without knowledge of reference structures for 30 RNAs and comparisons with other tools including MC-Sym, FARNA, and NAST. The bold fonts indicate the best results among all prediction results for each structure. The graph RMSD of our prediction better than any of three prediction results is indicated in bold. The last row indicates the number of RNAs predicted better than other tools. The best RMSD among all predictions (P3 in 5 groups and three other methods) is highlighted in gray. (N/A: program fails).

PDB	2gps	3gps	4gps	5gps	6gps	MC-Sym	FARNA	NAST
1RLG	**3.65**	**3.63**	**3.75**	**4.17**	**3.31**	5.97	6.31	5.94
1OOA	4.28	**3.14**	**3.56**	**3.57**	**2.81**	4.12	8.46	6.23
2IPY	2.56	2.62	2.76	2.91	2.41	**1.22**	2.09	3.47
2OZB	5.03	5.95	4.90	5.45	5.59	5.52	**4.27**	5.34
1MJI	**2.52**	**2.69**	**2.62**	**3.15**	**2.76**	5.33	5.08	N/A
2HW8	6.82	6.37	7.24	**5.40**	7.50	8.37	6.19	5.85
1I6U	3.27	**2.72**	**2.61**	**2.52**	**1.87**	2.88	5.85	4.25
1F1T	3.32	3.71	3.50	**2.68**	**2.82**	3.01	6.07	4.83
1ZHO	6.84	6.75	6.77	7.24	7.24	7.91	**5.75**	8.09
1S03	2.76	3.32	3.99	3.23	3.91	**1.73**	4.67	6.57
1XJR	**5.81**	**4.24**	**4.57**	**4.25**	**4.25**	6.84	9.72	9.21
1U63	**7.35**	**5.69**	**6.10**	**6.01**	**6.01**	14.22	14.82	N/A
2PXB	5.09	**2.75**	5.86	**3.85**	**4.11**	4.84	5.52	5.04
2OIU	8.31	**5.97**	**6.09**	7.72	7.00	**6.40**	14.55	N/A
1MZP	**4.68**	**5.76**	**5.50**	6.74	**6.01**	14.09	11.70	**6.14**
2HGH	**4.64**	**7.11**	**6.07**	**7.16**	**6.02**	13.98	11.58	7.64
1DK1	8.51	**8.11**	**5.86**	**6.43**	**6.63**	8.14	15.59	9.47
1MMS	**8.78**	**9.20**	**9.49**	**10.35**	**9.73**	18.00	18.31	11.13
1D4R	8.25	8.60	8.89	**7.27**	7.38	N/A	7.33	N/A
1KXK	8.47	7.18	6.21	7.59	6.78	**4.70**	7.21	7.04
1SJ4	9.89	10.01	10.55	11.14	10.64	N/A	**7.10**	N/A
1P5O	9.62	11.46	9.35	9.44	7.33	**6.69**	9.38	9.14
3D2G	14.52	13.90	15.71	18.34	16.51	**10.97**	16.67	N/A
2HOJ	**15.25**	17.34	**12.66**	**13.01**	**15.06**	16.34	17.64	N/A
2GDI	**10.43**	16.67	16.80	17.90	17.11	13.81	19.11	**12.90**
2GIS	16.33	16.86	17.01	17.43	16.43	19.04	**12.33**	N/A
1LNG	**16.28**	**8.66**	**14.98**	**14.56**	**10.17**	17.29	19.18	27.98
2LKR	17.07	20.83	**14.34**	16.90	16.88	**15.47**	25.42	16.35
1MFQ	**12.67**	**15.93**	14.81	**10.55**	**10.27**	35.28	16.48	27.76
1GID	31.52	33.21	30.61	28.24	**26.66**	N/A	**27.13**	61.03
#	11	14	15	16	16	7	5	2

Table S4. Graph results of MC/SA based on random moves with 10^4 steps without knowledge of reference structures for 30 RNAs and comparisons with other tools including MC-Sym, FARNA, and NAST. The graph RMSDs from native structure graphs and five representative graphs from Cluster 1 (lowest score, P3) to 5 (highest score), and the correlation coefficients r between graph RMSD from native structures and MC scores are shown. The graph RMSD of our prediction better than any of three prediction results (or other prediction results better than any of five representatives) is indicated in bold and the numbers are on the last row. The best RMSD among all graphs (five representative graphs and three other methods) is highlighted in gray. Corresponding cluster ID is indicated. A more thorough statistical analysis based on linear models to predict the best cluster ranks (the predicted cluster ID by the linear model EstR∼0.5572*IL + 1.3290*J + 1.2211*PK+1) did not reveal ways to improve results consistently, but this may be reassessed in the future with improvements in the scoring system for the cases with protein-binding cases and junction structures. (N/A: program fails).

PDB	C1	C2	C3	C4	C5	r	Cluster ID	EstR	MC-Sym	FARNA	NAST
1RLG	**4.17**	**3.55**	**3.74**	**3.66**	**4.54**	0.43	2	1	5.97	6.31	5.94
1OOA	**3.57**	4.36	5.85	8.20	10.74	0.93	1	1	4.12	8.46	6.23
2IPY	2.91	3.58	4.97	6.67	9.93	0.96	1	1	**1.22**	2.09	3.47
2OZB	5.45	4.75	4.86	7.19	8.93	0.46	2	1	5.52	**4.27**	5.34
1MJI	**3.15**	**2.75**	**3.95**	5.14	7.74	0.91	2	1	5.33	5.08	N/A
2HW8	**5.40**	6.49	6.61	9.25	6.24	0.16	1	1	8.37	6.19	5.85
1I6U	**2.52**	3.53	6.58	7.71	10.90	0.86	1	1	2.88	5.85	4.25
1F1T	**2.68**	4.30	5.83	7.90	11.39	0.83	1	1	3.01	6.07	4.83
1ZHO	7.24	**4.39**	**3.99**	6.20	7.18	0.20	3	1	7.91	5.75	8.09
1S03	3.23	3.48	5.76	8.98	11.77	0.79	1	2	**1.73**	4.67	6.57
1XJR	**4.25**	**4.69**	**6.50**	8.12	11.21	0.82	1	2	6.84	9.72	9.21
1U63	**6.01**	**7.27**	**4.93**	**5.24**	**7.57**	0.31	3	1	14.22	14.82	N/A
2PXB	**3.85**	**4.22**	6.82	12.67	14.86	0.89	1	1	4.84	5.52	5.04
2OIU	7.72	**3.85**	**4.96**	8.60	14.42	0.67	2	2	6.40	14.55	N/A
1MZP	6.74	**4.65**	6.33	6.83	8.48	0.55	2	1	14.09	11.70	6.14
2HGH	**7.16**	**5.62**	8.83	7.43	11.56	0.84	2	2	13.98	11.58	7.64
1DK1	**6.43**	10.32	8.96	10.62	16.44	0.56	1	2	8.14	15.59	9.47
1MMS	**10.35**	**8.04**	**7.55**	**8.14**	**7.85**	–0.29	3	2	18.00	18.31	11.13
1D4R	**7.27**	10.16	10.60	13.24	17.72	0.74	1	2	N/A	7.33	N/A
1KXK	5.29	4.78	8.90	11.15	15.81	0.88	1	2	**4.70**	7.21	7.04
1SJ4	11.14	9.06	11.80	7.90	10.71	0.13	4	2	N/A	**7.10**	N/A
1P5O	9.44	11.21	9.27	12.84	19.73	0.82	3	3	**6.69**	9.38	**9.14**
3D2G	18.34	12.61	13.47	**9.82**	12.30	–0.42	4	3	10.97	16.67	N/A
2HOJ	**13.01**	**12.57**	**15.65**	**11.91**	**11.55**	–0.47	5	3	16.34	17.64	N/A
2GDI	17.90	16.68	**11.34**	**12.22**	13.71	–0.44	3	3	13.81	19.11	12.90
2GIS	17.43	15.35	17.10	16.80	20.59	–0.01	2	4	19.04	**12.33**	N/A
1LNG	**14.56**	18.17	**12.57**	18.68	18.24	0.22	3	3	17.29	19.18	27.98
2LKR	16.90	21.15	18.28	22.17	17.73	0.17	1	4	**15.47**	25.42	**16.35**
1MFQ	**10.55**	**13.71**	18.25	**15.48**	20.40	0.63	1	4	35.28	16.48	27.76
1GID	28.24	29.80	29.86	**25.92**	**25.04**	0.42	5	5	N/A	27.13	61.03
#	16	12	10	8	5		22	17	5	3	2

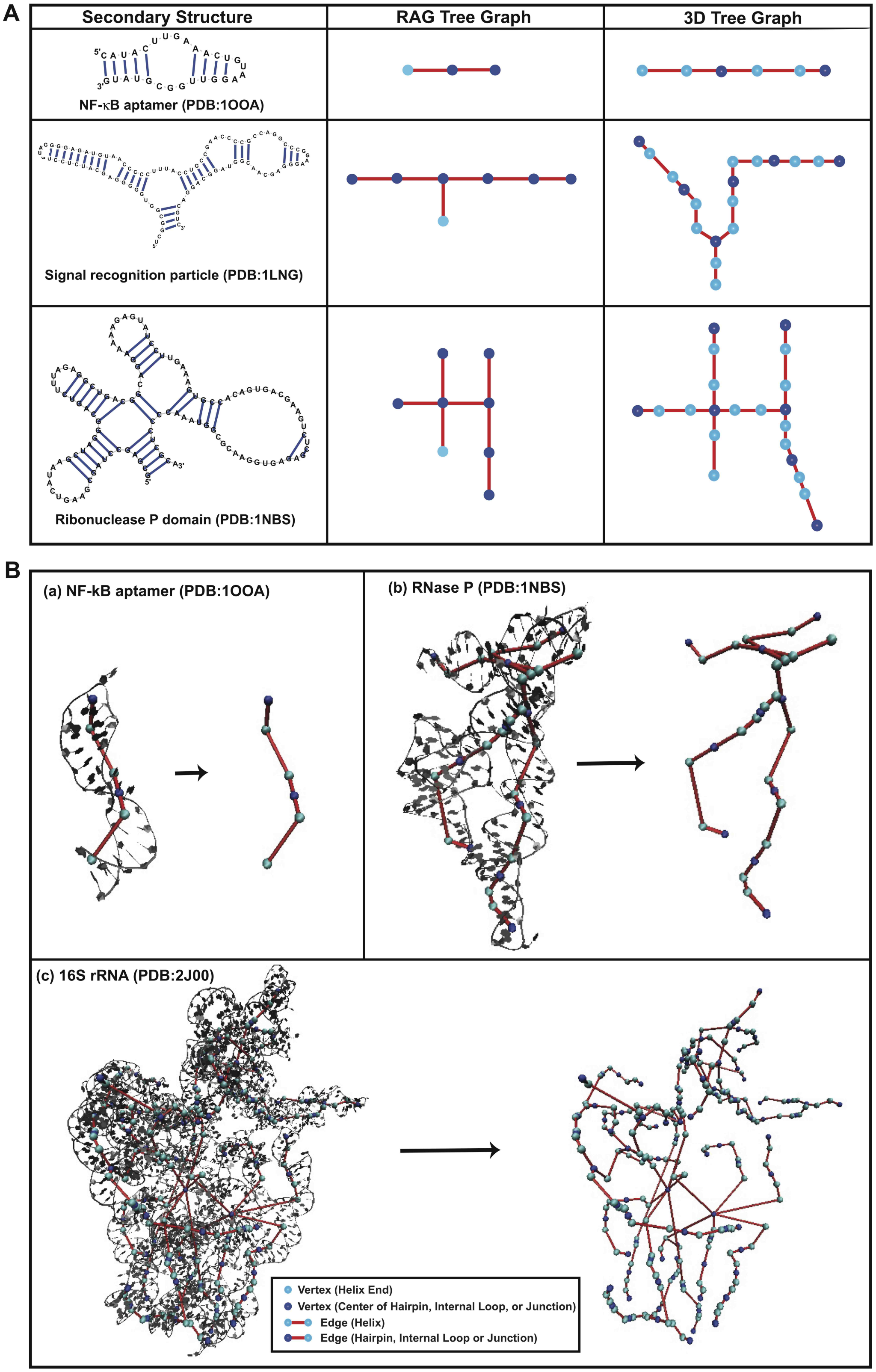
A
Secondary Structure
RAG Tree Graph
3D Tree Graph
NF-κB aptamer (PDB:1OOA)
Signal recognition particle (PDB:1LNG)
Ribonuclease P domain (PDB:1NBS)
5'
3'
B
(a) NF-kB aptamer (PDB:1OOA)
(b) RNase P (PDB:1NBS)
(c) 16S rRNA (PDB:2J00)
● Vertex (Helix End)
● Vertex (Center of Hairpin, Internal Loop, or Junction)
Edge (Helix)
Edge (Hairpin, Internal Loop or Junction)

Fig. S1. RNA graph representations: **(A)** RNA 2D structures, RAG, and refined RAG representations embedded in 3D space of RNA aptamer for transcription factor NF-κB (PDB: 1OOA), signal recognition particle (PDB: 1LNG), and RNase P (PDB: 1NBS); **(B)** RNA 3D structures and 3D graphs of (a) RNA aptamer for transcription factor NF-κB (PDB: 1OOA), (b) RNase P (PDB: 1NBS), and (c) 70S ribosomal RNA (PDB: 2J00).

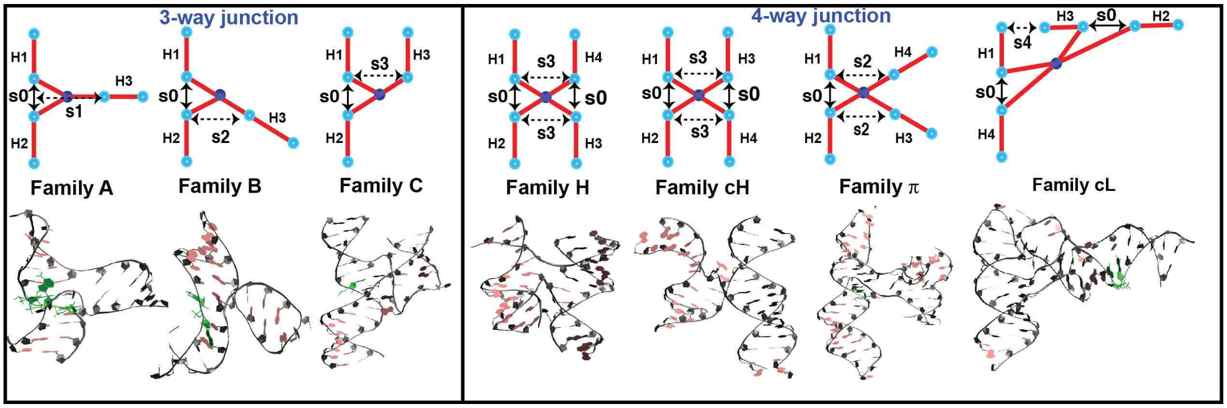

Fig. S2. Size measures of 3-way junction families – A (perpendicular), B (diagonal), and C (parallel) – and 4-way junction families – H (parallel), cH (crossed and parallel), π (diagonal), and cL (perpendicular), representing topologies predicted by RNAJAG. The junction distance parameter s0 denotes the distance between coaxial helices. The distance parameters s1, s2, and s3 measure distances between the non-coaxial helices, for perpendicular, diagonal, and parallel helices, respectively. We also calculate distance between disconnected perpendicular helices (H2 and H4) in cL 4-way junction family (denoted as s4).

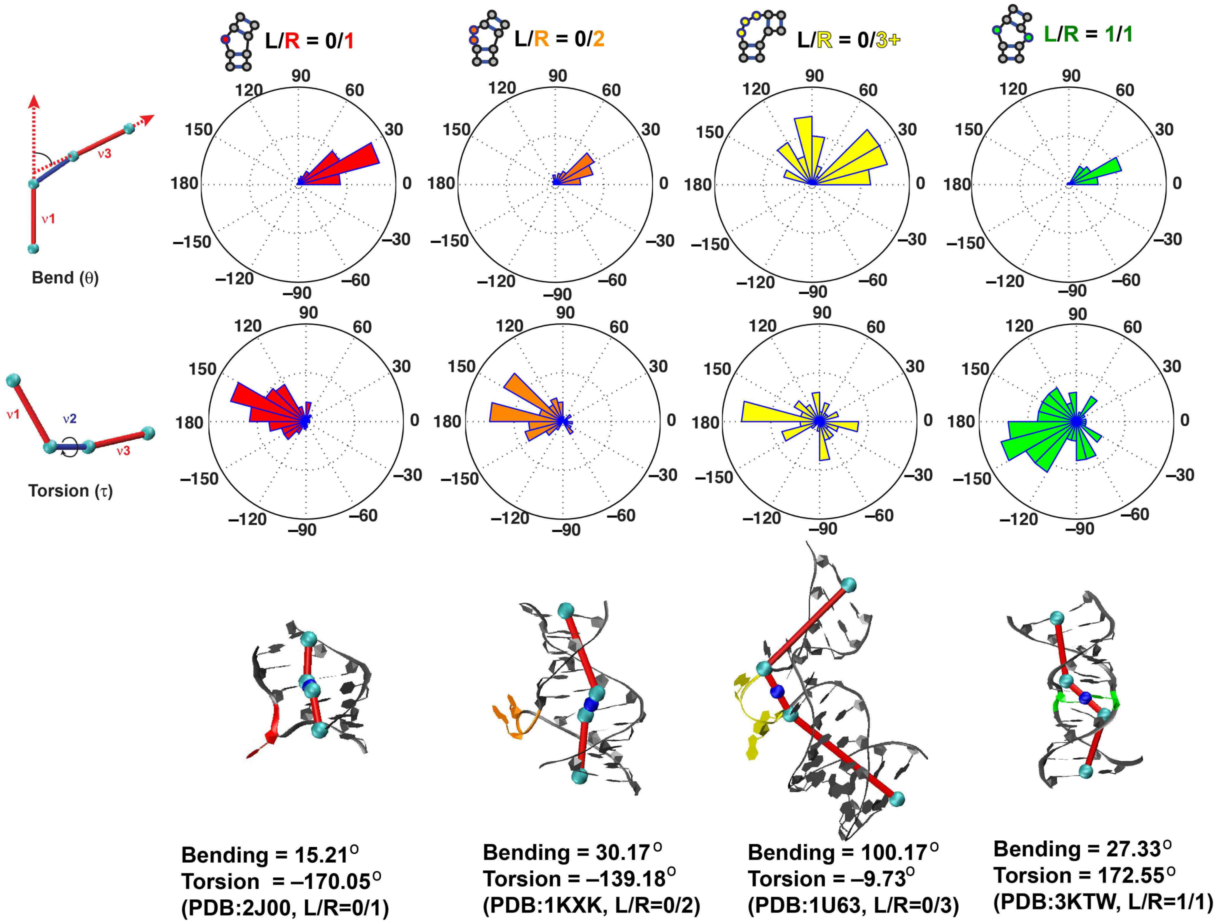

Fig. S3. Bending and torsion angles of internal loops. Distributions of bend and torsion angles of internal loops with L/R = 0/1, 0/2, 0/3+, and 1/1 are shown with RNA structures. Examples include: 2J00 (residues from G1193 to U1199 and from G1058 to C1063 with L/R = 0/1, as shown in red, bending = 15.21° and torsion = −170.05°), 1KXK (residues from U4 to C15 and from G57 to A66 with L/R = 0/2, as shown in orange, bending = 30.17° and torsion = −139.18°), 1U63 (residues from G2 to U16 and from A30 to C48 with L/R = 0/3, as shown in yellow, bending = 100.17° and torsion = −9.73°), and 3KTW (residues from C199 to C209 and from G214 to G224 with L/R = 1/1, as shown in green, bending = 22.33° and torsion = 172.55°).

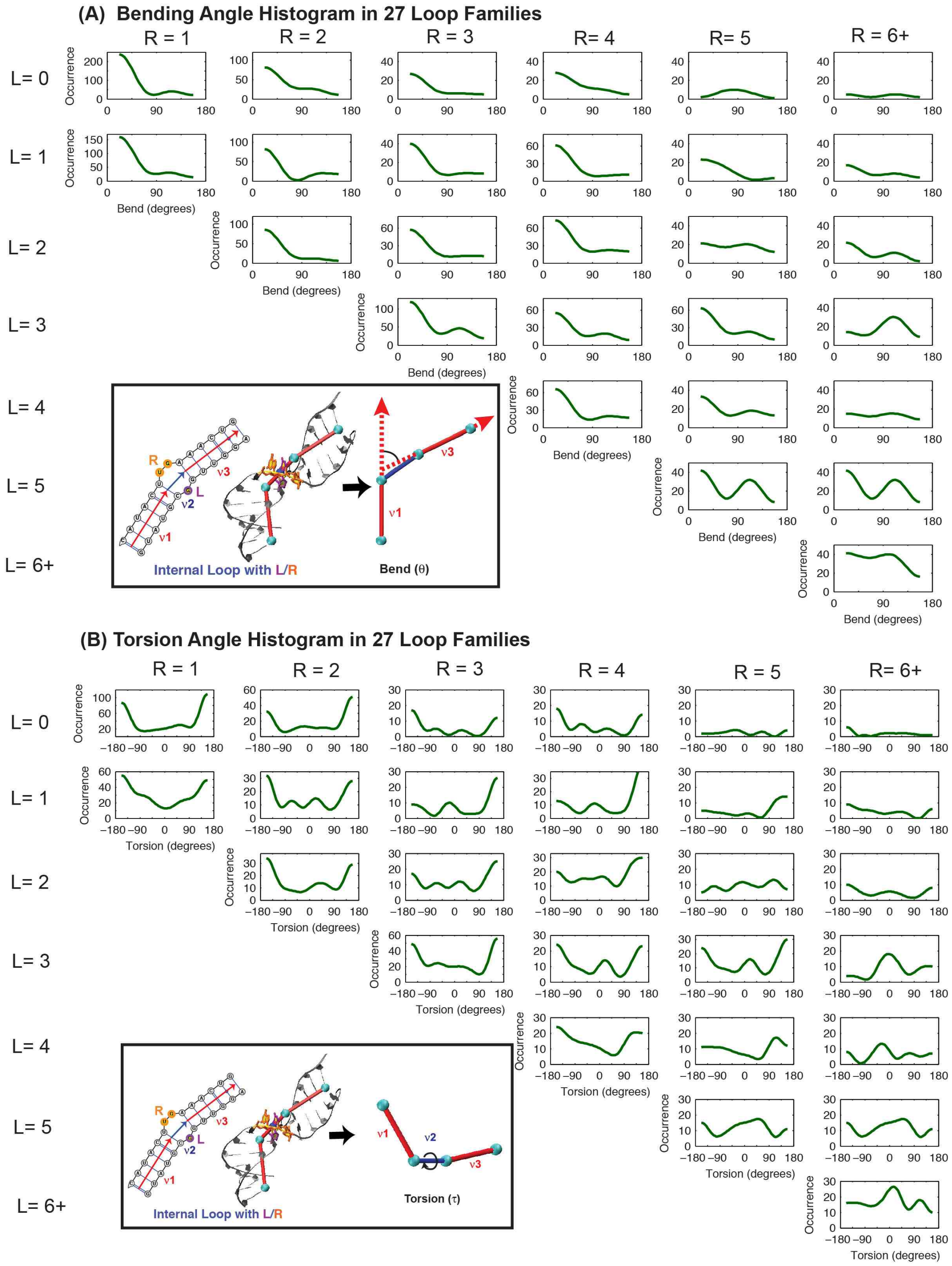

Fig. S4. Histogram of (A) bending and (B) torsion angles of 27 internal loop families. The 27 subplots correspond to the 27 internal loop families by L/R = 0/1, 0/2, 0/3, 0/4, 0/5, and 0/6+ (first row); 1/1, 1/2, 1/3, 1/4, 1/5, and 1/6+ (second row); 2/2, 2/3, 2/4, 2/5, and 2/6+(third row); 3/3, 3/4, 3/5, and 3/6+(fourth row); 4/4, 4/5, and 4/6+ (fifth row); 5/5, and 5/6+ (sixth row); and 6+/6+(seventh row). The L and R represent the nucleotide lengths of two single strands of internal loops where L is smaller than R.

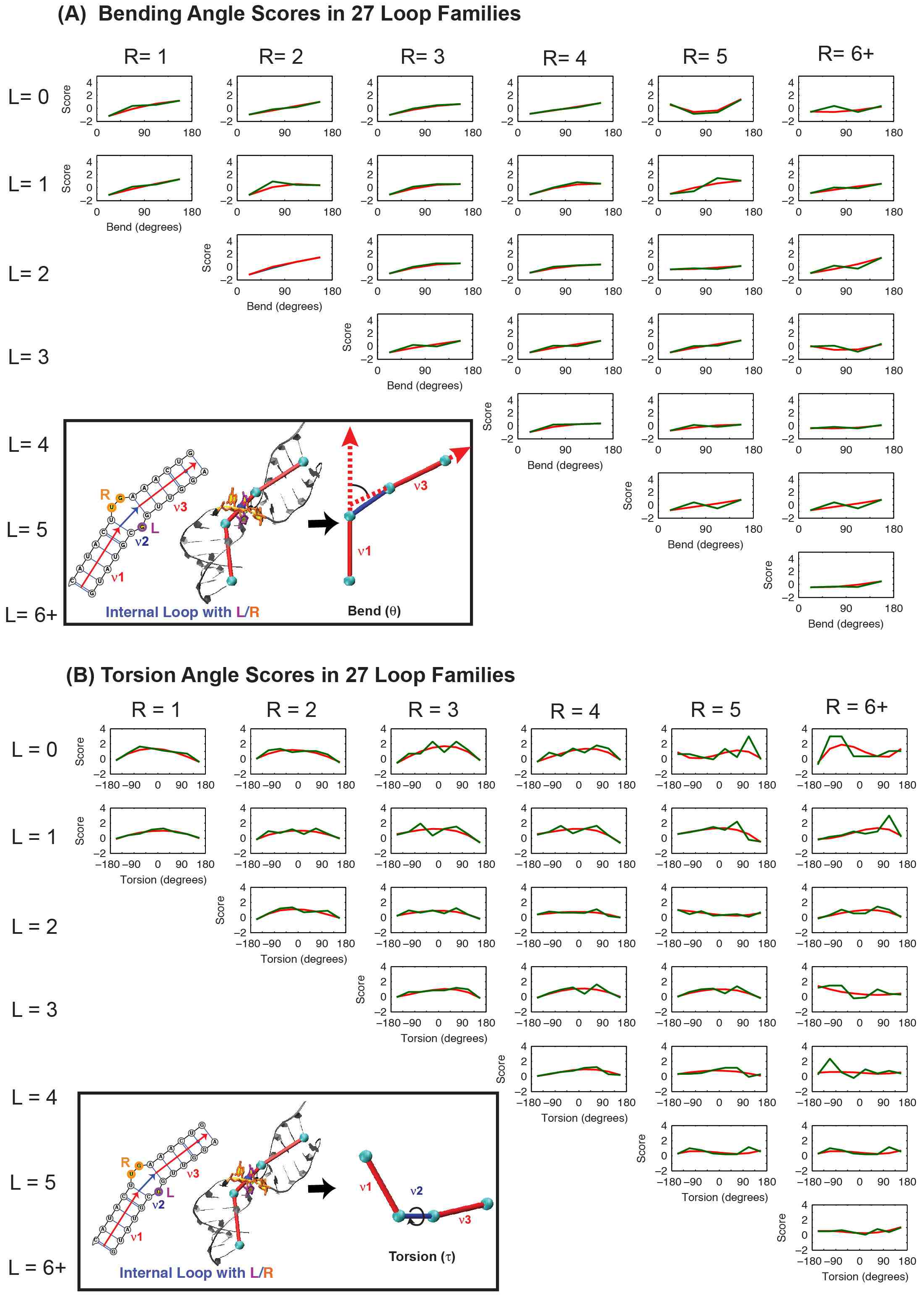

Fig. S5. Knowledge-based statistical potential for (A) bend and (B) torsion angles before (green) and after (red) optimization, along each bin of the discrete angular variable. The 27 subplots correspond to the 27 groups of internal loop defined in the text, respectively L/R = 0/1, 0/2, 0/3, 0/4, 0/5, and 0/6+ (first row); L/R = 1/1, 1/2, 1/3, 1/4, 1/5, and 1/6+ (second row); L/R = 2/2, 2/3, 2/4, 2/5, and 2/6+ (third row); L/R = 3/3, 3/4, 3/5, and 3/6+ (fourth row); L/R = 4/4, 4/5, and 4/6+ (fifth row); 5/5 and 5/6+ (sixth row); L/R = 6+/6+ (seventh row), where L and R are the nucleotide lengths of the single strands located on each side of the loop, and 6+ means more than or equal to 6.

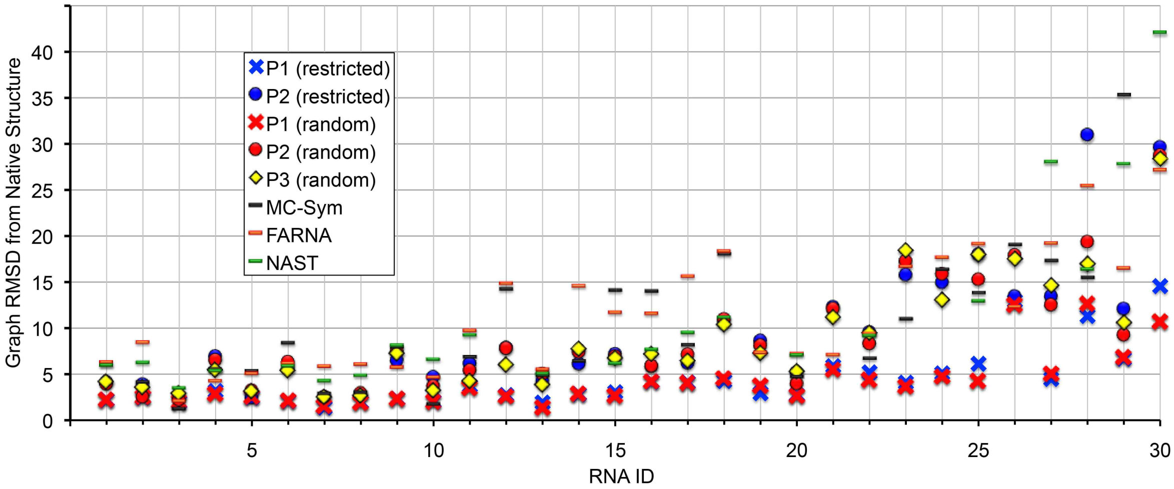

Fig. S6. Distribution of RMSD for the representative 30 RNAs using our graph predictions (P1–P3) after MC/SA based on two types of moves (restricted and random) and other 3D structure prediction programs.

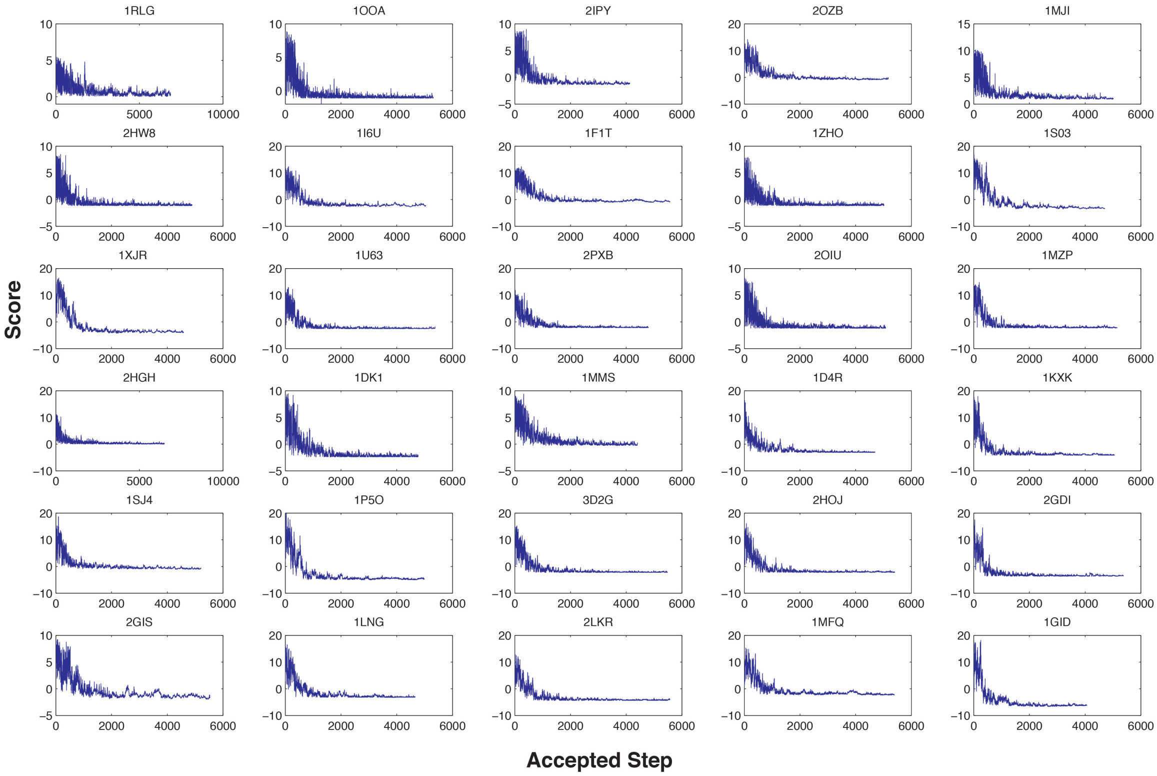

Fig. S7. The convergence of MC scores for 30 RNAs. The trajectories of scores along accepted steps among 10^4 MC steps after MC/SA based on restricted moves are shown.

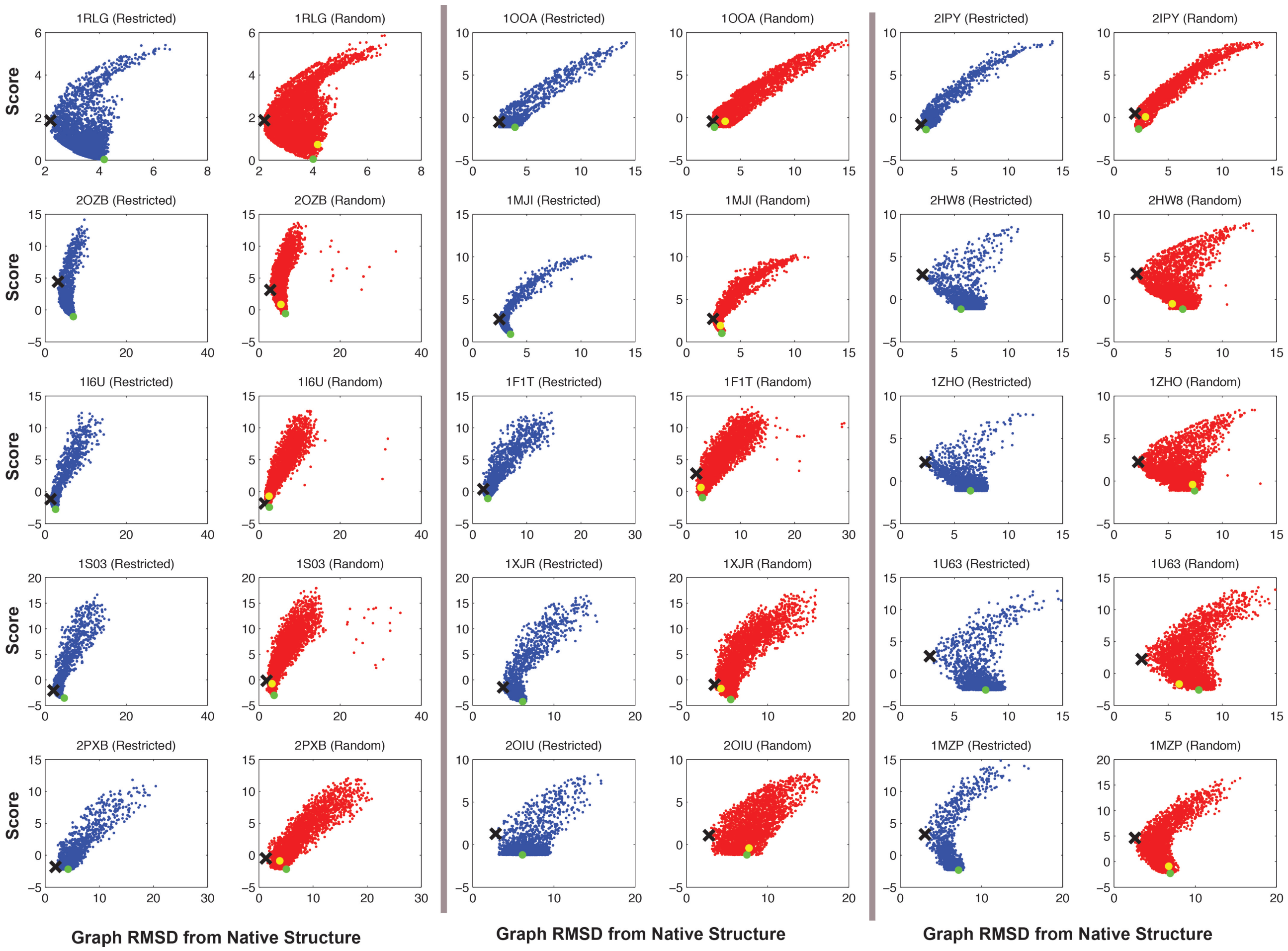

Fig. S8. The landscapes of MC scores against graph RMSD from native structure of accepted graphs after MC/SA based on restricted moves (which converge to one region, blue) and random moves (which explore multiple regions, red). Graphs selected by P1, P2, and P3 are indicated in each plot, as X, green dot, and yellow dot, respectively. (continued on the next page).

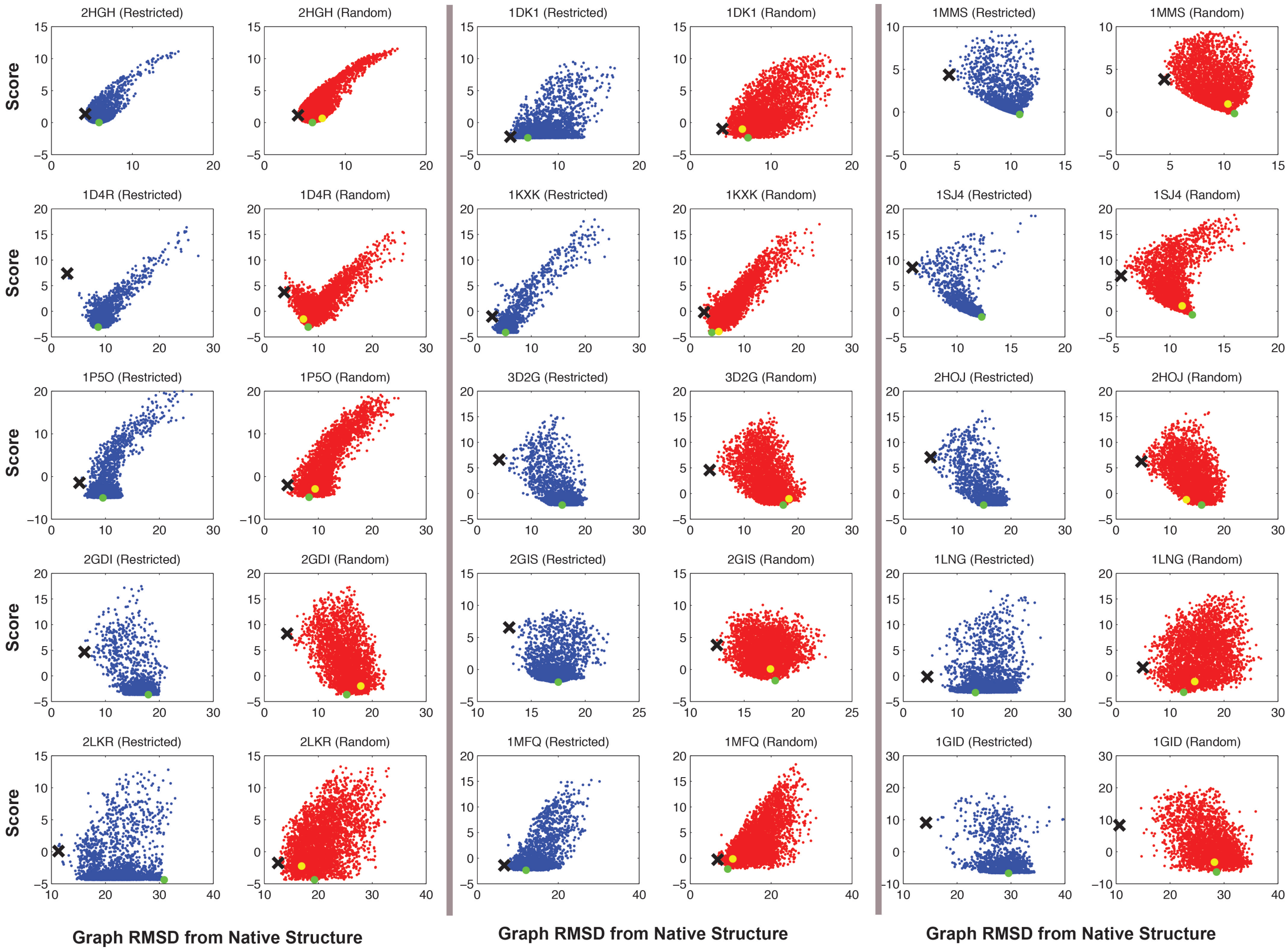

Fig. S8. The landscapes of MC scores against graph RMSD from native structure of accepted graphs after MC/SA based on restricted moves (which converge to one region, blue) and random moves (which explore multiple regions, red). Graphs selected by P1, P2, and P3 are indicated in each plot, as X, green dot, and yellow dot, respectively.

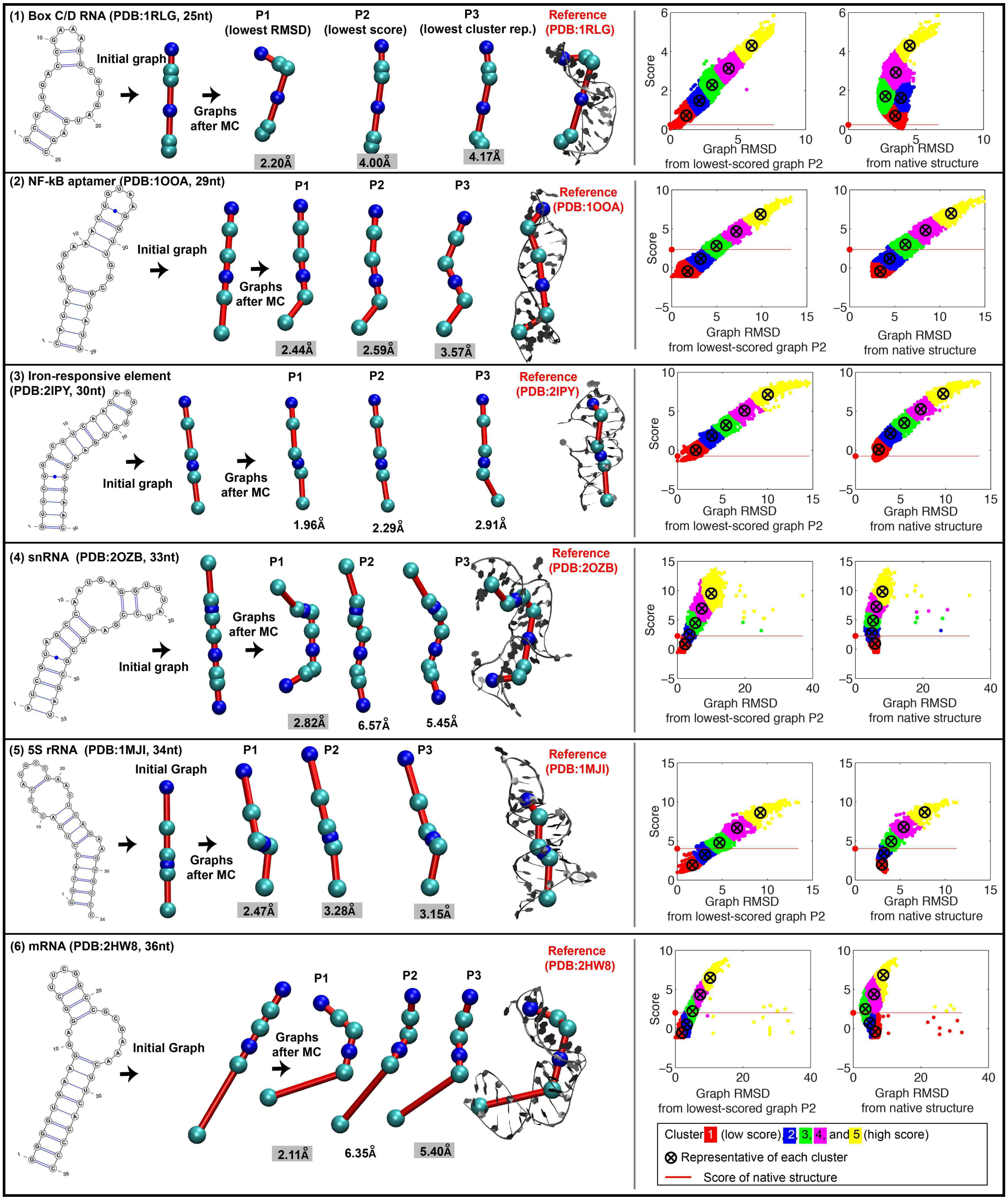

Fig. S9. Graph results for 30 RNAs. The input 2D structure, initial graphs before MC/SA, the lowest RMSD (P1), the lowest-scored (P2), the lowest representative among 5 clusters (P3) after MC/SA based on random moves, and reference graphs translated from solved structures, and landscapes are shown. Graphs selected by P1 and P2 based on restricted moves are similar to those by random moves shown here. (continued on the next page.)

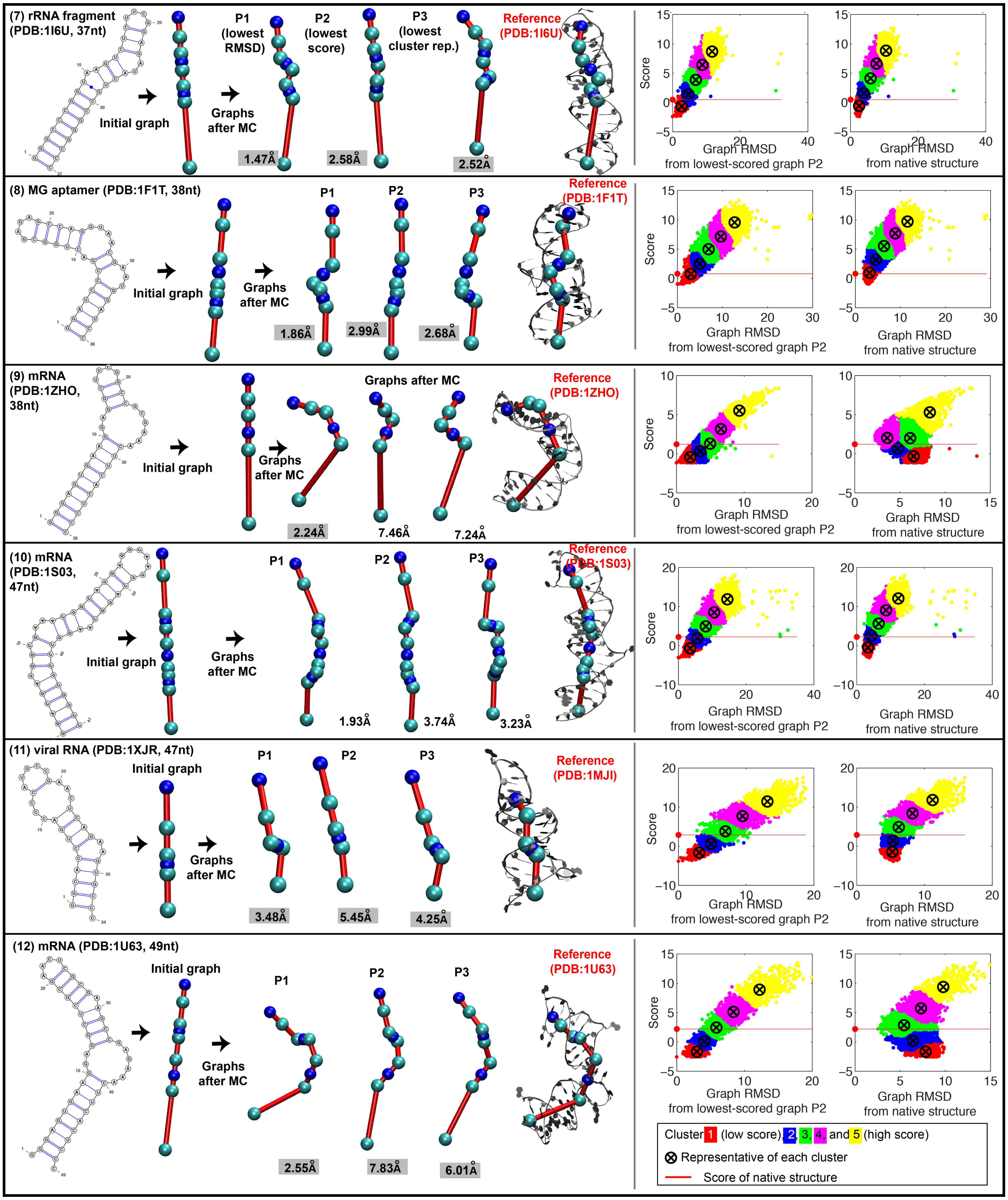

Fig. S9. Graph results for 30 RNAs. The input 2D structure, initial graphs before MC/SA, the lowest RMSD (P1), the lowest-scored (P2), the lowest representative among 5 clusters (P3) after MC/SA based on random moves, and reference graphs translated from solved structures, and landscapes are shown. Graphs selected by P1 and P2 based on restricted moves are similar to those by random moves shown here. (continued on the next page.)

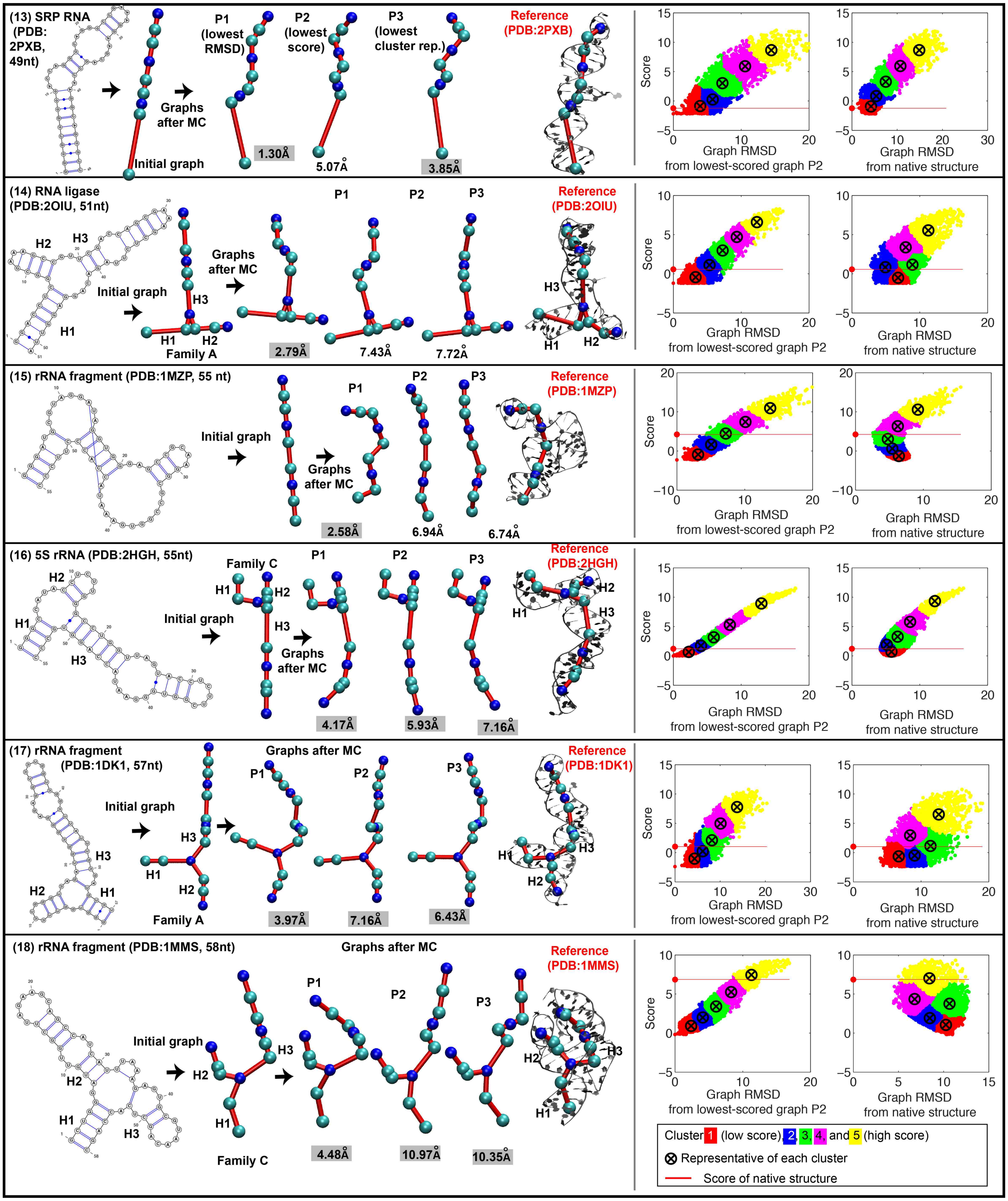

Fig. S9. Graph results for 30 RNAs. The input 2D structure, initial graphs before MC/SA, the lowest RMSD (P1), the lowest-scored (P2), the lowest representative among 5 clusters (P3) after MC/SA based on random moves, and reference graphs translated from solved structures, and landscapes are shown. Graphs selected by P1 and P2 based on restricted moves are similar to those by random moves shown here. (continued on the next page.)

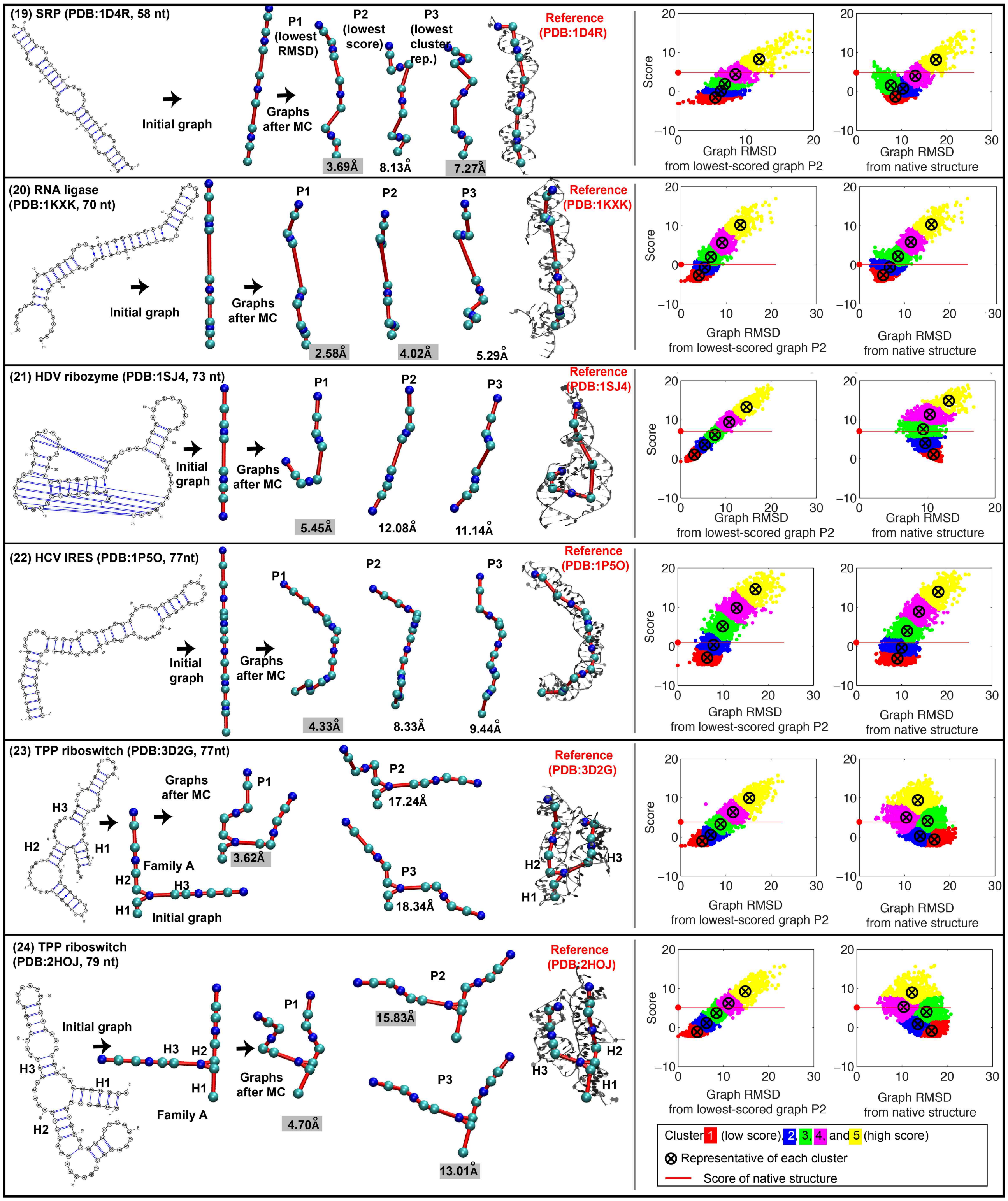

Fig. S9. Graph results for 30 RNAs. The input 2D structure, initial graphs before MC/SA, the lowest RMSD (P1), the lowest-scored (P2), the lowest representative among 5 clusters (P3) after MC/SA based on random moves, and reference graphs translated from solved structures, and landscapes are shown. Graphs selected by P1 and P2 based on restricted moves are similar to those by random moves shown here. (continued on the next page.)

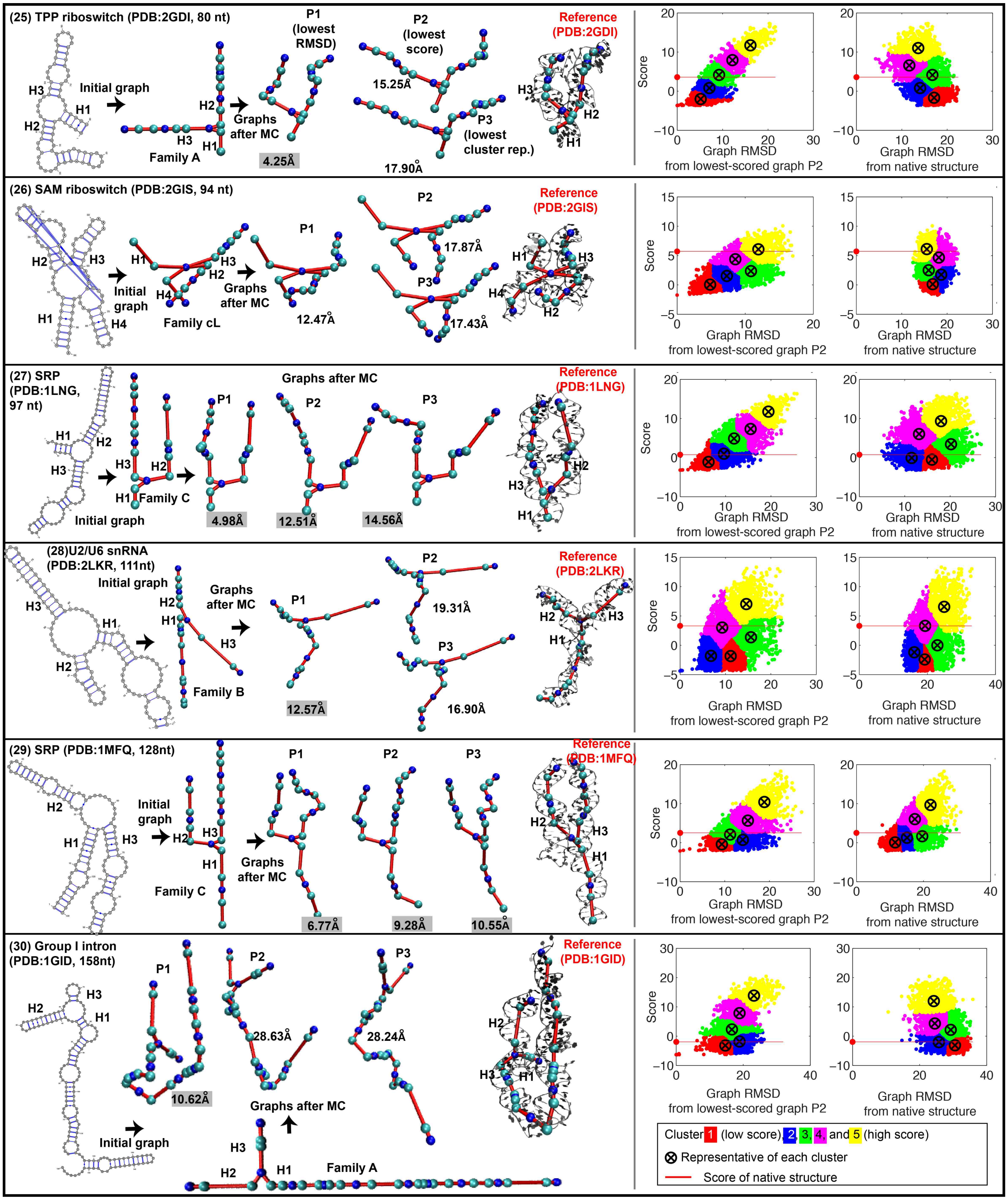

Fig. S9. Graph results for 30 RNAs. The input 2D structure, initial graphs before MC/SA, the lowest RMSD (P1), the lowest-scored (P2), the lowest representative among 5 clusters (P3) after MC/SA based on random moves, and reference graphs translated from solved structures, and landscapes are shown. Graphs selected by P1 and P2 based on restricted moves are similar to those by random moves shown here.

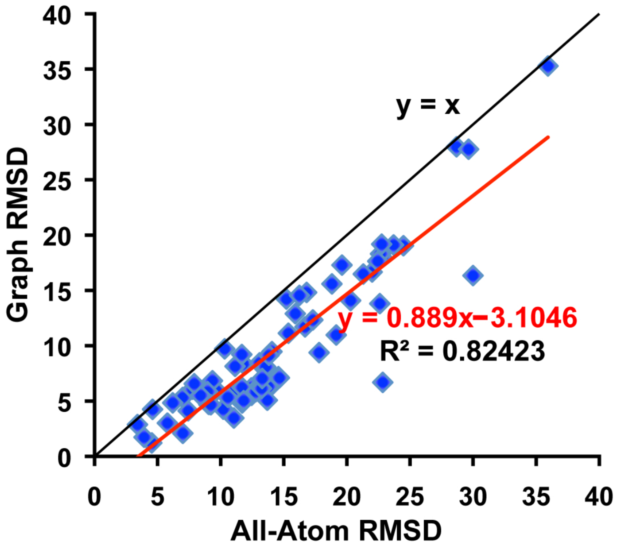

Fig. S10. Linear regression analysis of graph RMSD with respect to all-atom RMSD. A highly positive trend between graph and atom RMSD is observed with a slope value of 0.889.