Supplementary Information

Supplementary Figures

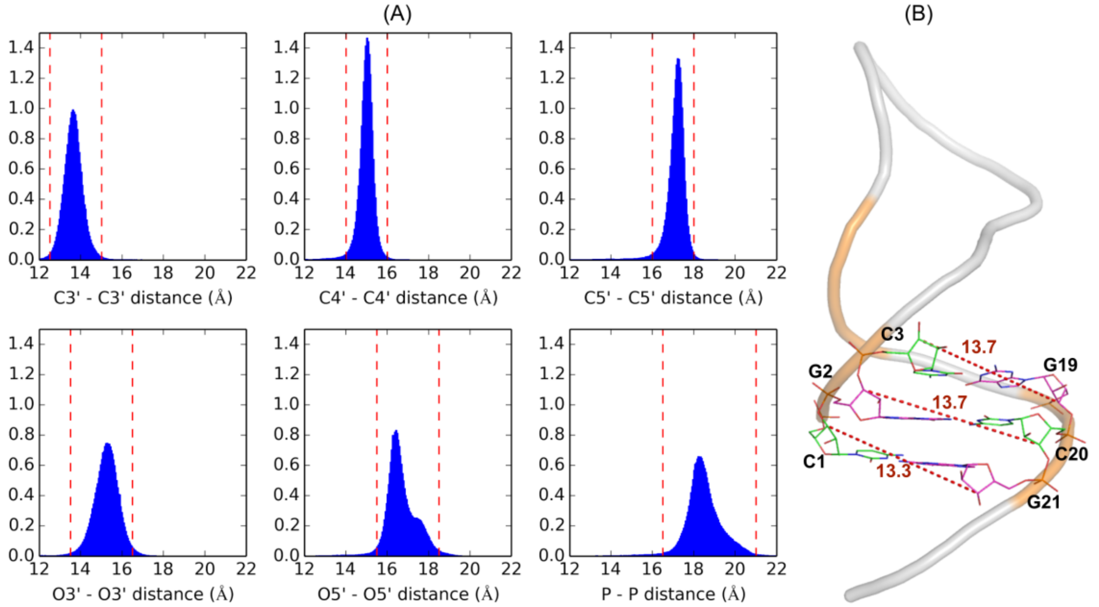
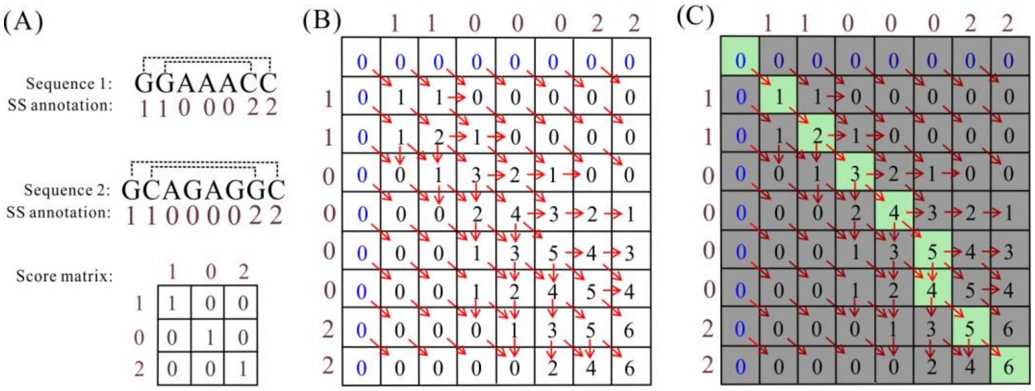


Fig. S1. Illustration of coarse-grained RNA secondary structure (SS) assignment in RNA-align, which assigns one of the three possible SS states (unpaired, paired with an upstream nucleotide, and paired with a downstream nucleotide) to each nucleotide. (A) Distance distribution of backbone atoms in Watson-Crick or Wobble base pairs. Vertical dash lines mark the upper and lower bound of atomic distances for considering two bases forming a base pair. By default, RNA-align only uses the C3’ atoms (upper left), whereas users have the option to use any one of the listed backbone atom type. (B) Part of structure from PDB ID 2ann Chain-B is shown as an example of SS assignment. Guanine, i.e. G, bases are shown in magenta lines while Cytosine, i.e. C, bases are shown in green lines. For two nucleotides to be considered as forming base pair, they must satisfy the following three conditions. First, the C3’ atom distance (red dash lines) should fall within the upper and lower bounds defined in Fig. S1A. Second, only G-C pair, G-U pair and A-U pair are allowed. Third, singleton pair is excluded; i.e. if neither nucleotide pair *i*-1 and *j*+1, nor nucleotide pair *i*+1 and *j*-1 satisfy the above two criteria, nucleotide *i* and *j* are never considered paired either. Based on these three criteria, RNA-align assigns that nucleotides C1, G2, and C3 form base pair with nucleotides G21, C20, and G19, respectively. This is consistent with the assignment based on full-atom structures, which should meet the following three conditions. First, only G-C pair, G-U pair and A-U pair are allowed. Second, the distance of the heavy atoms involved in hydrogen bond formation should be in [2,4] Å. For example, the distance between N1, N2, O6 of G and N3, O2, N4 of C, respectively, should be within this range if bases G and C form base pair. Third, the included angle should not exceed $\pi/3$. Included angle are formed by C4, N1 of base G and N3 of base C (C4/G-N1/G-N3/C) for GC pair (C4/A-N1/A-N3/U for AU pair; C6/U-N3/U-O6/G for GU pair). The programs to assign SS based on coarse-grain and full-atom structures are both available to download at https://zhanglab.ccmb.med.umich.edu/RNA-align/download.html.



	1	0	2
1	1	0	0
0	0	1	0
2	0	0	1

(B)

		1	1	0	0	0	2	2
	0	0	0	0	0	0	0	0
1	0	1	1	0	0	0	0	0
1	0	1	2	1	0	0	0	0
0	0	0	1	3	2	1	0	0
0	0	0	0	2	4	3	2	1
0	0	0	0	1	3	5	4	3
0	0	0	0	1	2	4	5	4
2	0	0	0	0	1	3	5	6
2	0	0	0	0	0	2	4	6

(C)

		1	1	0	0	0	2	2
	0	0	0	0	0	0	0	0
1	0	1	1	0	0	0	0	0
1	0	1	2	1	0	0	0	0
0	0	0	1	3	2	1	0	0
0	0	0	0	2	4	3	2	1
0	0	0	0	1	3	5	4	3
0	0	0	0	1	2	4	5	4
2	0	0	0	0	1	3	5	6
2	0	0	0	0	0	2	4	6

Fig. S2. Illustration of dynamic programming (DP) for secondary structure (SS) alignment. (A) For each position of the two RNAs, RNA-align assigns three state SS based on whether and how the nucleotides are paired in the RNA chain: 0 for unpaired, 1 for pairing with downstream base, and 2 for pairing with upstream base. A score matrix for aligning different SS type is defined: 1 for aligning identical SS type, and 0 for aligning different SS type. (B) The $(1 + L_2) \times (1 + L_1)$ dynamic programming matrix is defined, where L_1 =7 and $L_2 = 8$ are the length of the two RNAs. The first column and first row is initialized with zeros (blue). For the rest cells of the DP matrix (black), the value $m_{i,j}$ at the *i*th row and *j*th column can be iteratively calculated by $m_{i,j} = \max\{m_{i-1,j-1} + S_{ij}, m_{i-1,j} + g, m_{i,j-1} + g\}$, where $g = -1$ is gap penalty and S_{ij} is the score matrix value of aligning SS type at *j*th position in the first RNA and SS type at *i*th position of the second RNA. Red arrows denote which of the three cells the value of current cell (i, j) comes from: a down arrow for $m_{i,j}$ deriving from $m_{i-1,j}$, a right arrow for $m_{i,j}$ coming from $m_{i,j-1}$, and a diagonal arrow for $m_{i,j}$ from $m_{i-1,j-1}$. (C) Starting from the lower right cell, the alignment path (green) is derived by tracing back the arrows. The alignment in this case is:

```
GGAAA−CC
: : : : :   : :
GCAGAGGC
```

The above DP algorithm simplifies the standard Gotoh algorithm (Gotoh, 1982) for global alignment: in this simplified DP, $m_{i,j}$ only depends on $m_{i-1,j-1}$, $m_{i-1,j}$, and $m_{i,j-1}$, but not on $m_{i-k,j}$ or $m_{i,j-k}$ where $k \geq 2$. This allows RNA-align's DP subroutine to run approximately 1.5 times faster than the full Gotoh algorithm, with little difference in final alignment accuracy. After initial alignments are obtained, subsequent alignment-superposition iterations in RNA-align deploy a similar DP algorithm to derive new alignments from new superpositions, but with gap penalty $g = -0.6$ and alignment score $S_{ij} = \frac{1}{1+(d_i/d_0)^2}$.

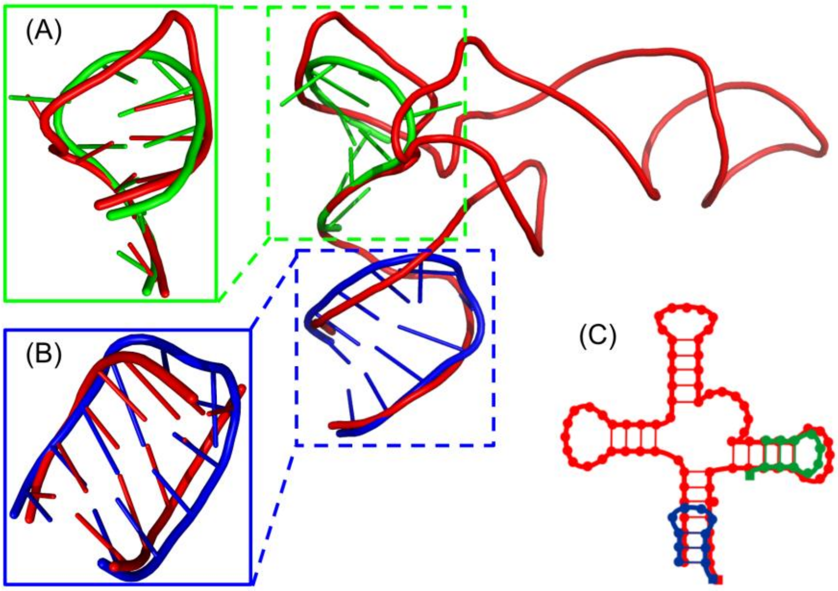
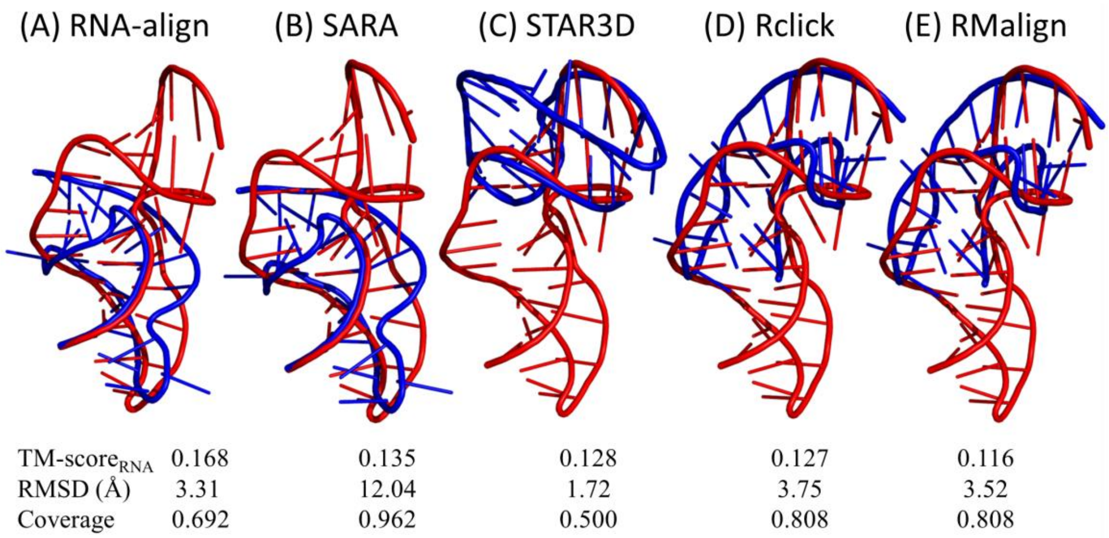


Fig. S3. RNA-align alignment of RNA structure from PDB ID 6enf Chain-x (red) to RNA structure from PDB ID 7msf Chain-S with (green, TM-score$_{\text{RNA}}$=0.378, inset A) and without (blue, TM-score$_{\text{RNA}}$=0.141, inset B) RNA secondary structure assignment. Only side chains of aligned helices and backbones are shown. (C) Diagram of secondary structure alignment.



TM-score$_{RNA}$	0.168	0.135	0.128	0.127	0.116
RMSD (Å)	3.31	12.04	1.72	3.75	3.52
Coverage	0.692	0.962	0.500	0.808	0.808

Fig. S4. Pseudoknotted RNA structure alignment guided by coarse-grain secondary structure assignment. Even though RNA-align's built-in SS assignment only depends on C3' atom distance and base type, it is accurate enough to ensure high quality final alignment: for the 235,641 random RNA pairs (Table S1), RNA-align using the current SS assignment for initial alignment achieve average TM-score$_{RNA}$ 0.21544, which is only 0.014% lower than that using the actual SS for initial alignment (average TM-score$_{RNA}$ 0.21547). This is in part because, different from other programs (Dror, et al., 2006; Ge and Zhang, 2015) that mainly relies on SS for initial alignment, RNA-align combines multiple initial alignments with iterative alignment-superposition procedure. This makes RNA-align less sensitive to minor SS assignment inaccuracy. For example, although RNA-align's SS assignment accuracies for two structures with pseudoknots, 2a43 Chain-A (blue) and 1rnk Chain-A (red) are only 92.3% and 94.1%, respectively, RNA-align generates alignment with TM-score$_{RNA}$ 0.168 (normalized by 2a43 Chain-A), which is higher than those from all four other programs for this RNA pair.

Supplementary Tables

Table S1. Average TM-score$_{RNA}$, average RMSD, average Coverage (number of aligned residues divided by query structure length) and total running time for the all-to-all alignment of 687 randomly selected RNA structures from PDB by RNA-align and five third-party programs, ranked in descending order of average TM-score$_{RNA}$. While RNA-align and RMalign can generate result for the full set of 235,641 RNA pairs, four third-party programs (SARA, STAR3D, ARTS, and Rclick) fail to generate alignments for some pairs. Therefore, we list the benchmark data in six blocks for the full set, the common set where RNA-align and one of the four third party programs can generate results, and the common set where all programs generate results. For the full set, if a program cannot generate result for a specific RNA pair, the TM-score$_{RNA}$ and Coverage for this pair is counted as zero for this program, while the RMSD is undefined (NA).

Dataset (number of pairs)	Program	TM-score$_{RNA}$	RMSD (Å)	Coverage	Time (hour)
Full set (235,641)	RNA-align	0.215	2.12	0.365	1.264
	RMalign	0.190	2.43	0.384	48.361
	SARA	0.153	NA	0.486	83.275
	Rclick	0.128	NA	0.272	84.063
	STAR3D	0.092	NA	0.197	113.300
	ARTS	0.071	NA	0.138	21.172
Common set between RNA-align and SARA (211,570)	RNA-align	0.216	2.26	0.370	1.135
	SARA	0.171	11.78	0.542	74.768
Common set between RNA-align and STAR3D (94,123)	RNA-align	0.258	3.05	0.446	0.503
	STAR3D	0.231	3.16	0.492	45.093
Common set between RNA-align and ARTS (81,249)	RNA-align	0.259	3.07	0.439	0.436
	ARTS	0.207	3.22	0.401	7.300
Common set between RNA-align and Rclick (235,626)	RNA-align	0.216	2.12	0.365	1.260
	Rclick	0.129	2.19	0.272	84.058
Common by all programs (76,067)	RNA-align	0.262	3.11	0.442	0.408
	RMalign	0.245	3.23	0.441	15.611
	STAR3D	0.212	3.21	0.480	36.758
	ARTS	0.211	3.23	0.401	6.835
	Rclick	0.202	3.31	0.360	27.140
	SARA	0.185	18.54	0.665	26.885

Supplementary Text

Text S1. Derivation of d_0 to scale TM-score$_{\text{RNA}}$

The normalization factor d_o in Eq. (2) is determined on 6,571,496 random RNA structure pairs with pairwise sequence identity less than 0.4. The raw TM-score$_{\text{RNA}}$ with a fix d_o=6 Å shows a power law dependence on length (TM-score$_{\text{RNA}}$= $2.08L^{-0.42}$, Fig. 1A black). Following the idea of TM-score (Zhang and Skolnick, 2004), d_o in TM-score$_{\text{RNA}}$ should have the general form of:

$$d_0 = a \cdot \sqrt[b]{L-c} - d \tag{S1}$$

Here, L is the number of nucleotides in the RNA, while a, b, c, and d are four parameters to be determined numerically. To make the power law dependency to a flat horizontal line (Fig. 1A), grid search is performed on the four parameters a, b, c, and d to minimize the following objective function:

$$|\beta| = \left| \frac{\sum_{n=1}^{N}(L_n - \bar{L}) \cdot (TM_n - \overline{TM})}{\sum_{n=1}^{N}(L_n - \bar{L})^2} \right| \tag{S2}$$

which is the absolute value for the slope of linear regression between L and TM-score$_{\text{RNA}}$. Here, N=6,571,496 is the number of RNA pairs. L_n and TM_n are the length and TM-score$_{\text{RNA}}$ for the shorter RNA in the *n*th RNA pair. $\bar{L}$ and $\overline{TM}$ are the average length and TM-score$_{\text{RNA}}$ on the whole dataset. In Eq. (S2), TM_n and $\overline{TM}$ depend on the choice of *a*, *b*, *c*, and *d*, which are finally determined as 0.6, 2, 0.5, and 2.5, respectively.

It should be noted that the d_0 and therefore the magnitude of TM-score$_{\text{RNA}}$ at a given length depend on the random data samples used for the parameterization. Here, the parameters for d_0 were determined by the random pairs of experimental RNA structures in the PDB. Since the experimental RNA structures for very short RNAs usually share similar hairpin folds, which should correspond to a random TM-score$_{\text{RNA}}$ value (~0.20) in the current protocol, the TM-score$_{\text{RNA}}$ value is particularly stringent (compared to the widely used RMSD) for the RNA pairs at the short length (e.g., $L < 30$). For example, an RNA structure pair of short length (e.g., $L = 20$) with a reasonable RMSD (e.g., 3 Å) may have a nearly random TM-score$_{\text{RNA}}$ value (~0.2) due to the fact that most of the randomly selected RNA structures at this length have the similar hairpin fold with a quite low RMSD. An alternative approach is to normalize the d_0 parameters using the RNA structures randomly generated (e.g., by self-avoided random walk), which will significantly increase the TM-score$_{\text{RNA}}$ value for random structure pairs of small RNAs. However, we believe that the current protocol to calculate d_0 based on experimental structures makes TM-score$_{\text{RNA}}$ more appropriately reflect the scale of RNA structure similarities in the PDB library.

Text S2. Calculation of posterior probability for a pair of RNAs belonging to the same Rfam family given the TM-score$_{\text{RNA}}$

963 RNA structures with non-identical sequences are collected from all 87 RNA families with at least one solved structure in Rfam database version 14.0. Each pair of RNAs can either belong to the same Rfam family (denoted as F) or different Rfam families (denoted as $\bar{F}$). RNAs from RF02540, RF02541 and RF02543 share high structure similarities and perform the same function as "large subunit ribosomal RNA". They are classified into three different families because they are from different domains of life (*Archaea*, *Bacteria*, and *Eukarya*, respectively). Due to their structure and function similarity, these three families are considered as the same family when calculating posterior probabilities. Similarly, the three Rfam families for "small subunit ribosomal RNA", RF01959, RF00177 and RF01960, are also considered as the same family.

The TM-score$_{\mathrm{RNA}}$, which is in the range of $(0,1]$, is discretized into 20 bins. The posterior probability of a structure pair belonging to the same family given the TM-score$_{\mathrm{RNA}}$ bin (TM) can be expressed by the Bayesian rule (Xu and Zhang, 2010):

$$P(F|TM) = \frac{P(TM|F) \cdot P(F)}{P(TM|F) \cdot P(F) + P(TM|\bar{F}) \cdot P(\bar{F})} \tag{S3}$$

Here, $P(F)$ and $P(\bar{F})$ are the prior probabilities of a random RNA pair belonging to the same and different families, respectively. These two prior probabilities can be calculated by

$$\begin{cases} P(F) = \dfrac{N_F}{N_F + N_{\bar{F}}} \\ P(\bar{F}) = \dfrac{N_{\bar{F}}}{N_F + N_{\bar{F}}} \end{cases} \tag{S4}$$

where N_F and $N_{\bar{F}}$ are the total number of the same- and different-family pairs, respectively

Meanwhile, the two conditional probabilities in Eq. (S3), $P(TM|F)$ and $P(TM|\bar{F})$, are the probabilities of observing the TM-score$_{\mathrm{RNA}}$, given that the pair belongs to the same and different families, respectively. These two probabilities can be expressed as

$$\begin{cases} P(TM|F) = \dfrac{N_F(TM)}{N_F} \\ P(TM|\bar{F}) = \dfrac{N_{\bar{F}}(TM)}{N_{\bar{F}}} \end{cases} \tag{S5}$$

where $N_F(TM)$ and $N_{\bar{F}}(TM)$ are the numbers of same- and different-family pairs in the given TM-score$_{\mathrm{RNA}}$ bin, respectively.

Using Eq. (S4) and Eq. (S5), Eq. (S3) can be simplified into:

$$\begin{aligned} P(F|TM) &= \frac{\dfrac{N_F(TM)}{N_F} \cdot \dfrac{N_F}{N_F + N_{\bar{F}}}}{\dfrac{N_F(TM)}{N_F} \cdot \dfrac{N_F}{N_F + N_{\bar{F}}} + \dfrac{N_{\bar{F}}(TM)}{N_{\bar{F}}} \cdot \dfrac{N_{\bar{F}}}{N_F + N_{\bar{F}}}} \\ &= \frac{N_F(TM)}{N_F(TM) + N_{\bar{F}}(TM)} \end{aligned} \tag{S6}$$

Since Eq. (S6) is easier to calculate than Eq. (S3) while being mathematically equivalent, we use Eq. (S6) to generate the data in Fig. 1B. Similarly, the probability of a RNA pair belonging to the same family, given the TM-score$_{\mathrm{RNA}}$ bin (TM), is computed by:

$$P(\bar{F}|TM) = \frac{N_{\bar{F}}(TM)}{N_F(TM) + N_{\bar{F}}(TM)} \tag{S7}$$

Reference

Dror, O., Nussinov, R. and Wolfson, H.J. The ARTS web server for aligning RNA tertiary structures. *Nucleic Acids Res* 2006;34:W412-W415.

Ge, P. and Zhang, S.J. STAR3D: a stack-based RNA 3D structural alignment tool. *Nucleic Acids Res* 2015;43(20).

Gotoh, O. An improved algorithm for matching biological sequences. *Journal of molecular biology* 1982;162(3):705-708.

Xu, J.R. and Zhang, Y. How significant is a protein structure similarity with TM-score=0.5? *Bioinformatics* 2010;26(7):889-895.

Zhang, Y. and Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* 2004;57(4):702-710.