rMSA: A Sequence Search and Alignment Algorithm to Improve RNA structure modeling

Supplementary Information

Table of Content

Supplementary Text

Supplementary Figures

Supplementary Tables

References

Supplementary Text

Text S1. Number of effective sequence (*Nf*)

The number of effective sequence is calculated as in our previous work [1].

$$Nf = \frac{1}{\sqrt{L}} \sum_{n=1}^{N} \frac{1}{1 + \sum_{m=1,m\neq n}^{N} I\left[S_{m,n} \geq S_{cut}\right]} \quad \ldots (S1)$$

In Equation S1, L is the length of target RNA; N is the total number of sequences in the MSA; m and n are the index of sequences in the MSA; $S_{m,n}$ is the sequence identity between the m-th and n-th sequences. When calculating the sequence identity, a gap is treated as an additional nucleotide type. S_{cut} is the sequence identity cutoff above which two sequences are considered redundant to each other. We used S_{cut}=80%, as described in a previous study [2]. I[ ] is an Iverson bracket operator, which equals to 1 if $S_{m,n} \geq S_{cut}$, or 0 otherwise. Mathematically, Nf is the number of non-redundant sequences in the MSA divided by square root of sequence length [3].

Text S2. Structure Conservation Index (*SCI*)

In addition to the MSAscore described by main text Equation 2, we also tested the usage of Nf (main text Equation 1) and the Structure Conservation Index (*SCI*) to select the final MSA from the set of 5 MSAs generated by rMSA for each target RNA. *SCI* was introduced by the work of RNAz [4], and is defined as:

$$SCI = \frac{E_{MSA}}{\frac{1}{N}\sum_{n=1}^{N} E_n} \quad \ldots \text{(S2)}$$

Here, N is the total number of sequences in the MSA, and E_n is the score of RNA secondary structure (rSS) predicted by minimal free energy method for the n-th sequence in the MSA, E_{MSA} is the score for the consensus rSS predicted by RNAalifold. A close to zero *SCI* indicates that RNAalifold is not able to find a consensus structure; a set of perfectly conserved structures has *SCI* close to 1. *SCI* may be slightly greater than 1, which means that the rSS is not only conserved across different sequences of the MSA, but also supported by consistent mutations, which contribute a covariance score to E_{MSA}.

Text S3. Evaluation metrics for rSS assignment

To evaluate the accuracy of base pair prediction by covariance programs, predicted pairs are ranked in descending order of covariance scores from PLMC and R-scape or the base pairing score reported by option -r of PETfold and option -p of RNAalifold. The top *Ln* base pairs are considered, where *Ln* is the number of base pairs in the experiment structure. The accuracy of rSS prediction for each RNA can then be quantified by F1-score and Mathews Correlation Coefficient (MCC):

$$F1 = \frac{2}{\frac{1}{PPV} + \frac{1}{TPR}} \quad \ldots \text{(S3)}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad \ldots \text{(S4)}$$

In the above equations, *TP* is the number of True Positive base pairs correctly predicted; *FP* is the number of False Positive base pairs predicted by covariance program and not in the experimental structure; *FN* is the number of False Negative base pairs in the experimental structure that are missed by rSS prediction; *TN* is the number of True Negative nucleotide pairs that are neither predicted nor in the experimental structure. *PPV* (Positive Predictive Value) and *TPR* (True Positive Rate) are the precision and sensitivity of prediction:

$$PPV = \frac{TP}{TP+FP} \quad \ldots \text{(S5)}$$

$$TPR = \frac{TP}{TP+FN} \quad \ldots \text{(S6)}$$

Since the number of predicted base pairs that we consider (*TP*+*FP*) is the same as the number of experimental base pairs, i.e., *Ln*=*TP*+*FN*, the above equations can be further simplified into:

$$F1 = PPV = TPR = \frac{TP}{Ln} \quad \ldots \text{(S7)}$$

$$MCC = \frac{TP{\cdot}TN-(Ln-TP)^2}{Ln(TN+Ln-TP)} \quad \ldots \text{(S8)}$$

The rSS in RNA refers to canonical base pairings, i.e., Watson-Crick and G:U Wobble base pairs. It is therefore equivalent to a contact map in protein and is different from protein secondary structure. Therefore, rSS prediction accuracy cannot be measured by metrics for protein secondary structure prediction such as Q3 and SOV. Similar to previous studies on covariance analysis in proteins [5-7] and RNAs [2], our assessment of covariance algorithm is on a given number of top predictions (*Ln*) rather than on predictions with score above a given threshold (e.g., score>0.5). This is because, even for the same target RNA and the same covariance algorithm, the prediction scores from alignments with different depths are not comparable. For example, on our dataset, the average PLMC scores for top *Ln* predicted pairs is 1.426 for rMSA, which is consistently higher than the average PLMC scores of 1.023, 0.553, 0.059, 0.012 for RNAcmap, Infernal, nhmmer, and blastn, respectively.

Text S4. Filtering of MSA for PETfold and RNAalifold

Unlike PLMC and R-scape, which can handle very deep MSAs, PETfold often fails to complete the computation for MSAs with lots of sequences, especially when the sequences have too many gaps. Indeed, when testing on rMSA and RNAcmap alignments, half of the RNAs in our benchmark dataset cannot generate PETfold prediction and report "Segmentation fault" instead. Similarly, when presented with very deep MSAs, RNAalifold often cannot predict any base pairs and reports “backtracking failed in repeat; string and structure have unequal length” instead. Therefore, for PETfold and RNAalifold, we filter the MSAs to exclude sequences for which >5% of its positions are gaps. If the resulting MSAs have >1500 sequences, only the top 1500 sequences are retained for PETfold and RNAalifold. This aggressive filtering allows PETfold and RNAalifold to complete prediction for 97.8% and 100% of the benchmark RNAs, respectively. For the remaining 8 RNAs whose alignments from rMSA (but not from other programs) cannot complete PETfold computation, their rMSA alignments are further reduced to the top 1000 sequences. This filtering of MSAs is an inevitable step to ensure that PETfold and RNAalifold can run through. It is not meant to improve the quality of MSAs. In fact, the F1-score of rSS prediction by PLMC, R-scape and R-scape --RAFSp dropped from 0.648, 0.575, and 0.561, respectively, for the original rMSA alignments to 0.500, 0.507, and 0.540 for filtered rMSA alignments.

Supplementary Figures

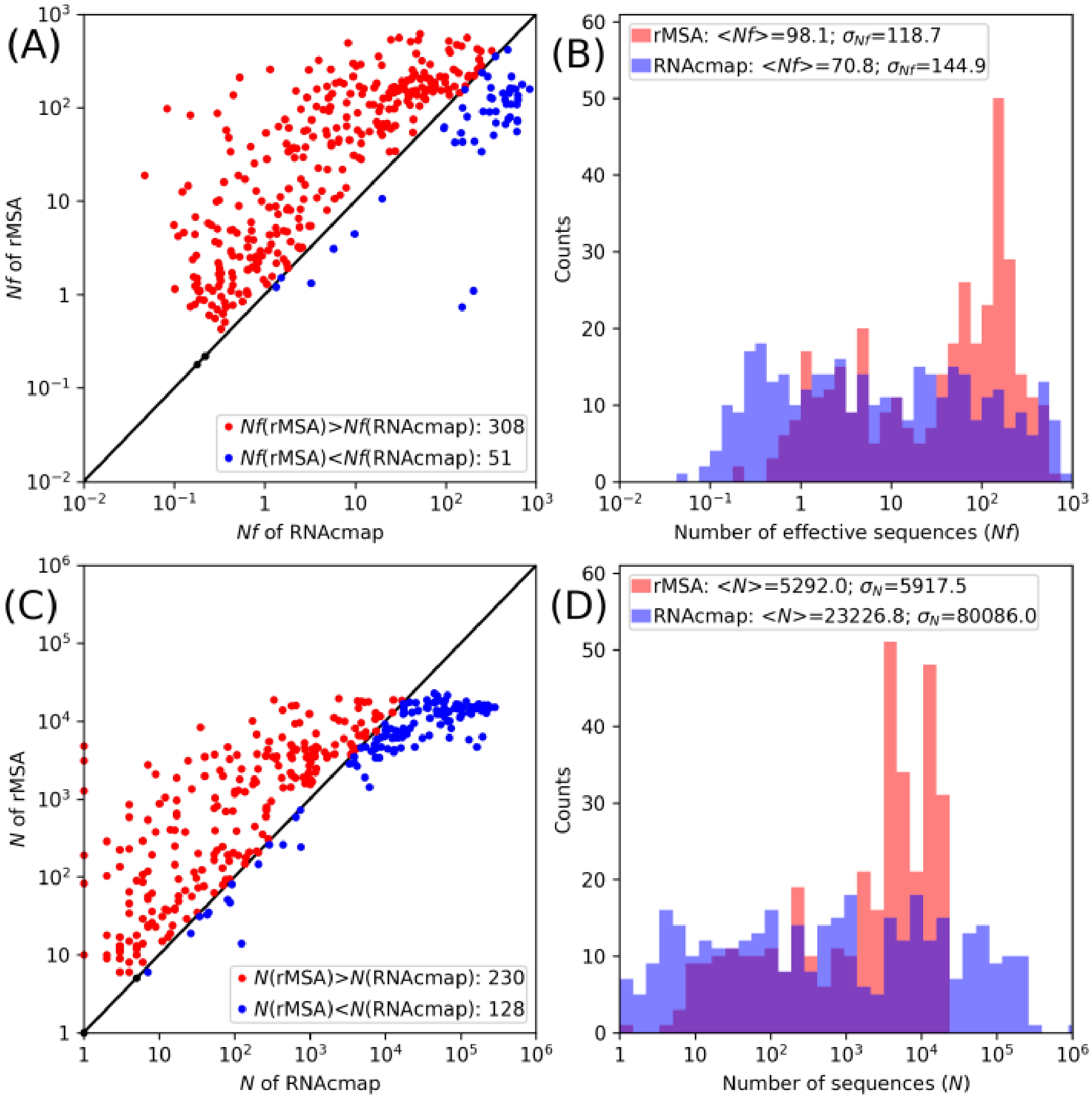

Figure S1. Comparison of alignment depths in $\log_{10}$ scale between rMSA and RNAcmap. **(A, C)** Head-to-head comparison of *Nf* **(A)** or *N* **(C)** of rMSA (y-axis) and that of RNAcmap (x-axis) for each target RNA. Figure legend shows the number of RNAs above (red) or below (blue) the diagonal line; RNAs on the diagonal line is in black. **(B, D)** Distribution of *Nf* **(B)** or *N* **(D)** for rMSA (red) and RNAcmap (blue). Figure legend shows the mean and standard deviations of *Nf* or *N*. Among the 51 RNAs whose rMSA *Nf* is lower than RNAcmap *Nf*, 34 and 8 are tRNAs and ribosomal large subunit RNAs, respectively.

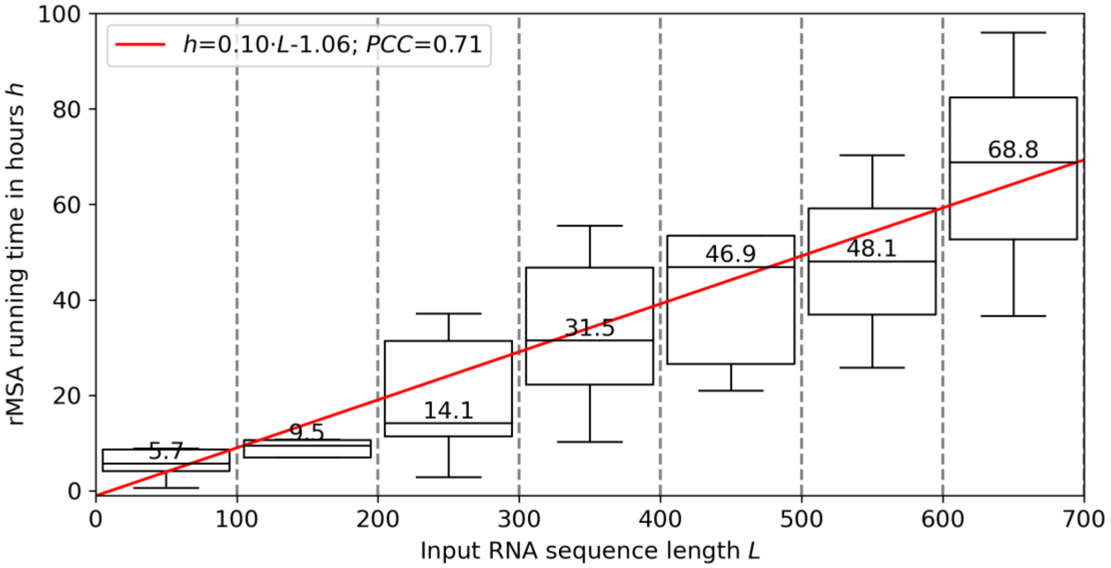

Figure S2. rMSA running time versus sequence length. The length of boxplot whiskers equals to the Interquartile range. Numbers on the top of the figures are the values for the least square linear fit (red line) and the Pearson's Correlation Coefficient (PCC).

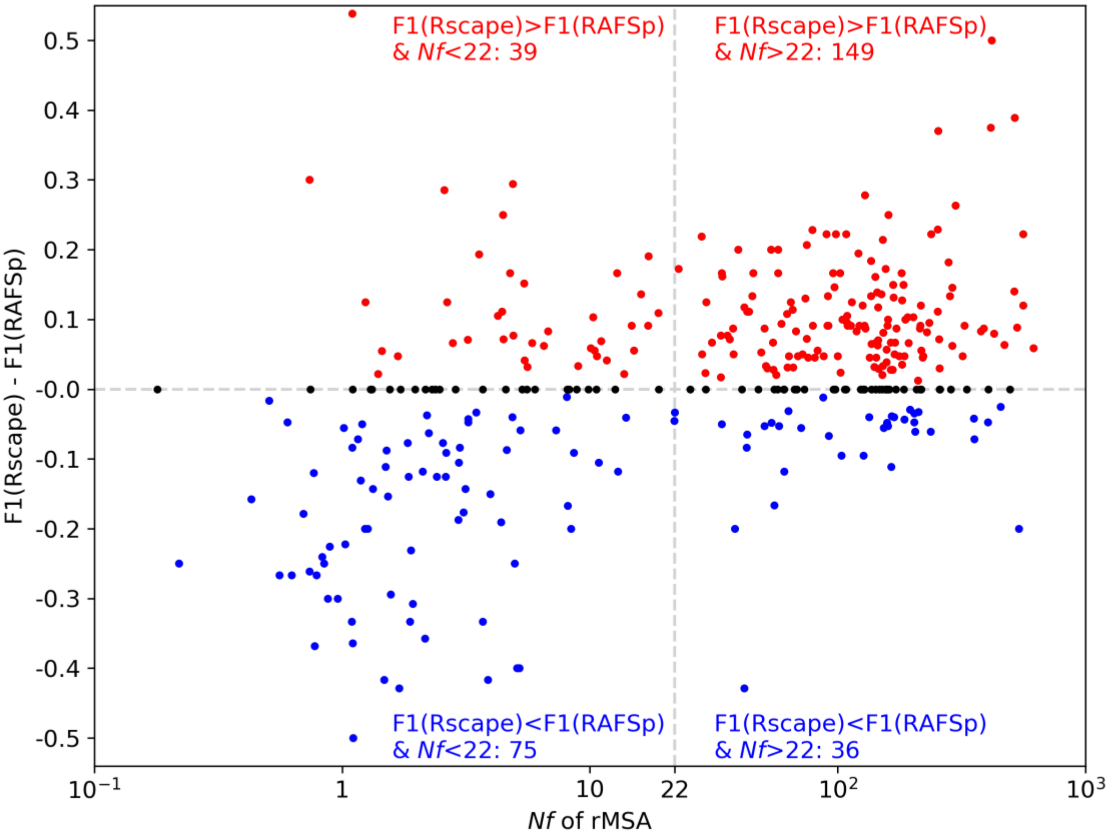

Figure S3. Scatter plot for F1-score of rSS prediction by R-scape using the default GTp statistics, denoted as F1(Rscape), minus that of R-scape --RAFSp, denoted as F1(RAFSp), versus the *Nf* of rMSA alignments for each target. The numbers in the plot show the number of targets in each of the four quadrants divided by F1(Rscape)-F1(RAFSp)=0 (horizontal grey dash line) and by *Nf*=22 (vertical grey dash line).

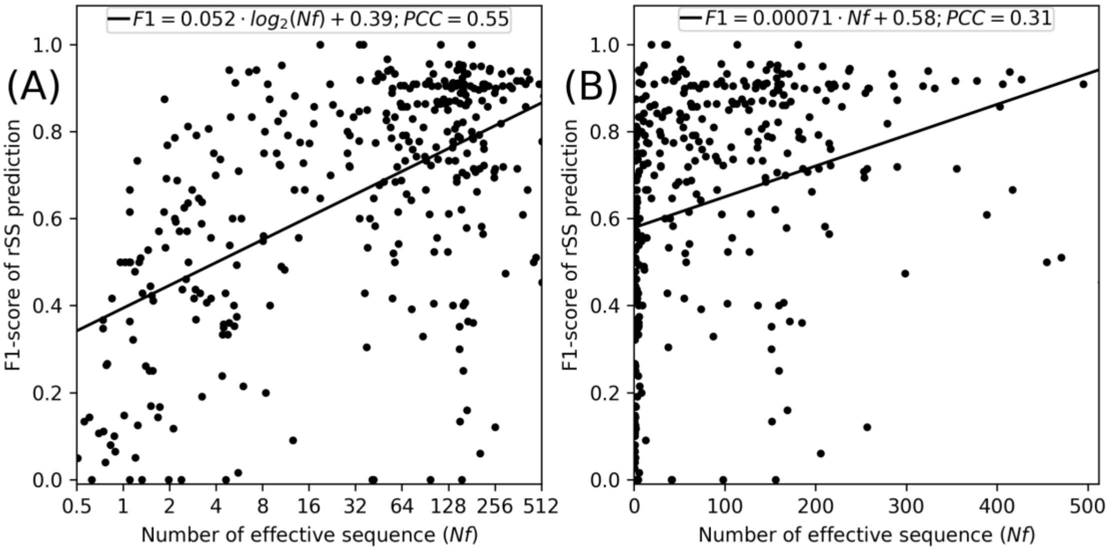

Figure S4. Scatter plot for F1-score of rSS prediction by PLMC versus rMSA *Nf* in $\log_2$ **(A)** and linear scale **(B)** for each target RNA. Numbers on the top of the figures are the values for the least square linear fit (black line) and the Pearson's Correlation Coefficient (PCC).

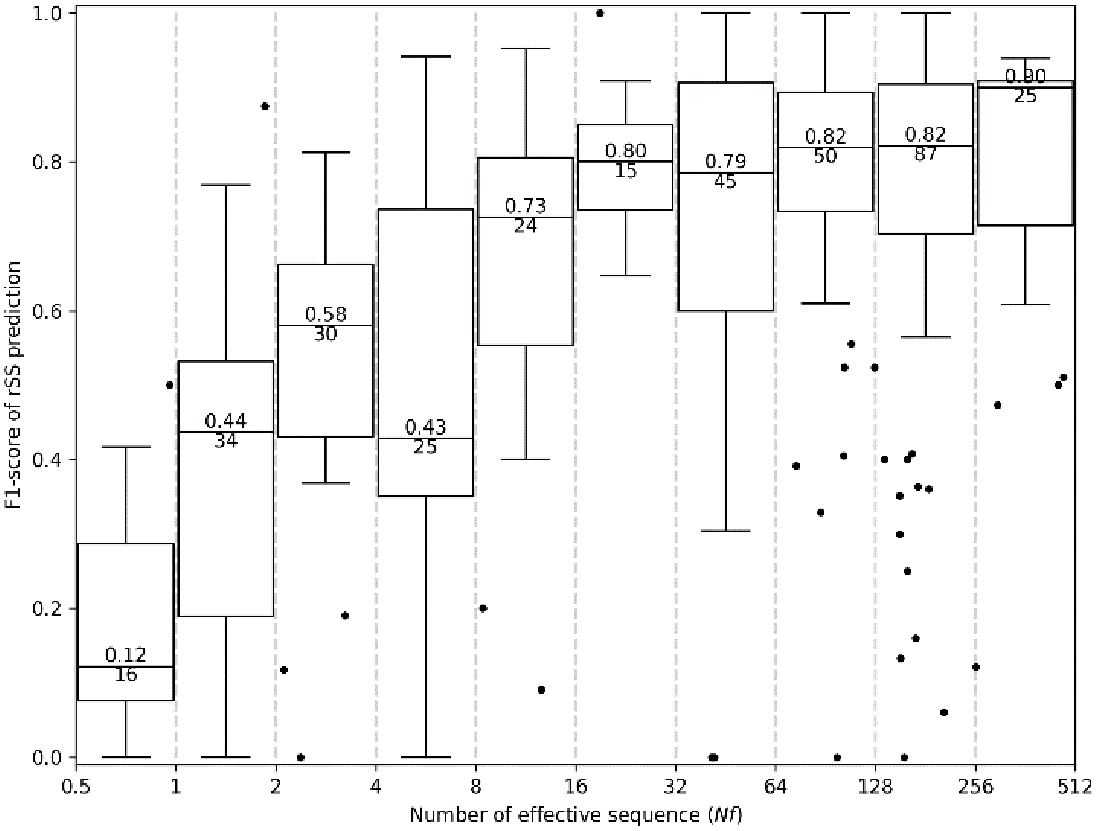

Figure S5. F1-score of rSS prediction by PLMC (y-axis) versus *Nf* of rMSA alignments (x-axis). The length of boxplot whiskers equals to the Interquartile range. Outliers are plotted as black dots. The values inside each box are the median F1-scores (above) and the number of rMSA alignments (below) for each *Nf* bin.

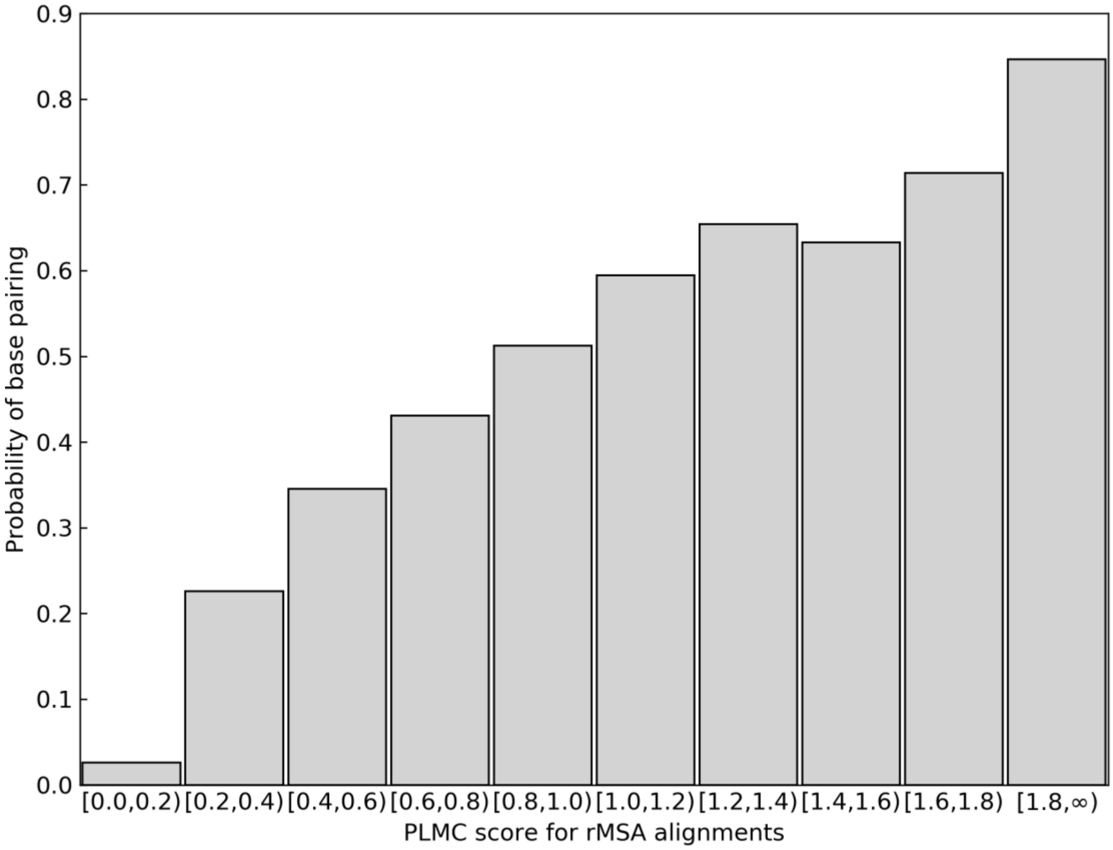

Figure S6. Probabilities of base pairing (*y*-axis) for pairs of nucleotides within different PLMC score bins (*x*-axis) for rMSA alignments of 361 target RNAs. The Pearson and Spearman correlation coefficients between base pairing probabilities and PLMC score are 0.969 (p-value=9.31E-8) and 0.988 (p-value=9.31E-8), respectively.

Supplementary Tables

Table S1. List of RNA structures in the benchmark dataset. Targets are ranked in descending order of length, which equals to the number of nucleotides with coordinates in the experimental structure. While this list may include several RNAs for the same RNA family, e.g., tRNA, all RNAs within the same family or among different families are all non-redundant with sequence identity <80%.

#	PDB	Chain	Length	Name
1	5g2x	A	692	Group IIA Intron
2	6c4h	A	636	23S rRNA
3	6hiw	CA	621	9S rRNA
4	6chr	A	621	Group IIB intron lariat
5	6hix	AA	591	12S rRNA
6	6n7r	R	558	U1 snRNA
7	1fg0	A	496	23S rRNA
8	5wlc	L0	488	5’ external transcribed spacer (5' ETS)
9	5zwn	P	480	U1 snRNA
10	6swe	2	460	16S rRNA
11	5j01	A	414	Group IIC intron lariat
12	6agb	A	369	Ribonuclease P RNA
13	6q9a	4	363	tmRNA
14	3bwp	A	356	Group IIC intron
15	3q1q	B	347	RNase P RNA
16	6ahu	A	341	H1 RNA (the RNA component of RNase P)
17	6w2s	0	339	CrPV 5'-UTR IRES
18	6q95	4	319	transfer-messenger RNA (tmRNA)
19	5u4j	a	307	16S rRNA
20	2a64	A	298	ribonuclease P RNA
21	6g90	1	293	U1 snRNA, U1 snRNA, U1 snRNA, U1 snRNA, U1 snRNA
22	6sxo	L5	268	28S ribosomal RNA including ES27L-B (2839-3265)
23	5flx	z	264	HCV-IRES
24	6frk	1	258	7SL RNA, cytoplasmic 1 (RN7SL1), SRP RNA
25	5a8l	A	253	28S rRNA
26	1x8w	B	247	Group I Intron RNA
27	5tc1	R	242	phage MS2 genome
28	5oql	2	230	U3 snoRNA
29	1y0q	A	229	Group I Intron ribozyme
30	6j6q	L	210	U2 snRNA
31	6p5n	1	207	IAPV-IRES
32	5zwm	H	206	U2 snRNA
33	5jup	EC	198	IRES
34	1u6b	B	195	Self-Splicing Group I Intron with Both Exons
35	6d90	4	194	CrPV-IRES
36	4gma	Z	192	Adenosylcobalamin riboswitch
37	6gyv	A	190	Lariat-capping ribozyme
38	5v3i	A	186	VS Ribozyme RNA
39	3jcs	3	184	26S gamma rRNA
40	6az3	4	183	rRNA delta
41	6j6g	D	179	U5 snRNA
42	3dil	A	173	Lysine riboswitch
43	5lj3	Z	171	U2 snRNA (small nuclear RNA)
44	5t2a	E	169	srRNA1

#	PDB	Chain	Length	Name
45	3p49	A	169	Glycine Riboswitch
46	5wlc	L2	169	U3 snoRNA
47	6ufh	A	165	ileS T-box
48	6qx9	1	164	U1 snRNA
49	6az3	7	163	rRNA 5.8S
50	6ft6	2	162	7S rRNA
51	3pdr	A	161	M-box Riboswitch RNA
52	4gxy	A	161	Adenosylcobalamin riboswitch
53	6ole	D	157	5.8S rRNA
54	5xxb	4	157	5.8S RNA
55	1u9s	A	155	Ribonuclease P
56	5xy3	4	154	5.8S rRNA
57	4v8p	B2	154	5.8S rRNA
58	3j79	C	151	5.8S rRNA
59	1nbs	B	150	Ribonuclease P RNA
60	6jdv	B	143	sgRNA
61	6id0	H	140	U2snRNA
62	3hhn	C	136	Class I ligase ribozyme, self-ligation product
63	3ndb	M	136	SRP RNA
64	6exn	2	136	U2 snRNA
65	6ahd	I	136	U4snRNA
66	3l3c	S	133	GLMS Ribozyme
67	4uyk	R	133	SRP RNA
68	2il9	A	128	Ribosomal binding domain of IRES RNA
69	6n5q	A	127	pir-miRNA-378a apical loop and one-base-pair fused to YdaO riboswitch
70	3v7e	C	126	SAM-I riboswitch aptamer with an engineered helix P3
71	5gan	V	124	U4 snRNA
72	2z75	B	123	glmS ribozyme RNA
73	5zlu	X	123	5S rRNA
74	1vq8	9	122	5S rRNA
75	5dm6	Y	122	5S rRNA
76	6t4q	C4	121	5S rRNA
77	5mmi	B	121	5S rRNA
78	4ybb	DB	120	5S rRNA
79	6ek0	L7	120	5S rRNA
80	1nbs	A	120	Ribonuclease P RNA
81	6qdv	2	120	U2 snRNA
82	4v8p	B3	120	5S RRNA
83	3jcs	8	119	5S ribosomal RNA
84	4kqy	A	119	YitJ S box/SAM-I riboswitch
85	6ues	A	119	Apo SAM-IV Riboswitch
86	5xxb	3	118	5S RNA
87	3j79	B	118	5S rRNA
88	5o60	B	118	5S rRNA
89	6rm3	L70	118	5S rRNA
90	6xyw	3	118	Plant mitochondrial rRNA
91	4qln	A	117	ydaO riboswitch
92	5xy3	3	117	5S rRNA
93	4qk9	A	116	C-di-AMP riboswitch
94	6s0x	B	115	5S rRNA
95	5wti	B	115	RNA
96	6v3a	B	115	5s rRNA
97	5f9r	A	115	sgRNA
98	6ahd	B	115	U5snRNA

#	PDB	Chain	Length	Name
99	5t5h	D	114	5S rRNA
100	6ha1	B	112	5S rRNA
101	5vt0	R	112	6S RNA derivative
102	6ol3	C	111	Adenovirus Virus-Associated (VA) RNA I apical and central domains
103	3jb9	P	111	U2 snRNA
104	6ny2	B	110	RNA
105	6ck5	A	108	PRPP riboswitch
106	3f2q	X	107	FMN riboswitch
107	4y1m	B	107	yybP-ykoY riboswitch
108	6dlr	A	106	PRPP Riboswitch
109	6eri	AB	106	4.5S rRNA
110	4wfl	A	105	Bacterial SRP Alu domain
111	3jb9	C	105	U5 snRNA
112	6j6g	E	103	U6 snRNA
113	4lck	C	102	T-box riboswitch stem I
114	4frn	A	102	Cobalamin riboswitch aptamer domain
115	2xxa	F	102	4.5S RNA
116	6mj0	A	101	Turnip yellow mosaic virus 3'UTR
117	4w90	C	101	Riboswitch a pseudo-dimeric RNA
118	5u33	B	101	sgRNA
119	6dmc	A	100	ppGpp Riboswitch
120	4y1j	A	100	yybP-ykoY riboswitch
121	3suh	X	100	Riboswitch
122	4rzd	A	98	PreQ1-III Riboswitch (Class 3)
123	6n2v	A	98	Mn riboswitch optimized construct
124	6az3	5	97	rRNA epsilon
125	6id1	F	97	U6snRNA
126	4oqu	A	97	SAM-I/IV riboswitch
127	6jxm	B	97	RNA
128	6joo	B	96	Guide RNA
129	3ktw	C	96	SRP RNA
130	6dvk	H	95	Computationally designed RNA
131	2ygh	A	95	SAM-I riboswitch
132	5ml7	A	95	23S ribosomal RNA
133	5b2o	B	94	Guide RNA
134	5x2h	B	93	sgRNA
135	3iwn	A	93	C-di-GMP riboswitch
136	1m5o	B	92	RNA Hairpin Ribozyme
137	4rum	A	92	NiCo riboswitch RNA
138	3cul	D	92	Aminoacyl-tRNA synthetase ribozyme
139	6mwn	A	91	Hepatitis A virus IRES domain V
140	5t83	A	90	Guanidine-I riboswitch
141	4lvw	A	89	THF riboswitch
142	5osg	2	88	18S rRNA
143	1wz2	C	88	tRNA
144	3owi	B	87	Domain II of glycine riboswitch
145	4v8b	AB	87	tRNA-LEU
146	3k0j	E	87	ThiM riboswitch
147	6c0f	6	87	ITS2
148	5aox	F	86	Alu Jo consensus RNA
149	3q1q	C	86	tRNA-PHE
150	4v8d	AB	85	tRNA-TYR
151	4mgn	A	85	glyQS T box riboswitch
152	5u3g	B	84	ykkC riboswitch

#	PDB	Chain	Length	Name
153	6b14	R	83	RNA aptamer
154	4yaz	A	83	3',3'-cGAMP riboswitch
155	6jq5	A	82	Hatchet Ribozyme
156	3w3s	B	82	selenocysteine tRNA
157	3am1	B	81	ASL-truncated tRNA
158	5lzd	y	81	Sec-tRNASec
159	3umy	B	80	23S rRNA
160	6ah3	T	80	pre-tRNA
161	3adb	C	79	selenocysteine tRNA
162	6cu1	A	79	YrlA effector-binding module
163	5on2	B	79	tRNA-LEU
164	3amt	B	78	tRNA-ILE
165	6rja	D	78	sgRNA
166	4v8n	AW	78	A-site tRNA ILE2 Agmatidine
167	6v5b	D	78	Pri-miR-16-2
168	5vpp	QV	78	P-site tRNA SufA6
169	6ugg	A	77	tRNA-ASP
170	3d2v	A	77	TPP-specific riboswitch
171	5ccb	N	77	tRNA3Lys
172	6svs	A	77	U:A-U-rich RNA triple helix with 11 consecutive base triples
173	5j8b	x	77	P-site tRNA
174	2zzm	B	77	tRNA-LEU
175	6ufm	A	77	RNA
176	6v3a	v	77	tRNA-MET
177	1h3e	B	77	tRNA-TYR(GUA)
178	6i0y	V	77	tRNA-PRO
179	1j2b	C	77	tRNA-VAL
180	5o2r	x	77	P-site tRNA-ILE
181	3a2k	C	77	bacterial tRNA
182	6q97	7	77	tRNA-VAL
183	5ah5	D	76	tRNA-LEU TAA isoacceptor
184	4jf2	A	76	PreQ1-II Riboswitch
185	6t4q	6	76	ICG tRNA Arg (P/P)
186	6az1	2	76	tRNA-PHE
187	1qf6	B	76	tRNA-THR
188	6r5q	2	76	P-tRNA
189	2iy5	T	76	tRNA-PHE
190	4v5l	AY	76	A-site tRNA G24A tRNA-TRP
191	5el6	3K	76	tRNA-LYS
192	6t7t	6	76	tRNA
193	4wj4	B	76	76mer-tRNA
194	5lzs	ii	76	tRNA
195	6r87	B	76	tRNA-ALA
196	4v5g	AY	76	A-site tRNA-THR
197	6q9a	7	76	P-site tRNA
198	2zue	B	75	tRNA-ARG
199	1g59	B	75	tRNA-GLU
200	4rdx	C	75	tRNA-HIS
201	5x6b	P	75	tRNA-CYS
202	1il2	C	75	Aspartyl Transfer RNA
203	6t4q	7	75	tRNA (E/E)
204	2csx	C	75	tRNA-MET
205	6tbv	PTR1	75	P-site tRNA-ARG
206	6tb3	n	75	tRNA

#	PDB	Chain	Length	Name
207	6qdw	v	75	tRNA-GLY
208	1gax	C	75	tRNA-VAL
209	6ek0	S6	75	tRNA-MET
210	2dr2	B	75	tRNA-TRP
211	4v7l	AY	75	tRNA-GLN
212	4lck	B	75	tRNA-GLY
213	3wqy	C	75	tRNA-ALA
214	4v9i	AY	75	A-site tRNA
215	6ip5	zu	75	P-site tRNA
216	1j1u	B	74	tRNA-TYR
217	3rg5	A	74	tRNA-SEC
218	1u0b	A	74	tRNA-CYS
219	4yye	C	74	tRNA
220	4prf	B	74	Hepatitis Delta virus ribozyme
221	3q3z	A	74	c-di-GMP-II riboswitch
222	5d8h	A	74	23S ribosomal RNA
223	3akz	E	74	tRNA-GLN
224	2d6f	F	74	tRNA
225	3j7a	7	74	tRNA
226	2dlc	Y	73	tRNA
227	5czz	B	73	RNA
228	6cae	1y	73	A-site and E-site tRNAs
229	5e6m	C	73	tRNA-GLY
230	4znp	A	73	pfI riboswitch
231	2der	C	73	tRNA
232	5wt3	C	73	tRNA
233	5ud5	D	72	tRNA-PYL
234	5wwr	C	72	tRNA
235	2zni	C	72	Bacterial tRNA
236	6lxd	D	72	Pri-miRNA
237	3la5	A	71	Adenosine Riboswitch
238	5kpy	A	71	5-hydroxytryptophan RNA aptamer
239	6az3	6	71	rRNA zeta
240	2du3	D	71	tRNA
241	2oiu	P	71	L1 Ribozyme RNA Ligase
242	5tpy	A	71	Exonuclease resistant RNA
243	6gaz	AV	71	P-site fMet-tRNAMet, mitochondrial
244	1kxk	A	70	ai5g group II Self-splicing intron
245	5ob3	A	69	RNA aptamer
246	3ivn	A	69	A-riboswitch
247	3eph	E	69	tRNA
248	6u8d	A	68	JIIIabc RNA
249	2qus	A	68	Hammerhead ribozyme
250	4pqv	A	68	XRN1-resistant flaviviral RNA
251	6p2h	A	68	2'-dG-II class of riboswitches
252	4wzj	V	68	U4 small nuclear RNA variant
253	4fe5	B	67	xpt-pbuX guanine riboswitch aptamer domain
254	3ski	B	67	2'-Deoxyguanosine riboswitch
255	6gaw	BB	67	tRNA-PHE, mitochondrial
256	5a8l	B	67	18S rRNA
257	1h4s	T	66	tRNA-PRO (CGG)
258	3r4f	A	66	pRNA
259	3egz	B	65	Tetracycline aptamer and artificial riboswitch
260	4xwf	A	64	pfl RNA

#	PDB	Chain	Length	Name
261	3nkb	B	64	Hepatitis delta virus ribozyme
262	5b63	D	64	tRNA-ARG
263	6rfl	U	63	chr17.trna16-GlnTTG
264	5t5a	A	62	Twister Sister (TS) Ribozyme
265	5btp	A	62	ZTP riboswitch
266	2czj	B	62	tmRNA
267	3icq	D	62	tRNA
268	2hvy	E	61	H/ACA RNA from RNA pseudouridine synthases
269	5ddp	A	61	L-glutamine riboswitch
270	1un6	E	61	5S rRNA
271	6db8	R	60	DIR2s RNA aptamer
272	3rw6	F	60	Constitutive transport element of Mason-Pfizer monkey virus RNA
273	6pmo	A	60	T-box riboswitch discriminator
274	4m4o	B	59	Aptamer minE
275	3lwr	D	58	H/ACA RNA
276	4v2s	Q	58	Bacterial small RNAs (sRNAs) rydC
277	1dk1	B	57	rRNA fragment
278	3j7y	B	57	mt-tRNA-VAL
279	4rge	A	56	env22 twister ribozyme
280	1ser	T	56	tRNA-SER
281	2nre	F	56	tRNA-LEU
282	5lyu	A	55	7SK RNA
283	4k27	U	55	Myotonic Dystrophy Type 2 RNA
284	1mzp	B	55	fragment of 23S rRNA
285	6lax	A	55	SAM-VI riboswitch
286	5uq8	x	55	mRNA
287	4pkd	V	54	U1 snRNA stem-loops 1 and 2
288	2fk6	R	53	tRNA-THR
289	3e5c	A	52	SMK box (SAM-III) Riboswitch
290	4enc	A	52	Fluoride riboswitch
291	4oji	A	52	Twister Ribozyme
292	5tf6	D	52	U6 snRNA
293	6aay	B	52	crRNA for Cas13b
294	2qwy	A	52	SAM-II riboswitch
295	6iv8	B	51	crRNA for Cas13d
296	3npq	A	51	S-Adenosylhomocysteine Riboswitch
297	5y7m	B	51	RNA fragments containing a K-turn motif
298	5xwy	B	51	crRNA for Cas13a
299	6e9e	B	51	crRNA
300	5y85	B	50	Four-way junctional Twister-Sister ribozyme
301	6qn3	A	50	Glutamine II Riboswitch
302	6ufj	A	50	Pistol ribozyme product
303	6r47	A	50	Pistol ribozyme
304	2pxd	B	49	4.5 S RNA
305	1u63	B	49	Fragment of mRNA for L1
306	6fz0	A	48	metY SAM V
307	4c7o	E	48	SRP RNA
308	5dqk	A	48	Hammerhead ribozyme
309	5ztm	C	48	Non-coding mRNA sequence roX2
310	5k7d	A	47	Pistol ribozyme
311	1s03	A	47	spc Operon mRNA
312	4o26	E	47	Telomerase TR
313	4qjh	B	47	Twister Ribozyme
314	5xtm	B	46	RNA fragment containing a K-turn motif

#	PDB	Chain	Length	Name
315	1xjr	A	46	s2m RNA
316	3iab	R	46	P3 domain of the RNA component of RNase MRP
317	2nue	C	46	RNA
318	4qjd	B	46	Twister RNA sequence
319	6d3p	A	45	Exoribonuclease-resistant RNA from Sweet clover necrotic mosaic virus
320	1p6v	B	45	tRNA domain of transfer-messenger RNA
321	4rmo	B	44	Antitoxin for CptIN Type III Toxin
322	3zp8	A	42	Hammerhead Ribozyme, Enzyme Strand
323	5m0h	A	42	ASH1 E3 (42 nt-TL/TLR)
324	5nwq	A	41	Guanidine III riboswitch
325	4pmi	A	40	Rev-Response-Element RNA
326	5kk5	B	40	crRNA
327	3p22	A	39	Core ENE hairpin from Kaposi's sarcoma-associated herpesviru PAN RNA
328	4kr6	C	39	Truncated tRNA
329	6d12	C	38	7SK RNA stem-loop 4
330	6e8s	A	38	iMango-III aptamer
331	1zho	B	38	mRNA
332	4pdb	I	38	SELEX RNA aptamer
333	1f1t	A	38	Malachite Green Aptamer RNA
334	1yls	B	38	RNA Diels-Alder ribozyme
335	6dtd	C	37	crRNA
336	1i6u	C	37	16s rRNA fragment
337	6sy4	C	37	TetR-binding aptamer K1
338	1kog	I	37	Threonyl-tRNA synthetase mRNA
339	5bjo	E	36	RNA aptamer
340	3gs5	C	36	RNA
341	6c65	A	36	Mango-II-A22U Fluorescent Aptamer
342	4x4p	B	36	G70A tRNA minihelix ending in CCAC
343	3ovb	D	35	tRNA mimic
344	6dcb	B	35	7SK RNA stem-loop 1 proximal
345	1et4	A	35	RNA Aptamer
346	2xdb	G	35	TOXI
347	5dea	A	35	sc1
348	6cf2	G	35	RNA aptamer
349	4v83	AV	35	domain 3 of PSIC IGR IRES RNA
350	2zh3	B	34	tRNA
351	4oog	D	34	RNase III cleavage product
352	4ato	G	33	TOXI
353	3gca	A	33	PreQ1 riboswitch
354	3fu2	B	33	PreQ1 riboswitch
355	5v3f	A	31	Fluorogenic RNA Mango
356	6h0r	A	31	SRS2 fragment of Rgs4 3' UTR
357	5voe	A	31	Aptamer 11F7t
358	1jbr	D	31	SRD RNA analog
359	2b63	R	31	RNA inhibitor of RNA polymerase II
360	3snp	C	30	Ferritin H IRE RNA
361	5y58	X	30	TLC1

Table S2. Average rSS prediction accuracies by different MSA construction and thermodynamics-based rSS prediction schema

rSS predictor	MSA [§]	F1	P-value	MCC	P-value
PETfold	rMSA	0.717	*	0.715	*
	rMSA (*Nf*)	0.716	4.03E-1	0.714	4.00E-1
	rMSA (*SCI*)	0.719	6.66E-1	0.717	6.65E-1
	RNAcmap	0.736	9.96E-1	0.735	9.96E-1
	Infernal	0.736	9.95E-1	0.734	9.95E-1
	nhmmer	0.730	9.45E-1	0.728	9.47E-1
	blastn	0.686	**1.17E-3**	0.684	**1.24E-3**
	RNAlien	0.695	**7.57E-3**	0.693	**8.03E-3**
	Single [‡]	0.690	**2.60E-3**	0.688	**2.75E-3**
RNAalifold	rMSA	0.692	*	0.690	*
	rMSA (*Nf*)	0.693	6.35E-1	0.691	6.33E-1
	rMSA (*SCI*)	0.700	9.36E-1	0.698	9.37E-1
	RNAcmap	0.712	9.80E-1	0.710	9.80E-1
	Infernal	0.724	9.99E-1	0.723	9.99E-1
	nhmmer	0.692	5.28E-1	0.691	5.36E-1
	blastn	0.684	2.62E-1	0.682	2.69E-1
	RNAlien	0.673	6.29E-2	0.671	6.58E-2
	Single [‡]	0.670	**4.51E-2**	0.668	**4.72E-2**

* All p-values are calculated by one-tail t-test to check if rMSA is better (higher F1 and higher MCC) than the respective MSA schema. P-values<0.05 are in bold.

† Apart from canonical base pairs, a covariance analysis can also report other pairwise interactions, such as the coupling between nucleotide pairs adjacent to each other in the sequence. To exclude these non-canonical interactions, the output of covariance analysis is filtered by the following steps before calculating the accuracy: firstly, only Watson-Crick (A:U and G:C) and Wobble (G:U) base pairs are included; secondly, the two nucleotides must be separated by at least 4 positions in the sequence; thirdly, if a base is predicted to simultaneously paired to another two or more bases, only a single base pair with the best covariance score is reported.

‡ "Single" means the input MSA only includes the target sequence. The rSS prediction for a single sequence differs between PETfold and RNAalifold, despite both using the thermodynamics parameters from ViennaRNA. This is because PETfold uses the maximum expected accuracy (MEA) model while RNAalifold uses the minimum free energy (MFE).

§ Version number of all MSA and rSS prediction programs are listed in **Table S3**.

Table S3. Version number of third-party programs used in this study.

Program purpose	Program [†]	Version
Multiple sequence alignment (MSA)	blastn	2.10.1+
	nhmmer	3.3
	Infernal	1.1.3
	RNAcmap	808941b63e2f964eb40a2d2b18aa3d5d0560109c
	RNAlien	1.8.0
RNA secondary structure (rSS) prediction	RNAfold	2.4.14
	RNAalifold	2.4.14
	PETfold	2.1
	PLMC	1a9a1e9228a2177c618c69040ea8cfc2d902d9df
	R-scape	1.2.3
Structure conservation index (SCI)	RNAz	2.1

† RNAcontact used for RNA contact prediction is not included in this table because it does not have a version number.

Table S4. Average alignment depths of different MSA construction schemes before and after the alignments are filtered for PETfold and RNAalifold.

MSA	Before filtering		After filtering	
	N	*Nf*	*N*	*Nf*
rMSA [†]	5292.0	98.1	1171.4	11.1
rMSA (*Nf*) [†]	5308.8	98.7	1164.1	11.0
RNAcmap	23226.8	70.8	769.6	18.0
Infernal	9177.0	21.9	644.2	6.3
nhmmer	3111.3	2.9	104.7	0.7
blastn	1014.1	0.3	10.6	0.1

[†] "rMSA" is the standard rMSA pipeline where the final rMSA alignment is selected by covariance score. "rMSA (*Nf*)" is a modified rMSA pipeline where the final alignment is selected by alignment depth. All but one RNA (5voe Chain A) have different alignments before versus after filtering.

Table S5. Impact of using different single sequence rSS predictions for covariance model (CM) construction on rMSA alignment quality, as measured by F1-score and MCC of rSS prediction using the resulting alignment. The rSS predictions for CM construction tested here includes minimum free energy (MFE, default in RNAfold) prediction and maximum expected accuracy (MEA, default in PETfold) prediction, both using the thermodynamics parameters from ViennaRNA 2.4. As references, the single prediction accuracies of RNAfold and PETfold are also listed.

Alignment method	Final rSS predictor	F1-score of final rSS prediction	MCC of final rSS prediction
rMSA using MFE for CM	PLMC	0.648	0.646
	PETfold	0.711	0.714
rMSA using MEA for CM	PLMC	0.655	0.652
	PETfold	0.713	0.716

Table S6. Average RNAcontact accuracies for different MSAs

MSA	F1	P-value	MCC	P-value
rMSA	0.296	*	0.233	*
RNAcmap	0.282	**4.32E-3**	0.220	**1.57E-2**
Infernal	0.283	**1.77E-3**	0.222	8.61E-2
nhmmer	0.257	**1.93E-13**	0.195	**1.30E-11**
blastn	0.234	**1.94E-24**	0.169	**1.64E-22**
Single †	0.236	**1.55E-25**	0.172	**7.42E-23**

* All p-values are calculated by one-tail t-test to check if rMSA is better (higher F1 and higher MCC) than the respective MSA schema. P-values<0.05 are in bold.

Reference

[1] Zhang C, Zheng W, Mortuza SM, Li Y, Zhang Y. DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. Bioinformatics. 2020;36:2105-12.

[2] Weinreb C, Riesselman AJ, Ingraham JB, Gross T, Sander C, Marks DS. 3D RNA and functional interactions from evolutionary couplings. Cell. 2016;165:963-75.

[3] Zhang C, Zheng W, Mortuza SM, Li Y, Zhang Y. DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. Bioinformatics. 2020;36:2105-12.

[4] Washietl S, Hofacker IL, Stadler PF. Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci U S A. 2005;102:2454-9.

[5] Shrestha R, Fajardo E, Gil N, Fidelis K, Kryshtafovych A, Monastyrskyy B, et al. Assessing the accuracy of contact predictions in CASP13. Proteins. 2019;87:1058-68.

[6] Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. P Natl Acad Sci USA. 2013;110:15674-9.

[7] Jones DT, Buchan DWA, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics. 2012;28:184-90.