Supplementary Material

Appendix I: Calculate base alignment probabilities

The initial base alignment probabilities for all possible aligned pairs of bases between two sequences are calculated by using the partition function based method (Miyazawa, 1995; Muckstein, *et al.*, 2002). $P_{AB}(a_i, b_k)$ is defined as the base alignment probability (or match probability) between base a_i from sequence A and base b_k from sequence B, which is obtained as follows:

$$P_{AB}(a_i,b_k)=\frac{Z(\Omega_{i,k})}{Z}=\frac{Z_{i,k}^{M}\hat{Z}_{i,k}^{M}}{Z}\exp(-\sigma\cdot\varepsilon(a_i,b_k))$$

where $\Omega_{i,k}$ is the class of alignments in which a_i is matched to b_k; $Z_{i,k}{}^{M}$ is the partition function of all alignments between the partial sequences, $[a_1 \ldots a_{i-1}]$ and $[b_1 \ldots b_{k-1}]$, ending with a (mis)match of (a_i, b_k), which can be calculated by the forward algorithm (Durbin, *et al.*, 1998); analogously, $\hat{Z}_{i,k}^{M}$ is the partition function of all alignments between the partial sequences, $[a_{i+1} \ldots a_m]$ and $[b_{k+1} \ldots b_n]$, beginning with a (mis)match of (a_i, b_k), which can be calculated by the backward algorithm (Durbin, *et al.*, 1998); $\varepsilon(a_i, b_k)$ is the (mis)match score of (a_i, b_k); σ is a constant related to the thermodynamic temperature (Muckstein, *et al.*, 2002), which is set to 1 in this work.

Appendix II: A heuristic method to estimate base pairing probabilities of a sequence that allows pseudoknot structures

We use a sampling strategy to estimate the pseudo-base-pairing probabilities. From the list of all possible stems, we randomly choose a stem and update the list by removing and trimming all stems conflicting to the chosen one, and repeat this process until no stems remain. All chosen stems are compatible with each other and form one potential secondary structure for this sequence. We repeat the sampling process S times to generate S structures. We then calculate the frequency of a base pair (a_i, a_j) in these S structures, weighted by the total stacking energy of the structure in which this base pair occurs, to approximate the initial base pairing probability of (a_i, a_j) as:

$$D_A(a_i,a_j)=\frac{\frac{1}{S}+\sum_{s}C_s(a_i,a_j)}{1+\sum_{s}e^{-E(s)}}$$

where $s \in \{S$ sampled structures$\}$; given any s,

$$C_s(a_i,a_j)=\begin{cases} e^{-E(s)} & (a_i,a_j)\,\textit{forms a base pair in } s; \\ 0 & (a_i,a_j)\,\textit{does not form a base pair in } s. \end{cases}$$

This heuristic sampling for estimating base pairing probabilities allows pseudoknots to occur in a structure. A theoretical method based on the partition function to calculate base pairing probabilities has been proposed (Dirks and Pierce, 2004), however the algorithm is computationally expensive with the time complexity of $O(L^5)$, where L is the length of sequence.

Appendix III: Sample compatible blocks to generate common structures

A probabilistic sampling approach is used to sample compatible blocks to generate common structures between two sequences:

a. Probabilistically choose a block (β) from the block list. The chance that a block is picked is defined by the probability:

$$p(\beta)=\frac{W(\beta)}{\sum_{\beta\in\{all\ blocks\}}W(\beta)}$$

The higher the conservation score of a block, the more likely it is picked to be part of a structure.

b. Update the list of blocks by deleting or trimming all blocks that conflict with the chosen block, recalculate conservation scores of modified blocks and recalculate $p(\beta)$ based on the updated conservation scores of all remaining blocks. We introduced a parameter, X, which limits the maximum number of crossovers (resulting in pseudoknots) allowed in a structure. The blocks that form crossovers with the previously chosen blocks will be eliminated if the maximum number of crossovers is reached (By default, no pseudoknots are allowed, i.e. $X = 0$.).

c. Repeat steps a and b until no blocks remain. All selected blocks are compatible with each other and form a common structure shared by the two sequences.

d. Single-stranded regions between adjacent blocks or between two arms of a block are realigned by the probability alignment algorithm described in the Appendix I.

This sampling process is repeated *S* times, where *S* is the sample size, and *S* common structures are ultimately generated in each iteration.

Appendix IV: A fast comparison approach on multiple sequences

To reduce runtime, we designed a fast comparison approach for structure sampling on multiple (*N*) sequences: in each iteration, instead of sampling common structures between all pairwise sequences as described in the paper, it only randomly picks one sequence (*A*) and samples common structures between the picked sequence *A* and all other sequences. *S* structures are sampled between *A* and any other sequence (*B*). The picked sequence *A* is involved in a total of $(N\text{-}1)\cdot S$ sampled structures, and its base pairing probabilities are calculated as:

$$D_A^r(a_i,a_j)=\frac{1}{N-1}\cdot\sum_B\left(\frac{D_A^{r-1}(a_i,a_j)\cdot T+\sum_s C_s^{AB}(a_i,a_j)}{T+S}\right)$$

Any unpicked sequence *B* is only involved in *S* sampled structures, and its base pairing probabilities are calculated as:

$$D_B^r(b_k,b_l)=\frac{D_B^{r-1}(b_k,b_l)\cdot T+\sum_s C_s^{AB}(b_k,b_l)}{T+S}$$

where $s\in\{S$ sampled structures between sequence A and $B\}$; $B\in\{N$ sequences$\}$, $B\neq A$; C_s^{AB} (a_i, a_j), C_s^{AB} (b_k, b_l) and *T* are the same as those defined in the paper.

The base alignment probabilities between the picked sequence *A* and any other sequence *B* are calculated using the same procedure described in the paper for two sequences. Because no structures are sampled between two unpicked sequences (*B* and *B'*), the base alignment probabilities between them remain unchanged:

$$P_{BB'}^r(b_k,b'_k)=P_{BB'}^{r-1}(b_k,b'_k)$$

The computation complexity of the fast approach is $O(m^2\cdot r\cdot S\cdot N)$ compared to $O(m^2\cdot r\cdot S\cdot N^2)$ for the slow approach, where *m* is the minimum of total stem numbers in all sequences, and *r* is the number of iterations.

REFERENCES

Bafna, V., Tang, H. and Zhang, S. (2006) Consensus folding of unaligned RNA sequences revisited, J Comput Biol, 13, 283-295.

Dirks, R.M. and Pierce, N.A. (2004) An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots, J Comput Chem, 25, 1295-1304.

Durbin, R., Eddy, S.R., Krogh, A. and Mitchison, G. (1998) Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press.

Miyazawa, S. (1995) A reliable sequence alignment method based on probabilities of residue correspondences, Protein Eng, 8, 999-1009.

Muckstein, U., Hofacker, I.L. and Stadler, P.F. (2002) Stochastic pairwise alignments, Bioinformatics, 18 Suppl 2, S153-160.

Supplementary Table 1. 10 RNA regulatory motif or ncRNA gene families for testing the performance of RNA Sampler and other programs

RNA family	Rfam ID	Rfam category	Seed[a]	Coverage[b]	Average length	Average perc identity	Note
Cobalamin	RF00174	Cis-reg→ riboswitch	171	158	204	46	Cobalamin riboswitch
gcvT	RF00504	Cis-reg→ riboswitch	117	116	101	51	gcvT element
glmS	RF00234	Cis-reg→ riboswitch	14	13	184	58	glmS ribozyme
Purine	RF00167	Cis-reg→ riboswitch	37	37	100	56	Purine riboswitch
RFN	RF00050	Cis-reg→ riboswitch	48	48	145	66	FMN riboswitch (RFN element)
Sbox	RF00162	Cis-reg→ riboswitch	71	71	110	67	SAM riboswitch (S box leader)
THI	RF00059	Cis-reg→ riboswitch	237	209	110	52	TPP riboswitch (THI element)
tRNA	RF00005	Gene→ tRNA	1114	397	71	43	tRNA
U1	RF00003	Gene→ snRNA→ splicing	54	54	155	59	U1 spliceosomal RNA
yybP-ykoY	RF00080	Cis-reg→ riboswitch	74	74	128	45	yybP-ykoY element

a. Seed: number of unique sequences in the Rfam seed alignment.

b. Coverage: number of unique sequences included in all the 100 test sets generated for each RNA family.

Supplementary Table 2. Comparison of performance between RNA Sampler and other programs on two-sequence sets of 10 RNA families at the base pair level

RNA family	RNA Sampler			CARNAC			Dynalign			RNAalifold			Stemloc		
	CC[a]	*SEN*[a]	*SPE*[a]	*CC*	*SEN*	*SPE*	*CC*	*SEN*	*SPE*	*CC*	*SEN*	*SPE*	*CC*	*SEN*	*SPE*
Cobalamin	**0.42**[b]	0.45	**0.40**	0.27	0.24	0.31	0.40	**0.46**	0.34	0.26	0.27	0.26	0.07	0.05	0.09
gcvT	0.48	0.47	0.48	0.35	0.30	0.44	**0.50**	**0.55**	0.46	0.30	0.30	0.30	0.48	0.43	**0.56**
glmS	0.63	0.64	0.61	0.60	0.53	**0.69**	**0.65**	**0.69**	0.61	0.50	0.49	0.51	0.53	0.47	0.61
Purine	0.73	0.81	0.65	0.52	0.45	0.64	**0.79**	**0.84**	**0.75**	0.72	0.72	0.73	0.77	0.80	0.74
RFN	**0.53**	0.65	**0.43**	0.37	0.36	0.40	0.52	**0.67**	0.40	0.50	0.63	0.40	0.38	0.42	0.34
Sbox	0.72	0.75	**0.69**	0.34	0.25	0.48	**0.73**	**0.79**	0.67	0.55	0.56	0.55	0.55	0.47	0.67
THI	**0.55**	**0.54**	**0.57**	0.37	0.28	0.51	0.51	0.51	0.52	0.47	0.44	0.50	0.48	0.41	**0.57**
tRNA	**0.82**	0.80	0.84	0.67	0.58	0.79	**0.82**	**0.82**	0.82	0.64	0.60	0.69	0.78	0.72	**0.87**
U1	0.57	0.62	0.52	0.33	0.28	0.41	**0.63**	**0.70**	**0.56**	0.53	0.57	0.49	0.07	0.06	0.08
yybp-ykoY	0.54	0.50	0.59	0.42	0.34	0.53	**0.59**	**0.59**	0.59	0.33	0.30	0.38	0.45	0.34	**0.66**
Average	0.60	0.62	**0.58**	0.42	0.36	0.52	**0.61**	**0.66**	0.57	0.48	0.49	0.48	0.46	0.42	0.52

a. *CC*, *SEN* and *SPE* are the average values on all pairwise combinations of unique sequences from the Rfam seed alignments that were included in the multiple-sequence sets for each RNA family.

b. Bold fonts represent the highest values of *CC*, *SEN* and *SPE* predicted for each RNA family.

Supplementary Table 3. Comparison of performance between RNA Sampler and other programs on multiple-sequence sets of 10 RNA families at the base pair level

RNA family	RNA Sampler			CARNAC			Dynalign[b]			FoldAlignM			RNAalifold[c]			Stemloc		
	CC[a]	*SEN*[a]	*SPE*[a]	*CC*	*SEN*	*SPE*	*CC*	*SEN*	*SPE*	*CC*	*SEN*	*SPE*	*CC*	*SEN*	*SPE*	*CC*	*SEN*	*SPE*
Cobalamin	**0.59**[d]	**0.58**	**0.61**	0.35	0.26	0.49	0.40	0.47	0.34	0.41	0.44	0.39	0.29	0.20	0.44	0.15	0.12	0.18
gcvT	**0.57**	0.53	**0.62**	0.46	0.37	0.60	0.50	**0.55**	0.46	0.49	0.50	0.47	0.33	0.24	0.47	0.43	0.40	0.47
glmS	**0.85**	**0.87**	**0.84**	0.70	0.65	0.76	0.65	0.69	0.61	0.66	0.68	0.64	0.50	0.39	0.64	0.70	0.64	0.76
purine	**0.83**	**0.91**	0.77	0.73	0.66	0.82	0.79	0.84	0.74	0.76	0.81	0.71	0.77	0.70	**0.85**	0.77	0.79	0.76
RFN	0.64	**0.76**	0.54	0.48	0.47	0.51	0.51	0.67	0.39	0.50	0.66	0.38	**0.67**	0.75	**0.59**	0.46	0.53	0.40
sbox	**0.78**	**0.80**	0.76	0.47	0.37	0.61	0.72	0.79	0.67	0.75	**0.80**	0.70	0.68	0.61	**0.77**	0.58	0.52	0.66
THI	**0.69**	0.64	0.74	0.46	0.34	0.64	0.52	0.52	0.52	0.67	**0.66**	0.67	0.54	0.35	**0.87**	0.59	0.53	0.68
tRNA	**0.94**	0.93	**0.95**	0.78	0.71	0.86	0.82	0.83	0.82	**0.94**	**0.94**	0.93	0.72	0.62	0.86	0.84	0.78	0.90
U1	**0.63**	0.66	**0.61**	0.47	0.41	0.54	**0.63**	**0.70**	0.56	0.60	0.68	0.53	0.60	0.60	0.60	0.23	0.22	0.25
yybp-ykoY	**0.67**	0.58	**0.79**	0.53	0.42	0.69	0.59	0.59	0.58	0.64	**0.60**	0.67	0.33	0.21	0.52	0.62	0.51	0.77
Average	**0.72**	**0.73**	**0.72**	0.54	0.47	0.65	0.61	0.67	0.57	0.64	0.68	0.61	0.54	0.47	0.66	0.54	0.50	0.58

a. *CC*, *SEN*, and *SPE* are the average values on 100 sequence sets generated for each RNA family. Each set consists of five sequences from the Rfam seed alignments.

b. *CC*, *SEN* and *SPE* by Dynalign are the average values on all unique pairwise predictions.

c. *CC*, *SEN* and *SPE* by RNAalifold are the average values on sequence sets that gave non-zero *CCs*, due to the large number of prediction failures on sequence sets of low identities.

d. Bold fonts represent the highest values of *CC*, *SEN* and *SPE* predicted for each RNA family.

Supplementary Table 4. Comparison of performance between RNA Sampler and other programs on multiple-sequence sets of 10 RNA families based on at the stem level (Bafna*, et al.*, 2006)

RNA family	RNA Sampler			CARNAC			Dynalign[b]			FoldAlignM			RNAalifold[c]			Stemloc		
	CC[a]	*SEN*[a]	*SPE*[a]	*CC*	*SEN*	*SPE*	*CC*	*SEN*	*SPE*	*CC*	*SEN*	*SPE*	*CC*	*SEN*	*SPE*	*CC*	*SEN*	*SPE*
Cobalamin	**0.63**[d]	**0.61**	**0.66**	0.38	0.23	0.65	0.44	0.48	0.41	0.46	0.46	0.46	0.32	0.22	0.50	0.16	0.12	0.23
gcvT	**0.65**	0.58	0.74	0.49	0.33	**0.77**	0.64	**0.66**	0.62	0.58	0.57	0.59	0.41	0.30	0.59	0.48	0.41	0.57
glmS	**0.86**	**0.91**	0.81	0.72	0.62	**0.85**	0.66	0.71	0.61	0.61	0.64	0.59	0.54	0.45	0.65	0.66	0.59	0.74
Purine	0.84	**0.98**	0.73	0.80	0.72	**0.92**	0.84	0.94	0.76	0.80	0.90	0.71	0.84	0.81	0.89	**0.85**	0.88	0.82
RFN	0.75	**0.96**	0.59	0.48	0.43	0.55	0.68	0.90	0.51	0.61	0.81	0.46	**0.81**	0.95	**0.70**	0.45	0.51	0.40
sbox	**0.84**	0.80	0.88	0.52	0.35	0.81	0.83	**0.83**	0.82	0.81	0.80	0.82	0.76	0.66	**0.89**	0.55	0.43	0.72
THI	**0.82**	0.80	0.85	0.56	0.39	0.83	0.65	0.65	0.65	0.79	**0.81**	0.78	0.65	0.45	**0.96**	0.70	0.61	0.80
tRNA	**0.96**	**0.95**	**0.97**	0.80	0.68	0.94	0.86	0.86	0.86	0.95	**0.95**	0.95	0.74	0.63	0.91	0.83	0.77	0.91
U1	0.69	0.78	0.61	0.55	0.48	0.64	**0.73**	**0.87**	0.61	0.70	0.85	0.58	0.72	0.75	**0.69**	0.26	0.25	0.28
yybp-ykoY	**0.69**	0.62	0.77	0.53	0.38	0.76	0.60	**0.61**	0.59	0.67	0.65	0.69	0.36	0.24	0.54	0.62	0.49	**0.78**
Average	**0.77**	**0.80**	0.76	0.58	0.46	**0.77**	0.69	0.75	0.64	0.70	0.74	0.66	0.62	0.55	0.73	0.56	0.51	0.63

a. *CC*, *SEN*, and *SPE* are the average values on 100 sequence sets generated for each RNA family. Each set consists of five sequences from the Rfam seed alignments. Single base pair stems in the Rfam structures are excluded from calculation.

b. *CC*, *SEN* and *SPE* by Dynalign are the average values on all unique two-sequence sets for each RNA family.

c. *CC*, *SEN*, and *SPE* by RNAalifold are the average values on sequence sets that gave non-zero *CCs*, due to the large number of prediction failures on sequence sets of low identities.

d. Bold fonts represent the highest values of *CC*, *SEN*, and *SPE* predicted for each RNA family.

Supplementary Table 5. Comparison of runtimes between RNA Sampler (fast and slow approaches) and other programs on multiple-sequence sets of 10 RNA families

RNA family	RNA Sampler (slow)	RNA Sampler (fast)	CARNAC	Dynalign[a]	FoldalignM	RNAalifold	Stemloc
	(s)	(s)	(s)	(s)	(s)	(s)	(s)
Cobalamin	179	94	1.81	8877	2694	0.14	157
gcvT	16	9	0.20	749	212	0.04	85
glmS	94	51	0.40	3518	2022	0.14	110
Purine	12	7	0.18	394	62	0.05	40
RFN	50	30	0.20	3348	622	0.08	11
Sbox	23	13	0.21	706	268	0.05	19
THI	27	15	0.30	585	403	0.08	27
tRNA	5	3	0.15	187	37	0.03	98
U1	89	51	0.27	5415	572	0.13	32
yybp-ykoY	37	22	0.27	1306	429	0.06	65

a. The runtime of Dynalign is the average of 10 randomly selected two-sequence sets in each RNA family. The runtimes of all other programs are based on the average of all 100 five-sequence sets in each RNA family.

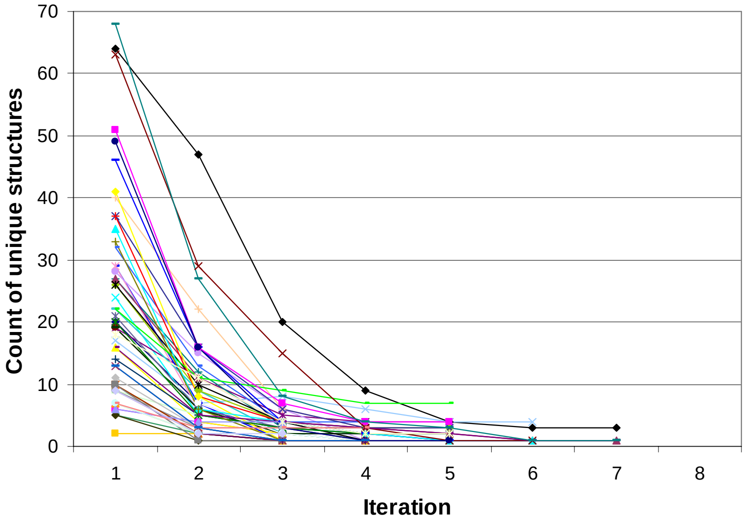
Supplementary Table 6. Comparison of performance and runtime between RNA Sampler (fast and slow approaches) and other programs on multiple-sequence sets of 10 RNA families at the base pair level

RNA family	RNA Sampler (slow)	RNA Sampler (fast)	RNA Sampler (slow)			RNA Sampler (fast)			Best among others programs[b]		
	runtime (s)	runtime (s)	*CC*[a]	*SEN*[a]	*SPE*[a]	*CC*	*SEN*	*SPE*	*CC*	*SEN*	*SPE*
Cobalamin	179	94	**0.59**[c]	**0.58**	**0.61**	0.52	0.49	0.55	0.41	0.47	0.49
gcvT	16	9	**0.57**	0.53	**0.62**	0.52	0.48	0.57	0.50	0.55	0.60
glmS	94	51	**0.85**	**0.87**	**0.84**	0.77	0.75	0.79	0.70	0.69	0.76
Purine	12	7	**0.83**	**0.91**	0.77	0.79	0.85	0.74	0.79	0.84	**0.85**
RFN	50	30	0.64	**0.76**	0.54	0.60	0.70	0.53	0.67	0.75	**0.59**
Sbox	23	13	**0.78**	**0.80**	0.76	0.73	0.73	0.74	0.75	**0.80**	**0.77**
THI	27	15	**0.69**	**0.64**	0.74	0.63	0.57	0.70	0.67	0.66	**0.87**
tRNA	5	3	**0.94**	0.93	**0.95**	0.91	0.89	0.93	**0.94**	**0.94**	0.93
U1	89	51	**0.63**	0.66	**0.61**	0.61	0.62	0.60	**0.63**	**0.70**	0.60
yybp-ykoY	37	22	**0.67**	0.58	**0.79**	0.59	0.50	0.70	0.64	**0.60**	0.77

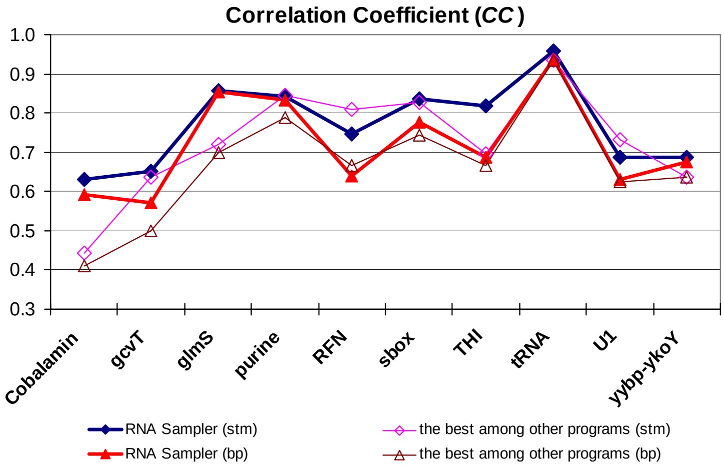
a. *CC*, *SEN*, and *SPE* are the average values on 100 sequence sets generated for each RNA family. Each set consists of five sequences from the Rfam seed alignments.

b. The best *CC*, *SEN* and *SPE* among CARNAC, Dynalign, FoldalignM, RNAalifold or Stemloc.

c. Bold fonts represent the highest values of *CC*, *SEN*, and *SPE* predicted for each RNA family.



Supplementary Fig. 1. Convergence tests of RNA Sampler on 55 pairwise combinations of 11 tRNA sequences**.** The sample size for each iteration is 100, i.e. *S* = *100*.



Supplementary Fig. 2. Correlation coefficient (*CC*) of predictions between RNA Sampler and other programs (CARNAC, Dynalign, FoldalignM, RNAalifold and Stemloc) based on exact base pair matches (▲RNA Sampler, △ the best among other programs) and overlapped stems (◆RNA Sampler, ◇the best among other programs). “The best among other programs” is the best performance among CARNAC, Dynalign, FoldalignM, RNAalifold and Stemloc on each RNA family. Detailed values are shown in Supplementary Table 3 and 4.

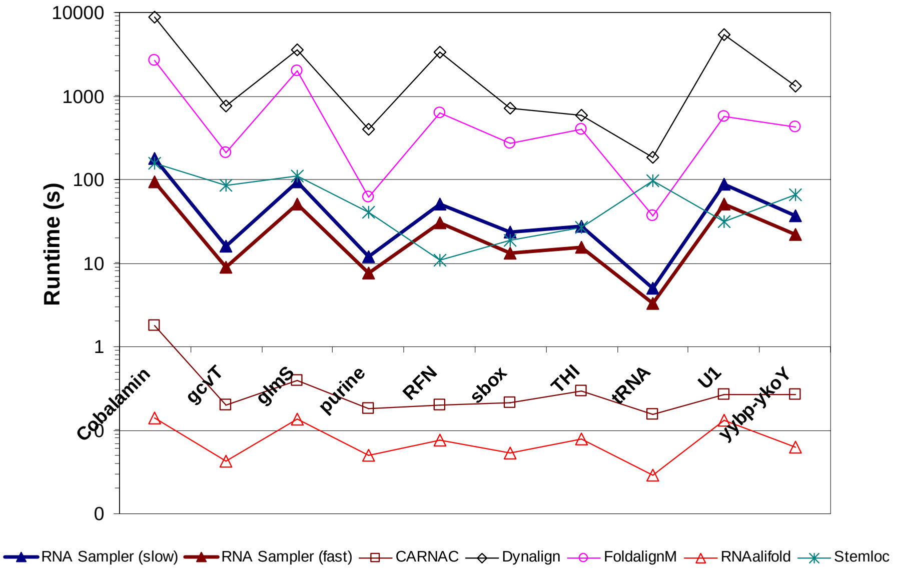
(a)

```
Published pseudoknot structure for α operon mRNA leader sequence in E.coli
>E.coli
GCCAATCTTTTGTATGTCTGTGCGTTTCCATTTGAGTATCCTGAAAACGGGCTTTTCAGCATGGAACGTACATATTAAATAGTAGGAGTGCATAGTGGCCCGTATAGCAGGCATTAACATTCCTGATCATAAGC
------------------<<<<<<<-<<<<<----------------------------->>>>>>>>>>>>--------------------------------------------------------------   18 -  30,  60 -  71
--------------------------------------<<<<----------------------------------------------------------------->>>>-----------------------   38 -  41, 107 - 110
-----------------------------------------------<<<<----------------------------------------------->>>>--------------------------------   47 -  51,  98 - 101
-------------------------------------------------------<<<<--------------------------------------------------------------->>>>--------   55 -  58, 122 - 125

Predicted pseudoknot structure by RNA Sampler (X = 3)
>E.coli
GCCAATCTTTTGTATGTCTGTGCGTTTCCATTTGAGTATCCTGAAAACGGGCTTTTCAGCATGGAACGTACATATTAAATAGTAGGAGTGCATAGTGGCCCGTATAGCAGGCATTAACATTCCTGATCATAAGC
--------<<<<<------------------------------------------------------------------------------------------------------------------->>>>>-    8 -  12, 128 - 132
------------------<<<<<<<<<<---------------------------------->>>>>>>>>>--------------------------------------------------------------   18 -  27,  62 -  71
-----------------------------------------<<<<<------->>>>>----------------------------------------------------------------------------   41 -  45,  53 -  57
----------------------------------------------<<<<<<<------------------------------------------->>>>>>>-------------------------------   46 -  52,  96 - 102
----------------------------------------------------------------------------------((((((((---------------------------))))))))---------   82 -  89, 117 - 124
```

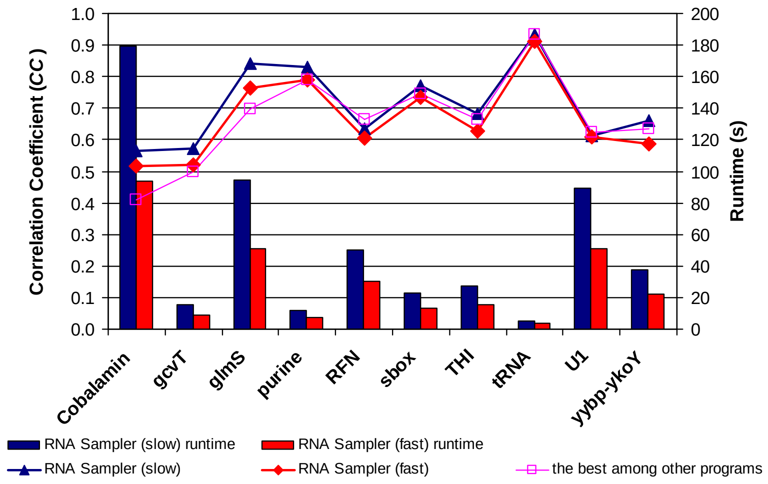
(b)

```
Published pseudoknot structure for S15 mRNA leader sequence in E.coli
> E.coli
UGUUAACCGUCUUGCGAUA..AACGUCGCGUAAAUUGUUUAACACUUUGCGUAACGUACACUGGGAUCGCUGAAUUAGAGAUCGGCGUCCUUUCAUUCUAUAUACUUUGGAGUUUUAAAAUGUCUCUAAGUACUGAAGCAACAGCUAAAAUCGUUU
-------------------..-<<<<-<<<<<<<----------->>>>>>>->>>>---------------------------------------------------------------------------------------------------
-------------------..-----------------------------------------<<<<<-<<<<<-------->>>>>>>>>>-----------------------------------------------------------------
-------------------..----------------------------------------------------<<<<<<<<---------------------------------------->>>>>>>>---------------------------

Predicted pseudoknot structure by RNA Sampler (X = 2)
> E.coli
UGUUAACCGUCUUGCGAUA..AACGUCGCGUAAAUUGUUUAACACUUUGCGUAACGUACACUGGGAUCGCUGAAUUAGAGAUCGGCGUCCUUUCAUUCUAUAUACUUUGGAGUUUUAAAAUGUCUCUAAGUACUGAAGCAACAGCUAAAAUCGUUU
-------------<<<<<-..-------------------------------------------------------------------------------------------------------------------------------->>>>>--
-------------------..------<<<<<<<----------->>>>>>>--------------------------------------------------------------------------------------------------------
-------------------..-----------------------------------------<<<<<-<<<<---------->>>>>>>>>-----------------------------------------------------------------
-------------------..-----------------------------------------------------<<<<<<<<--------------------------------------->>>>>>>>---------------------------


Published stem-loop structure for S15 mRNA leader sequence in E.coli
> E.coli
UGUUAACCGUCUUGCGAUA..AACGUCGCGUAAAUUGUUUAACACUUUGCGUAACGUACACUGGGAUCGCUGAAUUAGAGAUCGGCGUCCUUUCAUUCUAUAUACUUUGGAGUUUUAAAAUGUCUCUAAGUACUGAAGCAACAGCUAAAAUCGUUU
-------------------..-<<<<-<<<<<<<----------->>>>>>>->>>>---------------------------------------------------------------------------------------------------
-------------------..-----------------------------------------<<<<<-<<<<<-------->>>>>>>>>>-----------------------------------------------------------------
-------------------..---------------------------------------------------------------------------------<<<<<-<<<<----------->>>>>>>>-------------------------

Predicted stem-loop structure by RNA Sampler (X = 0)
> E.coli
UGUUAACCGUCUUGCGAUA..AACGUCGCGUAAAUUGUUUAACACUUUGCGUAACGUACACUGGGAUCGCUGAAUUAGAGAUCGGCGUCCUUUCAUUCUAUAUACUUUGGAGUUUUAAAAUGUCUCUAAGUACUGAAGCAACAGCUAAAAUCGUUU
-------------<<<<<-..-------------------------------------------------------------------------------------------------------------------------------->>>>>--
-------------------..------<<<<<<<----------->>>>>>>--------------------------------------------------------------------------------------------------------
-------------------..-----------------------------------------<<<<-<<<<<<-------->>>>>>>>>>-----------------------------------------------------------------
-------------------..------------------------------------------------------------------------------------<<<<<<<----------->>>>>>>--------------------------
```

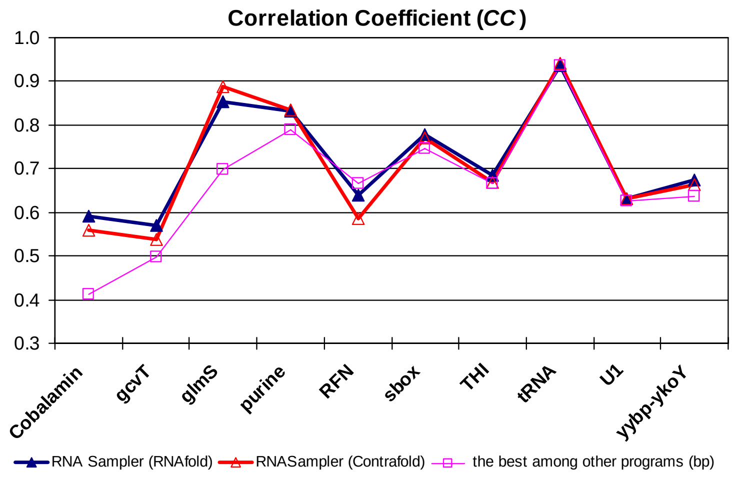
Supplementary Fig. 3. RNA Sampler predictions on multiple-sequence sets with pseudoknot structures. Only the predicted structure in *E.coli* is shown and compared to known structures. "**<<<--->>>**" represents stems; "**(((---)))**" represents conserved stems also predicted by comRNA but not in the known structure. All predictions are obtained using the following parameters: *r = 15*, *S = 100, sl = 4* and *d = 15.* (a). Predictions on α operon mRNA leader sequences with *X = 3*; (b). Predictions on S15 mRNA leader sequences. The pseudoknot structure was predicted with *X = 2*, and the stem-loop structure with *X = 0.*



Supplementary Fig. 4. Comparison of average runtime between RNA Sampler and other programs (CARNAC, Dynalign, FoldalignM, RNAalifold and Stemloc) on multiple-sequence sets of 10 RNA families. Detailed values are shown in Supplementary Table 5. The runtime of Dynalign is the average of 10 randomly selected two-sequence sets in each family.

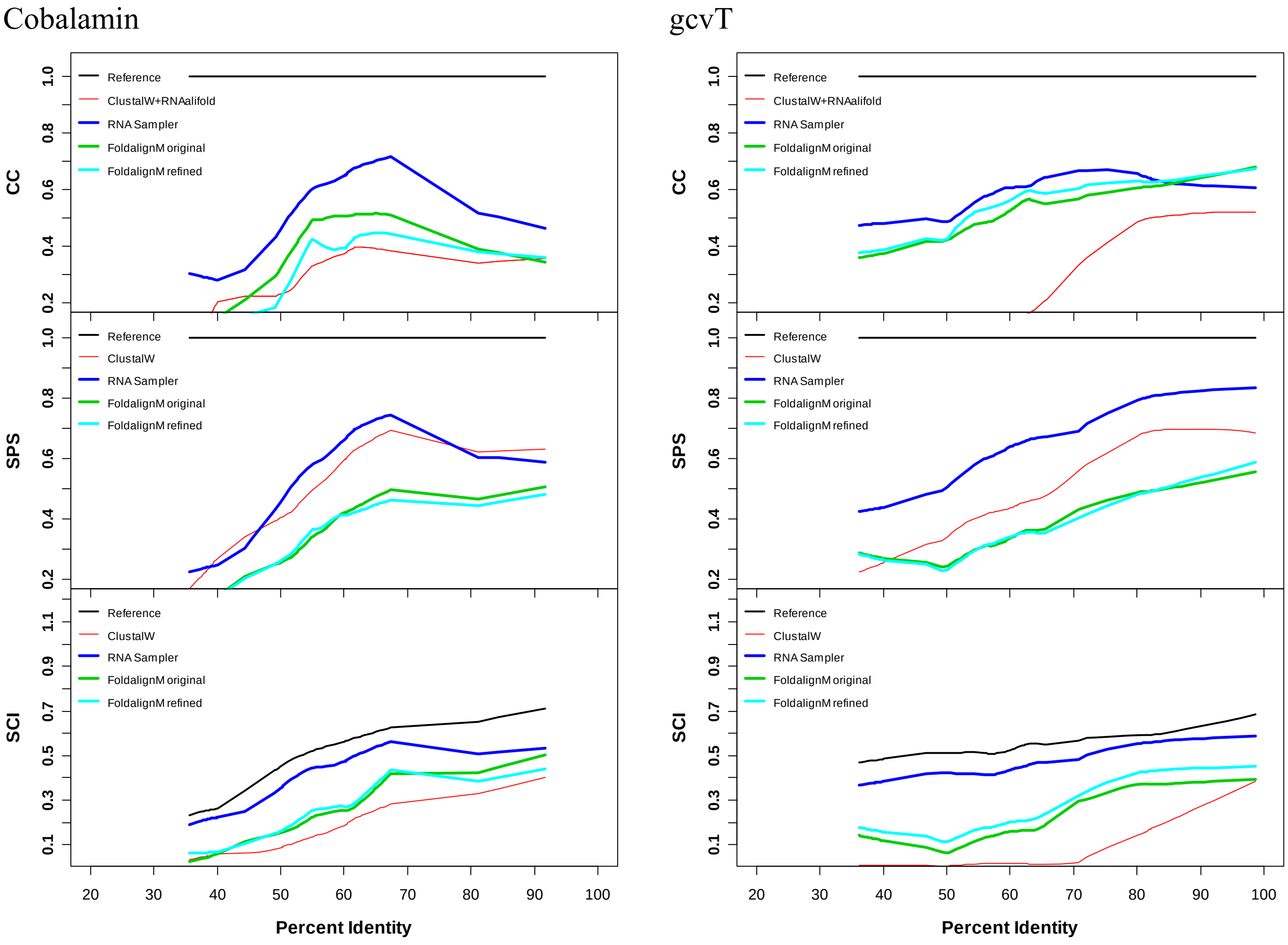


Supplementary Fig. 5. Comparison of performance and runtime between RNA Sampler (fast and slow approaches) and other programs (CARNAC, Dynalign, FoldalignM, RNAalifold and Stemloc). Detailed values are shown in Supplementary Table 6. "The best among other programs" is the best performance among CARNAC, Dynalign, FoldalignM, RNAalifold and Stemloc on each RNA family.



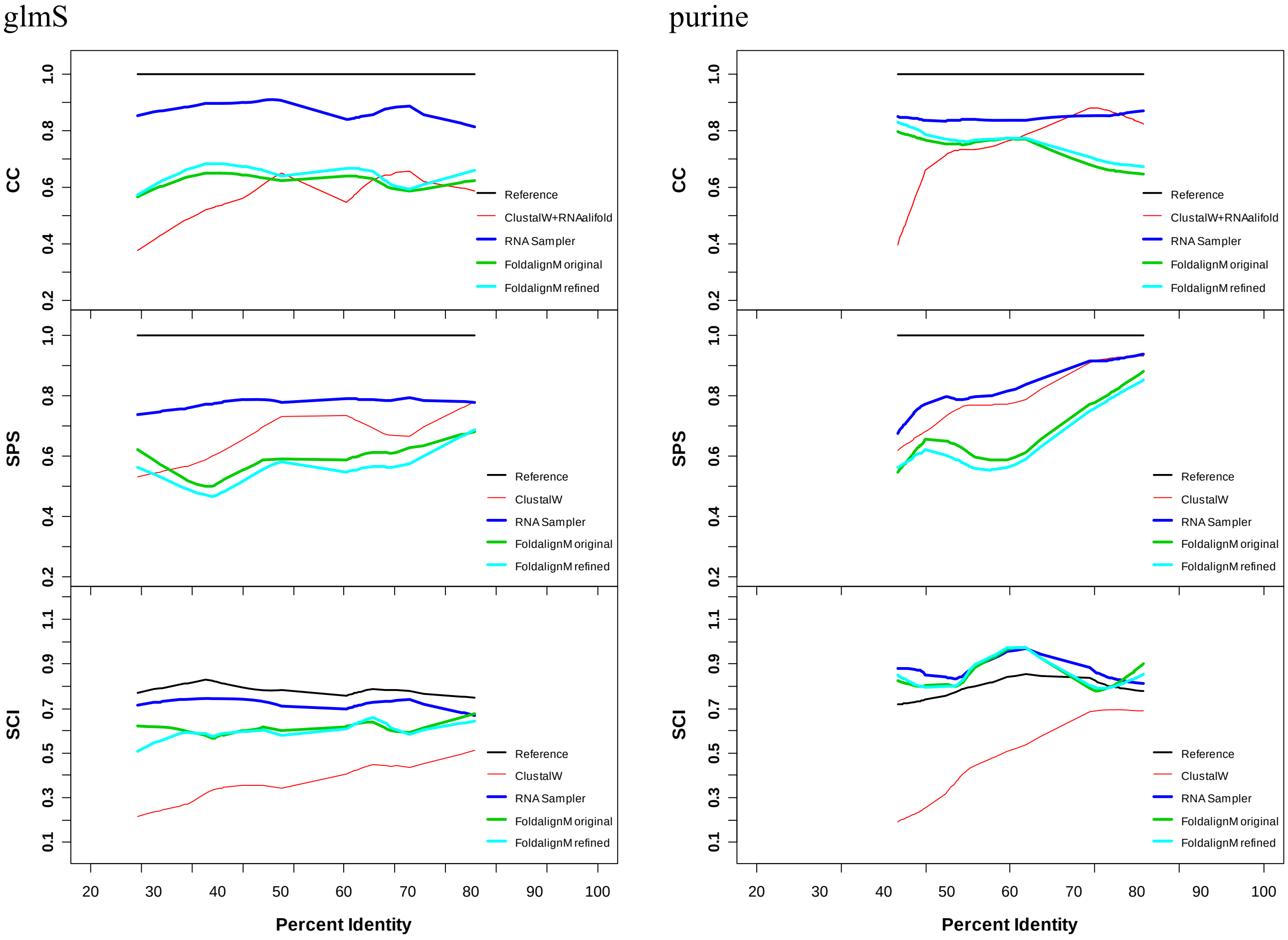
Supplementary Fig. 6. Comparison of performance between RNA Sampler with different initialization methods (base pairing probabilities calculated by RNAfold or by Contrafold) and other programs (CARNAC, Dynalign, FoldalignM, RNAalifold and Stemloc). "The best among other programs" is the best performance among CARNAC, Dynalign, FoldalignM, RNAalifold and Stemloc on each RNA family.

Supplementary Fig. 7. Comparison of *CC*, *SPS* and *SCI* among RNA Sampler, FoldalignM and RNAalifold (on ClustalW alignments) on individual RNA families: Cobalamin, gcvT, glmS, purine, RFN, sbox, THI, tRNA, U1, yybp-ykoY. The Rfam seed alignments and structures were used as benchmarking references. Only reliable alignments in the stem regions of the Rfam seed alignments were examined in calculating the *SPS* scores. The curves were generated using lowess (locally weighted regression) smoothing.



Cobalamin

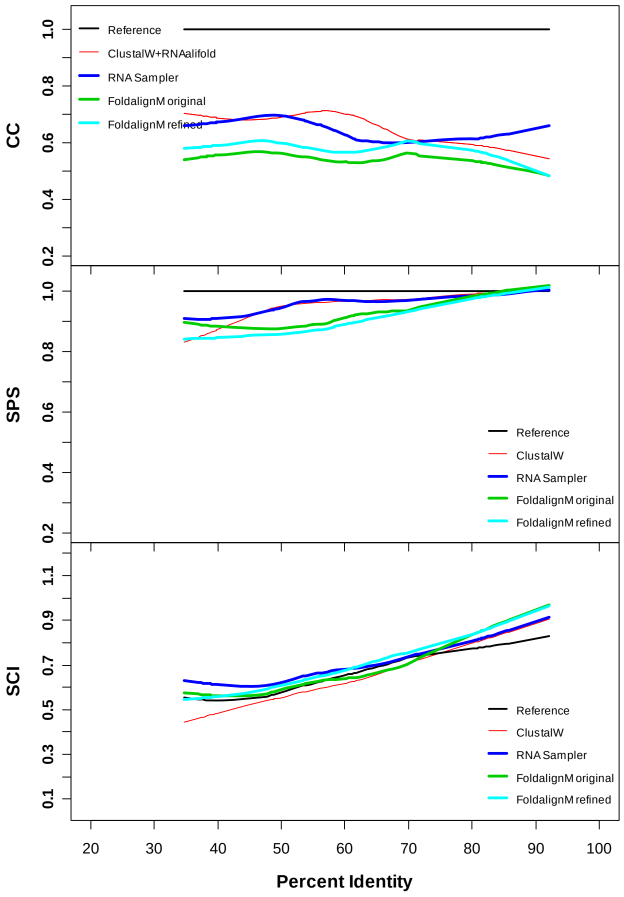
gcvT


glmS

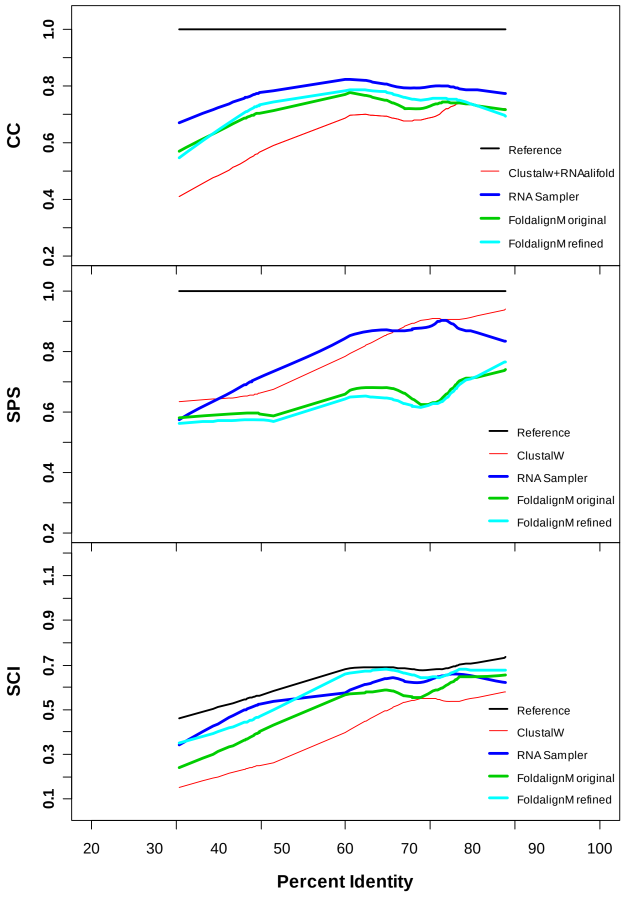
purine

CC
SPS
SCI
1.0
0.8
0.6
0.4
0.2
1.1
0.9
0.7
0.5
0.3
0.1
Reference
ClustalW+RNAalifold
RNA Sampler
FoldalignM original
FoldalignM refined
Reference
ClustalW
RNA Sampler
FoldalignM original
FoldalignM refined
20
30
40
50
60
70
80
90
100
Percent Identity

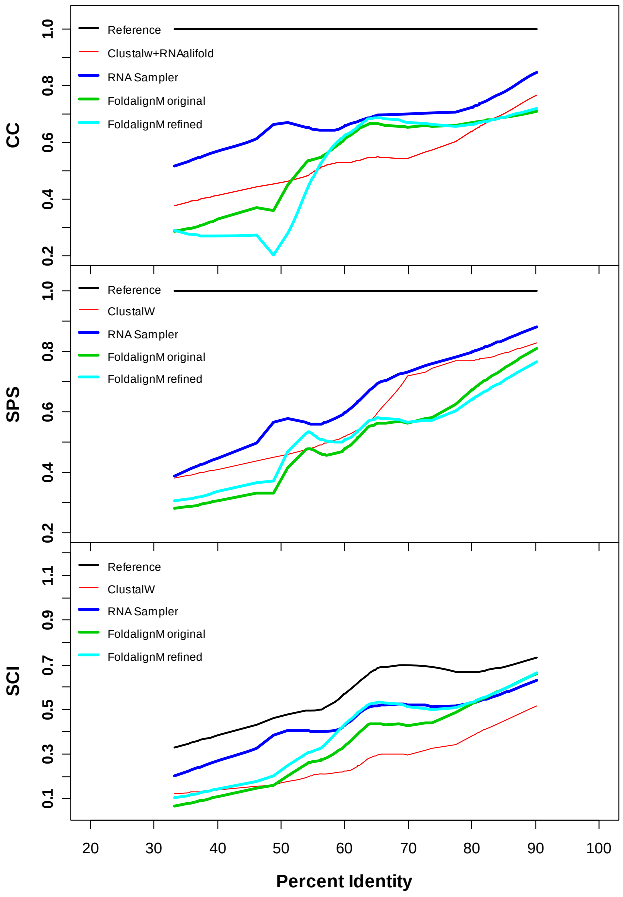
RFN


CC
SPS
SCI
1.0
0.8
0.6
0.4
0.2
1.1
0.9
0.7
0.5
0.3
0.1
Reference
ClustalW+RNAalifold
RNA Sampler
FoldalignM original
FoldalignM refined
Reference
ClustalW
RNA Sampler
FoldalignM original
FoldalignM refined
Reference
ClustalW
RNA Sampler
FoldalignM original
FoldalignM refined
20
30
40
50
60
70
80
90
100
Percent Identity

sbox

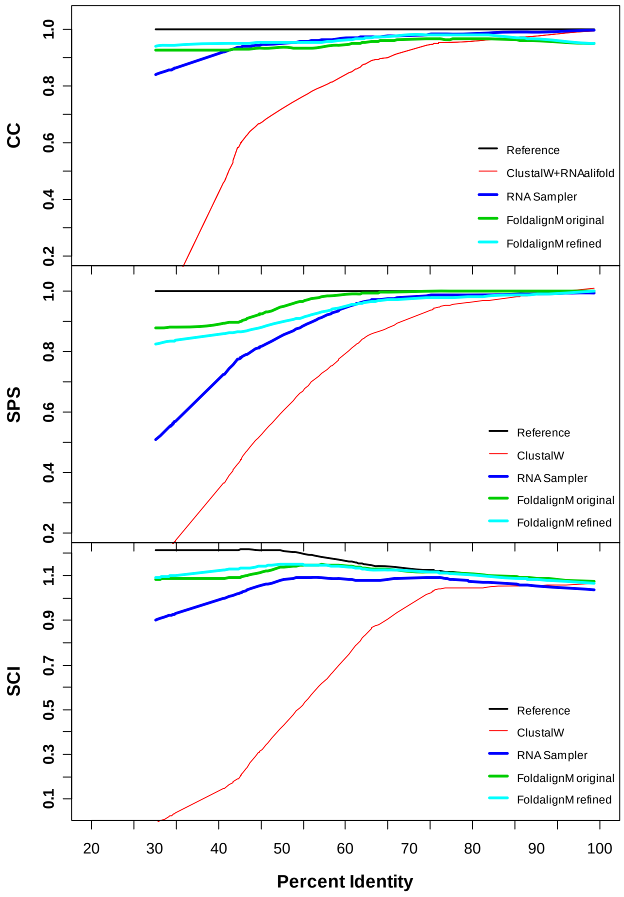

CC
SPS
SCI
1.0
0.8
0.6
0.4
0.2
1.1
0.9
0.7
0.5
0.3
0.1
Reference
Clustalw+RNAalifold
RNA Sampler
FoldalignM original
FoldalignM refined
Reference
ClustalW
RNA Sampler
FoldalignM original
FoldalignM refined
Reference
ClustalW
RNA Sampler
FoldalignM original
FoldalignM refined
20
30
40
50
60
70
80
90
100
Percent Identity

THI

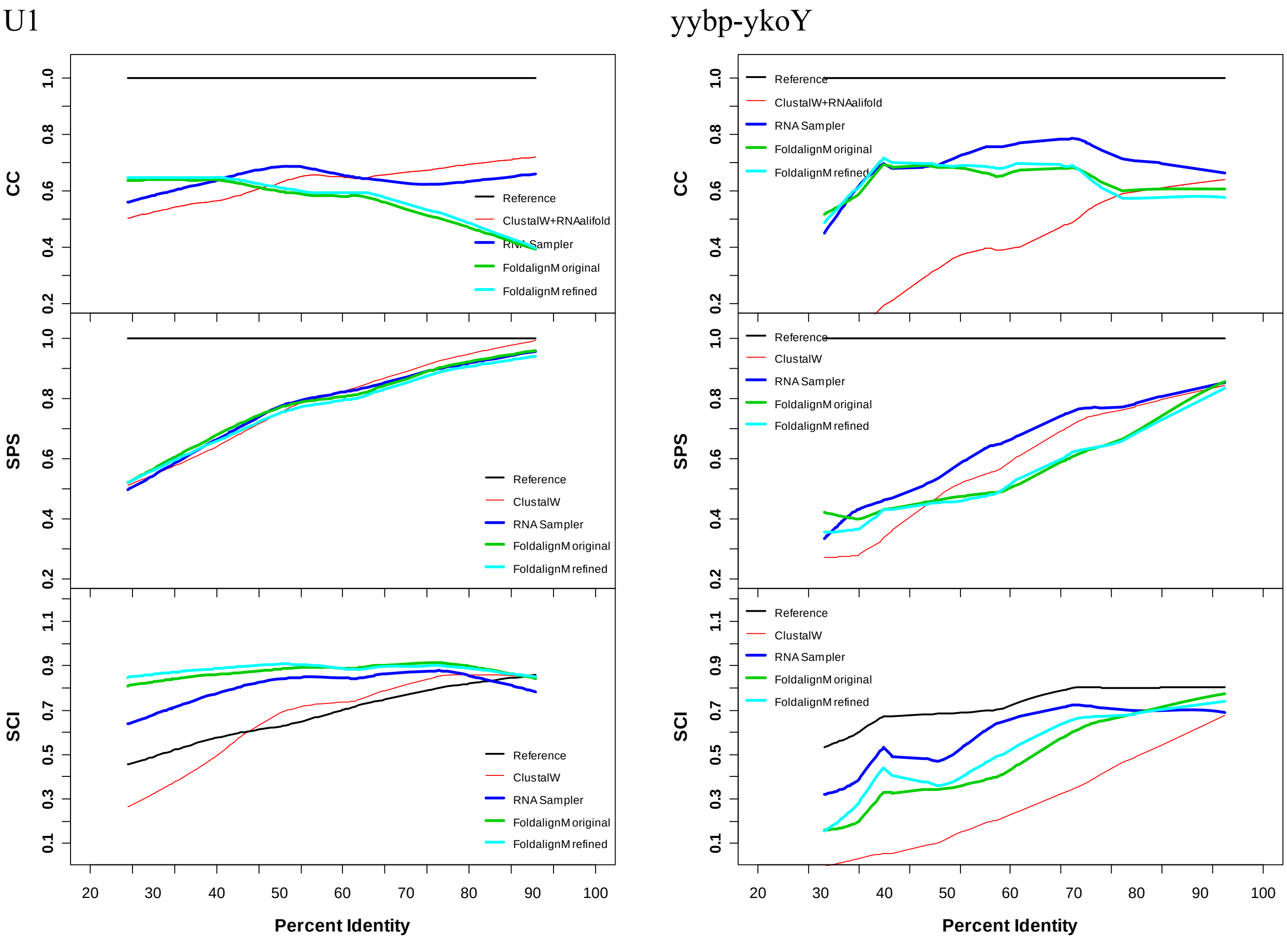


Reference
Clustalw+RNAalifold
RNA Sampler
FoldalignM original
FoldalignM refined
CC
0.2
0.4
0.6
0.8
1.0
Reference
ClustalW
RNA Sampler
FoldalignM original
FoldalignM refined
SPS
0.2
0.4
0.6
0.8
1.0
Reference
ClustalW
RNA Sampler
FoldalignM original
FoldalignM refined
SCI
0.1
0.3
0.5
0.7
0.9
1.1
20
30
40
50
60
70
80
90
100
Percent Identity

tRNA



Reference
ClustalW+RNAalifold
RNA Sampler
FoldalignM original
FoldalignM refined
CC
0.2
0.4
0.6
0.8
1.0
Reference
ClustalW
RNA Sampler
FoldalignM original
FoldalignM refined
SPS
0.2
0.4
0.6
0.8
1.0
Reference
ClustalW
RNA Sampler
FoldalignM original
FoldalignM refined
SCI
0.1
0.3
0.5
0.7
0.9
1.1
20
30
40
50
60
70
80
90
100
Percent Identity


U1
yybp-ykoY
CC
SPS
SCI
1.0
0.8
0.6
0.4
0.2
1.1
0.9
0.7
0.5
0.3
0.1
Reference
ClustalW+RNAalifold
ClustalW
RNA Sampler
FoldalignM original
FoldalignM refined
20
30
40
50
60
70
80
90
100
Percent Identity