Algorithm details

Self-Adaptation

Self-adaptation can be used to tune the probabilities associated with each variation operator concurrently with the process of evolution. Every solution in the population carries a set of variation probabilities and passes this information to the subsequent generation. The number of variation operators applied to each individual is determined by the formula

$$\hat{Y}'_{k+1} = Y_k + N(0,1) \quad (1)$$

where Y_k is the mean number of variation operators to apply at step k and $N(0,1)$ is a normal Gaussian distribution with mean $\mu = 0$, and standard deviation $\sigma = 1$. Given U structures in a bin,

$$Y_{k+1} = \max(\min(\hat{Y}'_{k+1}, U/2), 1) \quad (2)$$

The actual number of variation operators (Q_k) to apply at step k can be generated using a Poisson distribution with mean Y_k

$$Q_k = \text{Poisson}(Y_k) \quad (3)$$

Given a specific value for Q_k, variation operators must now be identified. This choice is made as a probability over the four possible variation operators.

Let m = the number of possible variation operators,

$\mu = 1/m$

$\sigma = 0.1 \cdot \mu$

$\mu \in [0,1]$

$\sigma \in [0,1]$

$a_{k,i}$ = probability of choosing operator i at time step k where $a_{k,i} \in [0,1] \forall i$ and

$$\sum_{i=1}^{n} a_{k,i} = 1 \quad (4)$$

Let $d_{k,i} = (a_{k,i} - \Delta)$

$\hat{d}_{k,i} = | d_{k,i} |$

$$x_{k,i} = m \cdot a_{k,i} + \Delta[N(0,1) \cdot \Delta \cdot \hat{d}_{k,i} \cdot sign(d_{k,i})] \quad (5)$$

where sign $(x) = 1$ if $x \geq 0$

sign $(x) = -1$ if $x < 0$.

The factor m applied to $a_{k,i}$ scales $x_{k,i}$ into the range [0,1] plus epsilon. The term $\Delta \cdot \hat{d}_{k,i} \cdot sign(d_k)$ is proportional to the difference between the current (at time step k) probability of choosing operator i and a uniform probability of choosing from the m operators. Thus, this term acts to drive $a_{k,i}$ back toward a uniform distribution. Finally,

$$\hat{x}_{k,i} = \max(x_{k,i}, \Delta_i) \quad (6)$$

where Δ represents a minimum probability threshold value to ensure the probability of choosing operator i is always non-zero, and rescaling,

$$a_{k+1,i} = \frac{\hat{x}_{k,i}}{\sum_{i=1}^{n} \hat{x}_{k,i}} \quad (7)$$

In addition to these methods of generating bins, the user can choose to either allow or avoid the placement of structures that overlap in sequence space in the same bin to avoid trivially redundant structures.

Fitness Components / Scoring

Nucleotide sequence similarity within a structural component measures the pairwise sequence similarity within each element of the RNA structure, as defined by the individual components in the RNAMotif descriptor. Using the algorithm ALIGN, the nucleotide strings in a

single bin are compared in a pairwise fashion to generate the best pairwise alignments over the set of symbols {A, G, C, U, □} using the match, mismatch, and gap penalty values shown in Figure S1 with an additional □12 penalty for gap opening. All component blocks (*b*) are scored separately for pairwise nucleotide similarity and associated with a sequence score (SS_b).

Component-based calculation of similarity offers distinct advantages in that the user may wish to specify an additional bonus for similarity in a particular structural component. In the experiments below, stems were given a weight of 1.2 and all single-stranded regions a weight of 1.0. The weights associated with all components in the pairwise comparison are summed for an overall score (CW_{tot}) using the equation

$$CW_{tot} = \sum_{b=1}^{n} CW_b \tag{8}$$

where *b* is the index for each component block. A weighted sequence score is then generated for each block

$$SS'_b = CW_b \square SS_b \tag{9}$$

and the weighted score is summed over all blocks

$$SS'_{tot} = \sum_{b=1}^{n} SS'_b \tag{10}$$

A final pairwise sequence score (SEQ) is generated using the equation

$$SEQ_{i,j} = \frac{\left(SS'_{tot} / CW_{tot}\right)}{\max(L_i, L_j)} \tag{11}$$

where *i* and *j* are structure indices and *L* is the length of the sequences being compared (Figure S2a).

The above calculation represents the sequence comparison of two structures in a bin. The overall fitness score for the sequence similarity of all structure pairs in a bin can be calculated by

summing the $SEQ_{i,j}$ scores and normalizing this value over the number of pairwise combinations (p) in a bin

$$SEQ_{tot} = \frac{\sum_{i=1}^{p} \sum_{j=1}^{p} SEQ_{i,j}}{p} \tag{12}$$

The range of minimum and maximum possible alignment scores in the RNAMotif output file is then calculated. This is done by determining the longest sequence for each structure block in the output file, and calculating scores for the theoretical conditions where each of the longest structure block was either paired with an identical copy of this sequence (the maximum sequence *similarity* score over the entire RNAMotif file), or with an equally long artificial "sequence" composed only of gaps (maximal *dissimilarity*). These maximum and minimum scores were used for normalization to the range [0,1] for all other sequence comparisons using the equation:

$$SEQ'_{tot} = \left[SEQ_{tot} \left(\frac{b - a}{d - c} \right) \right] + \left(\frac{a \cdot d - b \cdot c}{d - c} \right) \tag{13}$$

where $a = 0$, $b = 1$, c is the maximal dissimilarity score in the RNAMotif output file, d is the maximal similarity score in the RNAMotif output file and SEQ_{tot} is the total sequence score for all pairwise comparisons in a bin.

The second term in the fitness function, *structure component length similarity* (SCLS), is used to measure the similarity in terms of the lengths of all components in a structure. In the case, where a range of lengths is provided for any structure component in an RNAMotif descriptor, components of differing lengths can be generated. For each structure being compared, the length of each component is determined. The lengths of these individual structural components are compared on a pairwise basis for all structures in a bin. A structure score for each component block (ST) is calculated using the equation

$$ST_b = 1 - \left(\frac{\max(C_{1b}, C_{2b}) - \min(C_{1b}, C_{2b})}{\max B_b - \min B_b} \right) \tag{14}$$

where C_1 and C_2 are the structures being compared, max(C_{1b}, C_{2b}) is the maximum length sequence within component block b, min(C_{1b}, C_{2b}) is the minimum length sequence within block b, max B_b is the maximum length structure for block b found in the RNAMotif output file, and min B_b is the minimum length structure for block b found in the RNAMotif output file. In the condition where two structures contain missing components (represented as "**.**" in RNAMotif), each component is equated with a score of 0.

User-defined weights are associated with the importance of similar length for each structure component. For the experiments described below, these weights were all set to 1.0, except for length similarity in stems, which was set to 1.2. The component weight scores are summed over all component blocks b using the formula

$$CW_{tot} = \sum_{b=1}^{n} CW_b \tag{15}$$

A weighted sequence score is then generated for each block

$$ST'_b = CW_b \cdot ST_b \tag{16}$$

and the weighted scored summed over all blocks

$$ST_{tot} = \sum_{b=1}^{n} ST'_b \tag{17}$$

The pairwise structure component length similarity scores are summed to form an overall structure component length similarity score for each pair, with the sum then normalized over the number of blocks in the structure using the equation

$$SCLS_{i,j} = \frac{ST_{tot}}{CW_{tot}} \tag{18}$$

where i and j are structure indices. An example of this calculation is offered in Figure S2b.

To determine an overall score for a bin in terms of structure similarity, the SCLS scores are summed over all pairwise comparisons and normalized by the number of pairwise comparisons in the bin.

$$SCLS'_{tot} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} SCLS_{i,j}}{p} \quad (19)$$

Depending on the descriptor format, portions of the structures (and/or the entire structures) in the RNAMotif output file may contain scores for thermodynamic stability calculated using the function *efn*. When evaluating structure similarity using our algorithm, these *efn* values can also be used for structure comparison. To derive a *structure thermodynamic stability similarity* score (*EFN*) for a pair of structures, the difference in *efn* values between two portions of an overall structure is calculated and divided by the maximum of the two values (Figure S2c).

$$EFN_s = 1 - \frac{|(efn_{1s} - efn_{2s})|}{\max(|efn_{1s}|, |efn_{2s}|)} \quad (20)$$

where *s* is the portion of the structure receiving an *efn* score, efn_1 is the score for structure 1, efn_2 is the score for structure 2. Our fitness function minimizes the difference in *efn* value. Comparisons of structure components form an *efn* score (EFN_{tot}) for the pair, and are normalized over the number of portions in the structure receiving an *efn* score

$$EFN_{i,j} = \frac{\sum_{s=1}^{n} EFN_s}{s} \quad (21)$$

where *s* is the number of portions receiving an *efn* score, *i* and *j* are structure indices. In order to determine a total *EFN* score for a bin, the *EFN* scores for all pairwise combinations are summed and divided by the number of pairwise combinations

$$EFN'_{tot} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} EFN_{i,j}}{p} \quad (22)$$

These three fitness terms are combined to form a value representative of the overall worth of a given bin. The importance of each of these fitness terms is associated with a weight that can be user-defined. The total fitness (F_{bin}) of any given bin is therefore defined as the sum of its weighted component scores

$$F_{bin} = \frac{w_1(SEQ'_{tot}) + w_2(SCLS'_{tot}) + w_3(EFN'_{tot})}{w_1 + w_2 + w_3} \quad (23)$$

where w_i are the weights associated with the terms sequence alignment (SEQ'_{tot}), structure component length similarity ($SCLS'_{tot}$) and stem *efn* similarity (EFN'_{tot}). For the experiments presented here $w_1 = 0.8$, $w_2 = 0.2$, $w_3 = 0.0$.

Characteristics of the Evolutionary Process

Following selection, the first generation (G) of evolution is completed. The P saved bins from the first round of selection are used as "parents" to generate O offspring bins with variation in the manner described above. All P and O solutions are pooled into a single population, scored, and selected. This process is iterated until user-specified termination criteria are satisfied (e.g. user-defined G, CPU or clock time, or use of statistical methods to determine the appropriate number of generations when the expected change in fitness per generation is close to zero and past which further computation will not result in a large change in fitness but will take an unreasonable amount of time). The number of total function evaluations (E_{tot}) during the evolutionary process can be calculated as

$$E_{tot} = (P \times O \times G) + P. \quad (24)$$

In the cases presented below, we applied a rule such that no two bins in the population at any single generation may share identical structure sets.

Figure S1. Nucleotide association matrix for scoring pairwise threaded nucleotide similarity. Any nucleotide symbol paired against a gap (□) will receive an initial gap opening penalty of –12. In sequence alignments of RNAMotif structures, it is possible to have the condition where a gap can be paired with a gap. In this case, the position is given a score of 0.

	A				
A	5	**C**			
C	□4	5	**G**		
G	□4	□4	5	**U**	
U	□4	□4	□4	5	□
□	□4	□4	□4	□4	0

Figure S2a. Overview of *nucleotide sequence similarity* scoring. Two structures in a bin are chosen for pairwise nucleotide alignment. Components (h5, ss, h3) without any sequence in the structure (i.e. a missing bulge; represented as "**.**") are treated as gaps during sequence alignment. Using the scoring matrix in Figure S2, scores (*SS*) are generated for each component block (numbered 1-7) of the structure. User-defined weights (*CW*) are also associated with each component and the sums for these values are determined. The sequence scores are multiplied by the component scores to generate a weighted score (*SS'*). A final *SEQ* score is generated by dividing the sum of the weighted score by the sum of the component weights and dividing by the length of the longest structure.

RNAMotif output

```
gi|191071|-2.8|-2.2|1  72  20  ccc . cggg uguc ccug . ggg
gi|191071|-2.6|-1.4|1  285 18  ccc . cug uguc uag c ggg
```

Nucleotide Alignment

	$h5_1$	ss	$h5_2$	ss	$h3_2$	ss	$h3_1$
Structure 1	ccc	-	cggg	uguc	ccug	-	ggg
Structure 2	ccc	-	cug-	uguc	ua-g	c	ggg

Block Index	1	2	3	4	5	6	7		Sum
Sequence Score (*SS*)	15.0	0.0	-10.0	20.0	-19.0	-16.0	15.0	=	
Component Weight(*CW*)	1.2	1.0	1.2	1.0	1.2	1.0	1.2	=	7.8
Weighted SS (*SS'*)	18.0	0.0	-12.0	20.0	-22.8	-16.0	18.0	=	5.2

$$SEQ_{12} = \frac{\sum SS'/\sum CW}{\max(L_1, L_2)} = (5.2/7.8)/18 = 0.66/18 = 0.04$$

Figure S2b. Overview of *structure component length similarity* scoring. Two structures in a bin are chosen for pairwise nucleotide alignment. Components (h5, ss, h3) without any sequence in the structure (i.e. a missing bulge; represented as "**.**") are maintained during structure alignment. Scores (*ST*) are generated for each component block (numbered 1-7) of the structure. User-defined weights (*CW*) are also associated with each component and the sums for these values are determined. The structure scores are multiplied by the component scores to generate a weighted score (*ST'*). A final *SCLS* score is generated by dividing the sum of the weighted score by the sum of the component weights. For structure comparison, a "**.**" symbol is given a value of 0.

RNAMotif output

```
gi|191071|-2.8|-2.2|1  72  20  ccc . cggg uguc ccug . ggg
gi|191071|-2.6|-1.4|1  285 18  ccc . cug uguc uag c ggg
```

Nucleotide Alignment

	h5	ss	h5	ss	h3	ss	h3
Structure 1	ccc	.	cggg	uguc	ccug	.	ggg
Structure 2	ccc	.	cug	uguc	uag	c	ggg

Block Index	1	2	3	4	5	6	7	Sum
min B	0	0	3	3	3	0	0	
max B 3	3	6	6	6	3	3		
Structure Score (ST)	1	1	0.66	1	0.66	0.66	1	
Component Weight (CW)	1.2	1.0	1.2	1.0	1.2	1.0	1.2	7.8
Weighted Score (ST')	1.2	1.0	0.79	1	0.79	0.66	1.2	6.6

$SCLS_{12} = \sum ST' / \sum CW = 6.6/7.8 = 0.85$

Figure S2c. Overview of pairwise *structure thermodynamic stability similarity* scoring. Two structures in a bin are chosen for pairwise efn scoring. For each structure in the case above, two *efn* scores (EFN_a and EFN_b) are provided in the RNAMotif output file for different portions of each structure. The similarity between these values is compared pairwise, where *efn* similarity is maximized. The similarities are summed and divided by the number of *efn* components to generate a final *EFN* score for the pair.

RNAMotif output

```
gi|191071|-2.8|-2.2|1  72  20  ccc . cggg uguc ccug . ggg
gi|191071|-2.6|-1.4|1  285 18  ccc . cug uguc uag c ggg


EFN Structure 1a = blocks 3 through 5 = -2.2
EFN structure 1b = blocks 1 through 7 = -2.8
EFN Structure 2a = blocks 3 through 5 = -1.4
EFN Structure 2b = blocks 1 through 7 = -2.6
```

$$EFN_a = 1 - \left(\frac{|-2.2| - |-1.4|}{2.2} \right) = 0.64$$

$$EFN_b = 1 - \left(\frac{|-2.8| - |-2.6|}{|-2.8|} \right) = 0.93$$

$$EFN_{1,2} = \frac{EFN_a + EFN_b}{s} = \frac{1.57}{2} = 0.79$$