Integrating chemical footprinting data into RNA secondary structure prediction

Authors: Kourosh Zarringhalam, Michelle M. Meyer, Ivan Dotu, Jeffrey Chuang, Peter Clote

Published in *PLoS ONE*, vol. 7, no. 10, October 2012

OPEN ACCESS Freely available online

PLOS ONE

Integrating Chemical Footprinting Data into RNA Secondary Structure Prediction

Kourosh Zarringhalam[¤a], Michelle M. Meyer, Ivan Dotu, Jeffrey H. Chuang[¤b], Peter Clote*

Department of Biology, Boston College, Chestnut Hill, Massachusetts, United States of America

Abstract

Chemical and enzymatic footprinting experiments, such as shape (selective 2′-hydroxyl acylation analyzed by primer extension), yield important information about RNA secondary structure. Indeed, since the 2′-hydroxyl is reactive at flexible (loop) regions, but unreactive at base-paired regions, shape yields quantitative data about which RNA nucleotides are base-paired. Recently, low error rates in secondary structure prediction have been reported for three RNAs of moderate size, by including base stacking pseudo-energy terms derived from shape data into the computation of minimum free energy secondary structure. Here, we describe a novel method, RNAsc (*RNA soft constraints*), which includes pseudo-energy terms for each nucleotide position, rather than only for base stacking positions. We prove that RNAsc is *self-consistent*, in the sense that the nucleotide-specific probabilities of being unpaired in the low energy Boltzmann ensemble always become more closely correlated with the input shape data after application of RNAsc. From this mathematical perspective, the secondary structure predicted by RNAsc should be 'correct', in as much as the shape data is 'correct'. We benchmark RNAsc against the previously mentioned method for eight RNAs, for which both shape data and native structures are known, to find the same accuracy in 7 out of 8 cases, and an improvement of 25% in one case. Furthermore, we present what appears to be the first direct comparison of shape data and in-line probing data, by comparing yeast asp-tRNA shape data from the literature with data from in-line probing experiments we have recently performed. With respect to several criteria, we find that shape data appear to be more robust than in-line probing data, at least in the case of asp-tRNA.

Citation: Zarringhalam K, Meyer MM, Dotu I, Chuang JH, Clote P (2012) Integrating Chemical Footprinting Data into RNA Secondary Structure Prediction. PLoS ONE 7(10): e45160. doi:10.1371/journal.pone.0045160

Editor: Cynthia Gibas, University of North Carolina at Charlotte, United States of America

Received April 17, 2012; **Accepted** August 16, 2012; **Published** October 16, 2012

Funding: No current external funding sources for this study.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: peter.clote@bc.edu

¤a Current address: Department of Mathematics, University of Massachusetts Boston, Boston, Massachusetts, United States of America
¤b Current address: The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut, United States of America

Introduction

RNA is an important biomolecule, known to play both an *information carrying* and a *catalytic* role. RNA plays roles in numerous biological processes, including *retranslation* of the genetic code (selenocysteine insertion, ribosomal frameshift), transcriptional and translational gene regulation, temperature-dependent allosteric regulation, chemical modification of specific nucleotides in the ribosome, regulation of alternative splicing, apparent regulation of the formation of heterochromatin, etc. (See [1] for a recent review on the analysis of sequence and structure of such noncoding RNA.) Since the function of non-coding RNA largely depends on its structure and since it is believed that RNA plays many yet undiscovered roles in cellular processes, it is important to determine the structure of RNA.

A secondary structure for a given RNA nucleotide sequence $a_1, \ldots, a_n$ is a set S of base pairs (i,j), such that a_i, a_j forms either a Watson-Crick or GU (wobble) base pair, and such that there are no *base triples* or *pseudoknots* in S. In this context, a base triple in S consists of two base pairs (i,j), $(i,\ell) \in S$ or (i,j), $(k,j) \in S$. A pseudoknot in S consists of two base pairs (i,j), $(k,\ell) \in S$ with $i < k < j < \ell$. Although it is NP-hard [2] to compute the minimum free energy (MFE) tertiary (or even pseudoknotted) structure of RNA [3], the MFE secondary structure can be computed in time that is cubic in the input sequence length [4]. Moreover, it is widely believed that RNA folds in a hierarchical fashion [5–8], with the secondary structure acting as a scaffold for tertiary structure, although this is not universally accepted [9].

RNA secondary structure can be predicted by Zuker and Stiegler's algorithm [4], implemented in mfold [10], RNAfold [11], and RNAstructure [12]. This algorithm uses dynamic programming with free energy parameters from the Turner energy model [13] to compute the minimum free energy (MFE) structure.

A first step towards integrating chemical/enzymatic probing data was taken by Mathews et al. [14], where Zuker and Stiegler's algorithm was modified to support *hard constraints* reflecting the experimental data. In particular, given an RNA sequence, the software RNAstructure [14] computed the minimum free energy (MFE) secondary structure subject to user-defined constraints, such as stipulating that particular nucleotides remain unpaired, that pairs of specific nucleotides form a base pair, etc. Mathews et al. reported that the MFE structure prediction with (hard) constraints corresponding to chemical modification (1-cyclo-hexyl-3-(2-morpholinoethyl) carbodiimide metho-p-toluene sulfo-nate, dimethyl sulfate, and kethoxal) yielded an improvement in base-pair accuracy for 5S rRNA of *E. coli* from 26.3% to 86.8%

[14]. (See [15] for more remarks and a less optimistic evaluation of RNAstructure with hard constraints on 16S rRNA.)

Chemical/enzymatic probing data is probabilistic in nature, as exemplified in pars footprinting data [16]. Rarely is it absolutely clear that certain positions are unpaired, or that certain base pairs are formed; instead, there is a certain probability of these events. In moving away from error-prone *hard* constraints, Deigan et al. [15] took a second step of incorporating shape (selective 2′-hydroxyl acylation analyzed by primer extension) data [17,18], whose numerical values (continuously) range from 0 to approximately 2.2, by incorporating a *pseudo free energy* for base stacking into the Zuker algorithm. The pseudo free energy term in [15] was defined to be

$$\Delta G_{\text{SHAPE}}(i) = m\ln\left(\text{SHAPE } reactivity(i)+1\right)+b \qquad (1)$$

where $m = 2.6$ kcal/mol and $b = -0.8$ kcal/mol, for each position i occurring in a base pairing stack; if i is unpaired, then no pseudo free energy is added. (The position i is in a base pairing stack if $(i,j),(i+1,j-1)$ are base pairs, or if $(i,j),(i-1,j+1)$ are base pairs belonging to the secondary structure. For base pairs (i,j) that are surrounded by base pair neighbors $(i-1,j+1)$ and $(i+1,j-1)$, the pseudo-energy term is applied twice.) The resulting modified version of Zuker and Stiegler's algorithm, as implemented in RNAstructure was reported to yield secondary structure prediction accuracies of up to $96-100\%$ for three moderate-sized RNAs ($75-155$ nt) and for 16S rRNA (≈ 1500 nt). Wilkinson et al. [19] later described a model for the secondary structure of the HIV-1 genome, as computed by RNAstructure with shape pseudo energies defined in equation 1. If correct, this is a remarkable feat, given that the size of the HIV-1 genome is generally just under 10,000 nt (see http://www.hiv.lanl.gov), hence several times larger than the ribosome, whose crystal structure was only determined after years of painstaking work (the large unit, PDB code 1FFK [20], of the ribosome of *Haloarcula marismortui* consists of a 23S chain of length 2,922 nt and a 5S chain of 122 nt).

One issue with this approach is that it takes into consideration shape data only for base-stacked positions, i.e., a pseudo free energy term corresponding to shape data is applied at positions where a stacked base pair occurs, but not where nucleotides are unpaired. By ignoring shape data for unpaired nucleotide positions, this approach can thus bias structure prediction to form base pairs even at positions, which shape data may suggest are flexible. Indeed the expected distance of predicted base pairing probabilities computed by RNAstructure with shape values increases after the incorporation of the shape pseudo energy terms (see Table 1). (As later defined, RNAstructure and RNAsc both compute the probability $p_{i,j}$ that base pair (i,j) belongs to a structure in the low energy Boltzmann ensemble. Since the pseudo energy model for shape data incorporation is different in RNAstructure and RNAsc, the base pairing probabilities and Boltzmann low energy ensembles may be different.) In contrast to the pseudo energies of RNAstructure, our algorithm RNAsc, will always shift the distribution of conformations towards the shape measurements (see Methods for a mathematical proof).

Nonetheless, MFE dynamic programming methods that incorporate high throughput chemical/enzymatic footprinting data can yield important insights into the structure and function of RNA molecules, much faster than the labor-intensive X-ray diffraction methods.

The motivation for our work is to develop a method that incorporates chemical/enzymatic footprinting data in a *self-consistent* manner. In particular, given experimental data of the form $\mathbf{q}^s = (q_1^s, \ldots, q_n^s)$, where q_i^s is the experimental probability that the ith nucleotide is *unpaired* (or, more accurately, in a flexible region, as witnessed by high shape reactivity), our goal is to develop an algorithm incorporating footprinting data such that the *recalculated* probabilities $\mathbf{q}^* = (q_1^*, \ldots, q_n^*)$ are guaranteed to be closer to the experimental measurements. If our algorithm is self-consistent in this manner, then we have strong mathematical evidence that the partition function computation and hence the MFE computation are both as correct as is the shape data. In contrast to the pseudo energies of RNAstructure, we prove that our algorithm RNAsc is self-consistent, and on average, the ensemble of low energy secondary structures produced by our method yields a footprinting pattern that closely resembles the pattern from input experimental shape data. We benchmark our method against the RNAstructure program [19] on eight RNAs, for which shape data and native structures are both available. The secondary structure predictions from our method and from RNAstructure are fairly similar and both significantly improve secondary structure prediction without incorporation of footprinting data (e.g. mfold, RNAfold). However, the expected distance of the computed probabilities with the shape data is lower in our method for all the test cases. It is worth noting that the mistakes in the predicted secondary structure usually occur in positions where the shape data might be inaccurate, or where the native structure and shape data structures could be somewhat different, due to quite different temperatures required by each experimental protocol. Recent studies have shown that different experimental mapping approaches can provide complementary structural information [21]. Thus, we additionally performed in-line probing [22,23] on asp-tRNA, in order to compare the results of shape and in-line probing in the context of our algorithm. The source code of RNAsc as well as a web server is available at http://bioinformatics.bc.edu/clotelab/RNAsc/.

Methods

In-line probing experiments

DNA oligonucleotides for the sequence and its reverse complement were purchased from MWG Operon; remaining reagents were obtained from Sigma-Aldrich. DNA oligonucleotides were annealed to create templates for T7 polymerase transcription, and the transcription products were purified by denaturing PAGE and eluted in 10 mM Tris-HCl (pH 7.5 at 23°C), 200 mM NaCl and 1 mM EDTA. Following in-line probing protocols designed by the Breaker Lab [22,23], synthesized RNA molecules were dephosphorylated using alkaline phosphatase (Roche Diagnostics) and radiolabeled with [g-32P]ATP and T4 polynucleotide kinase (NEB) according to the manufacturers instructions. Spontaneous transesterification reactions using PAGE-purified, 5′ endlabeled RNAs were assembled as described in [23]. Incubations were performed for approximately 40 h at 25°C in 10-uL volumes containing 50 mM Tris-HCl (pH 8.3 at 23°C), 20 mM MgCl2, 100 mM KCl and ≈ 5 nM RNA. RNA fragments resulting from spontaneous transesterification were resolved by denaturing 10% PAGE, and imaged with a Molecular Dynamics STORM PhosphorImager. Quantification of gels were performed using SAFA (Semi-Automated Footprinting Analysis) [24]. In-line probing experiments were repeated an additional two times, resulting in gels with comparable data (data not shown). Fig. 1 is an image of the in-line probing gel for yeast asp-tRNA.

Computational methods

Briefly stated, our algorithm, RNAsc (*RNA soft constraints*), consists of a preprocessing step, that normalizes shape data to the

Table 1. Benchmark results.

Secondary structure prediction accuracy

RNA	len	test	(A)	(B)	(C)	RNA	len	test	(A)	(B)	(C)
asp-tRNA	75	sens.	1.00	1.00	0.76	phe-tRNA	76	sens.	1.00	0.75	0.95
		ppv	1.00	1.00	0.76			ppv	0.95	0.71	0.95
		ave ent.	0.21	0.17	0.27			ave ent.	0.2	0.17	0.46
		str. div.	19.53	17.17	22.60			str. div.	11.37	9.38	34.37
		edist.	23.7	61.77	24.9			edist.	29.51	61.77	33.68
HCV IRES	95	sens.	0.96	0.96	0.96	5S rRNA	120	sens.	0.94	0.94	0.26
		ppv	1.00	1.00	1.00			ppv	0.82	0.82	0.22
		ave ent.	0.05	0.06	0.27			ave ent.	0.30	0.17	0.27
		str. div.	3.20	3.57	21.45			str. div.	46.93	20.70	32.90
		edist.	31.36	52.48	36.53			edist.	42.57	54.01	46.41
P546	155	sens.	0.95	0.96	0.43	glycine	162	sens.	0.92	0.92	0.70
		ppv	0.96	0.98	0.44			ppv	0.84	0.84	0.61
		ave ent.	0.18	0.12	0.38			ave ent.	0.11	0.05	0.30
		str. div.	27.7	14.05	66.50			str. div.	15.14	5.13	44.16
		edist.	41.36	131.77	56.11			edist.	53.90	115.55	60.29

A comparison of three secondary structure prediction algorithms, using shape data from Deigan et al. [15] for the three RNA molecules, yeast aspartyl tRNA (asp-tRNA), hepatitis C virus internal ribosomal entry site (HCV IRES), and the P546 domain from the bI3 group I intron (P546), along with shape data from [26] for three additional RNA molecules, *E. coli* phenylalanine tRNA (phe-tRNA), *E. coli* 5S ribosomal RNA (5S rRNA), and *F. nucleatum* glycine riboswitch (glycine). The benchmark results are tabulated for (A) RNAsc+shape, (B) RNAstructure+shape, and (C) RNAstructure (with no shape data). Sensitivity $=\frac{TP}{TP+FN}$ is abbreviated by sens., positive predictive value $=\frac{TP}{TP+FP}$ is abbreviated by ppv. The average pointwise entropy, Morgan-Higgs structural diversity, and the expected distance of the computed probabilities to the probing data are abreviated by ave ent., str. div., and edist., respectively. Not shown: results for medloop and *V. vulnificus* adenine riboswitch (1Y26), for which all three methods have optima sensitivity and ppv values of 1.0.

doi:10.1371/journal.pone.0045160.t001

range [0,1], followed by a computation of the minimum free energy [resp. partition function], which incorporates pseudo-energy terms [resp. Boltzmann factors of pseudo-energy terms] for each nucleotide position. We begin by discussion of the normalization of shape data.

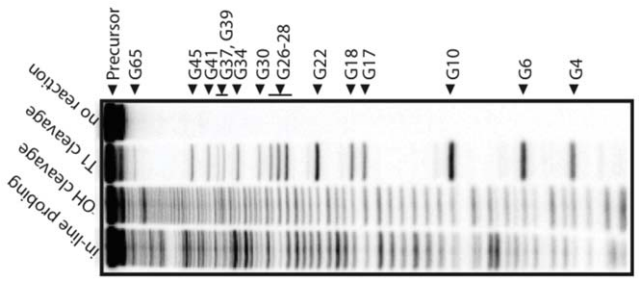

Figure 1. In-line probing. Spontaneous cleavage pattern resulting from in-line probing of yeast asp-tRNA, nucleotides with larger backbone flexibility will have higher rates of cleavage and thus bands of greater intensity. Lanes for no reaction, T1 RNase (cleavage following only guanosines), and partial hydroxyl cleavage (-OH, cleavage after each base) are indicated. Due to the high resolution of the gel, double bands appear for nucleotides 2–9. These bands correspond to RNA molecules where the $2'-3'$ cyclic phosphate intermediate has hydrolyzed to leave either no phosphate, or a mixture of 2′- and 3′-phosphate products which migrate more quickly on the gel. Quantifcation of these positions combined the bands corresponding to both products. The precursor RNA and T1 RNase cleavage products are marked. Not all guanosines show cleavage due to retention of secondary structure at 5 M urea and elevated temperature.

doi:10.1371/journal.pone.0045160.g001

Normalization of shape. In experiments reported by the Weeks Lab [25] as well as the Das Lab [26], shape reactivities range from 0 to roughly 2.2. Large reactivities suggest that the position is unpaired; small reactivities suggest that the position is base-paired. More specifically, nucleotides with shape reactivities ≥ 0.7 or $0.3-0.7$ are considered highly and moderately reactive, respectively [15]. The normalization is carried out in a piecewise linear fashion where 0.3 will be roughly mapped to 0.5. However, very low shape reactivities should not be mapped close to 0.5 either as it will bias the shape values toward unpaired nucleotides. For this reason the shape reactivity values <0.25 are linearly mapped to the interval [0.0,0.35), the reactivity values in [0.25,0.3) are linearly mapped to the interval [0.35,0.55), the reactivity values in [0.3,0.7) are linearly mapped to the interval [0.55,0.85), and lastly, the reactivities ≥ 0.7 are linearly mapped to the interval [0.85,1.0]. The selection of the threshold values are motivated by the moderate and high reactivity thresholds as reported in [15] and the examination of the cumulative distribution of the shape data (see File S1). The in-line probing data was normalized by mapping the outliers at the 0.05 and the 0.95 quantiles to 0.0 and 1.0 respectively and normalizing the rest of the data to [0.0,1.0] linearly. Fig. 2 shows a plot of the normalized and raw shape values as well as the normalization map.

Boltzmann weights. Let $a_1,\ldots,a_n$ be a fixed RNA sequence of length n, for which we are given normalized shape or in-line probing reactivity data $\mathbf{q}^s=(q_1^s,\ldots,q_n^s)$, where $q_i^s\in[0,1]$. For $x\in[0,1]$ and $i\in\{1,\ldots,n\}$, define the Boltzmann weight

$$w(x,i)=\exp((-\beta\cdot D(x,q^s{}_i))/RT) \qquad (2)$$

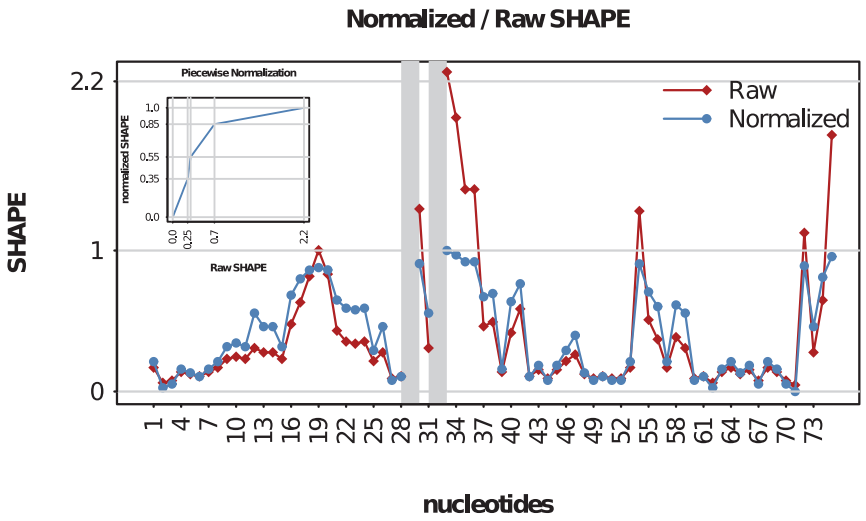

Figure 2. Normalization. Normalized (blue circles) and raw (red diamonds) shape values. Gray bars indicate the missing shape values. The subplots shows the piecewise normalization map.
doi:10.1371/journal.pone.0045160.g002

where β is a scaling parameter, and $D(x,q_i^s)=|x-q_i^s|$ measures the discrepancy between x and q_i^s. We will later incorporate Boltzmann weights in a *weighted* partition function Z^*, in a manner that reweights the ensemble of low energy conformations towards the shape data. When later used in recurrence relations for Z^*, the variable $x\in\{0,1\}$ is the indicator function for whether a position is unpaired (1) or paired (0) in a secondary structure under consideration. In the case of missing values, q_i^s may be assigned to 0.5, which represents no information about base pairing.

Weighting the partition function. In this section, we describe how to integrate Boltzmann weights into the computation of the partition function for secondary structures of a given RNA sequence.This allows us to compute the probability $p_{i,j}$ [resp. $p_{i,j}^*$] that (i,j) is a base pair in the Boltzmann ensemble of structures, where weights for shape or in-line probing have not [resp. have] been taken into consideration. As later explained, we will compare the probability $q_i^*=1-\sum_{i<j}p_{i,j}^*-\sum_{k<i}p_{k,i}^*$ with normalized shape reactivity q_i^s. Let $a[i,j]$ denote the subsequence $a_i,\ldots,a_j$ of a given, fixed RNA sequence $a_1,\ldots,a_n$ of length n. For $1\le i\le j\le n$, the McCaskill [27] partition function $Z(i,j)$ is defined by $Z(i,j)=\sum_S e^{-E(S)/RT}$, where the sum is taken over all secondary structures S of $a[i,j]$, $E(S)$ is the free energy of S with respect to the Turner energy model [13,28], $R=0.00198\ \frac{kcal}{mol\ K}$ is the universal gas constant, and T absolute temperature. The goal of the current paper is to integrate the previously defined weights into the partition function. We first require some notation. Here, we write Z^*,ZB^*, etc. instead of the more cumbersome notation $Z_{\mathbf{q}^s},ZB_{\mathbf{q}^s}$, etc. Thus Z^*,ZB^* etc. depend on the normalized footprinting data $\mathbf{q}^s=(q_1^s,\ldots,q_n^s)$, although $\mathbf{q}^s$ will not be explicitly mentioned.

Definition 1 (Weighted partition function). *Define*

- $Z^*(i,j)$: weighted partition function over all secondary structures of $a[i,j]$.
- $ZB^*(i,j)$: weighted partition function over all secondary structures of $a[i,j]$, which contain the base pair (i,j).
- $ZM^*(i,j)$: weighted partition function over all secondary structures of $a[i,j]$, subject to the constraint that $a[i,j]$ is part of a multiloop and has *at least* one component.
- $ZM1^*(i,j)$: weighted partition function over all secondary structures of $a[i,j]$, subject to the constraint that $a[i,j]$ is part of a multiloop and has *exactly* one component. Moreover, it is *required* that i base-pair in the interval $[i,j]$; i.e. (i,r) is a base pair, for some $i<r\le j$.

To compute partition function Z^*, we compute by dynamic programming $Z^*(i,j)$ for all $1\le i\le j\le n$ by increasing values of $j-i$. Structures on $a_i,\cdots,a_j$ can be subdivided into those for which j is unpaired in $[i,j]$, thus contributing $Z^*(i,j-1)$ times Boltzmann factor for j to be unpaired, and those for which j is paired with r for $r\in[i,j-\theta-1]$, thus contributing $Z^*(i,r-1)\cdot ZB^*(r,j)$ times Boltzmann factor for r,j to be paired. Subsequently $ZB^*(r,j)$ is computed by adding a contribution for all loops closed by base pair (r,j), i.e., hairpins, bulges, internal loops and multi loops whose latter contribution is recursively computed by jultiloop partition functions $ZM1^*(r,j)$ and $ZM^*(r,j)$. In essence, we apply Boltzmann weights to each nucleotide position k, while accounting for a distinct weight depending on whether k is paired or unpaired in the structure S under consideration: weight $\exp((-\beta\cdot D(1,q^s{}_i))/RT)$ if k is unpaired in S, weight $\exp((-\beta\cdot D(0,q^s{}_i))/RT)$ if k is base-paired in S. If all weights were set to 1, then the weighted partition function would be equivalent to the classic partition function. Similar forms of rearranging and reweighting of the partition function have been applied in the context of single stranded RNA binding proteins [29]. Details now follow. It will be expedient to define the function $F(i,j)=\Pi_{k=i}^j w(1,k)$, which represents the weight corresponding to a *loop* region in which $i,i+1,\ldots,j$ are unpaired. For $j<i$, $F(i,j)=1$, while for $i<j$,

$$F(i,j)=\exp\left(\frac{-\beta\times\sum_{k=i}^j D(1,q_k^s)}{RT}\right). \qquad (3)$$

In the base case, we define $Z^*(i,j)=F(i,j)$ and $ZB^*(i,j)=ZM1^*(i,j)=ZM^*(i,j)=0$ for all $i\le j\le i+\theta$, where θ is the minimum number of unpaired bases in a hairpin loop (generally $\theta=3$). In the inductive case, where $i+\theta<j$, we define

$$Z^*(i,j) = w(1,j)Z^*(i,j-1) + \sum_{r=i}^{j-\theta-1} w(0,r)w(0,j)Z^*(i,r-1)ZB^*(r,j). \quad (4)$$

Note that in the above equation $w(0,r)$ and $w(0,j)$ correspond to the weights for the nucleotides r and j being paired, but not necessarily to one another. If extra information on the pairing status of the nucleotides is available, (e.g., as in 'mutate and map' experiments [30]), these weights may be corrected accordingly to reflect the weight for the pairing of the rth and the jth nucleotides. Let $\mathcal{H}$ denote the free energy of a hairpin and let $\mathcal{I}$ denote the free energy of an internal loop (which combines the cases of stacked base pair, bulge and proper internal loop). The free energy for a multiloop containing N_b base pairs and N_u unpaired bases is given by the affine approximation $a+bN_b+cN_u$. The weighted partition function closed by base pair (i,j) is given by

$$\begin{aligned} ZB^*(i,j) &= e^{-\mathcal{H}(i,j)/RT}F(i+1,j-1) \\ &+ \sum_{i<\ell<r<j} w(0,\ell)w(0,r)F(i+1,\ell-1)F(r+1,j-1) \\ &\times e^{-\mathcal{I}(i,\ell,r,j)/RT}ZB^*(\ell,r) + e^{-(a+b)/RT} \\ &\times \sum_{r=i+1}^{j-\theta-2} w(0,r)ZM^*(i+1,r-1)ZM1^*(r,j-1), \end{aligned} \quad (5)$$

The weighted multiloop partition function with a single component and where position i is required to base-pair in the interval $[i,j]$ is given by

$$ZM1^*(i,j) = \sum_{r=i+\theta+1}^{j} w(0,r)F(r+1,j)ZB^*(i,r)e^{-c(j-r)/RT}. \quad (6)$$

Finally, the weighted multiloop partition function with one or more components, having no requirement that position i base-pair in the interval $[i,j]$ is given by

$$\begin{aligned} ZM^*(i,j) &= \sum_{r=i}^{j-\theta-1} w(0,r)F(i,r-1)ZM1^*(r,j)e^{-(b+c(r-i))/RT} \\ &+ e^{-b/RT}\sum_{r=i+\theta+1}^{j-\theta-1} w(0,r)ZM^*(i,r-1)ZM1^*(r,j) \end{aligned} \quad (7)$$

The weighted Boltzmann probability of base pair (i,j) is defined by

$$p^*_{i,j} = \frac{1}{Z^*(1,n)}\sum_{(i,j)\in S}\omega_{\mathbf{q}^s}(S)\cdot e^{-E(S)/RT} \quad (8)$$

where $\omega_{\mathbf{q}^s}(S) = \exp\left((-\beta d_{\mathbf{q}^s}(S))/RT\right)$ – see Methods. Following Zuker [31], the inner and outer partition function is computed, from which we easily obtain $p^*_{i,j}$.

The minimum free energy (MFE) structure can be computed by a modification of McCaskill's algorithm [27], where the weighted partition function is modified by replacing summations by minimizations, products by sums, and replacing the weights by $(x,i) = -RT\ln w(x,i) = \beta\cdot D(x,q^s_i)$. Although we did implement this algorithm, it does not include energy contributions for stacked, single-stranded nucleotides (dangles) or coaxial stacking, both known to be important in improving secondary structure prediction accuracy. For this reason, we modified the source code of RNAstructure, for both the MFE as well as the partition function computation which implements dangles and coaxial stacking. See File S1 for details. As in [15], the value of the scaling parameter β, is determined by a search to optimize positive predictive value and sensitivity.

Measures of uncertainty in the predicted low-energy ensemble of conformations. Pointwise entropy and Morgan-Higgs structural diversity [32] were used as measures of uncertainty in the prediction of the secondary structure. The poinwise entropy is defined as follows. For each *fixed* i in $1,\ldots,n$, define probability distribution $r^*_{i,j}$ on $j\in\{1,\ldots,n+1\}$ by setting $r^*_{i,j} = p^*_{i,j}$ for $1\le i<j\le n$, $r^*_{i,j} = p^*_{j,i}$ for $1\le j<i\le n$, and $r^*_{i,n+1} = 1-\sum_{j=1}^{n} p^*_{i,j}$. Pointwise entropy $H_i = -\sum_{j=1}^{n+1} r^*_{i,j}\ln r^*_{i,j}$ measures the *variability* in nucleotides found to be base-paired with i in the Boltzmann ensemble of low energy structures. The pointwise entropy without the probing data is computed similarly using the probabilities $p_{i,j}$. To reflect the nature of the probing data, we modified this definition as follows. Define the binary pointwise entropy at position i by $H_i = -r^*_{i,n+1}\ln(r^*_{i,n+1}) - (1-r^*_{i,n+1})\ln(1-r^*_{i,n+1})$. Binary entropy measures the uncertainty in the ith nucleotide being paired or unpaired, reflecting the signal detected by probing data. Similar computations were done with $p_{i,j}$ (the base pairing probabilities without the integration of the weights). The Morgan-Higgs structural diversity is defined by $\langle D_{mh}\rangle = n - \sum_{i=1}^{n}\sum_{j=0}^{n} p^{*2}_{i,j}$, where $p^*_{i,0}$ is defined by $p^*_{i,0} = 1-\sum_{j=1}^{n} p^*_{i,j}$. Similar computations were done with $p_{i,j}$.

RNAsc is guaranteed to improve agreement with SHAPE data

In this section, we show that on average, the ensemble of low energy secondary structures produced by our method yields a footprinting pattern that more closely resembles the pattern from input experimental shape data; in particular, we prove that the expected distance from (normalized) shape data for the ensemble of low energy structures (our algorithm) is strictly less than the expected distance from shape data for the Boltzmann ensemble of low energy structures (McCaskill's algorithm). First, we require some definitions. All secondary structures S considered in this section will be tacitly assumed to be secondary structures of the RNA molecule $a_1,\ldots,a_n$. Each secondary structure S can be assigned a binary sequence $\{b_i\}_{i=1}^n$ so that $b_i = 0$ if the nucleotide a_i is paired and $b_i = 1$ otherwise. Given experimental shape data yielding probabilities $\mathbf{q}^s = (q^s_1,\ldots,q^s_n)$, where q^s_i is the probability that nucleotide i is *unpaired*, the distance of S to $\mathbf{q}^s$ is defined by:

$$d_{\mathbf{q}^s}(S) = \sum_{i=1}^{n}|b_i - q^s_i|. \quad (9)$$

The shape *weight* of S is defined to be

$$\begin{aligned} \omega_{\mathbf{q}^s}(S) &= \prod_{i=1}^{n}\exp((-\beta|b_i - q^s{}_i|)/RT) \\ &= \exp\left((-\beta d_{\mathbf{q}^s}(S))/RT\right). \end{aligned} \quad (10)$$

The weighted partition function then becomes

$$Z^* = \sum_S \omega_{\mathbf{q}^s}(S)\exp(-E(S)/RT). \quad (11)$$

The *Boltzmann probability* $P(S)$ of secondary structure S is defined by

$$P(S) = \frac{\exp(-E(S)/RT)}{Z} \quad (12)$$

and the *weighted Boltzmann probability* $P^*(S)$ is defined by

$$P^*(S) = \frac{\omega_{\mathbf{q}^s}(S)\exp(-E(S)/RT)}{Z^*} \quad (13)$$

Define the *critical distance* d_c by

$$d_c = -\frac{1}{\beta}\left(RT\ln\left(\frac{Z^*}{Z}\right)\right). \quad (14)$$

Note that d_c does not depend on any particular secondary structure S, although it does depend on $n,T,\beta,\mathbf{q}^s$ and of course the input RNA sequence $a_1,\ldots,a_n$. It follows from definitions that for any secondary structure S,

$$d_{\mathbf{q}^s}(S) \le d_c \Leftrightarrow P^*(S) \ge P(S) \quad (15)$$

and strict inequalities hold as well. Indeed, since the exponential function is increasing, we have $d_{\mathbf{q}^s}(S) \le d_c$ if and only if

$$\exp(-\beta d_{\mathbf{q}^s}(S)/RT) \ge \exp\left(\ln\left(\frac{Z^*}{Z}\right)\right) = \frac{Z^*}{Z}.$$

Multiplying each side by $P(S)$, the above inequality can be written as

$$\omega_{\mathbf{q}^s}(S)P(S) \ge P(S)\frac{Z^*}{Z},$$

from which (15) follows. Similarly,

$$d_{\mathbf{q}^s}(S) > d_c \Leftrightarrow P^*(S) < P(S). \quad (16)$$

Next, define the expected distance $\langle D\rangle$ between $\mathbf{q}^s$, obtained by normalizing shape data, and the ensemble of low energy structures as follows:

$$\langle D\rangle = \sum_S P(S)d_{\mathbf{q}^s}(S). \quad (17)$$

Similarly, define the SHAPE *weighted expected distance* $\langle D^*\rangle$ between $\mathbf{q}^s$ and the ensemble of low energy structures by

$$\langle D^*\rangle = \sum_S P^*(S)d_{\mathbf{q}^s}(S). \quad (18)$$

Let $0 \le d_1 < d_2 < \cdots < d_N$ represent the sorted distances $d_{\mathbf{q}^s}(S)$ between all secondary structures of $a_1,\ldots,a_n$, for given normalized SHAPE data $\mathbf{q}^s$. Here N denotes the total number of secondary structures. Note that there may be many distinct secondary structures that have a given distance d_i to $\mathbf{q}^s$; i.e. possibly many distinct S for which $d_{\mathbf{q}^s}(S) = d_i$. Let i_0 be the largest index i such that $d_i \le d_c$; it follows that $d_{i_0} \le d_c$ and $d_{i_0+1} > d_c$. Let A [resp. B] consist of those secondary structures S, such that $d_{\mathbf{q}^s}(S) \le d_c$ [resp. $d_{\mathbf{q}^s}(S) > d_c$]; in other words

$$A = \{S : d_{\mathbf{q}^s}(S) = d_i, \textit{for some } i \in \{1,\ldots,i_0\}\ \}$$

$$B = \{S : d_{\mathbf{q}^s}(S) = d_i, \textit{for some } i \in \{i_0+1,\ldots,N\}\ \}.$$

THEOREM 1: For any given RNA sequence $a_1,\ldots,a_n$ and normalized SHAPE data $\mathbf{q}^*$, $\langle D^*\rangle < \langle D\rangle$.

PROOF:

$$\langle D\rangle - \langle D^*\rangle = \sum_{S\in A} d_{\mathbf{q}^s}(S)(P(S)-P^*(S)) + \sum_{S\in B} d_{\mathbf{q}^s}(S)(P(S)-P^*(S))$$

$$> \sum_{S\in A} d_{i_0}(P(S)-P^*(S)) + \sum_{S\in B} d_{i_0}(P(S)-P^*(S))$$

$$= d_{i_0}\cdot\sum_S (P(S)-P^*(S)) = d_{i_0}\cdot\left(\sum_S P(S) - \sum_S P^*(S)\right) = d_{i_0}\cdot 0 = 0.$$

To justify the inequality, note that for $S\in A$, $P(S)-P^*(S) \le 0$, hence for $i\in\{1,\ldots,i_0\}$, we have $d_i\cdot(P(S)-P^*(S)) \ge d_{i_0}\cdot(P(S)-P^*(S))$. On the other hand, for $S\in B$, $P(S)-P^*(S) > 0$, hence for $i\in\{i_0+1,\ldots,N\}$, we also have $d_i\cdot(P(S)-P^*(S)) \ge d_{i_0}\cdot(P(S)-P^*(S))$. Finally, the last line follows from the fact that P and P^* are both probability distributions, hence $\sum_S P(S) = 1 = \sum_S P^*(S)$. This completes the proof that $\langle D\rangle > \langle D^*\rangle$.

The above theorem can be generalized; however, we first require some notation. The weighted partition function Z^*, weighted Boltzmann probability $P^*(S)$, and weighted expected distance D^* were respectively defined in Equations (11),(13), and (18). When we wish to make the weighting parameter β explicit, we instead write Z_β, P_β and D_β. The following theorem shows that as the parameter β increases, the expected distance to normalized shape data decreases:

THEOREM 2: For any given RNA sequence $a_1,\ldots,a_n$, normalized SHAPE data $\mathbf{q}^*$ and $0 \le \beta_1 \le \beta_2$, $\langle D_{\beta_1}\rangle \ge \langle D_{\beta_2}\rangle$; moreover, strict inequalities hold as well.

The proof the the theorem can be found in File S1.

Quadratic time computation of expected distance from SHAPE data

Given RNAsc parameter $0 \le \beta$, recall that we defined the β-expected distance $\langle D_\beta\rangle$ between $\mathbf{q}^s$, obtained by normalizing SHAPE data, and the ensemble of low energy structures by

$$\langle D_\beta\rangle = \sum_S P_\beta(S)d_{\mathbf{q}^s}(S). \quad (19)$$

In the main text, we wrote $\langle D\rangle$, instead of $\langle D_0\rangle$ when $\beta = 0$.

In trying to compute $\langle D_\beta\rangle$ by definition, we seemingly require the sum over exponentially many secondary structures, or at least to approximate this sum by summing over a reprentative sample of structures, sampled from the low energy ensemble. This is not

necessary. Here, we show how to compute $\langle D_\beta \rangle$ from the base pairing probabilities $p_\beta(i,j)$, thus leading to a quadratic time algorithm.

By definition,

$$\langle D_\beta \rangle = \sum_{i=1}^{n} \left(\sum_S P_\beta(S) \cdot I[i \text{ unpaired}] \cdot (1 - q^s{}_i) + \sum_S P_\beta(S) \cdot I[i \text{ base} - \text{paired}] \cdot (q^s{}_i - 0) \right)$$

where I is denotes the indicator function. Now for any fixed $i = 1, \ldots, n$,

$$\sum_S P_\beta(S) \cdot I[i \text{ unpaired}] \cdot (1 - q_i^s) + \sum_S P_\beta(S) \cdot I[i \text{ base} - \text{paired}] \cdot q_i^s$$

is equal to

$$\sum_S P_\beta(S) \quad (I[i \text{ unpaired}] \cdot (1 - q^s{}_i) + I[i \text{ base} - \text{paired}] \cdot q^s{}_i) \tag{20}$$

Since $q_i = \sum_S P(S) \cdot I[i \text{ unpaired}]$, it follows that Equation (20) is equal to

$$q_i(1 - q_i^s) + q_i^s(1 - q_i). \tag{21}$$

It follows that

$$\langle D_\beta \rangle = \sum_{i=1}^{n} q_i(1 - q_i^s) + q_i^s(1 - q_i)$$

The values $q_i = 1 - \sum_{i<k} p_\beta(i,k) - \sum_{k<i} p_\beta(k,i)$ are computed in quadratic time from McCaskill's algorithm, and subsequently stored in an array. If follows that $\langle D_\beta \rangle$ can be computed in quadratic time.

Since RNAstructure of Deigan et al. [15] takes unnormalized SHAPE data in the range from 0 to 2.2, we define the expected distance $D^\dagger(S)$ between *unnormalized* shape data and structure S to be

$$\langle D^\dagger(S) \rangle = \sum_{i=1}^{n} P^\dagger(S) \cdot \quad (I[i \text{ unpaired in } S] \cdot (2.2 - s_i) + I[i \text{ base} - \text{paired in } S] \cdot (s_i - 0)) \tag{22}$$

where s_i denotes the unnormalized shape data at position i. The expected distance $D^\dagger$ between *unnormalized* SHAPE data and the ensemble of low energy structures computed by RNAstructure with incorporated shape data by

$$D^\dagger = \sum_S P^\dagger(S) \cdot D^\dagger(S). \tag{23}$$

Scrutiny of the proof just given yields an efficient computation of

$$\langle D^\dagger(S) \rangle = \sum_{i=1}^{n} q_i(2.2 - s_i) + s_i(1 - q_i). \tag{24}$$

Since the approach in [15] only considers stacked base pairs, it seems very likely that $\langle D^* \rangle < \langle D^\dagger \rangle$, where $\langle D^\dagger \rangle$ denotes the expected distance from SHAPE data for the Boltzmann ensemble of low energy structures after the incorporation of the SHAPE pseudo energy terms as in [15]. Indeed, the expected distance we obtain between *u*nnormalized input shape data $\mathbf{q}^s = (q_1^s, \ldots, q_n^s)$ and the computed probabilities $\mathbf{q}^* = (q_1^*, \ldots, q_n^*)$ demonstrates this fact (see Table 1).

Results

In this section we present the benchmarking results for our algorithm RNAsc, a novel algorithm that recalibrates probing data as probabilities of nucleotides being unpaired and integrates this information as 'soft constraints' into the computation of minimum free energy secondary structure (see Methods). Furthermore, we present a direct comparison of in-line probing data and shape data for yeast asp-tRNA.

Analysis of SHAPE and in-line probing for structure prediction

In order to directly characterize how well shape data reflects RNA secondary structure, we compared normalized SHAPE data with base pairing status, as determined from crystallographic or NMR structures. We define SHAPE *distance* to equal the difference between *normalized* SHAPE reactivity (see Methods), scaled from 0 to 1 (see section Normalization of shape) and binary base pairing status, with 0 for paired, 1 for unpaired, as derived from NMR or crystal structure. Using SHAPE data for *S. cerevisiae* apartyl-tRNA [25], HCV IRES [15], bI3 group I intron p456 [33], *E. coli* phenylalanine-tRNA [26], *E. coli* 5S RNA [26], and *Fusobacterium nucleatum* glycine riboswitch [26], we computed SHAPE distance at each nucleotide. We observed that at many positions the SHAPE distance has an absolute value greater than 0.5, thus indicating a significant difference between SHAPE reactivity and the actual secondary structure. We refer to these positions as discrepancies. Over the the set of RNAs we examined, between $24-35\%$ of the total data corresponded to such discrepancies (Fig. 3 and File S1). Many factors can account for these discrepancies, including differences between the crystal structure and the ensemble of structures in solution, potential tertiary contacts, and differential reactivity to the chemical agent.

To assess whether an alternative experimental method might yield data that more accurately reflects the secondary structure, we performed in-line probing on the *S. cereviseae* aspartyl-tRNA, for which shape data is available [25]. Like SHAPE, in-line probing is a measure of backbone flexibility, where nucleotides in loops and other unpaired regions are generally more reactive than those that are base-paired [34]. In-line probing takes advantage of the spontaneous transesterification reactions responsible for RNA degradation that occur only when the $2'-O$ from one nucleotide and the $5'-O$ of the next align in a 180 degree conformation around the phosphate. This conformation does not occur in the A-form helix, thus protecting linkages within the helix from cleavage. In-line probing and shape are thus likely to yield similar, but not equivalent data [35].

Our analysis indicates that in-line probing and shape reactivity profiles are quite distinct from one another. See Fig. 4 for a

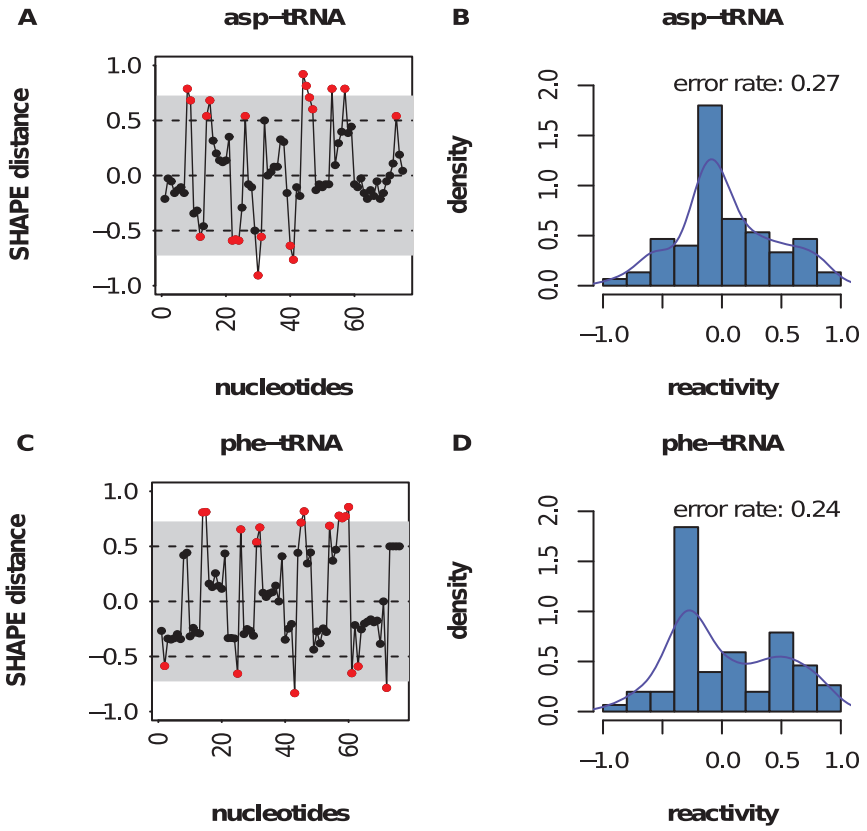

Figure 3. Shape discrepancies. Distribution of shape discrepancies in yeast asp-tRNA *(top)* and *E. coli* phe-tRNA *(bottom)*. shape data for asp-tRNA [resp. phe-tRNA] from the Weeks Lab [25] [resp. Das Lab [26]]. Using crystal structure as 'gold standard', red squares indicate locations where the absolute value of the difference of shape data and crystal structure (1 unpaired, 0 paired) exceeds 0.5. The plots on the right show the distribution of the discrepancy in shape as well as the error rate.
doi:10.1371/journal.pone.0045160.g003

comparison of shape and in-line probing profiles and File S1 for shape reactivity profiles of other RNA molecules.

The signal from in-line probing is significantly more diffuse than that from shape, and the error rate, as calculated above for shape, is significantly higher (27 vs. 36%). Thus shape is a better reflection of secondary structure than in-line probing, at least in the case of asp-tRNA.

Integrating shape and in-line probing data into our new algorithm RNAsc also shows that shape has an edge over in-line probing. The structures predicted by RNAsc for yeast asp-tRNA

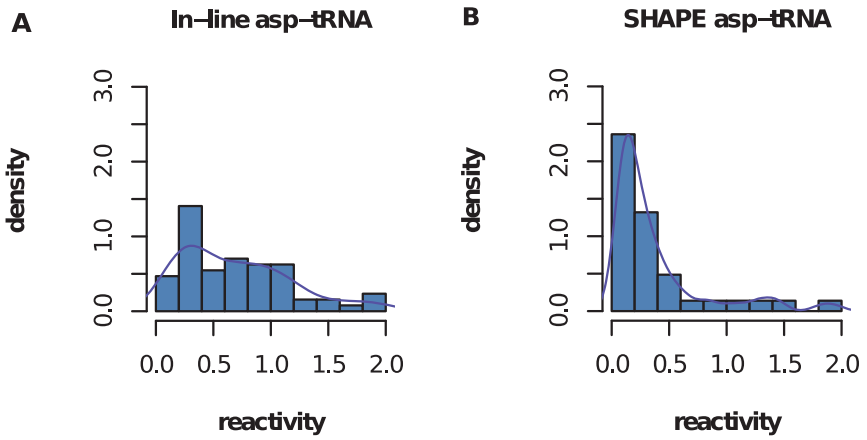

Figure 4. Comparison of In-line probing and shape. Distribution of reactivities of data from in-line probing *(A)* and shape *(B)*. In-line probing reactivities were determined using SAFA [24] and then normalized to range $[0,2.2]$, in order to be comparable with shape reactivities. Histograms suggest that in-line probing signal is more diffuse than that from shape. The fraction of base-pairs in asp-tRNA is 0.56 which could be used to estimate the threshold shape moderate reactivity.
doi:10.1371/journal.pone.0045160.g004

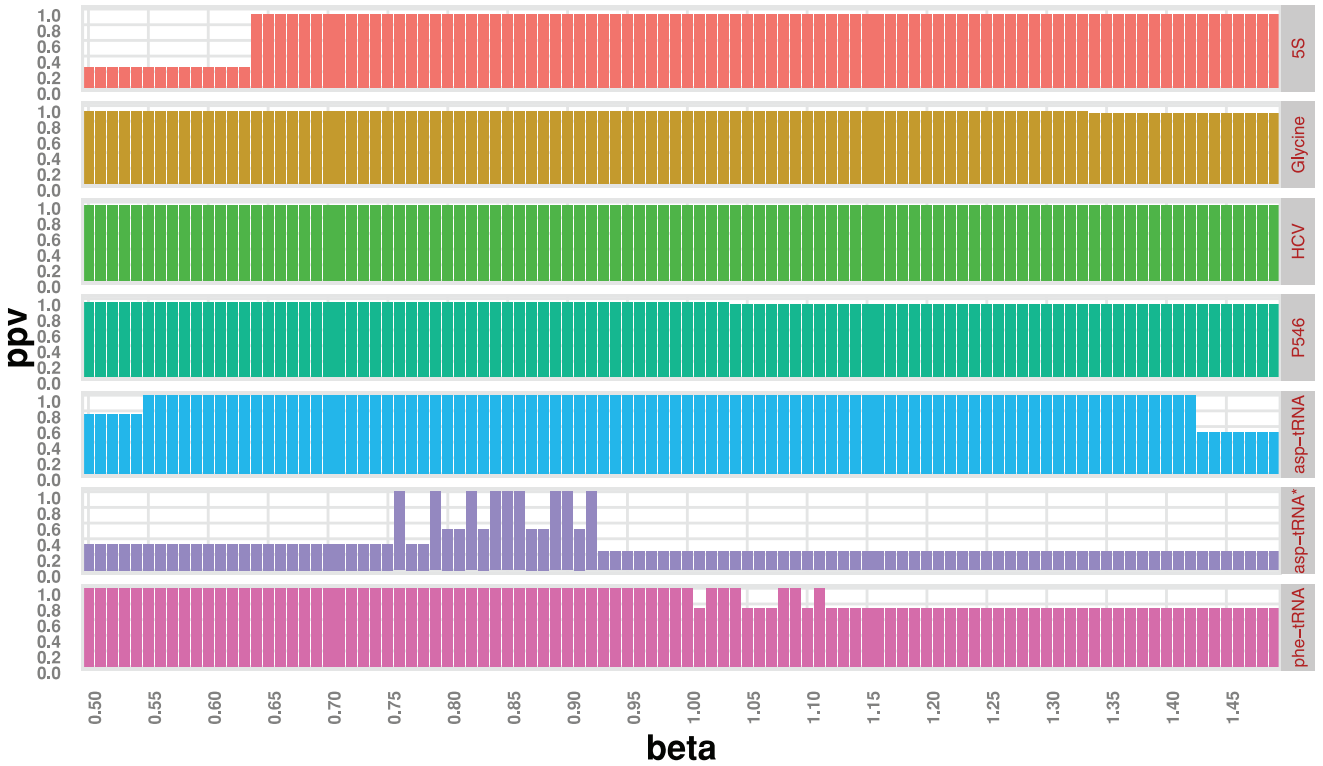

Figure 5. Optimal parameter value. The plots show heat maps displaying ppv ($\frac{TP}{TP+FP}$) as a function of parameter β for RNAsc with data from shape and in-line probing *(asp-tRNA*)*. Note the much larger area for good parameter choices when using shape data, rather than in-line probing data. This data suggests that shape data is more robust than in-line probing data, when used in computing MFE structure with RNAsc. Computations were done at 37°C.
doi:10.1371/journal.pone.0045160.g005

using in-line probing and shape data are both identical to the crystal structure. However, one measure of the robustness of the data in the context of our secondary structure prediction algorithm RNAsc is the range of the scaling parameter β over which the correct structure can be recovered. Recall that β is a weight parameter (see section Boltzmann Weights for details). We conducted a search for parameter β for yeast asp-tRNA, using both in-line probing data and SHAPE data. We found that when using in-line probing data, RNAsc produced the target structure for asp-tRNA only for a very narrow range of β, while when using shape data, this range was much larger (see Fig. 5). See Fig. 6 for a heat map of in-line vs. shape reactivity for asp-tRNA.

In a second analysis, we compared the pointwise entropy at each nucleotide using no data, shape data, and in-line probing

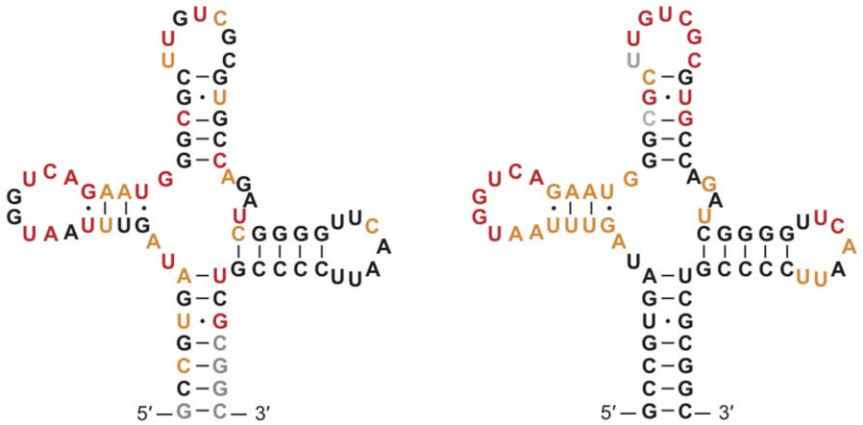

Figure 6. Heat maps of in-line probing and shape. Heat maps illustrating differences between in-line probing *(left)* and shape *(right)* analysis of the yeast asp-tRNA. Nucleotides are colored corresponding to cumulative activities described in Figure 3, where the least reactive 56% of bases are black (56% of bases are paired in the crystal structure), the most reactive 20% of bases are red, and the next most reactive 24% are yellow. Gray bases are bases for which there is no data available.
doi:10.1371/journal.pone.0045160.g006

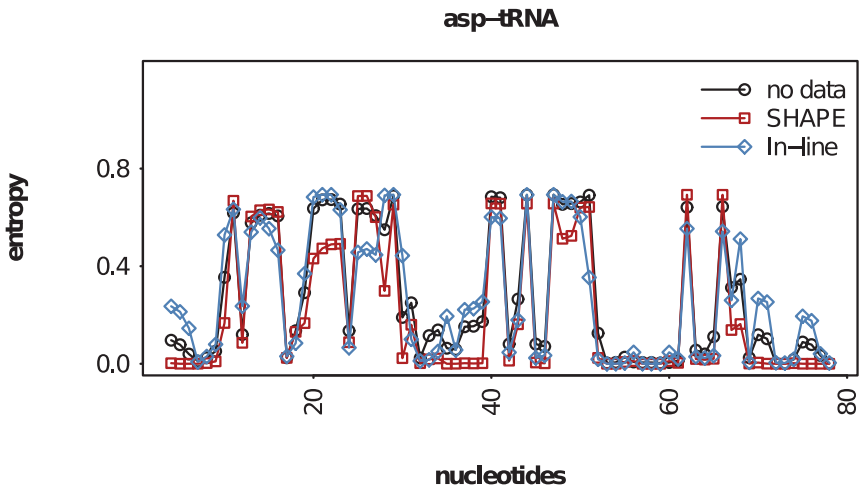

Figure 7. Pointwise entropies. Pointwise entropy of yeast asp-tRNA, computed from RNAsc using shape data (red squares), in-line probing (blue diamonds), and using no probing data (black circles). Average pointwise entropies: 0.210 (shape data), 0.267 (in-line probing), 0.269 (no data). As expected, by integrating either shape or in-line probing data into RNAsc, the variability (entropy) decreases; however, it appears that variability (entropy) is decreased more by shape than by in-line probing data – again, suggesting that shape data is more robust than in-line probing data when used with RNAsc.
doi:10.1371/journal.pone.0045160.g007

data (see Fig. 7). We observe that shape data decreases the average entropy more than in-line probing data. However, we also observe that there are positions where the in-line probing decreases the entropy more than shape, suggesting that combinations of different experimental approaches may be able to yield additional information.

Validation of RNAsc

Using SHAPE data from the Weeks Lab, we tested RNAsc on aspartyl-tRNA from *S. cerevisiae*, domain II of the hepatitis C virus internal ribosomal entry site (HCV IRES), and the P546 domain of the bI3 group I intron, from *E. coli*. Additionally, using shape data from the Das Lab, we tested RNAsc on *E. coli* phenylalanine tRNA (phe-tRNA), *E. coli* 5S ribosomal RNA (5S rRNA), and the glycine riboswitch from *F. nucleatum* with PDB code 3P49. As 'gold standard' structures, we used NMR structure for P546, and X-ray structures for remaining RNAs. Parameter used for RNAsc is $\beta = 0.89$, determined by search (see Fig. 5) to optimize sensitivity (proportion of true positives that are correctly identified) and positive predictive value (proportion of positive results that are true positives). Slippage of ± 1 [15,36] is *not* allowed, contrary to benchmarking results of some authors. Here, slippage [36] means that if base pair (i,j) is in the true structure, then the base pair (i,j) is counted as "correctly" predicted, if one of the base pairs $(i-1,j)$, $(i+1,j)$, $(i,j-1)$, $(i,j+1)$ appears in the predicted structure – we do *not* allow slippage in the results of this paper.

Table 1 presents a comparison of RNAsc with RNAstructure, including a comparison of structural variation in the ensemble of low energy structures. This variation is computed by pointwise entropy and Morgan-Higgs structural diversity (see Methods). The table shows that the low energy ensemble, as computed by RNAsc with integration of shape data, has intermediate variation between that computed by RNAstructure with and without shape data. The fact that RNAstructure with incorporated shape data computes an ensemble of structures with less variation appears to be expected, given the parameters used in the algorithm of Deigan et al. [15].

As explained in Deigan et al. [15], RNAstructure incorporates shape data by including a pseudo free energy term

$$\Delta G_{\text{SHAPE}}(i) = m \ln(\text{ SHAPE } \textit{reactivity} + 1) + b \qquad (25)$$

for a nucleotide position i. In the source code RNAstructure, it is clear that the pseudo free energy term $\Delta G_{shape}(i)$ is applied *only* for positions i involved in a stacked base pair. The optimal values for slope m and y-intercept b are obtained by grid search when maximizing structure prediction accuracy on certain known structures. Optimal slope and intercept values reported in [15] are $m = 2.6$ and $b = -0.8$ kcal/mol.

We now show that the smaller structural variation in the RNAstructure ensemble appears to be an artifact of the magnitude of parameters m,b. Consider the two most extreme cases: (1) position i in structure S is base-paired, but shape reactivity is a maximum, (2) position i in structure S is not paired, but shape reactivity is a minimum.

Suppose that position i is in a base-stacked region but the shape reactivity at position i is 2.2, a maximum, though there are sometimes shape reactivities larger than 2.2. With the default parameters for m,b, the pseudo free energy contribution of RNAstructure is $\Delta G_{shape}(i) = 2.6 \cdot \ln(2.2+1) - 0.8 = +2.22$, an energetic penalty. This penalty is quite large, given the fact that the largest (in absolute value) free energy contribution for base stacking is -3.3 kcal/mol [37]. Under the same assumptions, RNAsc would have a pseudo free energy of $\beta \cdot |q_i^s - 0| = 0.89 \cdot 1.0 = 0.89$, also an energetic penalty, yet much smaller than that of RNAstructure.

Suppose now that position i is in a loop region but the shape reactivity at position i is 0, the least possible value. Using the default parameters $m = 2.6, b = -0.8$ kcal/mol, the pseudo free energy contribution of RNAstructure, *if* applied in this case, would then be $\Delta G_{shape}(i) = 2.6 \cdot \ln(0+1) - 0.8 = -0.8$. This value, paradoxically, would be an energetic bonus, although the predicted structure disagrees with shape data! It is presumably for this reason that Deigan et al. do not apply any pseudo free energy term to nucleotide positions i located in a loop region. In contrast, under the same assumptions, RNAsc would have a pseudo free energy of $\beta \cdot |1 - q_i^s| = 0.89 \cdot (1-0) = 0.89$, again a

penalty – moreover, the same penalty of 0.89 kcal/mol is applied in each of the cases (1) and (2) just discussed.

From these illustrative examples, it is suggestive that structural *variability*, as measured by pointwise entropy and structural diversity, in the low energy ensemble calculated by RNAstructure is *higher* than that of the RNAsc low energy ensemble, due to the magnitude of the parameters m,b used in RNAsc.

Note that the average relative decrease in expected distance of the computed probabilities to shape data from RNAstructure to RNAsc is 48.9%. In fact the expected distance of the computed probabilities to shape *increases* for RNAstructure and *decreases* for RNAsc after the incorporation of shape in each case. Apart from the 'self-consistent' nature of our algorithm, not shared by RNAstructure, the demonstrable expected distance of the computed probabilities to shape data provided by our approach, indicates that we account more fully for the shape data. It is worth mentioning that for higher values of β the predicted Boltzmann probabilities q_i can be made to agree very closely with the experimental values q_i^* (strong self-consistency). Fig. 8 shows a plot of the expected distance of the computed probabilities to shape data for increasing values of β – see Methods for a proof. Note however that since the experimental probabilities (or normalized shape values) are generally not in perfect agreement with the native structure, we took the closeness of the predicted structure to the native structure as a measure for choosing the parameter β.

We believe RNAsc may be helpful long-term in elucidating the nature of discrepancies between shape and the native structure. As in any experimental protocol, there is a Gaussian error term; however, our data (not shown) indicates that shape discrepancy is positively correlated with high pointwise entropy. Indeed, it seems plausible that a region of the RNA molecule which fluctuates due to thermal motion, thus having higher pointwise entropy, might entail a more variable accessibility for the chemical probe NMIA, thus causing a greater shape discrepancy with the X-ray structure. The program RNAsc allows the user to determine such regions of high pointwise entropy, and to see the structure variability in that region by sampling. It may be possible to confirm or refute our hypothesis concerning the non-Gaussian nature of shape discrepancy ("error"), by performing additional shape probing experiments at lower temperatures. It follows that RNAsc could prove to be a valuable tool in this line of research.

Discussion

Widespread accessibility of quantitative RNA structural mapping techniques and medium- to high-throughput quantification of the data have motivated the development of computational tools to predict structures from such information. The integration of experimental data as "constraints" in the thermodynamic algorithm when computing minimum free energy (MFE) structure can significantly improve the accuracy of RNA structure prediction. However, such methods are also dependent on the quality of the data used for the constraints [26]. It is worth mentioning that the errors in our algorithm RNAsc are directly related to the errors in the experimental data. Fig. 9 shows shape distance to the native structure at the nucleotides where the secondary structure is predicted incorrectly for glycine riboswitch. As can be seen, the shape distances to the native structure are very large for 9 out of the 12 incorrectly predicted positions. Thus the

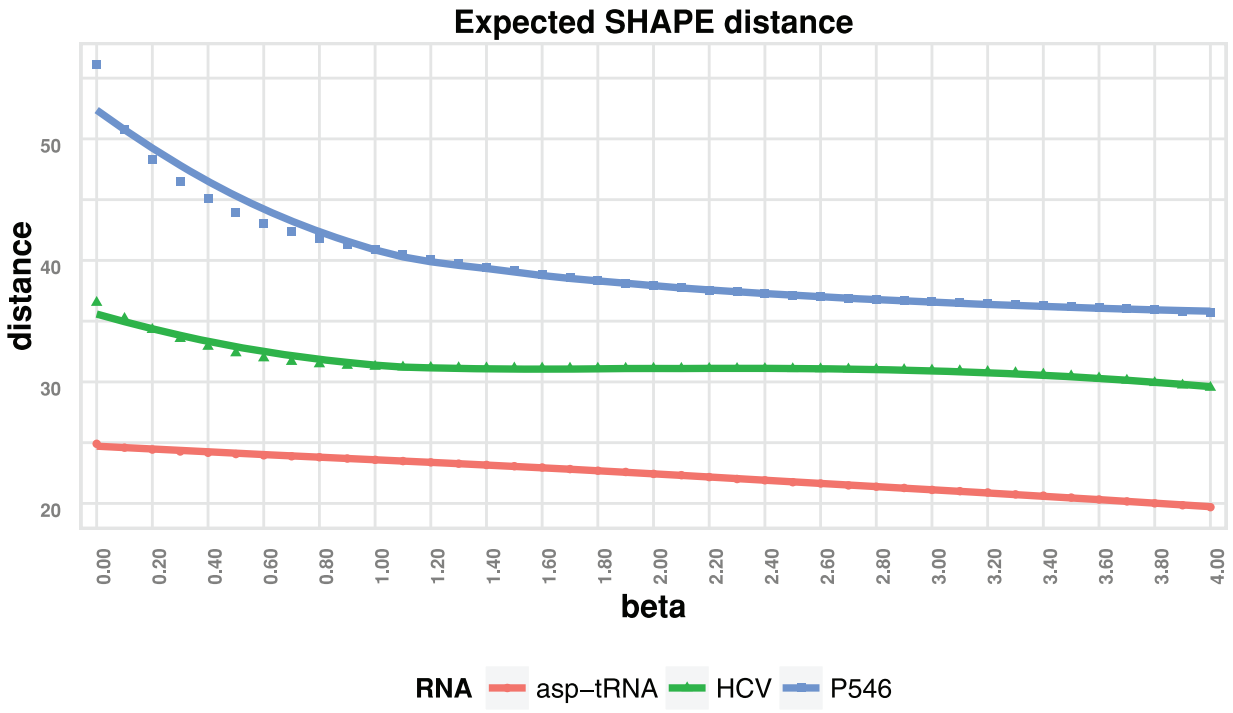

Figure 8. Expected distance of predicted probabilities with normalized shape data. The figure shows a plot of the expected distance $\langle D_\beta \rangle$ between normalized experimental shape values $q_1^*, \ldots, q_n^*$ and the low energy Boltzmann ensemble, as computed by RNAsc. The x-axis depicts increasing values of RNAsc parameter β, while the y-axis depicts expected distance $\langle D_\beta \rangle$. The curves confirm the statement of Theorem 2, which states that as β increases, the expected distance $\langle D_\beta \rangle$ decreases. The figure also shows that for higher values of β, q_i can be made to agree very closely q_i^*. The expected distances of the predicted probabilities with *u*nnormalized shape values for RNAstructure are 61.77, 52.48, and 131.77 for asp-tRNA, HCV, and P546 respectively using optimal parameter values ($b = -0.8$ and $m = 2.6$).
doi:10.1371/journal.pone.0045160.g008

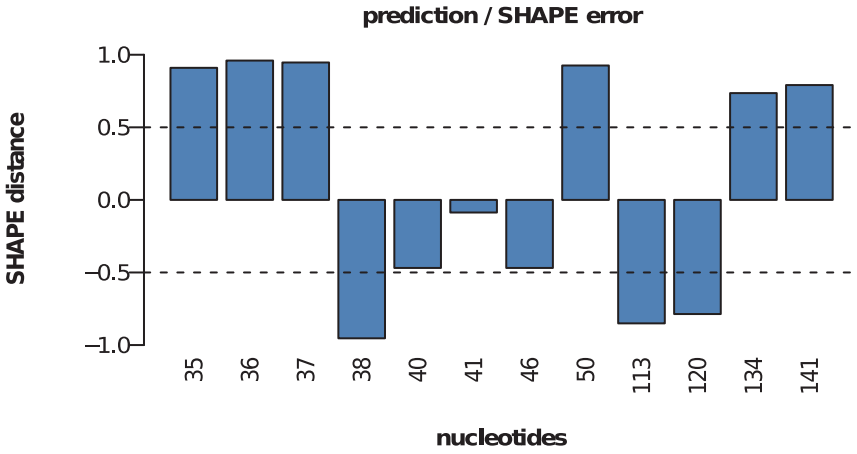

Figure 9. Errors in the prediction of the secondary structure of glycine riboswitch by RNAsc. On the x-axis, nucleotide positions are displayed, where the algorithm predicts the structure incorrectly. The y-axis represents the shape distance to the native structure at the given nucleotide. A shape distance with absolute value ≥ 0.5 indicates an error.
doi:10.1371/journal.pone.0045160.g009

prediction errors are due to the quality of the input data rather than limitations of the algorithm.

Two recent approaches towards overcoming this error include the iterative 'sample and select' approach of Quarrier et al. [38] and the 'mutate and map' strategy of Kladwang et al. [30]. The 'sample and select' strategy involves multiple mapping, followed by a simple filtering step, which removes the suboptimal structures (sampled from the low energy ensemble using the Sfold software [39]) that are incompatible with mapping data. In contrast, the 'mutate and map' strategy involves high-throughput structural probing of all single-nucleotide mutants, resulting in 2D shape data, followed by a computation of the minimum free energy structure, in which pseudo-energy base stacking terms have been added that correspond to Z-scores from 2D shape data. Although high-throughput 'mutate and map' strategies [30], using either shape -CE (capillary electrophoresis) or shape -Seq [40], provide very high secondary structure prediction accuracy, such methods also represent a significant increase in both experimental manipulation and cost that is often not warranted for more specific studies. Especially in such cases, we believe that our method, RNAsc, may be the tool of choice. On the other hand, the 'mutate and map' strategy can be normalized in such a way as to obtain base pairing probabilities. Since shape experiments can potentially probe tertiary interactions (as mentioned in the previous section), not only could we obtain probabilities for secondary interactions and canonical base pairs, but also for tertiary and long range interactions as well as non-canonical base pairs. These probabilities can later be used as input to algorithms such as Probknot [41] or even to a Maximum Weight Matching algorithm [42] to predict pseudoknotted structures and non-canonical base pairs. We are currently pursuing this line of research.

Supporting Information

File S1 Supplementary information.
(PDF)

Acknowledgments

We would like to thank D.H. Mathews for discussions and for making available the source code of RNAstructure [43], including the extension which incorporates base stacking pseudo-energies for shape data [15]. Thanks as well to R. Das for pointing us to the Stanford RNA Mapping Database http://rmdb.stanford.edu/and for a preprint of his paper on the 'mutate and map' strategy. We would like to thank the anonymous referees for helpful remarks.

Author Contributions

Conceived and designed the experiments: PC KZ MMM ID JHC. Performed the experiments: KZ MMM PC. Analyzed the data: KZ MMM ID JHC PC. Contributed reagents/materials/analysis tools: KZ MMM PC. Wrote the paper: KZ MMM ID JHC PC.

References

1. Washietl S (2010) Sequence and structure analysis of noncoding RNAs. Methods in molecular biology (Clifton, NJ) 609: 285–306.
2. Garey M, Johnson D (1990) Computers and Intractability: A Guide to the Theory of NPCompleteness. W.H. Freeman & Co., 338 pages pp. New York.
3. Lyngso RB, Pedersen CN (2000) RNA pseudoknot prediction in energy-based models. J Comput Biol 7: 409–427.
4. Zuker M, Stiegler P (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res 9: 133–148.
5. Tinoco JI, Bustamante C (1999) How RNA folds. Journal of Molecular Biology 293: 271–281.
6. Banerjee A, Jaeger J, Turner D (1993) Thermal unfolding of a group I ribozyme: The lowtemperature transition is primarily disruption of tertiary structure. Biochemistry 32: 153–163.
7. Cho SS Pincus DL, Thirumalai D (2009) Assembly mechanisms of RNA pseudoknots are determined by the stabilities of constituent secondary structures. Proc Natl Acad Sci USA 106: 17349–17354.
8. Bailor MH, Sun X, Al-Hashimi HM (2010) Topology links RNA secondary structure with global conformation, dynamics, and adaptation. Science 327: 202–206.
9. Wilkinson K, Merino E, Weeks K (2005) RNA SHAPE chemistry reveals nonhierarchical interactions dominate equilibrium structural transitions in tRNAAsp. J Am Chem Soc 127: 4659–4667.
10. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31(13): 3406–3415.
11. Hofacker I, Fontana W, Stadler P, Bonhoeffer L, Tacker M, et al. (1994) Fast folding and comparison of RNA secondary structures. Monatsch Chem 125: 167–188.
12. Mathews D, Turner D, Zuker M (2000) Secondary structure prediction. In: Beaucage S, Bergstrom D, Glick G, Jones R, editors, Current Protocols in Nucleic Acid Chemistry, New York: John Wiley & Sons. pp. 11.2.1–11.2.10.
13. Xia T, SantaLucia J, Burkard M, Kierzek R, Schroeder S, et al. (1999) Thermodynamic parameters for an expanded nearest-neighbor model for

formation of RNA duplexes with Watson-Crick base pairs. Biochemistry 37: 14719–35.
14. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, et al. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc Natl Acad Sci USA 101: 7287–7292.
15. Deigan KE, Li TW, Mathews DH, Weeks KM (2009) Accurate SHAPE-directed RNA structure determination. Proc Natl Acad Sci USA 106: 97–102.
16. Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, et al. (2010) Genome-wide measurement of RNA secondary structure in yeast. Nature 467: 103–107.
17. Merino EJ, Wilkinson KA, Coughlan JL, Weeks KM (2005) RNA structure analysis at single nucleotide resolution by selective 2′-hydroxyl acylation and primer extension (SHAPE). J Am Chem Soc 127: 4223–4231.
18. Wilkinson K, Merino E, Weeks K (2006) Selective 20-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. NATURE PROTOCOLS-ELECTRONIC EDITION- 1: 1610.
19. Wilkinson KA, Gorelick RJ, Vasa SM, Guex N, Rein A, et al. (2008) High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states. PLoS Biol 6: e96.
20. Ban N, Nissen P, Hansen J, Moore PB, Steitz TA (2000) The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. Science 289: 905–920.
21. Novikova IV, Hennelly SP, Sanbonmatsu KY (2012) Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. Nucleic Acids Research.
22. Mandal M, Boese B, Barrick J, Winkler W, Breaker R (2003) Riboswitches control fundamental biochemical pathways in Bacillus subtilis and other bacteria. Cell 113(5): 577–586.
23. Meyer M, Roth A, Chervin S, Garcia G, Breaker R (2008) Confirmation of a second natural preQ1 aptamer class in Streptococcaceae bacteria. RNA 14: 685–695.
24. Das R, Laederach A, Pearlman SM, Herschlag D, Altman RB (2005) SAFA: semi-automated footprinting analysis software for high-throughput quantification of nucleic acid footprinting experiments. RNA 11: 344–354.
25. Wilkinson K, Merino E, Weeks K (2005) RNA SHAPE chemistry reveals nonhierarchical interactions dominate equilibrium structural transitions in tRNAAsp transcripts. Journal of the American Chemical Society 127: 4659–4667.
26. Kladwang W, Vanlang CC, Cordero P, Das R (2011) Understanding the Errors of SHAPE-Directed RNA Structure Modeling. Biochemistry 50: 8049–8056.
27. McCaskill J (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers 29: 1105–1119.
28. Matthews D, Sabina J, Zuker M, Turner D (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol 288: 911–940.
29. Forties RA, Bundschuh R (2010) Modeling the interplay of single-stranded binding proteins and nucleic acid secondary structure. Bioinformatics 26: 61–67.
30. Kladwang W, Cordero P, Das R (2011) A mutate-and-map strategy accurately infers the base pairs of a 35-nucleotide model RNA. RNA 17: 522–534.
31. Zuker M (1989) On finding all suboptimal foldings of an RNA molecule. Science 244: 48–52.
32. Higgs PG (1996) Overlaps between RNA secondary structures. Physical Review Letters 76: 704–707.
33. Duncan C, Weeks K (2008) Shape analysis of long-range interactions reveals extensive and thermodynamically preferred misfolding in a fragile group i intron RNA. Biochemistry 47: 8504–8513.
34. Soukup G, Breaker R (1999) Relationship between internucleotide linkage geometry and the stability of RNA. RNA 5: 1308–1325.
35. Dann CE, Wakeman C, Sieling C, Baker S, Irnov I, et al. (2007) Structure and mechanism of a metal-sensing regulatory RNA. Cell 130: 878–892.
36. Mathews D (2005) Predicting a set of minimal free energy RNA secondary structures common to two sequences. Bioinformatics 15: 2246–2253.
37. Turner DH, Mathews DH (2010) Nndb: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic acids research 38: D280–2.
38. Quarrier S, Martin J, Davis-Neulander L, Beauregard A, Laederach A (2010) Evaluation of the information content of RNA structure mapping data for secondary structure prediction. RNA 16: 1108–1117.
39. Ding Y, Chan CY, Lawrence CE (2004) Sfold web server for statistical folding and rational design of nucleic acids. Nucleic Acids Res 32: 0.
40. Lucks JB, Mortimer SA, Trapnell C, Luo S, Aviran S, et al. (2011) Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). Proceedings of the National Academy of Sciences of the United States of America 108: 11063–11068.
41. Bellaousov S, Mathews DH (2010) Probknot: fast prediction of RNA secondary structure including pseudoknots. RNA (New York, NY) 16: 1870–1880.
42. Tabaska J, Cary R, Gabow H, Stormo G (1998) An RNA folding method capable of identifying pseudoknots and base triples. Bioinformatics 14: 691–699.
43. Reuter JS, Mathews DH (2010) RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics 11: 129.