Evaluation of the information content of RNA structure mapping data for secondary structure prediction

SCOTT QUARRIER,[1,3] JOSHUA S. MARTIN,[1,3] LAUREN DAVIS-NEULANDER,[2] ARTHUR BEAUREGARD,[1] and ALAIN LAEDERACH[1,2]

[1]Biomedical Sciences Program, University at Albany, Albany, New York 12208, USA
[2]Developmental Genetics and Bioinformatics, Wadsworth Center, Albany, New York 12208, USA

ABSTRACT

Structure mapping experiments (using probes such as dimethyl sulfate [DMS], kethoxal, and T1 and V1 RNases) are used to determine the secondary structures of RNA molecules. The process is iterative, combining the results of several probes with constrained minimum free-energy calculations to produce a model of the structure. We aim to evaluate whether particular probes provide more structural information, and specifically, how noise in the data affects the predictions. Our approach involves generating "decoy" RNA structures (using the sFold Boltzmann sampling procedure) and evaluating whether we are able to identify the correct structure from this ensemble of structures. We show that with perfect information, we are always able to identify the optimal structure for five RNAs of known structure. We then collected orthogonal structure mapping data (DMS and RNase T1 digest) under several solution conditions using our high-throughput capillary automated footprinting analysis (CAFA) technique on two group I introns of known structure. Analysis of these data reveals the error rates in the data under optimal (low salt) and suboptimal solution conditions (high $MgCl_2$). We show that despite these errors, our computational approach is less sensitive to experimental noise than traditional constraint-based structure prediction algorithms. Finally, we propose a novel approach for visualizing the interaction of chemical and enzymatic mapping data with RNA structure. We project the data onto the first two dimensions of a multidimensional scaling of the sFold-generated decoy structures. We are able to directly visualize the structural information content of structure mapping data and reconcile multiple data sets.

Keywords: RNA structure; chemical mapping; DMS, footprinting; secondary structure

INTRODUCTION

RNA is a multifaceted functional molecule that is capable of adopting a wide array of highly specific conformations that confer upon it the ability to carry out highly specialized functions in the cell (Dolnik 1999; Weinstock 2007; Morton 2008). When a novel RNA is identified, it is common practice to input its sequence into a folding program, such as mFold, to reveal structural motifs of interest (Reeder et al. 2006). The resulting structure provides a context upon which a hypothesis is generated, and new experiments are designed to probe the molecular details of the RNA. Because the RNA secondary structure is central to this process, experimental validation of the structural model is often desired. Structure mapping offers an experimentally straightforward approach for model validation (Tijerina et al. 2007; Wilkinson et al. 2008).

We use the term "structure mapping" to describe the broad range of chemical and enzymatic probes used in RNA structural analysis (Mitra et al. 2008). In particular, we focus on the subset of probes that modify the base of the nucleotide (including, but not limited to DMS, Kethoxal, CMCT [1-cyclohexyl-3-(morpholynoethyl) carbodiimide], and DEPC [diethyl pyrocarbonate]) (Ehresmann et al. 1987; Brunel and Romby 2000; Harkins 2001) as well as RNases (Donis-Keller et al. 1977; Lockard and Kumar 1981; Vary and Vournakis 1984) that selectively cleave single- or double-stranded RNA (e.g., T1, T2, U2, V1, and CL3). These probes are of particular interest to this study as they target the base, unlike probes such as the •OH (hydroxyl) radical and NMIA (N-methylisatoic anhydride) that cleave and modify the backbone of the RNA (Latham and Cech 1989; Wilkinson et al. 2005).

[3]These authors contributed equally to this work.

Reprint requests to: Alain Laederach, Developmental Genetics and Bioinformatics, Wadsworth Center, 150 New Scotland Avenue, Albany, NY 12208, USA; e-mail: alain@wadsworth.org; fax: (518) 486-4103.

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.1988510.

Structure mapping as described above determines whether a particular base is paired or not, but it does not identify the partner of the base in the pairing. By itself, this data may not contain sufficient information to identify the structure of the RNA. For this reason, structure determination requires combining the structure mapping data with a free-energy minimization approach (Mathews 2004; Hart et al. 2008; Wilkinson et al. 2008). This study investigates the complex relationship between the free-energy function and structure mapping data.

Traditionally, a constraint approach is used to incorporate structure mapping data into structure prediction. The state of the art for RNA structure prediction is to not allow chemically accessible nucleotides to be buried in a helix. Chemically modified nucleotides are constrained to be single stranded, at the end of a helix, in a GU pair, or adjacent to a GU pair (Mathews et al. 2004). Furthermore, for enzymatic data, it has been shown that constraining only nucleotides between two consistent cuts also improves structure prediction (Mathews et al. 1999). Although this approach will generally yield an improved structural prediction, we and others have found that even small deviations from perfect data, typical of structure mapping experiments, results in structures no better than the minimum free-energy (MFE) structure (Deigan et al. 2009). This result might suggest that the information content of structure mapping data is low, and that only perfect information will yield a correct structure. Since chemical reactivity is dependent on a variety of factors including 3D structure, base identity (Wilkinson et al. 2009), noncanonical base-pairing (Leontis et al. 2002), and solution conditions, it is difficult to carry out an experiment that yields perfect information. A recent study incorporating NMIA (a backbone probe) reactivity directly into the free-energy function, using as a training set the *E. coli* 16S rRNA, showed more than 30% improvement in sensitivity and positive predictive value (PPV) with the experimental data (Deigan et al. 2009).

To better characterize the relationship between the free-energy function and the structure mapping data, we propose to separate the data and energy minimization in our analysis. We leverage the ability of sFold to efficiently sample RNA secondary structures and then evaluate how well we identify the correct structure from an ensemble of decoys (Ding et al. 2004, 2005; Waldispuhl and Clote 2007). This approach allows us to evaluate different metrics for incorporating structure mapping data into RNA structure and determine the information content of the data independently of the free-energy function. Furthermore, this allows us to reconcile data that probe different nucleotides (e.g., DMS and T1 RNase) in a single computational framework.

Our results will be of particular interest to those undertaking structural determination experiments with multiple chemical and enzymatic probes, in that they allow a quantitative analysis of each probe's structural agreement with the other. In particular, if one probe systematically identifies a subset of structures different from the other probes, our method can help identify this potentially erroneous data and suggest a repeat of the experiment. Fundamentally, we offer a straightforward approach for reconciling multiple chemical and enzymatic mapping data sets and optimizing experimental protocols to improve structure prediction.

RESULTS

Perfect information

We began our investigation of the structural information content of structure mapping data by considering the case where perfect information is available, i.e., all base-paired and unpaired nucleotides are correctly identified by the data. To evaluate the quality of our prediction we consider two metrics, positive predictive value (PPV) and sensitivity. PPV is the percentage of predicted canonical base pairs that are found in the reference (crystal structure). Sensitivity is the percentage of reference (crystal) canonical base pairs that are found in the predicted structure. These metrics are commonly used to evaluate RNA secondary structure predictions (Mathews et al. 2004; Deigan et al. 2009). We aimed to predict RNA structures with the largest PPV and sensitivity using their crystal structure as a reference. The crystal structure (Fig. 1A) is a completely independent source of experimental information on the secondary structure of the RNA, and for this reason we chose to use it as a reference (Fig. 1B). For the purposes of this study, we only considered Watson–Crick canonical and G–U wobble base pairs as determined in the Nucleic Acid Database (Lemieux and Major 2002), excluding all tertiary contacts.

Our approach involved generating a large number of decoy structures using sFold (Ding et al. 2004) and then using structure mapping data to select the decoy with the lowest distance to the data (Fig. 1C). We refer to this novel approach as "sample and select," as opposed to more traditional constraint satisfaction (Mathews et al. 2004). This strategy requires defining a quantitative metric to evaluate the agreement between the structure mapping data and the different structures. For the case of perfect information, a simple "Manhattan" metric is sufficient, where the structure mapping data is represented as ones (meaning the base is paired) and zeros (base not paired). We computed such a vector corresponding to each decoy structure and calculated the distance as the sum of the absolute differences between the reference and decoy vectors.

We performed this calculation for 10^6 sFold-generated decoys of the P4P6 subdomain of the *T. thermophila* group I intron (Fig. 1). In Figure 2A, we plot the sum of the sensitivity and PPV as a function of the Manhattan distance. Several important results are revealed with this simple analysis:

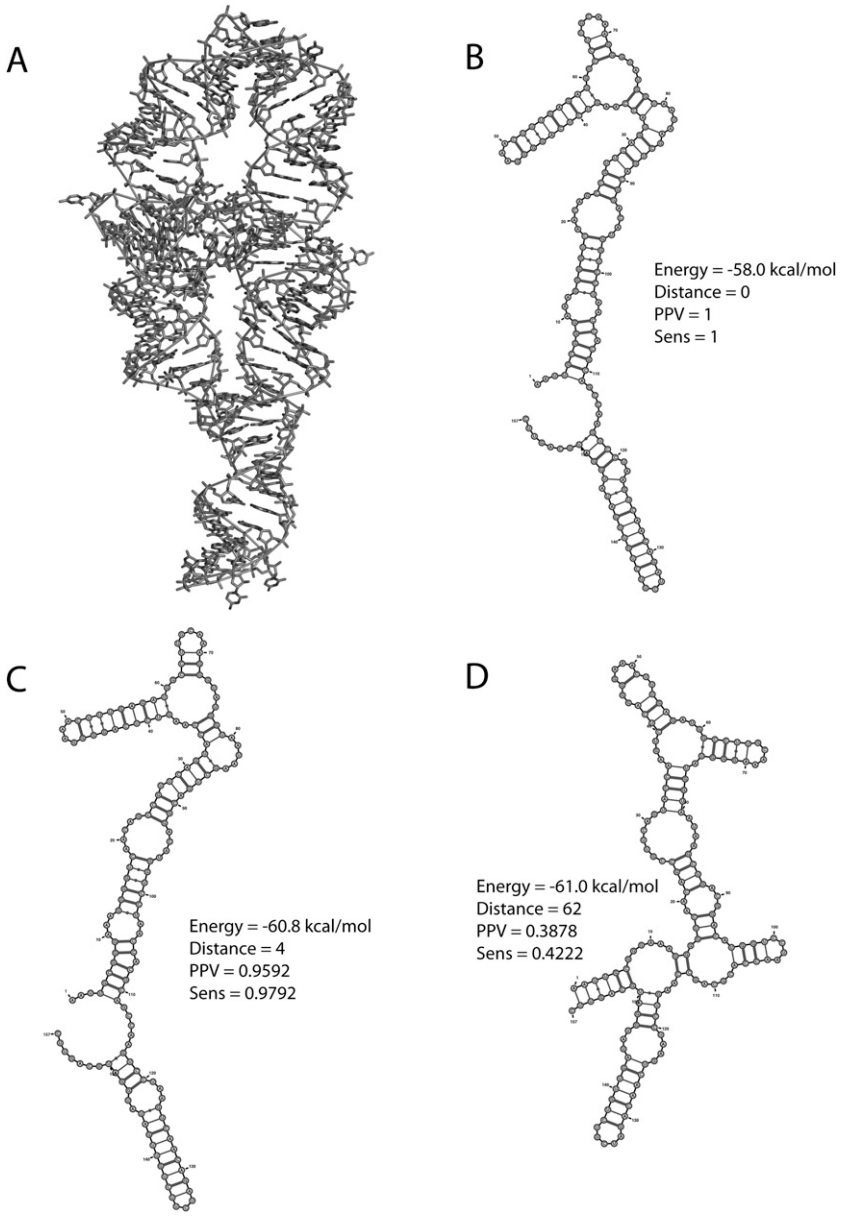

FIGURE 1. (*A*) Crystal structure representation of the P4P6 subdomain of the *T. thermophila* group I intron (PDB ID 1GID). In this study we used the secondary structure derived from the crystal structure as a reference as determined by NDB (Berman et al. 2002) excluding all tertiary contacts. (*B*) Secondary structure derived from the P4P6 crystal structure used as a reference. The PPV and sensitivity of the structure are 1 and it has a Manhattan distance of 0 since it is the reference structure. (*C*) sFold decoy with lowest Manhattan distance and greatest sum of PPV and sensitivity found in the 10^6 decoys that we generated for this analysis. (*D*) sFold decoy with the largest Manhattan distance and low PPV and sensitivity.

1. There is an inverse correlation between higher PPV and sensitivity and the Manhattan distance (Fig. 2A). Therefore, decoys with a large Manhattan distance will have lower PPV and sensitivity (Fig. 1D).
2. The sFold algorithm samples structures very near the correct crystal structure (for P4P6 a decoy with a Manhattan distance of 4 is sampled and is illustrated in Fig. 1C). Highly diverse structures are also sampled that are very different from the crystal structure, and thus have low PPV and sensitivity (Fig. 1D).
3. There is no correlation between the energy of the structure, PPV, and sensitivity (see Supplemental Fig. 1).
4. The Manhattan distance metric (Fig. 2A, *x*-axis) increases in units of two, since making or removing a base pair will always involve two bases. This also means that our distance 4 structure (Fig. 1C) is only two base pairs off from the correct crystal structure.

We performed similar calculations on five RNAs of known crystal structure and report the results in Table 1. We see that in all cases we sample structures with high PPV and sensitivity, and that these structures have a low Manhattan distance. For cases where we sample multiple equidistant structures (e.g., HCV IRES), the mean PPV and sensitivity of the decoys remains high and is above that of the MFE structure. In these examples we never sampled the correct structure (Manhattan distance = 0), but our results suggest, though do not prove, that perfect data will only "fit" one structure in all cases. The constraint approach (Table 1, far right column) does not identify the correct structure with perfect information. It should be noted that in these cases we use all of the data (paired and unpaired for all bases) and that constraint approaches are designed for incomplete data (e.g., only unpaired bases). It is likely that the poor performance of the constraint approaches in this case is due to overconstraining the structure. Table 1 also shows that the PPV and sensitivity values we obtain with our sampling methodology are on par with recent results incorporating SHAPE chemical mapping data in RNA structure prediction (Deigan et al. 2009). The results we present below using experimental structure mapping data show that the sample and select approach we propose allow us easily to reconcile multiple data sets and rapidly identify erroneous (or high error) data.

The advent of high-throughput techniques for obtaining structure mapping data (Mitra et al. 2008; Vasa et al. 2008) promises to provide a wealth of structural information on

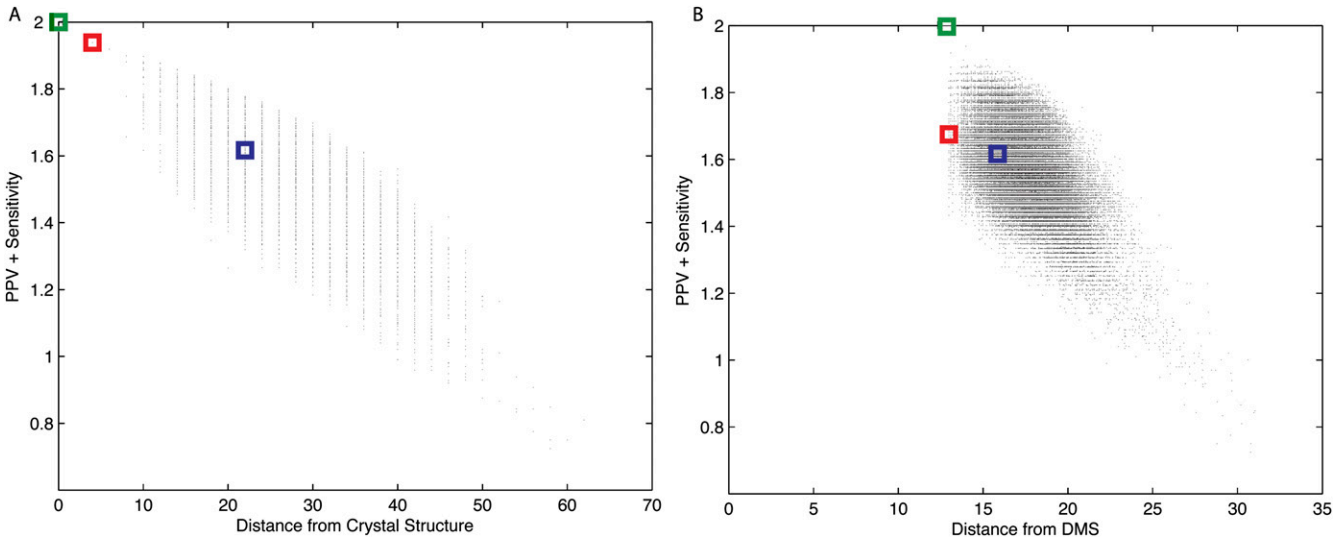

FIGURE 2. (*A*) Plot of the sum of PPV and sensitivity to the reference crystal structure as a function of the Manhattan distance to perfect information (ideal) data for 1,000,000 P4P6 decoys generated by sFold (Ding et al. 2005). (*B*) Similar plot computing the Manhattan distance to the experimentally obtained DMS chemical mapping data at 100 mM KCl for the P4P6 domain of the *Tetrahymena* group I intron. The lowest distance decoy is indicated with a red box, the MFE structure with a blue box, and the crystal (reference) is in green.

RNA. However, such data sets can have higher error rates due to the automated nature of the data fitting (Mitra et al. 2008). Furthermore, collecting structure mapping data under a wide range of solution conditions becomes trivial when using a multiplex capillary sequencer (Deigan et al. 2009). It is now possible to collect such data in vivo where solution conditions cannot be adjusted to optimize the reactivity of a probe to predict secondary structure (Adilakshmi et al. 2006). As a result it is important to consider the effects of inaccuracy in the structure mapping data on our predictions of RNA structure. We will also show that the error introduced by our not sampling the correct structure (Table 1) is very small compared with the error introduced by noise in experimental chemical mapping data.

Experimental data

We performed a series of structure mapping experiments with our CAFA technology (Mitra et al. 2008; Vasa et al.

TABLE 1. Prediction of RNA structure performance using perfect information

	Max PPV/sens[a]	Mean PPV/sens[b]	Number of decoys in top group[c]	Min Manh distance[d]	Total number of decoys	MFE PPV/sens[e]	SHAPE PPV/sens[f]	Constraint PPV/sens[g]
P4P6 (1GID)	0.96/0.98	0.96/0.98	1	4	10^6	0.82/0.80	0.98/0.96	0.96/0.60
Twort (1Y0Q)	0.92/0.98	0.92 ± 0.01/ 0.97 ± 0.02	2	12	10^6	0.66/0.77	N/A	0.75/0.43
Azoarcus (1ZZN)	0.97/0.93	0.97/0.93	1	10	10^5	0.83/0.79	N/A	0.68/0.80
tRNA (TRNA07)	0.80/1.0	0.80 ± 0.02/ 0.99 ± 0.02	2	10	10^5	0.65/1.0	1.0/1.0	0.77/0.95
HCV IRES (1P5O)	1.0/0.75	0.99 ± 0.02/ 0.72 ± 0.03	15	12	10^6	0.62/0.40	1.0/1.0	0.21/1.0

[a]Maximum PPV and sensitivity observed in the ensemble of structures.
[b]Mean PPV and sensitivity of decoys with lowest Manhattan distance to crystal structure.
[c]Number of decoys with minimum Manhattan distance to crystal structure.
[d]Minimum Manhattan distance.
[e]PPV and sensitivity of minimum free-energy structure.
[f]Values reported in Deigan et al. (2009) using SHAPE chemistry.
[g]Values using RNAFold (Bernhart et al. 2006) constraint approach with all ideal data.

2008) using a capillary sequencer to analyze fluorescently labeled cDNA obtained by reverse transcription of the chemically modified RNA. These experiments were designed to both characterize structure mapping data collected in this way and evaluate our sample and select approach as a method for determining structure. We aim to determine the operational characteristics for two commonly used chemical and enzymatic probes, DMS and T1 (Donis-Keller et al. 1977; Tijerina et al. 2007). To this end we chose to study the P4P6 domain of the *T. thermophila* group I intron and the *Twort* group I intron, both of which have been crystallized (Cate et al. 1996; Golden et al. 2005). We determined the reference secondary structure of both introns by analyzing the base-pairing interactions in their respective crystal structures (PDB ID GID and 1Y0Q), as the crystal structure is an independent measurement of structure (Berman et al. 2002). We only considered canonical Watson–Crick base-pairs and G–U wobbles, excluding any tertiary contacts from our analysis. Several comparative structures for our two introns of interest are published and generally agree with the crystal structure (maximum distance 12 using our Manhattan metric) (Cannone et al. 2002).

We conducted duplicates of experiments in which we probed the folded RNA either in the presence of 100 mM KCl or 10 mM $MgCl_2$. It is well established that in the presence of 100 mM KCl, the secondary structure of the RNA is formed, but only upon addition of $MgCl_2$ does the RNA fold into its native structure (Mathews et al. 1997; Laederach et al. 2006, 2007). Structure mapping experiments probe the accessibility of nucleotides, which is correlated to both secondary and tertiary structure (Vicens et al. 2007). Thus, we expect to see the best agreement between the chemical mapping data collected in the presence of 100 mM KCl and our reference structure. This is indeed the case for both P4P6 and the *Twort* ribozyme when probed with DMS, as the presence of $MgCl_2$ increases the error rate (Table 2).

We determined the experimental error rates reported in Table 2 by defining a threshold value above which a base is considered unpaired and below which it is considered paired. For the purpose of this study we used a threshold that minimizes the error rate as defined by Equation 1. In Figure 3 we plot histograms of the peak areas (as determined by single nucleotide peak fitting of the capillary trace) (Takamoto et al. 2004; Das et al. 2005; Mitra et al. 2008) for two of our P4P6 data sets. For the DMS data collected at 100 mM KCl, the choice of the threshold is relatively straightforward, as the distribution of the DMS reactivity is bimodal (Fig. 3A) and a relatively large range of thresholds yield low error rates (Fig. 3B). For data collected in the presence of 10 mM $MgCl_2$, however, the problem of determining a threshold is more complex due to the absence of a truly bimodal distribution in the peak areas (Fig. 3C). It is only when we determine the error rate as a function of the threshold that we see that there is a narrow window in the threshold values that yields a minimal error rate (shown as a dot–dash line in Fig. 3D). We used this optimal threshold to compute the error rates in Figure 2. However, this calculation requires a priori knowledge of the structure and is therefore not a viable option for RNA secondary structure prediction.

TABLE 2. RNA structure prediction using experimental data

	TP[a]	FP[b]	TN[c]	FN[d]	Error rate[e]	Mean PPV/sensitivity[f]	RNAfold PPV/sensitivity[g]	RNAstructure PPV/sensitivity[h]
P4P6 DMS 100 mM KCl	37–39	2	33	2	0.05–0.08	0.83 ± 0.03/ 0.84 ± 0.03	0.67–0.75/ 0.49–0.61	0.87/ 0.92–0.94
P4P6 DMS 10 mM $MgCl_2$	36	18–20	15–17	5	0.30–0.33	0.79 ± 0.04/ 0.78 ± 0.03	0.65–0.72/ 0.53	0.85–0.86/ 0.86–0.92
P4P6 T1 100 mM KCl	29–30	4–6	5–6	1–4	0.17–0.19	0.81 ± 0.07/ 0.93 ± 0.07	0.80/0.80	0.75–0.86/ 0.80–0.88
Twort DMS 100 mM KCl	27–31	12	26–27	9	0.22	0.75 ± 0.05/ 0.81 ± 0.06	0.44–0.79/ 0.46–0.57	0.85–0.89/ 0.65–0.76
Twort DMS 10 mM $MgCl_2$	34–40	12–14	23–26	6–10	0.24–0.27	0.65 ± 0.06/ 0.75 ± 0.04	0.67–0.80/ 0.42–0.53	0.87–0.88/ 0.64–0.68
Twort T1 100 mM KCl	25–28	9–10	1–2	0	0.25–0.26	0.69 ± 0.08/ 0.79 ± 0.06	0.80–0.84/ 0.69	0.9/ 0.72–0.80

[a]Number of bases correctly identified as base-paired.
[b]Number of bases identified as base-paired that are not base-paired in the reference crystal structure.
[c]Number of bases correctly identified as not base-paired.
[d]Number of bases identified as not base-paired that are base-paired in the reference crystal structure.
[e]Error computed as defined by Equation 1.
[f]Mean PPV and sensitivity of the top 100 decoys (out of 10^6) with lowest Manhattan distance to both data sets.
[g]RNAfold PPV and sensitivity for predicted structure using TN bases as constraints (unpaired).
[h]RNAstructure PPV and sensitivity for predicted structure using data thresholds that minimize error rate of the data.

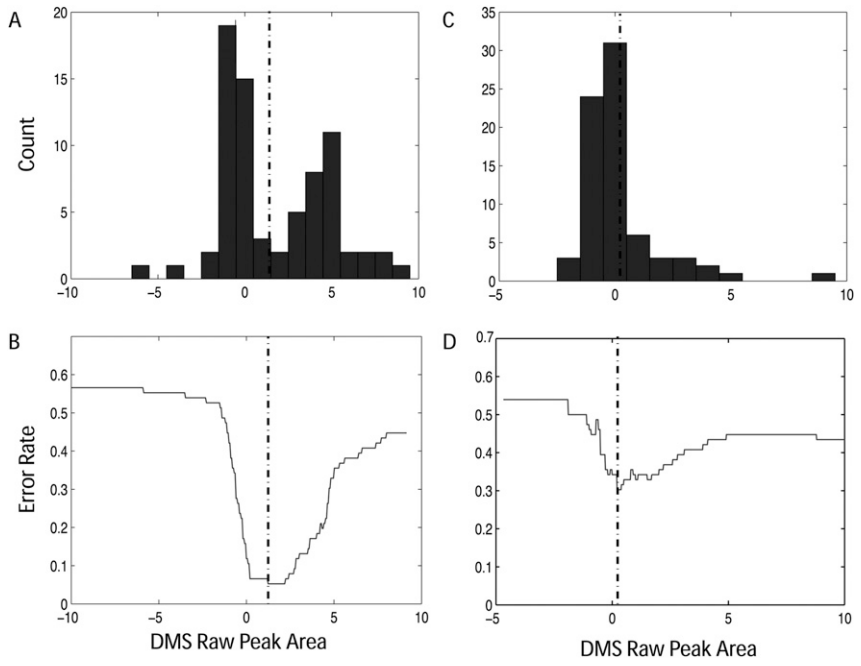

FIGURE 3. (*A*) Histogram of the raw DMS peak areas for adenines and cytosines for the *Tetrahymena* group I intron at 100 mM KCl. The distribution of peak areas is bimodal with several outliers. Some of the very negative and positive values are a result of RT stops that yield very large peak areas that is one source of noise in the data. (*B*) Error rate as a function of threshold when predicting paired/unpaired bases using as the reference the crystal structure (PDB ID 1GID). The dotted vertical line identifies an optimal threshold above which bases are considered unpaired (high DMS reactivity). (*C*) Same histogram as in *A*, however, for data collected in the presence of 10 mM $MgCl_2$. (*D*) Error rate as a function of threshold for data collected in the presence of 10 mM $MgCl_2$. An optimal threshold value can be found, but it is not apparent from the histogram (*C*).

One advantage of our sample and select approach is that it does not require defining a threshold, which can be difficult for data collected under non-ideal solution conditions (e.g., 10 mM $MgCl_2$) (Fig. 3B). Our method's performance is equal to (and in some cases better than) a constraint approach with the 100 mM KCl (low error) data, as can be seen in Table 2. However, it is with data where the false negative (FN) rates are high (10 mM $MgCl_2$) that our method shows the greatest improvement. Constraint approaches generally require choosing a high threshold so as to minimize the FN rate (Mathews et al. 2004). Although picking a higher threshold value does decrease the FN rate (and improve moderately the RNAfold prediction), our 10 mM $MgCl_2$ data has a high FN rate regardless of the choice of threshold. The high FN rate is due to strong RT stops that make background subtraction difficult (Mitra et al. 2008) and thus yield large positive peaks areas. This is an inherent feature of using a single primer and analyzing hundreds of nucleotides in a single reaction (Mitra et al. 2008). We used an automated approach for identifying these stops as well as background subtraction; nonetheless, in such large and comprehensive data sets, correctly identifying all RT stops and therefore eliminating false negatives is challenging to completely automate.

Visualizing decoy selection and structural information content

An interesting outcome of the Boltzmann sampling procedure for RNA is that the decoy structures generated form clusters (Ding et al. 2005). These are easily visualized by performing multidimensional scaling on the pairwise Manhattan distance matrix of all decoy structures. We performed this analysis on 5000 Boltzman sampled decoys generated by sFold (Ding et al. 2004) for P4P6 and project each decoy on the first two dimensions in Figure 4 as black dots. The first two components of our analysis capture 34% of the variance in the sample. We selected the 30 decoy structures that have the lowest Manhattan distance to the 100 mM KCl DMS (red), T1 data (Magenta), and 10 mM $MgCl_2$ DMS (blue) data. Figure 4A clearly illustrates that lower error data (red 100 mM KCl DMS) selects a majority of decoys near the reference crystal structure (green square), while the high error data (10 mM $MgCl_2$) selects decoys belonging to a different cluster. The T1 data (which only probes guanosines) selects a wider range of decoy structures (Magenta). It is important to note that in all cases the 30 structures we selected are within one unit of Manhattan distance to the top structure and, therefore, can be considered equidistant within noise.

To visualize all of the data collected for P4P6 (replicates included), we project our experimental data onto the two first dimensions of the multidimensional scaling calculation of the sFold decoys. The result of this calculation is illustrated in Figure 4B and allows one to visualize the experimental reproducibility in RNA multidimensional space. Colored squares and circles represent the full replicates of our data, with the green square indicating the "correct" crystal structure. The low-salt (100 mM KCl DMS data, red square and circle) is closest to the green reference structure. This visualization suggests that picking decoys in this multidimensional space near the experimental data may lead to better results than using our simple Manhattan metric. Although this is the case for the 10 mM $MgCl_2$ data on P4P6 (we see an improvement of PPV/sensitivity of 0.83 ± 0.02/0.81 ± 0.03 compared with 0.79 ± 0.04/0.78 ± 0.03 [Table 2] obtained with the Manhattan distance metric), the overall improvement does not seem to justify using this more complex distance metric. These visualizations remain powerful for the analysis of

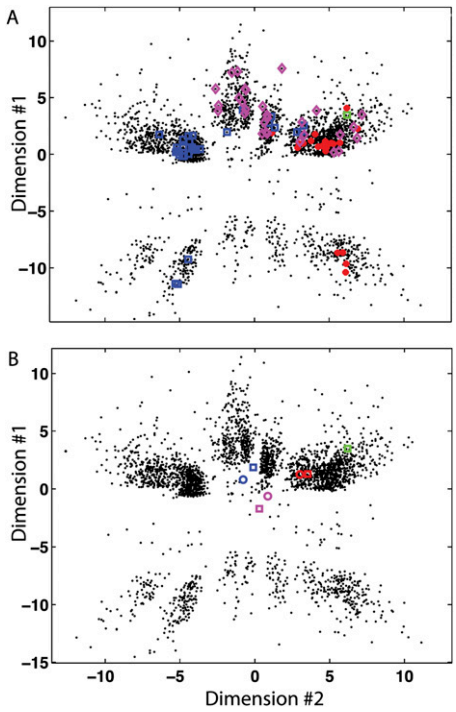

FIGURE 4. (*A*) Multidimensional scaling of 5000 P4P6 RNA decoys generated by sFold and projected onto the first two dimensions as black dots. The green square is the reference (crystal) structure projected onto the two dimensions. The red dots represent the top 30 equidistant structures to the P4P4 DMS data collected at 100 mM KCl, while the blue squares represent the same selection for the 10 mM $MgCl_2$ data computed using the Manhattan distance metric. The magenta diamonds represent the same selection for the 100 mM T1 data. (*B*) Same decoys as in *A*; however, in this case we directly projected the structure mapping data onto the first two dimensions. The green square represents the crystal (reference) structure, while the red data represent repeats of the 100 mM KCl DMS, blue 10 mM $MgCl_2$, and magenta 100 mM KCl T1 data.

structure mapping data and suggest possible novel algorithmic development for their analysis and interpretation.

DISCUSSION

Structure mapping is a proven technology for the analysis of RNA structure (Latham and Cech 1989; Shcherbakova and Brenowitz 2005, 2008; Tijerina et al. 2007; Deigan et al. 2009). The advent of high-throughput approaches for collecting such data promises to expand the importance of these techniques for the understanding of RNA structure and function in the cell. The data collected in this way present novel computational challenges in terms of their interpretation, since chemical reactivity to a probe is always a function of both secondary and tertiary structure. We are now in a position to more easily and rapidly probe RNA under different conditions and by using a wide variety of probes (Mitra et al. 2008). In this work we have attempted to quantitatively analyze the potential impact of RNA tertiary structure formation on secondary structure prediction and found that depending on the method used, the effects are important (Table 2). One solution is to incorporate the structure mapping data into the energy function for minimization and "learn" a set of parameters based on a reference structure. In terms of pure RNA secondary structure prediction, this approach produces remarkable results, correctly predicting most structures with 99% PPV and sensitivity on average (Deigan et al. 2009).

Our sample and select approach achieves similar performance to the above-mentioned method with perfect data. Decoy structures with near perfect sensitivity and PPV for most RNAs in our test set are sampled (first column of Table 1). As is illustrated in Figure 2, our approach identifies an ensemble of structures equidistant to a particular data set instead of a single structure. We also observe moderate improvements in PPV and sensitivity compared with the constraint approach when using our sample and select approach for structure prediction with experimental data (Table 2). We hypothesize that the uncoupling of the structure prediction and selection is more robust to experimental noise because the prediction relies more heavily on the free-energy rules than in the constraint approach. A single error in the experimental data cannot wreck a structure prediction with our sample and select approach.

We observe similarly good predictive performance using RNAstructure (Table 2) with our experimental data. RNAstructure does not impose a strict binary constraint based on chemical data, which clearly improves performance (Mathews et al. 2004). Nonetheless, with RNAstructure, the user must still define a threshold to determine which bases are unpaired. For the RNAs that we have collected data on (Table 2), we used the optimal threshold that minimizes the error in the data (as illustrated in Fig. 3) as input for RNAstructure. This is possible because the correct structure is known. If the structure of the RNA is unknown, different strategies for determining a threshold would be needed, which may lead to different results. The RNAstructure and RNAfold performance we report in Table 2 therefore represent a best-case scenario for these approaches. We specifically developed our sample and select approach to eliminate the need for data thresholding for RNA structure prediction.

A surprising result is that the top decoys selected, even with our very low error data, (5% error rate at 100 mM KCl) do not all fall within the same cluster (see red dots in the lower right cluster in Fig. 4A). There are two possible interpretations of this result. One is that the structural information content of structure mapping is low and that we cannot always identify a single unique structure from the data. Alternatively, it is possible that the nearest neighbor rules implemented in RNA structure prediction are inadequate at identifying the single structure. We will propose here that RNA does not adopt a single secondary structure conformation but instead populates an ensemble.

We will argue that the latter interpretation is more likely in this case and this is supported by our data.

A majority of computational approaches for RNA secondary structure prediction identify multiple alternative structures with very similar energies (Mathews 2004; Hart et al. 2008). The sFold Botzmann sampling we utilized in this study further illustrates the size of the conformational space of an RNA. Of the million decoys we generated for each molecule in our test set, only about 1600 (<0.16%) are structurally identical. This simple calculation puts into perspective the vastness of the conformational space accessible to an RNA polymer. Computational analysis of RNA structure therefore suggests that alternative RNA structures will also exist in solution.

The evidence for alternative secondary structures is not limited to computational results. Kinetic studies of the *Tetrahymena* group I intron reveal that mutations which disrupt RNA secondary structure have profound effects on the folding rates and are dependent on the initial conformational ensemble (Shcherbakova et al. 2004, 2008; Laederach et al. 2007). Furthermore, single molecule experiments reveal significant conformational heterogeneity in RNA over a wide variety of solution conditions (Bartley et al. 2003; Onoa et al. 2003; Zhuang 2005). In fact, most protocols for preparing RNA for structural characterization include several heating and cooling steps to maximize the homogeneity of the RNA (see Materials and Methods for the protocols used in our experiments and Uhlenbeck [1995]).

Our structure mapping data presents a significant dynamic range (Fig. 3). Although clearly bimodal for the 100 mM KCl solution condition (Fig. 3A), there remains important variability in the reactivity of individual nucleotides that are either paired or unpaired. If an ensemble of RNA structures exists in solution, the chemical reactivity observed for each nucleotide would be proportional to the ratio of unpaired to paired for that nucleotide. The fact that we observe reproducible differences in chemical reactivity for different nucleotides further suggests that the structure mapping experiments we performed are measuring a probability of base pairing. The ensemble of structures selected with our method therefore represents an experimentally constrained version of the sFold Boltzmann sampling.

Practically, our approach is most valuable as a tool for visualizing structure mapping data. Traditionally, the data is projected onto a single RNA structure as a visual validation for the prediction. The visualization proposed in Figure 4 offers a simple and informative representation of the data in terms of possible RNA conformations. Figure 4 clearly shows that the two data sets collected at 100 mM KCl (T1 and DMS, magenta and red, respectively) both identify the correct structural cluster containing the reference structure (green square). We clearly identify that the 10 mM $MgCl_2$ data (blue) selects decoys in an alternative cluster. Furthermore, we are able to visualize the relative uncertainty in the prediction from each data set from the relative spread of the points in the selection. When the data is projected onto the first two dimensions of the multidimensional scaling as in Figure 4B, it is easy to visualize multiple data sets and determine their relative agreement.

Standardized data sets such as the ones we have collected here are central to methodological development, and in the supplement of this manuscript we also provide the raw data from our experiments. We have also created a web server at http://cloud.wadsworth.org/mapfold that reproduces much of the computational functionality we have illustrated herein. Our goal is to facilitate the interpretation of structure mapping data and to relate it to structural ensembles.

MATERIALS AND METHODS

RNA transcription

The L–21 *T. thermophila* group I intron plasmid was provided by Nathan Boyd (Stanford University School of Medicine) and *Twort* intron plasmid was provided by Michael Brenowitz (Albert Einstein College of Medicine). Plasmids were transformed in DH5α electrocompetent cells and the DNA was purified with a Qiagen miniprep kit. The template DNA was amplified by PCR with the following primers:

P4P6: 5′-ACTCCAAAACTAATCAATATACTTTC-3′;
P4P6 forward: 5′-CCAAGTAATACGACTCACTATAGGAGGGAAAA-3′;
Twort: 5′-AATTATGTTACGGATAGGTTCGTTACTCC-3′; and
Twort forward: 5′-GCCAAGCTTAATACGACTCACTATAGAGC-3′,

Utilizing a T7 promoter, RNA was transcribed using MegaScript followed by MegaClear (Ambion) according to the manufacturer's protocol.

Folding conditions

Each experiment was performed in 100 mM KCl and in the presence or absence of 10 mM $MgCl_2$. The first condition contained the following: 10 μL of RNA (1 μM concentration), 2.5 μL of 10X CE (10 mM K^+ cacodylate at pH 7.3/0.1 mM EDTA) buffer, 1.25 μL of 2 M KCl (Ambion Buffer Kit), and 11.25 μL of dH_2O. (Final volume: 25 μL of final concentrations 0.4 μM RNA, 1X CE buffer, and 100 mM KCl.) The second condition contained the same as above but also included 1.25 μL of 100 mM $MgCl_2$ (diluted from 1M $MgCl_2$ Ambion Buffer Kit) and only 10 μL of dH_2O. (Final volume: 25 μL of final concentrations 0.4 μM RNA, 1X CE buffer, 100 mM KCl, and 10 mM $MgCl_2$.)

The RNA mixture, 25 μL ($MgCl_2$ free), was heated at 90°C for 2 min. Samples were removed from heat and cooled to room temperature for 15 min. Samples were then placed at 50°C for 10 min and $MgCl_2$ was added to a final concentration of 10 mM and incubation continued at 50°C for 15–30 min. Then samples were folded at 37°C and incubated for at least 1 h. Similarly, an unfolded sample was prepared in an identical way, but $MgCl_2$ was not added.

DMS modification

To previously folded RNA, 0.5 μL of DMS solution (15 μL of 100% ethanol, 3 μL of Dimethyl sulfate [Sigma Aldrich]) was

added and the reaction was incubated for 2 min at 37°C. To stop the reaction, 475 μL of quench solution (28% β-mercaptoethanol and 0.3 M Na-OAc) was added. For precipitation, 1 mL of 100% ethanol was added and incubated overnight at −80°C.

T1 digest protocol

To 5 ρmol of RNA, 5.8 μL of 1X TE (10mM Tris, 1mM EDTA at pH 8) buffer was added. Samples were heated at 95°C for 2 min, followed by a 10-min incubation at 50°C. While at 50°C, 1 μL (1 U) of RNase T1 enzyme (Ambion) was added and incubation continued for 1 min. Samples were removed from heat and placed on ice to stop reaction. Following the digestion, samples were treated twice with phenol/chloroform and RNA was precipitated using 3 μL of 3 M Na-OAc and 400 μL of 100% ethanol.

Primer extension of cleaved/modified RNA

After each chemical modification, RNA was resuspended in 9 μL of annealing buffer (50 mM Tris at pH 8.3, 60 mM NaCl, 10 mM DTT [Dithiothreitol, threo-2,3-dihydroxy-1,4-dithiolbutane]) and 1 μL of 5 μM Cy5 labeled primer (5′ Cy5 reverse primer defined above for P4P6 and *Twort*) was added. Samples were heated at 90°C for 3 min and then slowly cooled to 25°C for 1.5 h for primer annealing. Then, 9 μL of reverse transcription mix (4 μL of 5X FS [First Strand] buffer supplied with Superscript III, 1 μL of 0.1 M DTT, 2 μL of RNase Inhibitor, 2 μL of 10 mM dNTP mix) was added to each tube. For dideoxy sequencing reactions, 6 μL of 5 mM ddNTP (GE Healthcare) was added to this mixture. Tubes were incubated at 55°C for 5 min, and then 1 μL (200 U) of Superscript III (Invitrogen) enzyme was added. The final reaction volume was 20 μL, which was incubated for an additional 15 min at 55°C. Upon completion of the reverse transcription extension, the following process was used to degrade the RNA: 2 μL of 2 M NaOH was added to the reaction and then incubated at 95°C for 3 min. The solution was then neutralized by adding 2 μL of 2 M HCl, followed by 3 μL of 3 M Na-OAc to aid in cDNA precipitation. Lastly, 1 μL of 100 mM $MgCl_2$ and 90 μL of 100% ethanol was added to the tube. Finally, the pellet was rinsed using 70% ethanol. The dried cDNA pellet was then resuspended in 60 μL of SLS (Beckman Coulter) and analyzed using a CEQ 8000. All experiments were run entirely duplicated to estimate overall experimental error. Data traces were integrated using the CAFA and Shapefinder software (Mitra et al. 2008; Vasa et al. 2008).

Computational methods

All computations were carried out on a Mac OS X server (10.5) using a combination of Python, Matlab, and Perl for analysis. All optimizations were carried out using an Ansi-C–implemented version of a nonlinear large-scale bounded least-squares optimization routine based on the interior reflective Newton method (Coleman and Li 1996). Error rates reported in Table 1 were computed using Equation 1:

$$\text{Error} = \frac{FP + FN}{TP + FP + FN + TN}, \quad (1)$$

where FP, FN, TP, and TN are the false positive and negative and true positive and negative rates, respectively. The Manhattan distance metric is the sum of the absolute differences between all elements of two vectors of equal length. We used the standalone version of sFold available from http://sfold.wadsworth.org for generating decoy structures. Multidimensional scaling was performed in the Matlab using the Statistics toolbox. We created a web server interface that allows users to upload and visualize their structure mapping data in the context of sFold sampling at http://cloud.wadsworth.org/mapfold. Two-dimensional structure visualizations were generated using the VARNA visualization applet (Darty et al. 2009).

We used the "fold RNA single strand" command of RNAstructure while constraining unpaired bases using our experimental T1 and DMS data for Table 2. Only residues having DMS or T1 reactivity above the optimum threshold were constrained as being "unpaired" in RNAstructure (Mathews et al. 2004). We determined thresholds by minimizing the error rate as computed in Equation 1. We therefore used the correct structure to determine the threshold and thus provide input data for RNAstructure. We report in Table 2 the range of PPV and sensitivity for RNAfold and RNAstructure obtained for each of our independently collected data sets.

SUPPLEMENTAL MATERIAL

Supplemental material can be found at http://www.rnajournal.org.

ACKNOWLEDGMENTS

We thank Somdeb Mitra, Joerg Schlatterer, and Michael Brenowitz (all at Albert Einstein College of Medicine) for their assistance with the structure mapping data collection and providing us with the *Twort* plasmid. We also thank Nathan Boyd (Stanford University) for the *Tetrahymena* plasmid. We thank Ye Ding (Wadsworth Center) for providing the source code for sFold. We also thank Dave Mathews (University of Rochester) for his insightful comments while reviewing this manuscript. This work was supported by grants R00 GM079953 (NIGMS) and R21 MH087336 (NIMH) to A.L., and grant R37-GM39422 (NIH) to A.B. and L.D.-N.

Received November 5, 2009; accepted February 20, 2010.

REFERENCES

Adilakshmi T, Lease RA, Woodson SA. 2006. Hydroxyl radical footprinting in vivo: Mapping macromolecular structures with synchrotron radiation. *Nucleic Acids Res* **34:** e64. doi: 10.1093/nar/gkl1291.

Bartley LE, Zhuang X, Das R, Chu S, Herschlag D. 2003. Exploration of the transition state for tertiary structure formation between an RNA helix and a large structured RNA. *J Mol Biol* **328:** 1011–1026.

Berman HM, Westbrook J, Feng Z, Iype L, Schneider B, Zardecki C. 2002. The Nucleic Acid Database. *Acta Crystallogr D Biol Crystallogr* **58:** 889–898.

Bernhart SH, Hofacker IL, Stadler PF. 2006. Local RNA base pairing probabilities in large sequences. *Bioinformatics* **22:** 614–615.

Brunel C, Romby P. 2000. Probing RNA structure and RNA-ligand complexes with chemical probes. *Methods Enzymol* **318:** 3–21.

Cannone JJ, Subramanian S, Schnare MN, Collett JR, D'Souza LM, Du Y, Feng B, Lin N, Madabusi LV, Muller KM, et al. 2002. The comparative RNA web (CRW) site: An online database of comparative sequence and structure information for ribosomal,

intron, and other RNAs. *BMC Bioinformatics* **3:** 2. doi: 10.1186/1471-2105-3-2.
Cate JH, Gooding AR, Podell E, Zhou K, Golden BL, Kundrot CE, Cech TR, Doudna JA. 1996. Crystal structure of a group I ribozyme domain: Principles of RNA packing. *Science* **273:** 1678–1685.
Coleman TF, Li Y. 1996. An interior, trust region approach for nonlinear minimization subject to bounds. *SIAM J Optim* **6:** 418–445.
Darty K, Denise A, Ponty Y. 2009. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **25:** 1974–1975.
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.
Deigan KE, Li TW, Mathews DH, Weeks KM. 2009. Accurate SHAPE-directed RNA structure determination. *Proc Natl Acad Sci* **106:** 97–102.
Ding Y, Chan CY, Lawrence CE. 2004. Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res* **32:** W135–W141.
Ding Y, Chan CY, Lawrence CE. 2005. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA* **11:** 1157–1166.
Dolnik V. 1999. DNA sequencing by capillary electrophoresis (review). *J Biochem Biophys Methods* **41:** 103–119.
Donis-Keller H, Maxam AM, Gilbert W. 1977. Mapping adenines, guanines, and pyrimidines in RNA. *Nucleic Acids Res* **4:** 2527–2538.
Ehresmann C, Baudin F, Mougel M, Romby P, Ebel JP, Ehresmann B. 1987. Probing the structure of RNAs in solution. *Nucleic Acids Res* **15:** 9109–9128.
Golden BL, Kim H, Chase E. 2005. Crystal structure of a phage Twort group I ribozyme-product complex. *Nat Struct Mol Biol* **12:** 82–89.
Harkins EW. 2001. References to commonly used techniques. In *Current protocols in nucleic acid chemistry* (ed. SL Beaucage et al.), Appendix 3A. Wiley, New York.
Hart JM, Kennedy SD, Mathews DH, Turner DH. 2008. NMR-assisted prediction of RNA secondary structure: Identification of a probable pseudoknot in the coding region of an R2 retrotransposon. *J Am Chem Soc* **130:** 10233–10239.
Laederach A, Shcherbakova I, Liang M, Brenowitz M, Altman RB. 2006. Local kinetic measures of macromolecular structure reveal partitioning among multiple parallel pathways from the earliest steps in the folding of a large RNA molecule. *J Mol Biol* **358:** 1179–1190.
Laederach A, Shcherbakova I, Jonikas MA, Altman RB, Brenowitz M. 2007. Distinct contribution of electrostatics, initial conformational ensemble, and macromolecular stability in RNA folding. *Proc Natl Acad Sci* **104:** 7045–7050.
Latham JA, Cech TR. 1989. Defining the inside and outside of a catalytic RNA molecule. *Science* **245:** 276–282.
Lemieux S, Major F. 2002. RNA canonical and non-canonical base pairing types: A recognition method and complete repertoire. *Nucleic Acids Res* **30:** 4250–4263.
Leontis NB, Stombaugh J, Westhof E. 2002. The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res* **30:** 3497–3531.
Lockard RE, Kumar A. 1981. Mapping tRNA structure in solution using double-strand-specific ribonuclease V1 from cobra venom. *Nucleic Acids Res* **9:** 5125–5140.
Mathews DH. 2004. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA* **10:** 1178–1190.
Mathews DH, Banerjee AR, Luan DD, Eickbush TH, Turner DH. 1997. Secondary structure model of the RNA recognized by the reverse transcriptase from the R2 retrotransposable element. *RNA* **3:** 1–16.
Mathews DH, Sabina J, Zuker M, Turner DH. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* **288:** 911–940.
Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci* **101:** 7287–7292.
Mitra S, Shcherbakova IV, Altman RB, Brenowitz M, Laederach A. 2008. High-throughput single-nucleotide structural mapping by capillary automated footprinting analysis. *Nucleic Acids Res* **36:** e63. doi: 10.1093/nar/gkn267.
Morton NE. 2008. Into the post-HapMap era. *Adv Genet* **60:** 727–742.
Onoa B, Dumont S, Liphardt J, Smith SB, Tinoco I Jr, Bustamante C. 2003. Identifying kinetic barriers to mechanical unfolding of the *T. thermophila* ribozyme. *Science* **299:** 1892–1895.
Reeder J, Hochsmann M, Rehmsmeier M, Voss B, Giegerich R. 2006. Beyond Mfold: Recent advances in RNA bioinformatics. *J Biotechnol* **124:** 41–55.
Shcherbakova I, Brenowitz M. 2005. Perturbation of the hierarchical folding of a large RNA by the destabilization of its Scaffold's tertiary structure. *J Mol Biol* **354:** 483–496.
Shcherbakova I, Brenowitz M. 2008. Monitoring structural changes in nucleic acids with single residue spatial and millisecond time resolution by quantitative hydroxyl radical footprinting. *Nat Protoc* **3:** 288–302.
Shcherbakova I, Gupta S, Chance MR, Brenowitz M. 2004. Monovalent ion-mediated folding of the *Tetrahymena thermophila* ribozyme. *J Mol Biol* **342:** 1431–1442.
Shcherbakova I, Mitra S, Laederach A, Brenowitz M. 2008. Energy barriers, pathways, and dynamics during folding of large, multi-domain RNAs. *Curr Opin Chem Biol* **12:** 655–666.
Takamoto K, Chance MR, Brenowitz M. 2004. Semi-automated, single-band peak-fitting analysis of hydroxyl radical nucleic acid footprint autoradiograms for the quantitative analysis of transitions. *Nucleic Acids Res* **32:** e119. doi: 10.1093/nar/gnh117.
Tijerina P, Mohr S, Russell R. 2007. DMS footprinting of structured RNAs and RNA-protein complexes. *Nat Protoc* **2:** 2608–2623.
Uhlenbeck OC. 1995. Keeping RNA happy. *RNA* **1:** 4–6.
Vary CP, Vournakis JN. 1984. RNA structure analysis using T2 ribonuclease: Detection of pH and metal ion induced conformational changes in yeast tRNAPhe. *Nucleic Acids Res* **12:** 6763–6778.
Vasa SM, Guex N, Wilkinson KA, Weeks KM, Giddings MC. 2008. ShapeFinder: A software system for high-throughput quantitative analysis of nucleic acid reactivity information resolved by capillary electrophoresis. *RNA* **14:** 1979–1990.
Vicens Q, Gooding AR, Laederach A, Cech TR. 2007. Local RNA structural changes induced by crystallization are revealed by SHAPE. *RNA* **13:** 536–548.
Waldispuhl J, Clote P. 2007. Computing the partition function and sampling for saturated secondary structures of RNA, with respect to the Turner energy model. *J Comput Biol* **14:** 190–215.
Weinstock GM. 2007. ENCODE: More genomic empowerment. *Genome Res* **17:** 667–668.
Wilkinson KA, Merino EJ, Weeks KM. 2005. RNA SHAPE chemistry reveals nonhierarchical interactions dominate equilibrium structural transitions in tRNA(Asp) transcripts. *J Am Chem Soc* **127:** 4659–4667.
Wilkinson KA, Gorelick RJ, Vasa SM, Guex N, Rein A, Mathews DH, Giddings MC, Weeks KM. 2008. High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states. *PLoS Biol* **6:** e96. doi: 10.1371/journal.pbio.0060096.
Wilkinson KA, Vasa SM, Deigan KE, Mortimer SA, Giddings MC, Weeks KM. 2009. Influence of nucleotide identity on ribose 2'-hydroxyl reactivity in RNA. *RNA* **15:** 1314–1321.
Zhuang X. 2005. Single-molecule RNA science. *Annu Rev Biophys Biomol Struct* **34:** 399–414.

Evaluation of the information content of RNA structure mapping data for secondary structure prediction

Scott Quarrier, Joshua S. Martin, Lauren Davis-Neulander, et al.

RNA 2010 16: 1108-1117 originally published online April 22, 2010
Access the most recent version at doi:10.1261/rna.1988510