BIOCHEMISTRY
including biophysical chemistry & molecular biology

Article

pubs.acs.org/biochemistry

Principles for Understanding the Accuracy of SHAPE-Directed RNA Structure Modeling

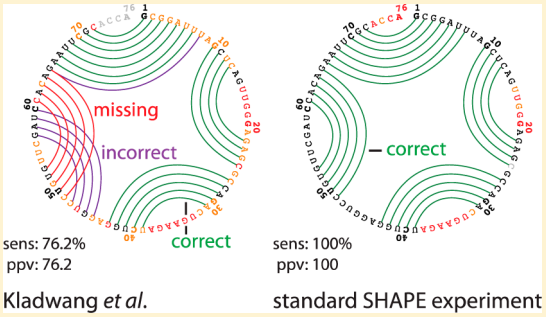
Christopher W. Leonard,[†] Christine E. Hajdin,[†] Fethullah Karabiber,[†] David H. Mathews,[∥] Oleg V. Favorov,[‡] Nikolay V. Dokholyan,[§] and Kevin M. Weeks*[,†]

[†]Department of Chemistry, [‡]Department of Biomedical Engineering, and [§]Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, North Carolina 27599-3290, United States

[∥]Department of Biochemistry and Biophysics, University of Rochester Medical Center, Rochester, New York 14642, United States

Supporting Information

ABSTRACT: Accurate RNA structure modeling is an important, incompletely solved, challenge. Single-nucleotide resolution SHAPE (selective 2′-hydroxyl acylation analyzed by primer extension) yields an experimental measurement of local nucleotide flexibility that can be incorporated as pseudo-free energy change constraints to direct secondary structure predictions. Prior work from our laboratory has emphasized both the overall accuracy of this approach and the need for nuanced interpretation of modeled structures. Recent studies by Das and colleagues [Kladwang, W., et al. (2011) *Biochemistry 50*, 8049; *Nat. Chem. 3*, 954], focused on analyzing six small RNAs, yielded poorer RNA secondary structure predictions than expected on the basis of prior benchmarking efforts. To understand the features that led to these divergent results, we re-examined four RNAs yielding the poorest results in this recent work: tRNA^{Phe}, the adenine and cyclic-di-GMP riboswitches, and 5S rRNA. Most of the errors reported by Das and colleagues reflected nonstandard experiment and data processing choices, and selective scoring rules. For two RNAs, tRNA^{Phe} and the adenine riboswitch, secondary structure predictions are nearly perfect if no experimental information is included but were rendered inaccurate by the SHAPE data of Das and colleagues. When best practices were used, single-sequence SHAPE-directed secondary structure modeling recovered ∼93% of individual base pairs and >90% of helices in the four RNAs, essentially indistinguishable from the results of the mutate-and-map approach with the exception of a single helix in the 5S rRNA. The field of experimentally directed RNA secondary structure prediction is entering a phase focused on the most difficult prediction challenges. We outline five constructive principles for guiding this field forward.



The functions of most RNA molecules are critically dependent on their structures, which are difficult to predict from first principles. An important first step in characterizing an RNA structure is to develop an accurate view of the pattern of base pairing or secondary structure. Recent work has emphasized that incorporation of nucleotide-resolution structural information obtained from chemical probing experiments dramatically improves the accuracy of secondary structure prediction.[1−4]

SHAPE (selective 2′-hydroxyl acylation analyzed by primer extension) is a chemical probing technology that measures local nucleotide flexibility in RNA as the ability of the ubiquitous 2′-hydroxyl group to form covalent adducts with electrophilic reagents.[5] SHAPE reagents react similarly with the four RNA nucleotides[6] and in a way that is strongly correlated with model-free measurements of molecular order.[7,8] Interactions that constrain nucleotide dynamics, including base pair formation, reduce the reactivity of the 2′-hydroxyl. SHAPE reactivities are thus roughly inversely proportional to the probability that a nucleotide forms a base pair. It is possible to devise a pseudo-free energy change term, ΔG_{SHAPE}[4,9] that, in conjunction with nearest-neighbor and other free energy terms,[10] can be used to direct prediction of RNA secondary structures. Overall, SHAPE-directed prediction has proven to yield significant improvements in RNA secondary structure prediction.

The field of experimentally directed RNA structure prediction is undergoing rapid advances, and it is important to benchmark these emerging methods using diverse and structurally challenging RNAs. Recent work by Das and colleagues evaluated SHAPE-directed secondary structure prediction using six small RNAs and also proposed a bootstrapping approach for the de novo identification of highly probable individual helices in the context of a larger structure prediction.[11] There are serious difficulties with both analyses. Bootstrapping entails resampling a given set of data with replacement and generally requires each data element to be independent of its order. For SHAPE data, both the measured

Received: June 7, 2012
Revised: September 12, 2012
Published: January 14, 2013

   dx.doi.org/10.1021/bi300755u | *Biochemistry* 2013, 52, 588−595

reactivities and their correct nucleotide positions are required for secondary structure modeling. Bootstrapping is thus inherently unsuitable for estimating helix-by-helix confidences for SHAPE (or any chemical probing) data (DOI: 10.1021/bi300756s). In this work, we examine the importance of experimental practices, data processing pipeline, and the need for consistent standards in evaluating RNA structure prediction. We conclude by outlining five principles that should guide future work designed to evaluate high-content, experimentally directed RNA structure prediction.

EXPERIMENTAL PROCEDURES

RNA Synthesis. All RNAs were synthesized from double-stranded DNAs generated by the polymerase chain reaction using single-stranded DNA templates (IDT) spanning the full length of the transcript, preceded by a 17-nucleotide T7 promoter and a 14-nucleotide 5′ cassette sequence, followed by a 43-nucleotide reverse transcription primer binding site.[12] RNAs were transcribed in 80 mM Hepes (pH 8.0), 40 mM DTT, 20 mM $MgCl_2$, 2 mM spermidine, 0.01% Triton X-100, 2 mM dNTP, and 0.1 mg/mL T7 polymerase at 37 °C for 3 h. Transcript RNA was precipitated and purified on 8% denaturing Tris-borate gels. Bands were visualized by UV shadowing, and RNAs were eluted overnight into water at 4 °C. Concentrations were calculated from the absorbance at 260 nm measured using a Nanodrop 2000c spectrophotometer.

RNA Constructs. Five RNAs were examined. We analyzed two variations of the *Vibrio cholera* cyclic-di-GMP riboswitch RNA, a "short P1" RNA corresponding to the RNA in the 3iwn crystal structure[13] and a "long P1" that both extends the P1 helix by 2 bp and truncates the U1A protein-binding site. The latter was the same sequence as the construct selectively used by Das and colleagues to evaluate the mutate-and-map method.[14] Sequences of *Escherichia coli* $tRNA^{Phe}$, the *Vibrio vulnificus* adenine riboswitch, and *E. coli* 5S rRNA corresponded to those evaluated in crystallographic studies[15,16] and used by Kladwang et al.[11]

SHAPE Experiments. Experiments with $tRNA^{Phe}$ were conducted using the folding buffer, SHAPE reagent concentrations, and other experimental conditions exactly as described by Kladwang et al.[11] or as outlined in standard SHAPE approaches.[9,12,17] For the approach of Das and co-workers, 2 pmol of RNA was denatured by heating at 95 °C for 1 min, snap-cooled on ice, and then refolding in 50 mM Hepes (pH 8.0) and 10 mM $MgCl_2$ for 30 min at 37 °C (labeled buffer A in Figure 1). We then added 20 μL of refolded RNA to the following volumes and concentrations of SHAPE reagents freshly dissolved in dry DMSO: 5 μL of 135 mM NMIA (Invitrogen, M25), 5 or 2 μL of 30 mM NMIA, or 5 or 2 μL of 30 mM 1M7. Reactions were performed at 24 and 37 °C with reaction times of 30 min for NMIA and 3 min for 1M7. No-reagent control reactions were performed identically, using neat DMSO rather than a reagent solution. For standard SHAPE experiments, the same RNA denaturing and refolding reactions were performed in 50 mM Hepes (pH 8.0), 200 mM potassium acetate (pH 7.7), and 3 mM $MgCl_2$ (labeled buffer B in Figure 1). Reactions were performed at 24 and 37 °C by addition of 10 μL of RNA to 1 μL of 30 mM 1M7 and incubation for 3 min. Reaction mixtures with the adenine or cyclic-di-GMP riboswitch contained 5 μM adenine or 10 μM cyclic-di-GMP (Biolog C057-01), respectively, and reactions were initiated via addition of 10 μL of refolded RNA to 1 μL of 30 mM 1M7; reaction mixtures were incubated for 3 min at 37 °C. In all cases, RNAs were recovered by the addition of 15 μL of water, 4 μL of 5 M NaCl, and 120 μL of 100% ethanol, followed by incubation at −80 °C for 10 min and centrifugation (14k rpm in a microfuge at 4 °C for 15 min). RNA was then resuspended in 10 μL of water. All RNAs were probed in two to four fully independent replicate experiments, in some cases performed months apart. Structure annotations shown in the individual figures correspond to single data sets, but all independently analyzed data sets yielded identical lowest-free energy structures. Data for the specificity domain of *Bacillus subtilis* RNase P were reported previously.[3]

Primer Extension. Reverse transcription reaction mixtures were prepared with the addition of 5 μL of 0.5 μM FAM-labeled (SHAPE-modified RNA trace) or JOE-labeled (sequencing trace) reverse transcription primer to 10 μL of the RNA solution. Primers were annealed by incubation at 65 °C for 3 min and at 42 °C for 2 min. Primer extension reactions were performed exactly as described previously[9,17] using SuperScript III (Invitrogen). Reverse transcription proceeded with incubation at 52 °C for 5 min and then at 65 °C for 5 min. SHAPE-modified samples were combined with sequencing reaction mixtures, precipitated with ethanol, resuspended in 10 μL of Hi-Di formamide (Applied Biosystems), heated with the tube cap open at 95 °C for 3 min, and resolved on an Applied Biosystems 3500 Genetic Analyzer capillary electrophoresis instrument.

Data Analysis and SHAPE-Directed RNA Structure Modeling. Raw capillary electrophoresis traces were processed using *ShapeFinder*[18] or a new custom software, *QuShape* (manuscript in press; software is available immediately at http://www.chem.unc.edu/rna/qushape). Structure predictions were identical for data processed by either approach. We attempted to process our experiments using HiTRACE[19] but were unable to run the publicly available version of the software. RNA structure prediction was performed using *RNAstructure*[20] versions 5.2−5.4 under either Mac OS 10.6.x or Unix following box-plot normalization, exactly as described previously.[4] ΔG_{SHAPE} parameters were 2.6 and −0.8 kcal/mol, the current default values in *RNAstructure*. We report the single lowest-free energy structure output by *RNAstructure* in each case. All data sets generated in this work are available at the SNRNASM community structure probing database.[21] SHAPE data reported by Kladwang et al.[11] were obtained from the RMDB (http://rmdb.stanford.edu) and normalized by the box-plot approach.[4] As reported previously,[11] we also could not identify an alternative mathematical manipulation that could convert the data of Das and colleagues into a form that would yield a fully correct secondary structure prediction for $tRNA^{Phe}$. RNA circle graphs were generated using *CircleCompare*, available as part of the *RNAstructure* package.

RESULTS AND DISCUSSION

In their evaluation of SHAPE-directed structure modeling, Das and colleagues reported an overall sensitivity (percentage of known base pairs predicted) for six small RNAs of 83% with a positive predictive value (percentage of predicted pairs in the accepted structure) of ∼80%.[11] Although these values represented substantial improvement over predictions achieved in the absence of SHAPE data (62 and 55%, respectively), we were surprised by these results because they were comparable to the very poorest predictions that we have obtained in extensive analyses focused on highly challenging RNAs. In addition, relatively poor SHAPE-directed predictions were

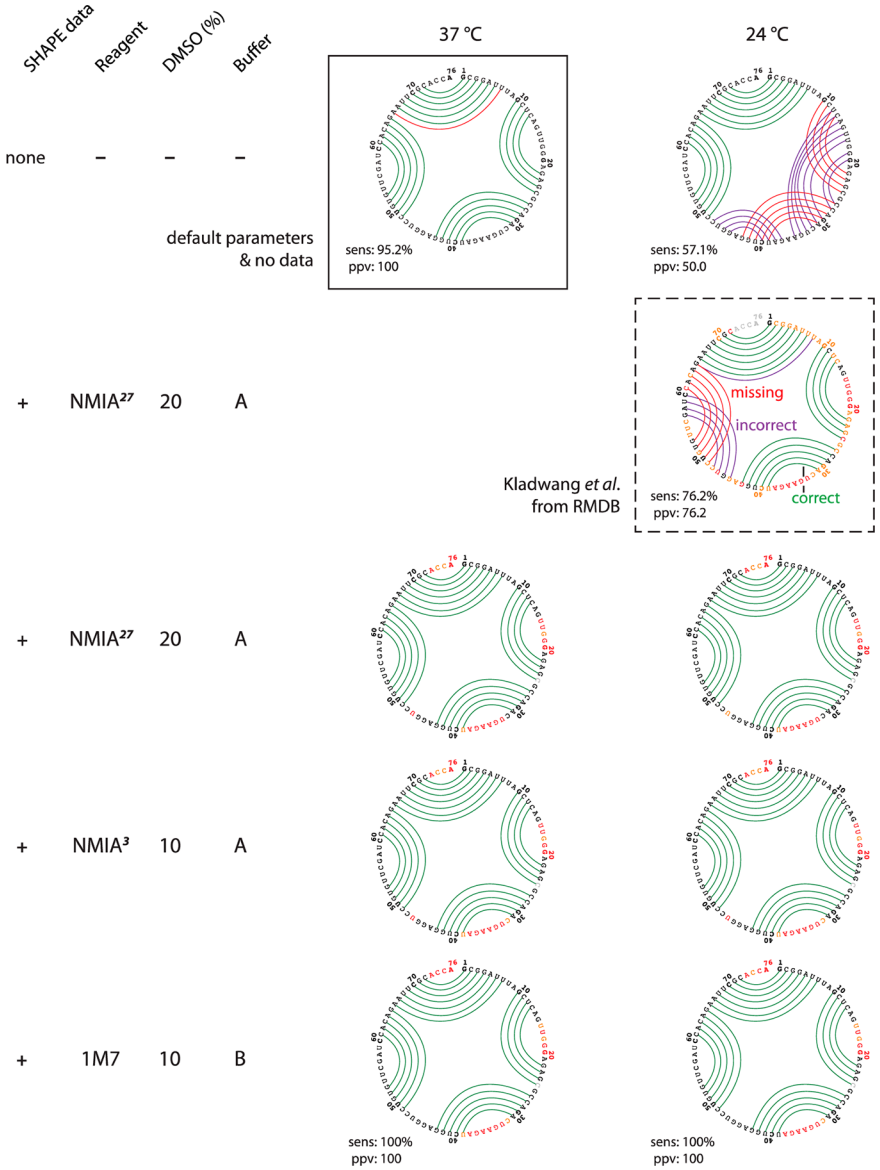


Figure 1. Evaluation of thermodynamically based and experimentally directed secondary structure modeling for tRNA^{Phe}. Secondary structure predictions were performed using *RNAstructure*[20] without experimental data (line 1) or with SHAPE data acquired as a function of experimental conditions. Experimental conditions were based either on those reported by Kladwang et al.[11,14] or on those developed by our laboratory.[9,12,17] Experimental variations included the reagent (NMIA or 1M7), the concentration of NMIA (either ~27 or 3 mM, indicated with superscripts 27 and 3, respectively; the 27 mM condition, used by Das and colleagues, results in the formation of a visible precipitate), the percentage of DMSO cosolvent (10 or 20%), and the solution buffer conditions (buffers A and B are reported in refs 11 and 12, respectively). Data of Kladwang et al., obtained from the RMDB, were processed with HiTRACE.[19] For all figures, if SHAPE data were used in the secondary structure prediction, nucleotides are colored by reactivity: low, medium, and high reactivities colored black, yellow, and red, respectively. Correct, missed and incorrect base pairs are shown with green, red and magenta arcs, respectively. The sensitivity (sens) and positive predictive value (ppv) are indicated for representative structures.

reported for *E. coli* tRNA^{Phe} and the *V. vulnificus* adenine riboswitch RNA even though, in our experience, the structures of RNAs with similar (simple) topologies are accurately predicted when SHAPE data are used to direct structure modeling.

We therefore performed SHAPE and used our data to direct secondary structure prediction for the four RNAs whose structures were predicted especially poorly.[11] We will emphasize the predicted lowest-free energy structure in each case. Throughout this work, structure predictions are presented in the form of RNA circle graphs.[20,22] In these plots, the RNA sequence is displayed around the circumference of a circle. If SHAPE data were used to direct the secondary structure prediction, the letters corresponding to each nucleotide are colored by their SHAPE reactivity (Figure 1). Base pairs are drawn as arcs; a series of parallel arcs indicates a helix. Base pairs and helices that are correct relative to the accepted structure are colored green, whereas missed (false negative) and incorrectly predicted (false positive) base pairs are colored red and magenta, respectively. A fully correct structure would therefore have only green arcs.

Case I: tRNA^{Phe} and the Adenine Riboswitch. tRNA^{Phe} and the adenine riboswitch will be discussed as a single case as similar issues were identified in structure prediction for both

dx.doi.org/10.1021/bi300755u | *Biochemistry* 2013, 52, 588−595

RNAs. If the sequence of $tRNA^{Phe}$ is submitted to *RNAstructure* (version 5.2 or 5.3) using default parameters and no experimental data, then the lowest-free energy predicted structure conforms almost exactly to the accepted structure with the exception of a single missed base pair (Figure 1, solid box). If the SHAPE data obtained by Das and colleagues are used to direct folding, one of the four helices in this RNA is missed (Figure 1, dashed box).[11] We obtained a similar result in our analysis of the structure of the adenine riboswitch. This structure is predicted perfectly without data. Use of the data generated by Das and colleagues reduced the prediction accuracy and yielded one false positive helix (Figure S1 of the Supporting Information).

These initial results were striking at two levels. First, these two RNAs have relatively simple topologies and are the kinds of RNA that are usually predicted with high accuracy by SHAPE-directed modeling. Second, to the best of our knowledge, these are unique examples in which the addition of nucleotide-resolution chemical probing information caused RNA secondary structure predictions to become less accurate.

We then performed SHAPE on $tRNA^{Phe}$ using the standard approach developed by our laboratory.[4,9,12,17,18,23] When these SHAPE data were used to direct structure prediction, the lowest-free energy structure for $tRNA^{Phe}$ coincided exactly with the accepted structure (Figure 1, bottom row). We therefore explored the differences between the standard SHAPE approach and the version used by Das and colleagues.

The experimental procedure used by Das and colleagues differed from our published approach in at least four ways. (1) Probing experiments were performed at 24 °C. At this temperature, the thermodynamic parameters for RNA loops and junctions are less accurate than at the standard temperature of 37 °C.[24] (2) The *N*-methylisatoic anhydride (NMIA) SHAPE reagent was used at a final concentration of 4.8 mg/mL. NMIA is not fully soluble at this concentration and forms a visible precipitate during the reaction. In addition, the reactivity of NMIA is sensitive to the specific ion environment[3] and NMIA preferentially reacts with nucleotides experiencing slow dynamics.[25,26] In our experience, 1-methyl-7-nitroisatoic anhydride (1M7)[3] is the probe of choice for this type of analysis; 1M7 yields more quantitatively accurate RNA structural information than NMIA. (3) Experiments were performed in 20% (v/v) DMSO cosolvent. DMSO denatures some nucleic acid structures at this concentration.[27,28] (4) Experiments were performed in a buffer different from that initially used for SHAPE-directed structure probing; however, especially with the 1M7 reagent,[3] we did not expect this to significantly change the quality of RNA structure prediction.

We systematically varied these parameters to understand the large differences in SHAPE-directed secondary structure prediction obtained by the two laboratories. Consistent with the known lower accuracy of current thermodynamic rules at temperatures other than 37 °C, the no-data prediction accuracy for $tRNA^{Phe}$ was notably poorer at 24 °C than at 37 °C (in Figure 1, first row, compare boxed and unboxed structures). We then performed SHAPE experiments under exactly the conditions reported by Das and colleagues, including the high concentration of NMIA, 20% DMSO, and 24 °C. The SHAPE reactivity pattern had a notably higher fraction of moderately and highly reactive nucleotides than those obtained under our standard conditions (Figure 1, yellow and red positions, in row 3). Nevertheless, these SHAPE data resulted in a predicted RNA secondary structure that agreed with the accepted model. SHAPE data obtained under both standard conditions and those of Das and colleagues for the adenine riboswitch also resulted in a structure that agreed with the accepted model (Figure S1 of the Supporting Information).

Although we strongly recommend the use of the 1M7 reagent, use of fully soluble reagent concentrations, and maintaining organic cosolvent concentrations below 10%, at least in the cases of $tRNA^{Phe}$ and the adenine riboswitch, SHAPE-directed RNA secondary structure prediction proved to be robust under these experimental conditions. In summary, formally identical experiments in our lab could not reproduce the poor secondary structure predictions reported by Das and colleagues.

This analysis of $tRNA^{Phe}$ and the adenine riboswitch RNAs suggests that differences in SHAPE-directed secondary structure modeling accuracy reflected differences in data processing approaches. The Das lab used HiTRACE,[19] which appears to undercorrect for background and assumes that signal decay in the primer extension step is the same for all RNAs, which is unlikely. The approach implemented in HiTRACE ultimately yielded over-reactive SHAPE profiles with few or no unreactive positions (Figure 1, dashed box, and Figure S2 of the Supporting Information) and disrupted the otherwise strong relationship between SHAPE reactivity and the probability that a nucleotide is base paired.

Case II: Cyclic-di-GMP Riboswitch. One of the objectives of the analysis of RNA secondary structure modeling undertaken by Das and colleagues was to compare single-sequence SHAPE-directed prediction to an information-rich mutate-and-map approach in which SHAPE data are collected for a large group of sequence variants in which all possible nucleotides in a given RNA are mutated at least once.[14,29] Das and colleagues used different RNA constructs for the cyclic-di-GMP (c-di-GMP) riboswitch to evaluate single-sequence SHAPE-directed structure prediction and the mutate-and-map approach. These constructs differed in the length of their P1 helices. The first construct corresponded closely to that used in two independent crystallographic studies[13,30] and featured a short, bulged P1 helix of 4 bp (in Figure 2, P1 is emphasized with brackets). In both crystals, the P1 helix forms extensive crystal contacts with the net effect of substantially stabilizing this helix; the adjacent CAC bulge also forms crystal contacts that ultimately define its local conformation. As noted by the authors of one of the crystallographic studies, these features, shared between two distinct crystal structures, "are indicative of a helix that possesses some measure of instability ... consistent with its anticipated function as a molecular switch."[31] In evaluating the mutate-and-map approach, Das and colleagues used a different cyclic-di-GMP RNA in which helix P1 was extended by 2 bp to yield a much more stable 6 bp helix. This change from a short to a long P1 helix has a dramatic effect on the structure of the RNA and increases the affinity for the cyclic-di-GMP ligand by a large factor.[31]

In the absence of SHAPE data, the structures of both short and long P1 forms of the cyclic-di-GMP riboswitch RNA are predicted fairly well with the exception of the region at or near P1; prediction sensitivities are ~81 and 75%, respectively (Figure 2, row 1). We analyzed both the short and long P1 variants by SHAPE and used our data to guide RNA secondary structure prediction. In the case of the short P1 variant, incorporation of SHAPE data modestly improved the accuracy relative to the no-data case. Most nucleotides whose pairing partners were predicted incorrectly are involved in the likely

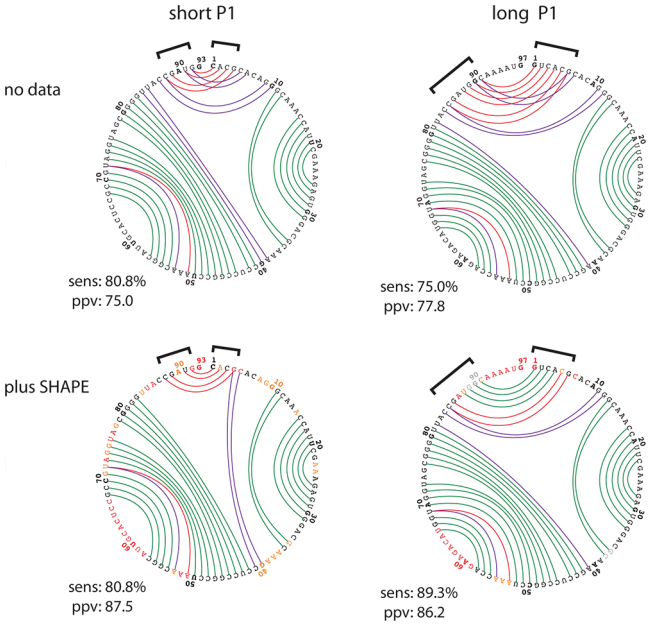


Figure 2. SHAPE-directed secondary structure modeling for the cyclic-di-GMP riboswitch as a function of the length of helix P1. The position of helix P1 is emphasized with brackets.

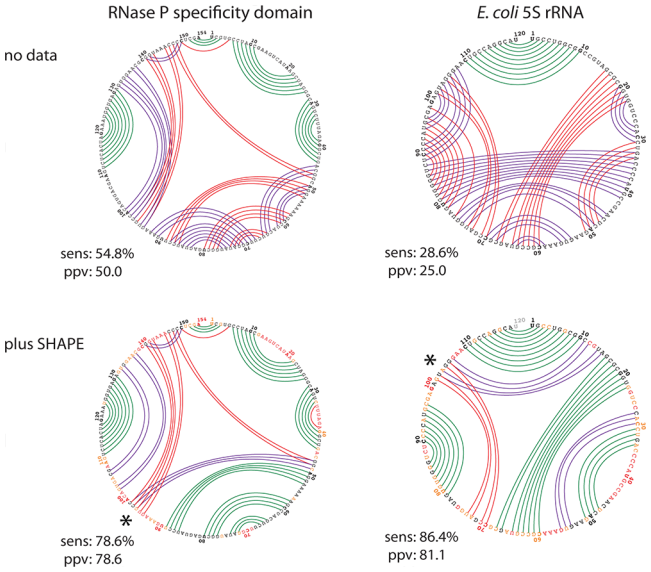


Figure 3. SHAPE-directed modeling for two highly challenging RNAs, the specificity domain of RNase P and *E. coli* 5S rRNA. The single helix, whose misprediction dominates modeling errors in each case, is highlighted with an asterisk.

dynamic P1 helix. When SHAPE data were used to direct folding of the more stable RNA containing the longer P1 helix, the resulting lowest-free energy structure has an overall sensitivity of 89%, mispairs three base pairs in P1, and is exactly the same as that predicted by the Das group using the mutate-and-map approach (Figure 2, row 2).

These results emphasize that differences in RNA sequence outside the core region of interest can have large effects on the stability of a given structure, its fundamental SHAPE reactivity, and the resulting secondary structure prediction. In summary, for the cyclic-di-GMP RNA, single-sequence SHAPE-directed structure prediction and mutate-and-map modeling, the latter requiring two orders-of-magnitude more distinct probing experiments, yielded identical secondary structure predictions if the same RNAs and same scoring rules were used.

Case III: *B. subtilis* RNase P Specificity Domain and *E. coli* 5S rRNAs. SHAPE-directed RNA secondary structure prediction yields highly successful predictions in many cases, including for the 1542-nucleotide *E. coli* 16S rRNA[4] and for $tRNA^{Phe}$ and the adenine and cyclic-di-GMP riboswitch RNAs as described here (Figures 1 and 2 and Figure S1 of the Supporting Information). However, there are a few RNAs that remain refractory to concise SHAPE-directed prediction.[4,5,9] To date and in ongoing work, we have evaluated dozens of RNAs spanning thousands of nucleotides. Our two consistently poorest predictions are those for the RNase P specificity domain[3,9] and for the *E. coli* 5S rRNA (Figure 3).

In the absence of experimental data, the *RNAstructure* algorithm predicts the structures of these two RNAs with sensitivities of 55 and 29%, respectively. Our SHAPE-directed prediction yielded substantial improvements, resulting in sensitivities of 79% for the RNase P domain and 86% for the 5S rRNA (Figure 3). The SHAPE-directed models for the RNase P and 5S rRNA thus represent large improvements, but in strong contrast to the small errors observed for many other RNAs, prediction errors are significant. In both cases, the major prediction error stems from misassignment of a single helix (Figure 3, row 2, emphasized with asterisks). Incorrect prediction of one helix causes errors that propagate through each structure. Both of these RNAs function only in the context of binding by obligate protein cofactors, and the RNase P domain required 80 mM $SrCl_2$ to form crystals,[32] features that may partially explain the challenge of predicting structures for these RNAs. In both cases, at least one strand of the misassigned helix contains nucleotides with relatively high SHAPE reactivities, suggestive of semistable or a mixture of conformations. These two RNAs thus represent challenges to SHAPE-directed secondary structure modeling but, notably, involve significant extenuating circumstances.

Case IV: Small Training Sets. All approaches for using experimental information to direct RNA secondary structure modeling require some kind of parametrization of the experimental data. In our experience, almost any RNA can be induced to fold properly with appropriate parameter choices. Thus, it is critical to guard against overoptimization. Das and colleagues have focused on and drawn strong conclusions from a small data set of six RNAs.[11,33] We therefore examined the role that optimization over a small data set might have in prediction accuracies.

Our first-generation parameters were optimized using ~2500 nucleotides in the *E. coli* 23S rRNA and generally work well for many different classes of RNA, including the 1542-nucleotide *E. coli* 16S rRNA[4] and diverse small RNAs (in Table 1, compare no data and global parameters columns). We used a leave-one-out jackknife approach to optimize parameters for a group of five RNAs that include four of the six RNAs in the Das training set, plus the bI3 P546 domain in place of the *Tetrahymena* example. We readily obtained a set of parameters that yielded near-perfect predictions for all five RNAs (in Table 1, see the small training set columns). The near-perfect predictions include those for the cyclic-di-GMP riboswitch and 5S rRNAs, for which we report errors (Figures 2 and 3).

In summary, the conclusions of recent papers[11,14,33] comparing SHAPE-directed modeling with other approaches would have been different if higher-quality data had been

dx.doi.org/10.1021/bi300755u | *Biochemistry* 2013, 52, 588−595

Table 1. Effect of Optimizing Chemical Probing Parameters over Small Data Sets[a]

		without SHAPE data			with SHAPE data			with SHAPE data		
		no data			global			small training set		
RNA	length (no. of nucleotides)	sens	ppv	geo	sens	ppv	geo	sens	ppv	geo
adenine riboswitch	71	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
$tRNA^{Phe}$	76	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**95**	**98**
cyclic-di-GMP riboswitch	97	55	49	52	89	86	88	**96**	**93**	**95**
5S rRNA, *E. coli*	120	29	25	27	86	81	84	**91**	**91**	**91**
P546 domain, bI3 intron	155	43	44	44	**96**	**98**	**97**	**95**	**96**	**96**
average		65	64	64	94	93	94	96	96	96

[a]Global parameters were $m = 2.6$ and $b = -0.8$, which give high-quality models for both small RNAs and kibobase length rRNAs.[4] The small training set parameters were $m = 1.3$ and $b = -0.3$. Accuracies of >90% are shown in bold. Prediction accuracies are given as the sensitivity (sens), the positive predictive value (ppv), and their geometric mean (geo).

obtained and if parameters for the SHAPE approach had been defined using the same small data set approach used to optimize the mutate-and-map[11,14] or DMS[33] method. In fact, this exploratory analysis (Table 1) suggests single-sequence SHAPE would have outperformed both mutate-and-map and DMS mapping.

Addendum. Das and colleagues recently published new data for SHAPE analyses of small RNAs, in which they have dramatically improved their workflow for processing capillary electrophoresis data.[33] Inspection of the new data shows that both background subtraction and signal decay corrections are more accurate. These data now show that >90% of all accepted base pairs are recovered for the same set of RNAs,[33] a substantial improvement. For example, with a more accurate data analysis pipeline, Das and colleagues now predict the structure of $tRNA^{Phe}$ nearly perfectly (Figure 4), corroborating the conclusions outlined above.

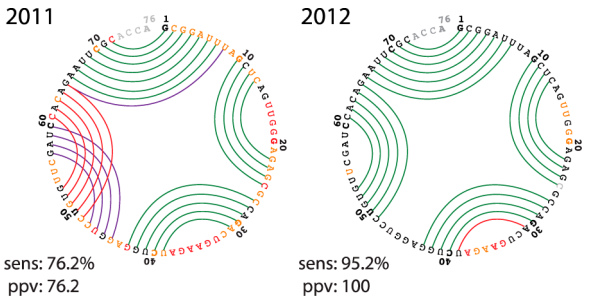


Figure 4. Comparison of SHAPE-directed modeling of $tRNA^{Phe}$, as published by Das and colleagues in September and December 2011[11,14] (left), vs that submitted in July 2012[33] (right). Models are annotated using the scheme in Figure 1.

PERSPECTIVE

Status of RNA Secondary Structure Prediction. Experimentally directed secondary structure prediction is emerging as a powerful approach for accurately modeling many RNA secondary structures, at least given the current, relatively small database of RNAs with well-defined structures. When single-nucleotide-resolution SHAPE data, in conjunction with nearest neighbor and other thermodynamic parameters, are used to drive secondary structure prediction, the median recovery of accepted base pairs exceeds 90% (refs 4 and 5, Figures 1−3, and Figure S1 of the Supporting Information). However, a few RNAs remain challenging to single-sequence prediction, and there are specific classes of important RNAs, including very highly structured RNAs and large RNAs containing pseudoknots, for which additional refinements to current algorithms are required to achieve accurate predictions.[4,9]

A high degree of nuance and care is required to fully analyze, understand, and minimize potential errors in secondary structure modeling.[5,34] In particular, the choice of experimental conditions, accurate data processing, and identical scoring rules are crucial (Figures 1−4, Figure S1 of the Supporting Information, and refs 11 and 14). In some cases, differences observed between an experimentally supported model and an accepted structure may, in fact, reflect bona fide structural differences reflecting thermodynamically accessible states, crystallization conditions, and contributions of (missing) protein cofactors.

Principles To Guide Evaluations of Secondary Structure Modeling. The field of experimentally directed RNA secondary structure modeling is entering a phase focused on refining structures for especially challenging targets. Current frontier challenges include independent benchmarking of RNAs with well-defined accepted structures, the potential for a priori identification of those helices that are the most well-defined by a given set of experimental information, and accurate modeling of long and full length RNA transcripts. The following principles should be emphasized as the field of experimentally directed RNA secondary structure prediction focuses on addressing the remaining challenges in modeling "hard" RNAs.

(1) Do No Harm. RNA structure modeling is sensitive to the precise sequence, specific solution environment, and data analysis pipeline. Many experimental details are likely to be important, and data need to be of high quality and processed accurately. There is often a trade-off between high throughput and structural accuracy. Any new experimentally directed prediction approach should be compared with folding analyses performed both in the absence of data and with conventional, more highly curated methods. The observation that prediction quality decreases with the use of experimental information or upon changes to the data analysis pipeline provides a strong cautionary signal.

(2) Evaluate Others As You Would Have Others Evaluate You. Comparisons between evolving modeling methods are important and appropriate; however, the same RNA sequences and the same scoring rules should be used in each case. Using less stable RNAs, more stringent rules, or inaccurate data processing when evaluating different algorithms

dx.doi.org/10.1021/bi300755u | *Biochemistry* 2013, 52, 588−595

and modeling approaches will not provide the information needed to advance RNA secondary structure prediction.

(3) Value Concision. One of the most important recent insights in RNA secondary (and tertiary) structure modeling is that addition of experimental information can dramatically improve the quality of the resulting structure predictions. There is a critical balance to be struck between creating experimental approaches that are information-rich yet remain tractable and readily implementable by nonexpert laboratories. High value should be placed on methods that scale gracefully to large RNAs and that can interrogate authentic biological transcripts.

(4) Recognize That All That Glitters (structurally) Is Not Gold. Ideally, there would exist a large database of complex RNAs of diverse lengths, whose in-solution structures were well-established. Unfortunately, there are very few accepted RNA secondary structures that meet this criterion. Every high-resolution RNA structure is the product of careful sequence selection and intense experimental optimization, and there is abundant evidence that conditions used in high-resolution crystallography and nuclear magnetic resonance studies impose large constraints on RNA structure, ultimately stabilizing some local conformations that may not be dominant in solution and limiting the classes of RNA amenable to study.[31,35−38] In some cases, an RNA may exist in an equilibrium between multiple structures, and it is an over-simplification to focus on a single low-free energy structure. Ultimately, nuance is required to evaluate the final (often few) distinctions between modeled and accepted structures in complex RNAs.

(5) Appreciate the Size of the RNA World. RNAs with higher-order structure likely span a broad continuum. Some structures are compact, involve many noncanonical tertiary interactions, and are highly stable. Most high-resolution structures in current databases fall into this category. Many complex RNAs, with significant underlying structure that affects their biological functions,[39,40] are not amenable to current high-resolution structure determination approaches, however. Accurate refinement of secondary structure models for these dynamic, but clearly structured, RNAs is important. This critical goal will be met only by including both diverse classes of RNAs and, especially, large RNAs in the training sets used to develop structure modeling algorithms. Structures of large messenger and noncoding RNAs are unlikely to be as well-defined as those of RNAs that can be studied by atomic-resolution approaches, and again, nuance will be required to interpret the successes and limitation of large-scale modeling approaches.

■ ASSOCIATED CONTENT

Supporting Information

Two figures. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: weeks@unc.edu. Phone: (919) 962-7486.

Author Contributions

C.W.L. and C.E.H. contributed equally to this work.

Funding

This work was supported by grants from the National Institutes of Health (AI068462 to K.M.W., GM080742 to N.V.D., and GM076485 to D.H.M.) and the National Science Foundation (MCB-1121024 to K.M.W.).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We are indebted to D. Turner for a thoughtful reading of the manuscript and to R. Das and colleagues for making structure probing data available at the RMDB.

■ REFERENCES

(1) Mathews, D. H., Sabina, J., Zuker, M., and Turner, D. H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol. 288*, 911−940.

(2) Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M., and Turner, D. H. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. U.S.A. 101*, 7287−7292.

(3) Mortimer, S. A., and Weeks, K. M. (2007) A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J. Am. Chem. Soc. 129*, 4144−4145.

(4) Deigan, K. E., Li, T. W., Mathews, D. H., and Weeks, K. M. (2009) Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. U.S.A. 106*, 97−102.

(5) Weeks, K. M., and Mauger, D. M. (2011) Exploring RNA structural codes with SHAPE chemistry. *Acc. Chem. Res. 44*, 1280−1291.

(6) Wilkinson, K. A., Vasa, S. M., Deigan, K. E., Mortimer, S. A., Giddings, M. C., and Weeks, K. M. (2009) Influence of nucleotide identity on ribose 2′-hydroxyl reactivity in RNA. *RNA 15*, 1314−1321.

(7) Gherghe, C. M., Shajani, Z., Wilkinson, K. A., Varani, G., and Weeks, K. M. (2008) Strong correlation between SHAPE chemistry and the generalized NMR order parameter (S^2) in RNA. *J. Am. Chem. Soc. 130*, 12244−12245.

(8) McGinnis, J. L., Dunkle, J. A., Cate, J. H., and Weeks, K. M. (2012) The mechanisms of RNA SHAPE chemistry. *J. Am. Chem. Soc. 134*, 6617−6624.

(9) Low, J. T., and Weeks, K. M. (2010) SHAPE-directed RNA secondary structure prediction. *Methods 52*, 150−158.

(10) Turner, D. H., and Mathews, D. H. (2010) NNDB: The nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res. 38*, D280−D282.

(11) Kladwang, W., VanLang, C. C., Cordero, P., and Das, R. (2011) Understanding the errors of SHAPE-directed RNA structure modeling. *Biochemistry 50*, 8049−8056.

(12) Wilkinson, K. A., Merino, E. J., and Weeks, K. M. (2006) Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE): Quantitative RNA structure analysis at single nucleotide resolution. *Nat. Protoc. 1*, 1610−1616.

(13) Kulshina, N., Baird, N. J., and Ferre-D'Amare, A. R. (2009) Recognition of the bacterial second messenger cyclic diguanylate by its cognate riboswitch. *Nat. Struct. Mol. Biol. 16*, 1212−1217.

(14) Kladwang, W., VanLang, C. C., Cordero, P., and Das, R. (2011) A two-dimensional mutate-and-map strategy for non-coding RNA structure. *Nat. Chem. 3*, 954−962.

(15) Byrne, R. T., Konevega, A. L., Rodnina, M. V., and Antson, A. A. (2010) The crystal structure of unmodified tRNAPhe from *Escherichia coli*. *Nucleic Acids Res. 38*, 4154−4162.

(16) Serganov, A., Yuan, Y. R., Pikovskaya, O., Polonskaia, A., Malinina, L., Phan, A. T., Hobartner, C., Micura, R., Breaker, R. R., and Patel, D. J. (2004) Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs. *Chem. Biol. 11*, 1729−1741.

(17) McGinnis, J. L., Duncan, C. D., and Weeks, K. M. (2009) High-throughput SHAPE and hydroxyl radical analysis of RNA structure and ribonucleoprotein assembly. *Methods Enzymol. 468*, 67−89.

(18) Vasa, S. M., Guex, N., Wilkinson, K. A., Weeks, K. M., and Giddings, M. C. (2008) ShapeFinder: A software system for high-

throughput quantitative analysis of nucleic acid reactivity information resolved by capillary electrophoresis. *RNA 14*, 1979−1990.
(19) Yoon, S., Kim, J., Hum, J., Kim, H., Park, S., Kladwang, W., and Das, R. (2011) HiTRACE: High-throughput robust analysis for capillary electrophoresis. *Bioinformatics 27*, 1798−1805.
(20) Reuter, J. S., and Mathews, D. H. (2010) RNAstructure: Software for RNA secondary structure prediction and analysis. *BMC Bioinf. 11*, 129.
(21) Rocca-Serra, P., Bellaousov, S., Birmingham, A., Chen, C., Cordero, P., Das, R., Davis-Neulander, L., Duncan, C. D., Halvorsen, M., Knight, R., Leontis, N. B., Mathews, D. H., Ritz, J., Stombaugh, J., Weeks, K. M., Zirbel, C. L., and Laederach, A. (2011) Sharing and archiving nucleic acid structure mapping data. *RNA 17*, 1204−1212.
(22) Nussinov, R., Pieczenik, G., Griggs, J. R., and Kleitman, D. J. (1978) Algorithms for loop matchings. *SIAM J. Appl. Math. 35*, 68−82.
(23) Merino, E. J., Wilkinson, K. A., Coughlan, J. L., and Weeks, K. M. (2005) RNA structure analysis at single nucleotide resolution by selective 2′-hydroxyl acylation and primer extension (SHAPE). *J. Am. Chem. Soc. 127*, 4223−4231.
(24) Lu, Z. J., Turner, D. H., and Mathews, D. H. (2006) A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation. *Nucleic Acids Res. 34*, 4912−4924.
(25) Gherghe, C. M., Mortimer, S. A., Krahn, J. M., Thompson, N. L., and Weeks, K. M. (2008) Slow conformational dynamics at C2′-endo nucleotides in RNA. *J. Am. Chem. Soc. 130*, 8884−8885.
(26) Mortimer, S. A., and Weeks, K. M. (2009) C2′-endo nucleotides as molecular timers suggested by the folding of an RNA domain. *Proc. Natl. Acad. Sci. U.S.A. 106*, 15622−15627.
(27) Escara, J. F., and Hutton, J. R. (1980) Thermal stability and renaturation of DNA in dimethyl sulfoxide solutions: Acceleration of the renaturation rate. *Biopolymers 19*, 1315−1327.
(28) Nichols, N. M., Tabor, S., and McReynolds, L. A. (2008) RNA ligases. *Curr. Protoc. Mol. Biol.* No. Chapter 3, 3.15.11−13.15.14.
(29) Kladwang, W., Cordero, P., and Das, R. (2011) A mutate-and-map strategy accurately infers the base pairs of a 35-nucleotide model RNA. *RNA 17*, 522−534.
(30) Smith, K. D., Lipchock, S. V., Ames, T. D., Wang, J., Breaker, R. R., and Strobel, S. A. (2009) Structural basis of ligand binding by a c-di-GMP riboswitch. *Nat. Struct. Mol. Biol. 16*, 1218−1223.
(31) Smith, K. D., Lipchock, S. V., Livingston, A. L., Shanahan, C. A., and Strobel, S. A. (2010) Structural and biochemical determinants of ligand binding by the c-di-GMP riboswitch. *Biochemistry 49*, 7351−7359.
(32) Krasilnikov, A. S., Yang, X., Pan, T., and Mondragon, A. (2003) Crystal structure of the specificity domain of ribonuclease P. *Nature 421*, 760−764.
(33) Cordero, P., Kladwang, W., VanLang, C. C., and Das, R. (2012) Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference. *Biochemistry 51*, 7037−7039.
(34) Weeks, K. M. (2010) Advances in RNA structure analysis by chemical probing. *Curr. Opin. Struct. Biol. 20*, 295−304.
(35) Edwards, A. L., Reyes, F. E., Heroux, A., and Batey, R. T. (2010) Structural basis for recognition of S-adenosylhomocysteine by riboswitches. *RNA 16*, 2144−2155.
(36) Kang, M., Peterson, R., and Feigon, J. (2010) Erratum: Structural insights into riboswitch control of the biosynthesis of queuosine, a modified nucleotide found in the anticodon of tRNA. *Mol. Cell 39*, 653−655.
(37) Zhang, Q., Kang, M., Peterson, R. D., and Feigon, J. (2011) Comparison of solution and crystal structures of preQ1 riboswitch reveals calcium-induced changes in conformation and dynamics. *J. Am. Chem. Soc. 133*, 5190−5193.
(38) Dibrov, S., McLean, J., and Hermann, T. (2011) Structure of an RNA dimer of a regulatory element from human thymidylate synthase mRNA. *Acta Crystallogr. D67*, 97−104.
(39) Warf, M. B., and Berglund, J. A. (2010) Role of RNA structure in regulating pre-mRNA splicing. *Trends Biochem. Sci. 35*, 169−178.
(40) Low, J. T., Knoepfel, S. A., Watts, J. M., ter Brake, O., Berkhout, B., and Weeks, K. M. (2012) SHAPE-directed discovery of potent shRNA inhibitors of HIV-1. *Mol. Ther. 20*, 820−828.