Published in final edited form as:
Biochemistry. 2013 January 29; 52(4): 588–595. doi:10.1021/bi300755u.

Principles for understanding the accuracy of SHAPE-directed RNA structure modeling

Christopher W. Leonard[1,5], **Christine E. Hajdin**[1,5], **Fethullah Karabiber**[1], **David H. Mathews**[4], **Oleg Favorov**[2], **Nikolay V. Dokholyan**[3], and **Kevin M. Weeks**[1,*]

[1]Department of Chemistry, University of North Carolina, Chapel Hill, NC 27599-3290

[2]Department of Biomedical Engineering, University of North Carolina, Chapel Hill, NC 27599-3290

[3]Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, NC 27599-3290

[4]Department of Biochemistry and Biophysics, University of Rochester Medical Center, Rochester, NY 14642

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 some apparent discrepancies between modeled and accepted structures. Recent studies by Das and colleagues [Kladwang *et al., Biochemistry 50*:8049 (2011) and *Nat. Chem. 3*:954 (2011)], focused on analyzing six small RNAs, yielded poorer RNA secondary structure predictions than expected based on 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 non-standard 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 Das and colleagues SHAPE data. When best practices were used, single-sequence SHAPE-directed secondary structure modeling recovered ~93% of individual base pairs and greater than 90% of helices in the four RNAs, essentially indistinguishable from 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.

[*]correspondence, weeks@unc.edu, 919-962-7486.
[5]Contributed equally.

Supporting Information
Two figures. This material is available free of charge via the Internet at http://pubs.acs.org. All SHAPE data from this work are available at the SNRNASM community database: http://snrnasm.bio.unc.edu.

Introduction

The functions of most RNA molecules are critically dependent on their structures, which are difficult to predict from first principles. A critical 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 reactivity of the 2'-hydroxyl, and 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 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.[12] In this communication, 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 secondary structure prediction.

Experimental

RNA synthesis

All RNAs were synthesized from double-stranded DNAs generated by PCR using single-stranded DNA templates (IDT) spanning the full length of the transcript, preceded by a 17-nt T7 promoter and a 14-nt 5' cassette sequence, followed by a 43-nt reverse transcription primer binding site.[13] 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, 0.1 mg/mL T7 polymerase at 37 °C for 3 hr. 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 absorbance at 260 nm measured using a Nanodrop 2000c spectrophotometer.

RNA constructs

Five RNAs were examined. We analyzed two variations of the *V. cholera* cyclic-di-GMP riboswitch RNA, a "short P1" RNA corresponding to the RNA in the 3iwn crystal structure[14] and a "long P1" that both extends the P1 helix by two base pairs 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.[15] Sequences of *E. coli* $tRNA^{Phe}$, the *V. vulnificus* adenine riboswitch, and *E. coli* 5S rRNA corresponded to those evaluated in crystallographic studies[16,17] and used by Kladwang *et al.*[11]

SHAPE experiments

Experiments with $tRNA^{Phe}$ were carried out 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.[9,13,18] For the Das and coworkers approach, 2 pmol RNA was denatured by heating 95 °C for 1 min, snap cooled on ice, then refolded in 50 mM Hepes (pH 8.0), 10 mM $MgCl_2$ for 30 min at 37 °C (labeled Buffer A in Figure 1). We then added 20 μL refolded RNA to the following volumes and concentrations of SHAPE reagents freshly dissolved in dry DMSO: 5 μL 135 mM NMIA (Invitrogen, M25), 5 μL or 2 μL 30 mM NMIA, or 5 μL or 2 μL 30 mM 1M7. Reactions were performed at 24 °C and at 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), 3 mM $MgCl_2$ (labeled Buffer B in Figure 1). Reactions were performed at 24 °C and at 37 °C by adding 10 μL RNA to 1 μL of 30 mM 1M7 and incubating for 3 min. Reactions with the adenine and c-di-GMP riboswitches contained 5 μM adenine or 10 μM cyclic-di-GMP (Biolog C057-01), respectively, and were initiated adding 10 μL refolded RNA to 1 μL 30 mM 1M7; reactions were incubated for 3 min at 37 °C. In all cases, RNAs were recovered by the addition of 15 μL 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 water. All RNAs were probed in 2–4 fully independent replicate experiments, in some cases performed months apart. Structure annotations shown in the individual figures correspond to single datasets but all independently analyzed datasets yielded identical lowest free energy structures. Data for the specificity domain of *B. subtilis* RNase P were reported previously.[3]

Primer extension

Reverse transcription reactions were prepared with the addition of 5 μL 0.5 μM FAM-labeled (SHAPE-modified RNA trace) or JOE-labeled (sequencing trace) reverse transcription primer, as appropriate, to 10 μL 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[9,18] using SuperScript III (Invitrogen). Reverse transcription proceeded with incubation at 52 °C for 5 min, then at 65 °C for 5 min. SHAPE-modified samples were combined with sequencing reactions, precipitated with ethanol, resuspended in 10 μL Hi-Di formamide (Applied Biosystems), heated with 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*[19] or a new custom software, *QuShape* (manuscript submitted; software is available immediately at: www.chem.unc.edu/rna/qushape). Structure predictions were identical for data processed by either approach. We attempted to process our experiments using HiTRACE[20] but were unable to run the publicly available version of the software; multiple attempts to solicit assistance from the authors were unsuccessful. RNA structure prediction was performed using *RNAstructure*[21] versions 5.2–5.4 under either Mac OS 10.6.× or Unix following box-plot normalization, exactly as described.[4] $\Delta G_{\mathrm{SHAPE}}$ 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 datasets generated in this work are available at the SNRNASM community structure probing database.[22] 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,[11] we also could not identify an alternative mathematical manipulation that could convert the Das and colleagues data into a form that would yield a fully correct secondary structure prediction for tRNA$^{\mathrm{Phe}}$. 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 reported for *E. coli* tRNA$^{\mathrm{Phe}}$ 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.[21,23] 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 shown in green, whereas missed (false negative) and incorrectly predicted (false positive) base pairs are shown in red and magenta, respectively. A fully correct structure would therefore have only green arcs.

Case I: tRNA$^{\mathrm{Phe}}$ and the adenine riboswitch

tRNA$^{\mathrm{Phe}}$ and the adenine riboswitch will be discussed as a single case as similar issues were identified in structure prediction for both RNAs. If the sequence of tRNA$^{\mathrm{Phe}}$ is submitted to the *RNAstructure* program (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 (Supporting Figure 1).

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.[4,9,13,18,19,24] 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, lowest 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 37 °C temperature.[25] (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, NMIA reactivity is sensitive to the specific ion environment[3] and preferentially reacts with nucleotides experiencing slow dynamics.[26,27] 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% (vol/vol) DMSO co-solvent. DMSO denatures some nucleic acid structures at this concentration.[28,29] (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 nodata prediction accuracy for tRNA^{Phe} was notably poorer at 24 °C than at 37 °C (Figure 1, first line, 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 at 24 °C. The SHAPE reactivity patterns 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, this SHAPE data resulted in a predicted RNA secondary structure that agreed with the accepted model. SHAPE data obtained under both standard and the Das and colleagues conditions for the adenine riboswitch also resulted in a structure that agreed with the accepted model (Supporting Figure 1).

Although we strongly recommend the use of the 1M7 reagent, use of fully soluble reagent concentrations, and maintaining organic co-solvent concentrations below 10%, at least in the cases of tRNA^{Phe} and the adenine riboswitch, SHAPE-directed RNA secondary structure prediction proved robust under these experimental conditions. In sum, 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 the program HiTRACE,[20] which appears to under-correct 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 Supporting Figure 2) 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.[15,30] 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[14,31] and featured a short, bulged P1 helix of four base pairs (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."[32] In evaluating the mutate-and-map approach, Das and colleagues used a *different* c-di-GMP RNA in which the P1 helix was extended by two base pairs to yield a much more stable six base pair helix. This change from a short to a long P1 helix has a dramatic effect on the structure of the RNA and increases affinity for the c-di-GMP ligand by a large factor.[32]

In the absence of SHAPE data, the structures of both short and long P1 forms of the c-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, line 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 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 predicted by the Das group using the mutate-and-map approach (Figure 2, line 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 sum, for the c-di-GMP RNA, single-sequence SHAPE-directed structure prediction and mutate-and-map modeling – the latter requiring two orders of magnitude greater number of 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-nt *E. coli* 16S rRNA[4] and for tRNA^{Phe} and the adenine and cdi- GMP riboswitch RNAs as described here (Figures 1, 2 and S1). 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 52 and 26%, respectively. Our SHAPE-directed prediction yielded substantial improvements, resulting in sensitivities of 78% 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 mis-assignment of a single helix (Figure 3, row 2, emphasized with asterisks). Incorrect prediction of one helix causes errors that propagate throughout 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,[33] features that may partially explain the challenge of predicting structures for these RNAs. In both cases at least one strand of the mis-assigned helix contains nucleotides with relatively high SHAPE reactivities, suggestive of semistable or 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 over-optimization. Das and colleagues have focused on and drawn strong conclusions from a small dataset of six RNAs.[11,34] We therefore examined the role that optimization over a small dataset might have on prediction accuracies.

Our first-generation parameters were optimized using ~2,500 nts in the *E. coli* 23S rRNA and generally work well for many different classes of RNA, including the 1,542-nt *E. coli* 16S rRNA[4] and diverse small RNAs (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 4 of the 6 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 (Table 1, see ‘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 sum, the conclusions of recent papers[11,15,34] comparing SHAPE-directed modeling with other approaches would have been different if higher quality data had been obtained and if parameters for the SHAPE approach had been obtained using the same small dataset approach used to optimize the mutate-and-map[11,15] or DMS[34] methods. 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.[34] Inspection of the new data shows that both background subtraction and signal decay corrections have been improved. Das and colleagues now report recovery of greater than 90% of all accepted base pairs for the same set of RNAs,[34] 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.

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 and Figures 1–3 and S1). However, a few RNAs remain challenging to single-sequence prediction, including the *B. subtilis* RNase P specificity domain and *E. coli* 5S rRNA, 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 are required to fully analyze, understand, and minimize potential errors in secondary structure modeling.[5,35] In particular, choice of experimental conditions, accurate data processing, and identical scoring rules are crucial (Figures 1–4 and S1 and Refs. 11, 15). 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 well. 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 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 non-expert

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 NMR 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.[32,36–39] In some cases, an RNA may exist in an equilibrium between multiple structures and it is an oversimplification 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 non-canonical 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 many biological functions[40,41] – 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 only be met 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 non-coding 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by grants from the NIH (AI068462 to K.M.W., GM080742 to N.V.D., and GM076485 to D.H.M.) and the NSF (MCB-1121024 to K.M.W.). 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 DH, Sabina J, Zuker M, Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. 1999; 288:911–940. [PubMed: 10329189]

2. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc. Natl. Acad. Sci. USA. 2004; 101:7287–7292. [PubMed: 15123812]

3. Mortimer SA, Weeks KM. A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. J. Am. Chem. Soc. 2007; 129:4144–4145. [PubMed: 17367143]

4. Deigan KE, Li TW, Mathews DH, Weeks KM. Accurate SHAPE-directed RNA structure determination. Proc. Natl. Acad. Sci. USA. 2009; 106:97–102. [PubMed: 19109441]

5. Weeks KM, Mauger DM. Exploring RNA structural codes with SHAPE chemistry. Acc. Chem. Res. 2011; 44:1280–1291. [PubMed: 21615079]

6. Wilkinson KA, Vasa SM, Deigan KE, Mortimer SA, Giddings MC, Weeks KM. Influence of nucleotide identity on ribose 2'-hydroxyl reactivity in RNA. RNA. 2009; 15:1314–1321. [PubMed: 19458034]

7. Gherghe CM, Shajani Z, Wilkinson KA, Varani G, Weeks KM. Strong correlation between SHAPE chemistry and the generalized NMR order parameter (S2) in RNA. J. Am. Chem. Soc. 2008; 130:12244–12245. [PubMed: 18710236]

8. McGinnis JL, Dunkle JA, Cate JH, Weeks KM. The mechanisms of RNA SHAPE chemistry. J. Am. Chem. Soc. 2012; 134:6617–6624. [PubMed: 22475022]

9. Low JT, Weeks KM. SHAPE-directed RNA secondary structure prediction. Methods. 2010; 52:150–158. [PubMed: 20554050]

10. Turner DH, Mathews DH. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucl. Acids Res. 2010; 38:D280–D282. [PubMed: 19880381]

11. Kladwang W, VanLang CC, Cordero P, Das R. Understanding the errors of SHAPE-directed RNA structure modeling. Biochemistry. 2011; 50:8049–8056. [PubMed: 21842868]

12. Ramachandran S, Ding F, Weeks KM, Dokholyan NV. Statistical analysis of SHAPE-directed RNA secondary structure modeling. Biochemistry. 2012; 51 under revision.

13. Wilkinson KA, Merino EJ, Weeks KM. Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. Nature Protoc. 2006; 1:1610–1616. [PubMed: 17406453]

14. Kulshina N, Baird NJ, Ferre-D'Amare AR. Recognition of the bacterial second messenger cyclic diguanylate by its cognate riboswitch. Nature Struct. Mol. Biol. 2009; 16:1212–1217. [PubMed: 19898478]

15. Kladwang W, VanLang CC, Cordero P, Das R. A two-dimensional mutate-and-map strategy for non-coding RNA structure. Nature Chem. 2011; 3:954–962. [PubMed: 22109276]

16. Byrne RT, Konevega AL, Rodnina MV, Antson AA. The crystal structure of unmodified tRNAPhe from Escherichia coli. Nucl. Acids Res. 2010; 38:4154–4162. [PubMed: 20203084]

17. Serganov A, Yuan YR, Pikovskaya O, Polonskaia A, Malinina L, Phan AT, Hobartner C, Micura R, Breaker RR, Patel DJ. Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs. Chem. Biol. 2004; 11:1729–1741. [PubMed: 15610857]

18. McGinnis JL, Duncan CD, Weeks KM. High-throughput SHAPE and hydroxyl radical analysis of RNA structure and ribonucleoprotein assembly. Methods Enzymol. 2009; 468:67–89. [PubMed: 20946765]

19. Vasa SM, Guex N, Wilkinson KA, Weeks KM, Giddings MC. ShapeFinder: a software system for high-throughput quantitative analysis of nucleic acid reactivity information resolved by capillary electrophoresis. RNA. 2008; 14:1979–1990. [PubMed: 18772246]

20. Yoon S, Kim J, Hum J, Kim H, Park S, Kladwang W, Das R. HiTRACE: high-throughput robust analysis for capillary electrophoresis. Bioinformatics. 2011; 27:1798–1805. [PubMed: 21561922]

21. Reuter JS, Mathews DH. RNA structure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics. 2010; 11:129. [PubMed: 20230624]

22. Rocca-Serra P, Bellaousov S, Birmingham A, Chen C, Cordero P, Das R, Davis-Neulander L, Duncan CD, Halvorsen M, Knight R, Leontis NB, Mathews DH, Ritz J, Stombaugh J, Weeks KM, Zirbel CL, Laederach A. Sharing and archiving nucleic acid structure mapping data. RNA. 2011; 17:1204–1212. [PubMed: 21610212]

23. Nussinov R, Pieczenik G, Griggs JR, Kleitman DJ. Algorithms for loop matchings. SIAM J. Appl. Math. 1978; 35:68–82.

24. Merino EJ, Wilkinson KA, Coughlan JL, Weeks KM. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). J. Am. Chem. Soc. 2005; 127:4223–4231. [PubMed: 15783204]

25. Lu ZJ, Turner DH, Mathews DH. A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation. Nucl. Acids Res. 2006; 34:4912–4924. [PubMed: 16982646]

26. Gherghe CM, Mortimer SA, Krahn JM, Thompson NL, Weeks KM. Slow conformational dynamics at C2'-endo nucleotides in RNA. J. Am. Chem. Soc. 2008; 130:8884–8885. [PubMed: 18558680]

27. Mortimer SA, Weeks KM. C2'-endo nucleotides as molecular timers suggested by the folding of an RNA domain. Proc. Natl. Acad. Sci. USA. 2009; 106:15622–15627. [PubMed: 19717440]

28. Escara JF, Hutton JR. Thermal stability and renaturation of DNA in dimethyl sulfoxide solutions: acceleration of the renaturation rate. Biopolymers. 1980; 19:1315–1327. [PubMed: 7397315]

29. Nichols NM, Tabor S, McReynolds LA. RNA ligases. Curr. Protoc. Mol. Biol. Chapter. 2008; 3:3.15.11–13.15.14.

30. Kladwang W, Cordero P, Das R. A mutate-and-map strategy accurately infers the base pairs of a 35-nucleotide model RNA. RNA. 2011; 17:522–534. [PubMed: 21239468]

31. Smith KD, Lipchock SV, Ames TD, Wang J, Breaker RR, Strobel SA. Structural basis of ligand binding by a c-di-GMP riboswitch. Nature Struct. Mol. Biol. 2009; 16:1218–1223. [PubMed: 19898477]

32. Smith KD, Lipchock SV, Livingston AL, Shanahan CA, Strobel SA. Structural and biochemical determinants of ligand binding by the c-di-GMP riboswitch. Biochemistry. 2010; 49:7351–7359. [PubMed: 20690679]

33. Krasilnikov AS, Yang X, Pan T, Mondragon A. Crystal structure of the specificity domain of ribonuclease P. Nature. 2003; 421:760–764. [PubMed: 12610630]

34. Cordero P, Kladwang W, VanLang CC, Das R. Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference. Biochemistry. 2012; 51:7037–7039. [PubMed: 22913637]

35. Weeks KM. Advances in RNA structure analysis by chemical probing. Curr. Opin. Struct. Biol. 2010; 20:295–304. [PubMed: 20447823]

36. Edwards AL, Reyes FE, Heroux A, Batey RT. Structural basis for recognition of S-adenosylhomocysteine by riboswitches. RNA. 2010; 16:2144–2155. [PubMed: 20864509]

37. Kang M, Peterson R, Feigon J. Erratum: Structural insights into riboswitch control of the biosynthesis of queuosine, a modified nucleotide found in the anticodon of tRNA. Mol. Cell. 2010; 39:653–655.

38. Zhang Q, Kang M, Peterson RD, Feigon J. Comparison of solution and crystal structures of preQ1 riboswitch reveals calcium-induced changes in conformation and dynamics. J. Am. Chem. Soc. 2011; 133:5190–5193. [PubMed: 21410253]

39. Dibrov S, McLean J, Hermann T. Structure of an RNA dimer of a regulatory element from human thymidylate synthase mRNA, Acta Crystall. D. Biol. Crystall. 2011; 67:97–104.

40. Warf MB, Berglund JA. Role of RNA structure in regulating pre-mRNA splicing. Trends Biochem. Sci. 2010; 35:169–178. [PubMed: 19959365]

41. Low JT, Knoepfel SA, Watts JM, ter Brake O, Berkhout B, Weeks KM. SHAPE-directed discovery of potent shRNA inhibitors of HIV-1. Mol. Ther. 2012; 20:820–828. [PubMed: 22314289]

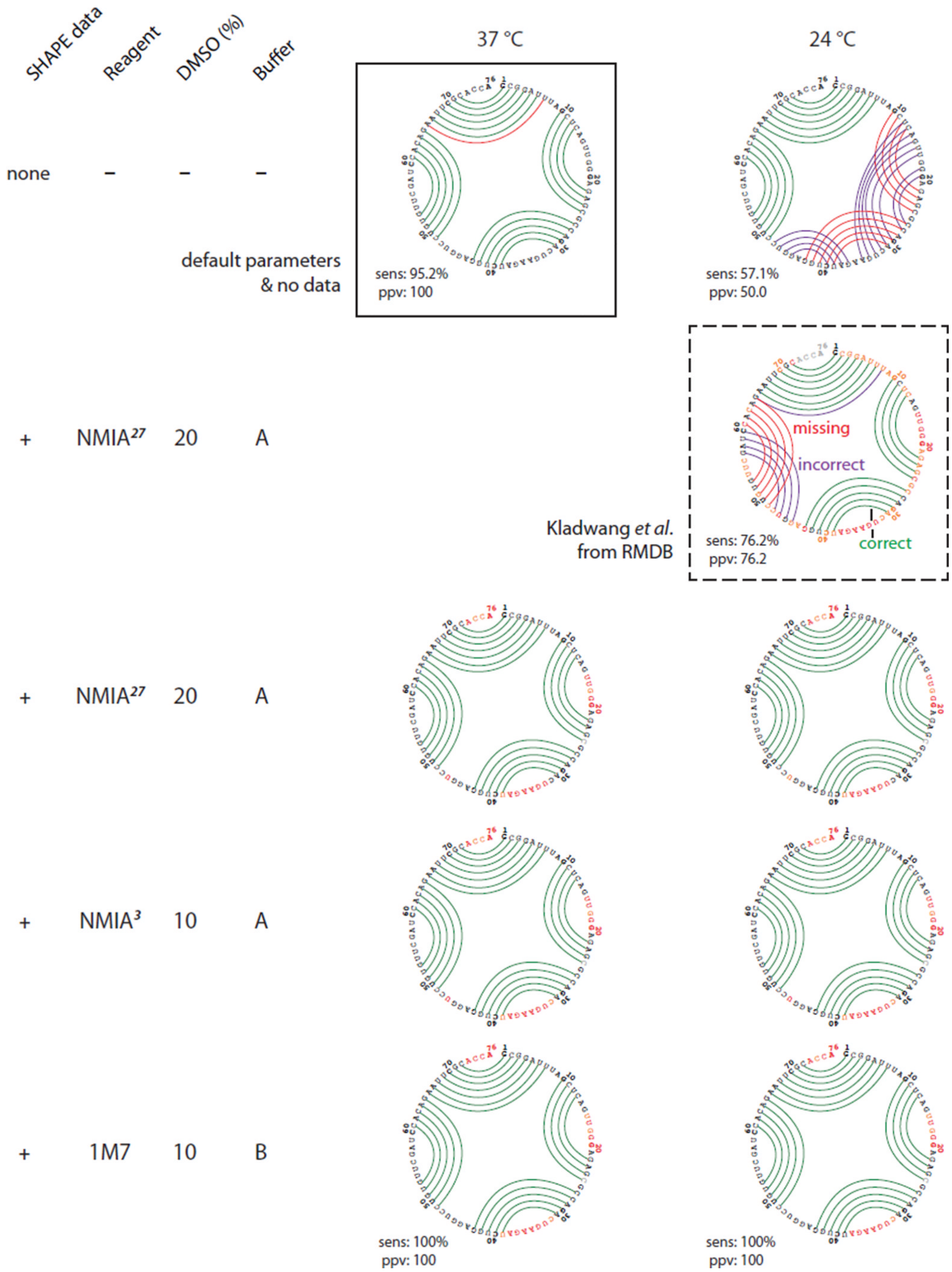

Figure 1.

Evaluation of thermodynamically based and experimentally directed secondary structure modeling for tRNA^{Phe}. Secondary structure predictions were performed using *RNAstructure*[21] 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.*[11,15] or on those developed by our laboratory.[9,13,18] Experimental variations included reagent (NMIA or 1M7), concentration of NMIA (either ~27 or 3 mM, indicated with superscripts of *27* and *3*, respectively; the 27 mM condition results in formation of visible precipitate), percentage of DMSO co-solvent (10 or 20%), and solution buffer conditions (buffers A and B are reported in Refs. 11 and 13, respectively).

Kladwang *et al*. data, obtained from the RMDB, were processed with HiTRACE.[20] If SHAPE data were used in the secondary structure prediction, nucleotides are colored by reactivity: low, medium, and high reactivities are indicated in black, yellow, and red, respectively. Sensitivity (sens) and positive predictive value (ppv) are indicated for representative structures.

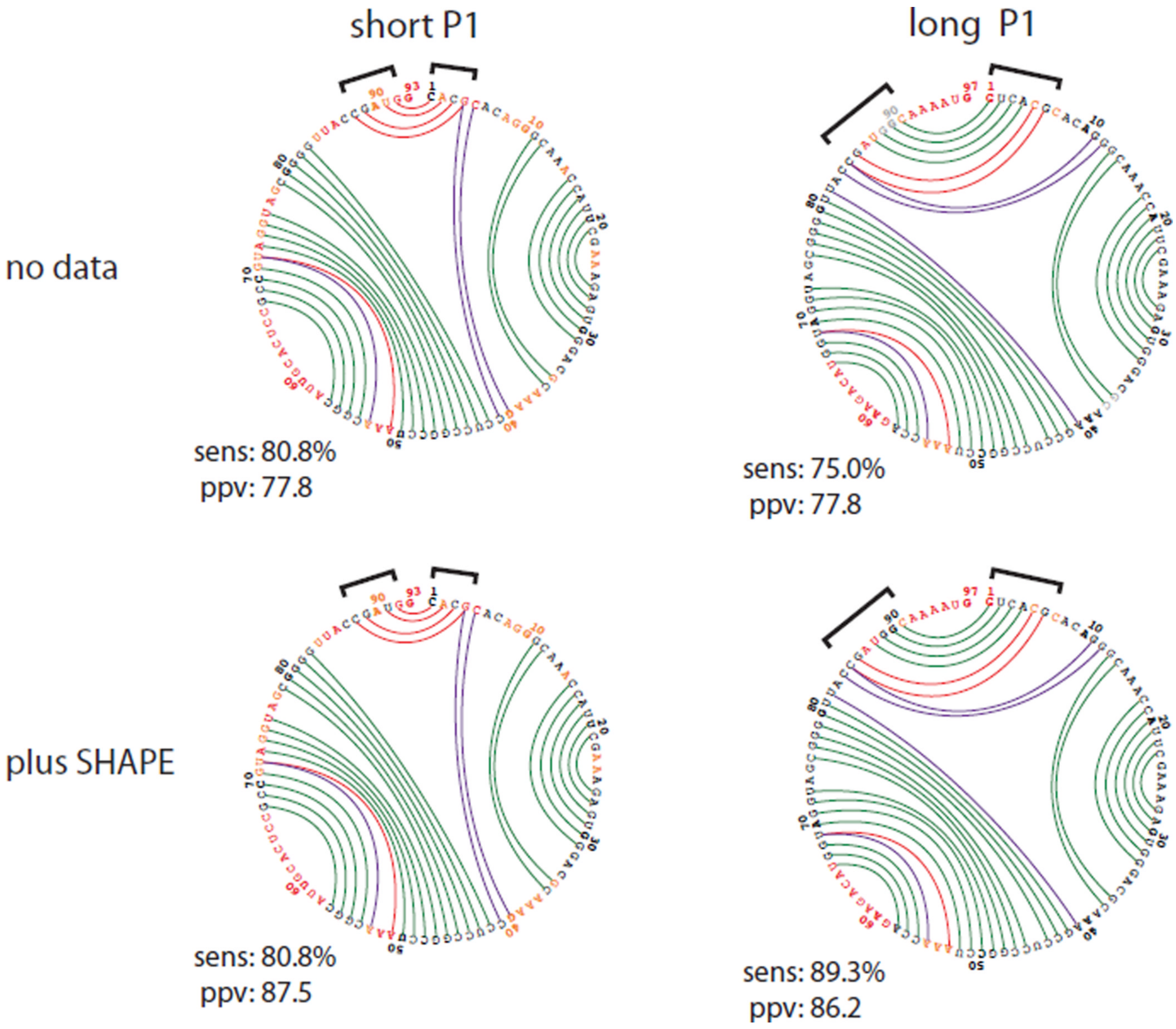

Figure 2.

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

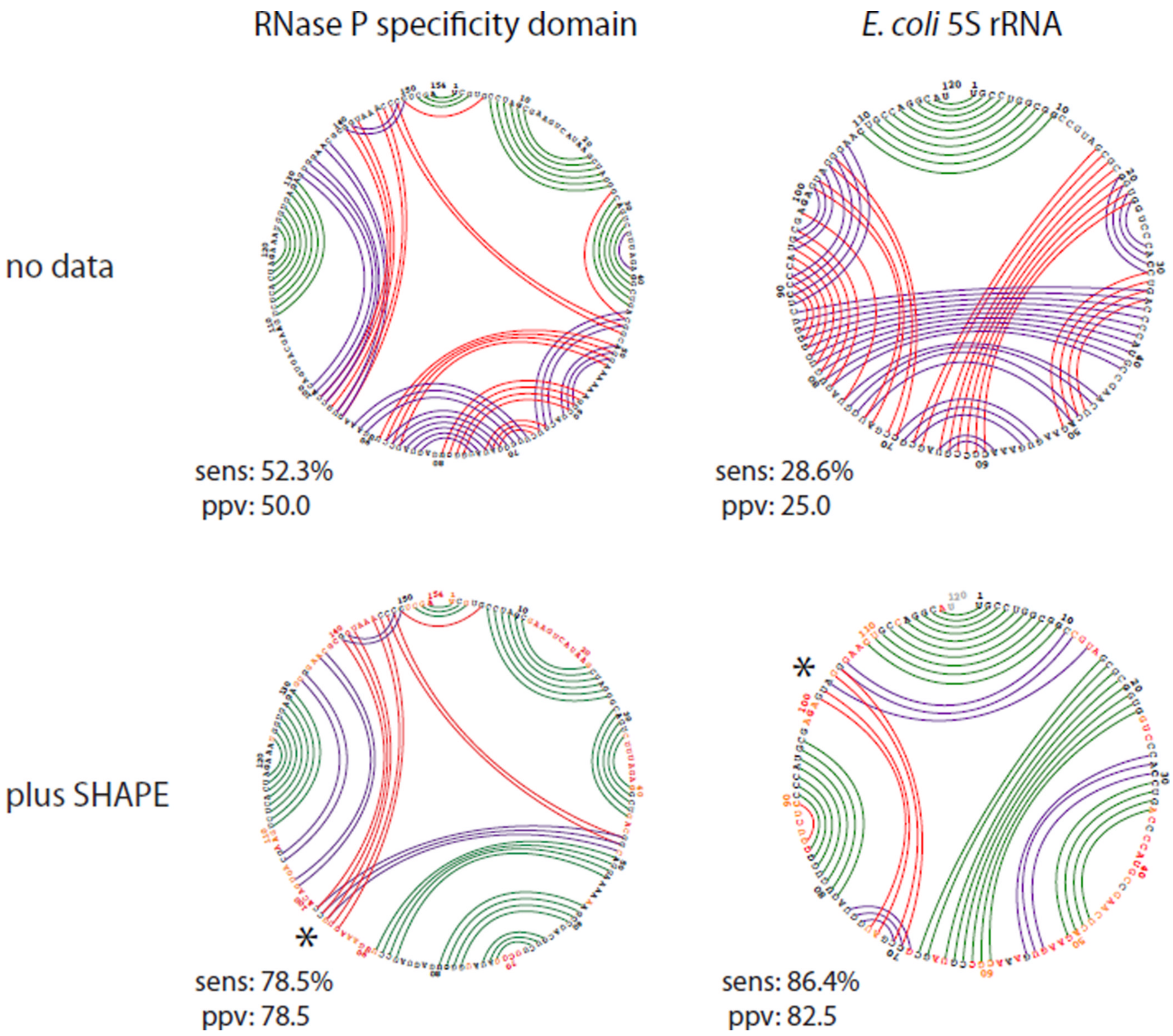

Figure 3.

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

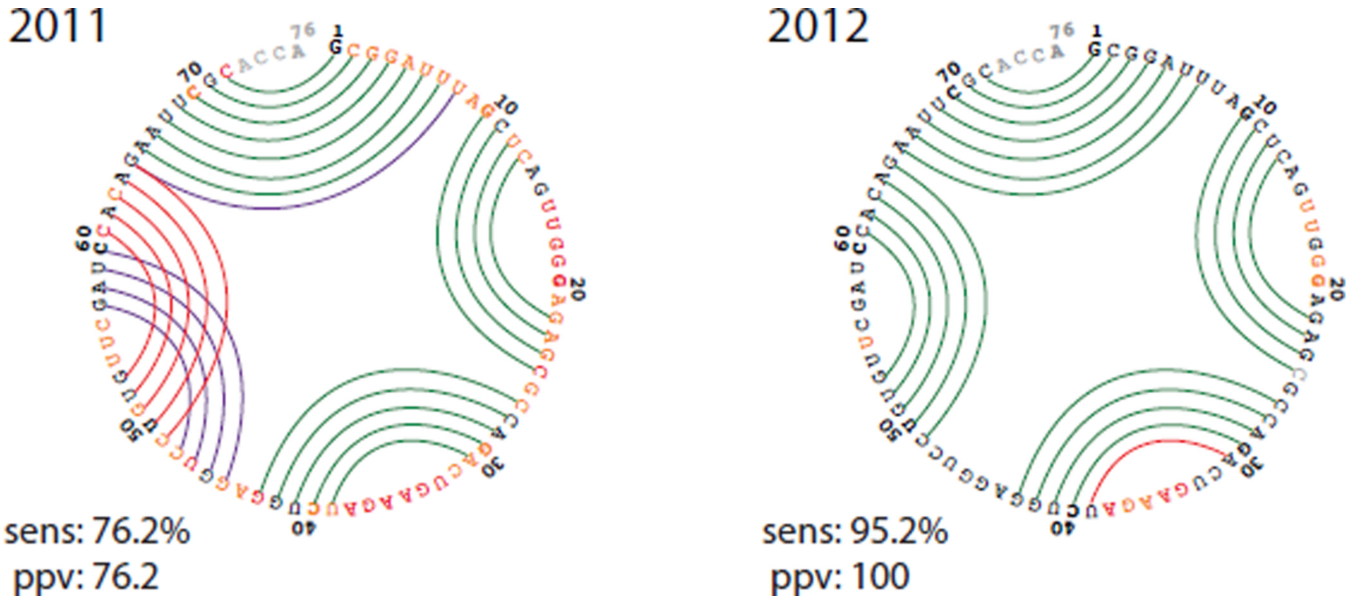
2011
1
10
20
30
40
50
60
70
76
GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA
sens: 76.2%
ppv: 76.2
2012
1
10
20
30
40
50
60
70
76
GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA
sens: 95.2%
ppv: 100

Figure 4.

Comparison of SHAPE-directed modeling of tRNA^{Phe}, as published by Das and colleagues in September and December 2011[11,15] (left) versus submitted July 2012[34] (right). Models are annotated using the scheme in Figure 1.

Table 1

Effect of optimizing chemical probing parameters over small datasets

SHAPE data:		−			+			+		
Parameters[a]:		**No data**			**Global**			**Small training set**		
RNA	**Length**	**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	26	25	26	86	82	84	**92**	**100**	**96**
P546 domain, bI3 intron	155	43	44	44	**96**	**98**	**97**	**95**	**96**	**96**
Average		65	64	64	94	93	94	97	97	97

[a] Global parameters were $m = 2.6$ and $b -0.8$, which give high quality models for both small RNAs and for kibobase length ribosomal RNAs.[4] The small training set parameters were $m = 1.3$ and $b = -0.3$.

Accuracies greater than 90% are highlighted in bold. Prediction accuracies are given as sensitivity (sens), positive predictive value (ppv) and their geometric mean (geo).