BIOCHEMISTRY
including biophysical chemistry & molecular biology

Article

pubs.acs.org/biochemistry

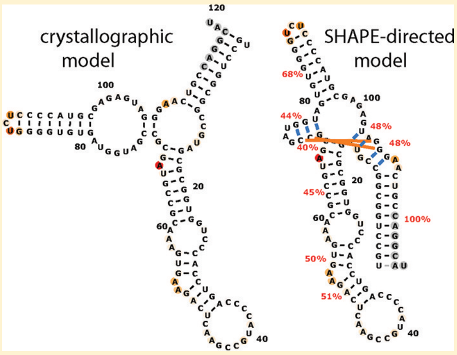
Understanding the Errors of SHAPE-Directed RNA Structure Modeling

Wipapat Kladwang,[†] Christopher C. VanLang,[‡] Pablo Cordero,[§] and Rhiju Das*,[†,§,∥]

Departments of [†]Biochemistry, [‡]Chemical Engineering, [§]Biomedical Informatics, and [∥]Physics, Stanford University, Stanford, California 94305, United States

Supporting Information

ABSTRACT: Single-nucleotide-resolution chemical mapping for structured RNA is being rapidly advanced by new chemistries, faster readouts, and coupling to computational algorithms. Recent tests have shown that selective 2′-hydroxyl acylation by primer extension (SHAPE) can give near-zero error rates (0–2%) in modeling the helices of RNA secondary structure. Here, we benchmark the method using six molecules for which crystallographic data are available: tRNA(phe) and 5S rRNA from *Escherichia coli*, the P4–P6 domain of the *Tetrahymena* group I ribozyme, and ligand-bound domains from riboswitches for adenine, cyclic di-GMP, and glycine. SHAPE-directed modeling of these highly structured RNAs gave an overall false negative rate (FNR) of 17% and a false discovery rate (FDR) of 21%, with at least one helix prediction error in five of the six cases. Extensive variations of data processing, normalization, and modeling parameters did not significantly mitigate modeling errors. Only one varation, filtering out data collected with deoxyinosine triphosphate during primer extension, gave a modest improvement (FNR = 12%, and FDR = 14%). The residual structure modeling errors are explained by the insufficient information content of these RNAs' SHAPE data, as evaluated by a nonparametric bootstrapping analysis. Beyond these benchmark cases, bootstrapping suggests a low level of confidence (<50%) in the majority of helices in a previously proposed SHAPE-directed model for the HIV-1 RNA genome. Thus, SHAPE-directed RNA modeling is not always unambiguous, and helix-by-helix confidence estimates, as described herein, may be critical for interpreting results from this powerful methodology.



The continuing discoveries of new classes of RNA enzymes, switches, and ribonucleoprotein assemblies provide complex challenges for structural and mechanistic dissection (see, e.g., refs 1–4). While crystallographic, spectroscopic, and phylogenetic analyses have led to a deeper understanding of several key model systems, the throughput or applicability of these methods is limited, especially for noncoding RNAs that switch between multiple states in their functional cycles.[5–8] In recent years, several laboratories have revisited a widely applicable chemical approach for attaining nucleotide-resolution RNA structural information, variously called "footprinting" or "chemical structure mapping". Recent advances have included novel chemical modification strategies, faster data analysis software, accelerated readouts via capillary electrophoresis, and multiplexed purification by magnetic beads.[9–14]

Despite these advances, chemical mapping data are not expected to generally give structure models accurate at nucleotide resolution. To a first approximation, the protection of an RNA nucleotide from chemical modification indicates that it forms some interaction with a partner elsewhere in the system. However, these data by themselves do not provide enough information to define the interaction partner. Instead, the mapping data can be used to test, refine, or guide structure hypotheses derived from manual inspection or automated algorithms.[15–17] The accuracy of this approach is necessarily limited by uncertainties in the modeling, including incomplete treatment of noncanonical base pairs, base–backbone interactions, and pseudoknotted folds,[17] and imperfect correlations of chemical modification rates to structural features. Indeed, there are notable historical examples of chemical data giving misleading structural suggestions, including blind modeling work on tRNA[18,19] and 5S rRNA.[20,21]

It was therefore exciting when recent studies of 2′-OH acylation (the SHAPE method) coupled to the *RNAstructure* algorithm reported secondary structure inference with unprecedented sensitivity (98–100% helix recovery).[17] The work acknowledged several uncertainties. Measurements were taken on ribosomal RNA without protein partners, which may not form the same structures as crystallized protein-bound complexes. For other test cases, the assumed experimental structures were derived from phylogenetic analysis (P546 domain from the *bI3* group I intron), NMR data (HCV IRES), or crystals of constructs with modifications not present in the SHAPE-probed constructs (tRNA^{Asp}). A "gold-standard" benchmark of SHAPE-directed secondary structure inference on RNAs with corresponding crystallographic models remains unavailable. We present herein

Received: April 6, 2011
Revised: July 29, 2011
Published: August 15, 2011

dx.doi.org/10.1021/bi200524n | *Biochemistry* 2011, 50, 8049–8056

SHAPE data, secondary structure inference, and analysis of systematic and statistical errors for six such RNAs containing a total of 661 nucleotides and 42 helices. Our results provide a rigorous appraisal of the strengths and limitations of this promising chemical–computational technology.

■ EXPERIMENTAL PROCEDURES

Preparation of Model RNAs. The DNA templates for each RNA (Table S1 of the Supporting Information) consisted of the 20-nucleotide T7 RNA polymerase promoter sequence (TTCTAATACGACTCACTATA) followed by the desired sequence. Double-stranded templates were prepared by polymerase chain reaction assembly of DNA oligomers up to 60 nucleotides in length (IDT, Integrated DNA Technologies, Coralville, IA) with Phusion DNA polymerase (Finnzymes) and purified with AMPure magnetic beads (Agencourt, Beckman Coulter) following the manufacturer's instructions. Sample concentrations were estimated on the basis of UV absorbance at 260 nm measured on Nanodrop 100 or 8000 spectrophotometers. Verification of template length was accomplished by electrophoresis of all samples and 10 and 20 bp ladder length standards (Fermentas) in 4% agarose gels (containing 0.5 mg/mL ethidium bromide) and 1× TBE (100 mM Tris, 83 mM boric acid, and 1 mM disodium EDTA).

In vitro RNA transcription reactions were conducted in 40 μL volumes with 10 pmol of DNA template, 20 units of T7 RNA polymerase (New England Biolabs), 40 mM Tris-HCl (pH 8.1), 25 mM $MgCl_2$, 2 mM spermidine, ATP, CTP, GTP, and UTP (1 mM each), 4% polyethylene glycol 1200, and 0.01% Triton X-100. Reaction mixtures were incubated at 37 °C for 4 h and monitored by electrophoresis of all samples along with 100–1000-nucleotide RNA length standards (RiboRuler, Fermentas) in 4% denaturing agarose gels [1.1% formaldehyde; run in 1× TAE (40 mM Tris, 20 mM acetic acid, and 1 mM disodium EDTA)], stained with SYBR Green II RNA gel stain (Invitrogen) following the manufacturer's instructions. RNA samples were purified with MagMax magnetic beads (Ambion), following the manufacturer's instructions, and concentrations were measured by absorbance at 260 nm on Nanodrop 100 or 8000 spectrophotometers.

Chemical Probing Measurements. Chemical modification reaction mixtures consisted of 1.2 pmol of RNA in a volume of 20 μL with 50 mM Na-HEPES (pH 8.0) and 10 mM $MgCl_2$ and/or ligand at the desired concentration (see Table S1 of the Supporting Information), and 5 μL of SHAPE modification reagent. The modification reagent was 24 mg/mL *N*-methylisatoic anhydride (NMIA) freshly dissolved in anhydrous dimethyl sulfoxide (DMSO). The reaction mixtures were incubated at 24 °C for 15–60 min, with shorter modification times for the longer RNAs to maintain overall modification rates of <30%. In control reactions (for background measurements), 5 μL of deionized water or DMSO was added instead of modification reagent, and the mixture was incubated for the same amount of time. For experiments testing DMSO effects, higher concentrations of NMIA in DMSO were prepared and 2 μL of the modification reagent was added to the 20 μL reaction mixture. Reactions were quenched with a premixed solution of 5 μL of 0.5 M Na-MES (pH 6.0), 3 μL of 5 M NaCl, 1.5 μL of oligo-dT beads [poly(A) purist (Ambion)], and 0.25 μL of 0.5 μM 5′-rhodamine-green-labeled primer (AAAAAAAAAAAAAAAAAAAAGTTGTTGTTGTTGTTTCTTT) complementary to the 3′ end of the RNAs (also used in our previous studies[13,14]), and 0.05 μL of a 0.5 μM Alexa-555-labeled oligonucleotide (used to verify normalization). The reaction mixtures were purified by magnetic separation, rinsed with 40 μL of 70% ethanol twice, and allowed to air-dry for 10 min while remaining on a 96-post magnetic stand. The magnetic bead mixtures were resuspended in 2.5 μL of deionized water.

The resulting mixtures of modified RNAs and primers bound to magnetic beads were reverse transcribed by the addition of a premixed solution containing 0.2 μL of SuperScript III (Invitrogen), 1.0 μL of 5× SuperScript First Strand buffer (Invitrogen), 0.4 μL of dNTPs at 10 mM each (dATP, dCTP, and dTTP, with either dGTP or dITP[22]), 0.25 μL of 0.1 M DTT, and 0.65 μL of water. The reaction mixtures (total volume of 5 μL) were incubated at 42 °C for 30 min. RNA was degraded by the addition of 5 μL of 0.4 M NaOH and incubation at 90 °C for 3 min. The solutions were neutralized by the addition of 5 μL of an acid quench (2 volumes of 5 M NaCl, 2 volumes of 2 M HCl, and 3 volumes of 3 M sodium acetate). Fluorescent DNA products were purified by magnetic bead separation, rinsed twice with 40 μL of 70% ethanol, and air-dried for 5 min. The reverse transcription products, along with magnetic beads, were resuspended in 10 μL of a solution containing 0.125 mM Na-EDTA (pH 8.0) and a Texas Red-labeled reference ladder (whose fluorescence is spectrally separated from the rhodamine green-labeled products). The products were separated by capillary electrophoresis on an ABI 3100 or ABI 3700 DNA sequencer. Reference ladders were created using an analogous protocol without chemical modification and the addition of, for example, 2′,3′-dideoxy-TTP in an amount equimolar with respect to the amount of dTTP in the reverse transcriptase reaction.

The HiTRACE software[23,24] was used to analyze the electropherograms. Briefly, traces were aligned by automatically shifting and scaling the time coordinate, based on cross correlation of the Texas Red reference ladder coloaded with all samples. Sequence assignments for bands, verified by comparison to sequencing ladders, permitted the automated peak fitting of the traces to Gaussians.

Likelihood-Based Processing of SHAPE Data. Quantified SHAPE data were corrected for attenuation of longer reverse transcriptase products due to chemical modification, normalized, and background-subtracted. Rather than using an approximate exponential correction and background scaling,[25] we used a likelihood framework to determine the final, corrected SHAPE reactivities (see also ref 26). Furthermore, a likelihood-derived analysis was implemented to average replicate SHAPE data sets across several experiments. Both of these procedures are described in detail in the methods of the Supporting Information. The algorithms are available in the functions *overmod_and_background_correct_logL.m* and *get_average_standard_state.m* within the freely available HiTRACE software package.[24] Final averaged data and errors have been made made publicly available in the Stanford RNA Mapping Database (http://rmdb.stanford.edu). The accession IDs are TRNAPH_SHP_0001, TRP4P6_SHP_0001, 5SRRNA_SHP_0001, ADDRSW_SHP_0001, CIDGMP_SHP_0001, and GLYCFN_SHP_0001.

Computational Modeling. The *Fold* executable of the RNAstructure package (version 5.3) was used to infer SHAPE-directed secondary structures. The entire RNA sequences (Table S1 of the Supporting Information), including added flanking sequences, were used for all calculations. The flag "-T 297.15" set the temperature to match our experimental conditions (24 °C). The flags "-sh", "-sm", and "-si" were used to input the SHAPE data file, slope m, and intercept b. The latter parameters define the pseudoenergy formula $\Delta G_i = m \log(S_i + 1) + b$, where S_i is the SHAPE reactivity. In the RNAstructure implementation, these pseudoenergies are applied to each nucleotide that forms an edge base pair and doubly applied to each nucleotide that forms an internal base pair. Boltzmann probability calculations used the *partition* executable with the same flags.

Nonparametric bootstrapping analysis was conducted as follows. Given normalized SHAPE data S_i for nucleotides i = 1, 2, ..., N, a bootstrap replicate was generated by choosing N random indices i' from 1 to N, with replacement[27,28,50] (i.e., some nucleotide positions are not represented, and some are present in multiple copies; for the latter, SHAPE pseudoenergies were scaled proportionally). The resulting data sets $S_{i'}$ contained the same number of data points and carried any systematic errors present in the original data set. Secondary structure models directed by these data were analyzed in MATLAB to assess the frequency of each base pair arising in the replicates; the maximum bootstrap value across the base pairs of each helix was taken as the bootstrap value for the helix. The bootstrapping analysis is being made available on an automated server at http://rmdb.stanford.edu/structureserver.

Additional calculations were conducted with the fold() routine of the ViennaRNA package (version 1.8.4, equivalent to the "RNAfold" command lines)[29] extended to accept SHAPE data and calculate pseudoenergies with the same formula used in RNAstructure; calculations were facilitated through Python bindings available through the software's convenient SWIG (Simplified Wrapper and Interface Generator) interface. Secondary structure figures were prepared with VARNA.[30]

Assessment of Accuracy. A crystallographic helix was considered correctly recovered if more than 50% of its base pairs were observed in a helix by the computational model. (In practice, 34 of 35 such helices retained all crystallographic base pairs.) Note that, unlike prior work, helix slips of ±1 were not considered correct [i.e., the pairing (i,j) was not allowed to match the pairing $(i,j-1)$ or $(i,j+1)$].

RESULTS

Accuracy of Modeling without Experimental Data. The benchmark herein (Table S1 of Supporting Information) collects a diverse set of noncoding RNA domains, containing two classic RNA folding model systems, unmodified tRNA$^{\mathrm{phe}}$ from *Escherichia coli*[31] and the P4–P6 domain of the *Tetrahymena* group I ribozyme;[32] a functional RNA that has been a frequent test case for modeling algorithms, *E. coli* 5S rRNA;[15,16,20,21] and three ligand-bound domains from bacterial riboswitches for adenine, cyclic di-GMP, and glycine.[33−39] For the last RNA (glycine riboswitch from *Fusobacterium nucleatum*), crystallographic data were not available at the time of modeling but released at the time of submission of the manuscript; it served as a blind test within our benchmark.

As a control, we first applied the *RNAstructure*[15,16] algorithm *Fold* without any experimental data to the benchmark set (Figure S1 of the Supporting Information). Here and below, we discuss modeling errors in terms of the false negative rate (FNR; fraction of crystallographic helices that were missed) and false discovery rate (FDR; fraction of predicted helices that were incorrect). The values are summarized, along with the related statistics of sensitivity and positive predicted value, in Table 1. To highlight features of the RNAs' global folds, we present results in terms of helices rather than individual base pairs. For the sake of completeness, FNR, FDR, sensitivity, and positive predictive values at the base pair level are also compiled in Table S2 of the Supporting Information.

Table 1. Accuracy of Secondary Structure Recovery by *RNAstructure* with and without SHAPE Data

		no. of helices[a]				
			RNAstructure		with SHAPE	
RNA	no. of nucleotides	Cryst	TP	FP	TP	FP
tRNA$^{\mathrm{phe}}$	76	4	2	3	3	1
P4–P6 RNA	158	11	10	1	9	1
5S rRNA	118	7	1	9	6	3
adenine riboswitch	71	3	2	3	3	1
cyclic di-GMP ribosw.	80	8	6	2	6	2
glycine riboswitch	158	9	5	3	8	1
total	661	42	26	21	35	9
false negative rate[b]			38.1%		16.7%	
false discovery rate[c]			44.7%		20.5%	
sensitivity[d]			61.9%		83.3%	
positive predictive value[e]			55.3%		79.5%	

[a] Abbreviations: Cryst, number of helices in the crystallographic model; TP, true positives; FP, false positives. [b] False negative rate = 1 − TP/Cryst. [c] False discovery rate = FP/(TP + FP). [d] Sensitivity = (1 − false negative rate) = TP/Cryst. [e] Positive predictive value = (1 − false discovery rate) = TP/(TP + FP).

Without any data, the *RNAstructure* algorithm missed 16 of 42 helices, giving an FNR of 16/42 (38%). The models mispredicted an additional 21 helices, giving an FDR of 21/(26 + 21) (45%) (Table 1). These error rates are significantly worse than their ideal values (0%) and confirm the known inaccuracy of current secondary structure prediction methods without experimental guidance (see, e.g., refs 16 and 17).

Accuracy of Modeling with SHAPE Data. We then acquired SHAPE data for each RNA in 50 mM Na-HEPES (pH 8.0), 10 mM $\mathrm{MgCl_2}$, and saturating concentrations of ligand (for the three riboswitch domains), using the modification reagent *N*-methylisatoic anhydride (NMIA). Quantitation of data for each RNA involved correction for attenuation of long products, background subtraction, and averaging of 12−28 replicates (Table S1 of the Supporting Information) guided by a likelihood framework (methods). The data were in excellent agreement with the expected structures [Figures 1 and 2 (left panels)]. Strong SHAPE reactivities occur mainly at nucleotides that are outside Watson−Crick helices observed in crystallographic models. On the basis of prior work,[17] we expected that inclusion of these data as a pseudoenergy term in the *RNAstructure* algorithm would substantially improve the accuracy of computational models, with a helix-level FNR as low as 0−2%. The improvement was indeed significant, but not to the extent expected (Figure 2, right panels; Table 1). The FNR decreased from 38 to 17% (missing 7 of 42 helices), and the FDR decreased from 45 to 21% (misprediction of 9 helices). In five of the six RNAs, the calculations failed to recover all the crystallographic helices.

Evaluating Sources of Systematic Error. The results described above give a somewhat less optimistic picture of SHAPE-directed modeling than previously published measurements.[17]

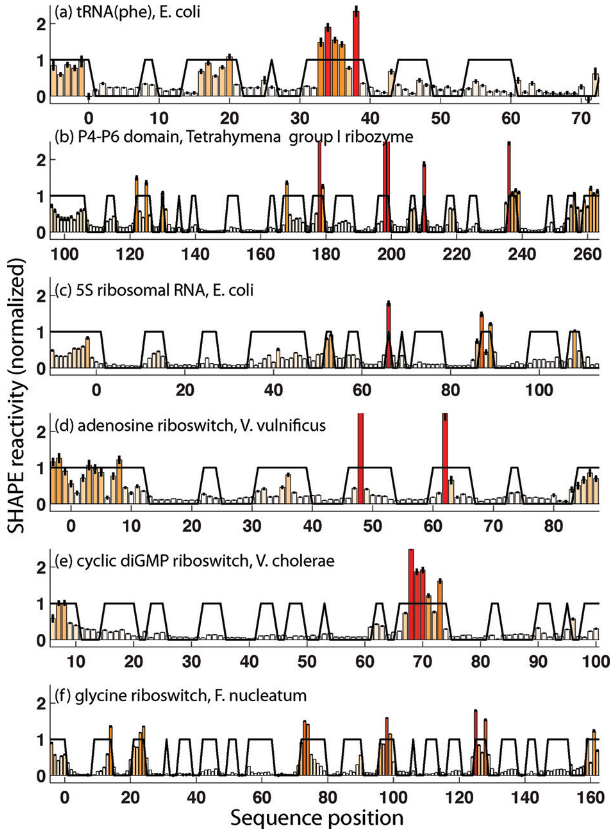


Figure 1. SHAPE reactivities measured at single-nucleotide resolution for six noncoding RNAs of known structure. Black lines mark residues that are paired or unpaired in the crystallographic secondary structures with values of 0.0 or 1.0, respectively.

The differences between SHAPE benchmarks can be most simply ascribed to different test RNAs. Nevertheless, we investigated several other possible systematic explanations for the error rates (FNR and FDR of 17 and 21%, respectively) in our test set. First, we used herein an evaluation scheme to define helix recovery more stringent than those used in previous work,[15−17] which permitted helix register slips of ±1 (see Experimental Procedures). Using those less stringent criteria gave similar FNR and FDR values of 14 and 18%, respectively. Second, we checked for experimental artifacts. Filtering out nucleotides whose SHAPE pseudoenergy errors exceeded 0.4 kcal/mol gave similar FNR and FDR values [14 and 18%, respectively (Table 2)]. Third, to test the quality of our lab's experimental procedures and data processing, we acquired SHAPE measurements on an RNA with a previously published SHAPE-directed model, the hepatitis C virus internal ribosomal entry site domain II. The resulting secondary structure (Figure S2 of the Supporting Information) agreed with prior independent work.[17] Fourth, primer extension with dNTPs containing dITP instead of dGTP reduces errors in quantitating "compressed" bands near G nucleotides[14,22,40] but gives added variance at C nucleotides due to reverse transcriptase pausing (Figure S3 of the Supporting Information and ref 14). Using only data collected with dGTP gave helix-level FNR and FDR values of 12 and 14%, respectively (Table 2), an improvement, but still higher than values of 0−2% achieved for previous test RNAs.[17] The FNR and FDR increased when we used only data collected with dITP (26 and 28%, respectively). Fifth, as an additional check on experimental artifacts, we acquired SHAPE data for all the RNAs with the newly developed 2′-OH acylating reagent 1-methyl-7-nitroisatoic anhydride (1M7);[41] the FNR and FDR for models based on these data were identical to the measurements with the more widely used NMIA (Table 2).

Sixth, model accuracy might be unduly sensitive to the highest or lowest reactivities in the SHAPE data. However, capping "outliers" (see the methods of the Supporting Information); changing the cutoffs for capping; removing outliers; only including high-reactivity data; and excluding SHAPE data for nucleotides near the 5′ and 3′ ends of the RNA did not improve the accuracy (Table 2). Seventh, the pseudoenergy for base pairing is derived from SHAPE data by a logarithmic formula [$\Delta G = m \log(1.0 + \text{SHAPE}) + b$]. Optimizing parameters m and b did not affect the FNR and improved the FDR only slightly [from 21 to 18% (Table 2)]. Eighth, choices in normalizing SHAPE data can affect the modeling, but varying the normalization by factors between 0.5- and 2-fold did not significantly improve the accuracy (Table 2). Ninth, we explored whether energy inaccuracies stem from *RNAstructure*'s thermodynamic parameters, SHAPE data, or both. Comparing energies of crystallographic and model structures indicated that both thermodynamic and SHAPE energies are imbalanced to favor incorrect models [by averages of 1.7 and 1.3 kcal/mol, respectively (Table S3 of the Supporting Information)]. Additionally, shifting the Boltzmann weight balances by increasing the modeling temperature from 24 to 37 °C did not change the error rates (Table 2). Tenth, we additionally tested for algorithm biases by recomputing models in *ViennaRNA*[29] rather than *RNAstructure*, but overall, the FNR and FDR both increased [to 26 and 28%, respectively (Table 2)].

Evidence against Crystal versus Solution Structure Discrepancies. Having found no straightforward explanation for SHAPE-directed modeling errors from systematic errors in experimental data acquisition, data processing, or modeling protocols, we investigated whether there might be differences between these RNAs' secondary structures in available crystals and under our experimental solution conditions, as occurred in prior work with extracted rRNA.[17] Several lines of evidence disfavor this hypothesis in our cases. For tRNA(phe), the P4−P6 domain, the 5S rRNA, and the purine and cyclic di-GMP riboswitch, independent crystallographic models of several variants indicate that the RNAs' secondary structures agree with phylogenetic analysis and are furthermore robust to different conditions, binding partners, and crystallographic contexts (Table S1 of the Supporting Information). In addition, while flanking sequences added to constructs (Table S1 of the Supporting Information) might disrupt the target domains, we designed these sequences to prevent such pairings and checked this lack of pairings by calculations with and without SHAPE data (Figure S1 of the Supporting Information and Figure 2).

Misfolding to kinetically trapped secondary or tertiary structures could lead to differences in solution chemical mapping data compared to those expected from crystallographic structures. To test this possibility, we acquired data for the RNAs after incubating them in 10 mM Na-MES (pH 6.0) and 10 mM $MgCl_2$ for 30 min at 50 °C (refolding conditions developed for large ribozymes[42,43]); the resulting reactivities were indistinguishable from those of RNAs without the refolding treatment (see, e.g., Figure S3 of the Supporting Information for tRNA data). Similarly, we tested for adverse effects of dimethyl sulfoxide (DMSO, used to solubilize the

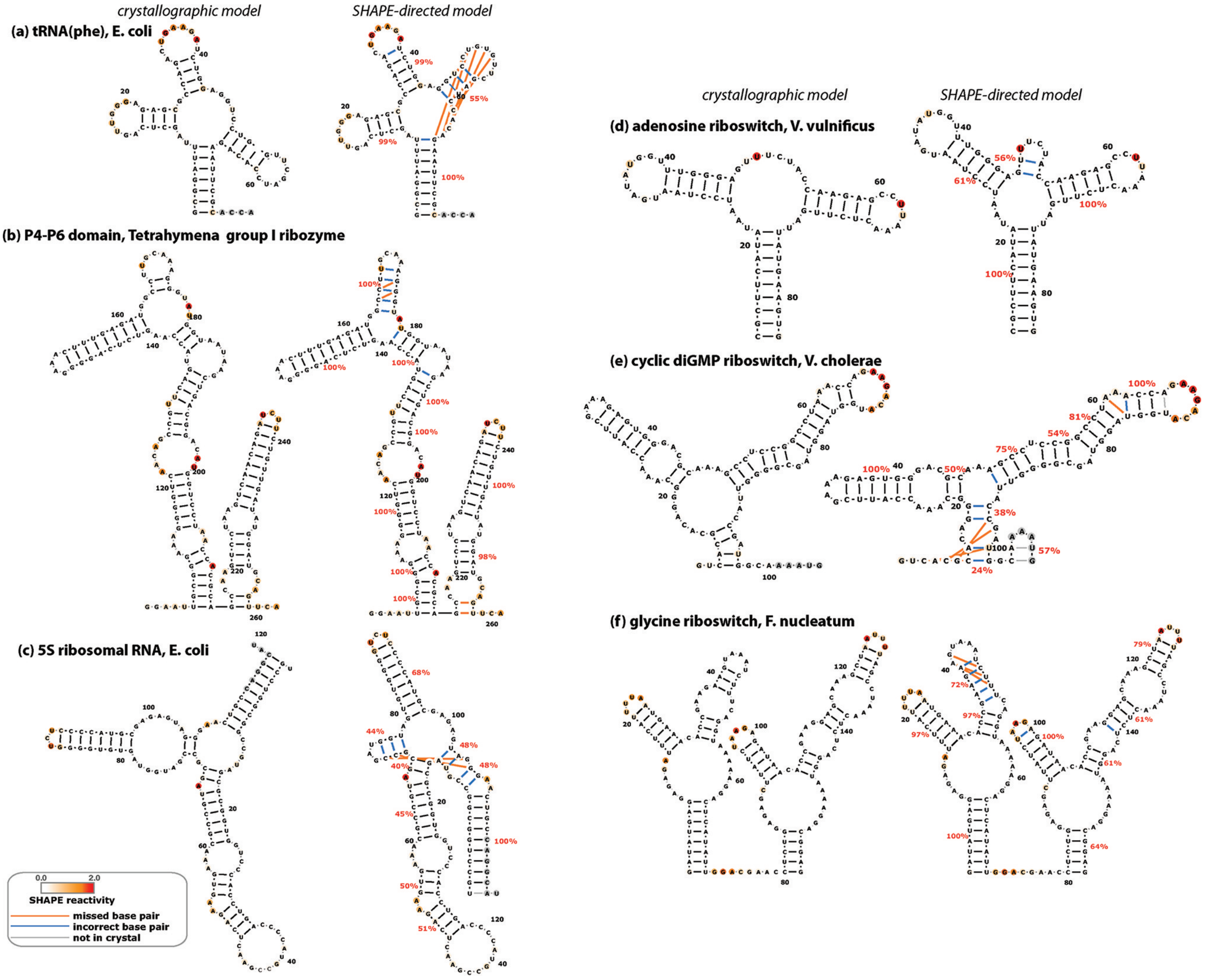

crystallographic model
SHAPE-directed model
(a) tRNA(phe), E. coli
(b) P4-P6 domain, Tetrahymena group I ribozyme
(c) 5S ribosomal RNA, E. coli
(d) adenosine riboswitch, V. vulnificus
(e) cyclic diGMP riboswitch, V. cholerae
(f) glycine riboswitch, F. nucleatum
0.0
2.0
SHAPE reactivity
missed base pair
incorrect base pair
not in crystal

Figure 2. Crystallographic (left) and SHAPE-directed (right) secondary structure models for a benchmark of noncoding RNAs. SHAPE reactivities are shown as colors on bases and match colors in Figure 1. Cyan lines mark incorrect base pairs. Orange lines mark crystallographic base pairs missing in each model. Gray lines mark base pairs in regions outside the crystallized construct. Helix confidence estimates from bootstrap analyses are given as red percentages. For the sake of clarity, flanking sequences (see Table S1 of the Supporting Information) are not shown.

SHAPE reagent)[44] by repeating measurements at lower DMSO concentrations (10% vs 25%); SHAPE data were indistinguishable under the two conditions (Figure S3 of the Supporting Information gives tRNA data).

In addition to these results disfavoring differences in crystal and solution structures, our solution measurements gave positive evidence of the RNAs folding into the correct tertiary conformations. The P4−P6 domain and the 5S rRNA gave changes in their metal core and loop E regions, respectively, upon addition of Mg^{2+}, as expected from prior biophysical analysis (e.g., refs 45−48). The three riboswitches gave SHAPE changes with and without their ligands (Figure S4 of the Supporting Information). Most strongly, we have subjected each of these RNAs to the mutate-and-map method, a two-dimensional extension of chemical mapping,[13,14] and observed near-complete recovery of the crystallographic helices (98% sensitivity[49]), indicating that the dominant solution structure matches the structure determined by crystallography.

Assessing Information Content and Confidence by Bootstrapping. A final explanation for the errors of SHAPE-directed structure models could be that the experimental data have insufficient information content to define the secondary structure. That is, the data, while accurately reflecting each RNA's solution conformation, are also consistent with incorrect secondary structures with similar calculated energy. Indeed, the minimum energy model can be highly sensitive to small changes in the SHAPE data (see the tRNA example in Figure S5a−c of the Supporting Information), and in some cases, the incorrect lowest-energy SHAPE-directed model is within 1 kcal/mol of the crystallographic model [see $tRNA^{phe}$ and the cyclic di-GMP riboswitch (Table S3 of the Supporting Information)]. Unfortunately, quantitatively interpreting energy differences between models [as well as partition function-based base pair probabilities, which are skewed to high values (see Figure S5a of the Supporting Information)] is currently complicated by the nonphysical nature of the SHAPE pseudoenergies. For example, a useful confidence value should be a good approximation of the actual modeling accuracy. In contrast, the mean base pair probability value over all predicted helices is 88%, suggesting a false discovery rate of 12% (100% − 88%), underestimating the actual error rate of 21%.

Table 2. Effects of Variations of Data Processing or Modeling on the Accuracy of SHAPE-Directed Secondary Structure Modeling

variation in modeling[a]	TP[b]	FP[b]	FNR (%)	FDR (%)
no SHAPE data (control)	26	21	38.1	44.7
SHAPE-directed, default parameters	**35**	**9**	**16.7**	**20.5**
remove residues with large errors[c]	36	8	14.3	18.2
use only data collected with dITP during primer extension	31	12	26.2	27.9
use only data collected with dGTP during primer extension	37	6	11.9	14.0
use 1M7 instead of NMIA reagent	35	9	16.7	20.5
cap outliers[d] at cutoff value	35	9	16.7	20.5
cap outliers[d] at 2.0	35	9	16.7	20.5
remove five additional residues from 5′ and 3′ ends	35	9	16.7	20.5
remove residues with SHAPE < 0.5	32	14	23.8	30.4
optimized m and b in pseudoenergy relation[e]	35	8	16.7	18.6
adjust normalization 2-fold	35	8	16.7	18.6
adjust normalization 1.5-fold	35	9	16.7	20.5
adjust normalization 0.75-fold	34	12	19.0	26.1
adjust normalization 0.5-fold	31	14	26.2	31.1
RNAstructure T = 37 °C (not 24 °C)	35	9	16.7	20.5
ViennaRNA[f] instead of *RNAstructure*	32	10	23.8	23.8

[a] All variations are described relative to "default conditions" (bold) using *RNAstructure* version 5.3. [b] The total number of crystallographic helices is 42. TP, true positives; FP, false positives. [c] Any residues whose estimated measurement error of SHAPE reactivity would give errors of more than ±0.4 kcal/mol if included in a base pair, using the SHAPE pseudoenergy relation. [d] Outliers were defined as in the normalization procedure: those with values above a cutoff equal to 1.5 times the interquartile range. [e] Pseudoenergy applied to base-paired nucleotides given by $m\log(1.0 + \text{SHAPE}) + b$. Default parameters in *RNAstructure* are as follows: $m = 2.6$, and $b = -0.8$. The combinations of m and b that gave optimal accuracies for this benchmark were 3.0 and −0.6, respectively. [f] *ViennnaRNA* version 1.8.4, using the default parameter set of ref 15.

We therefore estimated the helix-by-helix confidence of SHAPE-directed models through a nonparametric bootstrapping procedure, inspired by techniques developed to evaluate phylogenetic trees from multiple-sequence alignments.[27,28,50] We generated 400 mock replicates of each data set by resampling with replacement the SHAPE data for individual residues, generating secondary structure models directed by these mock data sets, and evaluating the frequency with which each predicted helix appeared in these replicates (Figure S5b of the Supporting Information and percentage values in Figure 2). One-fourth of the modeled helices (11 of 44) appeared with bootstrap values of <55%, suggesting insufficient information for confident determination of their structure; 7 of these 11 helices were indeed incorrect. Encouragingly, the 33 helices with bootstrap values of >55% included only two errors, of which one was a single-nucleotide register shift. Further, these bootstrap values are robust to small changes in the SHAPE data (see the tRNA example in Figure S6d,e of the Supporting Information). Finally, the overall mean of the helix bootstrap values was 77%. This result predicts a false discovery rate of 23% (100% − 77%), in accord with the actual rate of 21%. Bootstrap analysis therefore appears to be well-suited for evaluating confidence in SHAPE-directed models.

Bootstrap Analysis of an Independent Test Case: The HIV-1 Genome Model. As a final demonstration of the utility of bootstrapping confidence estimation, we investigated the information content of an external data set. Recent application of the SHAPE method to the 9173-nucleotide RNA genome extracted from the NL4-3 HIV-1 virion gave a secondary structure hypothesis containing 429 helices,[51] and the quantitated SHAPE reactivity data have been published. Employing these data and previously used modeling constraints (including division of the modeled genome into five separate domains), the current version of *RNAstructure* (version 5.3) largely recovers the prior working model (Table S4 of the Supporting Information). Furthermore, bootstrapping revealed additional useful information. The 57-nucleotide 5′ TAR element, two helices with lengths greater than 10 bp in the *gag-pol* region, and the signal-peptide stem at the 5′ end of *gp120* give bootstrap values of >95%. These regions are thus high-confidence features of the SHAPE-directed model. Overall, however, 236 of 429 helices in the model have bootstrap confidence estimates of <50%. (If base pairs across the five assumed domains are permitted, more helices are found with such low bootstrap values.) The bootstrap value averaged over all predicted helices is 49%; excluding 59 stems in the prior model that are not recovered with the current version of *RNAstructure* gives a similar value of 55%. These results suggest that much of the HIV-1 secondary structure remains uncertain, even in regions that are strongly protected from SHAPE modification (Figure S7 of the Supporting Information). These low-confidence regions either form single structures that are poorly constrained by the SHAPE data or interconvert between multiple well-formed structures in solution. A tabulation of the helix-by-helix confidence estimates in Table S4 of the Supporting Information should help guide further dissection of these uncertain regions by other chemical and structural approaches.

DISCUSSION

With recent experimental and computational acceleration, nucleotide-resolution chemical mapping permits the characterization of noncoding RNAs at an unprecedented rate. Nevertheless, the resulting data are not always sufficient to determine the molecule's secondary structure, especially if additional tertiary interactions are present. The helix-level error rates found in this study of six highly structured RNAs (false negative rate and false discovery rate of 17 and 20%, respectively) are significantly better than those of models generated without data (38 and 45%, respectively) but higher than those for prior SHAPE modeling test cases (FNR of 0−2%). The modeling inaccuracy found herein is similar to error rates (FNR of ~24%) found in benchmarks with other chemical modifiers, including dimethyl sulfate, kethoxal, and carbodiimide,[16] albeit on different RNAs and with different modeling protocols. Side-by-side tests on the same model RNAs will be necessary to rigorously compare conventional chemical approaches with SHAPE-based methods.

As with all structure characterization methods, SHAPE-directed models cannot be considered "determined structures" but instead are useful hypotheses, especially if accompanied by confidence estimates. This work proposes a bootstrapping analysis for SHAPE-directed modeling that provides such confidence values for novel RNAs. In addition to giving correct predictions for helix accuracy in six crystallized RNAs, bootstrapping analysis of the HIV-1 RNA genome finds numerous regions with high degrees of uncertainty in the RNA's current SHAPE-directed working model. More information-rich multidimensional methods, such as NMR and the mutate-and-map chemical approaches,[13,14] should be able to test these predictions and, more generally, help attain accurate models of noncoding RNAs.

ASSOCIATED CONTENT

Ⓢ Supporting Information

Methods for likelihood-based data processing, four tables with detailed benchmark information and systematic error analyses, and eight supporting figures. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: rhiju@stanford.edu. Phone: (650) 723-5976. Fax: (650) 723-6783.

ACKNOWLEDGMENTS

We thank authors of *RNAstructure* and *ViennaRNA* for making the source code freely available; M. Elazar and J. Glenn for the gift of 1M7; and J. Lucks, D. Mathews, K. Weeks, and the Das lab for comments on the manuscript.

REFERENCES

(1) Cruz, J. A., and Westhof, E. (2009) The dynamic landscapes of RNA architecture. *Cell 136*, 604−609.

(2) Gesteland, R. F., Cech, T. R., and Atkins, J. F. (2006) *The RNA world: The nature of modern RNA suggests a prebiotic RNA world*, Cold Spring Harbor Laboratory Press, Plainview, NY.

(3) Noller, H. F. (2005) RNA structure: Reading the ribosome. *Science 309*, 1508−1514.

(4) Pedersen, J. S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E. S., Kent, J., Miller, W., and Haussler, D. (2006) Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol. 2*, e33.

(5) Collins, K. (2006) The biogenesis and regulation of telomerase holoenzymes. *Nat. Rev. Mol. Cell Biol. 7*, 484−494.

(6) Staley, J. P., and Guthrie, C. (1998) Mechanical devices of the spliceosome: Motors, clocks, springs, and things. *Cell 92*, 315−326.

(7) Panning, B., Dausman, J., and Jaenisch, R. (1997) X chromosome inactivation is mediated by Xist RNA stabilization. *Cell 90*, 907−916.

(8) Winkler, W. C., and Breaker, R. R. (2003) Genetic control by metabolite-binding riboswitches. *ChemBioChem 4*, 1024−1032.

(9) Regulski, E. E., and Breaker, R. R. (2008) In-line probing analysis of riboswitches. *Methods Mol. Biol. 419*, 53−67.

(10) Wilkinson, K. A., Gorelick, R. J., Vasa, S. M., Guex, N., Rein, A., Mathews, D. H., Giddings, M. C., and Weeks, K. M. (2008) High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states. *PLoS Biol. 6*, e96.

(11) Mitra, S., Shcherbakova, I. V., Altman, R. B., Brenowitz, M., and Laederach, A. (2008) High-throughput single-nucleotide structural mapping by capillary automated footprinting analysis. *Nucleic Acids Res. 36*, e63.

(12) Das, R., Karanicolas, J., and Baker, D. (2010) Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat. Methods 7*, 291−294.

(13) Kladwang, W., and Das, R. (2010) A mutate-and-map strategy for inferring base pairs in structured nucleic acids: Proof of concept on a DNA/RNA helix. *Biochemistry 49*, 7414−7416.

(14) 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.

(15) 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.

(16) 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.

(17) 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.

(18) Levitt, M. (1969) Detailed molecular model for transfer ribonucleic acid. *Nature 224*, 759−763.

(19) Sussman, J. L., and Kim, S. (1976) Three-dimensional structure of a transfer RNA in two crystal forms. *Science 192*, 853−858.

(20) Brunel, C., Romby, P., Westhof, E., Ehresmann, C., and Ehresmann, B. (1991) Three-dimensional model of *Escherichia coli* ribosomal 5 S RNA as deduced from structure probing in solution and computer modeling. *J. Mol. Biol. 221*, 293−308.

(21) Leontis, N.B., and Westhof, E. (1998) The 5S rRNA loop E: Chemical probing and phylogenetic data versus crystal structure. *RNA 4*, 1134−1153.

(22) Mills, D. R., and Kramer, F. R. (1979) Structure-independent nucleotide sequence analysis. *Proc. Natl. Acad. Sci. U.S.A. 76*, 2232−2235.

(23) Das, R., Laederach, A., Pearlman, S. M., Herschlag, D., and Altman, R. B. (2005) SAFA: Semi-automated footprinting analysis software for high-throughput quantification of nucleic acid footprinting experiments. *RNA 11*, 344−354.

(24) Yoon, S., Kim, J., Hum, J., Kim, H., Park, S., Kladwang, W., and Das, R. (2011) HiTRACE: High-Throughput Robust Analysis for Capillary Electropherograms. *Bioinformatics 27*, 1798−1805.

(25) 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.

(26) Aviran, S., Trapnell, C., Lucks, J. B., Mortimer, S. A., Luo, S., Schroth, G. P., Doudna, J. A., Arkin, A. P., and Pachter, L. (2011) Modeling and automation of sequencing-based characterization of RNA structure. *Proc. Natl. Acad. Sci. U.S.A. 108*, 11069−11074.

(27) Efron, B., and Tibshirani, R. J. (1998) *An Introduction to the Bootstrap*, Chapman & Hall, Boca Raton, FL.

(28) Efron, B., Halloran, E., and Holmes, S. (1996) Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci. U.S.A. 93*, 13429−13434.

(29) Hofacker, I. L. (2004) RNA secondary structure analysis using the Vienna RNA package. *Current Protocols in Bioinformatics*, Chapter 12, Unit 12, p 12, Wiley, New York.

(30) Darty, K., Denise, A., and Ponty, Y. (2009) VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics 25*, 1974−1975.

(31) 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.

(32) Cate, J. H., Gooding, A. R., Podell, E., Zhou, K., Golden, B. L., Kundrot, C. E., Cech, T. R., and Doudna, J. A. (1996) Crystal structure of a group I ribozyme domain: Principles of RNA packing. *Science 273*, 1678−1685.

(33) Mandal, M., and Breaker, R. R. (2004) Adenine riboswitches and gene activation by disruption of a transcription terminator. *Nat. Struct. Mol. Biol. 11*, 29−35.

(34) 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.

(35) Sudarsan, N., Lee, E. R., Weinberg, Z., Moy, R. H., Kim, J. N., Link, K. H., and Breaker, R. R. (2008) Riboswitches in eubacteria sense the second messenger cyclic di-GMP. *Science 321*, 411−413.

(36) 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.

(37) 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.

(38) Mandal, M., Lee, M., Barrick, J. E., Weinberg, Z., Emilsson, G. M., Ruzzo, W. L., and Breaker, R. R. (2004) A glycine-dependent riboswitch that uses cooperative binding to control gene expression. *Science 306*, 275−279.

(39) Butler, E. B., Xiong, Y., Wang, J., and Strobel, S. A. (2011) Structural basis of cooperative ligand binding by the glycine riboswitch. *Chem. Biol. 18*, 293−298.

(40) 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.

(41) 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.

(42) Russell, R., and Herschlag, D. (1999) New pathways in folding of the *Tetrahymena* group I RNA enzyme. *J. Mol. Biol. 291*, 1155−1167.

(43) Russell, R., Das, R., Suh, H., Travers, K. J., Laederach, A., Engelhardt, M. A., and Herschlag, D. (2006) The paradoxical behavior of a highly structured misfolded intermediate in RNA folding. *J. Mol. Biol. 363*, 531−544.

(44) Hickey, D. R., and Turner, D. H. (1985) Solvent effects on the stability of A7U7p. *Biochemistry 24*, 2086−2094.

(45) Takamoto, K., Das, R., He, Q., Doniach, S., Brenowitz, M., Herschlag, D., and Chance, M. R. (2004) Principles of RNA compaction: Insights from the equilibrium folding pathway of the P4-P6 RNA domain in monovalent cations. *J. Mol. Biol. 343*, 1195−1206.

(46) Correll, C. C., Freeborn, B., Moore, P. B., and Steitz, T. A. (1997) Metals, motifs, and recognition in the crystal structure of a 5S rRNA domain. *Cell 91*, 705−712.

(47) Lemay, J. F., Penedo, J. C., Tremblay, R., Lilley, D. M., and Lafontaine, D. A. (2006) Folding of the adenine riboswitch. *Chem. Biol. 13*, 857−868.

(48) Kwon, M., and Strobel, S. A. (2008) Chemical basis of glycine riboswitch cooperativity. *RNA 14*, 25−34.

(49) Kladwang, W., VanLang, C. C., Cordero, P., and Das, R. (2011) Two-dimensional chemical mapping for non-coding RNAs: The mutate-and-map strategy. *Nat. Chem.*, in revision.

(50) Felsenstein, J. (1985) Confidence limits on phylogenies: An approach using the bootstrap. *Evolution 39*, 783−791.

(51) Watts, J. M., Dang, K. K., Gorelick, R. J., Leonard, C. W., Bess, J. W. Jr., Swanstrom, R., Burch, C. L., and Weeks, K. M. (2009) Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature 460*, 711−716.