Quantitative Dimethyl Sulfate Mapping for Automated RNA Secondary Structure Inference

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

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

S Supporting Information

ABSTRACT: For decades, dimethyl sulfate (DMS) mapping has informed manual modeling of RNA structure in vitro and in vivo. Here, we incorporate DMS data into automated secondary structure inference using an energy minimization framework developed for 2′-OH acylation (SHAPE) mapping. On six noncoding RNAs with crystallographic models, DMS-guided modeling achieves overall false negative and false discovery rates of 9.5% and 11.6%, respectively, comparable to or better than those of SHAPE-guided modeling, and bootstrapping provides straightforward confidence estimates. Integrating DMS−SHAPE data and including 1-cyclohexyl(2-morpholinoethyl) carbodiimide metho-*p*-toluene sulfonate (CMCT) reactivities provide small additional improvements. These results establish DMS mapping, an already routine technique, as a quantitative tool for unbiased RNA secondary structure modeling.

Understanding the many biological functions of RNAs, from genetic regulation to catalysis, requires accurate portraits of the RNAs' folds. Among biochemical tools available for interrogating RNA structure, chemical mapping or "footprinting" uniquely permits rapid characterization of any RNA or ribonucleoprotein system in solution at single-nucleotide resolution (see, e.g., refs 1 and 2). Chemical mapping is being advanced by several groups through new approaches for chemical modification, coupling to high-throughput readouts, rapid data processing, high-throughput mutagenesis, and incorporation into structure prediction algorithms.[3−7]

Perhaps the most widely used RNA chemical probe is dimethyl sulfate (DMS).[8−11] DMS modification of the Watson−Crick edge of adenosines or cytosines (at N1 or N3, respectively) blocks reverse transcription, so that reactivities can be obtained by primer extension at single-nucleotide resolution. Nucleotides that appear to be most strongly protected or reactive to DMS can be inferred to be base-paired or unpaired, respectively. This qualitative or "binary" information can be used for RNA structure modeling by manual or automatic methods.[10,12] More recently developed methods, such as selective 2′-hydroxyl acylation with primer extension (SHAPE),[6] give reactivities that correlate with Watson−Crick base pairing for all nucleotide types, providing more data points than DMS. Indeed, when incorporated into free energy minimization algorithms as energetic bonuses, called pseudoenergies, SHAPE data can recover RNA secondary structures with a high level of accuracy.[11] Further, nonparametric bootstrapping (repeating the algorithms on data sets resampled with replacement) can identify regions with poor confidence.[13] Nevertheless, this pseudoenergy framework has not been leveraged for prior chemical approaches such as DMS mapping, despite the wide use of these data in in vitro, in vivo, and in virio contexts.[9,12,14,15]

We present herein a benchmark of pseudoenergy-guided secondary structure modeling based on DMS data for six noncoding RNAs: unmodified *Escherichia coli* tRNA^{phe},[16] the P4−P6 domain of the *Tetrahymena* group I ribozyme,[17] *E. coli* 5S rRNA,[12] and three ligand-bound domains from bacterial riboswitches (the *Vibrio vulnificus add* adenine riboswitch,[18] the *Vibrio cholerae* cyclic di-GMP riboswitch,[19] and the *Fusobacterium nucleatum* glycine riboswitch[20]). In all cases, crystallographic data, confirmed by solution analyses with the two-dimensional mutate-and-map approach,[21] have provided "gold-standard" secondary structures (Table S1 of the Supporting Information) for evaluating the method's accuracy. The challenging nature of this benchmark is confirmed by the poor accuracy of the *RNAstructure* algorithm without data (Table 1). These models miss 38% of true helices [false negative rate (FNR)], and 45% of the returned helices are incorrect [false discovery rate (FDR)].

We measured DMS reactivities and estimated errors, inferred from three to eight replicates for each of the six RNAs (Figures S4−S9 and Table S1 of the Supporting Information). Analogous to prior SHAPE studies,[11,13] we incorporated these DMS data into *RNAstructure* by transforming them into pseudoenergies, giving favorable energies or penalties depending on whether paired nucleotides were DMS-protected or reactive, respectively. We tested pseudoenergy frameworks based on both a previous ad hoc formula and an empirically derived statistical potential [inspired by techniques in three-dimensional structure prediction (see Methods and Figure S1 of the Supporting Information)]. The two methods gave consistent secondary structures. Because primer extension primarily reads out DMS reactivity at adenosines and cytosines, we excluded reactivities at other bases when performing structure modeling. DMS-guided modeling of the six ncRNAs gave an FNR of 9.5% and an FDR of 11.6% (Table 1 and Figure 1; see also Table S2 of the Supporting Information), more than 3-fold better than without the data. These error rates

Received: July 2, 2012
Revised: August 12, 2012
Published: August 22, 2012

Table 1. Performance of Free Energy Minimization Guided by Reactivity-Derived Pseudoenergies from DMS and SHAPE Chemical Modifications[a]

		no data		DMS		SHAPE		DMS and SHAPE	
	cryst.	TP	FP	TP	FP	TP	FP	TP	FP
tRNA^{phe}	4	2	3	4	0	4	0	4	0
adenine riboswitch	3	2	3	3	1	3	1	3	1
cdGMP riboswitch	8	6	2	6	0	8	0	8	0
5S rRNA	7	1	9	6	3	6	3	6	3
P4–P6 RNA	11	10	1	10	1	9	1	9	1
glycine riboswitch	9	5	3	9	0	8	0	9	0
total	42	26	21	38	6	38	5	39	5
FNR		38.1%		9.5%		9.5%		7.1%	
FDR		44.7%		11.6%		13.6%		11.4%	
sensitivity		61.9%		90.5%		90.5%		92.9%	
PPV		55.3%		88.4%		86.4%		88.6%	

[a] Abbreviations: TP, true positives; FP, false positives; cryst., number of helices in the crystallographic model; FNR, false negative rate (1 − TP/total); FDR, false discovery rate [FP/(TP + FP)]; sensitivity, 1 − FNR; PPV, positive predictive value (1 − FDR).

are lower than those previously achieved by SHAPE-directed modeling (FNR of 17% and FDR of 21% on the same RNAs[13]). Furthermore, the DMS-guided FNR and FDR values are equal to and lower than, respectively, values for SHAPE-based measurements in which primer extension was conducted without deoxyinosine triphosphate (FNR of 9.6% and FDR of 13.6%) to avoid known artifacts.[13]

We were surprised that DMS mapping gave similar or better information content, compared to SHAPE data, as the latter provides reactivities at approximately twice the number of nucleotides per RNA. [Indeed, restricting the algorithm to use SHAPE data at adenines and cytosines gave worse models (see Table S3 of the Supporting Information).] An explanation for our results derives from distinct SHAPE and DMS signatures at nucleotides that are not in Watson–Crick secondary structure but that nevertheless form noncanonical interactions [see, e.g., A37 in the *F. nucleatum* glycine riboswitch (Figure 2A)]. These nucleotides appear to be protected from the SHAPE reaction and thus receive pseudoenergies that incorrectly reward their pairings inside Watson–Crick secondary structure. However, these same nucleotides can expose their Watson–Crick edges to solvent and react strongly with DMS, signifying that they are outside Watson–Crick helices. The DMS-guided modeling can thus return the correct secondary structure in regions where the SHAPE data cannot distinguish Watson–Crick from non-Watson–Crick base pairs (compare panels B and C of Figure 2).

Reactivity histograms (Figure 2D,E) further support the enhanced predictive power of DMS vis-à-vis SHAPE. DMS mapping better distinguishes between nucleotides inside Watson–Crick helices and nucleotides outside helices [see also the receiver operating characteristic curve and quantitation (Figure S2 of the Supporting Information)].

Like SHAPE-guided modeling, DMS-directed structure inference still produces errors (Table 1), e.g., for the central junction of the 5S rRNA (Figure 1). Some of these errors may be resolved through better incorporation of the DMS-derived pseudoenergies at, e.g., isolated, or "singlet", base pairs. Nevertheless, as with SHAPE modeling, these erroneous regions can be pinpointed by estimating helix-by-helix confidence values through nonparametric bootstrapping (Methods of the Supporting Information and ref 13; see also Figure S3 of the Supporting Information). For example, this procedure gives a high degree of confidence (≥90%) at almost all helices in the correctly recovered structure of the glycine riboswitch but low levels of confidence (<50%) throughout the imperfect 5S rRNA DMS model (Figure 1).

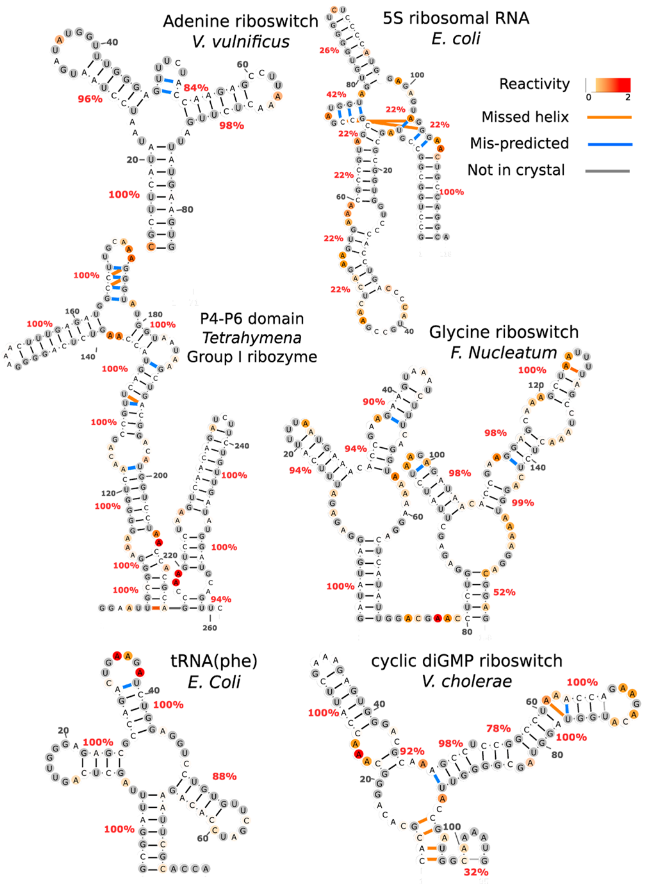

Figure 1. Pseudoenergy-guided secondary structure models using DMS data on six noncoding RNAs. DMS data and secondary structure models for *E. coli* tRNA^{phe}, the P4–P6 domain of the *Tetrahymena* group I ribozyme, *E. coli* 5S rRNA, the *V. vulnificus add* adenine riboswitch, the *V. cholerae* cyclic di-GMP riboswitch, and the *F. nucleatum* glycine riboswitch. Missed base pairs are highlighted with blue lines; mispredicted base pairs are denoted with orange lines. Helix bootstrap confidence values are colored red. G and U nucleotides that do not give DMS signals in primer extension and nucleotides with unavailable reactivities are colored gray.

For many applications, DMS and SHAPE measurements can be acquired in parallel, so we sought to determine if their combination might improve automated secondary structure inference. Application of both sets of pseudoenergies gave a slight improvement in the algorithm's accuracy (FNR of 7.1% and FDR of 11.4%). In addition, we performed measurements with a reagent that primarily modifies Waston–Crick edges of guanosine and uracil, 1-cyclohexyl(2-morpholinoethyl) carbodiimide metho-*p*-toluene sulfonate (CMCT).[22] Incorporation of these data into *RNAstructure* gave poorer accuracy modeling than the DMS- or SHAPE-guided modeling described above [FNR of 14.3% and FDR or 18.2% (see Table S4 of the Supporting Information)], consistent with weaker discrimination between paired and unpaired residues (Figures S1 and

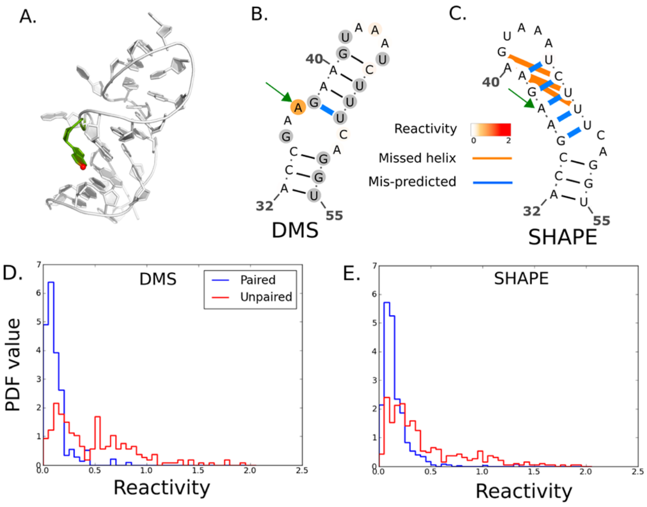

Figure 2. DMS vis-à-vis SHAPE for secondary structure inference. (A) The P3 hairpin of the glycine riboswitch is correctly predicted by DMS-guided modeling, but not by SHAPE. A37 (green) has its Watson−Crick edge exposed, making its N1 atom (red sphere) accessible to DMS modification that guides *RNAstructure* to the correct helix (B). However, A37 (green arrow) is stabilized by local interactions, protecting it from SHAPE modification, resulting in an incorrect SHAPE-predicted helix (C). (D and E) Reactivity histograms for DMS (D) and SHAPE (E) for all chemical mapping data on the six noncoding RNAs.

S2 of the Supporting Information). Integrating CMCT with DMS and/or SHAPE data did not improve accuracy (Table S2 of the Supporting Information). CMCT gives weak reactivities in bases that are unpaired but still stacked (e.g., see ref 23), reducing its information content for discriminating unpaired and paired nucleotides.

The benchmark results presented here establish that chemical mapping with DMS can achieve prediction accuracies comparable to those of the SHAPE protocol using pseudoenergies to guide free energy minimization. DMS has been extensively used both in vitro and in vivo, for time-resolved RNA folding, precise thermodynamic analysis, and mapping RNA−protein interfaces.[9,12,14,15,22] Sophisticated techniques for optimizing the reaction rate and its quenching have been developed.[9,24] Applying automated structure modeling, as demonstrated herein, will allow researchers to better take advantage of this large body of previous work. Furthermore, future studies may find it advantageous to perform both DMS and SHAPE approaches in parallel. Along with bootstrapping,[13] comparison of separate DMS-guided versus SHAPE-guided secondary structure models will permit rapid assessment of systematic errors and thus provide more accurate inferences.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Supporting methods, figures, and model accuracy tables. This material is available free of charge via the Internet at http://pubs.acs.org. Single nucleotide-resolution data are available at http://rmdb.stanford.edu.

■ AUTHOR INFORMATION

Corresponding Author

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

Funding

This work is supported by the Burroughs-Wellcome Foundation (CASI to R.D.), a Hewlett-Packard Stanford Graduate Fellowship (to P.C.), and the National Institutes of Health (T32 HG000044 to C.C.V. and R01 GM102519 to R.D.).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank the authors of *RNAstructure* for making the source code freely available and members of the Das lab for comments on the manuscript.

■ REFERENCES

(1) Black, D. L., and Pinto, A. L. (1989) *Mol. Cell. Biol. 9*, 3350−3359.

(2) Moazed, D., and Noller, H. F. (1991) *Proc. Natl. Acad. Sci. U.S.A. 88*, 3725−3728.

(3) Mitra, S., Shcherbakova, I. V., Altman, R. B., Brenowitz, M., and Laederach, A. (2008) *Nucleic Acids Res. 36*, e63.

(4) Yoon, S., Kim, J., Hum, J., Kim, H., Park, S., Kladwang, W., and Das, R. (2011) *Bioinformatics 27*, 1798−1805.

(5) Kladwang, W., Cordero, P., and Das, R. (2011) *RNA 17*, 522−534.

(6) Wilkinson, K. A., Merino, E. J., and Weeks, K. M. (2006) *Nat. Protoc. 1*, 1610−1616.

(7) Lucks, J. B., Mortimer, S. A., Trapnell, C., Luo, S., Aviran, S., Schroth, G. P., Pachter, L., Doudna, J. A., and Arkin, A. P. (2011) *Proc. Natl. Acad. Sci. U.S.A. 108*, 11063−11068.

(8) Peattie, D. A., and Gilbert, W. (1980) *Proc. Natl. Acad. Sci. U.S.A. 77*, 4679−4682.

(9) Tijerina, P., Mohr, S., and Russell, R. (2007) *Nat. Protoc. 2*, 2608−2623.

(10) Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M., and Turner, D. H. (2004) *Proc. Natl. Acad. Sci. U.S.A. 101*, 7287−7292.

(11) Deigan, K. E., Li, T. W., Mathews, D. H., and Weeks, K. M. (2009) *Proc. Natl. Acad. Sci. U.S.A. 106*, 97−102.

(12) Leontis, N. B., and Westhof, E. (1998) *RNA 4*, 1134−1153.

(13) Kladwang, W., VanLang, C. C., Cordero, P., and Das, R. (2011) *Biochemistry 50*, 8049−8056.

(14) Lempereur, L., Nicoloso, M., Riehl, N., Ehresmann, C., Ehresmann, B., and Bachellerie, J. P. (1985) *Nucleic Acids Res. 13*, 8339.

(15) Wells, S. E., Hughes, J. M., Igel, A. H., and Ares, M. (2000) *Methods Enzymol. 318*, 479−493.

(16) Byrne, R. T., Konevega, A. L., Rodnina, M. V., and Antson, A. A. (2010) *Nucleic Acids Res. 38*, 4154−4162.

(17) Cate, J. H., Gooding, A. R., Podell, E., Zhou, K., Golden, B. L., Kundrot, C. E., Cech, T. R., and Doudna, J. A. (1996) *Science 273*, 1678−1685.

(18) 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) *Chem. Biol. 11*, 1729−1741.

(19) Smith, K. D., Lipchock, S. V., Livingston, A. L., Shanahan, C. A., and Strobel, S. A. (2010) *Biochemistry 49*, 7351−7359.

(20) Butler, E. B., Xiong, Y., Wang, J., and Strobel, S. A. (2011) *Chem. Biol. 18*, 293−298.

(21) Kladwang, W., VanLang, C. C., Cordero, P., and Das, R. (2011) *Nat. Chem. 3*, 954−962.

(22) Planning, S. (2000) in *Current protocols in nucleic acid chemistry* (Beaucage, S. L., et al., Eds.) Chapter 6, pp 1−21, Wiley, New York.

(23) Sripakdeevong, P., Kladwang, W., and Das, R. (2011) *Proc. Natl. Acad. Sci. U.S.A. 108*, 20573−20578.

(24) Das, R., Karanicolas, J., and Baker, D. (2010) *Nat. Methods 7*, 291−294.