Published in final edited form as:

Biochemistry. 2012 September 11; 51(36): 7037–7039. doi:10.1021/bi3008802.

Quantitative DMS mapping for automated RNA secondary structure inference

Pablo Cordero¹, Wipapat Kladwang², Christopher C. VanLang³, and Rhiju Das^{1,2,4,*}

¹Department of Biomedical Informatics, Stanford University, Stanford CA 94305

²Department of Biochemistry, Stanford University, Stanford CA 94305

³Department of Chemical Engineering, Stanford University, Stanford CA 94305

⁴Department of Physics, Stanford University, Stanford CA 94305

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) map-ping. On six non-coding RNAs with crystallographic models, DMS-guided modeling achieves overall false negative and false discovery rates of 9.5% and 11.6%, comparable or better than SHAPE-guided modeling; and bootstrapping provides straightforward confidence estimates. Integrating DMS/SHAPE data and including CMCT reactivities give small additional improvements. These results establish DMS mapping – an already routine technique – as a quantitative tool for unbiased RNA 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 "foot-printing" uniquely permits rapid characterization of any RNA or ribonucleoprotein system in solution at single-nucleotide resolution [see, e.g. ref. (1, 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 di-methyl 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 most strongly protected or reactive to DMS can be inferred to be base-paired or un-paired, respectively – qualitative or 'binary' information that 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 high accuracy (11). Further, non-parametric bootstrapping – repeating the algorithms on data sets

*To whom correspondence should be addressed. rhiju@stanford.edu. Phone: (650) 723-5976. Fax: (650) 723-6783. .

SUPPORTING INFORMATION Supporting methods, figures, and model accuracy tables are available free of charge at <http://pubs.acs.org>.

resampled with replacement – can identify regions with poor confidence (13). Nevertheless, this pseudo-energy frame-work 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 pseudo-energy-guided secondary structure modeling based on DMS data for six non-coding RNAs: unmodified *E. 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 *V. vulnificus* *add* adenine riboswitch (18), *V. cholerae* cyclic di-GMP riboswitch (19), and *F. 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 (Supporting Table S1) 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 (Supporting Figures S4 to S9 & Table S1). Analogous to prior SHAPE studies (11, 13), we incorporated these DMS data into *RNAstructure* by transforming them into pseudo-energies, giving favorable energies or penalties depending on whether paired nucleotides were DMS-protected or reactive, respectively. We tested pseudo-energy frameworks based on both a previous *ad hoc* formula and an empirically derived statistical potential (inspired by techniques in 3D structure prediction; see Supporting Methods and Figure S1). 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 FDR of 11.6% (Table 1 and Figure 1, see also Table S2), more than three-fold better than without the data. These error rates are lower than those previously achieved by SHAPE-directed modeling [FNR: 17%; FDR: 21% on the same RNAs (13)]. Furthermore, the DMS-guided FNR and FDR values are equal to and lower, respectively, than values for SHAPE-based measurements in which primer extension was carried out without deoxyinosine triphosphate (FNR: 9.6%, FDR: 13.6%) to avoid known artefacts (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 Supporting Table S3.) 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 non-canonical interactions (see, e.g., A37 in the *F. nucleatum* glycine riboswitch; Figure 2A). These nucleotides appear protected from the SHAPE reaction and thus receive pseudo-energies 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 Figures 2B and 2C).

Reactivity histograms (Figures 2G and 2H) 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 receiver operating characteristic curve and quantitation; Supporting Fig. S2.)

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 non-parametric bootstrapping [Supplemental Methods and ref. (13); see also Supporting Figure S3]. For example, this procedure gives high confidence (90%) at almost all helices in the glycine riboswitch but low confidence values (<50 %) throughout the 5S rRNA DMS model (Figure 1).

For many applications, DMS and SHAPE measurements can be carried out in parallel, so we sought to determine if their combination might improve automated secondary structure inference. Application of both sets of pseudo-energies 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 Watson/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 above (FNR of 14.3%, FDR of 18.2%; see Supporting Table S4), consistent with weaker discrimination between paired/unpaired residues (Supporting Figures S1 and S2). Integrating CMCT with DMS and/or SHAPE data did not improve accuracy (Supplemental Table S2). 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 herein establish that chemical mapping with dimethyl sulfate (DMS) can achieve prediction accuracies comparable to the SHAPE protocol using pseudo-energies 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 enable 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 vs. SHAPE-guided secondary structure models will permit rapid assessment of systematic errors and thus provide more accurate inferences.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

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

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.).

REFERENCES

1. Black DL, Pinto AL. Molecular and cellular biology. 1989; 9:3350–3359. [PubMed: 2552294]
2. Moazed D, Noller HF. Proc. Natl. Acad. Sci. U. S. A. 1991; 88:3725–3728. [PubMed: 2023922]
3. Mitra S, Shcherbakova IV, Altman RB, Brenowitz M, Laederach A. Nucleic acids research. 2008; 36:e63. [PubMed: 18477638]

4. Yoon S, Kim J, Hum J, Kim H, Park S, Kladwang W, Das R. *Bioinformatics*. 2011; 27:1798–1805. [PubMed: 21561922]
5. Kladwang W, Cordero P, Das R. *RNA*. 2011; 17:522–534. [PubMed: 21239468]
6. Wilkinson KA, Merino EJ, Weeks KM. *Nature protocols*. 2006; 1:1610–1616.
7. Lucks JB, Mortimer SA, Trapnell C, Luo S, Aviran S, Schroth GP, Pachter L, Doudna JA, Arkin AP, Hage J, Meeus M. *Proc. Natl. Acad. Sci. U. S. A.* 2011; 108:11063–11068. [PubMed: 21642531]
8. Peattie DA, Gilbert W. *Proc. Natl. Acad. Sci. U. S. A.* 1980; 77:4679–4682. [PubMed: 6159633]
9. Tijerina P, Mohr S, Russell R. *Nature Protocols*. 2007; 2:2608–2623.
10. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. *Proc. Natl. Acad. Sci. U. S. A.* 2004; 101:7287–7292. [PubMed: 15123812]
11. Deigan KE, Li TW, Mathews DH, Weeks KM. *Proc. Natl. Acad. Sci. U. S. A.* 2009; 106:97–102. [PubMed: 19109441]
12. Leontis NB, Westhof E. *RNA*. 1998; 4:1134–1153. [PubMed: 9740131]
13. Kladwang W, VanLang CC, Cordero P, Das R. *Biochemistry*. 2011; 50:8049–8056. [PubMed: 21842868]
14. Lempereur L, Nicoloso M, Riehl N, Ehresmann C, Ehresmann B, Bachellerie JP. *Nucleic Acids Res.* 1985; 13:8339. [PubMed: 2417197]
15. Wells SE, Hughes JM, Igel AH, Ares M. *Methods in Enzymology*. 2000; 318:479–493. [PubMed: 10890007]
16. Byrne RT, Konevega AL, Rodnina MV, Antson AA. *Nucleic Acids Res.* 2010; 38:4154–4162. [PubMed: 20203084]
17. Cate JH, Gooding AR, Podell E, Zhou K, Golden BL, Kundrot CE, Cech TR, Doudna JA. *Science*. 1996; 273:1678–1685. [PubMed: 8781224]
18. Serganov A, Yuan Y-R, Pikovskaya O, Polonskaia A, Malinina L, Phan AT, Hobartner C, Micura R, Breaker RR, Patel DJ. *Chemistry & biology*. 2004; 11:1729–1741. [PubMed: 15610857]
19. Smith KD, Lipchock SV, Livingston AL, Shanahan CA, Strobel SA. *Biochemistry*. 2010; 49:7351–7359. [PubMed: 20690679]
20. Butler EB, Xiong Y, Wang J, Strobel SA. *Chemistry & biology*. 2011; 18:293–298. [PubMed: 21439473]
21. Kladwang W, VanLang CC, Cordero P, Das R. *Nature chemistry*. 2011; 3:954–962.
22. Planning, S. *Current protocols in nucleic acid chemistry* edited by Serge L Beaucage et al Chapter 6. 2000. p. 1-21.
23. Sripakdeevong P, Kladwang W, Das R. *Proceedings of the National Academy of Sciences*. 2011; 108:20573–20578.
24. Das R, Karanicolas J, Baker D. *Nature methods*. 2010; 7:291–294. [PubMed: 20190761]

FIGURE 1.

Pseudo-energy-guided secondary structure models using DMS data on 6 non-coding 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, *V. cholerae* cyclic di-GMP riboswitch, and *F. nucleatum* glycine riboswitch. Missed base pairs are highlighted in blue lines; mis-predicted base pairs are indicated by orange lines. Helix bootstrap confidence values are shown in red. G and U nucleotides which don't give DMS signals in primer extension and nucleotides with unavailable reactivities are colored gray.

FIGURE 2. DMS vis-à-vis SHAPE for secondary structure inference

(A) The L3 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 N-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). Reactivity histograms for DMS and SHAPE, for all chemical mapping data on the 6 non-coding RNAs are shown in (D) and (E) respectively.

Table 1

Performance of free energy minimization guided by reactivity-derived pseudo-energies from DMS and SHAPE chemical modifications.

	Cryst	No data	DMS			SHAPE			DMS+SHAPE		
			T	P	FP	T	P	FP	T	P	FP
tRNA ^{phe}	4	2	3	4	0	4	0	4	0	4	0
adenine rbs.	3	2	3	3	1	3	1	3	1	3	1
cdGMP rbs.	8	6	2	6	0	8	0	8	0	8	0
5S rRNA	7	1	9	6	3	6	3	6	3	6	3
P4IP6 RNA	11	10	1	10	1	9	1	9	1	9	1
glycine rbs.	9	5	3	9	0	8	0	9	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%		

Abbreviations: TP, true positives; FP false positives; Cryst., number of helices in crystallographic model; FNR, False negative rate = $1 - TP/Total$; FDR, False discovery rate = $FP/(TP + FP)$; Sensitivity = $(1 - FNR)$; PPV, Positive predictive value = $(1 - FDR)$