BIOCHEMISTRY
including biophysical chemistry & molecular biology

Article

pubs.acs.org/biochemistry

Statistical Analysis of SHAPE-Directed RNA Secondary Structure Modeling

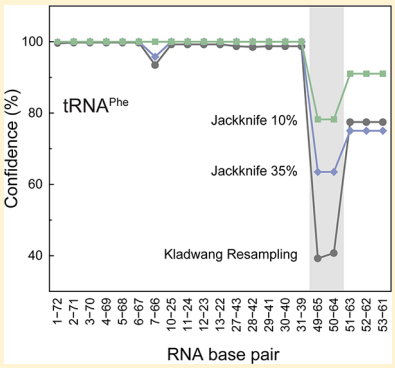
Srinivas Ramachandran,[†] Feng Ding,[†,§] Kevin M. Weeks,[‡] and Nikolay V. Dokholyan*[,†]

[†]Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, United States

[‡]Department of Chemistry, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-3290, United States

[§]Department of Physics and Astronomy, Clemson University, Clemson, South Carolina 29634, United States

ABSTRACT: The ability to predict RNA secondary structure is fundamental for understanding and manipulating RNA function. The information obtained from selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE) experiments greatly improves the accuracy of RNA secondary structure prediction. Recently, Das and colleagues [Kladwang, W., et al. (2011) *Biochemistry 50*, 8049−8056] proposed a "bootstrapping" approach for estimating the variance and helix-by-helix confidence levels of predicted secondary structures based on resampling (randomizing and summing) the measured SHAPE data. We show that the specific resampling approach described by Kladwang et al. introduces systematic errors and underestimates confidence in secondary structure prediction using SHAPE data. Instead, a leave-data-out jackknife approach better estimates the influence of a given experimental data set on SHAPE-directed secondary structure modeling. Even when 35% of the data were left out in the jackknife approach, the confidence levels of SHAPE-directed secondary structure prediction were significantly higher than those calculated by Das and colleagues using bootstrapping. Helix confidence levels were thus underestimated in the recent study, and the resampling approach implemented by Kladwang et al. is not an appropriate metric for evaluating SHAPE-directed secondary structure modeling.



Despite an explosion in discoveries of RNAs and their functional roles in biology, accurate knowledge of the structures of these molecules is incomplete.[1] Knowledge of information encoded in RNA structures, especially in RNA secondary structures (the pattern of base pairs), is necessary for understanding and manipulating RNA function. Computational RNA secondary structure prediction methods[2,3] have been widely used to generate structural hypotheses in RNA research. Secondary structures predicted from sequence alone often have significant errors, however, including both falsely predicted and missing base pairs.[1,4,5] Incorporation of experimental structural information derived from chemical probing experiments can significantly improve secondary structure predictions.[6] For example, the comprehensive and quantitative information available from selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE) probing experiments greatly improves the accuracy of RNA secondary structure prediction.[5,7,8]

With widespread adoption of SHAPE and other experimental approaches for directing RNA secondary structure prediction, rigorous *a priori* estimation of the confidence level of a given RNA secondary structure prediction would constitute a major and welcome advance. Recently, Das and colleagues proposed a "bootstrapping" (resampling) approach for estimating the variance and confidence level of predicted secondary structures based on resampling of measured SHAPE data.[9] On the basis of their statistical study, Das and colleagues suggested that the confidence level of SHAPE-derived RNA secondary structure prediction is ∼77%. Follow-up analysis of the work of Das and colleagues revealed that important components of their experimental work were not consistent with recommended practices in using and evaluating SHAPE technologies.[9] In this work, we show that the specific resampling approach developed by Das and colleagues is unphysical, introduces systematic error into the resampled data, and results in a large underestimation of the confidence of SHAPE-directed secondary structure prediction. As detailed here, a leave-data-out jackknife approach more accurately estimates the influence of a given experimental data set on SHAPE-directed secondary structure modeling.

MATERIALS AND METHODS

SHAPE-Directed Secondary Structure Prediction. We compared the resampling approach of Kladwang et al. and our jackknife approach on four RNA molecules: tRNA^{Phe} (76 nucleotides), the adenine (71 nucleotides) and cyclic-di-GMP (97 nucleotides) riboswitches, and 5S rRNA (120 nucleotides). The SHAPE data for these RNAs are presented in the preceding paper[9] and were used to direct secondary structure prediction using RNAstructure as described previously.[5]

Received: June 7, 2012
Revised: December 23, 2012
Published: January 4, 2013

   dx.doi.org/10.1021/bi300756s | *Biochemistry* 2013, 52, 596−599

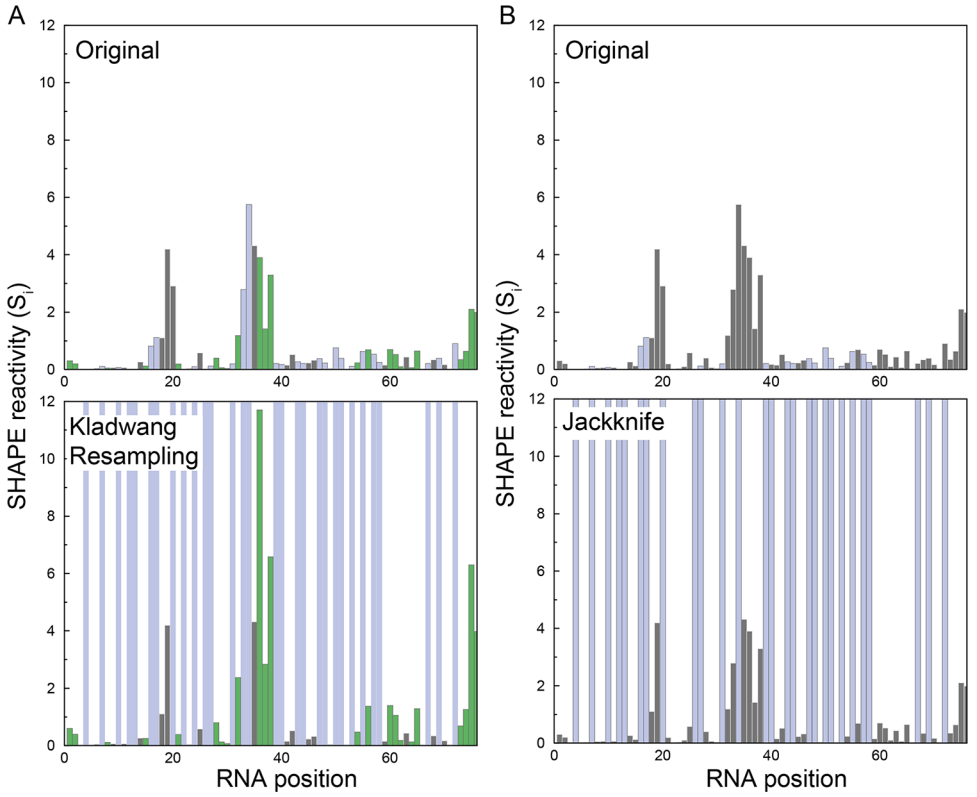


Figure 1. Generation of decoy SHAPE data sets via resampling—bootstrap and jackknife approaches. (A) $tRNA^{Phe}$ SHAPE reactivities determined using a standard[9] approach (top) and representative decoy data set generated by the specific resampling algorithm of Kladwang et al.[9] (bottom). The SHAPE reactivities not present in the decoy data sets (top) are shown as blue bars in the original data set and are shaded blue in the decoy data set (bottom). Kladwang resampling also results in modified SHAPE reactivities (green bars; notice the increase in SHAPE reactivities at these positions in the decoy data set). (B) SHAPE reactivities (top) and representative decoy data set as generated by jackknifing (bottom). The jackknife approach results in SHAPE reactivities that are not picked in the decoy data set (blue bars).

Kladwang Resampling. For RNA molecules with nucleotide positions 1, 2, 3, ..., N, bootstrap decoys were generated as described previously.[9] To summarize, N nucleotide positions were randomly picked with repetition. Thus, some nucleotide positions were picked multiple times, whereas others were not picked at all. If a nucleotide position was picked three times, for example, the SHAPE reactivity of that position was multiplied by 3. In total, 400 such decoy SHAPE data sets were generated, and the RNA secondary structure was predicted for each decoy SHAPE data set using RNAstructure.

Jackknife. The jackknife decoy was generated by picking $(1 - f)N$ nucleotide positions randomly, where f is the fraction of data omitted. We performed jackknife analysis using f values of 0.1 and 0.35. Thus, for a 100-nucleotide RNA molecule, the jackknife decoy data set will contain SHAPE values for 90 and 65 nucleotides, respectively, picked randomly from the 100 values. Four hundred such decoy data sets were generated, and the RNA secondary structure was predicted for each decoy SHAPE data set using RNAstructure. For both jackknife and bootstrapping approaches, we calculated the percentage of these 400 decoy RNA secondary structures that contained each base pair observed in the secondary structure predicted by the original SHAPE data. The generation of a decoy data set for each approach is illustrated in Figure 1. Control simulations performed with 100 decoy data sets converged with those using 400 decoys; we report the results of the 400-member decoy data sets here. Performing a jackknife analysis with 35% omitted data represents an extreme case, because it is uncommon to miss so many SHAPE measurements. Thus, the jackknifing results presented here represent an overestimation of the variability of SHAPE-directed structure modeling.

RESULTS AND DISCUSSION

SHAPE reactivity at a position (S_i) in an RNA reports the conformational flexibility of a given nucleotide and is inversely related to the propensity of the position to form a base pair or tertiary contact.[7,10] The measured SHAPE reactivity can be used as an experimentally based correction to bias an RNA secondary structure prediction.[5] To estimate confidence values for individual helices in a given secondary structure prediction, Das and colleagues proposed a bootstrapping approach for resampling the measured SHAPE data.[9] This approach was motivated by the demonstrated utility of bootstrapping analyses in evaluation of phylogenetic trees.[11] Bootstrapping entails reshuffling a given set of data with repetition; critically, for bootstrapping to be statistically justified, the calculated quantity usually needs to be independent of the order of the shuffled data, an assumption that is generally valid in the construction of phylogenetic trees.

In the approach taken by Das and colleagues, SHAPE reactivities were generated for an RNA molecule by "resampling with repetition" the original SHAPE reactivities. Given SHAPE reactivities S_i for nucleotides i = 1, 2, 3, ..., N, resampling with repetition entails picking N indices randomly, whereby an index value can be selected multiple times for a

dx.doi.org/10.1021/bi300756s | Biochemistry 2013, 52, 596−599

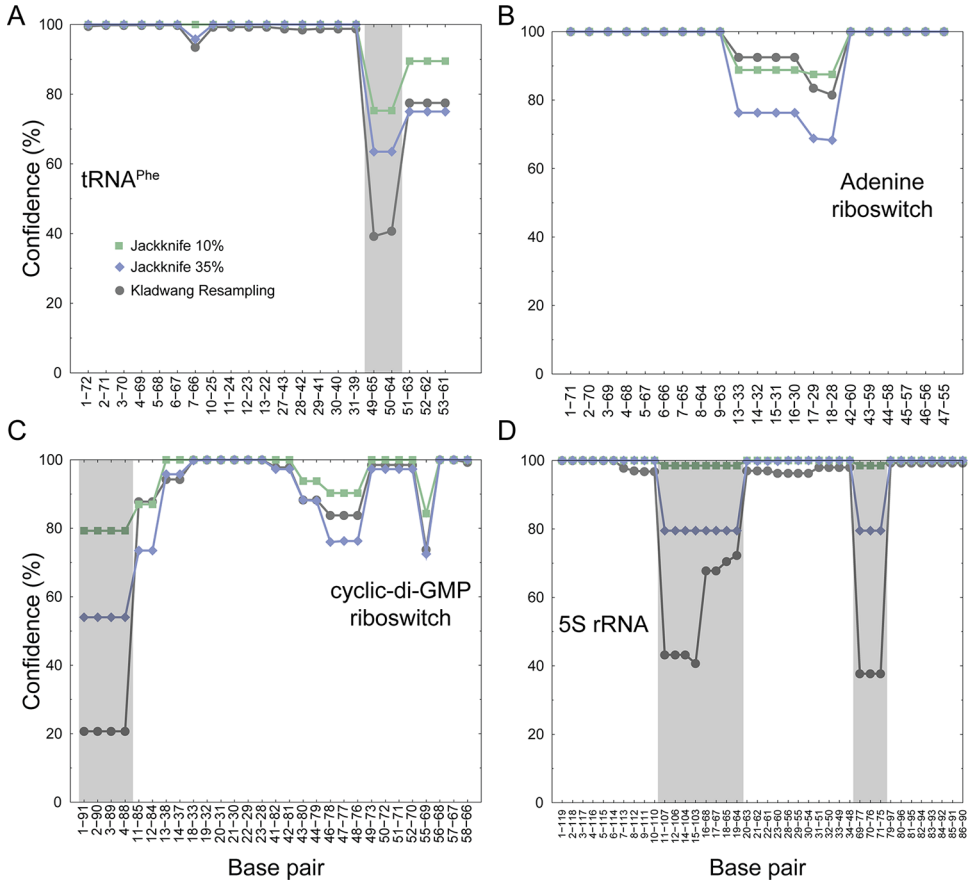


Figure 2. Confidence estimation for SHAPE-directed secondary structure modeling. Confidence estimates for SHAPE-predicted base pairs, calculated using the specific resampling algorithm of Kladwang et al.[9] (gray circles) and using jackknife procedures that omit either 10% (green squares) or 35% (blue diamonds) of the data for (A) $tRNA^{Phe}$, (B) adenine riboswitch, (C) cyclic-di-GMP riboswitch, and (D) 5S rRNA. Gray shading emphasizes regions in which the resampling–bootstrap approach underestimated confidence values as compared to removing (an extreme) 35% of the experimental SHAPE data.

given position, such that some nucleotides are not picked at all and some are picked multiple times. This specific implementation is not appropriate for treatment of SHAPE data as SHAPE reactivities are not independent: both the actual reactivity and the position associated with a given reactivity are required for secondary structure modeling. As SHAPE data are thus inherently unsuitable for "resampling with replacement", Das and colleagues retained the positional information of SHAPE reactivity after resampling (hence, there was no shuffling per se).[9] In addition, when a position was sampled multiple times during bootstrapping, the SHAPE reactivity of the position was increased by the amount equal to the original reactivity of the position each time (R. Das, personal communication). Hence, the relative error introduced at each position in this resampling approach is nS_i, where n is the number of repetitions (0, 1, 2, ...) (Figure 1A,B). This treatment of repetition is intrinsically different from that of the repetition introduced by a standard bootstrapping approach as implemented for multiple-sequence alignments, which increases the weight of the repeated column in the alignment and thus enhances both sequence similarity and differences. In contrast, in the approach used by Das and colleagues, repetition of high SHAPE values increases the probability of breaking a base pair, whereas repetition of low SHAPE values does not increase the probability of forming a base pair.

This problem is most significant in regions of low, but non-zero, SHAPE reactivities that are often base-paired and was compounded by SHAPE signal processing errors in the prior work.[9] Several occurrences of multiplying the SHAPE reactivity value under this approach would effectively destabilize a helix in which the original SHAPE data would otherwise clearly score as unreactive, and likely paired, overall. Thus, the bootstrapping approach as implemented by Das and colleagues[9] introduces systematic error because the errors are integer multiples of the original data itself. In addition, the decoy data sets generated were not the same size because, after each round of shuffling to retain positional information, the number of data points differs from the number in the original data set. Because the errors added to the data set are integer multiples of the data itself, the level of perturbation is high, resulting in many data sets that are highly dissimilar to the experimental data (Figure 1A). Overall, this approach results in an underestimation of the confidence of secondary structure prediction from the SHAPE data.

An alternate statistical method for estimation of the influence of a specific experimental data set on secondary structure modeling is the jackknife approach. Resampling by the jackknife approach has also been used to estimate the confidence of phylogenetic trees.[12] With jackknifing, the fluctuation is introduced simply through omission of a fraction of the data, and the number of data points in each decoy data set is identical (Figure 1B). The key variable in jackknife resampling is the fraction of excluded data.

To quantify the extent of underestimation of the confidence of SHAPE prediction by the method used by Das and

dx.doi.org/10.1021/bi300756s | *Biochemistry* 2013, 52, 596–599

colleagues, we used the jackknife method to evaluate the recovery of SHAPE-predicted base pairs for four RNA molecules ($tRNA^{Phe}$, the adenine and cyclic-di-GMP riboswitches, and 5S rRNA) by generating decoys with up to 35% of the SHAPE data randomly removed. SHAPE data were measured using the documented approach for performing the SHAPE experiment,[8,13] which is substantially different from that introduced by Das and colleagues.[9] Experimentally, it is uncommon to have 35% of SHAPE data missing, and few experimentalists would try to predict a structure with such a high level of missing information; we therefore also performed a jackknife analysis with 10% of the data omitted, the latter corresponding to a more realistic practical occurrence.

In general, both the resampling approach of Kladwang et al. and jackknifing identified the same helices as being less well-defined by the SHAPE data (Figure 2). However, the approach of Kladwang et al. underestimated the confidence of the helices in three of the SHAPE-directed structure models, those of $tRNA^{Phe}$, the cyclic-di-GMP riboswitch, and 5S rRNA. Estimated confidence values based on bootstrapping were lower than confidence estimates obtained using decoy data sets that were missing (an extreme) 35% of the experimental data (Figure 2). In contrast, using the jackknife approach with a more physically and experimentally realistic 10% of the data omitted, the majority of the base pairs have 100% prediction confidence values, and the lowest is ∼80%, supporting the general robustness of the SHAPE-directed secondary structure models.

A second problem with the specific resampling approach created by Das and co-workers lies in their helix-by-helix interpretation. The bootstrap resampling of the SHAPE data introduces noise that perturbs the relative free energy of each RNA structure. Longer helices with lower free energies are less sensitive to this perturbation and are more likely to have high confidence values in prediction, especially given that Das and colleagues define the bootstrap value of a helix as the maximum of the bootstrap values of base pairs across that helix.[9] By this definition, longer helices have more base pairs and are more likely to have at least one base pair with a high "bootstrap value". In contrast, shorter and less stable helices are more sensitive to perturbations and thus more prone to break under perturbation. Therefore, the estimated bootstrap value of a helix primarily reflects the stability of the helix rather than the underlying SHAPE data.

In summary, the specific bootstrapping procedure proposed by Das and colleagues for resampling the SHAPE data is unphysical, introduces systematic error into the resampled data, and results in an underestimation of the confidence of SHAPE-directed secondary structure modeling. The calculated confidence levels obtained via this approach are not an appropriate metric for estimation of the accuracy of experimentally directed RNA secondary structure prediction. Instead, this work supports use of the jackknife approach for generating resampled SHAPE data and estimating the sensitivity of predicted secondary structures to the underlying SHAPE data set. The more general issue of *a priori* identification of individual highly probable helices within a given experimentally directed RNA structure model remains a major research challenge.

AUTHOR INFORMATION

Corresponding Author

*Address: 3097 Genetic Medicine Building, Campus Box 7260, Chapel Hill, NC 27599. E-mail: dokh@unc.edu. Phone: (919) 843-2513.

Author Contributions

S.R. and F.D. contributed equally to this work.

Funding

This work was supported by National Institutes of Health Grants GM080742 to N.V.D. and AI068462 to K.M.W.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Shapiro, B. A., Yingling, Y. G., Kasprzak, W., and Bindewald, E. (2007) Bridging the gap in RNA structure prediction. *Curr. Opin. Struct. Biol. 17*, 157−165.

(2) Mathews, D. H. (2006) Revolutions in RNA secondary structure prediction. *J. Mol. Biol. 359*, 526−532.

(3) 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.

(4) Mathews, D. H., and Turner, D. H. (2006) Prediction of RNA secondary structure by free energy minimization. *Curr. Opin. Struct. Biol. 16*, 270−278.

(5) 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.

(6) 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.

(7) Merino, E. J., Wilkinson, K. A., Coughlan, J. L., and Weeks, K. M. (2005) RNA structure analysis at single nucleotide resolution by selective 2′-hydroxyl acylation and primer extension (SHAPE). *J. Am. Chem. Soc. 127*, 4223−4231.

(8) 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.

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

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

(11) Efron, B., Halloran, E., and Holmes, S. (1996) Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci. U.S.A. 93*, 7085−7090.

(12) Soltis, P. S., and Soltis, D. E. (2003) Applying the bootstrap in phylogeny reconstruction. *Statistical Science 18*, 256−267.

(13) 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.