# Structure probing data enhances RNA-RNA interaction prediction

Milad Miladi [1], Soheila Montaseri [1], Rolf Backofen [1,2,*], and Martin Raden [1,*]

[1]Bioinformatics Group, Department of Computer Science, University of Freiburg, Georges-Koehler-Allee 106, D-79110 Freiburg, Germany and
[2]Center for Biological Signaling Studies (BIOSS), University of Freiburg, Germany
*To whom correspondence should be addressed.
mmann,backofen@informatik.uni-freiburg.de

**Abstract**

**Summary:** Structure probing data has been shown to improve thermodynamics-based RNA structure prediction. However, this type of data has not been used to improve the prediction of RNA-RNA interactions. This is even more promising as the type of information (chemical reactivity as provided by SHAPE) is closely tied to the accessibility of nucleotides, which is an essential part for scoring RNA-RNA interactions. Here we show how that such experimental data can be incorporated seamlessly into accessibility-based RNA-RNA interaction prediction approaches, as implemented in IntaRNA. This is possible via the computation and use of unpaired probabilities that incorporate the structure probing information. We show that experimental SHAPE data can significantly improve RNA-RNA interaction prediction. We evaluate our approach by investigating interactions of the spliceosomal U1 RNA with its target splice sites. When SHAPE data is used, known target sites are predicted with increased precision and specificity.
**Availability:** https://github.com/BackofenLab/IntaRNA
**Supplementary material:** https://github.com/BackofenLab/IntaRNA-benchmark-SHAPE
**Contact:** {mmann,backofen}@informatik.uni-freiburg.de
**Keywords:** RNA-RNA interaction prediction, accessibility, RNA structure probing, RNA secondary structure, chemical footprinting, SHAPE

# 1 Introduction

The function of many if not most non-coding (nc)RNA molecules is to act as platforms for inter-molecular interaction, which depends on their structure and sequence. A large number of ncRNAs regulate their target RNA molecules via base-pairing. For instance, small (s)RNAs regulate the translation of their target genes by direct RNA-RNA interactions with the respective messenger (m)RNAs (Wright *et al.*, 2013). To predict such interactions, knowledge about potential interaction sites is needed, i.e. regions not involved in intra-molecular base pairing. State-of-the-art RNA-RNA interaction prediction tools like IntaRNA (Busch *et al.*, 2008; Mann *et al.*, 2017; Raden *et al.*, 2018a) compute unpaired probabilities to gain this accessibility information. While correct within their thermodynamic models, such probabilities do not incorporate all cellular constraints and dynamics that define accessible regions and thus the likelihood for interaction.

The accuracy of RNA structure prediction can be improved when experimental structure probing data such as SHAPE[1] is incorporated (Hajdin *et al.*, 2013; Sükösd *et al.*, 2013; Lotfi *et al.*, 2015). To this end, SHAPE information[2] is converted to pseudo-energy terms (Zarringhalam *et al.*, 2012; Deigan *et al.*, 2009; Washietl *et al.*, 2012) to guide thermodynamic RNA structure prediction methods (Lorenz *et al.*, 2016a,b; Montaseri *et al.*, 2017; Spasic *et al.*, 2018).

As SHAPE reactivity is related to the accessibility of nucleotides, it is even more promising to use such experimental data for improving the accuracy of RNA-RNA interaction prediction. For that reason, we introduce a seamless incorporation of SHAPE data into accessibility-based prediction approaches such as IntaRNA within this manuscript.

We show that SHAPE-guided accessibility prediction improves RNA-RNA interaction prediction. To this end, we study the probabilities of U1 interacting with its pre-mRNA target sites. U1 is involved in pre-mRNA splicing by recognizing the 5' site of introns via inter-molecular base pairing (Hertel and Graveley, 2005). Due to the dynamics and constraints imposed by the spliceosome, it is generally challenging to avoid false positive interaction predictions, which are either wrong predictions of U1's recognition site with (random) regions of the mRNA or predicted interactions of other accessible U1 regions with the mRNA. For that reason, we used U1 as an example to show that *in vivo* probing data effectively reduces false positive predictions in RNA-RNA interaction prediction.

# 2 Methods

Given two RNA molecules with nucleotide sequences $S^1, S^2 \in \{A,C,G,U\}^*$, we define interaction $I$ between $S^1$ and $S^2$ as a set of inter-molecular base pairs (i.e. $I = \{\ (i,j)\ |\ i \in [1,|S^1|] \wedge j \in [1,|S^2|]\}$), that are complementary (i.e. $\forall (i,j) \in I : \{S^1_i, S^2_j\} \in \{\{A,U\},\{C,G\},\{G,U\}\}$) and non-crossing (i.e. $\forall (i,j) \neq (i',j') \in I : i < i' \rightarrow j > j'$). Furthermore, any position forms at most one inter-molecular base pair (i.e. $\forall (i,j),(i',j') : i = i' \leftrightarrow j = j'$). For any interaction $I$, the hybridization energy $E^{hyb}(I)$ can be computed using a standard Nearest-Neighbor energy model (Turner and Mathews, 2010).

The accessibility-based free energy of an interaction $I$ is defined by

$$E(I) = E^{hyb}(I) + ED^1(I) + ED^2(I), \tag{1}$$

where the $ED^{1,2}(\geq 0)$ terms represent the energy (penalty) needed to make the respective interacting subsequences of $S^{1,2}$ unpaired/accessible (Mückstein *et al.*, 2006; Raden *et al.*, 2018b; Wright *et al.*, 2018).

To compute $ED$ terms, we need the left-/right-most base pair of $I$ given by $(l^1, r^2) = \arg\min_{(i,j)\in I}(i)$ and $(r^1, l^2) = \arg\max_{(i,j)\in I}(i)$, respectively. Both base pairs define the interacting subsequences, i.e. $S^1_{l^1..r^1}$ and $S^2_{l^2..r^2}$. Based on that, the penalty terms are given by

$$ED^*(I) \quad = \quad -RT \log\left(Pr^{ss}(S^*_{l^*..r^*})\right) \qquad \text{with } * \in \{1,2\} \tag{2}$$

where $R$ is the gas constant, $T$ is the temperature, and $Pr^{ss}$ denotes the unpaired probability of a given subsequence, which can be efficiently computed (Bernhart *et al.*, 2006; Mückstein *et al.*, 2006).

As discussed above, SHAPE reactivity data can be incorporated into thermodynamic prediction tools via pseudo energy terms (Lotfi *et al.*, 2015; Deigan *et al.*, 2009) as incorporated into the Vienna RNA package (VRNA) (Lorenz *et al.*, 2016b). The latter enables SHAPE-guided computation of unpaired probabilities, i.e. the $Pr^{ss}$ terms from Eq. 2. While SHAPE-guided energy evaluations can not be compared to unconstrained energy values (due to the pseudo-energy terms), unpaired probabilities are compatible, since they are reflecting the accessible structure space rather than individual structures. Thus, SHAPE-constrained $Pr^{ss}_{\text{SHAPE}}$ values can be directly used within the $ED$ computation (Eq. 2), which provides a constrained accessibility-based interaction energy (Eq. 1) without further methodical changes. This approach is implemented in the recent version of IntaRNA e.g. available via Bioconda (Grüning *et al.*, 2018).

To assess the effect of SHAPE data, we define the *spot probability* $Pr^{\text{spot}}$ of an interaction site of interest. A *spot* is defined by a pair of indices $k,l$ for $S^1, S^2$, resp., and $Pr^{\text{spot}}(k,l)$ as the partition function quotient

$$Pr^{\text{spot}}(k,l) = \sum_{I' \in \mathcal{I}^*} exp(-E(I')/RT) / \sum_{I \in \mathcal{I}} exp(-E(I)/RT), \tag{3}$$

where $\mathcal{I}$ denotes the set of all possible interactions and $\mathcal{I}^* \subseteq \mathcal{I}$ the subset of interactions that cover the spot, i.e. position $k,l$ are within the respective interacting subsequences[3] $S^1_{l^1..r^1}$ and $S^2_{l^2..r^2}$ (see above).

[1] Selective 2′-Hydroxyl Acylation analyzed by Primer Extension (SHAPE) (Wilkinson *et al.*, 2006).

[2] For simplicity we refer to probing experiments of all reagents (SHAPE, DMS) as SHAPE.

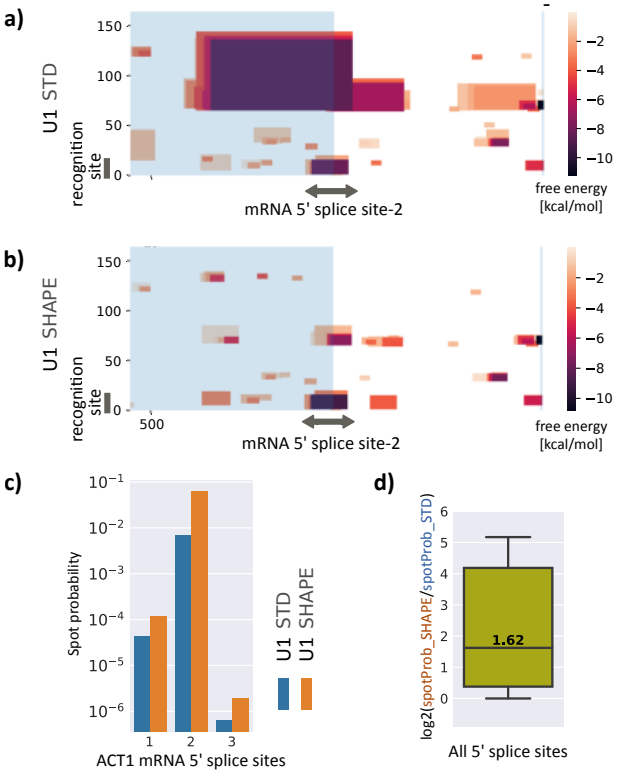


Figure 1: RNA-RNA interaction prediction between spliceosomal RNA U1 with ACT1 mRNA of *Arabidopsis thaliana*. IntARNA interactions predicted between U1 (y-axis) and the region around the second intron splice site of ACT1 mRNA using **(a)** unconstrained (STD) and **(b)** SHAPE-constrained accessibility estimates for U1. **(c)** Spot probabilities of U1 recognition site (spot index = 8) interacting with the three 5' splice sites of ACT1 mRNA (spot = 1st intron index), with (orange) and without (blue) SHAPE constraints. **(d)** Relative splice-site interaction probability (log scale) of all studied mRNAs and spots.

## 3 Results

SHAPE data for U1 was obtained from *in vivo* DMS-seq RNA structure probing of *Arabidopsis thaliana* (Ding *et al.*, 2014). The pre-mRNA sequences for 5 genes including ACT1, which have been previously validated to perform U1-dependent splicing (Yeh *et al.*, 2017), were extracted for the analyses. Figures 1a,b) exemplify the effect of SHAPE-constrained predictions using IntaRNA 2.2.0, VRNA v2.4.7 and pseudo energies following Zarringhalam *et al.* (2012). Without SHAPE constraints, the splice site is predicted to interact with various regions of U1 with high probability (i.e. low energy). In contrast, when using SHAPE-corrected accessibility terms, the splice site is predicted to be the dominant target of U1's recognition site. This interaction, for instance, is shifted upwards from rank 9 (standard prediction) to 3 (SHAPE-constrained) among all predicted interactions of U1 with the ACT1 mRNA. Figure 1c provides the interaction probabilities of U1's recognition site with all three 5' splice sites of ACT1. All splice sites are predicted with increased probability when SHAPE data was used. As shown in Fig. 1a,b), this effect results from a decreased number of wrong low energy interactions, i.e. false positive predictions. Over all mRNAs, the probabilities of correct splice site recognition were increased on average by a factor of 3.08 (Figure 1d). The supplementary material provides further details on data extraction, analysis procedure and the evaluation of all studied mRNAs.

[3]Note, interactions $I \in \mathcal{I}^*$ covering a spot at $k, l$ do not necessarily contain the base pair $(k, l)$, i.e. $k$, $l$ or both can be unpaired.

## 4 Conclusion

Most of the non-coding RNAs perform their function via molecular interactions for which experimental data is still sparse. Prediction of RNA-RNA interaction has proven to be quite useful for detecting targets of sRNA especially in prokaryotes (Backofen and Hess, 2010). However, the false positive rate is still quite high, making RNA-RNA interaction prediction alone too error-prone for eukaryotes.

The only possibility to reduce errors is to combine interaction prediction with other type of data. Here, *in vivo* structure probing data seems especially suited as it represents a multitude of factors that guide RNA structure formation; like the binding of other molecules or kinetic effects. We have shown that SHAPE data indeed improves RNA-RNA interaction prediction accuracy. To this end, we have successfully extended IntaRNA to incorporate SHAPE data in its accessibility computation and to compute spot probabilities of interaction sites. The predicted interaction probabilities of splicesomal U1 RNA with its known target splice sites were significantly improved. This results from a decreased number of false positive (wrong low energy) predictions.

Recently, structure probing has been complemented by next-generation sequencing to quickly obtain single or transcriptome-wide probing data (Kutchko and Laederach, 2017; Choudhary *et al.*, 2017). This produces large data sets that demand for fast methods incorporating the probing data, which is met by our introduced extension of IntaRNA.

## Acknowledgements

We thank Dr. Ronny Lorenz for discussions on SHAPE integration.

## Funding

This work was supported by Bundesministerium fr Bildung und Forschung [031A538A RBC, 031L0106B] and Deutsche Forschungsgemeinschaft [BA 2168/14-1, BA 2168/16-1].

## References


Backofen, R. and Hess, W. R. (2010). Computational prediction of sRNAs and their targets in bacteria. *RNA Biol*, **7**(1), 33–42.

Bernhart, S. H. *et al.* (2006). Local RNA base pairing probabilities in large sequences. *Bioinformatics*, **22**(5), 614–615.

Busch, A. *et al.* (2008). IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, **24**(24), 2849.

Choudhary, K. *et al.* (2017). Comparative and integrative analysis of RNA structural profiling data: current practices and emerging questions. *Quantitative Biology*, **5**(1), 3–24.

Deigan, K. E. *et al.* (2009). Accurate shape-directed RNA structure determination. *Proceedings of the National Academy of Sciences*, **106**(1), 97–102.

Ding, Y. *et al.* (2014). In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature*, **505**(7485), 696.

Grüning, B. *et al.* (2018). Bioconda: A sustainable and comprehensive software distribution for the life sciences. *bioRxiv*. accepted for publication in Nature Methods.

Hajdin, C. E. *et al.* (2013). Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proceedings of the National Academy of Sciences*, **110**(14), 5498–5503.

Hertel, K. J. and Graveley, B. R. (2005). RS domains contact the pre-mRNA throughout spliceosome assembly. *Trends in biochemical sciences*, **30**(3), 115–118.

Kutchko, K. M. and Laederach, A. (2017). Transcending the prediction paradigm: novel applications of SHAPE to RNA function and evolution. *Wiley Interdisciplinary Reviews: RNA*, **8**(1).

Lorenz, R. *et al.* (2016a). RNA folding with hard and soft constraints. *Algorithms for Molecular Biology*, **11**(1), 8.

Lorenz, R. *et al.* (2016b). SHAPE directed RNA folding. *Bioinformatics*, **32**(1), 145–147.

Lotfi, M. *et al.* (2015). RNA secondary structure prediction based on SHAPE data in helix regions. *Journal of theoretical biology*, **380**, 178–182.

Mann, M. *et al.* (2017). IntaRNA 2.0: enhanced and customizable prediction of RNA-RNA interactions. **45**(W1), W435–W439.

Montaseri, S. *et al.* (2017). Evaluating the quality of SHAPE data simulated by k-mers for RNA structure prediction. *Journal of Bioinformatics and Computational Biology*, **15**(06), 1750023.

Mückstein, U. *et al.* (2006). Thermodynamics of RNARNA binding. *Bioinformatics*, **22**(10), 1177.

Raden, M. *et al.* (2018a). Freiburg RNA tools: a central online resource for RNA-focused research and teaching. *Nucleic Acids Research*, page gky329. eprint ahead of print.

Raden, M. *et al.* (2018b). Interactive implementations of RNA structure and RNA-RNA interaction prediction approaches for example-driven teaching. *PLOS Comp Biol*. (accepted).

Spasic, A. *et al.* (2018). Modeling RNA secondary structure folding ensembles using SHAPE mapping data. *Nucleic Acids Research*, **46**(1), 314–323.

Sükösd, Z. *et al.* (2013). Evaluating the accuracy of SHAPE-directed RNA secondary structure predictions. *Nucleic acids research*, **41**(5), 2807–2816.

Turner, D. H. and Mathews, D. H. (2010). NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res*, **38**(Database issue), D280–2.

Washietl, S. *et al.* (2012). RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. *Nucleic acids research*, **40**(10), 4261–4272.

Wilkinson, K. A. *et al.* (2006). Selective 2'-hydroxyl acylation analyzed by primer extension (shape): quantitative RNA structure analysis at single nucleotide resolution. *Nature protocols*, **1**(3), 1610.

Wright, P. R. *et al.* (2013). Comparative genomics boosts target prediction for bacterial small RNAs. **110**(37), E3487–96.

Wright, P. R. *et al.* (2018). Structure and interaction prediction in prokaryotic RNA biology. *Microbiol Spectrum*, **6**(2).

Yeh, C.-S. *et al.* (2017). The conserved AU dinucleotide at the 5' end of nascent U1 snRNA is optimized for the interaction with nuclear cap-binding-complex. *Nucleic acids research*, **45**(16), 9679–9693.

Zarringhalam, K. *et al.* (2012). Integrating chemical footprinting data into RNA secondary structure prediction. *PLOS ONE*, **7**(10), 1–13.