Integration of accessibility data from structure probing into 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: Experimental structure probing data has been shown to improve thermodynamics-based RNA secondary structure prediction. To this end, chemical reactivity information (as provided e.g. by SHAPE) is incorporated, which encodes whether or not individual nucleotides are involved in intra-molecular structure. Since inter-molecular RNA-RNA interactions are often confined to unpaired RNA regions, SHAPE data is even more promising to improve interaction prediction. Here we show how 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 a spliceosomal U1 snRNA transcript with its target splice sites. When SHAPE data is incorporated, known target sites are predicted with increased precision and specificity.
Keywords: RNA-RNA interaction prediction, accessibility, RNA structure probing, RNA secondary structure, chemical footprinting, SHAPE
Availability: https://github.com/BackofenLab/IntaRNA

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 [1]. To predict such interactions, regions not involved in intra-molecular base pairing have to be identified. This "free-to-interact" potential of a region, i.e. its unpaired probabilities, is computed by assessing the fraction of structures where the region is free (unpaired) within the overall structure ensemble of an RNA (see [11] for a detailed introduction). The probabilities are used by state-of-the-art prediction tools like IntaRNA [8] to account for the regions' accessibility. 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 improves when experimental structure probing data, such as SHAPE [13], is incorporated [4]. This is done by converting the chemically sensed reactivity values to pseudo-energy terms. Pseudo-energies are combined with structure formation energies from thermodynamic models, that are used for RNA structure prediction [7, 9, 12]. As SHAPE[1] reactivity is associated with the accessibility of nucleotides, the use of such experimental data is even more promising to improve the accuracy of RNA-RNA interaction prediction methods. For that reason, we introduce a seamless incorporation of SHAPE data into accessibility-based prediction approaches such as IntaRNA.

Recently, structure probing has been complemented by next-generation sequencing techniques to efficiently obtain transcriptome-wide reactivity information [6, 2]. This produces large data sets that demand for fast methods incorporating the probing data, which is met by our extension of IntaRNA introduced in the following.

For a given RNA-RNA interaction I (see supplementary material for detailed formalisms), its accessibility-based free energy is defined by $E(I) = E^{hyb}(I) + ED^1(I) + ED^2(I)$. Therein, $E^{hyb}(I)$ provides the hybridization energy from intermolecular base pairing while the $ED^{1,2}$ terms represent the energy (penalty) needed to make the respective interacting subsequences unpaired/accessible [10] . ED terms are defined by unpaired probabilities Pr^{ss} of the subsequences via $ED(I) = -RT \log(Pr^{ss})$, where R is the gas constant and T the temperature. Detailed introductions on ED computation are provided e.g. in [11, 14]. Computation of unpaired probabilities can be guided by SHAPE data [7]. While SHAPE-guided energy evaluations can not be compared to unconstrained energy values (due to the introduced pseudo-energy terms), unpaired probabilities are compatible, since they are reflecting the accessible structure space rather than individual structures. Thus, SHAPE-constrained Pr^{ss}_{SHAPE} values can be directly used for ED and thus E computation while preserving comparability of the resulting energies.

[1]For simplicity, SHAPE refers to any structure probing experiment (e.g. SHAPE, DMS).

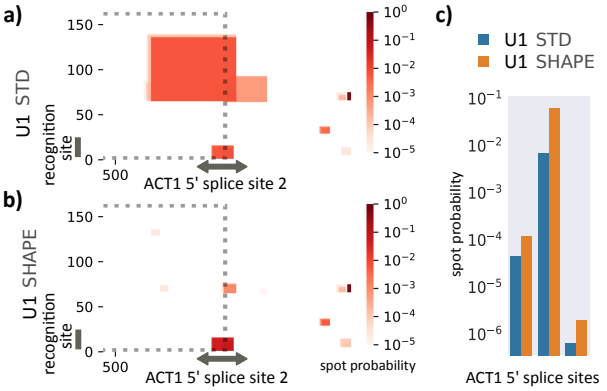


Figure 1: RNA-RNA interaction prediction between spliceosomal RNA U1 with ACT1 mRNA of *Arabidopsis thaliana*. Interaction probabilities predicted between U1 (y-axis) and the region around the second intron splice site of ACT1 coding sequence mRNA using **(a)** unconstrained (STD) and **(b)** SHAPE-constrained accessibility estimates for U1. The dotted lines enclose U1 interactions with exon 2. **(c)** Spot probabilities of U1 recognition site (spot index = 8) interacting with the 5' splice sites of ACT1 (spot = 1st intron index), with SHAPE constraints (orange) and without (blue).

Now, we show that SHAPE-guided accessibility prediction improves RNA-RNA interaction prediction. To this end, we study the probabilities of U1 small nuclear (sn)RNA interacting with pre-mRNA target sites, which is an established example of inter-molecular RNA interaction essential for RNA-splicing in eukaryotes. U1 is involved in pre-mRNA splicing by recognizing the 5' site of introns via inter-molecular base pairing [5]. Due to the dynamics and constraints imposed by the spliceosome, it is generally challenging to avoid false positive interaction predictions, which are either predictions of U1's recognition site with random regions of the mRNA or predicted interactions of other U1 accessible 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.

SHAPE data for U1 was obtained from *in vivo* DMS-seq RNA structure probing of *Arabidopsis thaliana* [3]. We selected the U1 homolog transcript bearing the largest secondary structure distance between the unconstrained and SHAPE-constrained structure prediction. The pre-mRNA sequences for 5 genes were extracted, which have been previously validated to perform U1-dependent splicing [15]. Figures 1a,b) exemplify the effect of SHAPE-constrained predictions for ACT1 mRNA. Without SHAPE constraints, the splice site is predicted to interact with various regions of U1 with high probability; see supplementary material for formalisms. In contrast, the splice site's interaction with U1's recognition site is dominant with high specificity in SHAPE-guided IntaRNA predictions. Furthermore, SHAPE-guided prediction has a higher

precision. Among all predicted interactions of U1 with the ACT1 mRNA, the ranking of the known interaction is improved from 9 to 3 in SHAPE-guided mode. When investigating the accessibility profile of U1 (supplementary material), this mainly results from an increased SHAPE-guided unpaired probability Pr^{ss}_{SHAPE} of U1's recognition site and thus reduced ED penalties.

The interaction probability of U1's recognition site with all three 5' splice sites of ACT1's coding sequence are increased when SHAPE data is incorporated (Fig. 1c). This effect results from a decreased number of false positive predictions (Fig. 1a,b). Following this trend, the probabilities of splice site recognition are improved among all the mRNAs and are on average 3.08 times higher (SHAPE/STD). Further details about data, evaluation, and analyses are provided in the supplementary material.

Acknowledgements

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

Funding

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

References

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

[2] Krishna Choudhary, Fei Deng, and Sharon Aviran. Comparative and integrative analysis of RNA structural profiling data: current practices and emerging questions. *Quantitative Biology*, 5(1):3–24, 2017.

[3] Yiliang Ding, Yin Tang, Chun Kit Kwok, Yu Zhang, Philip C Bevilacqua, and Sarah M Assmann. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature*, 505(7485):696, 2014.

[4] Christine E Hajdin, Stanislav Bellaousov, Wayne Huggins, Christopher W Leonard, David H Mathews, and Kevin M Weeks. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proceedings of the National Academy of Sciences*, 110(14):5498–5503, 2013.

[5] Klemens J Hertel and Brenton R Graveley. RS domains contact the pre-mRNA throughout spliceosome assembly. *Trends in biochemical sciences*, 30(3):115–118, 2005.

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

[7] Ronny Lorenz, Dominik Luntzer, Ivo L. Hofacker, Peter F. Stadler, and Michael T. Wolfinger. SHAPE directed RNA folding. *Bioinformatics*, 32(1):145–147, 2016.

[8] Martin Mann, Patrick R. Wright, and Rolf Backofen. IntaRNA 2.0: enhanced and customizable prediction of RNA-RNA interactions. *NAR*, 45(W1):W435–W439, 2017.

[9] Soheila Montaseri, Fatemeh Zare-Mirakabad, and Mohammad Ganjtabesh. Evaluating the quality of SHAPE data simulated by k-mers for RNA structure prediction. *Journal of Bioinformatics and Computational Biology*, 15(06):1750023, 2017.

[10] Ulrike Mückstein, Hakim Tafer, Jörg Hackermüller, Stephan H. Bernhart, Peter F. Stadler, and Ivo L. Hofacker. Thermodynamics of RNA-RNA binding. *Bioinformatics*, 22(10):1177, 2006.

[11] Martin Raden, Mostafa Mahmoud Mohamed, Syed Mohsin Ali, and Rolf Backofen. Interactive implementations of RNA structure and RNA-RNA interaction prediction approaches for example-driven teaching. *PLOS Comp Biol*, 14(8):e1006341, 2018.

[12] Aleksandar Spasic, Sarah M Assmann, Philip C Bevilacqua, and David H Mathews. Modeling RNA secondary structure folding ensembles using SHAPE mapping data. *Nucleic Acids Research*, 46(1):314–323, 2018.

[13] Kevin A Wilkinson, Edward J Merino, and Kevin M Weeks. Selective 2'-hydroxyl acylation analyzed by primer extension (shape): quantitative RNA structure analysis at single nucleotide resolution. *Nature protocols*, 1(3):1610, 2006.

[14] Patrick R. Wright, Martin Mann, and Rolf Backofen. Structure and interaction prediction in prokaryotic RNA biology. *Microbiol Spectrum*, 6(2), 2018.

[15] Chung-Shu Yeh, Shang-Lin Chang, Jui-Hui Chen, Hsuan-Kai Wang, Yue-Chang Chou, Chun-Hsiung Wang, Shih-Hsin Huang, Amy Larson, Jeffrey A Pleiss, Wei-Hau Chang, et al. 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, 2017.

Integration of accessibility data from structure probing into RNA-RNA interaction prediction

–

Supplementary material

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

[1]Bioinformatics Group, Department of Computer Science, University of Freiburg, Germany, and

[2]Center for Biological Signaling Studies (BIOSS), University of Freiburg, Germany

Contents

1 Integrating SHAPE data into accessibility-based RNA-RNA interaction prediction **2**

2 Spot probabilities of RNA-RNA interaction sites **3**

3 Sequence data and evaluation scripts **3**

4 DMS-seq reactivity data extraction **4**

5 U1 accessibilities with and without SHAPE reactivities **4**

6 Pre-mRNA U1 interaction predictions details **5**

1 Integrating SHAPE data into accessibility-based RNA-RNA interaction prediction

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. $I = \{\ (i,j) \mid i \in [1, |S^1|] \wedge j \in [1, |S^2|]\}$), 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.*, 2018; 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

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

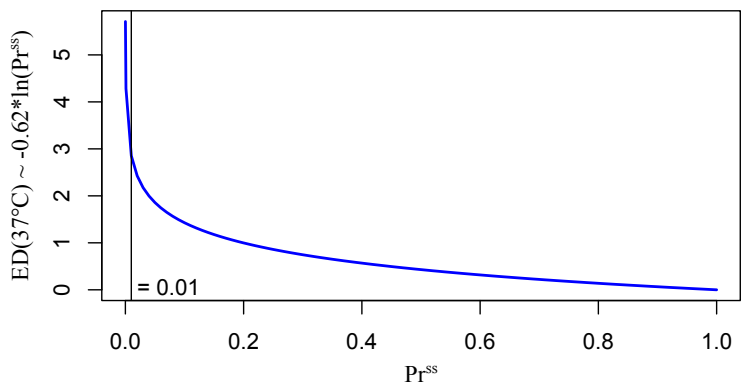
where R is the gas constant, T is the temperature, and Pr^{ss} denotes the unpaired probability of a given subsequence, which can be efficiently computed (Bernhart *et al.*, 2006).

As discussed above, SHAPE reactivity data can be incorporated into thermodynamic prediction tools via pseudo energy terms (Deigan *et al.*, 2009) as has been integrated in Vienna RNA package(Lorenz *et al.*, 2016). 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}_{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).

IntaRNA interfaces all three pseudo-energy conversion methods (Zarringhalam *et al.*, 2012; Deigan *et al.*, 2009; Washietl *et al.*, 2012) currently implemented within the Vienna RNA package. We observed best performance with Zarringhalam's method (data not shown), which we attribute to the incorporation of reactivity on both paired and unpaired terms and was used for this study

Note, since unpaired probabilities are incorporated on negated log-scale only (compare Eq. 2), small or intermediate changes of high probability values are

expected to show only minor effects in the RNA-RNA interaction prediction but should still guide or fine-tune the prediction, see Fig. 1. On the contrary, if the Pr^{ss}_{SHAPE} values deviate much (orders of magnitude) from Pr^{ss} (e.g. rendering presumably unpaired regions inaccessible since they might be already blocked by other substrates or vice versa as an indirect consequence e.g. of binding) or if the probability values are very small, strong prediction effects are expected.



Supplementary Figure 1: Relation of ED penalties and unpaired probabilities Pr^{ss} at temperature $T = 37°\text{C}$, i.e. $RT \sim 0.62$.

2 Spot probabilities of RNA-RNA interaction sites

To assess the effect of SHAPE data, we define the *spot probability* Pr^{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[1] $S^1_{l^1..r^1}$ and $S^2_{l^2..r^2}$ (see above).

3 Sequence data and evaluation scripts

All sequences used within this study as well as the used evaluation scripts are available online at

[1]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.

https://github.com/BackofenLab/IntaRNA-benchmark-SHAPE

while the integration of the IntaRNA approach is available in version $\geq$ 2.2.0 at

https://github.com/BackofenLab/IntaRNA

This study used IntaRNA 2.2.0, ViennaRNA v2.4.7 and pseudo energy computation for SHAPE data following Zarringhalam *et al.* (2012).

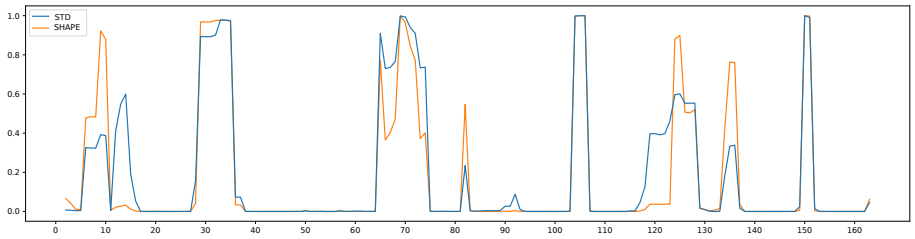
4 DMS-seq reactivity data extraction

We used the recommended settings from (Ding *et al.*, 2015) to map the sequencing reads and compute reactivities of the annotated transcripts using the Galaxy tools Afgan *et al.* (2018) Bowtie-2 and StructureFold (Langmead and Salzberg, 2012; Tang *et al.*, 2015). We selected the U1 homolog transcript bearing the largest secondary structure distance between the unconstrained and SHAPE-constrained structure prediction (using RNAfold).

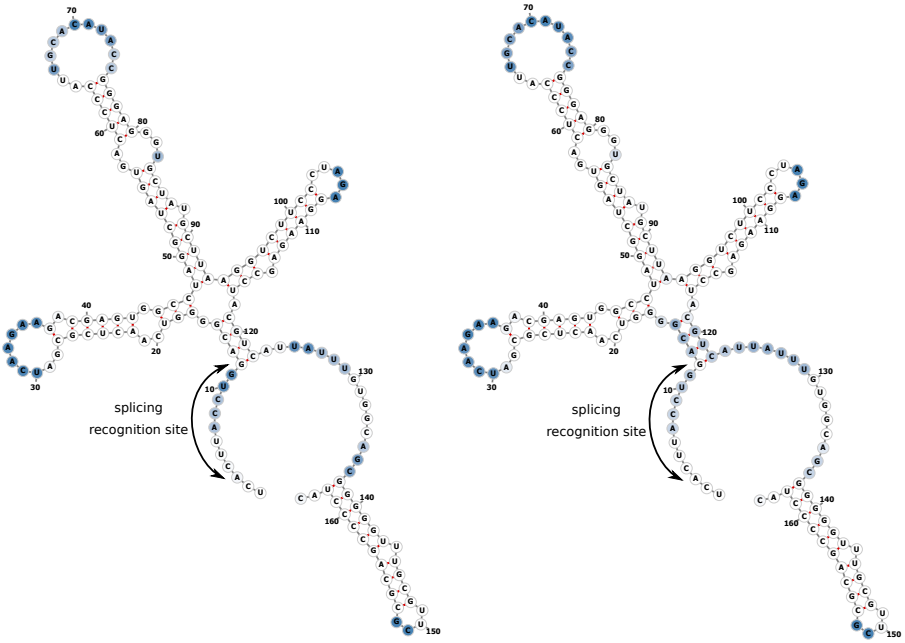
5 U1 accessibilities with and without SHAPE reactivities

Supplementary Figure 2 depicts the accessibility of the U1 RNA in terms of position-wise unpaired probabilities computed with SHAPE data and without. Without SHAPE data, the potential of the recognition site for inter-molecular interaction is underestimated, as can be seen when comparing to the SHAPE-guided unpaired probabilities. Furthermore, the subsequent region is wrongly assumed to be accessible (while known to be blocked by intra-molecular helix formation). The probabilities correspond to the 3-mer accessibilities (RNAplfold -u 3), such that for each position i the probability that 3-mer [i-2, i] is unpaired.

This can also be seen from the accessibility-annotated structure plots in Supplementary Fig. 3. The accessibility ("unpairedness") of the recognition site is much better reflected by SHAPE-guided terms compared to standard computations.



Supplementary Figure 2: Position-wise accessibility (y-axis) of U1 with SHAPE data (orange) and without (blue, STD).



(a) SHAPE (b) STD

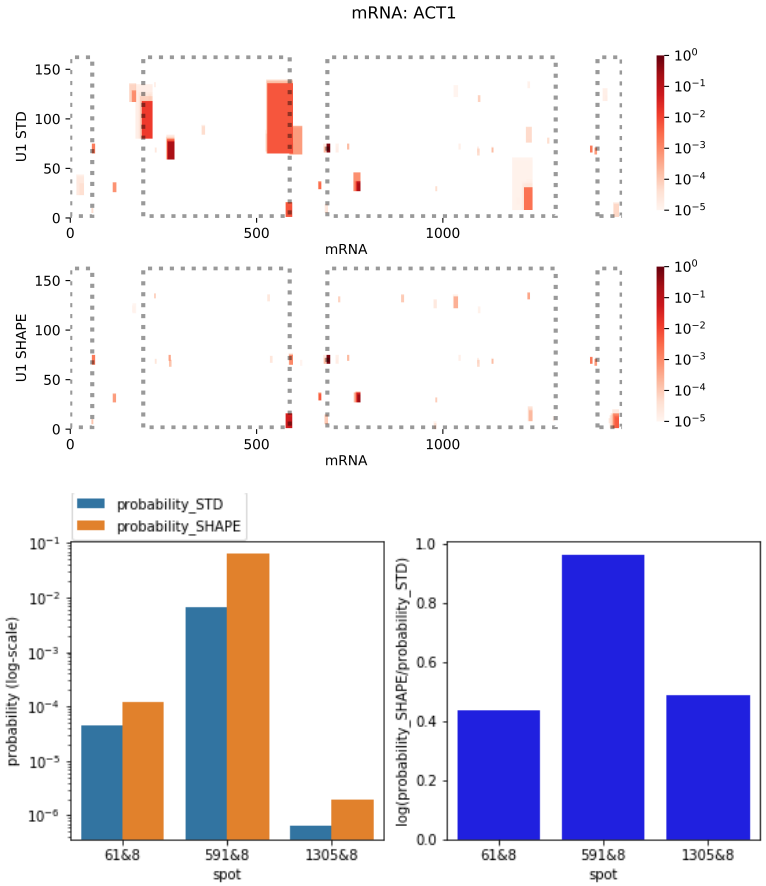
Supplementary Figure 3: Accessibilities mapped to the U1 secondary structure and color-coded, more accessible positions have darker colors. Visualization using Forna web-server (Kerpedjiev *et al.*, 2015)

6 Pre-mRNA U1 interaction predictions details

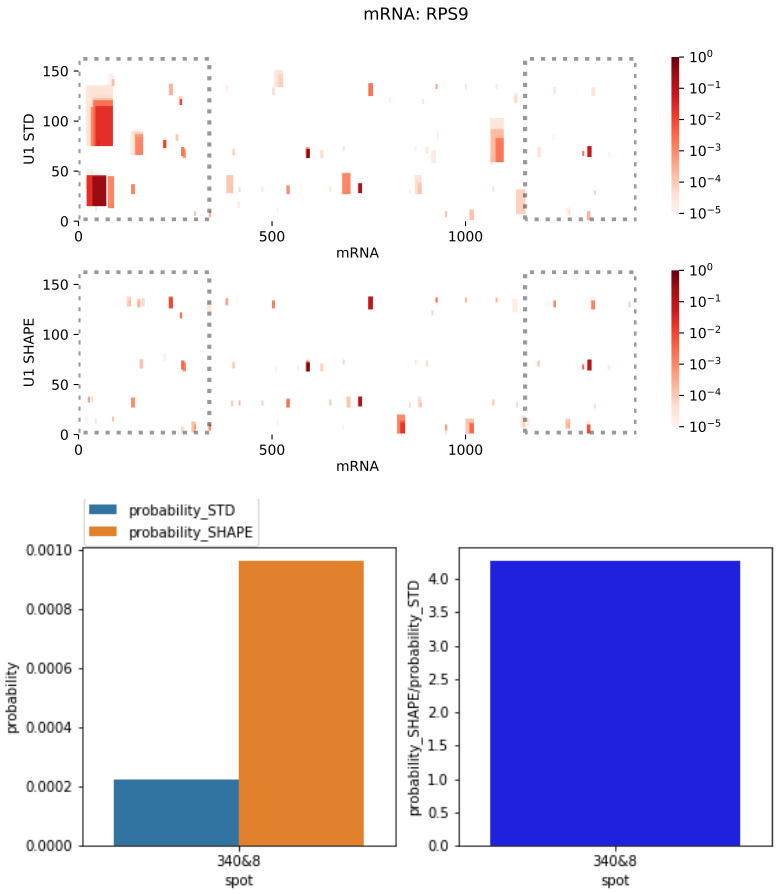
In the following, we provide the interaction prediction heatmap for all studied mRNAs. Index combinations that mark interactions of U1 with coding exons are enclosed by gray dotted boxes. The coloring represents the respective spot probabilities. Each heatmap is complemented with a visualization of the spot probabilities of U1's recognition site interacting with each CDS 5'-splice site using SHAPE data and without.

Predicted interaction sites are drawn in colored boxes where darker boxes relate to higher interaction (spot) probabilities. The x-axis corresponds to pre-mRNA indices while the y-axis represents positions of U1. For the latter, the U1 recognition site for intronic 5' splice sites is at position $\tilde{4}$-11. Thus, interactions of that site correspond to the bottom of the graph. The top graph represents interaction sites without SHAPE constraints while the bottom graph depicts the altered prediction when U1 SHAPE constraints are considered within IntaRNA's accessibility computation.

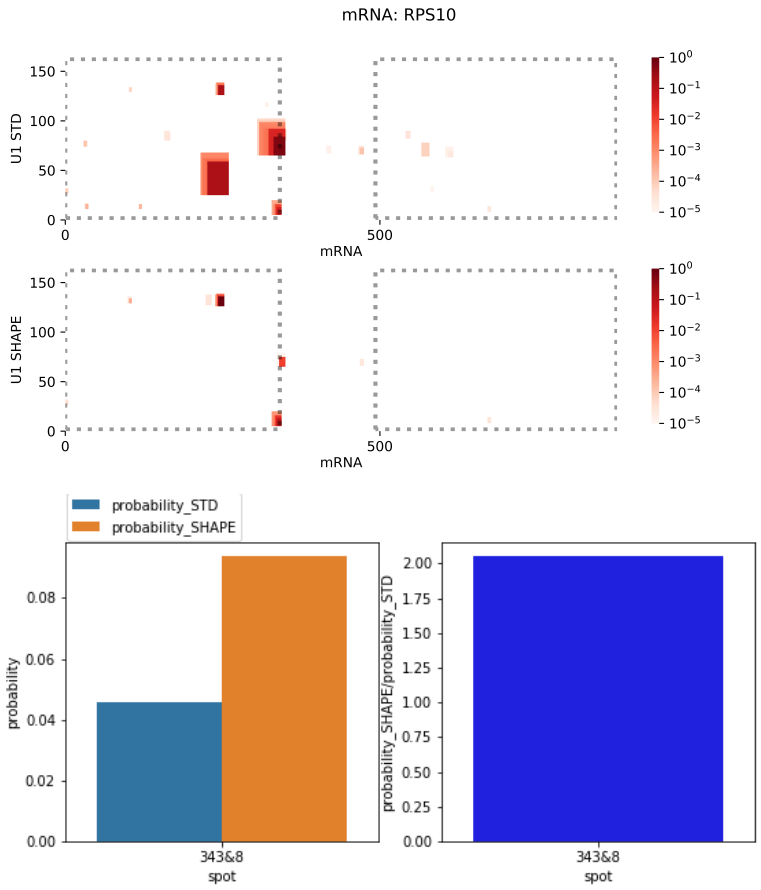
The left barplot shows the predicted interaction probabilities of all intronic 5' splice sites with the recognition site of U1 (called a spot). Blue bars represent the unconstrained probabilities while orange bars depict the probabilities when U1 SHAPE constraints are used. The right bar plot provides the ratio of both probabilities for each interaction site (spot). Please note, representations of multiple spots are in log-scaling to enable depiction.



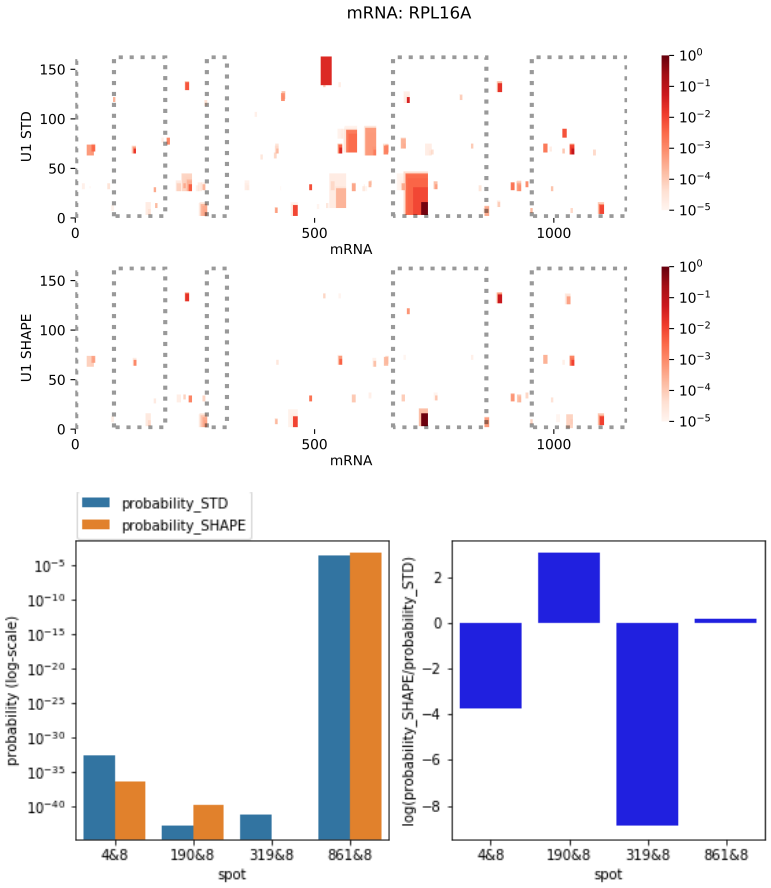
Supplementary Figure 4: U1-ACT1 interaction prediction with and without U1 probing data. Note, bar plots are in $\log_{10}$-scale.



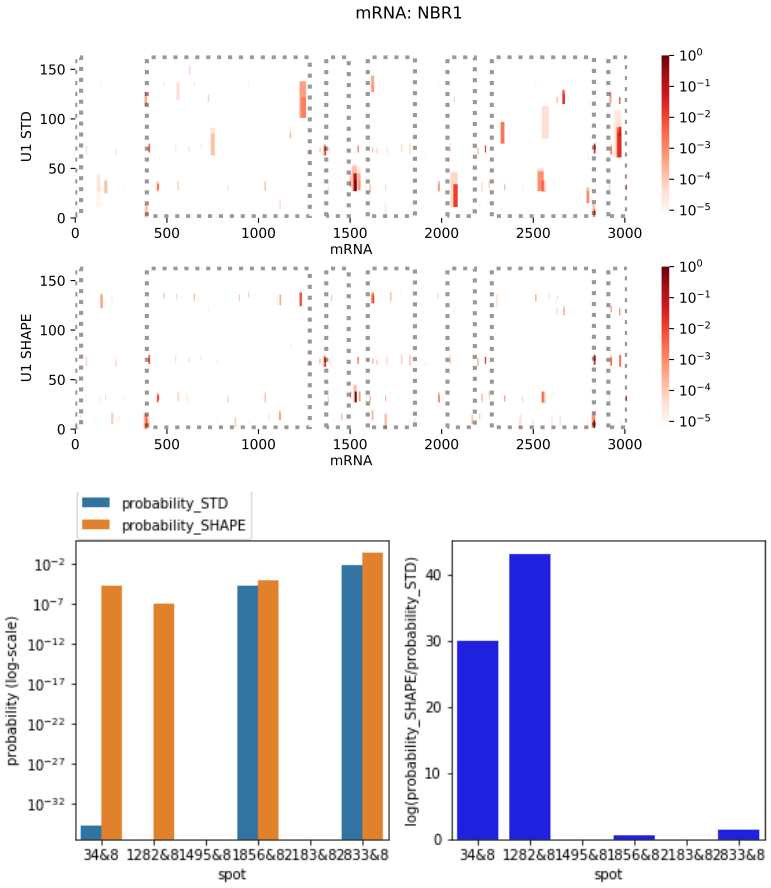
Supplementary Figure 5: U1-RPS9 interaction prediction with and without U1 probing data



Supplementary Figure 6: U1-RPS10 interaction prediction with and without U1 probing data



Supplementary Figure 7: U1-RPL16A interaction prediction with and without U1 probing data. Note, bar plots are in $\log_{10}$-scale.



Supplementary Figure 8: U1-NBR1 interaction prediction with and without U1 probing data. Note, bar plots are in $\log_{10}$-scale.

References

Afgan, E. *et al.* (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*, page gky379.

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

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.* (2015). Genome-wide profiling of in vivo RNA structure at single-nucleotide resolution using structure-seq. *Nature protocols*, **10**(7), 1050.

Grüning, B. *et al.* (2018). Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods*. online publication ahead of print.

Kerpedjiev, P. *et al.* (2015). Forna (force-directed rna): Simple and effective online RNA secondary structure diagrams. *Bioinformatics*, **31**(20), 3377–3379.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, **9**(4), 357.

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

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

Raden, M. *et al.* (2018). Interactive implementations of RNA structure and RNA-RNA interaction prediction approaches for example-driven teaching. *PLOS Comp Biol*, **14**(8), e1006341.

Tang, Y. *et al.* (2015). StructureFold: genome-wide RNA secondary structure mapping and reconstruction in vivo. *Bioinformatics*, **31**(16), 2668–2675.

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.

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

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